BIG DATA, AI AND THE FUTURE OF EVALUATION

Ref. Ares(2021)3693531 - 04/06/2021

DIRECTORATE-GENERAL REGIONAL AND URBAN POLICY – "EVALUATION NETWORK MEETING"

WebEx Meeting, 1 July 2021

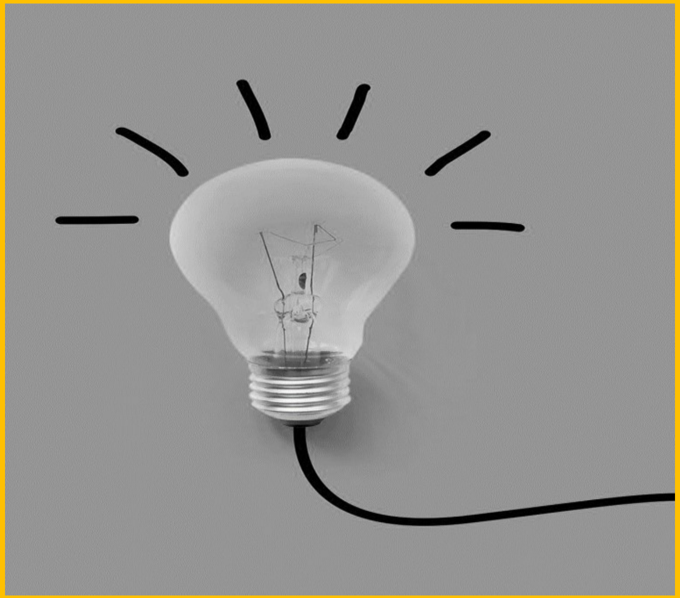
Agenda

* Thursday July 1st, 2021 *

WebEx meeting

Frans L Leeuw

Maastricht University



OUTLINE

1. Introduction
2. Basic concepts: Big Data, AI /machine learning, big data ecosystem
3. SIDE 1 of the Coin: What have BD/ AI to offer to evaluations of policies and programs?
4. Pitfalls, risks and limitations of BD/AI
5. SIDE 2 of the coin: How *[therefore]* to evaluate the use of BD/AI in policy programs and interventions?
6. Conclusions

1

Concepts: big data and evaluation, AI and the big data ecosystem (BD/AI)

Table 1. Big data characteristics

3 Vs	Volume	Vast amount of data that has to be captured, stored, processed and displayed
	Velocity	Rate at which the data is being generated, or analyzed
	Variety	Differences in data structure (format) or differences in data sources themselves (text, images, voice, geospacial data)
5 Vs	Veracity	Truthfulness (uncertainty) of data, authenticity, provenance, accountability
	Validity	Suitability of the selected dataset for a given application, accuracy and correctness of the data for its intended use
7 Vs	Volatility	Temporal validity and fluency of the data, data currency and availability, and ensures rapid retrieval of information as required
	Value	Usefulness and relevance of the extracted data in making decisions and capacity in turning information into action
10 Vs	Visualization	Data representation and understandability of methods (data clustering or using tree maps, sunbursts, parallel coordinates, circular network diagrams, or cone trees)
	Vulnerability	Security and privacy concerns associated with data processing
	Variability	the changing meaning of data, inconsistencies in the data, biases, ambiguities, and noise in data

Big Data need Analytics.
Statistical programs/ software including **Machine Learning** and **Artificial Intelligence** are often applied.

Artificial Intelligence applies different types of machine learning like **supervised and unsupervised learning** and **deep learning**.

Supervised learning involves machine learning algorithms that learn under the presence of a supervisor or guidance (e.g., regression).

Unsupervised Learning uses machine learning algorithms to analyze and cluster unlabeled data sets, which algo's discover hidden patterns in data without the need for human intervention (hence, "unsupervised").

Deep learning has been specifically modeled after the human brain. It works with algo's of a brain-like logical structure, called artificial neural networks.

TOGETHER THIS IS REFERRED TO AS THE *BIG DATA ECOSYSTEM*

2

WHAT HAVE BD/ AI TO OFFER TO EVALUATIONS OF POLICIES AND PROGRAMS?

THE FIRST SIDE OF THE COIN

Types of big data and evaluation applications

Big data [tools]	Examples of Evaluation applications
1. Social media analysis	➢ Opinions and sentiments
2. Satellites, drones and sensors (incl I of Things)	➢ Use of services (water, power); ➢ Crowd management ➢ Creating digital twins
3. Radio call-in programs	➢ Identifying sources of conflict
4. Gdelt and lookalikes	➢ Global Data on Events, Location and Tone "monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages". A telescope on society.
5. Mobile phones	➢ Effectiveness of phone messages to influence health behavior ➢ Identifying poverty hotspots

FOCUSING ON EVALUATION TOPICS DIRECTLY **RELATED TO EU/MS PROGRAMS AND ACTIVITIES:**

EXAMPLES regarding

- *Mobility and infrastructure;*
- *auditing/ inspecting neighborhoods;*
- *digitized educational assessments;*
- *social enterprise networks;*
- *textmining, natural language processing and 100.000 plus....documents*

EXAMPLE 1: INFRASTRUCTURE FOR TRANSPORT AND MOBILITY AND BIG DATA

Check for updates

OPEN **Countrywide population movement monitoring using mobile devices generated (big) data during the COVID-19 crisis**

Semmelweis University, Budapest, Hungary

In: *Nature,* 2021 (11).

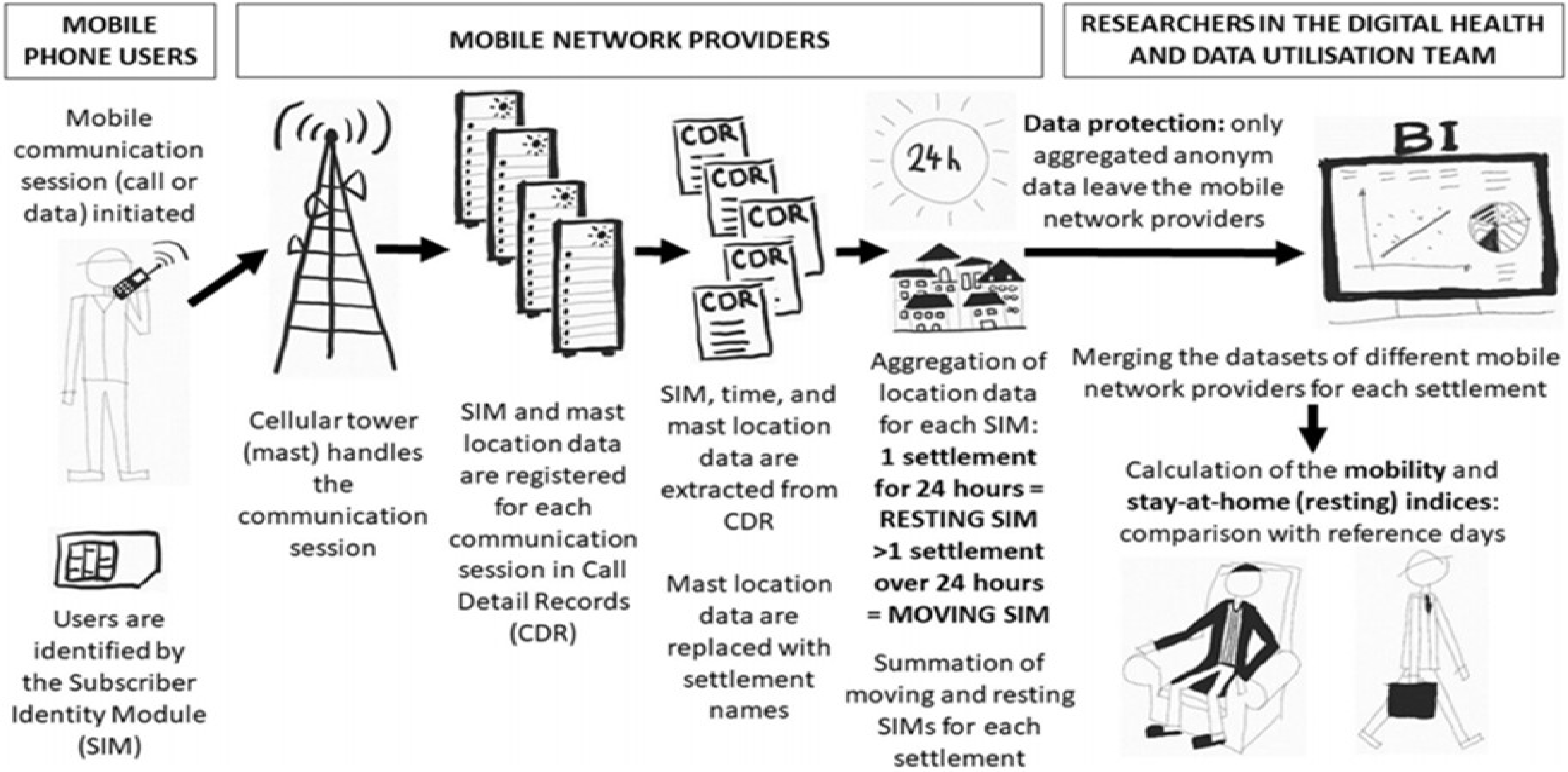


Figure 1. Schematic presentation of the methodology.

EXAMPLE 2: NEIGHBOURHOOD AUDITING IN A VIRTUAL WAY

The Promise, Practicalities, and Perils of Virtually Auditing Neighborhoods Using Google Street View

By
MICHAEL D. M. BADER,
STEPHEN J. MOONEY,
BLAKE BENNETT,
and
ANDREW G. RUNDLE

Google's Street View product offers an opportunity to cost-effectively measure neighborhood conditions across metropolitan areas.[1] Street View displays photographs with detailed geographic data from neighborhoods across the world. One can conduct audits analogous to in-person neighborhood audits using the photographs compiled in Street View. The systematic collection and analysis of Street View data have been used to assess the relationships between neighborhood conditions and individual outcomes (Odgers et al. 2012) and to study the relationships between social and economic change and changing physical conditions of neighborhoods (Hwang and Sampson 2014).

EXAMPLE 3: EVALUATING NEW EDUCATIONAL DIGITAL ASSESSMENTS EU WIDE

LOGIN

HOME WHY ASSESSMENT WHY DFA NEWS & PUBLICATIONS ABOUT PROJECT OUTCOMES

Testing and supporting the use of digital formative assessment

The teacher's role is shifting from a transmitter of knowledge to a learning coach, helping students understand and steer their own learning. Formative assessment enables the teacher to decide how best to help students develop their learning path. Assess@Learning (A@L) aims to facilitate and guide this change of role through digital formative assessment.

READ MORE

EXAMPLE 4 SOCIAL NETWORK ANALYSIS, REGIONAL DEVELOPMENT & A SMALL SIDE STEP TO DENMARK

Article

Evaluation
2016, Vol. 22(4) 435–450

DOI: 10.1177/1356389016667888
evi.sagepub.com
SAGE

Network analysis as a method of evaluating enterprise networks in regional development projects

Tamás Lahdelma
Urban Research TA Ltd, Finland

Seppo Laakso
Urban Research TA Ltd, Finland

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Interaction data from the Copenhagen Networks Study

Piotr Sapiezynski[1], Arkadiusz Stopczynski[1], David Dreyer Lassen[2] & Sune Lehmann[1,2*]

We describe the multi-layer temporal network which connects a population of more than 700 university students over a period of four weeks. The dataset was collected via smartphones as part of the *Copenhagen Networks Study*. We include the network of physical proximity among the participants (estimated via Bluetooth signal strength), the network of phone calls (start time, duration, no content), the network of text messages (time of message, no content), and information about Facebook friendships. Thus, we provide multiple types of communication networks expressed in a single, large population with high temporal resolution, and over a period of multiple weeks, a fact which makes the dataset shared here unique. We expect that reuse of this dataset will allow researchers to make progress on the analysis and modeling of human social networks.

EXAMPLE 5: REVIEWING AUDIT REPORTS AND SYNTHEZING EVALUATIONS

Natural language processing (NLP) is (...) concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them.

Our slogan:

We automate the boring stuff,
so you can audit the exiting stuff!

Introduction to Text Mining for Auditors

The Cool Stuff of Natural Language Processing and Machine Learning

Jan Roar Beckstrom
deputy director general/chief data scientist
Office of the Auditor General of Norway

Article

International Journal of Qualitative Methods
Volume 18: 1–9
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1609406919887021
journals.sagepub.com/home/ijq
SAGE

Natural Language Processing (NLP) in Qualitative Public Health Research: A Proof of Concept Study

William Leeson[1], Adam Resnick[1], Daniel Alexander[1], and John Rovers[2]

Machine Learning in Evaluative Synthesis – Lessons from Private-Sector Evaluation in the World Bank Group*

Leonardo Bravo[1], Ariya Hagh[2], Yuan Xiang[3], Jos Vaessen[4]

Article

Evaluation
1–25
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1356389020905322
journals.sagepub.com/home/evi
SAGE

Applying Big Data visualization to detect trends in 30 years of performance reports

Eran Raveh, Yuval Ofek, Ron Bekkerman and Hertzel Cohen
University of Haifa, Israel

3

PITFALLS, RISKS AND CHALLENGES USING BIG DATA/AI

- Bias 1 – only includes subjects using the apps/ the ‘digital divide’ issue;
- Bias 2 – legacy problems contributing to social bias (racial, economic and gender)
- First-the-data, then-(maybe)-the-Problem… incl. Big Data Hubris (other types of data are no longer seen as relevant)

- Ground truthing but how true is groundtruthing?
- ‘Where has theory gone’? And what about underlying assumptions when working with BD/AI?
- What about causality, incl. the blackbox character of AI
- Valid operationalizations of criteria to make BD/AI understandable, safe, transparent, and ‘believable‘.
- Problems regarding the ownership of Big Data
- Privacy and security issues, incl dangers of hacking/cyber crime

GIVEN THESE CHALLENGES AND THE RAPID DIFFUSION OF AI/BD IN SOCIETY, NOW THE OTHER SIDE OF THE COIN

4

HOW TO EVALUATE THE USE OF BD/AI IN DEVELOPING AND IMPLEMENTING POLICY PROGRAMS, INTERVENTIONS ETC?

STEPS HELPING TO EVALUATE THE USE, VALIDITY AND IMPACT (ON PEOPLE, CLIENTS, PATIENTS ETC) WHEN WORKING WITH BD and AI

1. *Specifiying the problem(s) that have to be adressed*

2. Search and articulate the assumptions underlying working w BD/ AI: the theory-driven approach.


POLICING & SOCIETY

Routledge
Taylor & Francis Group

Policing and Society

An International Journal of Research and Policy

ISSN: 1043-9463 (Print) 1477-2728 (Online) Journal homepage: http://www.tandfonline.com/loi/gpas20

Algorithmic prediction in policing: assumptions, evaluation, and accountability

Lyria Bennett Moses & Janet Chan

3. Distinguish between evaluating (AI) algorithms as such working within and between computers and evaluating (AI) algo's implementation in practical situations like health care, education, transport, justice etc. *Let data scientists etc do the first type, focusing on the performance of AI within the digital/ computerized world sui generis. We as evaluators can do the second type (with their support).*

JAMIA Open, 3(3), 2020, 326–331
doi: 10.1093/jamiaopen/ooaa033
Advance Access Publication Date: 8 September 2020
Perspective

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Perspective

Evaluating artificial intelligence in medicine: phases of clinical research

Yoonyoung Park[1], Gretchen Purcell Jackson[2,3], Morgan A. Foreman[1], Daniel Gruen[1], Jianying Hu[4] and Amar K. Das[1]

[1]Center for Computational Health, IBM Research Cambridge, Cambridge, Massachusetts, USA, [2]Center for AI, Research, and Evaluation, IBM Watson Health, Cambridge, MA, USA, [3]Department of Pediatric Surgery, Vanderbilt University Medical Center, Nashville, Tennessee, USA and [4]Center for Computational Health, IBM T.J. Watson Research Center, Yorktown Heights, New York, USA

Corresponding Author: Yoonyoung Park, IBM Research, 75 Binney Street, Cambridge, MA 02142, USA (yoonyoung.par-k@ibm.com)

Received 11 September 2019; Revised 15 May 2020; Editorial Decision 25 June 2020; Accepted 1 July 2020

IF TIME PERMITS MY ALMOST FINAL SLIDE SHOWS HOW EVALUATING AI/BD IN PRACTICE RESEMBLES THE MONITORING AND IMPACT EVALUATION OF MEDICAL DRUGS AND DEVICES.

Table 1. Evaluation for AI software compared to the approval processes of drug and devices for healthcare

Study phases	Drug	Device	AI in healthcare	Examples of study methods
Phase 0 Discovery and invention	Compound development In vitro/animal tests	User needs and workflow assessment Prototype design and development	User needs and workflow assessment Data quality check Algorithm development and performance evaluation Prototype design	Ethnographic studies to identify user needs, laboratory studies on limited data sets to measure algorithm prediction accuracy
Phase 1 Safety and dosage	Determine optimal dose Identify potential toxicities	Quality control Design updates	In silico algorithm performance optimization Usability tests	Determination of thresholds to balance sensitivity and specificity for a particular clinical use case, scenario-based testing to assess cognitive overload
Phase 2 Efficacy and side effects	Early efficacy tests Adverse event identification	Proof-of-concept tests Potential harm identification Design and quality improvement	Controlled algorithm performance/efficacy evaluation by intended users in medical setting Interface design Quality improvement	Retraining and reassessing model performance with larger real-world data sets, measurement of the efficiency of information delivery and workflow integration with representative users, pilot study of predictive algorithm in a clinical setting
Phase 3 Therapeutic efficacy	Clinical trial Adverse event identification	Clinical trial Adverse event identification	Clinical trial Adverse events identification	Randomized controlled trial to test whether delivery AI-based decision support affects clinical outcomes and/or results in user overtrust
Phase 4 Safety and effectiveness	Postmarketing surveillance	Postapproval studies	Postdeployment surveillance	Measurement of algorithmic performance drift

SOME CONCLUSIONS

Compared to 5 – 10 years ago, evaluators pay more attention to Big Data/ AI.

Focus is more on description /stocktaking (behavior, attitudes, trends, comparisons) than on contribution or causal impact analysis.

Preventing Big Data Hubris and working in a theory-informed way are important ingredients of BD/AI-evaluations but are still difficult to find.

Working with Big Data/ AI is (still, but maybe forever) confronted with also other pitfalls, risks and difficulties. These make that evaluating how Big Data and Artificial Intelligence work out in practice and in combination with H(uman) I(ntelligence) is very relevant indeed.

To do this work, an interdisciplinary approach and collaboration (with data-scientists and others) is required.

The persistent focus on bias in BD/AI is relevant indeed, but is also an indicator of some shortsightedness, if one takes on board the finding that HI (Human Intelligence/ Decision-making) is confronted with some 20 plus cognitive distortions and biases.

A final question… when do you think this was written?

Evaluation should be a mechanism of progress both within and across AI research projects. For the individual, evaluation can tell us how and why our methods and programs work and, so, tell us how our research should proceed. For the community, evaluation expedites the understanding of available methods and, so, their integration into further research.

AI Magazine Volume 9 Number 4 (1988) (© AAAI)

How Evaluation Guides AI Research

Paul R. Cohen and Adele E. Howe