כנס מחקר: מידע וידע 2013 אוניברסיטת חיפה

Big Data & Data Science in Academia – *Why*, *When* and *What* Aspects

Elan Sasson
Post-Doctoral Fellow
Tel Aviv University

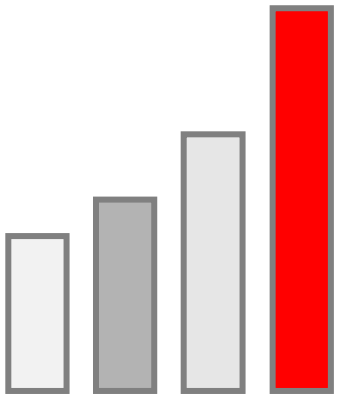
Adir Even & Nava Pliskin
Ben-Gurion University of the Negev

Big Data

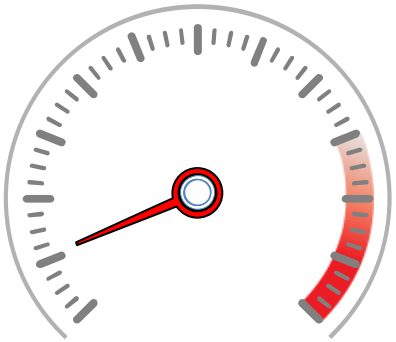
Big Data refers to datasets that grow so large that it is difficult to capture, store, manage, share, analyze and visualize with the typical database software tools.

Digital Universe ~ 40 Zettabytes(10^{21}) 2020 , only 0.5% is analyzed (IDC 2012)

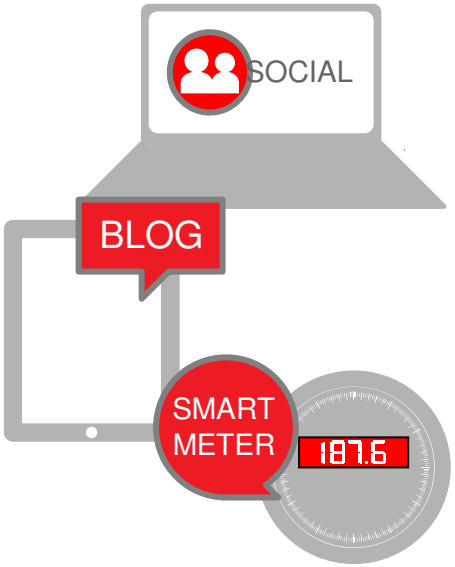
What makes it Big Data?



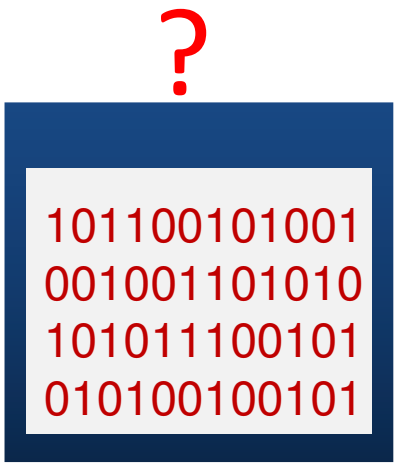
VOLUME



VELOCITY



VARIETY



VERACITY

V^4

Volume: Gigabyte(10^9), Terabyte(10^{12}), Petabyte(10^{15}), Exabyte(10^{18}), Zettabytes(10^{21})

Variety: Structured, semi-structured, unstructured; Text, image, audio, video, record

Velocity: Dynamic, sometimes time-varying

Veracity: Uncertainty of data

What is Big Data?

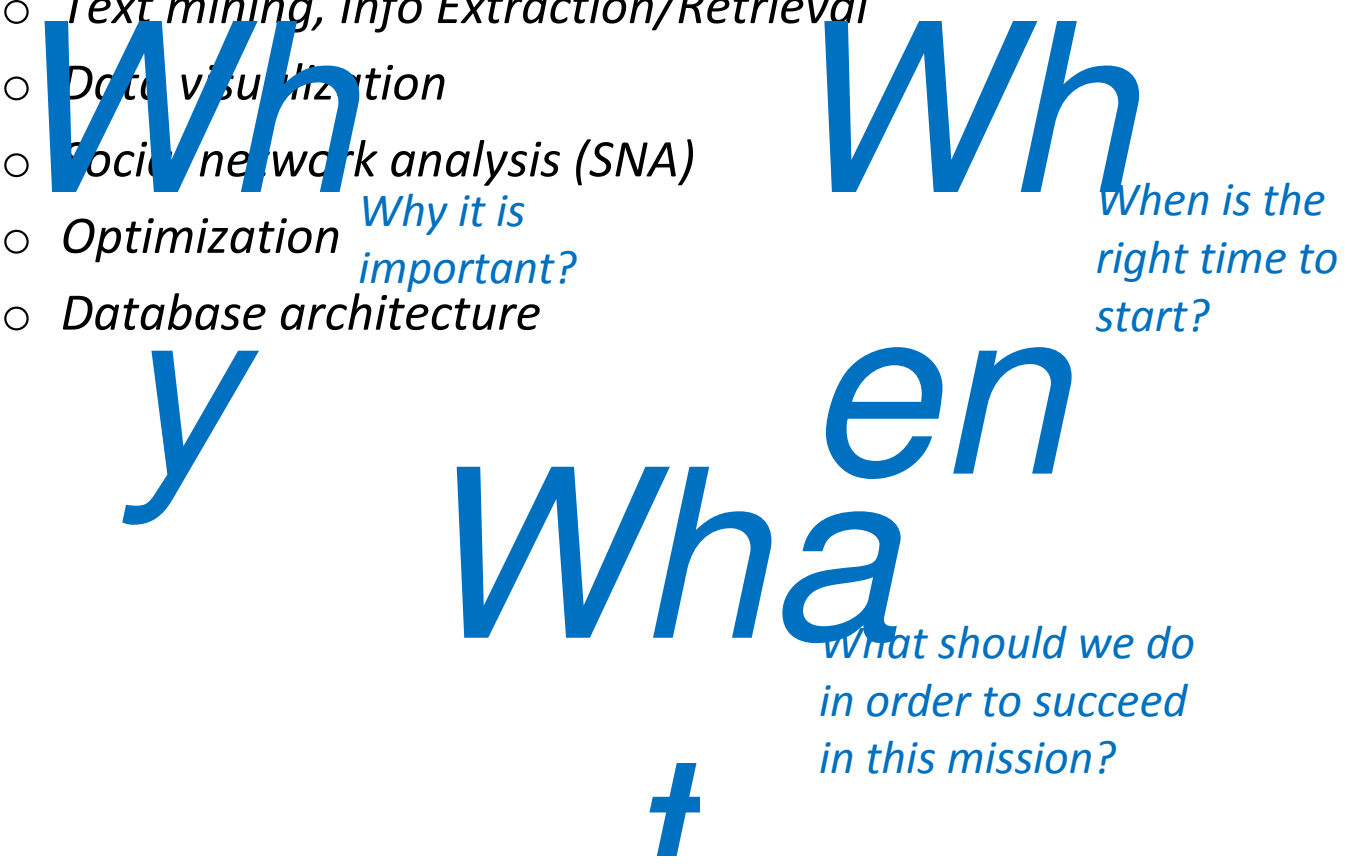
- High *volume*, *velocity* and/or *variety* information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation (Gartner, 2012)
- Large pools of *unstructured* and *structured* data that can be captured, communicated, aggregated, stored, and analyzed which are now becoming part of every *sector* and function of the global economy (Mckinsey, 2011)
- A new generation of *technologies* and *architectures*, designed to *economically* extract value from large volumes of a wide variety of data, by enabling high-velocity capture, discovery and analysis (IDC, 2011)

How can we get started
teaching a wide-ranging
interdisciplinary field that

1) has not established itself as
a solid & deep academic discipline?
2) is so clouded in hype?

The connection between Big Data & Data Science versus academia is undeniably evident

- *Statistics*
- *Machine learning*
- *Data mining*
- *Text mining, Info Extraction/Retrieval*
- *Data visualization*
- *Social network analysis (SNA)*
- *Optimization*
- *Database architecture*



Why

Big Data and Data Science Hype

- *The "Geek Chic" - in the headlines everywhere*
- ***"Data Scientist – The Sexiest Job of the 21st Century"*** (*HBR, 2013*)
- ***"Data Crunchers Now the Cool Kids on Campus"*** *(WSJ, 2013)*

Data Science Education

- *Students pursuing the field of Data Science*
- *To educate in what is likely to be a new profession in the industry*
- ***"…shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data…"*** *(McKinsey, 2011)*
- *To create effective workforce of big-data problem solvers*
- *To teach a set of skills that is highly on demand in business*

Why

Data Science has the potential to be a research discipline

- *There is actually a new field emerging that needs to be taught*
- *There is an opportunity to help define the new field*
- ***'Data Science: An action plan for expanding the technical areas of the fields of statistics'*** *(Bill Cleveland, 2001)*
- *There is a difference between statistics in industry and academia «clean data sets …?.»*

Data Science is not just happening in industry

- *Scholars struggle with large amounts of data and computation challenges*
  - *in diverse research fields as physics, biology, genomics, urban planning, medicine..*
- *They need access to tools developed in industry*
- *They need to be trained in best-practice and design patterns which emerged from the industry*

W

How Data Science is grappling with ethics….

hy

Corporations (Google, Facebook…. and government (NSA….)

Academia should

- ***Initiate ethics guidelines to be integrated into public policy***
- *Should take an active and leading role in educating students about «user-level data privacy concerns»:*
- *Sample data ethically - sensitive data should not be held at all*
- *Limit use - use of data for unspecified purposes should be forbidden*
- *Specify purpose - data that no longer serve a purpose should be destroyed*

When

New academic programs in Data Science

- *Universities are building related new courses and curricula*
- *New academic centers (e.g., NYU, Columbia, Stanford…)*

Ben-Gurion University

- *Recently launched 2 graduate (M.Sc.) programs*

Tel Aviv University

- *Recently launched graduate (M.Sc.) program*

When

BGU undergraduate elective: Big Data Analytics

- *Given last fall*
- *Enrolled 22 engineering students*
- *No prerequisites (other than mandatory IS courses)*

Challenges

- *Teaching a course where the topic is still getting defined in «both» industry and academia*
- *Mitigating the tension between industry and academia:*
  - *Chaotic landscape of tools and best practices versus established research fields*
- *Finding textbook or well defined body of knowledge*
- *Trying to figure out how to put all together*

Wh

Major central questions

- *What is a Data Science?*
- *What a Data Scientist does?*
- *What does Data Science mean?*
- *Is Big Data = Data Science?*
- *Is Data Science the science of Big Data?*

en

- ***Working definitions***
  - *Data Science:*

    the study of the space of problems solved with data and extreme-scale analytics
  - *Data Scientist (noun):*

    better at statistics than any software engineer and

    better at software engineering than any other statistician

    Josh Wills

Wh

Big Data Analytics – Data Scientist tasks

Business Insights

Data-driven decision making en

Building data models

Visualization, Dashboards KPIs

Statistics (basic) , Correlations

Programing, Map-Reduce transformation

Schema less non-relational data bases

Data mining & machine learning, text mining, IE, NLP, IR

Graphs mining and large scale networks analytics

Visualization techniques

Story telling…

…….

Course final project – TM DM in R and Rattle, SNA (Gephi – Open Source)

My DS definitions keep expanding toward Version 2.0…

Wh at

The most challenging aspect …

- *Diversity of expertise and knowledge required*
- *No one person can be the project Data Scientist → team work*
- *Interdisciplinary teams with mixture of backgrounds and skills*
- *Learning environment: ideal for training next generation Data Scientists?*

The traditional teaching environment should be revised

- *Modify and update the concept of lecturers as sole knowledge providers to knowledge mediators*
- *Co-lecture and co-teach with academic experts in specific related fields*
- *Co-operate with industry experts and software vendors*

Wh

The traditional learning environment should also be revised

at

- *Educate students to be consumers of knowledge*
- *Expand the learning experience with related online courses (MOOC)*
  - *Coursera, MIT Open Courseware, Stanford online courses, Kahn Academy…*
- *Establish*
  - *interactive user forums and a collaborative community learning environment for students, teaching assistants, and lecturers from all over the world*
  - *Self-pace learning environment (e.g.. R on the web)*

From sage on the stage to guide on the side *(Alison King)*

It is necessary to quickly modify and update our core curriculum and courses to provide students the skill sets needed to become data scientists

Q&A

sasson.elan@gmail.com