Horizon 2020 2014-15 Work Programme

ICT-16 2015: Big Data research

Stefano Bertolo

European Commission, DG CONNECT
Unit G3 – Data Value Chain

OUTLINE

1. Budget, deadlines
2. Big Data is not just a buzzword
3. The Work Programme
4. What we will be looking for

Timeline, Project types, budget*

Call published: 15 October 2014
Deadline: 17:00 14 April 2015

a) budget 38M€; Research and Innovation Actions

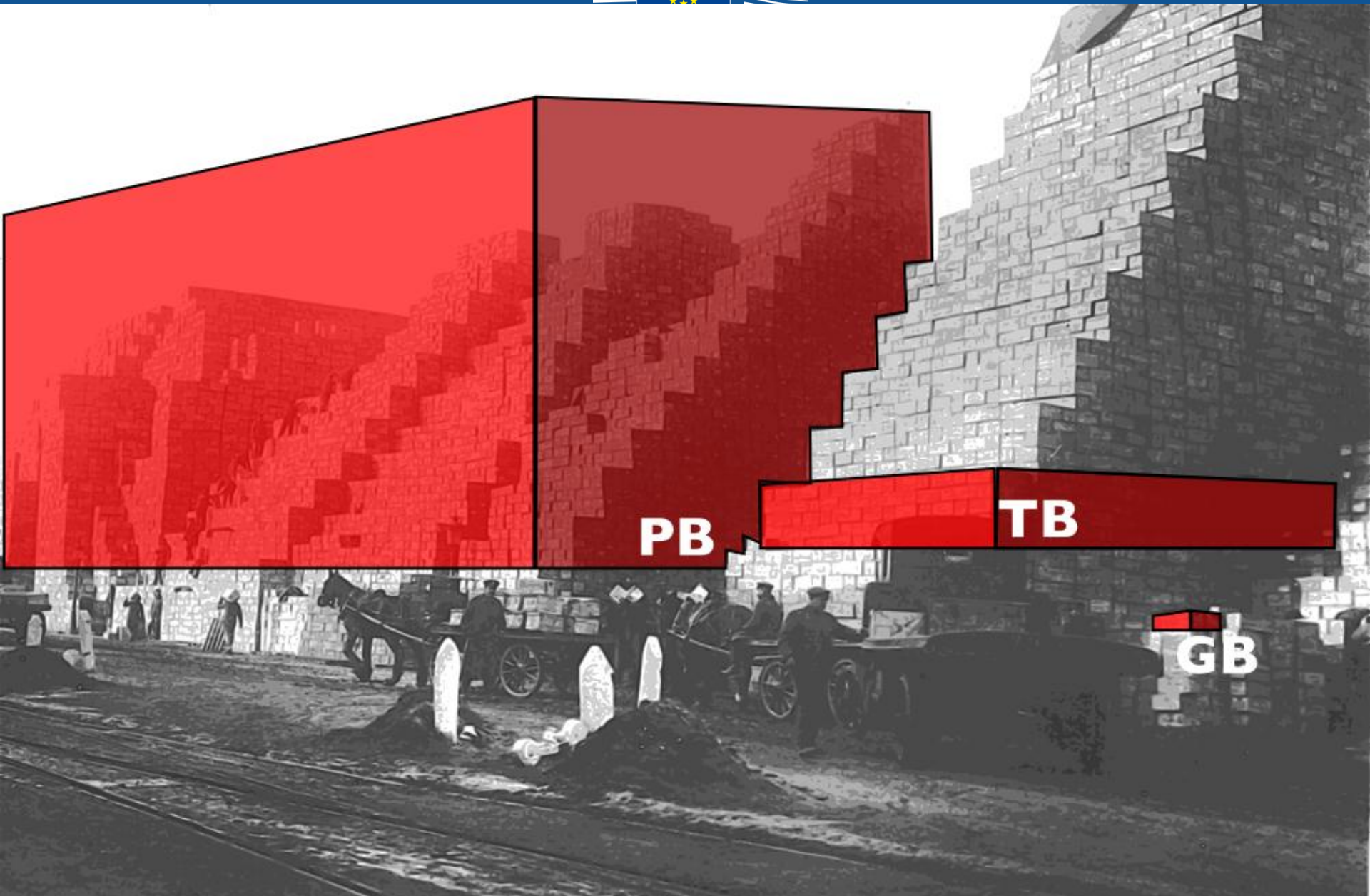
b) Budget 1M€; Coordination and Support Actions

***indicative and subject to separate 2015 decision**

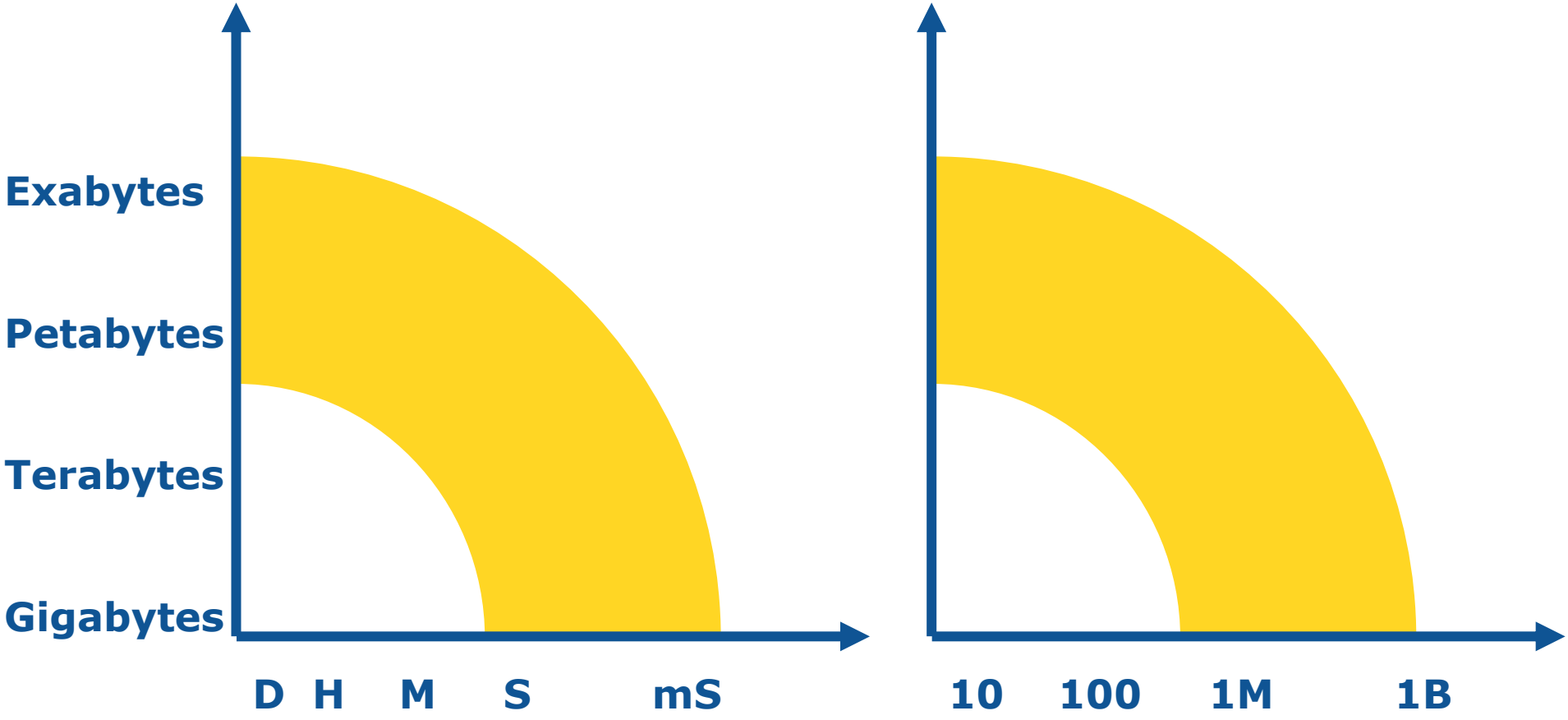
Big Data will get bigger

"big data" is when the size of the data itself becomes part of the problem

"big data" is data that becomes large enough that it cannot be processed using conventional methods


PB
TB
GB

Volume, Velocity, Variety



a) Research and Innovation actions 38M€

- **Research and Innovation**
  - Data quality
  - Data analytics
  - Prediction
  - Visualization
- **Benchmarks**
  - Industrial relevance
  - Sustainability strategy

a) Research and Innovation

- **Data structures**
  - Examples (not to be duplicated): HyperLogLog, CountMinSketch, DHT
  - Memory hierarchy conscious data structures
  - Energy conscious data structures
  - Streams
- **Algorithms**
  - Energy conscious algorithms (reversible computing?)
  - Memory hierarchy conscious data structures
  - Parallelism (exploitation of multi cores; functional programming)
  - Streaming analytics
  - Real time
  - Data curation
  - Prediction

a) Research and Innovation

- **Visualization**
  - Not just pretty pictures
  - Usability (on various platforms)
  - Measurable impact on business processes
  - Importance of experimental protocol, population of subject matter experts

a) Benchmarks

- **Industrial relevance**
  - Not (only) about 'cool' technology, but about problems (European) industry has
  - Examples (to be learned from, not to be duplicated):
    - http://www.tpc.org
    - http://prof.ict.ac.cn/BigDataBench/
    - https://amplab.cs.berkeley.edu/benchmark/#
    - http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/spec/index.html
    - http://ldbc.eu
- **Sustainability strategy**
  - Benchmarks help industry make investment decisions
  - Benchmark organizations support decision process in long term

b) Challenges and Prize schemes 1M€

Support actions to define challenges and prize schemes for verifiable performance in tasks requiring extremely large scale prediction and deep analysis. Compact consortia are required to organise and run well-publicised fast turn-around prediction competitions based on European datasets of a significant size. Proposals in this category are expected to be short in duration and are not required to provide sustainability strategies past the end of the project.

b) Challenges and Prize schemes

Verifiable performance

- Example http://sortbenchmark.org
- Must agree in advance on details of environment, data
- Must agree to the publishing of results

b) Challenges and Prize schemes

Extremely large scale prediction/analysis

- Identify prediction tasks that are relevant to those who benefit from the predictions (not only those who write the prediction tools)
- Examples on http://www.kaggle.com
- Must obtain relevant large scale datasets (this requires a detailed plan, not just a vague hope for the future)
- Datasets must be of European industrial interest

b) Challenges and Prize schemes

Compact consortia

- **Minimum necessary to run successful prize**
- **Skills and appetite for logistics/plumbing**
- **Skills and appetite for community, communications work**
- **Job of consortium is to administer the prize (not do research on its own)**

b) Challenges and Prize schemes

Prize money

- Must be explicitly budgeted as cost of activity
- Can only be awarded to entities that are eligible for funding under H2020
- Skills and appetite for community, communications work
- Job of consortium is to administer the prize (not do research on its own)

Expected impact

- *Ability to track publicly and quantitatively progress in the performance and optimization of very large scale data analytics technologies in a European ecosystem consisting of hundreds of companies; the ability to track this progress is crucial for industrial planning and strategy development.*

Expected impact

- *Advanced real-time and predictive data analytics technologies thoroughly validated by means of rigorous experiments testing their scalability, accuracy and feasibility and ready to be turned over to thousands of innovators and large scale system developers.*

Expected impact

- Demonstrated ability of developed technologies to keep abreast of growth in data volumes and variety by validation experiments.
- Demonstration of the technological and value-generation potential of the European Open Data documenting improvements in the market position and job creations of hundreds of European data intensive companies.

.

What we will be looking for

"When art critics get together they talk about Form and Structure and Meaning. When us artists get together, we talk about where to get cheap turpentine"

– Pablo Picasso

What we will be looking for

The availability for testing and validation purposes of extremely large and realistically complex European data sets and/or streams is a strict requirement for participation as is the availability of appropriate populations of experimental subjects for human factors testing in the domain of usability and effectiveness of visualizations

What we will be looking for: availability of data sets

Proposal must clearly state (ideally in a dedicated, easy to find, table):

- **Which datasets: what type of data do they contain? how big are they now? How fast do they grow?**
- **When: must be available at the very beginning of the project. Additional data at M6, M12, M18…?**
- **From whom: which partner(s) of the consortium have which access rights? (relevant if customer data)**

What we will be looking for: European data sets

- Can I use Twitter, Flickr, …? **No**. Not primarily
- Ecologically valid (i.e. operational) data from European companies
- Data sets from EU open data portals
- Data sets from other EU public sources, e.g. Copernicus http://www.copernicus.eu see http://europa.eu/rapid/press-release_IP-13-1067_en.htm

What we will be looking for: experimental subjects

- Usefulness of system proposed should be an hypothesis to be **experimentally validated**
- Define **explicitly** protocol of experiment
- Present **explicit** analysis of required sample size http://en.wikipedia.org/wiki/Statistical_power
- Describe plans to recruit required number of subjects of appropriate types (i.e. subject matter experts, **not** just a few willing graduate students)

One more thing…

Some principles valid for all ICT objectives (ICT-15, ICT-16, ICT-17, ICT-22a)

What we will be looking for: user-defined challenges

- **FP7:**
  - first researchers identify topic interesting to them,
  - then they look for narrative/qualitative validating use cases
- **H2020:**
  - first industry identifies a problem they have and cannot solve with current methods,
  - then they recruit for scientists to find one or more solutions,
  - then they define expected impact as a measurable outcome
  - then they try the solutions (run rigorous experiments)
  - then they report the measured outcomes (good or bad)

What we will be looking for: operational capacity

- **Can each partner commit the right people?**
  - Problem of swans turning into ducklings
- **Does each partner have the right resources?**
  - Computation infrastructure, data
- **Does each partner have the right skills?**
  - Experimental protocols, usability, privacy, …
- **Does each partner have enough bandwidth?**
  - Current level of commitment in other projects/activities

What we will be looking for: data management plan

- **How to manage proprietary data**
  - Make sure industrial partners are comfortable sharing
  - Ensure protection of personal data
- **How to manage/share non-proprietary data**
  - In H2020 non-proprietary data under Open Access regime
  - Choose formats wisely for maximal reuse
  - Establish credible post-project future for datasets

Thank You!

-

Questions?