Introduction to Data Science

Sam Oh & Wonhong Jang
SKKU iSchool

What is Data?

- factual information (as measurements or statistics) used as a basis for reasoning, discussion, or calculation

- information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

- information in numerical form that can be digitally transmitted or processed

(Source: Merriam-Webster Dictionary)

What is Data?

- Data is a set of values of qualitative or quantitative variables; restated, pieces of data are individual pieces of information. Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs or images.

- Data as an abstract concept can be viewed as the lowest level of abstraction, from which information and then knowledge are derived.

(Source: Wikipedia)

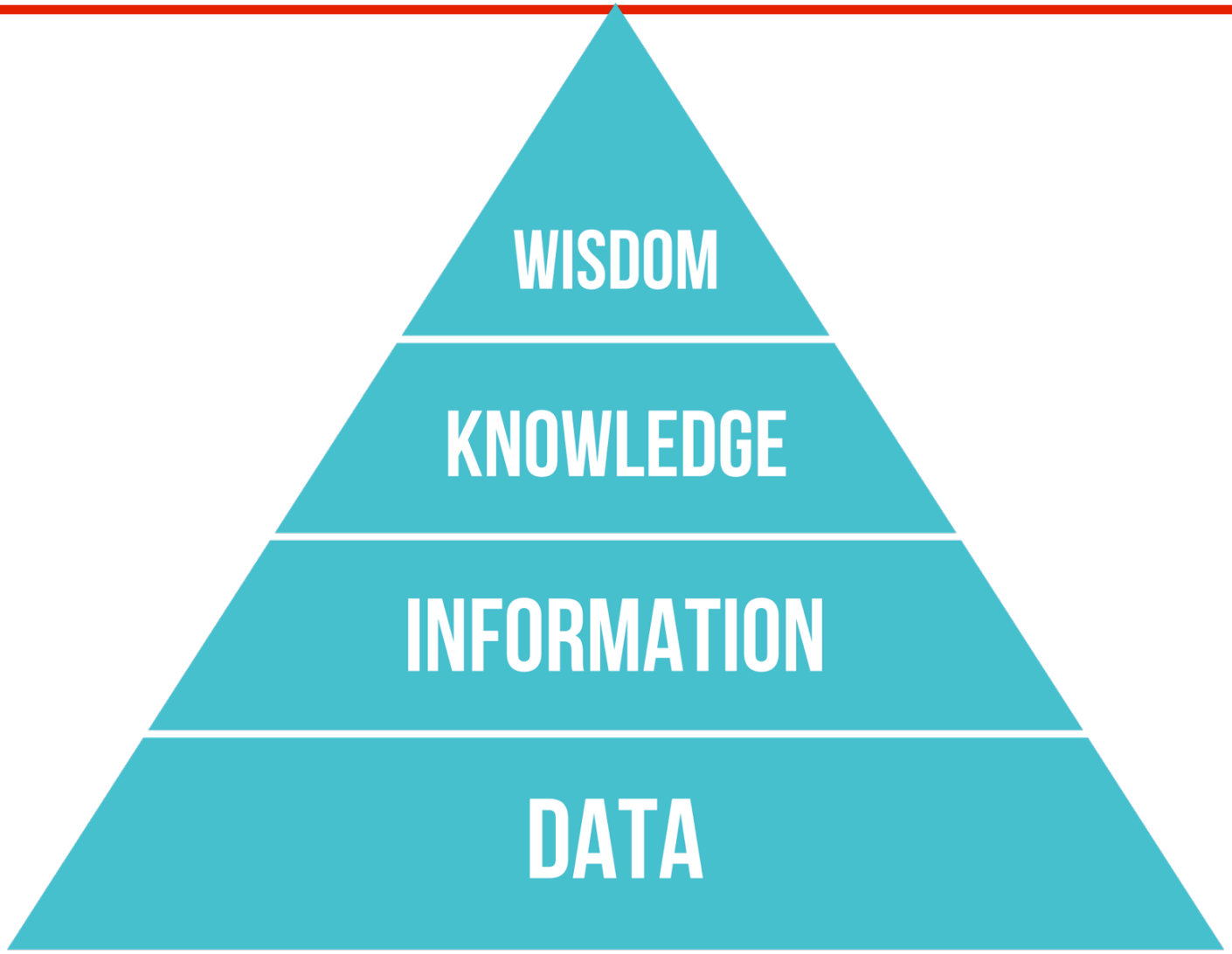
What is Raw Data?

- Raw data is a term for data collected from a source
- Raw data has not been subjected to processing or any other manipulation, and are also referred to as primary data

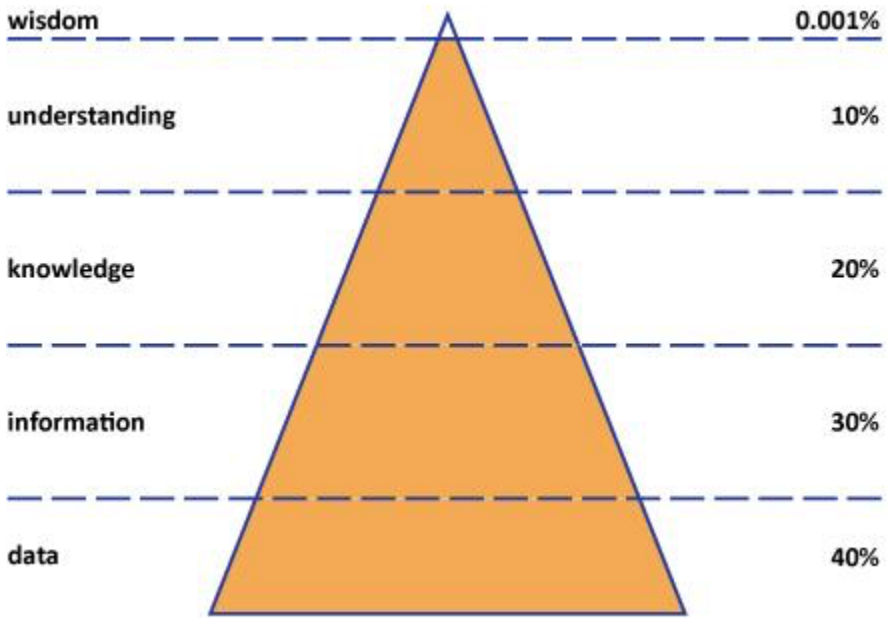
(Source: Wikipedia)

Sir Tim Berners-Lee: Raw data, now! (Wired)

Data to Wisdom



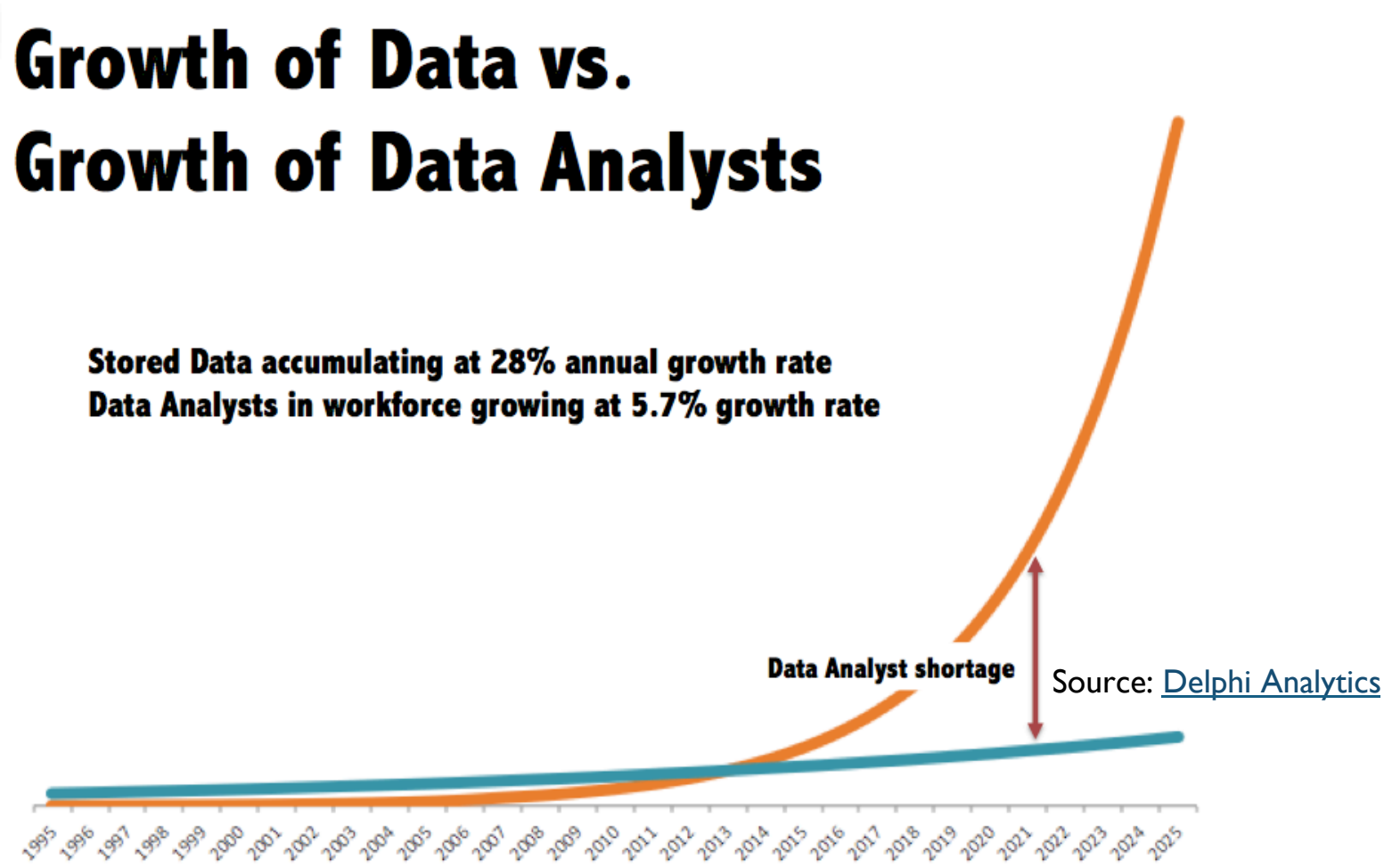
Data to Wisdom



The volume of data, information, knowledge, etc in an organisation

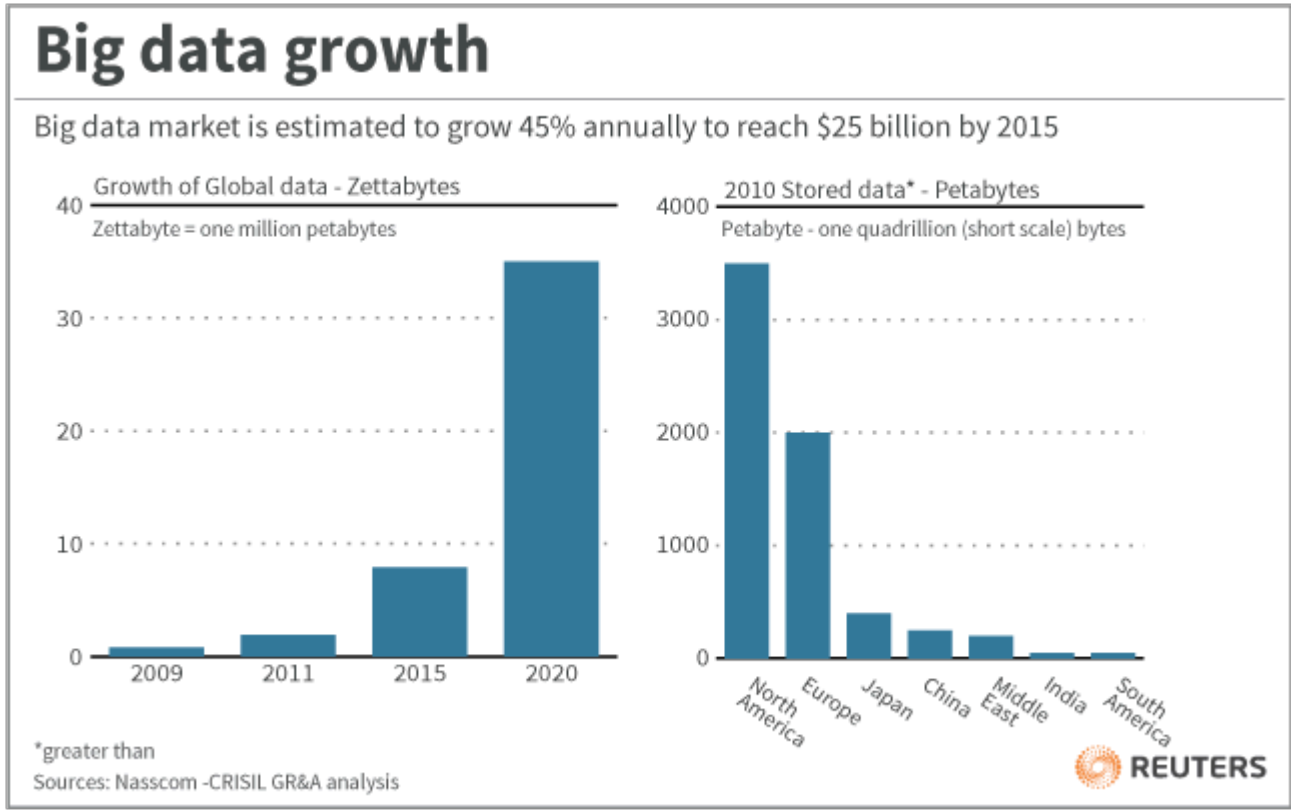
(Source: Russell L. Ackoff, "From data to Wisdom", 1990)

Why Data Science Matters?



Source: Delphi Analytics

Why Data Science Matters?



Reuters graphic/Catherine Trevethan 05/10/12

What is Big Data?

- Big data essentially means datasets that are too large for traditional data processing systems, and therefore require new processing technologies

(Source: Data Science for Business)

- Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured

(Source: SAS)

- Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate

(Source: Wikipedia)

Why Big Data?

- The global big data market will reach $46.34 billion by 2018

(Source: http://www.marketsandmarkets.com/)

- 90% of the data in the world today has been created in the last two years alone

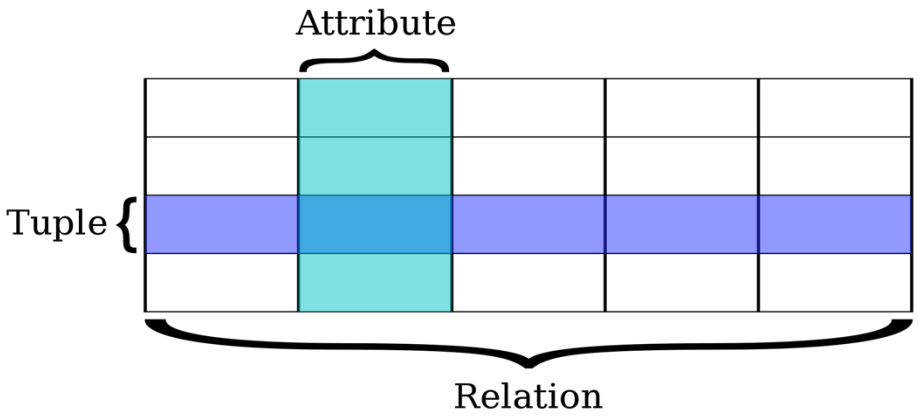
(Source: "If you think Big Data's Big Now, Just Wait", Ron Miller, TechTrends, 2014)

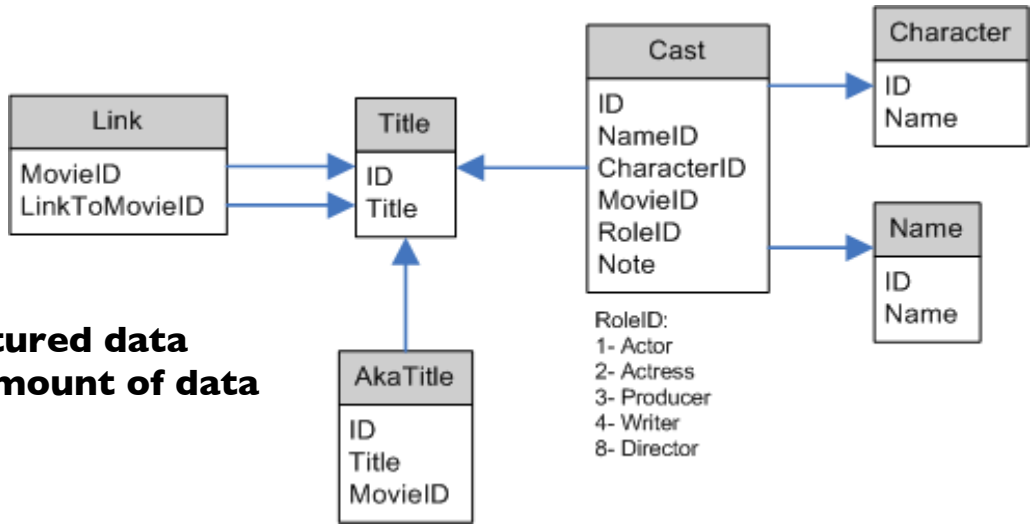
Why are the Problems of Big Data?

- "Through 2015, more than 85% of Fortune 500 organizations will fail to effectively exploit Big Data for competitive advantage."

(Source: Gartner)

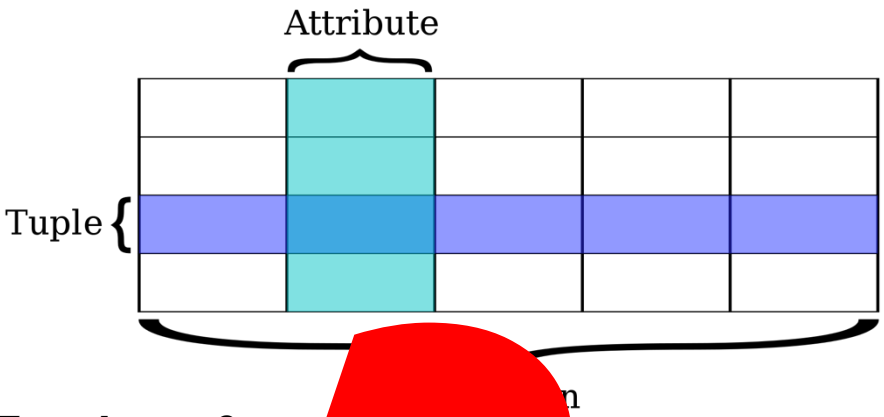
Traditional Data Processing Method



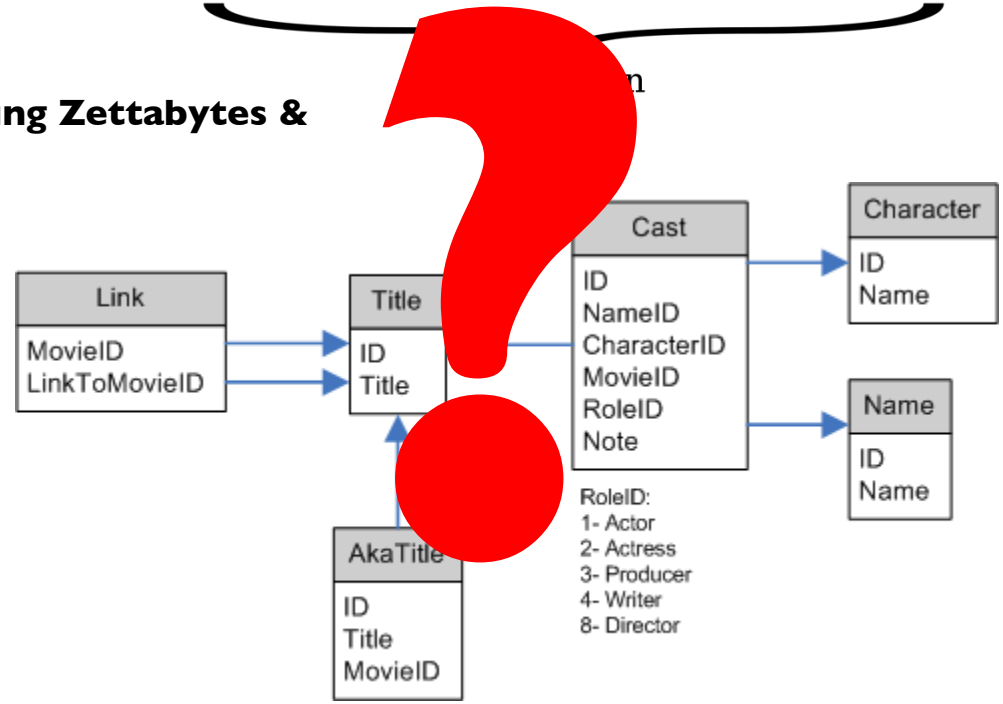


Mainly handled structured data
Limited in handling amount of data

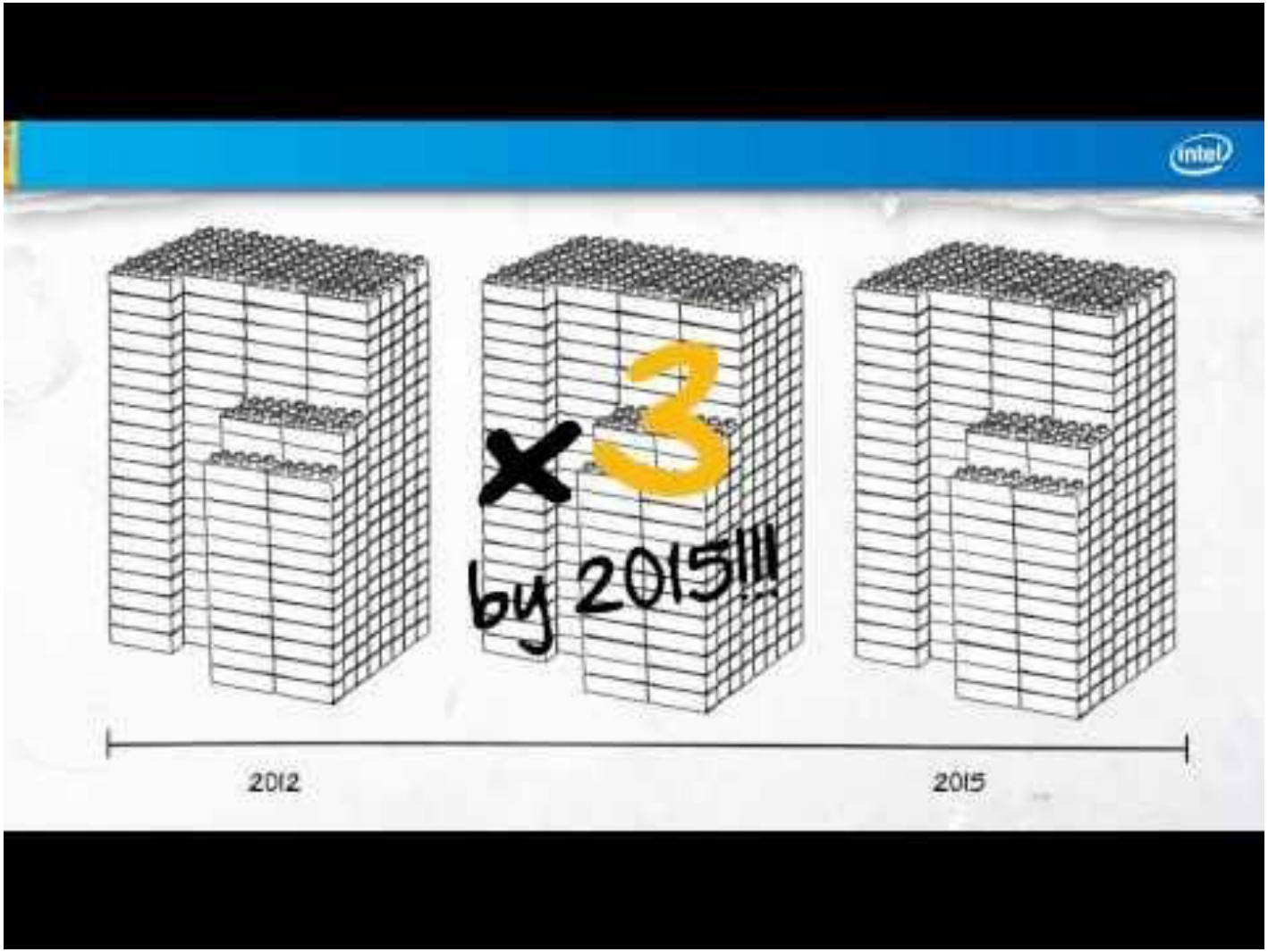
Traditional Data Processing Method



Not efficient in handling Zettabytes & unstructured data

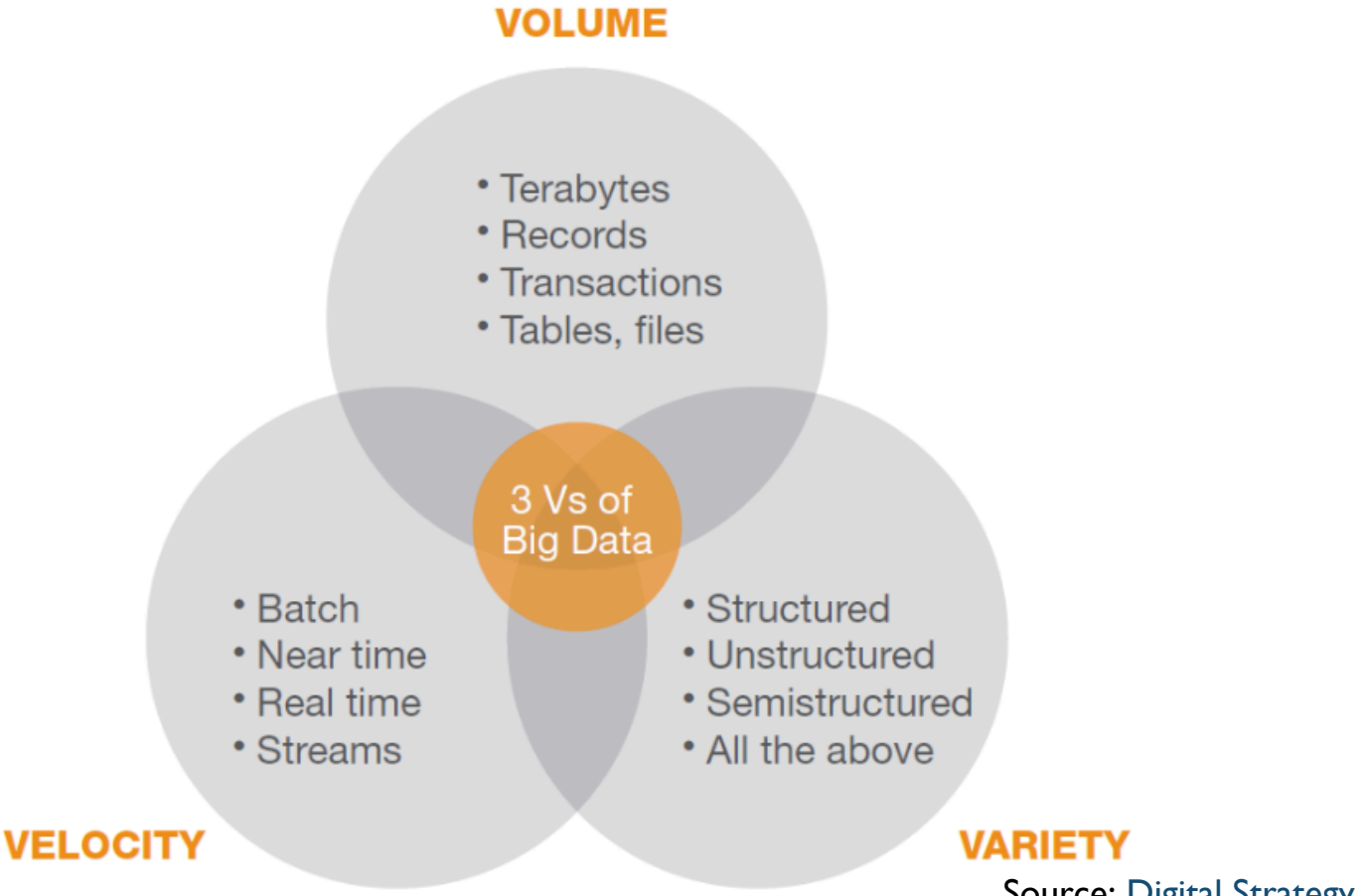


What is Big Data? (by Intel)



Characteristics of Big Data

- **3Vs of Big Data**



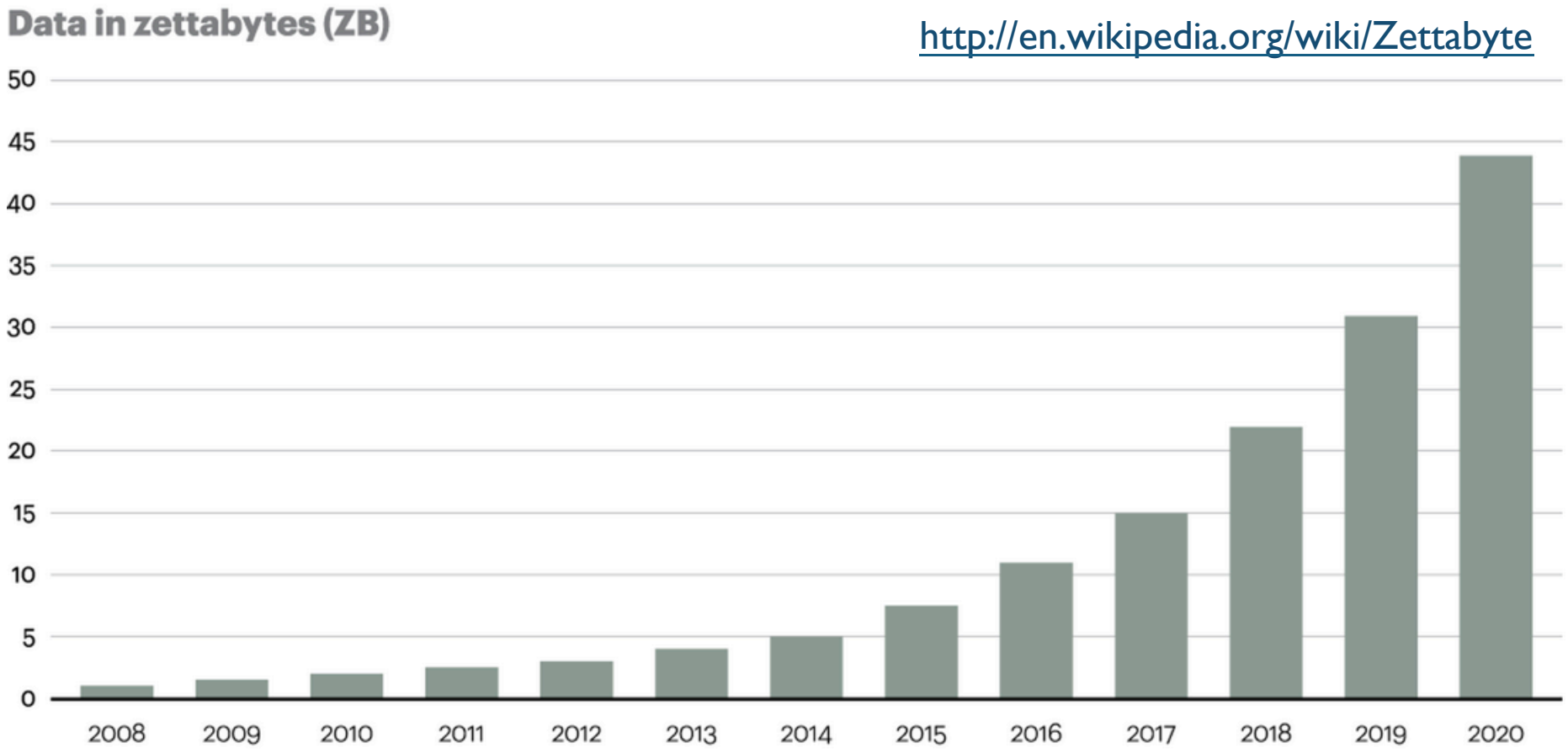
Source: Digital Strategy

Characteristics of Big Data

- **3Vs of Big Data**
  - **Volume**: Many factors contribute to the increase in data volume
  - **Velocity**: Data is streaming in at unprecedented speed and must be dealt with in timely manner
  - **Variety**: Data today comes in all types of formats

Source: SAS

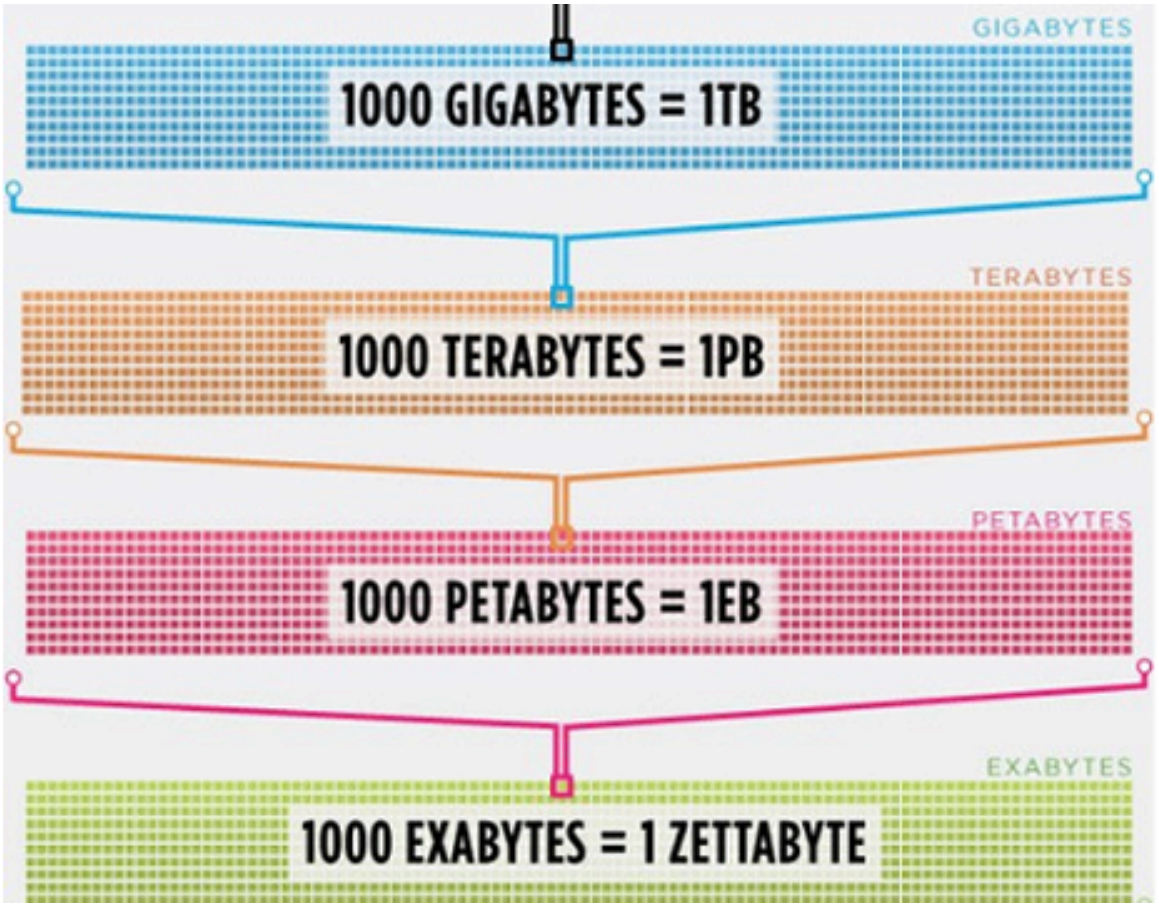
Big Data - Volume



http://en.wikipedia.org/wiki/Zettabyte

Source: Oracle, 2012

Big Data - Volume

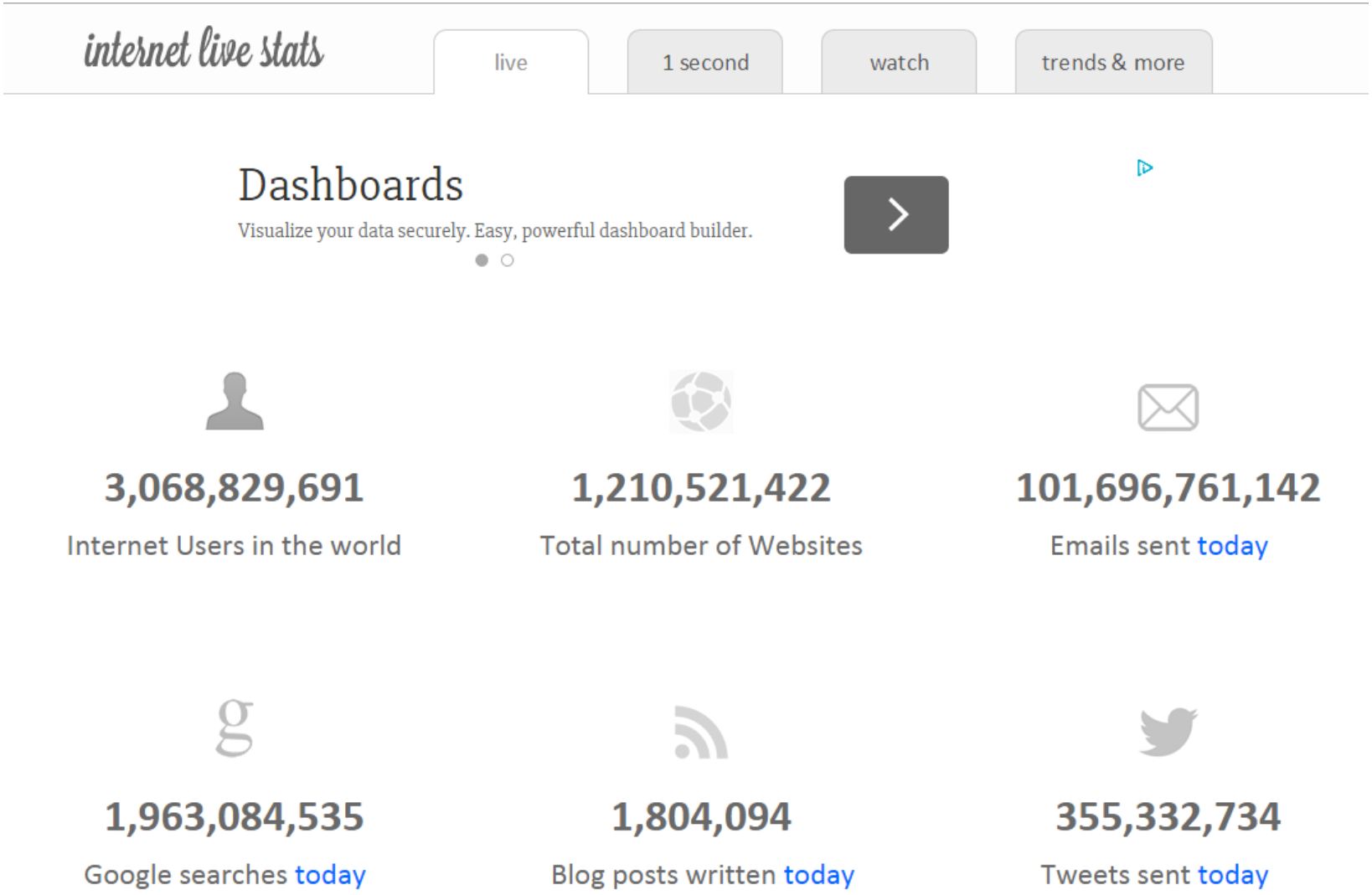


Big Data - Velocity

Source: Ravi Kalakota

http://www.internetlivestats.com/

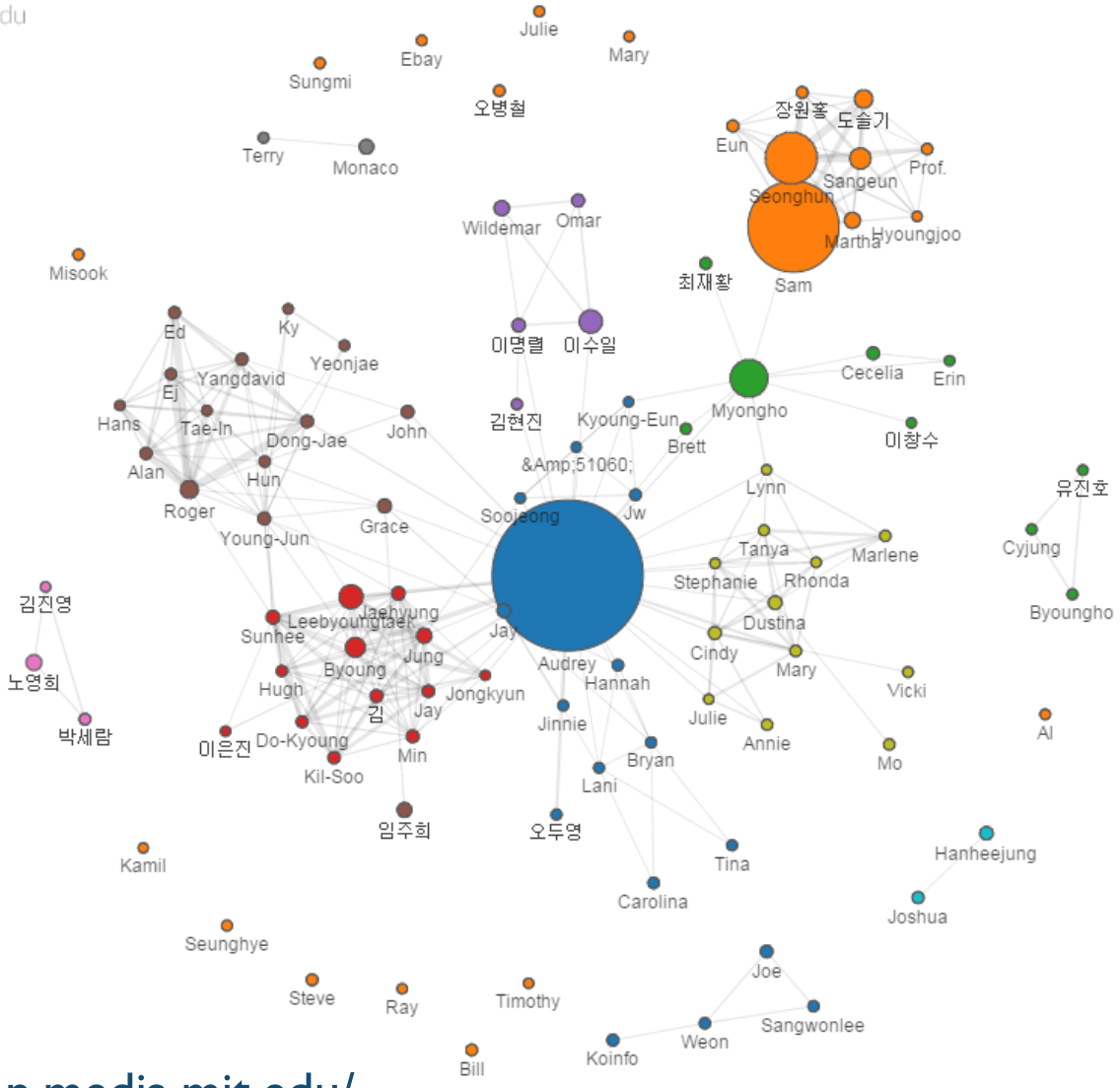
8,621 Tweets sent in 1 second



http://www.internetlivestats.com/

Personal Email (8 years) Analysis

immersion.media.mit.edu



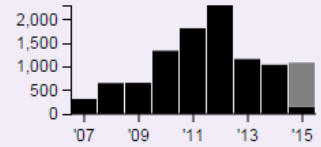
Joseph Lee

467 collaborators

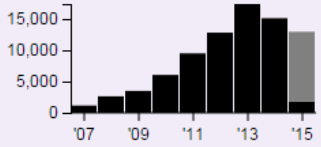
79,949 emails

My Stats | Top Collaborators

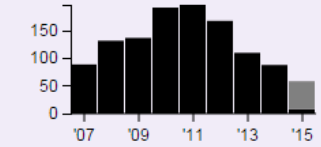
Emails Sent



Emails Received



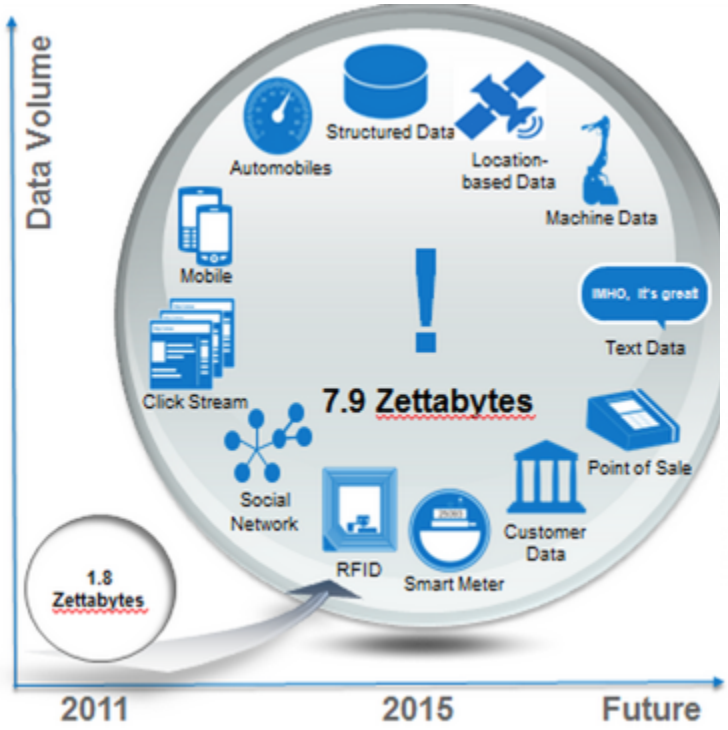
New Collaborators

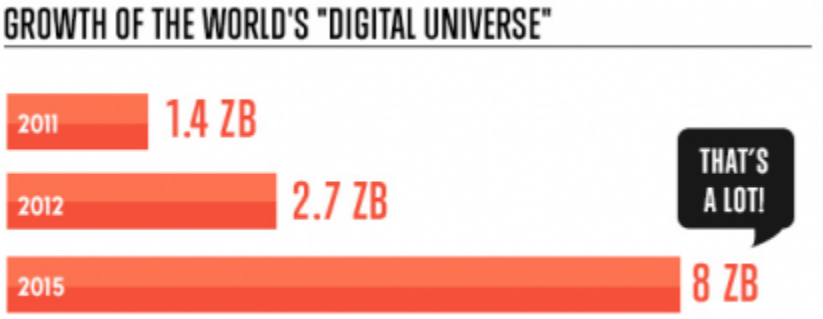


https://immersion.media.mit.edu/

8.0 years
23 Feb 2007 - 19 Feb 2015

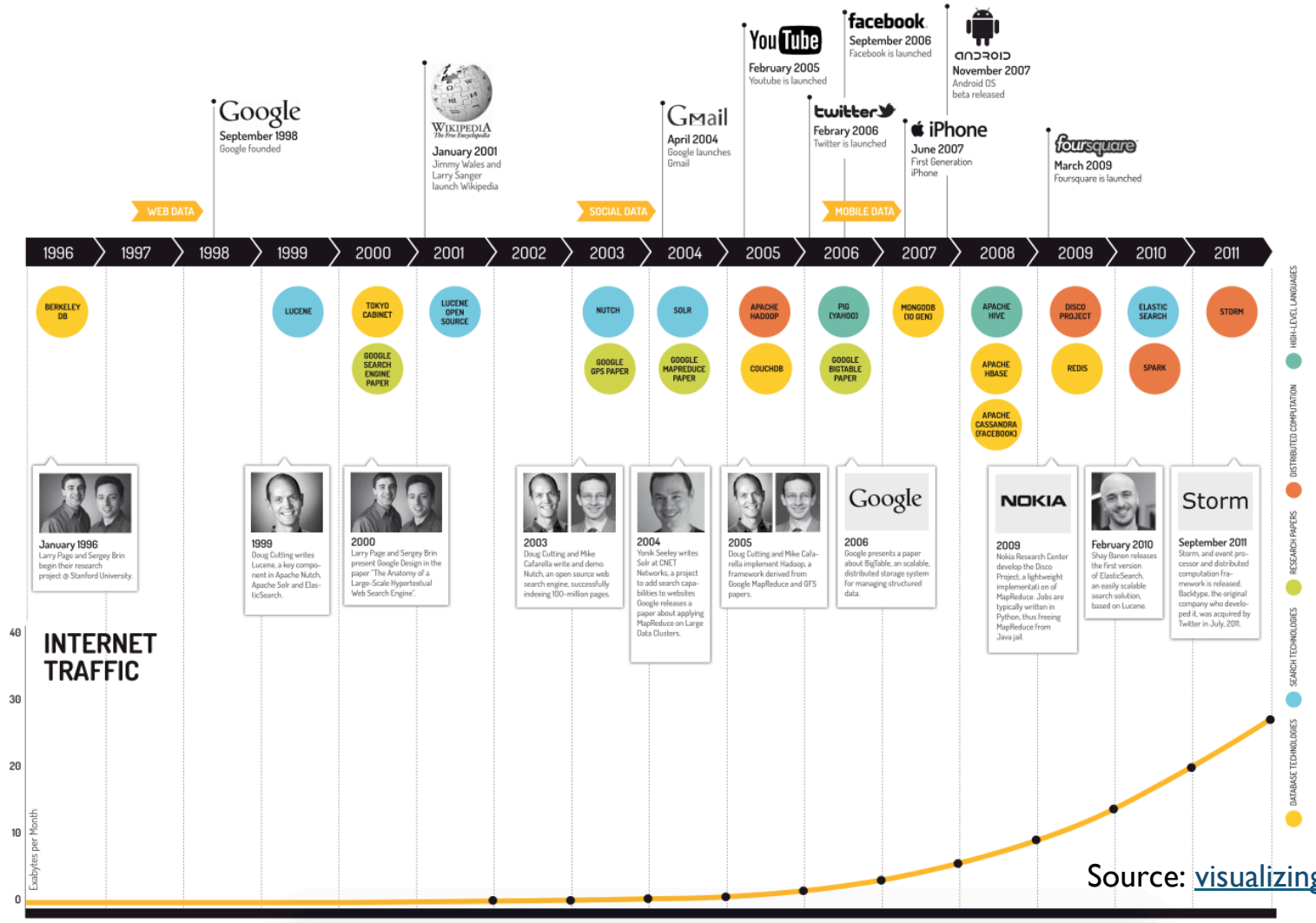
Big Data - Variety





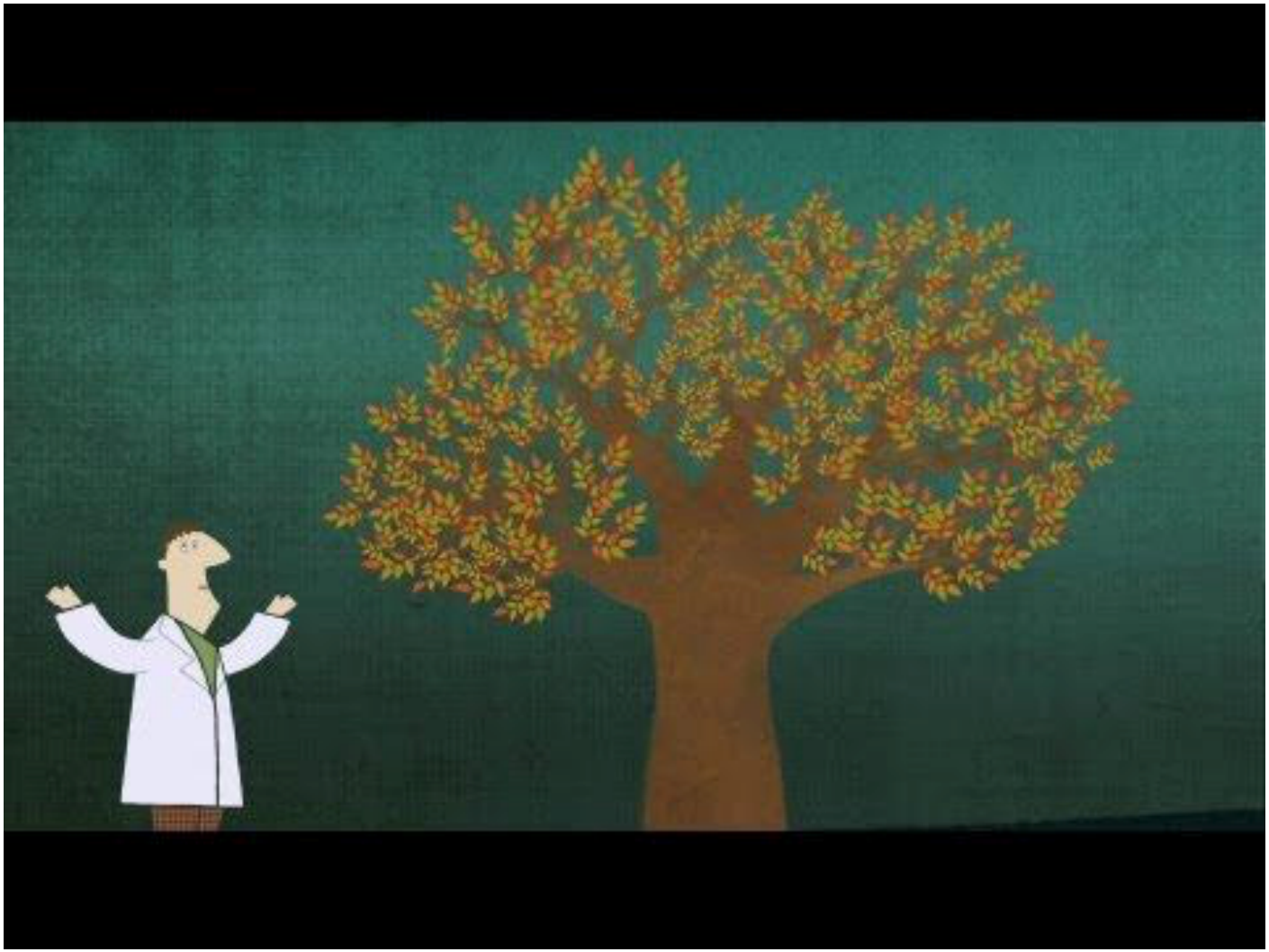
Source: Rose Business Technologies

History of Big Data



Source: visualizing.org

History of Big Data (by TED)



Big Data Sets Example (#1 in Cloud Market)

Active Big Data Industries

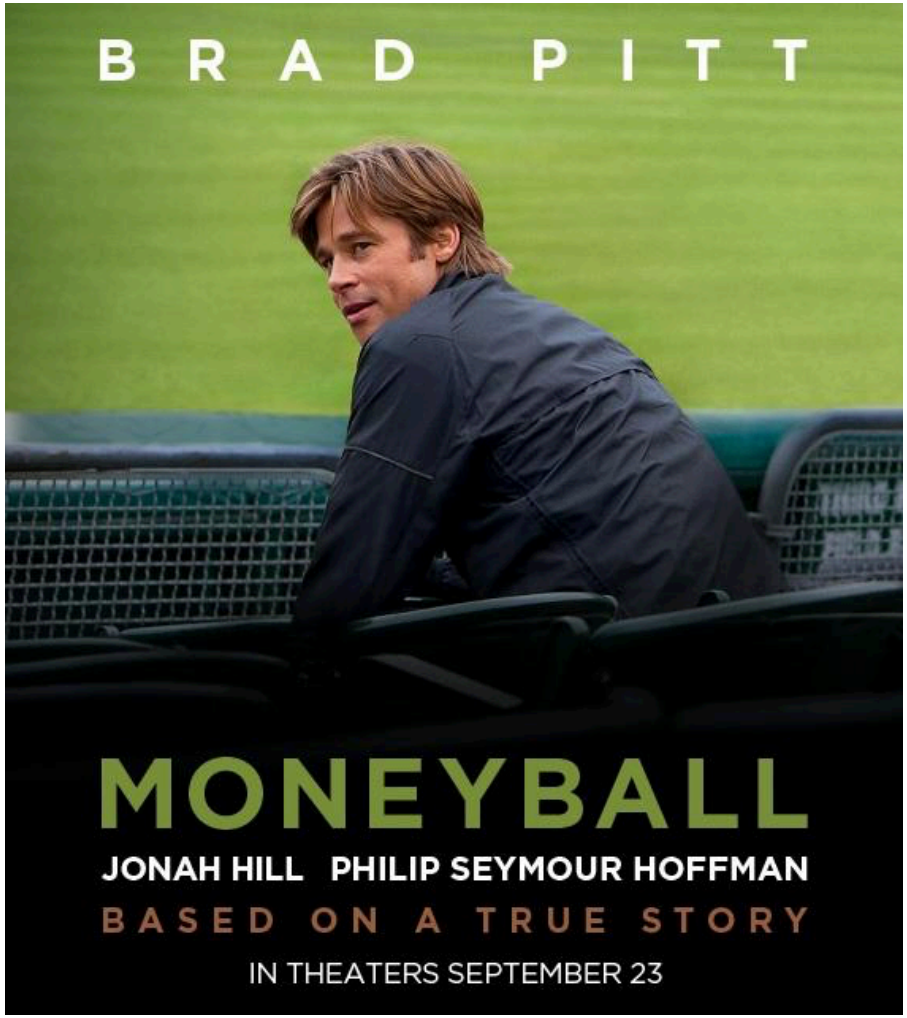
- Retail
- Telecommunications
- Consulting
- Healthcare
- Air transportation
- Construction
- Food products
- Steel and Manufacturing in general
- Industrial instruments
- Automobile industry
- Customer care
- Publishing
- Logistics

Big Data Use Cases

Google was able to spot trends in the Swine Flu epidemic roughly two weeks before the Center for Disease Control by analyzing searches that people were making in different regions of the country.

Source: Mike Loukides, "What is data science ?"

Big Data Use Cases



Big Data Use Cases

Google Books

About Google Books | General Help | Partner Program | **Library Project** | Perspectives

Library Partners | Screenshots | Librarian Help

Google Books Library Project – An enhanced card catalog of the world's books

We're working with several major libraries to include their collections in Google Books and, like a card catalog, show users information about the book, and in many cases, a few snippets – a few sentences to display the search term in context.

What does a Google Books Library Project book look like?

When you click on a search result for a book from the Library Project, you'll see basic bibliographic information about the book, and in many cases, a few snippets – a few sentences showing your search term in context. If the book is out of copyright, you'll be able to view and download the entire book. In all cases, you'll see links directing you to online bookstores where you can buy the book and libraries where you can borrow it.

Full View

View the entire book

Limited View

View a limited number of pages

Snippet View

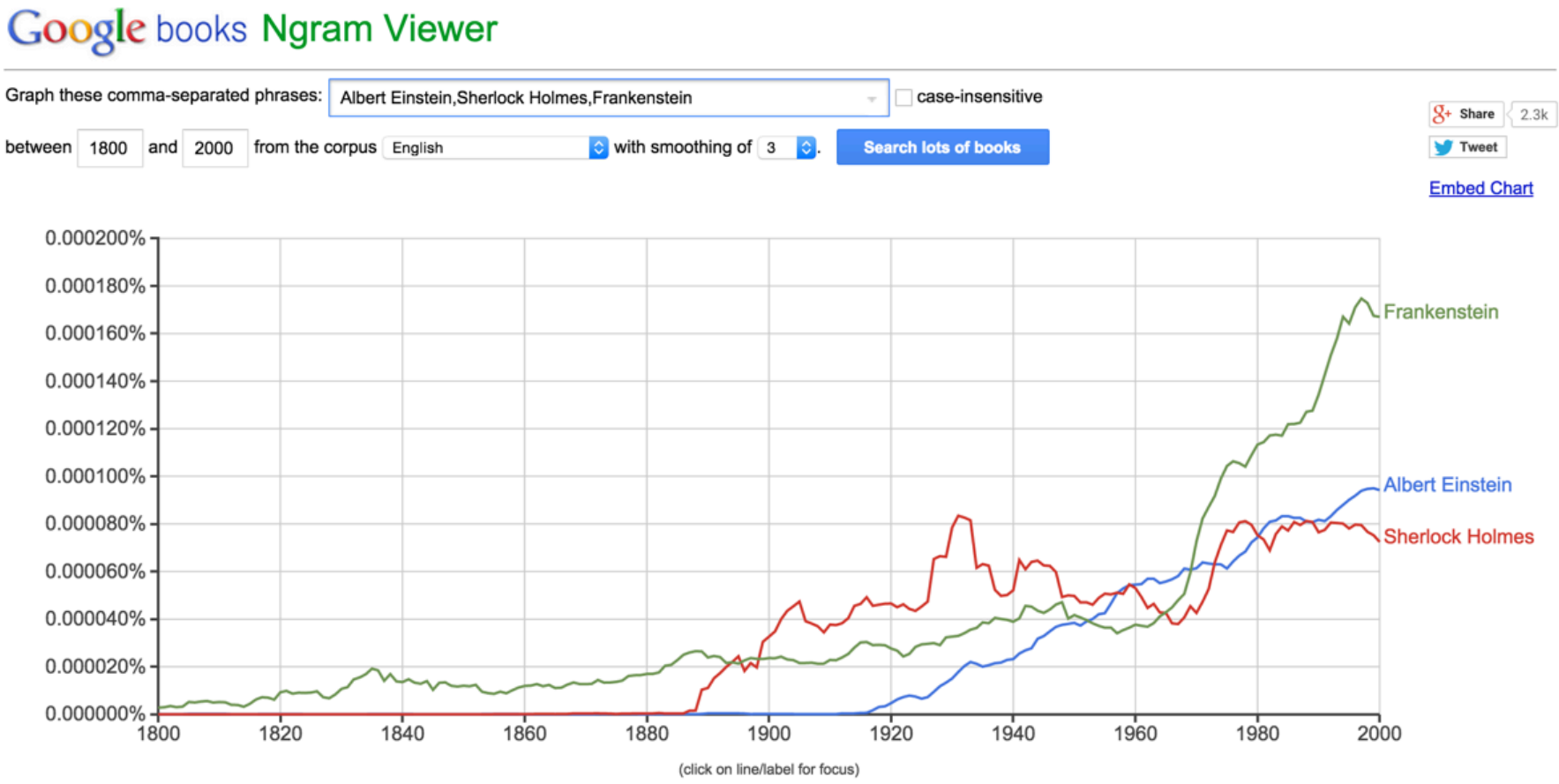
View a few sentences surrounding the search term

No Preview Available

View basic information about the book

Big Data Use Cases

Google Ngram Viewer



Big Data Use Cases

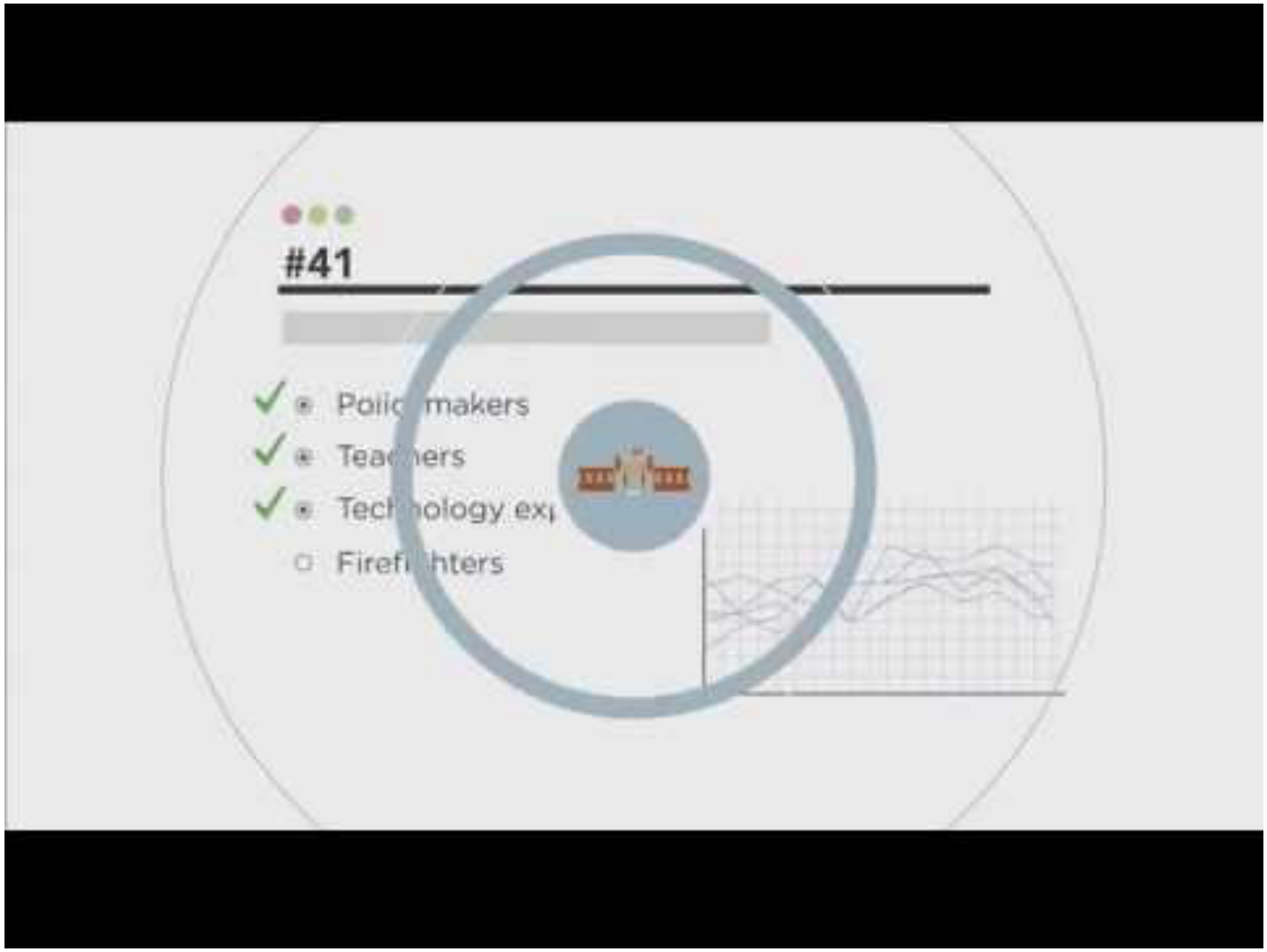
By World Future Forum

Big Data Use Cases

Daily Life Related

Big Data Use Cases

Education Related



Big Data Use Cases

Source: Google

Raoul Rañoa / @latimesgraphics

Big Data Use Cases

What is Data Science?

- A recent term that has multiple definitions but is generally accepted as a discipline that incorporates statistics, data visualization, computer programming, data mining, machine learning and database engineering to solve complex problems

  (Source: Data Scientist - The definitive guide to becoming a data scientist)

- Data science is the extraction of knowledge from data

  (Source: Wikipedia)

What is Data Science?

- The study of the generalizable extraction of knowledge from data

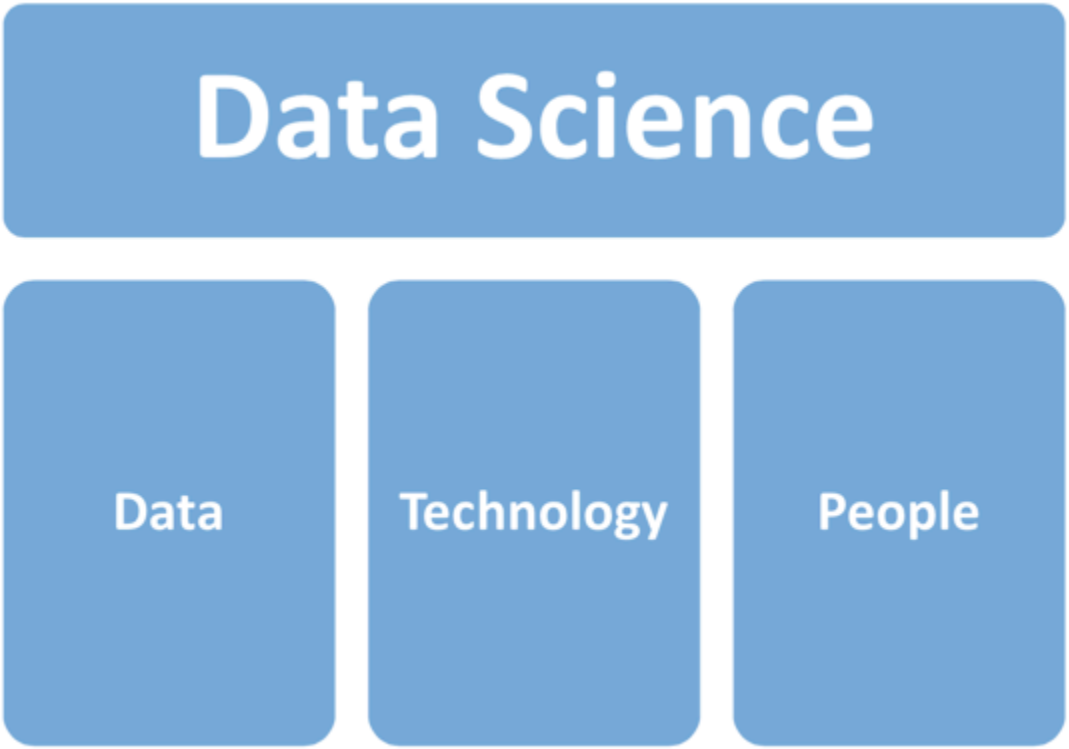
(Vasant Dhar, 2013)

- Data science involves principles, processes, and techniques for understanding phenomena via the automated analysis of data

(Provost & Fawcett, 2013)

What is Data Science?

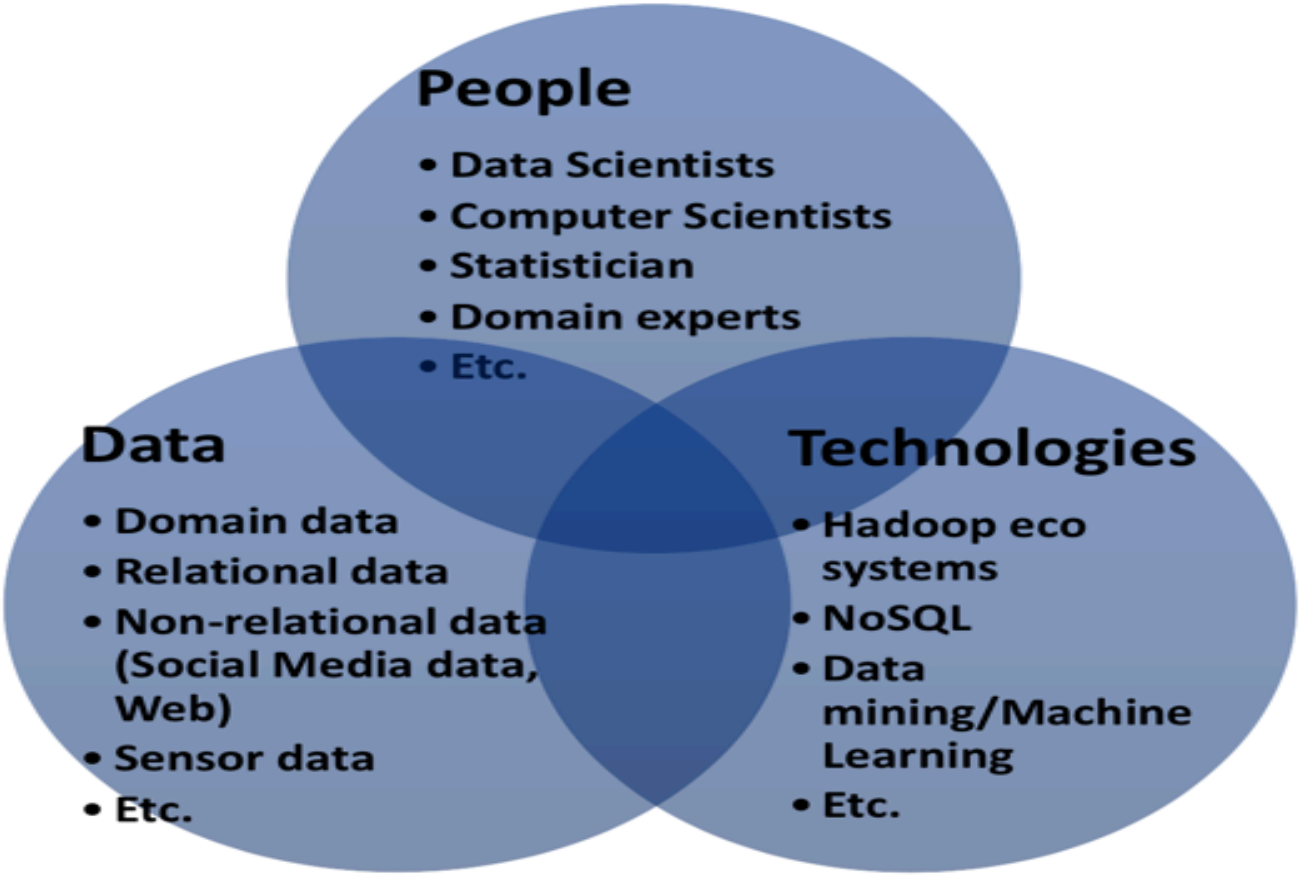
- Three pillars of Data Science



(Source: Il-Yoel Song, 2014)

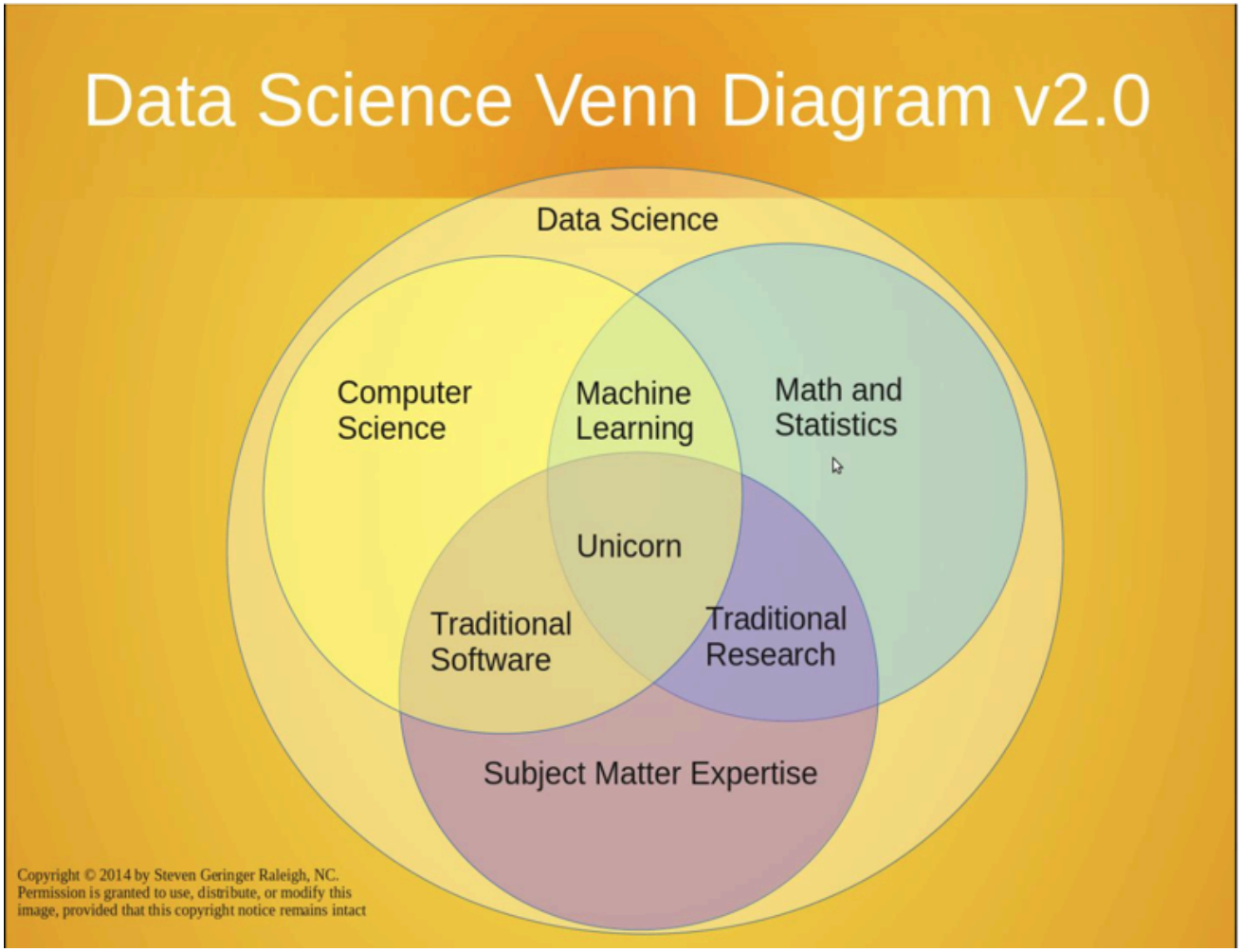
What is Data Science?

- Three pillars of Data Science



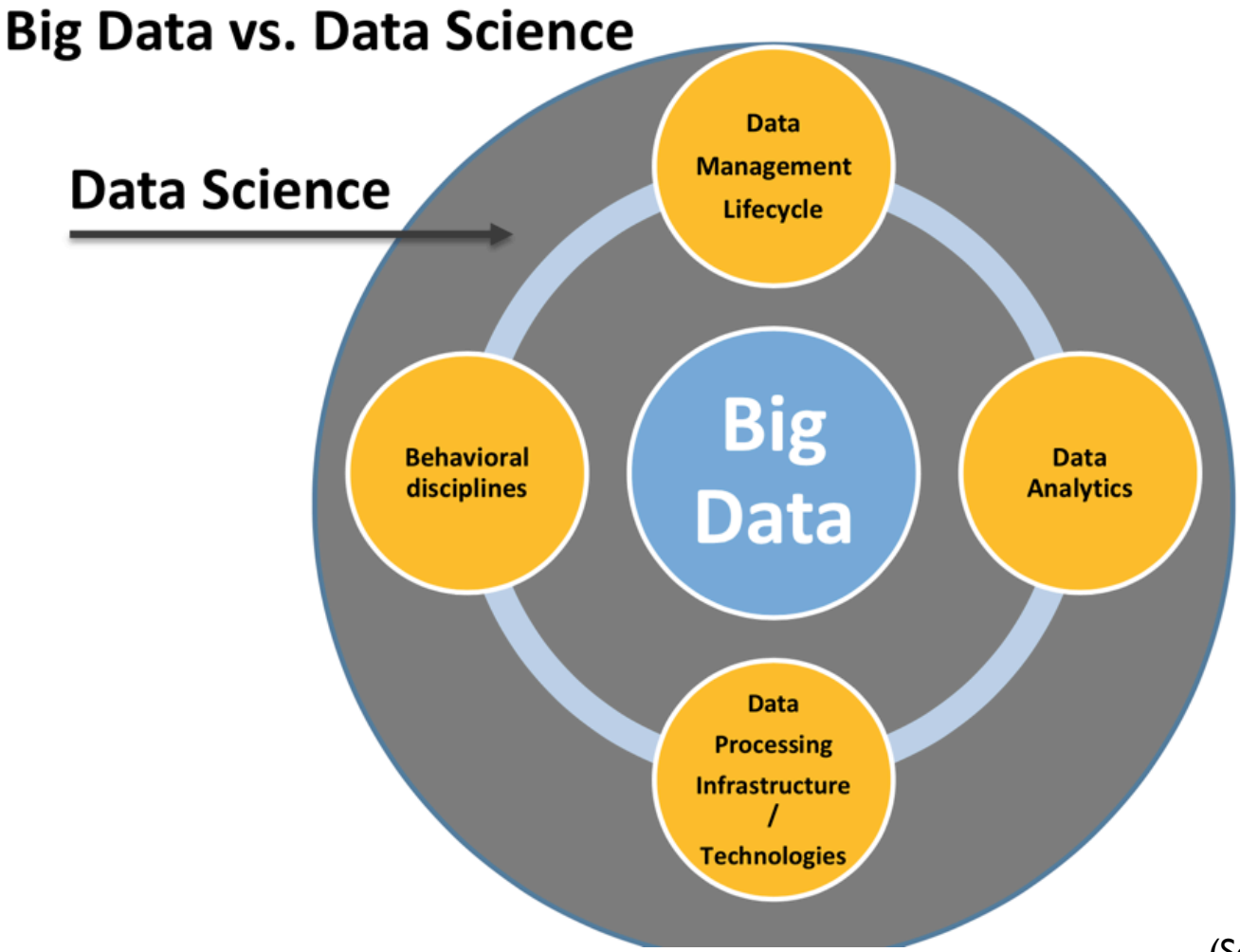
(Source: Il-Yoel Song, 2014)

What is Data Science?



(Source: Steve Geringer, 2014)

Data Science vs Big Data



(Source: Il-Yoel Song, 2014)

What is Data Scientist ?

- A data scientist is an individual, organization or application that performs statistical analysis, data mining and retrieval processes on a large amount of data to identify trends, figures and other relevant information

(Source: Techopedia)

What Data Scientists Do?

- Extract useful knowledge from data to solve business problems
- Get the right requirements
- Ask the right question with business goals and metrics in mind
- Explore solution spaces iteratively without pre-determined end in mind
- Work with domain experts

(Source: Il-Yoel Song, 2014)

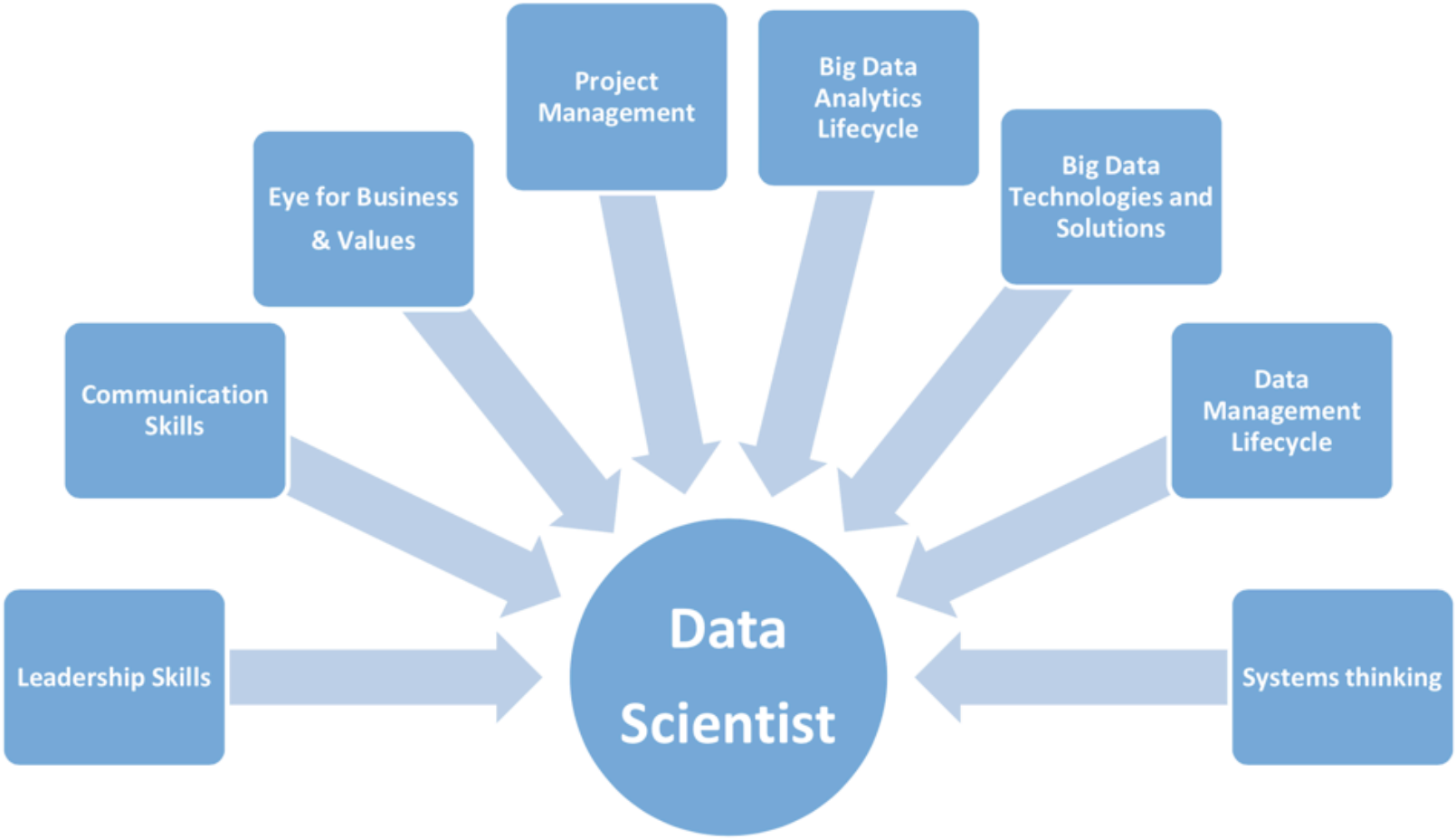
What Data Scientists Do?

- Extract relevant data, use/reuse/merge data
  - Separate noise from signals in big data
- Select right technologies and tools
- Perform analytics, evaluate, and visualize
  - Discover hidden insights
  - Improve decision-making
  - Automate business processes
- Automate the data-driven decision-making

(Source: Il-Yoel Song, 2014)

What Data Scientists Do?

Who is a Data Scientist?



(Source: Il-Yoel Song, 2014)

How to become a Data Scientist?

- Data scientists need to know how to code
- Data scientists need to be comfortable with mathematics & statistics
- Data scientists need to know machine learning & software engineering

❖ **Learning data science can be really hard**

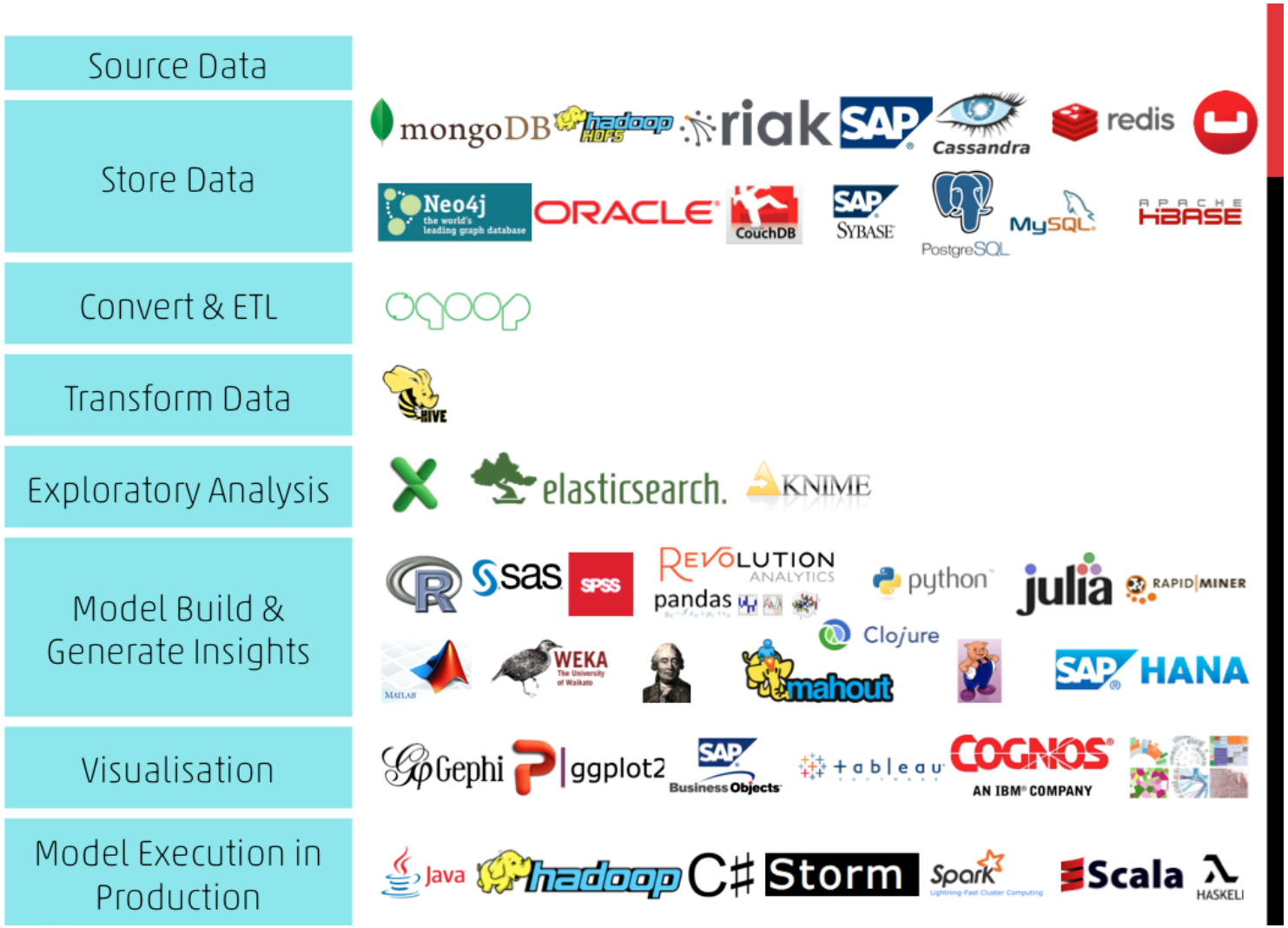
Source: Ryan Orban, Slideshare

Data scientist toolkit

- Java, R, Python, Clojure, Haskell, Scala…
- Hadoop, HDFS&MapReduce, Spark, Storm…
- HBase, Pig&Hive, Shark, Impala, Cascalog…
- ETL, Webscrapers, Flume, Sqoop, Hume…
- SQL, RDBMS, DW, OLAP…
- Knime, Weka, RapidMiner, SciPy, NumPy…
- D3.js, Gephi, Tableu, Flare, Shiny, ggplot2…
- SAS, SPSS, Matalab…
- NoSQL, MongoDB, Couchbase, Cassandra…
- And more...

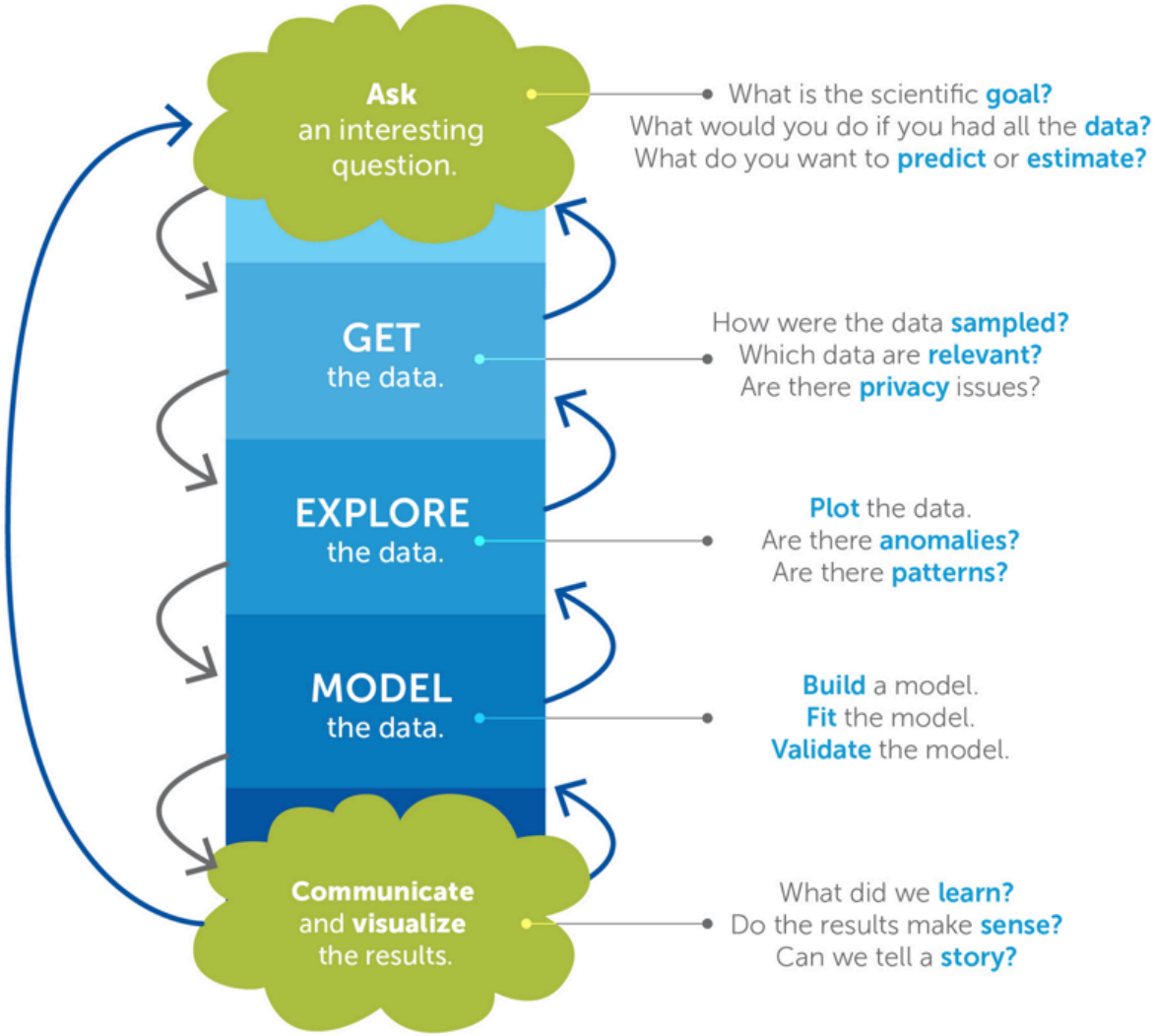
Data Scientist Toolkit

Data Scientist Toolkit



Data Scientist Interview

Data Science Process



(Source: Opera solutions)

What is Data Product ?

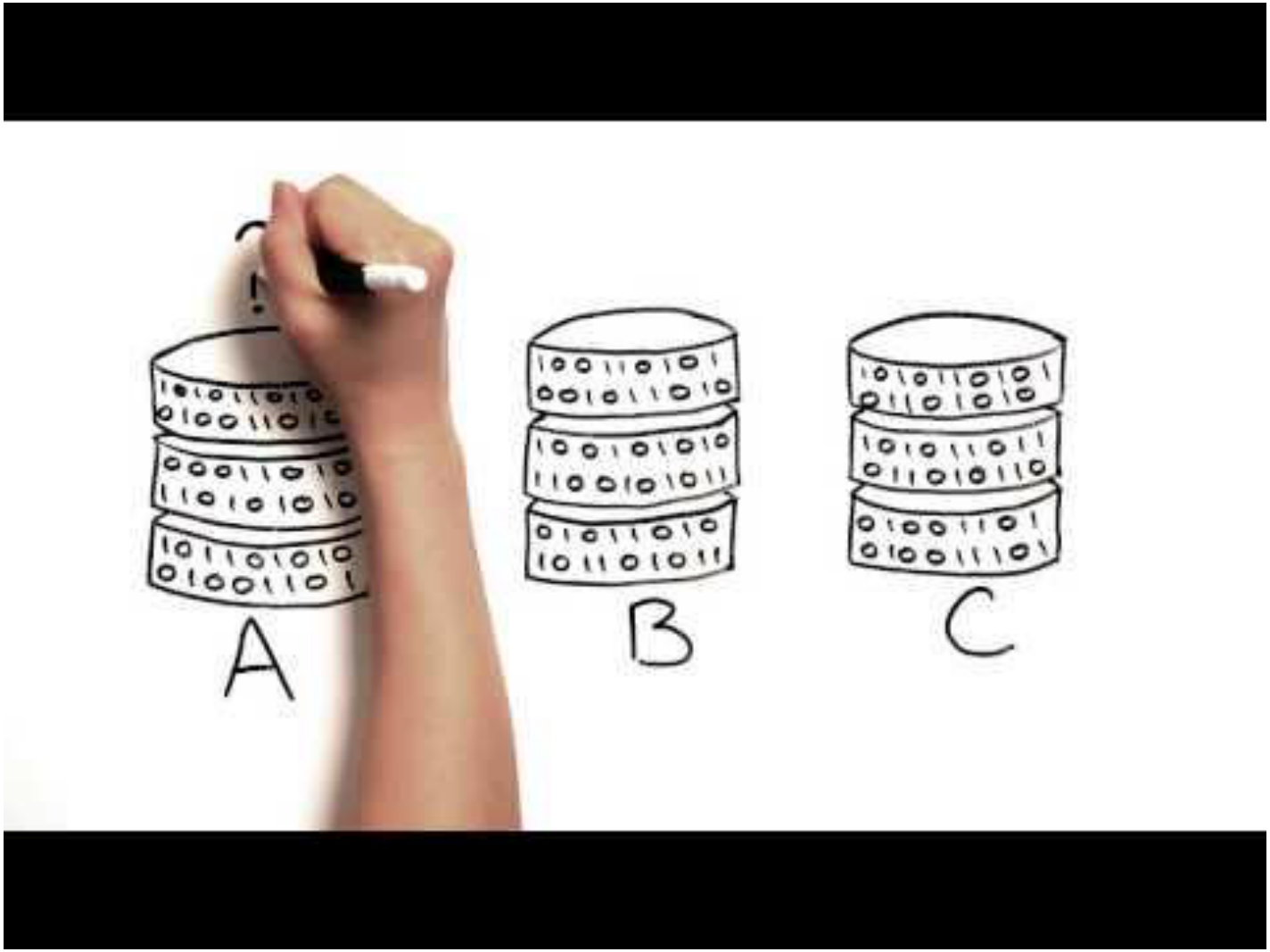
A data product is the production output from a statistical analysis. Data products automate complex analysis tasks or use technology to expand the utility of a data informed model, algorithm or inference

(Source: coursera.org)

Data Product examples

- References
  - The 10 Coolest Big Data Products Of 2014
  - 16 Top Big Data Analytics Platforms

Data Analytics



Q & A

Sir Tim Berners-Lee: The next Web of open, linked data (TED)