

PRACTITIONER'S TOOLKIT

BEING SMART WITH DATA, USING INNOVATIVE SOLUTIONS



***Europe Direct is a service to help you find answers
to your questions about the European Union.***

Freephone number (*):

00 800 6 7 8 9 10 11

(* The information given is free, as are most calls
(though some operators, phone boxes or hotels may charge you).

More information on the European Union is available on the internet (<http://europa.eu>).

Luxembourg: Publications Office of the European Union, 2017

ISBN 978-92-79-66962-0

doi:10.2767/797031

© European Union, 2017

Reproduction is authorised provided the source is acknowledged.

Cover picture: © European Union

The European Network of Public Employment Services was created following a Decision of the European Parliament and Council in June 2014 (DECISION No 573/2014/EU). Its objective is to reinforce PES capacity, effectiveness and efficiency. This activity has been developed within the work programme of the European PES Network. For further information: <http://ec.europa.eu/social/PESNetwork>.

This activity has received financial support from the European Union Programme for Employment and Social Innovation "EaSI" (2014-2020). For further information please consult: <http://ec.europa.eu/social/easi>

LEGAL NOTICE

This document has been prepared for the European Commission however it reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

PRACTITIONER'S TOOLKIT

**BEING SMART WITH DATA,
USING INNOVATIVE
SOLUTIONS**

Written by Dr. Willem Pieterse,
the Center for e-Government Studies,
in collaboration with ICF



MARCH 2017

Contents

| | |
|--|-----------|
| CHAPTER 1. INTRODUCTION | 6 |
| 1.1 What is this toolkit about? | 6 |
| 1.2 Who this toolkit is for | 8 |
| 1.2.1 PES with little or no experience working with data | 8 |
| 1.2.2 PES with some or intermediate levels of experience | 8 |
| 1.2.3 PES with advanced levels of experience | 8 |
| 1.2.4 Tactical or operational managers | 8 |
| 1.3 Strategic managers | 9 |
| 1.4 Scope of this toolkit | 9 |
| 1.5 Reading guide | 9 |
| CHAPTER 2. GETTING STARTED WITH DATA | 10 |
| 2.1 Creating a plan | 10 |
| 2.1.1 Deductive approaches | 10 |
| 2.1.2 Inductive approaches | 11 |
| 2.2 Creating your data team | 12 |
| 2.2.1 Leadership | 12 |
| 2.2.2 Data team members | 13 |
| 2.3 Setting up the data infrastructure | 14 |
| 2.4 Creating a data-catalogue | 15 |
| 2.5 Costs and budgeting | 16 |
| CHAPTER 3. ORGANISING DATA | 19 |
| 3.1 Cleaning & Sanitising | 19 |
| 3.2 Describing data & data characteristics | 20 |
| 3.3 Quality control | 21 |
| 3.4 Integrating data sources | 21 |
| 3.5 Security and Data Protection | 22 |
| CHAPTER 4. ANALYSING DATA | 25 |
| 4.1 Overview | 25 |
| 4.2 Statistics | 26 |
| 4.3 Data mining & KDD | 29 |

| | |
|--|-----------|
| 4.4 Advanced Analytics | 31 |
| 4.4.1 Artificial Intelligence | 31 |
| 4.4.2 Machine Learning | 33 |
| 4.4.3 Deep Learning | 34 |
| 4.5 Combinations & Derivations | 35 |
| CHAPTER 5. PRESENTING & REPORTING | 36 |
| 5.1 Why move away from traditional reports? | 36 |
| 5.2 (Interactive) Visualisations | 37 |
| 5.3 Interactive Tools & Dashboards | 38 |
| 5.4 Open data | 40 |
| CHAPTER 6. EVALUATION & CONTINUATION | 41 |
| 6.1 Evaluation | 41 |
| 6.2 Continuation and scale-up of pilots | 43 |
| APPENDICES | 45 |
| Appendix 1 Safe Harbor De-identification types | 45 |



Chapter 1.

Introduction

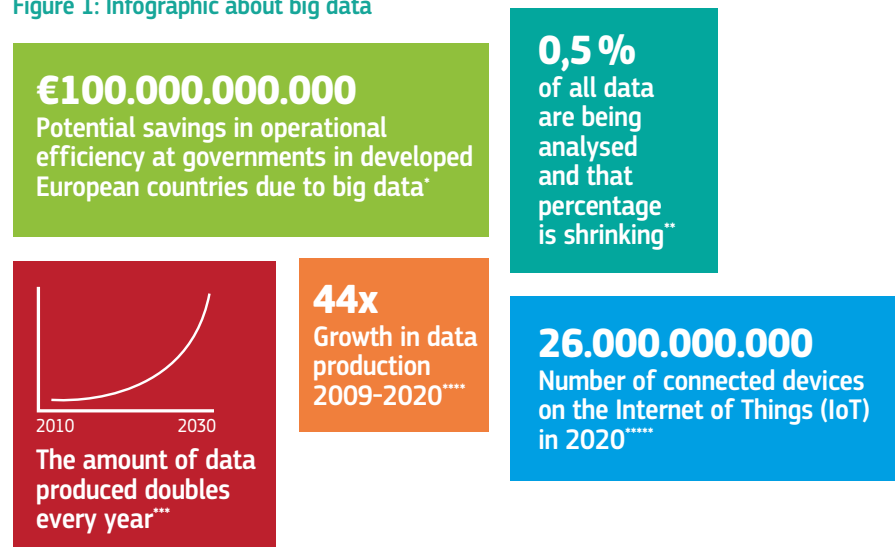


The role and importance of data in our society is growing fast. Not only do we collect more and more data, but advancements in computing power as well as the tools and algorithms to analyse data allow organisations to use data in entirely new ways. This also applies to Public Employment Services (PES). A few PES are exploring the use of Big Data to improve efficiency and effectiveness of processes, improve customer satisfaction and/or innovate in order to transform how the PES functions. This toolkit could help them on this journey. Most PES, however, are at the start of this journey and this toolkit can also help PES who want to start using their data in better and smarter ways.

1.1 What is this toolkit about?

This toolkit is about data and the use of data to create a better functioning PES. Among the chief reasons to have a toolkit about data is that PES, like most other organisations, are collecting more and more data. And as data is being stored in production systems and/or data warehouses, the possibility arises to use this data for all kinds of purposes. The infographic below illustrates a) the tremendous growth in (world-wide) data production, b) the potential (cost) benefits of using so-called 'big' data and c), the anticipated future growth of data production (e.g. due to the growth of the Internet of Things).

Figure 1: Infographic about big data



* http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode

** <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>

*** <https://www.technologyreview.com/business-report/big-data-gets-personal/>

**** <http://www.gartner.com/newsroom/id/2684616>

***** <http://www.emc.com/about/news/press/2011/20110628-01.htm>

With the expected future growth of data production, several challenges arise:

- ▶ How to unlock the potential of data we already have to improve efficiency, effectiveness, customer satisfaction or any other goal?
- ▶ How to set-up the infrastructure in the organisation now so that we gain experience with data analytics before we 'drown in a sea of data'?
- ▶ How to integrate data-analytics into the DNA of the organisation so that organisational decision making improves and organisational agility increases?

This toolkit aims to help organisations, specifically PES, to get started with finding answers to these questions.

Even though the topic of this toolkit is 'data', the true goal is to help PES transform their existing data into Information, Knowledge, and subsequently Wisdom. In each stage of transformation of data, the data is enriched, as the schema below illustrates.

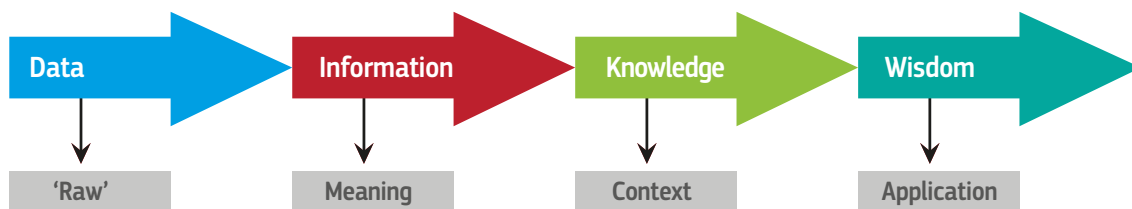


Figure 2: Data transformation process

| STEP | TRANSFORMATION | EXPLANATION | EXAMPLE(S) |
|-------------|--------------------|---|--|
| Data | None ('Raw' data) | This is plainly 'data', the numbers how you would extract those from any system. | 63 [just the plain number] |
| Information | Adding Meaning | When transforming data into information, we add basic meaning to the information. Very often this happens by adding units, variables and definitions. | 63% of lower educated clients are unhappy with PES service levels |
| Knowledge | Adding context | When adding context, we are able to make sense of the data, it starts telling a story. | 63% of lower educated clients are unhappy with PES service levels, compared to 48% of higher educated clients. |
| Wisdom | Adding application | Turning knowledge into action is the last step of the process. By combining data points, you can create actionable results. | 63% of lower educated clients are unhappy with PES service levels, compared to 48% of higher educated clients. This correlates with their evaluation of language difficulty on the PES website. ✓ Changing language level could solve this problem |



The toolkit is based on the thematic review workshop on ‘modernising PES through data and IT systems’¹. The workshop revealed a need to understand the topic of ‘data’ in more detail and explore how PES can benefit from advancements in this field. Core concept in this toolkit is the concept of ‘smart data’. By using the concept ‘smart data’, we want to prevent bias towards data being a normative goal in itself. Smart data is seen as the sum of:

- ▶ **(Big) Data** [the data itself]
- ▶ **Utility** [the potential utility derived from the data]
- ▶ **Semantics** [the semantic understanding of the data]
- ▶ **Data Quality** [the quality of the data collected]
- ▶ **Security** [the ways data are managed securely]
- ▶ **Data Protection** [how privacy and confidentiality are guarded]

These different topics are woven throughout the body of this toolkit. Similarly, to stay focused, we streamline this toolkit along the lines of the PDCA (Plan, Do, Check, Act) Cycle. The first content chapter [2] focuses on how to get started and create a plan. The following chapters [3-5] focus on the actual ‘doing’. While we focus throughout the toolkit and the proper checks and balances, chapter 6, is specifically devoted to the topic of checking and evaluating. While this is a practitioners’ toolkit, most content in this toolkit is actionable, but once again, specific action points after the analytical process are discussed in the final chapter.

1.2 Who this toolkit is for

The primary audience for this toolkit consists of PES who have little to no experience working with (big) data. As a practitioners’ toolkit, managers on tactical and operational levels may benefit the most from the content in this toolkit. For example managers who have been tasked with analytics or data science. However, there will be uses for other audiences as well. Below we lay out how the different audiences could benefit from this toolkit.

1 The Thematic Review Workshop took place in Croatia in 6-7 July 2016; it was developed under the PES Mutual Learning Programme.

1.2.1 PES with little or no experience working with data

This toolkit can serve as a starting point for those PES who want to get started with data. It gives an overview of relevant actions and provides practical tips on how to get started and where.

PES with little or no experience are advised to start reading at: – **Chapter 2 Getting started with data**

1.2.2 PES with some or intermediate levels of experience

For PES with some experience with data, especially the sections on advanced analytics and the various examples provided throughout this toolkit may be of interest. Furthermore, the sections on reporting and presenting data may provide new insights. Lastly, the parts on data security and protection could serve as a good refresh on matters related to security.

PES with some experience, are advised to start reading at: – **Chapter 3 Organising data**

1.2.3 PES with advanced levels of experience

PES that have much experience working with data may still benefit from this toolkit. On the one hand, the toolkit may provide some new insights, especially regarding more novel developments in the advanced analytics section. Furthermore, the toolkit can serve as an introduction for new employees who join data teams. Lastly, even advanced PES may find interesting examples from other PES throughout this toolkit.

The most relevant section for PES with advanced levels of experience will be section 4.4.

1.2.4 Tactical or operational managers

Tactical or operational managers tasked with setting up data functions and/or capabilities within PES may benefit the most from this toolkit. The entire toolkit should be relevant for this audience. The toolkit will help you get familiar with much of the jargon used in the world of data analytics and should give you enough guidance to get started. While this toolkit is meant to provide a generic introduction and practical tips and tricks, most sections will provide links to other resources that could help you further.

1.3 Strategic managers

Biggest value for strategic managers of this toolkit is twofold. The first is that it helps higher level executives in familiarising themselves with the possibilities of data analytics. The second is that the toolkit can support strategic managers in their decision making processes around data analytics, for example in terms of hiring a data team, and the goals where data analytics could make a difference.

Specific strategic points are added throughout this toolkit that are relevant for strategic managers (See pages 10, 11, 13, 15, 16, 20, 22, 23 and 31). Strategic manager will benefit the more from reading chapters 2 & 6.

1.4 Scope of this toolkit

This toolkit is about the use of data that PES collect in their systems or through other methods. Data currently stored in production systems or data warehouses are examples of this type of data. We also include data that are available in a PES that do not stem from primary processes, such as research data (e.g. data-sets from survey research), data shared from other organisations (e.g. educational data) or even reports, books, etc. In sum; we focus on all data the PES already has and less so on data the PES would need to collect to achieve certain goals. Conducting research (e.g. how to conduct surveys, interviews, focus-groups, etc.) is *not* part of this toolkit.

In terms of analytics; the focus is on the more 'novel' types of analytics and specifically those more commonly associated with big data. Where appropriate, we will discuss more 'traditional' types of analytics (such as more common statistical methods), and tools (such as Excel, SPSS, SAS, etc.), but given the abundance of (online) resources, we will link to those resources instead of providing more comprehensive information in this toolkit.

The same applies to the presentation of information. While traditional research reports (e.g. print, pdfs, etc.) are still widely used and can be good ways to present data and while (static) graphics created in – for example – spreadsheet software can provide compelling insights, we again choose to focus on the more novel solutions to present data. This includes tools to create interactive graphs and online dashboards.

1.5 Reading guide

The content of this toolkit is organised as follows:

1. Getting started with analytics. Creating a plan, as well as a data team. Setting up the data infrastructure and the creation of an inventory or data catalogue.
2. Organising data. Cleaning and describing data. Ensuring the quality of data, integration of data sources. How to security data and protect privacy and confidentiality.
3. The actual analysis of data. Different types of data analysis including statistical methods, data mining and advanced topics such as artificial intelligence and machine learning.
4. Presenting and reporting of data. What are novel ways to present results and what are considerations when reporting outcomes? Also discussion of open data.
5. Evaluation, continuation and implementation. How to go from small pilot or experiment to a broader implementation? What are the key technical and organisational considerations?

Chapter 2.

Getting started with data

This chapter deals with the question of how to get started. What are the main things to think about when an organisation wants to start doing more with the existing data? What kind of skills and knowledge would you need?

The following topics form the heart of this chapter:

- ▶ Provide overview of ‘things to think of when starting to work with data’
- ▶ Help create a team of people who can work with data
- ▶ Setting up the organisational infrastructure

The following strategic questions are answered in this chapter:

- ▶ What is the best way to get started and what can I expect from working with data?
- ▶ Whom should be in charge and where in the organisation should we position the data function?
- ▶ What can I do to ensure success?

The following tactical questions are answered in this chapter:

- ▶ What are the things we need to do actually start working with data?

2.1 Creating a plan

The most important question to answer when starting to work with data is: ‘why?’ Depending on the answer to that question the approach taken when implementing analytics and working with data will be different.

- ▶ Is there a concrete (organisational) problem that needs to be solved?

- ▶ Does the organisation have quantifiable goals and data available to track progress towards meeting those goals?
- ▶ Does the organisation simply want to learn about working with data so that experience is gained that could be useful for the future?
- ▶ Does the organisation want to innovate and see what improvements can be made from tinkering with (data) resources available?

When the ‘why’ question of the organisation looks like the former two questions, the organisation is better off starting with a *deductive approach* to the use of data. When it looks like latter two questions, an *inductive approach* may be better. The *utility* derived from each approach will be different, as well as the composition of the data team involved and the place of this *team* in the organisation.

Smart data = (Big) Data + *Utility* + Semantics + Data Quality + Security + Data Protection

Utility refers to the usefulness of smart data. The more it aids in solving business problems and/or the increase of knowledge in the organisation, the higher the utility is. However, depending on the approach taken when starting with analytics (deductive or inductive) expectations regarding utility can be different and should be managed as such.

2.1.1 Deductive approaches

In deductive approaches, the organisation has a (relatively) clear understanding of what it wants to do, for example what the problem is that needs to be solved.

For instance, a PES may find that a new automated vacancy matching system yields lower rates of successful matches that the previous manual matching

Figure 3: Deductive research approaches

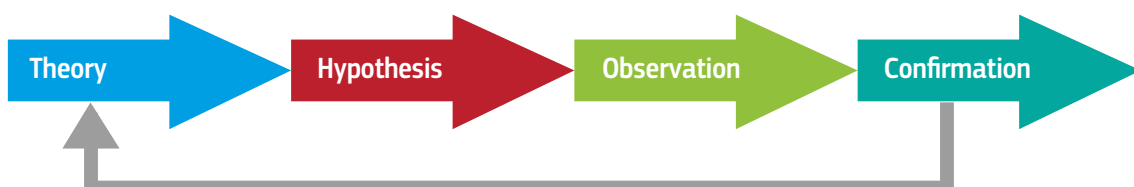
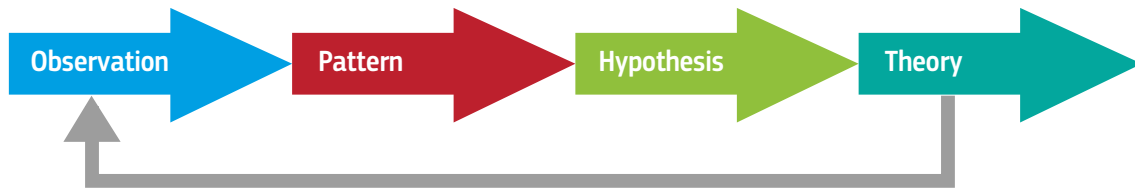


Figure 4: Inductive research approaches



process. In this case, the PES could formulate the expectation or *theory* that the new matching system is working improperly. This theory could be translated into a series of concrete expectations or *hypotheses* that could be tested using data (e.g. 'The matching algorithms do not include certain important variables' or 'the weight of certain variables in calculating matches is too high/low'). If such hypotheses exist, a *data team* could start working on testing these hypothesis and work closely with the business process owners in the PES to collect and analyse data.

Concept: Data team

In the context of this toolkit, we define the *data team* as the group of people within the PES tasked with data analytics. This could overlap (partially) with existing research teams (whose focus typically is more on *statistics* (see section 4.2) and/or *data mining* (see section 4.3)). Focus of the data team is on the analysis of (big) data that can be extracted from systems or imported from other sources.

In such a situation, the analytics function would be more embedded within the relevant business processes to ensure smooth communication to solve the problem at hand. Furthermore, higher levels of validity of and confidence in the results are required if outcomes are being used in production processes.

Concept: Data analytics²

The whole process of formulating data goals, collecting, organising, analysing and presenting (big) data.

² Some equate data analytics to data analysis, we decide to differentiate for the sake of clarity (and in line with many publications dealing with big data). We see 'analysis' as the process of analysing data, whereas we see analytics as a much broader term involving the setting of goals, collection, organisation, analysis and presentation of data.

2.1.2 Inductive approaches

The set-up of inductive approaches is different. There is no clear understanding of a problem and the focus is on:

1. Learning from the data and/or finding ways to generate value from this data.
2. Learning about analytics and gaining knowledge about the process and (smart) data applications.

Therefore, in inductive approaches there are fewer (if any) explicit expectations or hypothesis regarding outcomes. The focus is on letting the data speak and discovering interesting patterns in the data. For example, suppose a PES gets high levels of unstructured emails from clients (for example with questions, complains, comments), a data team could start analysing these emails to see if there are any interesting patterns in this data, for example regarding:

- ▶ The word choice of people (that could be used to change the tone of voice in communication or the use of certain synonyms in written communication).
- ▶ The questions people have (that could be used to change the way information is being displayed on websites).
- ▶ The language (skill) level of certain groups of job seekers (that could be used to create customer segments that could be targeted using different communication styles).

Only after certain patterns have been discovered in the data after analysing high numbers of observations, certain hypothesis and eventually theories could be formulated regarding the discovered patterns (which in subsequent rounds of deductive analysis could be further tested).

In such a situation, the analytics function would take more the shape of a *laboratory* or *experimental unit*. This means it operates more on a stand-alone basis, has more freedom to experiment and is less tied

Deductive vs Inductive

| | DEDUCTIVE | INDUCTIVE |
|------------------------------|---|--|
| Primary focus | Solving business problems | Learning and innovation |
| Management focus | Integration with the business, bridging parts of the organisation. Introducing the team to the organisation (so they understand the processes). | Shielding the data team from the business, to make sure they can focus on their work. Making sure data is accessible and the team works with the data. |
| Position in the organisation | Close to/part of business processes | Independent/removed from the business processes |
| Team values | Validity, robustness, value driven | Creativity, making mistakes, trial and error |
| Team composition | Focused, mostly on data engineers and scientists | Broad, including social scientists & people with creative profiles |

to specific business processes. The following table compares differences between the different approaches.

PES wanting to start with analytics are typically better of starting with inductive approaches. It typically allows for smaller scale experiments that allow both the data team, and the PES, to grow accustomed to working with data and slowly turn into a data driven organisation. Creating a small team that operates relatively independently from the organisation in order to prove value in the long term is a good starting point. Once the team has experience and shown value, the team could be brought into the organisation more and start shifting to more deductive approaches.

2.2 Creating your data team

Crucial to the success of working with data, is the composition of the data team. Relevant questions in this respect are; what is the approach we are taking (inductive, deductive, or a combination)? How many resources can we make available? What is the time pressure to deliver results? In this section we discuss team leadership (and variations therein) and several tiers of potential positions within the team.

2.2.1 Leadership

In smaller scale settings, the team will be smaller and team leadership will generally have a less senior position. In this case a senior data-scientist or manager of data analytics would be a fitting role to lead the team. As for the position in the organisation, a role under the following functional leadership role is possible:

- ▶ **CDO (Chief Data Officer) [or equivalent]**
In practice, only large (data) mature organisation will have a leading data position. The CDO is responsible for governance and utilisation of data across the entire organisation. This means that de CDO oversees all data initiatives and coordinates all analytics activities within the organisation.
- ▶ **CIO (Chief Innovation Officer) [or equivalent]**
The first interpretation of CIO is that of Innovation officer. This role is concerned with innovation and change management within the organisation. If the data team is positioned under the Innovation Officer, the focus of the data team will most likely be on more inductive approaches, trying to create innovative data-driven solutions.
- ▶ **CIO (Chief Information Officer) [or equivalent]**
The second interpretation is that of Information officer. This role is reserved for the highest ranking officer who is responsible for information technology and computer systems inside the organisation. If positioned under a CIO, the data team will probably be focused more on supporting the technology role in the organisation and hence have a more deductive orientation.
- ▶ **CTO (Chief Technology Officer) [or equivalent]**
The CTO is in charge of (the broader) technology used by the organisation, but could also focus on core technologies if technology is important in client facing processes (with the increasing levels of automation used in PES, that seems to apply here). If positioned under a CTO, the data team will probably focus on deductive, client oriented and technology related issues.

▶ **Director of R&D [or equivalent]**

The last role is that of the director of R&D. While there is some overlap with the Chief Innovation Officer's role, this CIO is typically occupied with more short term change management and the implementation of innovations. The director of R&D is typically charged with longer term research and development. If positioned here, the data team will be more experimental and focused on the development of longer term innovations.

Strategic insight

The purpose of this overview of roles is not to prescribe where the data team should be positioned. Rather, it is meant to raise the awareness that the position in the organisation will impact the expectations one should have of the team and what the focus of the team will have. This could impact hiring of team members as well.

2.2.2 Data team members

Depending on the (desired) size, workload and focus of the team, different types of team members are needed to start a successful data analytics practice. We divide these types in three tiers:

▶ **Key Roles**

These are the *must have* members as you start building your data team.

▶ **Secondary roles**

These are the *good to have* roles and will become more relevant as the team grows in size.

▶ **Tertiary roles**

These are the *nice to have* roles that add value to the team, but are less critical than the others. They will most likely become relevant once the team reaches high levels of maturity and has a large size.

We can define three key roles in working with the actual data:

▶ **Data engineers**

The data engineer typically sets up and works with the data infrastructure. That means that they set up databases and work on Extraction, Transformation and Loading (ETL, see below) tasks, they support data analysts and data scientists in their roles, and lastly they make

sure the system works smoothly and performs well. Very often they have a background in software engineering.

▶ **Data analysts**

Data Analysts are the professionals in their organization who query and process data. Furthermore, they typically create data reports and summaries, visualizations. This role is more closely associated with Data Mining and Statistical Analysis.

▶ **Data scientists**

The data scientist is the most important role in the context of this toolkit. Key role of the data scientist is to generate valuable and actionable insights from the data and help solve problems in the organization using data. Data scientists apply (mostly) advanced analytics, such as machine learning, to data.

Secondary roles:

▶ **Social scientists**

Team members with a social science background (e.g. Sociology, psychology, communication, marketing, etc.) can perform two important roles on the team. The first is to help create theories and hypotheses can be answered/ tested using deductive approaches. The second is to help make sense of the outcome of analysis when more inductive approaches are taken. In this way, social scientist hold a key role in translating organisational goals and/or problems into the actual 'data work' and subsequently translate the outcome of the 'data work' into implications and actions for the organisation.

▶ **Software engineers**

Even though there is a wealth of tools available, to organise, analyse and present data, very often an organisation will discover that the tools do not fit their needs entirely. Software engineers built custom service solutions (in conjunction with the rest of the team) that help maximize results across the organization. Difference with data engineers is that the software engineer in this context typically works on more front end (or customer) facing solutions (for example dashboards or mobile applications to access data and outcomes).

Some tertiary roles (that we won't discuss into detail):

▶ **Graphical and/or interface designers**

To help design useful and usable applications and visualisations.

- ▶ **Philosophers**
To help interpret data and help create theories and hypotheses.
- ▶ **Mathematicians & Statisticians**
To aid in the further development of complicated mathematical models and/or complicated statistical analyses.

2.3 Setting up the data infrastructure

Once a team is in place, it is necessary to create the data infrastructure on which the data team can work. The most important challenge here is to create (a) database(s) from which the data team can pull the data needed to perform their analyses. In computing, Extract, Transform, Load (ETL) refers to the process in database usage and especially in data warehousing of performing:

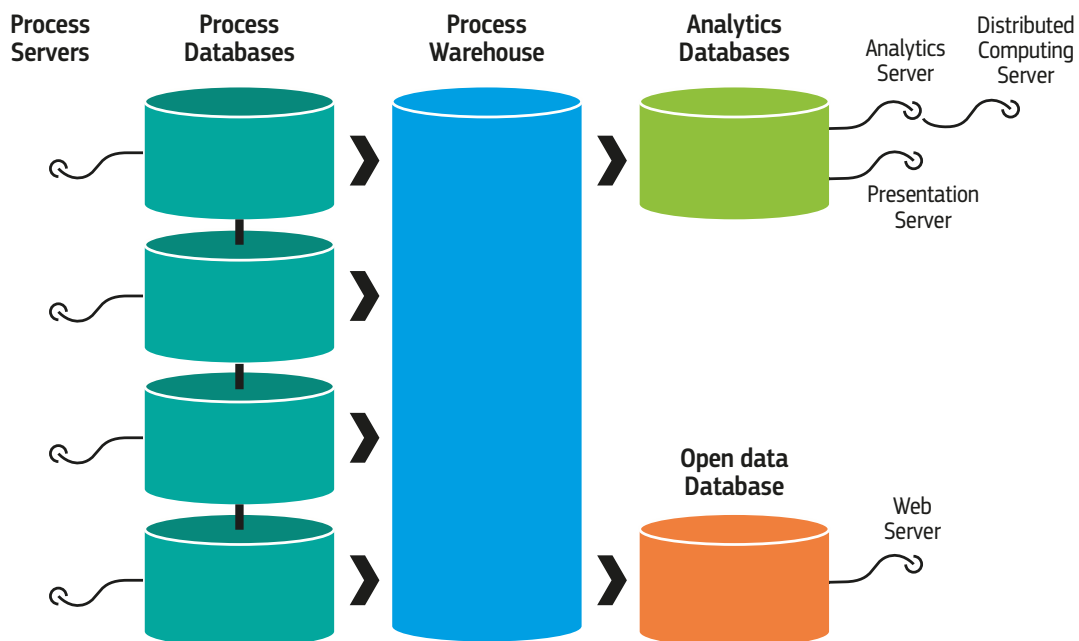
- ▶ **Extraction.** The process of extracting data from different data sources
- ▶ **Transformation.** Transformation the data for storing it in the proper format or structure for the purposes of querying and analysis
- ▶ **Loading** – loads it into the final target (database, more specifically, operational data store, data mart, or data warehouse, or (in this case) an analytics database from which analytics are performed).

Many configurations of ETL are possible and exist in practice. Most of these approaches have pros and cons:

- ▶ **Working from production systems/data bases**
In this setup, data engineers would collect data directly from production systems (e.g. ‘live’ profiling or matching systems). While this allows the team to work with the most current data. The downside is that it creates security risks and the risk of ‘messing with live data’. Furthermore, if data from multiple processes are needed, the team has to extract and transform data multiple times.
- ▶ **Working from a central data warehouse**
If the organisation has (one or more) central (virtual³) warehouse(s), it is not uncommon for data teams to work directly with the data from the warehouse. As the data in a warehouse typically is not used in production settings, production processes should not be impacted. However, this setup does create problems that are (especially) relevant in the government context. The first is that in many cases (such as legal archive reasons) a data warehouse serves as an important data store or archive.

3 It is fairly common to have a virtual database that uses wrappers to gather data from multiple different data sources.

Figure 5: Potential database architecture



Getting started

| PACKAGE/TOOL | EXPLANATION | LINK |
|----------------|--|---|
| Hadoop | Apache Hadoop is a software library that enables for the distributed processing of large data sets across clusters of computers using simple programming models. | http://hadoop.apache.org |
| Spark | Apache Spark is an engine for (distributed) large scale data processing. | http://spark.apache.org/ |
| Scriptella ETL | Scriptella ETL Scriptella is an open source ETL tool. | http://scriptella.org/ |

Working with data from the warehouse directly could pose a risk regarding the integrity of this archive. The second is that the warehouse may contain sensitive data or PII [see below] that should not be used for analytics purposes.

► **Creating a dedicated analytics database**

The third and last scenario is that of (a) dedicated database(s) solely for the purpose of analytics.

In this case, relevant data are being pulled from the warehouse (and other sources) and – after being sanitised – loaded into an analytics database. This database would then be the primary source for the data team to work from. This database could be supported by several types of servers to aid in the analytics:

► **Analytics Servers**

Servers whose sole purpose it is to run (computational) analytics. These servers typically have lots of processing power to run heavy and complicated analytics.

► **Distributed Computing Servers**

When datasets become too large, it may no longer be feasible to run on them on a single server. In this situation it is common to set up a distributed computing environment. What this entails is (conceptually) straightforward: a dataset is being broken down in smaller pieces, send to different servers (computers), analysed on these servers, and results of the analysis are bundled back into one single outcome. Many commercial distributed environments are available (such as Amazon Web Services (AWS), Google Compute Engine, and many others), but depending on the sensitivity of the data, a PES could consider setting up a distributed environment.

► **Presentation and visualisation servers**

These servers typically run applications to show results (such as dashboards) and/or the data catalogue.

Concept: PII

PII stands for Personal Identifiable Information. This is information that helps identify individual people. Classic examples of PII are names, addresses and unique person identifiers such as social security or 'citizen numbers'. One problem with PII and analytics is that, as more and more different types of data are being combined, it becomes increasingly possible to – indirectly – pinpoint individuals. It is important, especially as data are being opened (see section 5.4) to ensure that individuals cannot be identified.

2.4 Creating a data-catalogue

In order for any organisation to start answering question, it is important to know what information the organisation already has. This prevents the organisation from collecting the same information again and it yields an overview of data stored within the organisation that can be used for other purposes.

A data catalogue is an instrument that provides an overview of (all) data present in the organisation, it (semantically) explains the data and variables, and adds meta-data to these data sets. This means that the catalogue contains descriptions of the variables and nature of the data.

Concept: Data catalogue

A data catalogue is an overview of the existing data in databases and provides descriptions – using meta data – of the nature and state of the data, such as the base tables, volume, definitions, properties, synonyms, annotations and tags. As such, the data catalogue is an important part of the *semantics* part of the Smart Data equation.

Concept: Meta-data

Meta data is data that describes other data. For example, meta-data about a vacancy database could include descriptions of all variables stored in the database, the nature of these variables (e.g. Are they text, numbers (and what kind of number format), etc.) and such things as the number of records. In essence, the data-catalogue is an overview of all the meta-data.

**Smart data = (Big) Data + Utility
+ Semantics + Data Quality + Security
+ Data Protection**

Semantics refers to the meaning of data. Knowing exactly what data you have and what the data can tell you about reality is an important aspect of smart data. Semantics here does not just refer to descriptions of data and variables, but also to their meaning in real life. For example, does a measure or proxy of *customer satisfaction* really signify satisfaction of clients in real life?

While the primary function of a data catalogue is to document the data stored in the organisations databases, more types of data could be included in a data catalogue, such as:

- ▶ Research data (such as survey data sets).
- ▶ Externally available (analytics) data (such as what is collected through Google analytics or other trackers).
- ▶ Other relevant data (such as social media use, relevant (technical) documents).

A data catalogue can be compared to an index in a library. A library is a collection of books, written by different people on different subjects at different points in time of different lengths for different audiences. An index in the library contains the overview of exactly what you can find in the library. Furthermore, the index allows you to find the resources you need.

While a static data-catalogue is possible (literally a catalogue describing data), most organisations choose to create a database based data-catalogue with a searchable front-end (often using web technologies). While technically an open data catalogue, the catalogue at <http://catalog.data.gov/dataset> provides a good example of what a typically data catalogue looks like.

The process of creating a data catalogue is pretty straightforward. Members of the data team will have to document all available data in the organisation and use meta-data to describe this data. For example, they have to:

- ▶ Create an overview of all data sources, databases and datasets available and the nature of these data sets (e.g. what kind of databases do we have?).
- ▶ Document and describe all variables within these datasets.
- ▶ Document the number of records and changes in these records (e.g. How often are databases updated).

2.5 Costs and budgeting

We finish this chapter by focusing on the costs associated with starting an analytics practice. Budgeting for analytics activities proves to be a challenging task. A study by Gartner⁴ found that more than half of all analytics projects failed because they were not completed within budget or on schedule. Key reason for this is that it is very difficult to estimate the costs of early analytics projects, for example because:

- ▶ Early on, it is often unclear which information the organisation has and how useful this information is.
- ▶ Organising and cleaning data take up considerable amounts of time (and resources) and organisations tend to underestimate the amount of time it takes to get all data organised.

4 <http://www.gartner.com/newsroom/id/2637615>

- ▶ Especially with inductive projects, the expected outcomes are difficult to foresee, this creates many uncertainties regarding timeline and costs.

Nevertheless, it is possible to create a budget and allocate costs for execution of analytics. The cost of any analytics endeavour typically breaks down in the following:

- ▶ The set-up cost for the relevant infrastructure (databases, analytics servers, data extraction, transformation & loading, other hardware and software). The good news regarding these costs is that, while still significant, the cost of acquiring, storing and managing data keeps on going down. Especially using scalable solutions (see 2.3) it is possible to set up a relatively affordable infrastructure.
- ▶ The recurring/ongoing costs for the infrastructure (power, licensing fees, maintenance, etc.). Total sum of these costs depends on many factors such as the use of freely available software, or commercial software, the size of the server park, etc.
- ▶ Personnel costs. This entails the costs of the data-team such as hiring and salary costs. While analytics is driven by technology, human labour still makes up a large portion of the total cost of any analytics initiatives (in most cases the bulk). Reason for this is that organizing, sanitizing, cleaning and subsequently analysing and interpreting of data is a very labour intensive process (also see chapter 3). The more ambitious the project, the higher the labour costs will be.

However, the degree to which all costs occur depend on a number of factors, most notably the following choices will impact the cost:

- ▶ **In-house or outsourcing?**
The list of costs above assumes the organisation will want to own their own infrastructure and have all personnel on staff. There are, however a number of alternative scenarios:
- **Use of external or cloud infrastructure?**
Many organisations host their analytics servers elsewhere, often in the cloud. Benefits of using external providers for storage and/or analytics solutions is that a) you only pay for the capacity needed, b) it is easy to scale the capacity needed up or down and c) you don't have to worry (as much) about administration, maintenance, etc. However, there are a number of drawbacks, the first is a lower level of control over the infrastructure (which could for

example conflict with the needs of the organisation when certain requirements cannot be met [for example when a certain tool does not work on the platform provided]). The second is that certain legal requirements could prohibit the organisation from storing data outside of the organisation, the government and/or the country. It is wise to consider the specific legal requirements (see also section 3.5) before making this decision.

- **Internal data team or external service provider?**
Organisations that want to engage in analytics activities face an important (cost related) decision when it comes to the personnel aspect; hire a data-team or outsource the personnel aspect. Several (consulting) providers provide analytics services that could take care of the personnel requirement (and often the infrastructure as well). Furthermore, the organisation could consider using freelance personnel or virtual (job) marketplaces, but given the sensitivity of the data often involved these options are often not realistic. The benefits of using an external provider are that a) it could be cost effective if analytics are only needed on a project basis (and not continuously), b) you don't have to worry about training or capacity. Downsides are that a) the organisation builds little experience and knowledge about analytics internally, b) costs can be (much) higher in the long run. So if the organisation plans to seriously build analytics capabilities, creating the capacity internally seems the right course of action.
- **Hybrid or not?**
Lastly, the organisation could choose a mix of doing part of the infrastructure internally (such as data storage) and part externally (such as running certain analysis on an external server provided by a third party) and hire part of the data team and use an external provider for specific expertise areas. In practice, most advanced organisations use some kind of hybrid version. Reasons for this are that a) analytics needs can become very specific (or advanced) and using an external provider is more cost-effective than building the capability internally for just one specific type of analysis b) depending on the nature and scope of the analytics activities, the organisation may only need a certain amount of capacity to 'run the normal business' but during peak times or for specific activities, extra capacity could be needed in which case partnering with a third party is reasonable.

Furthermore, the following are relevant considerations when it comes to costs:

- ▶ As mentioned above, many projects overrun their budgets. This means that there is a tendency to underestimate costs. For this reason it is advisable to either lower project requirements or not be overly optimistic when starting with analytics.
- ▶ Costs of analytics do not scale linearly. Not only is there typically a large start-up cost involved, but as analytics activities become more

advanced, so does the complexity of the work. Especially when a multitude of data sources are being used and models become more complicated, costs can rise at a higher than linear rate.

- ▶ The best way to keep costs manageable is to start small. Smaller datasets can be analysed on relatively cheap hardware and smaller teams are needed for smaller projects. In that sense, if budgets are non-existent, or small, it is advisable to start a smaller analytics practice and grow this over time as the team proves its value.



PRACTICAL EXAMPLE

In May 2016, the Executive Office of the United States President released a report on the opportunities of Big Data. The report contains a case study on the potential of Big Data for employment. As key problem in employment related to data, the report recognises that 'traditional hiring practices may unnecessarily filter out applicants whose skills match the job opening'. To solve this problem, big data is seen as an opportunity: 'Big data can be used to uncover or possibly reduce employment discrimination'. For example, big data analytics can be used:

- ▶ To prevent 'affinity bias' or 'like me bias' in the hiring process (for example where hiring managers tend to select candidates like them or whom they like).
- ▶ To find potential job-candidates who otherwise might have been overlooked based on the more traditional educational or workplace-experience related job-requirements. For example, by looking at the skills and knowledge areas that have made other employees successful, a matching system could use 'pattern matching' to recognize the characteristics that made current employees successful and thus need to be looked for in future employees.
- ▶ Large data analytics systems could help prevent biases often seen in traditional hiring practices that could lead to discrimination. An algorithm could be designed to not look at factors like age, gender, race or any factor whereas it is much more difficult to block such (implicit) factors as a human.
- ▶ Beyond supporting or recommending matching/hiring decisions, advanced algorithms create the possibility of solving long-term employment challenges related to discrimination, such as the wage gap or occupational segregation, for example by going beyond formal job qualifications, but finding the person for the job based on cultural or other factors.
- ▶ Using data-analytics new kinds of 'candidate scores' or matching scores can be created by using diverse and new sources of information on job candidates. The report mentions how one employment research firm found that distance employees commute to work to be one of the strongest predictors of how long customer service employees will stay with their jobs. Such variables and data could be used to improve matching algorithms for specific (customer service) job vacancies.
- ▶ Finally, machine learning based algorithms could help decide what kinds of employees are likely to be successful by reviewing the past performance of existing employees of certain companies or job seekers who worked for certain firms or by analysing the preferences of hiring managers as shown by their past decisions. This could also apply to such things as employee turnover and the likelihood that certain people will retain jobs in certain industries.



Chapter 3.

Organising data

In this chapter we focus on the organisation of the data. We discuss how data can be cleaned and sanitised and how existing data from within and outside of the organisation can be extracted, transformed and loaded into a database that can be used for analytics.

The following topics form the main content of this chapter:

- ▶ Provide overview of what organising data entails.
- ▶ Focus on areas such as security, data protection and privacy.

Strategic questions this chapter answers:

- ▶ How can we guarantee security?

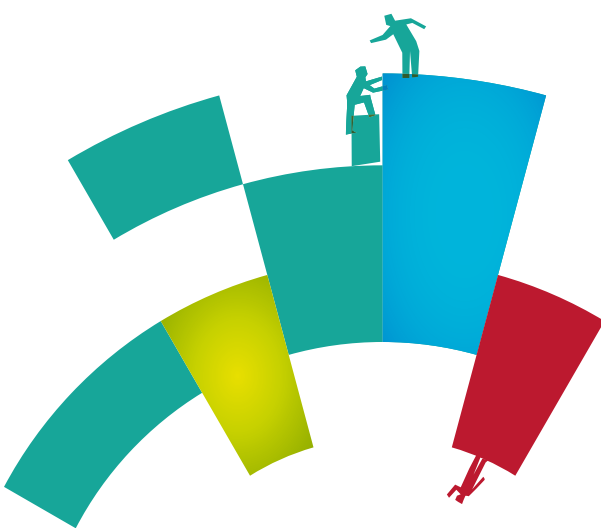
Tactical questions this chapter answers:

- ▶ How to organise and protect data?
- ▶ What are reasonable expectations regarding data organising.

3.1 Cleaning & Sanitising

Organising data is a complicated process that requires a lot of time and devotion. Data scientists can spend up to 90% of their time organising data. This has several reasons:

1. The data are in an undesired format. For example, databases can store times and dates using many different formats (e.g. think about the European DD/MM/YY format versus the US MM/DD/YY format). Importing different data sources in one database can create inconsistencies in certain variables that need to be fixed. Furthermore, data might have to be reorganised so that analytics can be run.
2. The data need to be cleaned. Many databases contain noise that could be relevant for one process, but irrelevant for analytical processes. Cleaning the analytics database helps reduce *noise* (and could aid in database performance). In a data-catalogue, certain variables that are less relevant for analytics could be marked and deleted when creating a dedicated analytics database.
3. The data need to be sanitised. This refers to the process of ridding the database from unwanted information. For example, personal identifiable information (PII) (see section 2.3) needs to be removed from the datasets. When variables are documented properly, sanitisation can be done using scripts. When that is not the case, the data-team will have to check the data manually and sanitise when needed.



Getting started

The following might be useful to get started on organising and cleaning data:

| PACKAGE/TOOL | EXPLANATION | LINK |
|----------------------------------|---|---|
| Encog Machine Learning Framework | Encog is an advanced machine learning framework that supports a variety of advanced algorithms, as well as support classes to normalize and process data. | http://www.heatonresearch.com/encog/ |

4. The data need to be transformed. Even though transformation is part of the ETL process (see section 2.3). Additional transformation can be needed while preparing for data analytics. For example, decimal places could have to be fixed or floating numbers could have to be converted to integers. Another example is to transform unstructured data into structured data.

Concept: Structured vs. Unstructured data

Structured data is data with a high level of organisation (and formatting). For example, a table with records, variables and labelled data (such as a table with jobseekers' demographic information) is structured information. In its most simple form, unstructured data is data lacking such organisation. A database with PDFs or photocopies of jobseekers' resumes is an example of unstructured data.

5. Missing data might have to be imputed. Imputation is the process of substituting missing data with calculated values. In databases where data is missing, scientists may choose to run algorithms to *impute* these missing data points. One way to do this is by looking at patterns in the data. For example, if people with similar characteristics consistently have the same value regarding a certain variable, the likelihood increases that the missing values are similar.

It is extremely important that data are cleaned and sanitised properly. For this reason, PES starting to work with data could create peer review processes to make sure data are being reviewed after being organised by other team members.

**Smart data = (Big) Data + Utility
+ Semantics + Data Quality + Security
+ Data Protection**

Data Quality refers to various aspects of the data itself. These are a) completeness (do I have enough data about everybody to make claims), b) cleanness (are the data well cleaned, sanitised and maintained), c) have high levels of validity and d) impact (have they been analysed in such a way that they retain their validity and create relevant meaning for the organisation).

3.2 Describing data & data characteristics

Descriptive analytics help us understand the data. When 'describing' data, we look at how the data is distributed (e.g. normal, power-law, linear), what the key characteristics of the data are (e.g. the mode, median and mean) and we check for such things as the outliers in the data. Descriptives are important because they:

- ▶ Help us understand the nature of the data (e.g. what is the nature of the variables).
- ▶ Help us draw initial conclusions about the data (e.g. Based on the distribution of data we can observe that certain jobseekers have certain characteristics).
- ▶ Help us prepare for further analysis (e.g. by removing outliers that distort the data).

Furthermore, descriptive statistics are another way to check the quality of the data and the organisation of the data. For example, it could show inconsistencies in the data collected and areas where data have been improperly transformed.

Getting started

The following methods, tools and/or applications can be used to calculate descriptive analytics:

- ▶ Most spreadsheet software (e.g. Microsoft Excel, LibreOffice Calc) include basic functions for most descriptive analytics
- ▶ Most statistical software (e.g. SPSS, SAS) have dedicated functions for descriptive analysis
- ▶ Modelling languages (e.g. R & Python) have packages for descriptive analysis.

3.3 Quality control

High quality data is an important element of *smart data*. Ensuring quality and having the proper checks and balances in place can help a PES being confident in the quality of the available data. The most important way to control quality is to have control mechanisms during every single step of the data process. This means that during collection, organisation, analysis, presentation, and evaluation activities have to be deployed to check for the quality of the data.

These are ways to control the quality of the data during various stages of the process. Examples include:

When preparing for data collection:

- ▶ Checking the quality of theories or hypotheses using expert evaluations or literature reviews.

When collecting data:

- ▶ Taking multiple measurements, observations or samples
- ▶ Using standardised methods, instruments and protocols when collecting data (which could be included in the data catalogue).

When organising data:

- ▶ Setting up validation rules when entering data
- ▶ Using strict protocols when cleaning data
- ▶ Documenting database structures properly (and reviewing this)
- ▶ Creation of automated scripts for organisation tasks (and checking their performance using reviews)
- ▶ Having peer reviews of organisation methods and practices.

When analysing data and presenting results:

- ▶ Model testing using multiple (independent) samples

- ▶ Code and model reviews either by team members or external experts
- ▶ Combining model fit measures with theoretical evaluations
- ▶ Automated testing of code before models are being executed or any code is being published or pushed to a production environment.

3.4 Integrating data sources

Focus of this section is not on the integration of data in business processes. Rather, the focus is on the combination of different types of data from within and between organisations.

Data integration is the process of combining data housed in different sources and giving users a unified view of the data. While in many organisations, a data warehouse will already have integrated the most important data sources, there are many scenarios in which data will have to be integrated or combined to create a unified view necessary for analytics, such as:

▶ Missing data

Imagine job seekers registering online, but not completing all information during the registration process (e.g. their educational background). These data are missing in the database and this could be a problem if we want to do education related analysis. If, during the registration process, the job seeker has also uploaded a resume, we may be able to extract the data from the resume and integrate the two data sources in order to create a complete record.

▶ Triangulation

Triangulation refers to the combination of different data sources to check for the quality of one (or both). For example, data scientists could use predictive modelling to predict satisfaction of jobseekers with a matching tool and use survey data to check the results of the model.

▶ Increased quantity

The most important reason, however, to integrate data sources is to have more data available. For example, data on coaching or training outcomes could be used to enhance the profile of a job seeker.

While the actual implementation of data integration is a technical question, the main challenge of data integration is organisational in nature. It involves such activities as:

- ▶ Convincing different stakeholders of opening up ‘their’ data silos.
- ▶ Getting cooperation from IT departments to actually get the data.
- ▶ Coordinate privacy and security risks with relevant stakeholders.
- ▶ Working with relevant stakeholders to understand the nature of the data (i.e. Add to the data-catalogue).
- ▶ Making agreements around updates of the data or SLAs (i.e. how often and to what standards data are being shared).

Concept: Service Level Agreement (SLA)

An agreement specifying the quality of services delivered from one (part of an) organisation to another. For example about the uptime of servers or the refresh rate of data.

The stronger the mandate of the leader of the data team and the higher the position in the organisation, the more (formal) organisational power the team has in getting the data it needs. This is an important consideration when starting with data analytics. A more experimental team focused on inductive approaches, may have more difficulty in integrating data sources from across the organisation versus a team with a position more closely tied to business processes.

This applies even more strongly to data integration across organisations. If the success of the data team depends on data from other organisations, it will need a ‘stronger’ position in the organisation, preferably with support from the highest levels of leadership in the organisation. Logical data sharing partners for PES include:

Other governments:

- ▶ **Ministries of education (or similar)**
For example regarding data about the future workforce, which could be helpful in predictive models for future matching applications or unemployment forecasting.
- ▶ **Tax agencies (or similar)**
Regarding financial data about job seekers (e.g. For benefit fraud detection).
- ▶ **Social security institutions (or similar)**
Regarding social security or benefit information. This is especially relevant when there is a legal obligation for data collection and/or sharing.
- ▶ **Statistics bureaus**
For various types of information such as population mobility (which could be used to fine-tune job recommendations) or household developments (which could impact the labour force).
- ▶ **Regional or local governments**
For data regarding specific local or regional circumstances (e.g. Local employment initiatives).

Businesses:

- ▶ For example regarding job developments (are business going to add or remove positions?), their future needs.

Other organisations:

- ▶ Such as foundations working in the labour market (for example, overviews of activities could help in interpreting specific labour market fluctuations).

3.5 Security and Data Protection

Security and Data Protection are two other key elements from the Smart Data equation. Protecting (user) data and having good security should be among the highest priorities of both the data team, as well as the leadership of the team and the parts of the organisation involved in that data used by the team. Several types of security are important:



PRACTICAL EXAMPLE

The X-Road is Estonia’s infrastructure that connects databases from a multitude of governmental agencies. It is best described as a distributed service bus which allows databases to interact, making integrated e-services possible. This, however, also creates opportunities for data integration that can be used for data analytics.

Currently, 219 databases are connected to X-Road and these result in over 1 700 services being offered. By integrating data sources, citizens only have to supply many pieces of information once and it already allows fraud prevention through analytics.

Please, find further information in the following [link](#).

Physical security

Where are the data stored? How easily can these machines be accessed? Storing data in safe locations where only authorised personnel has access is one of the key steps to ensure. The following can help improving physical security:

- ▶ Have strict access controls to rooms containing data servers
- ▶ Have protocols for the use of (sensitive) data on desktops, laptops and other devices
- ▶ Have guidelines for the use of mobile devices (e.g. Laptops, tables, phones) outside of secure areas
- ▶ Have protocols for the use of removable media (e.g. USB drives, removable hard drives, etc.).

Virtual access security

Once being close to a machine (or accessing it remotely), how easy is it to gain access? Are (safe) passwords in place? Is data encrypted? The following can help in improving virtual access security:

- ▶ Develop guidelines for the encryption of data (especially on those devices with easier physical access)
- ▶ Develop policies for the use of firewalls and ant-virus software
- ▶ Have strict protocols regarding passwords, password sharing and password changes
- ▶ Limit the use of APIs (and other ways to access data) to non-sensitive data and/or open data.

**Smart data = (Big) Data + Utility
+ Semantics + Data Quality + Security
+ Data Protection**

Security refers to the ways the data are being securely stored and managed. This applies not only to the *physical* security (who can access servers and related systems?) but also *virtual* security (who has access to data secured on systems?). Security is important to make sure systems are being hacked and/or data does not 'leak' or is being stolen.

In addition, the following can help with security in general:

- ▶ Makes sure software is always up to date
- ▶ Having a regular security meeting to discuss and refresh members of the data-team's memories on security related matters
- ▶ Make security training a standard part of (new) data-team members, so that a culture of security awareness is instilled from the start.

The second topic we are discussing in this toolkit is that of data protection.

**Smart data = (Big) Data + Utility
+ Semantics + Data Quality + Security
+ Data Protection**

Data protection refers to the ways privacy and confidentiality are safeguarded. Are PII replaced by other identifiers? Can data be traced (in combination or not) to individuals? Are users aware of the use of their data? Protecting data properly will protect (vulnerable) individuals and minimise organisational risks.

For purposes of this toolkit, we break down data protection in two topics; privacy and confidentiality. While privacy applies to the person that needs to be protected, confidentiality applies to the person's data. When data can be used to identify a person, *privacy* issues may arise. When data about a person can be used maliciously (for example by judging a person using data that was supposed to be collected anonymously), *confidentiality* issues can arise.

Guarding the privacy of individuals and the confidentiality of their data is important to ensure no harm is done to any individual or organisation. Main consideration regarding privacy consists of the applicable laws and regulations. On an EU level, the following are important:

- ▶ **Directive 95/46/EC | Protection of personal data.**
Directive 95/46/EC sets up a regulatory framework which seeks to strike a balance between a high level of protection for the privacy of individuals and the free movement of personal data within the European Union (EU). To do so, the Directive sets strict limits on the collection and use of personal data and demands that each Member State set up an independent national body responsible for the supervision of any activity linked to the processing of personal data [quoted from URL below]
URL: <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV%3A114012>
- ▶ **Regulation (EU) 2016/679 | General Data Protection Regulation**
This Regulation is set to replace directive 95/46/EC. It was adopted on 27 April 2016 and enters into application on 25 May 2018.
More info: http://ec.europa.eu/justice/data-protection/index_en.htm

Besides applicable European regulations, every single member states will have their own applicable laws and regulations. These should be consulted before starting analytics projects. When data across multiple countries are being collected, laws from multiple countries may apply. Special care should also be given to the (cloud) storage of data in countries other than the home country and/or outside of the EU.

Next to abiding by the law, the following good practices can help to establish good data practices and ensure privacy protection:

▶ **Implement solid de-identification protocols and capabilities.**

This consists of the removal (or replacement) of PII (see section 2.3), as well as having checks and balances in place that ensure that no PII enters the process at any time and/or individuals can be identified using analytics (e.g. By combining a multitude of variables, it could be possible to narrow data sets down to individuals). The U.S. Department of Health and Human Services (HHS) describes two (commonly accepted) methods to de-identify information⁵:

• **Expert Determination method:**

In this scenario qualified experts apply statistical or scientific principles to render information to be not individually identifiable.

• **Safe Harbor method:**

The removal of 18 types of data related to individuals from the dataset completely (see Appendix 1) for an overview.

▶ **Always use privacy impact assessments (PIA).**

PIAs are tools used to identify and mitigate privacy risks. Section 3, article 33 of the new EU General Data Protection Regulation [Data protection impact assessment and prior authorisation] already stipulates ‘controllers and processors to carry out a data protection impact assessment prior to risky processing operations’. However, a good practice could be to assess the privacy impact for *any* project related to individual people and/or cases. At the very least, such a PIA should:

- Assess whether the information used complies with all (privacy-related) legal and regulatory requirements.
- Make an inventory of potential risks of working with PII.
- Assess processes for handling information to reduce or mitigate potential privacy risks.
- Investigate the consent methods (see below) used to ask individuals’ permission for the use of their data.
- Record the outcomes of the assessment and make them available.
- Implement solutions for any risks or problems discovered.

▶ **Implement ‘privacy by design’ principles (also required according to article 23 of the new EU General Data Protection Regulation).**

This means that any project, process or service needs to be designed from the start to adhere to the strictest possible privacy considerations. For example, when creating an analytics database, PII should never be included in such a database in the first place. This prevents the data team to work with PII in the first place.

To safeguard confidentiality, two actions are important

- ▶ Use consent procedures when collecting information. For example ask people for consent to use their data when they register as unemployed, or when they fill out surveys.
- ▶ Actively inform individuals of the purposes for which their data is being used (very often this happens in conjunction with the consent procedure).

⁵ See <http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>

Chapter 4.

Analysing data

In the fourth chapter we discuss the analyses of the data. We discuss more traditional types of analytics (such as statistical methods and data mining), but also discuss novel and innovative types of analytics such as machine learning and artificial intelligence. The focus of these innovative types is not necessarily on how PES are using these types, but on potential use cases for the future.

The following topics form the main content of this chapter:

- ▶ Provide overview of analytical techniques and tools
- ▶ Provide guidance on what goals can be achieved using what methods.

These strategic question are being answered in this chapter:

- ▶ How can analytics help achieving specific strategic organisational goals?

The following tactical questions are central in this chapter:

- ▶ What are types of analytics that are available and how can they help PES?
- ▶ Which analytics to choose and what are their pros/cons?

4.1 Overview

Before we start our overview of techniques to work with data, we first give an overview of some common analytical approaches and their differences/overlap⁶. As the graph makes clear, there is an abundance of methods, tools and approaches available to transform data into value. The specific use case of each approach depends on the goal (see Chapter 2) the PES wants to achieve.

⁶ Goal is not to be complete in this overview, but to provide an overview of relevant approaches and to show their relations.

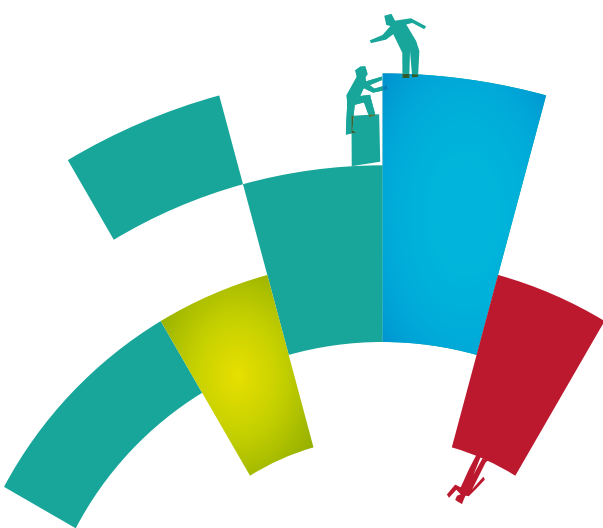
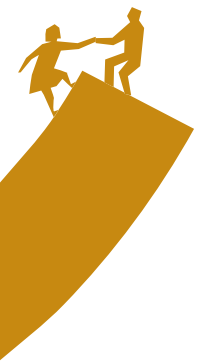
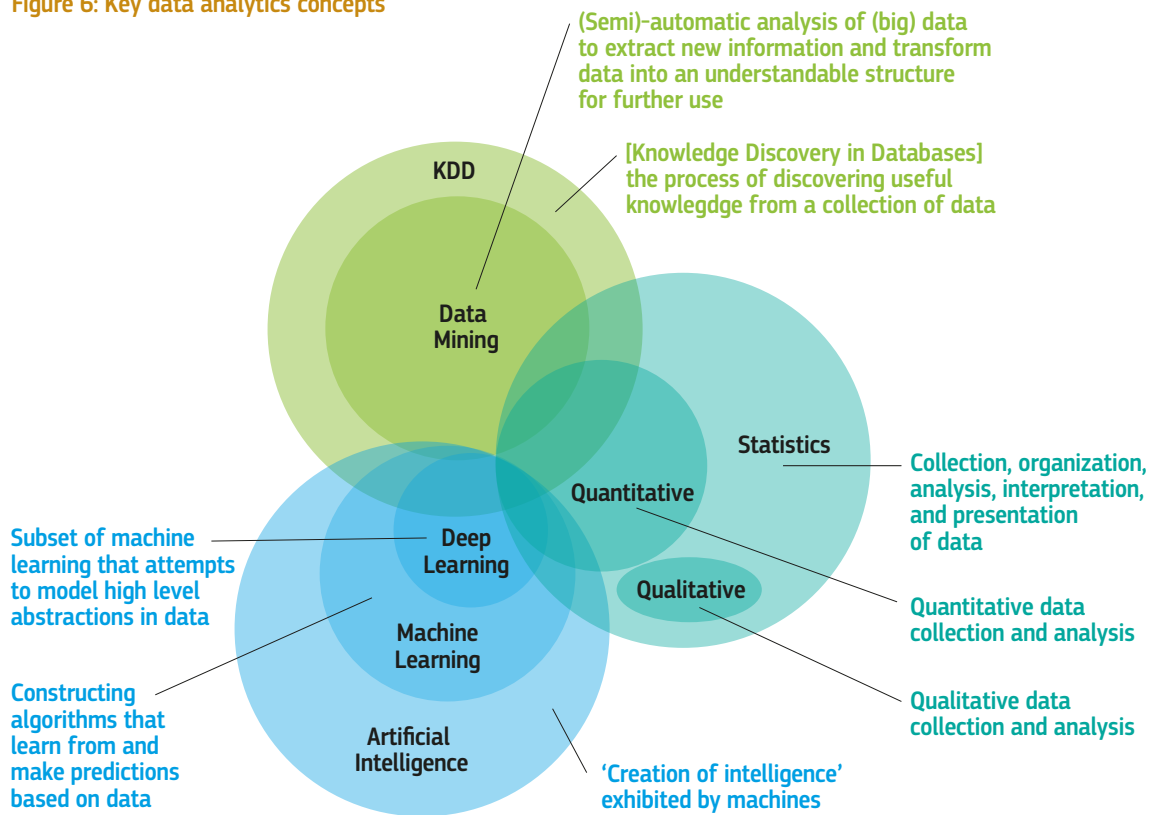


Figure 6: Key data analytics concepts



4.2 Statistics

Statistics refers to (traditional types) of research to collect data from people or other units of research. Typically, statistics is seen as merely 'traditional social science research' or just the 'statistical analysis of data'. While both are not entirely correct, we will try and explain the concept in more detail in this section. For the purpose of this toolkit, our focus is on the data collection and analysis part and we limit ourselves to statistics in the more traditional social science sense. Statistical analysis is being used in pretty much every other method of analyses, hence our focus.

We focus on two key approaches within statistics:

1. Quantitative methods

These are methods typically used to collect and analyse data that has to be 'asked' from entities and cannot be gathered using other methods. Put simply, quantitative methods are used to test hypotheses. Most commonly, quantitative methods are associated with surveys and questionnaires, however there are many more quantitative methods of collecting data, such as large scale observations or quantifications

of qualitative data (for example, if many interviews are being held, certain responses could be quantified and used for quantitative analysis).

Quantitative methods are, even in the age of big data, still very useful. While most data pulled from databases are objective, factual data, quantitative methods can be used to measure more subjective elements. The most typical example is customer satisfaction surveys. In the context of innovation, surveys are a good way to test assumptions derived from theories or get structured (quantified) feedback on prototypes or ideas. Other methods, apart from surveys, are:

► A/B tests⁷

Where different groups of respondents are being exposed to different versions of a product (or 'treatment'). Measuring behaviour after exposure can help infer effects about the treatment. This could be used by PES to experiment with different versions of websites or other online tools.

⁷ Depending on the design of the study, most of the methods explained can take a more qualitative or quantitative nature. We mention them there where they – in our view – are used more commonly.

The following table gives more detail of each approach and potential use cases for PES.

| APPROACH | DESCRIPTION | POTENTIAL USE CASE |
|--|--|---|
| Statistics | Collection, organisation, analysis, interpretation and presentation of data | Traditional types of research to collect data from people or other units of research. |
| Example(s) | Surveys, observations, interviews and the subsequent analysis of this data. | |
| Quantitative statistics ⁸ | The use of quantitative methods to collect and analyse data. Surveys are a common method to collect quantitative data. | Quantitative methods of data collection are typically used to collect data that has to be 'asked' from entities and cannot be gathered using other methods. |
| Example(s) | Surveys to measure customer satisfaction. | |
| Qualitative statistics | Use of qualitative methods to gather data that focuses on 'quality', rather than 'quantity'. Common methods are interviews, focus groups, etc. | Qualitative methods of data collection are typically used to gain a deeper 'descriptive' understanding of a topic. |
| Example(s) | Interviews to understand why people contact a PES. | |
| Data mining | Data mining is the application of specific algorithms in order to extract patterns from data. | Data mining is used to condense large amounts of data and/or transform them. |
| Example(s) | Mining usage statistics of service channels and showing trends/developments. | |
| KDD (Knowledge Discovery in Databases) | KDD is the process of discovering knowledge and patterns in large amounts of data. | KDD is used to build upon data mining and transform data into knowledge. |
| Example(s) | Fraud detection of benefits. | |
| Artificial intelligence | Goal of AI is to create intelligence exhibited by machines. | Create smarter technologies that can make decisions or support decision making. |
| Example(s) | Smart job search systems that work based on customer profiles. | |
| Machine learning | Algorithms that learn from data processed to make predictions and improve outcomes. | Machine learning is used to create better functioning algorithms and models by learning from ongoing analysis. |
| Example(s) | Improvement of matching systems by analysing reasons for previous matches and/or mismatches. | |
| Deep learning | Algorithms that learns and models based on high level abstract and layers or manifestations of unstructured data (such as written text, pictures, videos and many combinations of data sources). | Deep learning is used to explore data that is highly unstructured and abstracted and tries to create abstractions from this data. |
| Example(s) | Drawing inferences from writing styles and formatting of resumes to improve vacancy matching or training. | |

► Eye tracking

Where respondents use a device that tracks their eye movements. This helps understand how respondents navigate products, which parts draw attention, etc. This could be used in development stages of new (online) tools to help understand how people navigate pages.

⁸ One could argue that most approaches are quantitative and therefore fall in this bucket. However, we use quantitative statistics in a 'social science' context, i.e. Referring to the analysis is quantitative data collected through social science research methods, such as surveys.



PRACTICAL EXAMPLE

The US State of New Mexico Department of Workforce Solutions (DWS) noticed that many benefits applications made mistakes (purposefully or not) while applying for (unemployment) benefits, resulting in improper benefits payments.

DWS partnered with Deloitte to conduct a two stage project. The first was to use quantitative statistical analysis to model suspect behaviours. The next step was to gently nudge individuals into more desirable behaviours. The key to this was to design and test communications and notifications for claimants at three moments: 1) during the vetting process for eligibility, 2) when individuals report work and earnings, and 3) while determining an action plan to seek new employment.

By field testing different types of communications, DWS was able to analyse the best working solution and subsequently implement this. DWS was able to substantially influence claimants' behaviour. The State successfully increased accurate reporting while reducing improper payments.

(see <https://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/business-analytics-case-studies.html>)

2. Qualitative methods

These are methods typically used to gain a deeper 'descriptive' understanding of a topic. Put simply, qualitative methods are commonly used to *create* hypotheses. Most often, people associate qualitative methods with (group) interviews, but as well as with quantitative methods, there are many more approaches. Relevant examples in this context are:

► Think aloud methods

Where people are asked to perform a task and explain their thought process while performing these tasks. This could be used to gain deeper insights into the choices people make when using (online) tools or applications (such as mobile apps).

► Plus-minus methods

Where respondents are asked to mark part of a product (e.g. an online tool, or physical product) they like or dislike. Subsequently respondents are asked to elaborate on their choices. Often these are used to test brochures and other physical products, but could be used in online, product environments as well.

Like quantitative methods, qualitative research remains valuable in the age of big and/or smart data. The key function of qualitative methods is to help make sense of the world and/or get a deeper understanding of phenomenon that simply cannot be generated through other means of analysis. In innovation settings, qualitative methods are most commonly used in conjunction with other methods, throughout the innovation process.

Getting started

The following methods, tools and/or applications can be used to collect and analyse (statistical) data:

- There are many online (free) survey tools available, such as SurveyMonkey [surveymonkey.com]. Non-free products (such as Qualtrics [qualtrics.com]) often have more functionality and/or better support
- Most spreadsheet software (e.g. Microsoft Excel, LibreOffice Calc) include basic statistical functions
- Dedicated statistical software (e.g. SPSS, SAS, PSPP) can be used to perform (complex) statistical analysis on (mostly) quantitative data
- Various tools exist for the transcription and analysis of qualitative data. Examples are:

| PACKAGE/TOOL | EXPLANATION | LINK |
|----------------|--|---|
| QDA Miner Lite | Free computer assisted qualitative analysis software. Can be used for the analysis of textual data such as interview and news transcripts, open-ended responses, etc. | https://provalisresearch.com/products/qualitative-data-analysis-software/freeware/ |
| Weft QDA | Free tool to analyse qualitative data such as interview transcripts, fieldnotes and other documents. | http://www.pressure.to/qda/ |
| Transana | Software to a) various types of data in a single analysis, b) categorize and code segments of the data, c) explore coded data through text reports, graphical reports, and searches. | https://www.transana.com/ |

- Open Source and Free dedicated tools and packages can be found in the table below:

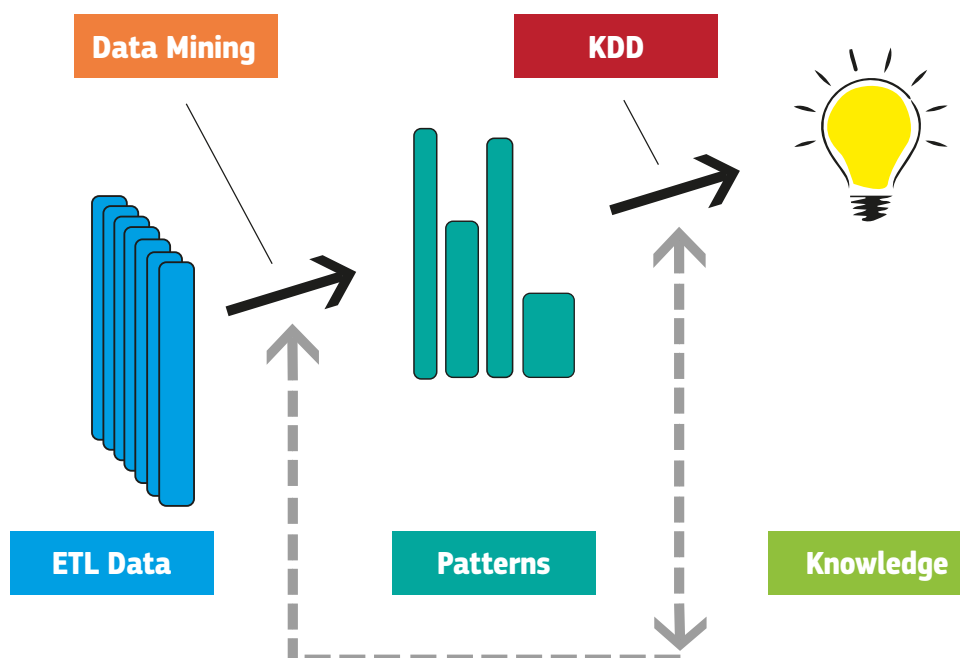
| PACKAGE/TOOL | EXPLANATION | LINK |
|--------------|---|---|
| R | R is a free software environment for analytics and visualisations. It is one of the most popular programming language for analytics. It is highly modular and a large support/ documentation community online. | www.r-project.org |
| Python | Python is a programming language that is being used heavily for analytics (sometimes in conjunction with R). Like R, it is very modular and various packages exist for specific types of analytics (or visualisations). | https://www.python.org |
| DataMelt | DataMelt is a free mathematics software. It can be used for numeric computation, statistics, symbolic calculations, data analysis and data visualization. | http://jwork.org/dmelt/ |

4.3 Data mining & KDD

Data Mining and Knowledge Discovery in Databases (KDD) are closely related (yet different) types of analytics. Their relationship can be seen in figure 7. We could argue that KDD is a way of turning *information* (see figure 7) gathered through data mining into valuable *knowledge*.

Data mining refers the application of specific algorithms in order to extract patterns from data. Data mining was invented when datasets became too large to be analysed by humans so researchers invented ways to condense large datasets and extract useful types of information. Therefore the key difference (in this context) between the statistical methods discussed in the previous section is that data mining is aimed at the automation of the analysis and presentation of results.

Figure 7: KDD process



Two common applications of data mining are:

► **Automated prediction of trends and behaviours.**

For example, based on previous purchases, marketers can estimate the likelihood of customers buying other products or when they are likely to buy the same product again.

Within PES, this could for example be used to:

- Estimate the likelihood that (certain) job seekers find (certain) jobs in a certain period of time.
- Estimate the probability of benefit fraud occurring among certain groups of people with benefits.
- Unemployment forecasting based on historical data.

► **Automated discovery of previously unknown patterns.**

By combing different variables and many types of data, data mining can be used to discover patterns in data that were previously unknown. For example, this type of data mining is used in marketing to discover if customers who purchase certain products are also buying other products (for Example, Amazon uses this to give recommendations to customers: ‘people who bought this product, also bought that product’).

Within PES, this could for example be used to:

- If job seekers with certain similar aspects on their resumes are more likely to find jobs quicker.
- If certain combinations of job seeker characteristics would also make them a good fit for vacancies not directly fitting with their past experiences.

Nowadays, data mining is often used in conjunction with more advanced types of analytics (as we will discuss below). For example, certain probabilities of occurrences discovered using data mining can be used as inputs for machine learning models. Likewise, machine learning could be used to improved algorithms used for data mining.

KDD in many ways is a follow-up step to data mining. It is important to mention here as it illustrates two points:

- While data mined can have value in itself, the true value lies in the interpretation of data and its transformation into knowledge.
- It requires extra effort to turn data into interpretative knowledge (and actionable wisdom).

To move from data mining to KDD, PES can do two things:

- Enrich the data mined (e.g. by combining variables) so that more value is created or patterns become more obvious. For example, by combining unemployment trends with labour market seasonality, it becomes possible to correct for seasonal variations in unemployment and assess the true trend (if any).
- Use experts to interpret results. Making sense of results can be extremely difficult without the proper subject matter expertise and knowledge of the context in which the analysis takes place. It is for these reasons that having social scientists (and other experts) on the data team can enhance its value manifold.

Getting started

The following methods, tools and/or applications can be used for data mining and/or KDD:

| PACKAGE/TOOL | EXPLANATION | LINK |
|--------------|--|---|
| RapidMiner | Open Source platform to mine data (and run data science applications). | https://rapidminer.com/ |
| Data Applied | Cloud based, free data mining and visualisation platform. | http://www.data-applied.com/ |
| Orange | Open source data mining and visualisation platform (with machine learning capabilities). | http://orange.biolab.si/ |

4.4 Advanced Analytics

4.4.1 Artificial Intelligence

Artificial intelligence (AI) is used to create smarter technologies that can make decisions or support decision making. The main goal of AI is to create technologies that are so smart that they can think and act like humans. The 'Turing test' is the benchmark for AI after which it can be considered human smart.

Artificial Intelligence is a broad concept that encompasses machine learning, deep learning and intersects with other types of analytics, such as data mining and statistics. In this toolkit we restrict ourselves to the 'intelligent' applications of AI where the application exhibits certain levels of smartness based on learning and creativity.

The Turing test

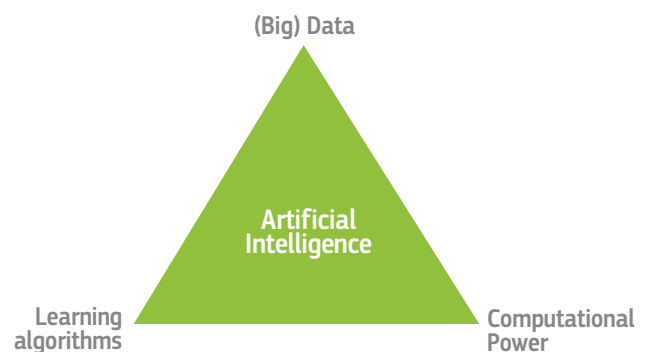
The Turing test was developed by mathematician Alan Turing in 1950 in order to test the ability of a machine to exhibit intelligent behaviour. The common version of the Turing test requires a computer to pretend to be a human. A human interrogates the machine (without knowing it is one) and subsequently decides whether he/she was talking to a human or machine. If the machine wins, it has passed the Turing test.

Understanding AI works best by looking at real world applications. Perhaps the earliest application of AI that received world-wide attention was when IBM's chess computer Deep Blue beat world champion chess player Gary Kasparov in 1996. This computer was able to process and analyse enormous amounts of data (previous chess games) and based on Kasparov's playing style reason which move was best out of all available moves. This ability of 'reasoning' is a key characteristic of AI.

More recently (March 2016), AlphaGo (by Google DeepMind) was the first artificial intelligence application to beat a human at the game Go. This is an important feat for a number of reasons. The first is that, compared to chess, Go is a much more complicated game with many more moves. Analysing

all the potential moves of a game of Go requires tremendous processing power and the ability of AlphaGo to crunch the numbers is a testament to the progress made in computational power. The second is that AlphaGo beat a human against the expectations of (many) AI experts who did not expect the application to be sophisticated enough in terms of reasoning and decision making. If you want to see more current examples of how AI is developing, <https://aiexperiments.withgoogle.com/> is a good website to get inspiration.

Figure 8: Components of Artificial Intelligence



These two examples, and the progress they showcase, help us understand the more foundational properties of AI (see figure 8). Furthermore, it helps us understand what a PES needs in order to start using AI:

- ▶ Large amounts of data that are available for AI to process.
- ▶ The availability of sufficient amounts of processing power.
- ▶ Algorithms that can learn from the data and become increasingly smart.

| EXAMPLES OF AI | POTENTIAL FOR PES |
|--|--|
| Chat bots are more and more commonplace in the private sector. These bots are able to provide basic customer service and help clients solve basic problems. | Similarly, bots could be used in service settings. They could ease the burden of call centre agents by responding to simple inquiries leaving to the complex and ambiguous tasks for human agents. |
| Virtual Assistants such as Siri, OkGoogle, Cortana, that assist people throughout their days by given (localised and personalised) recommendations, pro-actively serve information and answer questions. | Virtual Assistants that guide the job seeker through the entire process of registering to finding a job and exiting the system. It reminds job seekers of when to do things, gives them advice on how to do things and answers questions about the process (e.g. Give writing tips when a jobseeker creates a resume or application letter). |
| Video Games where artificial intelligence controls the behaviour of 'opponents' in shooter games. These opponents can 'learn' from the players' behaviour and subsequently change their routines. | Serious gaming where jobseekers can practice job interviews and based on AI the interviewer asks relevant questions (and subsequently gives feedback). |

Examples and potential for PES

In the table above, we describe several existing applications of AI and how PES could develop similar applications of AI.

Getting started

With the novelty to AI and the lack of experiences of PES (or other governments) that could be readily implemented within PES, it seems advisable to start small. Smaller scale experiments with AI allow PES to explore the possibilities, reduce the amount of data needed and the complexity of the algorithms. The following tools could be of help.

| PACKAGE/TOOL | EXPLANATION | LINK |
|--------------|---|---|
| Open Cog | OpenCog is an open-source software project aimed creating general Artificial Intelligence applications. | http://opencog.org/ |
| Watson AI | Watson is being marketed by IBM as relatively generic application to business and governments and is used in various areas such as (structured and unstructured) analytics, virtual assistant, data integration and search. | http://www.ibm.com/watson/ |

4.4.2 Machine Learning

Machine learning is used to create better functioning algorithms and models by learning from ongoing analysis. Machine learning is a subset of artificial intelligence and there is disagreement about the exact difference between the two concepts. We see the difference as machine learning being mostly used to analyse large volumes of data, discover patterns in these data and subsequently learn from the data. Artificial intelligence goes one step further

and includes systems that can make decisions, combine elements, reason and thus show behaviour comparable to human thinking. Herein also lies a key difference between machine learning and data-mining/KDD. In machine learning there is a clear emphasis in learning from the data and the applied analysis for future iterations.

Several types of machine learning exist; in the table below we list a number of common types of machine learning:

| TYPE | EXPLANATION | PES POTENTIAL APPLICATION |
|-----------------------|--|--|
| Supervised Learning | The team feeds clearly labelled data into the computer with desired outcomes and the machine learns how to manipulate inputs into outputs. | Learn how jobseekers best match certain jobs based on pre-defined inputs. |
| Unsupervised Learning | In this type of learning, the machine tries to create inferences by itself based on unlabelled or unstructured data. | Analyse job-seekers resumes, create certain skill categories based on these resumes and match these skills to vacancies. |
| Reinforced Learning | The machine must achieve a goal and has create its own feedback mechanisms to assess whether it is getting closer to its goal (such as self-driving cars). | Running simulations to minimise unemployment over time. |

Machine learning is being widely applied in commercial settings and healthcare. Virus scanners on computers are based on machine learning and so are smart thermostats that learn about your heating preferences and adjust heating cycles accordingly.

In the following table we list some more applications from machine learning from other domains and how these could be applied within PES.

| EXAMPLES OF MACHINE LEARNING | POTENTIAL FOR PES |
|---|--|
| Recommender systems from companies such as Netflix and Amazon that become smarter by learning from users' behaviours and people similar to the user. | Improvement of job matching systems based on technology similar to (commercial) recommender systems. For example, based on successful matches of (previous) jobseekers, with similar characteristics, more tailored recommendations for vacancies could be made. |
| Fraud detection systems such as used by banks, insurance, and credit card companies rely on machine learning to detect potentially fraudulent transactions. | Detection of potentially fraudulent benefit applications by analysing patterns in past known applications and learning the difference between fraudulent and non-fraudulent applications. |
| Chat application Skype ⁹ uses machine learning to translate voice or instant messaging conversations in a multitude of languages in real time. Its machine learning algorithms make it better as it gets used more frequently. | Facilitate international job interviews or cross-regional interviews in countries with multiple languages. It could also be used for job (language) training and/or counselling. |

9 See <https://www.skype.com/en/features/skype-translator/>

Examples and potential for PES

Within PES, machine learning has not been widely applied. One notable exception is the application of learning algorithms at the Flemish PES (VDAB) (see below).

Getting started

Several of the tools already mentioned can be used for machine learning (such as R & Python). The table below lists more relevant tools.

| PACKAGE/TOOL | EXPLANATION | LINK |
|----------------------------------|---|---|
| Apache Mahout | An environment for quickly creating scalable machine learning applications as well as a free library of machine learning algorithms. | https://mahout.apache.org/ |
| OpenNN | A C++ library implementing neural networks. | http://www.opennn.net/ |
| TensorFlow | A software library for machine learning. | https://www.tensorflow.org/ |
| Encog Machine Learning Framework | Encog is an advanced machine learning framework that supports a variety of advanced algorithms, as well as support classes to normalize and process data. | http://www.heatonresearch.com/encog/ |



PRACTICAL EXAMPLE

The Flemish PES (VDAB) is working to improve job matching by using Big Data to recommend vacancies to jobseekers when they access the VDAB's vacancy system. In 2016 a *Recommender system* was developed based on a twofold objective: finding out which users are interested in which vacancies (by looking at what they click on, what they read, open and look at); and predicting a jobseeker's interest in other vacancies (by looking at what similar users have looked at, analysing behaviours). VDAB seeks to make both accurate and extended recommendations to jobseekers through this system, looking to open the pool of jobs that a jobseeker could find interesting based on a range of preferences that are expressed in the vacancy.

The *Recommender system* is currently being tested on a set of sub-users. This system has been developed by VDAB's Innovation Lab. More information can be found about the Lab on a specific fiche accessible in the [PES Practices website](#).

is a subset of artificial intelligence. In our view¹⁰, the key difference between machine learning and deep learning is that deep learning focuses heavily on unstructured and abstract data as well as the combination of many layers of data. Machine learning tends to focus on structured data and discovering patterns in data that are well organised.

Deep learning is for example used to automatically organise and tag photos. Companies like Google and Facebook can recognise people and locations in photos and can use this information to tag and categorise the photos.

Thinking along these lines, a potential application for deep learning in PES is to have algorithms learn from job seekers' CVs or resumes and discover useful information. For example, formatting styles, fonts used, colours and pictures could tell us something about the job seeker that could be useful when recommending jobs or to help them optimise their resumes. Similarly, recordings (with consent and for

4.4.3 Deep Learning

Deep learning is used to explore data that is highly unstructured and abstracted and tries to create abstractions from this data. Deep learning is a subset of machine learning (which, as explained above)

¹⁰ Once again, many different interpretations exist, so the reader may have come across different definitions. We have tried to create an easy to understand common definition.

| PACKAGE/TOOL | EXPLANATION | LINK |
|----------------|--|---|
| Deeplearning4j | Open-source, distributed deep learning framework. | https://deeplearning4j.org/ |
| Keras | Python library for development and evaluation of deep learning models. | https://keras.io/ |

research purposes) could be used to analyse tone of voice and emotions and this could be used to personalise service delivery processes.

Getting started

As with machine learning, many of the tools already mentioned can be used for deep learning (such as R & Python). The table above lists more relevant tools and/or specific packages.

4.5 Combinations & Derivations

In the previous section we have described what currently (in our view) the most important and promising types of analytics are for PES. However, many sub-types, combinations and derivations of these main classes exist. In this section we briefly mention several of these types of analytics. For each type, we define and explain the concept, describe how it relates to other types of analytics, and outline how it may be of value for PES.

1. Predictive/Prescriptive analytics

Based on models, trying to extrapolate from previous data points to future data points. Many recommender systems are based on predictive analytics and so are well known examples such as weather forecasts.

Within PES, this could, for example be used in:

- ▶ Unemployment or labour market forecasting
- ▶ Job seeker profiling applications (e.g. By estimating job seekers developments and training needs)
- ▶ Matching applications (e.g. By predicting the ease with which vacancies can be filled).

2. Natural language processing (NLP)

NLP refers to a broad class of methods to interpret 'normal people' or natural language (and translate that in other 'types' of language). This is used by speech recognition and translation software.

Within PES, this could be used to:

- ▶ Better understand customer service communication (and for example create content that better aligns with clients' language)
- ▶ Interpret jobseekers' resumes, and better match the languages used by jobseekers and employers
- ▶ Create better classification schemes (e.g. ESCO) by mapping jargon and technical terms to human language.

3. Image recognition

This is a special class of machine/deep learning focused on understanding or learning from (digital) images. As mentioned above, the underlying deep learning algorithms are used to tag or categorize content).

Within PES, this could be used to:


- ▶ Analyse profile pictures used for resumes and make recommendations for job seekers' pictures to better match certain jobs.

4. Speech recognition

Closely related to natural language processing, speech recognition is used to understand spoken language. Combined with natural language processing and other types of machine learning and AI, this could be used to create social robots and/or chat bots. Currently, speech recognition is used to transcribe spoken communication. This helps create archives of communication and allows organisations to understand content (such as questions customers have and their accompanying emotions).

Within PES, this has the following potential applications:

- ▶ Understanding tone of voice and emotions in customer service interactions to better understand perceived problems and obstacles
- ▶ Allow communication to be continued and stored on other channels
- ▶ Better understand word choices and use these to update web and written content.



Chapter 5. Presenting & Reporting

In this chapter we focus on presenting results and creating reports. We do not focus on traditional (written) reports (given the abundance of resources on reporting). We do focus on (interactive) visualisations and dashboards as well as sharing data with the public using open data. Throughout, we discuss pros and cons and considerations.

The following topics form the main content of this chapter:

- ▶ Understand why traditional reports may not be the best way to present results from smart data analytics
- ▶ Provide overview of (novel) ways to present and report data
- ▶ Considerations when reporting data and who to open results up to.

Strategic questions this chapter answers:

- ▶ How can I get the best possible insights to help my (strategic) work?

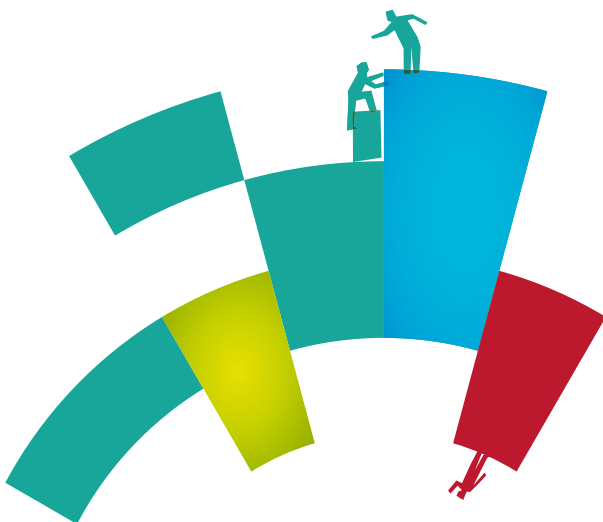
Tactical questions this chapter answers:

- ▶ What are available ways to present and report data?
- ▶ What are the use cases for each of those and how will they help PES?

5.1 Why move away from traditional reports?

While the traditional report is still widely used and in many cases a good way to convey information and report research findings, many organisations are moving away from written reports when reporting about smart data. Several reasons for this exist:

- ▶ **Keep results up to date**
One problem with traditional reports is their static nature. These creates problems with dynamic data and especially high velocity data that creates a continuous stream of insights with no natural stopping points. To benefit from the possibility to refresh results, more dynamic ways to report outcomes are desired.
- ▶ **Say more with less**
Text is often not the best way to describe results and the static nature of many graphs in reports do not allow to dig deeper into the data. This is a problem solved by interactive visualisations that allow to filter, sort and search through the data and show those insights that matter.



▶ **Be more appealing and actionable**

One of the bigger problems with traditional reports is their linear nature and the often dry nature of writing that does not compel readers to read the entire report, let alone follow up on its recommendations. This is a problem that many interactive tools try to solve by offering personalised insights and offer more dynamic routes to explore outcomes.

5.2 (Interactive) Visualisations

We start this overview with a short discussion of the role of visualisations. Visualisations are a powerful way to show findings and interesting patterns in data. Key benefit of visualisations is that they allow to:

- ▶ More easily show relationships and developments (over time) than using text.
- ▶ Provide a better way of conveying and remembering information than using text for many people (depending on their learning styles).

However, visualisations also have drawbacks, most notably:

- ▶ Only visual representations may lead to false conclusions, for example when a graph suggests a relationship between variables while in real life there is no significant relationship.
- ▶ Complex graphs can be overwhelming and distract from the message that the sender wants to convey. Especially the 'volume' aspect of big data can cause problems when trying to visualise too many data points at once.
- ▶ Very often it still requires skills and (domain) expertise to interpret results. When just visualisations are presented it can be difficult to judge the value of the results presented.
- ▶ While great for presenting results and data visualisations are not the best vehicle to raise concerns, points for discussion and describe context.

In general, when information is very complex and much contextual information is needed simple visualisations as a stand-alone way to report information are not the best way to convey a message. For these reasons, the following seem good use cases to use (interactive) visualisations:

- ▶ In production environments when the information is simple enough to be understood without too much contextual information. For example, visual

elements could be used to show the extent to which job seekers match to certain jobs.

- ▶ In situations when summaries of information are presented (for example in evaluations of pilots). Condensed versions of information (like summaries) reduce the complexities of information and make it easier to use visualisations).
- ▶ When room for contextual information is available (e.g. in dashboards or interactive web pages) more complex information can be transmitted using visualisations, provided the contextual information allows to interpret information correctly.
- ▶ When experts (e.g. from the data team) are available to help interpret information, more complex visualisations can be used and the experts can help resolve any ambiguities.

Getting started

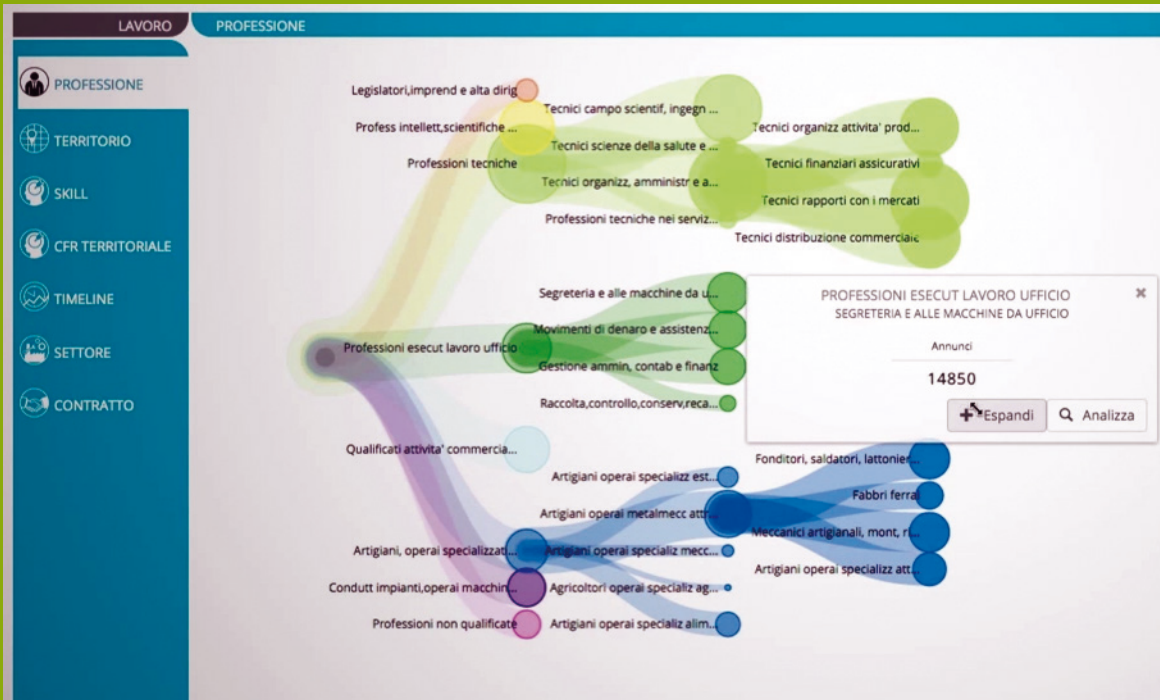
There are many ways to create following methods, tools and/or applications can be used to calculate descriptive analytics:

- ▶ Most spreadsheet software (e.g. Microsoft Excel, LibreOffice Calc) include basic graph functionality as well as rudimentary capabilities for interactivity.
- ▶ Most statistical software (e.g. SPSS, SAS) have rudimentary visualisation capabilities that allow for basic manipulations.
- ▶ Many packages or programming languages can be used for data visualisations. Some of the most common ones are the jQuery, Chart.js or libraries for JavaScript. For some (interactive) examples see <http://blocks.org/mbostock>.
- ▶ Many online services are built on top of these packages (and other code) and allow to create visualizations online without the need for coding (e.g. Datawrapper.de).
- ▶ Certain dedicated software tools exist for data visualisations. Tableau (.com) is a well-known example, but others exist, such as Sisense (.com). Very often these tools blur the line between 'simple' visualisation tools and (online) dashboards that can be used to analyse and manipulate data.



PRACTICAL EXAMPLE

WollyBI (wollybi.com) is a spin-off from the Department of Statistics and Quantitative Methods (CRISP), University of Milan-Bicocca. Together with regional public employment service, CRISP created the WollyBI platform as a means to visually explore the labour market in various regions in Italy based on vacancies, location and skills. It currently allows to – in a user friendly manner – visualise various analyses of over 1.5 million job vacancies.



5.3 Interactive Tools & Dashboards

In this section we discuss (novel) ways to present data using interactive (web) tools and online dashboards. The benefit of these tools is that they allow the user to interact with the data and thus allow the user to a) personalise the data to fit his/her needs (e.g. Through searching, sorting, and filtering), b) explore patterns more easily while going from one section to another (related) section and c) get help and contextual information more easily (e.g. Through embedded help functions).

We define a Dashboard in this context as ‘an interactive tool, most often based on web technologies, working directly on top of data sources that allow to manipulate and visualise information and provide additional textual and contextual information’.

Interactive tools and dashboard aim to resolve many of the issues existing with stand-alone visualisations. For example:

- ▶ Dashboards typically allow for more sophisticated types of data manipulations such as sorting, searching and filtering.
- ▶ Dashboards can include additional information explaining data points that help interpret information.
- ▶ Dashboards can include contextual information that help understand the setting in which the data was collected and analysed. This can increase understanding of the data.
- ▶ Dashboards can include recommendations or conclusions that build upon the data presented.
- ▶ Dashboards allow to link between different sections and thus allow for more dynamic routing of information.

- ▶ Dashboards can link to information outside of the dashboard, allowing to link to other relevant subjects.
- ▶ Dashboards allow to integrate communication tools (e.g. Chat applications) or link to communication tools, so that the user of the dashboard can contact the data team or other support staff to aid with the use of the dashboard and/or interpretation of information.

In many cases dashboards are cloud based or through other means accessible on the internet or intranet. This allows to share information easily and allow flexible access policies. Furthermore, because dashboards typically work on top of analytics data sources they allow for easy discrete or continuous data updates.

However, dashboards could suffer from the following drawbacks:

- ▶ Because of the flexible nature of the dashboards and the many features than can be added, the

risk exists of trying to add 'everything' to a dashboard leading to something called 'feature creep', namely adding every single possible data point and type of information. This could severely distract from the goal which should be achieved.

- ▶ Maintaining access rights requires resources and solid policies for this have to be created.
- ▶ When dashboards are open (i.e. without user authorisation or accessible without credentials), extra care needs to be taken to prevent misinterpretation and abuse of data.
- ▶ Even with the possibilities to explain and add context, interpreting data remains complicated and even with the best dashboard, experts may still be needed. The risk exists that dashboards are used to replace instead of *complement* experts.

Getting started

The following tools can help getting started with more interactive visualisations or dashboards.

| PACKAGE/TOOL | EXPLANATION | LINK |
|---------------|--|---|
| Shiny | Web application to turn (R Based) analytics into interactive applications. | http://shiny.rstudio.com/ |
| Google Charts | Web application to create (interactive) graphs from various types of data. | https://developers.google.com/chart/ |
| Tableau | One of the most popular applications to create interactive graphs or dashboards. Available desktop, server and cloud solution. [has a free trial, afterwards paid for] | http://tableau.com |

5.4 Open data

A special class of reporting is that of open data. Open data is slightly outside of the scope of this toolkit and for this reason, we only discuss the concept briefly. In open data, data sources (such as raw data or analysed data) are opened up for broader audiences to use. Four main reasons for open data exist:

- ▶ **Create transparency**
By opening up data about governments and processes, public organisations allow the public to see what governments do, how they function and how money is being spend.
- ▶ **Accountability**
A second reason, and tied to the first, is that of accountability. Using open data, citizens can check upon the promises and actions of governments.
- ▶ **Improve services or processes**
Using open data, external parties can help governments improve processes (e.g. by finding data patterns or results not previously discovered by governments themselves). Furthermore, it creates the opportunity for external parties to compete with PES and do certain things more effectively and/or efficiently.
- ▶ **Innovate**
Using open data, third parties could develop new, innovative applications that could provide new or additional services for job seekers or employers. This could indirectly benefit PES.

While open data have these potential benefits, the following points, directly tied to the topic of this toolkit need to be kept in mind:

- ▶ Develop good policies for privacy and confidentiality when opening up data to the public. Also bear in mind that, through combinations of variables and smart inferences, it sometimes is possible to identify (possible) individuals even when PII is removed from datasets. For this reason, open data should be screened extra carefully.
- ▶ Security is always important and when creating an open data infrastructure, the PES basically creates a (vulnerable) entrance to its data systems. To minimise security risks, it is advisable to create a separate infrastructure for open data that is not connected to critical systems or infrastructures.
- ▶ To make sure open data is used and interpreted properly, it is important that open data are documented and described properly.

The following examples of open data can be inspirational:

- ▶ **LMI for All** (see <http://www.lmiforall.org.uk>)
Online data portal, containing sources of high quality, reliable labour market information (LMI) from the UK.
- ▶ **Data.gov** (<http://data.gov>)
Open data portal from the United States Government. Contains a wide variety of open data sets and a good searchable interface.
- ▶ **European Union Open Data Portal** (<https://data.europa.eu/euodp/en/data>)
The EU Open Data Portal is an access point to a growing range of data produced by the institutions and other bodies of the European Union.
- ▶ **Canadian Open Data Portal** (<http://open.canada.ca/en/open-data>)
Is interesting because it showcased applications developed based on the open data available in the portal.

Chapter 6. Evaluation & Continuation

In the sixth and final chapter we focus on aspects pertaining to evaluation, and continuation. Evaluating both the outcomes and the process are often neglected parts of any analytical process. However, they are extremely important in judging the quality of both process and outcomes and therefore are a crucial step in deciding what to do with the outcomes of an analytics process. For example; should results of a pilot be implemented organisation wide? Only a thorough evaluation can help answer this question properly.

The following topics form the main content of this chapter:

- ▶ Help understand the importance of evaluation
- ▶ Create an ongoing data analytics infrastructure and culture.

In this chapter, we answer the following strategic questions:

- ▶ How do we ensure our analytics are correct?
- ▶ What do I need to do to make this part of (ongoing work in) the PES?

The following tactical questions are answered:

- ▶ How do I evaluate the process and the outcomes?
- ▶ What can I do to ensure continuation?
- ▶ How do I scale up from pilots/small scale projects to something larger?

6.1 Evaluation

Note: this section does not focus on evaluation as a research method. It focuses on the evaluation of any analytics process.

Evaluation is the last step in any (stand-alone) analytics process and should be a part of any ongoing or continuous activity. There are many reasons why evaluations are very important, such as:

- ▶ Learning from the experience in order to improve future activities or prevent mistakes from happening again.
- ▶ Assessing the quality of the results so that they can be judged by their true value.
- ▶ Assessing the quality of the process to verify whether no mistakes have kept in and the results are reflective of the desired situation.
- ▶ Documenting the experiences so that other teams can replicate the process and/or draw the same learnings.



| | | EVALUATION FOCUS | |
|------------------|------------|--------------------------|----------------------------|
| EVALUATION FOCUS | | Process oriented | Outcome oriented |
| | Continuous | Process flow evaluations | System outcome evaluations |
| | Ad-hoc | Evaluations | Assessments |

- ▶ Understanding whether the results of the analytics activities are worth the (financial) investments.
- ▶ Determining whether one time projects or pilots can or need to be implemented throughout the organisation or be scaled up.

Given the many reasons to conduct evaluations it is important to include evaluations in any data related activity and this entails:

- ▶ Incorporation of evaluation as a step or activity in any data analytics process
- ▶ Dedication of resources to conduct the evaluation.

Not every evaluation is the same though and we can distinguish between four different types of evaluations¹¹ based on the combination of the evaluation focus (what is being evaluated) and evaluation moment (when is the evaluation taking place). The following figure gives an overview. Different types of evaluations can also take place simultaneously—they are not mutually exclusive.

Each of these evaluations has the following characteristics:

- ▶ **Process flow evaluations**
These are ongoing evaluations of analytical processes and aim to answer such questions as: Is the analytical system working as expected? Are there any errors? Are the results within anticipated margins and are their large variations in the time it takes to run models? Process flow evaluations are typically used to determine if the process itself is functioning well. For example, when a recommender system takes much longer to generate a recommendation in certain situations, there could be a bug or error in the model or data input.
- ▶ **System outcome evaluations**
These evaluated the outcomes of an analytical model or system continuously. They aim to answer questions as: are the outcomes of the

analytics as expected? Do (predictive) analytics match actual outcomes or can we triangulate outcomes to other data-sources and assess their validity?

For example, when a PES implements a predictive analytics tool to predict the likelihood of job seekers finding a new job within a certain amount of time, system outcome evaluations will focus on measuring whether the predicted result (eventually) will lead to job seekers finding jobs.

- ▶ **Process performance evaluations**
In these evaluations, usually when a project or pilot is concluded or a process is evaluated more thorough, the process itself is being evaluated. Usually, these evaluations are broader than process flow evaluations and process performance evaluations, in the context of analytics, focus on the entire analytical process and could include such questions as: was the team working well together, was the process effective and efficient, are we happy with the tools and methods used? Are outcomes being used properly?

For example, when a PES conducts a pilot to implement a new profiling system, a process performance evaluation could consist of interviews with members of the data team and PES employees using the new profiling system to evaluate the process.

- ▶ **Outcome assessments**
This fourth type of evaluation focuses on either the actual outcomes of a pilot or one-time project or the (overall or cumulative) outcomes of an ongoing process at clearly defined points in time. This helps answer questions such as: re analytics creating correct outcomes over longer periods of time (e.g. When adjusting for seasonal fluctuations), are the outcomes of a pilot satisfactory? Are the outcomes of a new analytical tool/method more robust, valid or reliable compared to the old method (and should we therefore implement the new method?).

¹¹ Also see the Analytical Paper on his topic, Pieterse (2016).

For example, one could compare the quality of human generated vacancy matches with those of a new analytics platform after a certain amount of time and certain number of matches made. The comparison of the two ways of matching could tell the organisation which one performs better (and should therefore be implemented).

Getting started

The following can help in getting started with evaluations:

- ▶ Include evaluations in any *analytics* plan or proposal. This ensures commitment to evaluation from the start. This should also include the exact evaluation moment. For continuous evaluations, many moments to assess these evaluations could be scheduled (for example recurring evaluation meetings).
- ▶ Try to connect evaluations to KPIs or practical relevant outcomes in the organisation. For example, certain performance goals for a new tool can be specified and evaluations can be used to measure progress towards these tools. This way, evaluations become a more practical and useful tool.
- ▶ Similarly to (for example) security training, (new) data team members could be trained regularly on the importance of evaluations.
- ▶ Especially concerning more important initiatives, it could be helpful to get the help of external auditors or evaluation agencies. Furthermore, external auditors can be useful in general to use (randomly) on projects to guarantee integrity and guard bias within the data team.

6.2 Continuation and scale-up of pilots

While ongoing evaluation is incredibly important and an organisation should keep a pulse of any ongoing analytical activities, one types of evaluation is very important in terms of the decision the organisation has to make:

For any data analytics project that starts small, at some point in time the organisation has to decide whether this project (whether it is a research activity, pilot or experiment) will move from being a small scale pilot to being rolled out in the entire organisation.

This decision regarding continuation or full scale-up or implementation in the organisation is important

for the organisation. It will likely impact how (parts) of the organisation are working, could impact service delivery for jobseekers and/or employers and could consume valuable resources in the organisation. Besides these organisational aspects, there are several technical considerations when scaling up initiatives. Here are some examples of both:

Technical considerations:

- ▶ From a data standpoint, one has to be absolutely certain that the results are valid, reliable and can be generalised (if that is the aim). This typically requires a lot of testing and running models to assure model fit is right.
- ▶ When scaling up, the assumption is that the new tool or product has obvious (and measured) benefits outweighing the previous tool or product. A cost-benefit analysis can help in this decision making process. Similarly, this can help in deciding whether to continue certain analytics projects.
- ▶ Broader implementation requires that the tool is capable of being scaled-up. This is not always the case. For example, code may not always be very efficient which is not necessarily a problem working with small datasets in smaller settings, but could lead to performance issues in large production environments. A technical assessment and requirements analysis can help in determining what needs to happen in terms of product (performance) requirements and stability before a tool is ready for a production environment.
- ▶ The interface of any product or tool could work well in a laboratory setting but not necessarily in a production environment. Highly trained members of a data-team used to working with data and tools have different needs than PES staff members working with production tools. For example in terms of: a) the ease of UI/UX, b) explanations/context surrounding datapoints, c) help/support functionality. User interface experts can help readying an experimental tool for a production environment.

Organisational considerations:

- ▶ Very often an expansion or scale-up of a project or implementation of a product requires training and support for new staff members. This requires resources and time that may not be directly available and need to be planned.
- ▶ When the purpose of a data-team is solely to focus on experimenting and innovation,



PRACTICAL EXAMPLE

The Finish PES developed a new statistical profiling tool that was implemented in the organisation in 2007. The profiling tool was part of an integrated IT system that calculated a risk estimate for the jobseeker at registration using administrative data. The new model was found to be 90 per cent effective at estimating the likelihood of a jobseeker being unemployed for over 12 months. However, case workers did not think the tool was useful and did not trust the results from the tool. As a result the tool was withdrawn from the production environment (see Kureková, 2014).

the data-team may not always be the most logical place to 'own' a production tool or facility. If that is the case, a consideration becomes who the functional owner of the tool should be (and who is responsible for support, further development and maintenance).

- ▶ Perhaps most importantly, and often overlooked, is dealing with resistance in the organisation, as well as creating a 'data-driven' or innovation oriented culture in the organisation. Without the support from the employees in the organisation, any innovation is doomed to fail. Proper communication plans and cultural initiatives are essential in the scale-up or implementation of any new tool. Current attitudes and cultures should be taken into account when making implementation decisions.

In sum, while an analytics activity may result in a useful result, or even application that could be used within the organisation, there usually is a fairly long way to go before results can be used in the everyday work of the PES. Planning the implementation process carefully and taking the considerations above into account can help greatly in creating a smooth process.

Appendices

Appendix 1 | Safe Harbor De-identification types

This list gives an overview of the HHS types of PII that need to be removed from data sets following the Safe Harbor method. For the full text, consult <http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>

- a) Names
- b) Address/geographic information (such as street address, city, county, precinct, (full) ZIP code)
- c) All elements of dates (except year) directly related to an individual (such as birth date, admission date, discharge date, death date)
- d) Telephone numbers
- e) Vehicle identifiers and serial numbers, including license plate numbers
- f) Fax numbers
- g) Device identifiers and serial numbers
- h) Email addresses
- i) Web Universal Resource Locators (URLs)
- j) Social security numbers [or any other national equivalent]
- k) Internet Protocol (IP) addresses
- l) Medical record numbers
- m) Biometric identifiers, including finger and voice prints
- n) Health plan beneficiary numbers
- o) Full-face photographs and any comparable images
- p) Account numbers
- q) Any other unique identifying number, characteristic, or code
- r) Certificate/license numbers

HOW TO OBTAIN EU PUBLICATIONS

Free publications:

- one copy:
via EU Bookshop (<http://bookshop.europa.eu>)
- more than one copy or posters/maps:
from the European Union's representations (http://ec.europa.eu/represent_en.htm);
from the delegations in non-EU countries (http://eeas.europa.eu/delegations/index_en.htm);
by contacting the Europe Direct service (http://europa.eu/eurodirect/index_en.htm) or
calling 00 800 6 7 8 9 10 11 (freephone number from anywhere in the EU) (*).

(* The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (<http://bookshop.europa.eu>).

