
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?

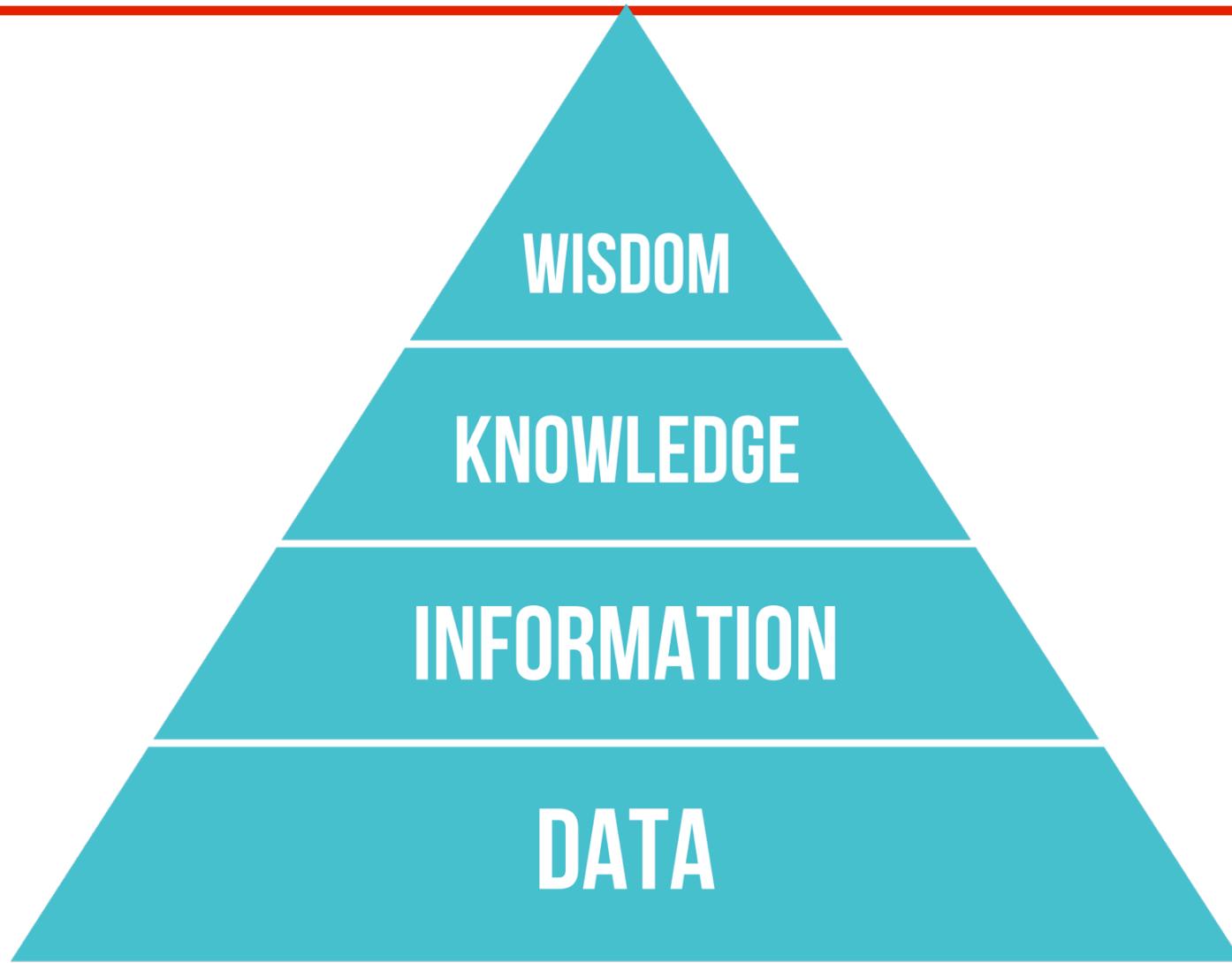
- Raw data is a term for data collected from a source
- Raw data has not been subjected to processing or any other manipulation, and are also referred to as primary data

(Source: Wikipedia)

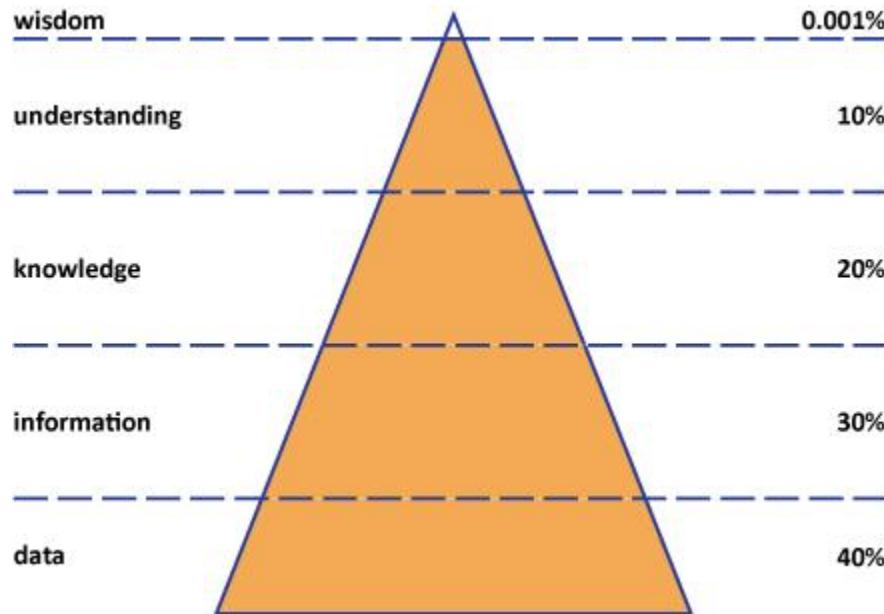


[Sir Tim Berners-Lee: Raw data, now! \(Wired\)](#)

Data to Wisdom



Data to Wisdom



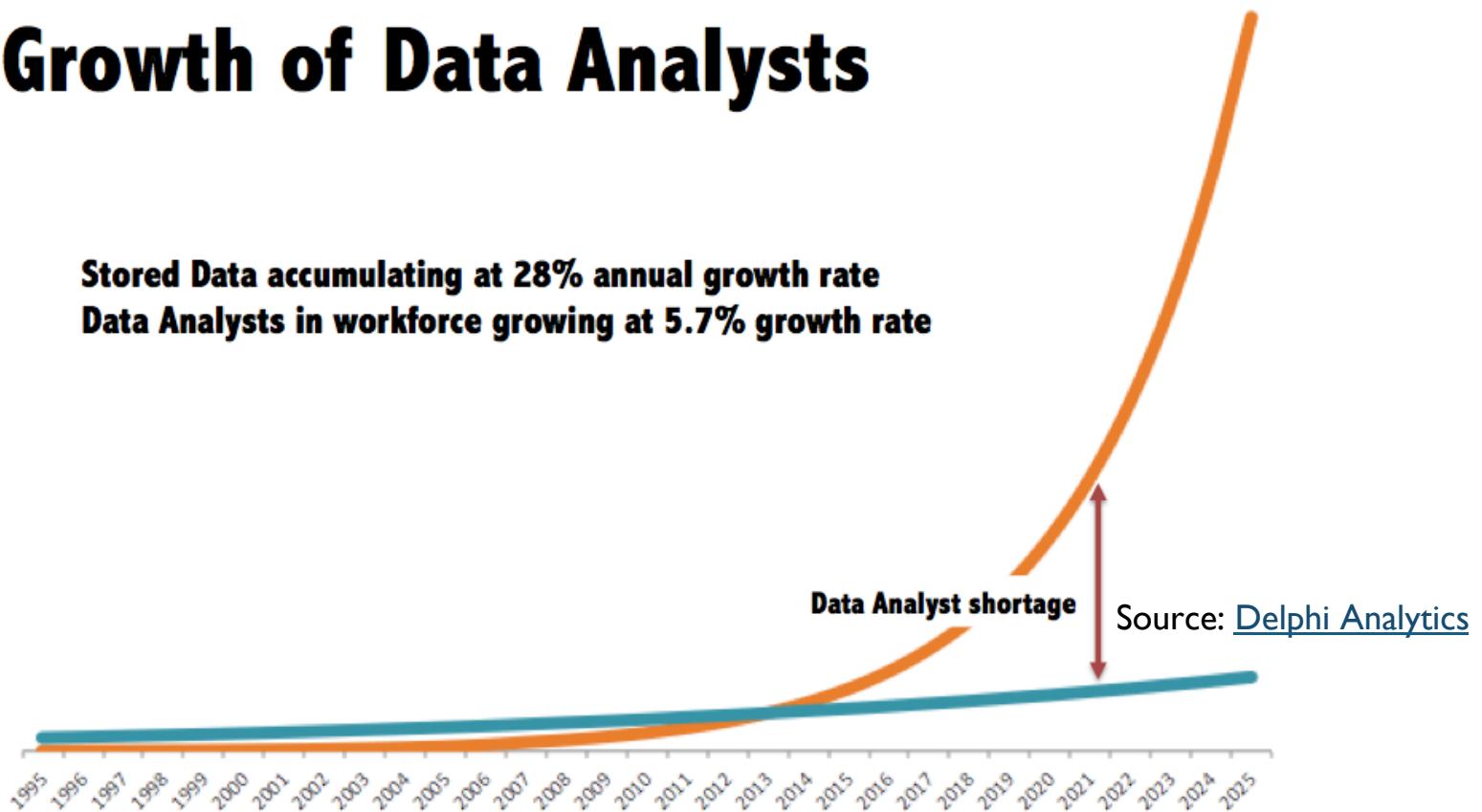
The volume of data, information, knowledge, etc in an organisation

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

Why Data Science Matters?

Growth of Data vs. Growth of Data Analysts

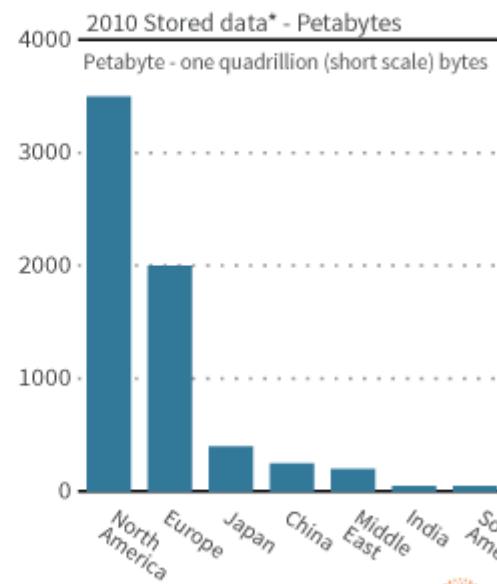
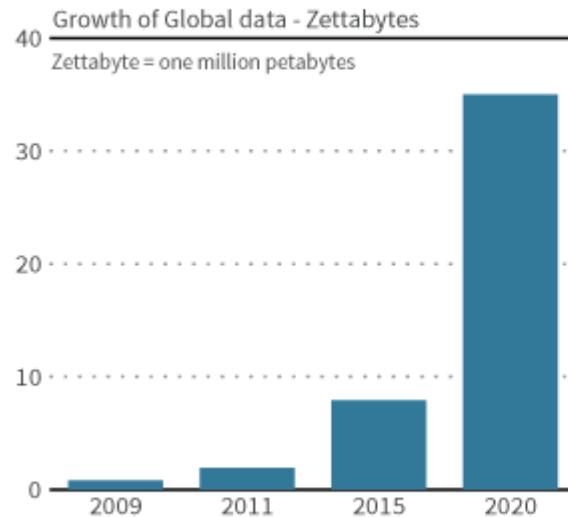
Stored Data accumulating at 28% annual growth rate
Data Analysts in workforce growing at 5.7% growth rate



Why Data Science Matters?

Big data growth

Big data market is estimated to grow 45% annually to reach \$25 billion by 2015



*greater than

Sources: Nasscom -CRISIL GR&A analysis

 **REUTERS**

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

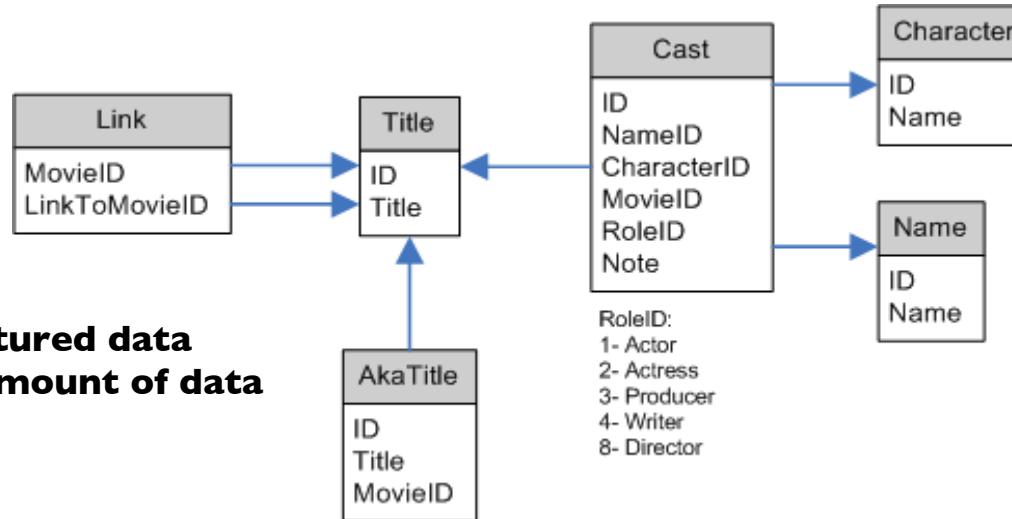
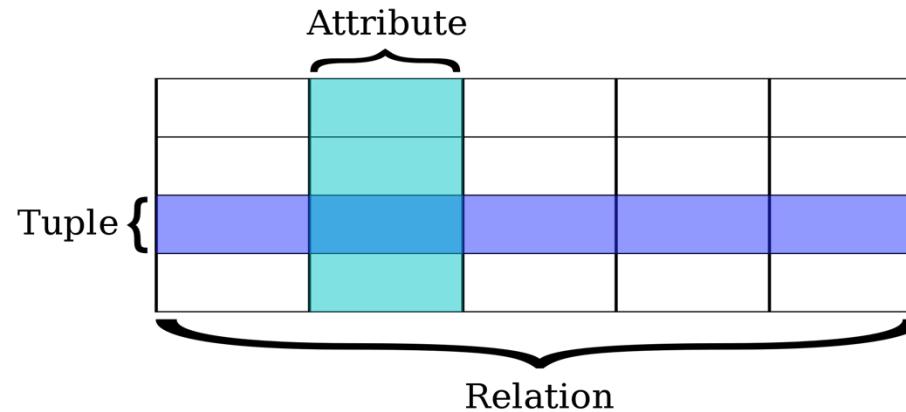
(Source: “If you think Big Data’s Big Now, Just Wait”, Ron Miller, TechTrends, 2014)

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.”

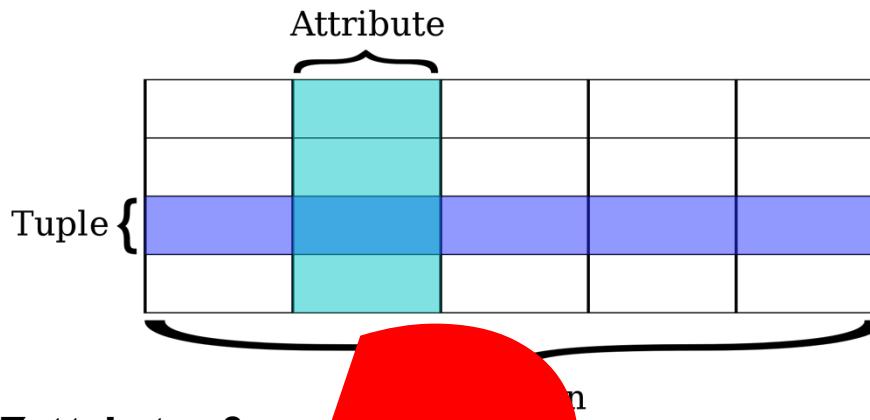
(Source: [Gartner](#))

Traditional Data Processing Method

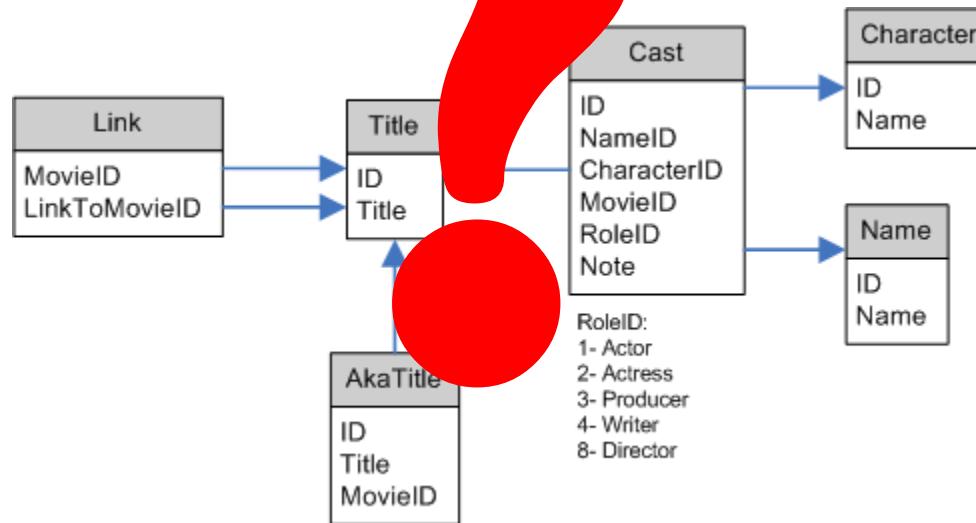


Mainly handled structured data
Limited in handling amount of data

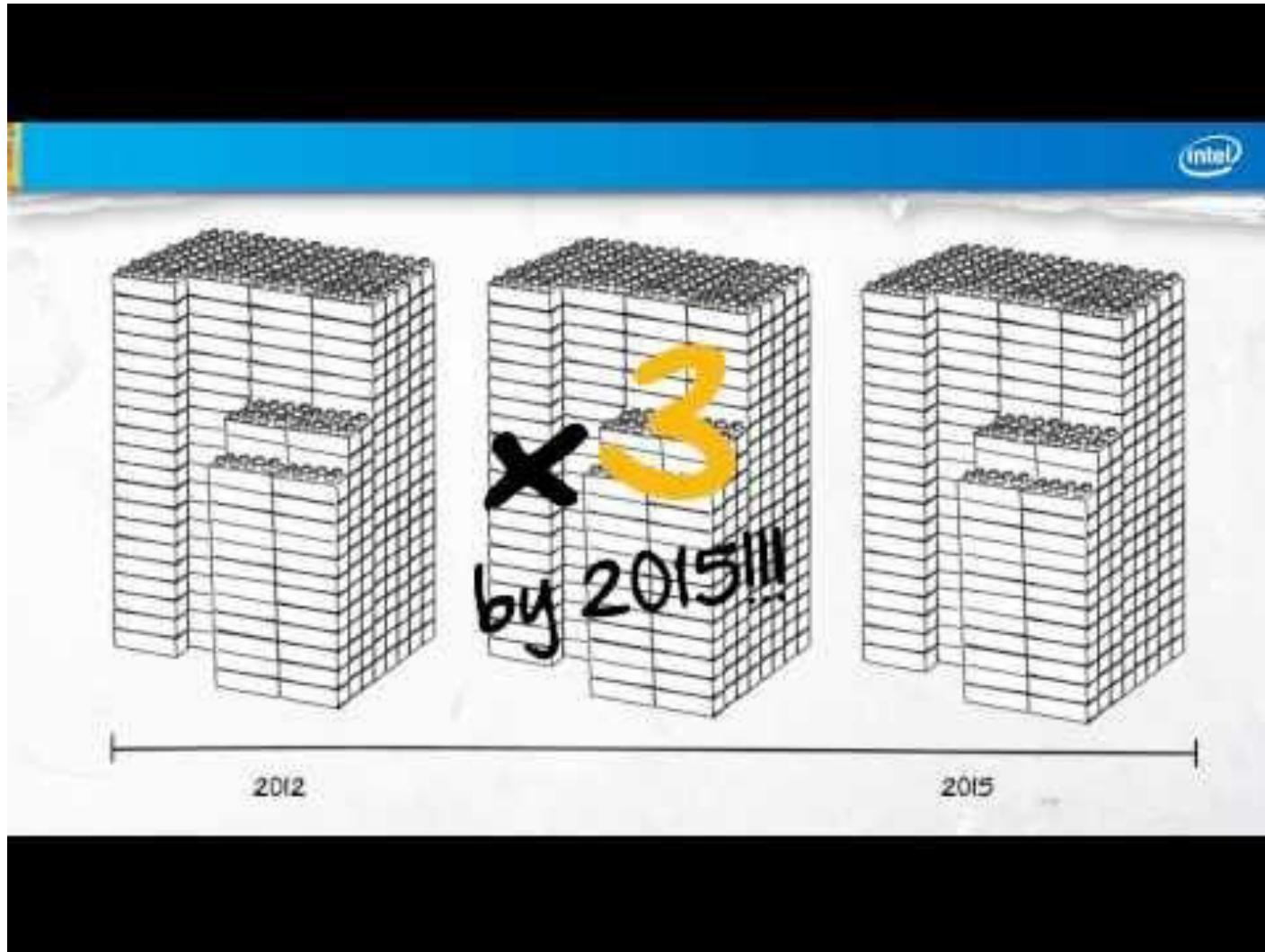
Traditional Data Processing Method



Not efficient in handling Zettabytes & unstructured data

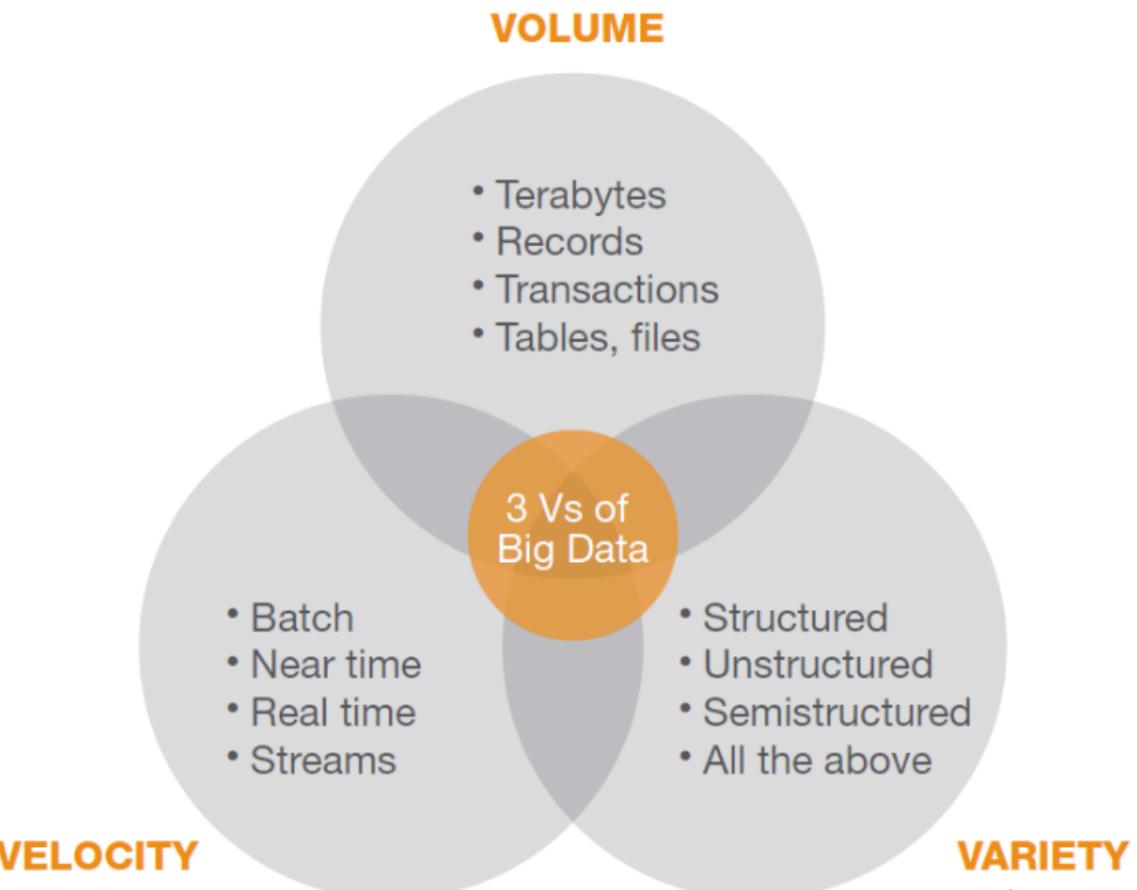


What is Big Data? (by Intel)



Characteristics of Big Data

- **3Vs of Big Data**



Source: [Digital Strategy](#)

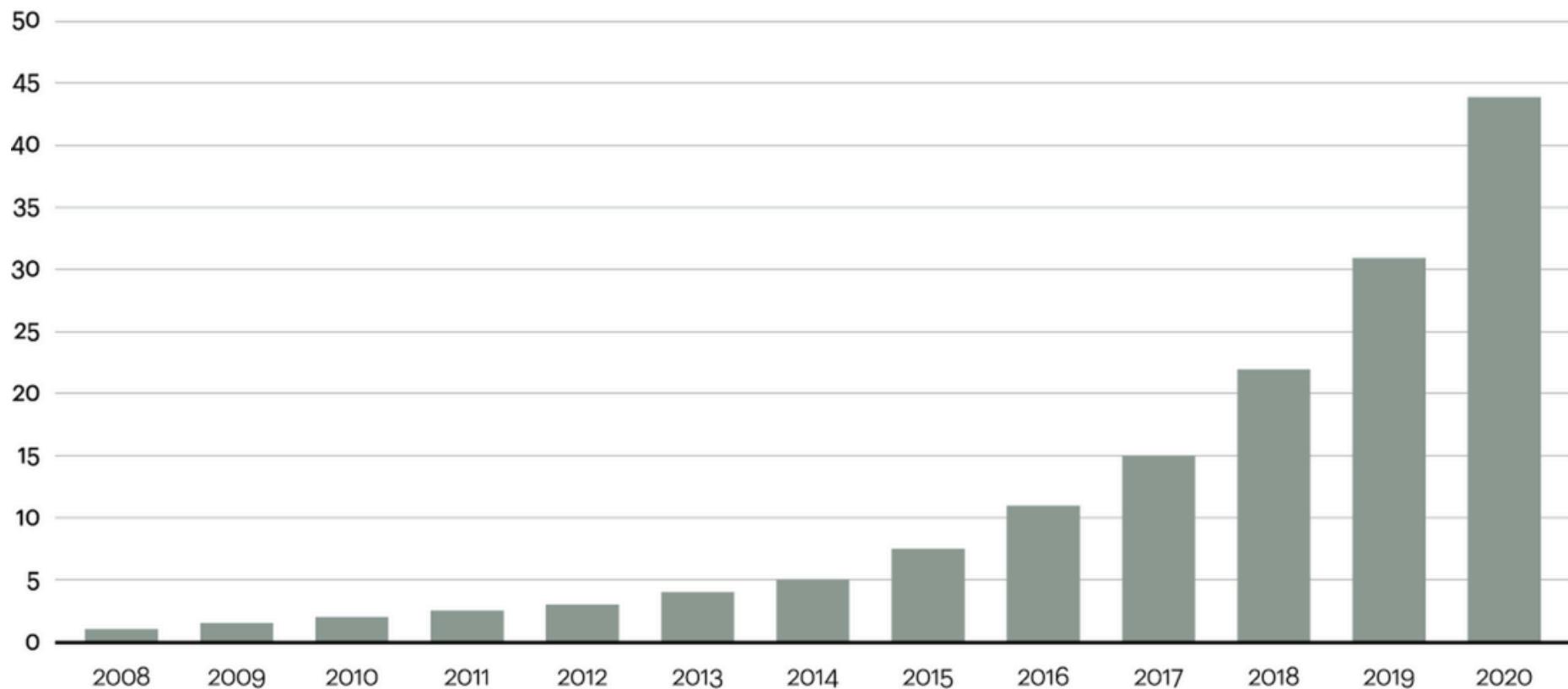
Characteristics of Big Data

- **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

Big Data - Volume

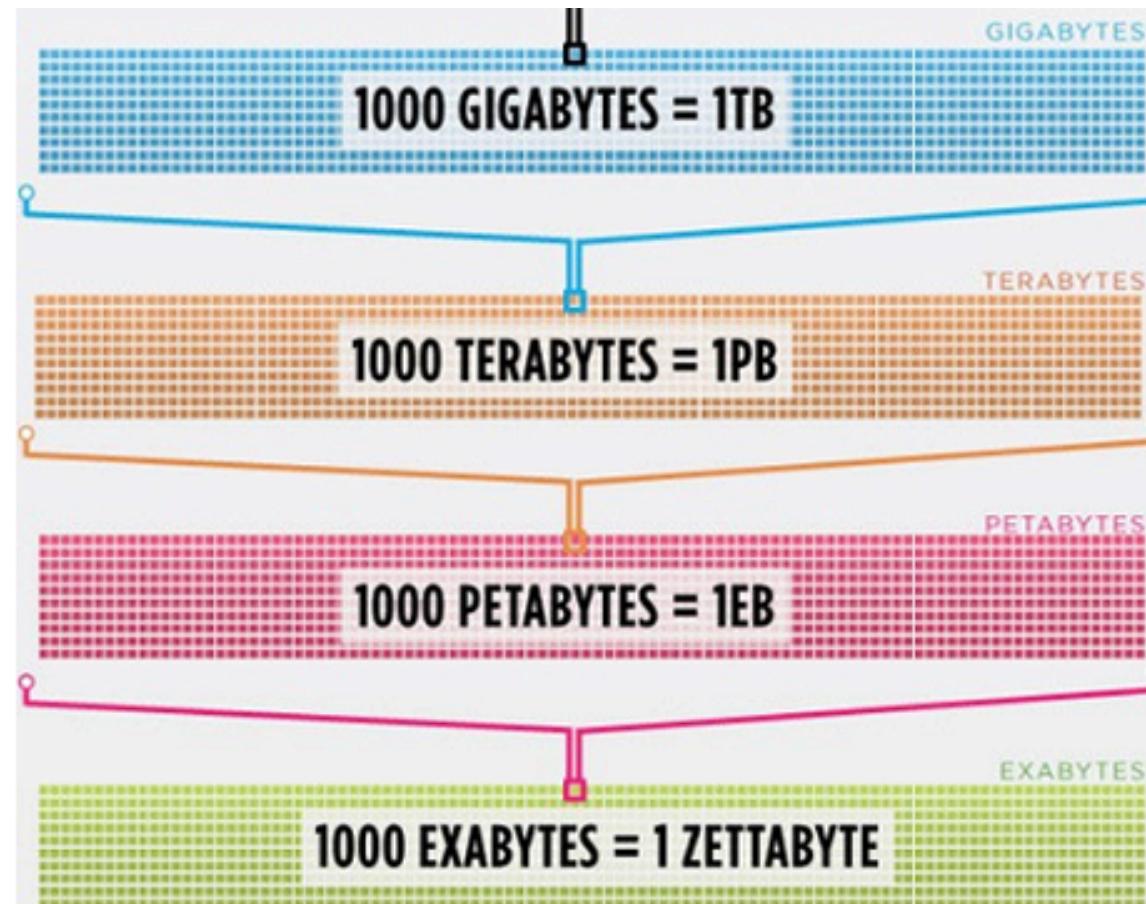
Data in zettabytes (ZB)

<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

internet live stats

live

1 second

watch

trends & more

Dashboards

Visualize your data securely. Easy, powerful dashboard builder.



3,068,829,691

Internet Users in the world



1,210,521,422

Total number of Websites



101,696,761,142

Emails sent **today**



1,963,084,535

Google searches **today**



1,804,094

Blog posts written **today**

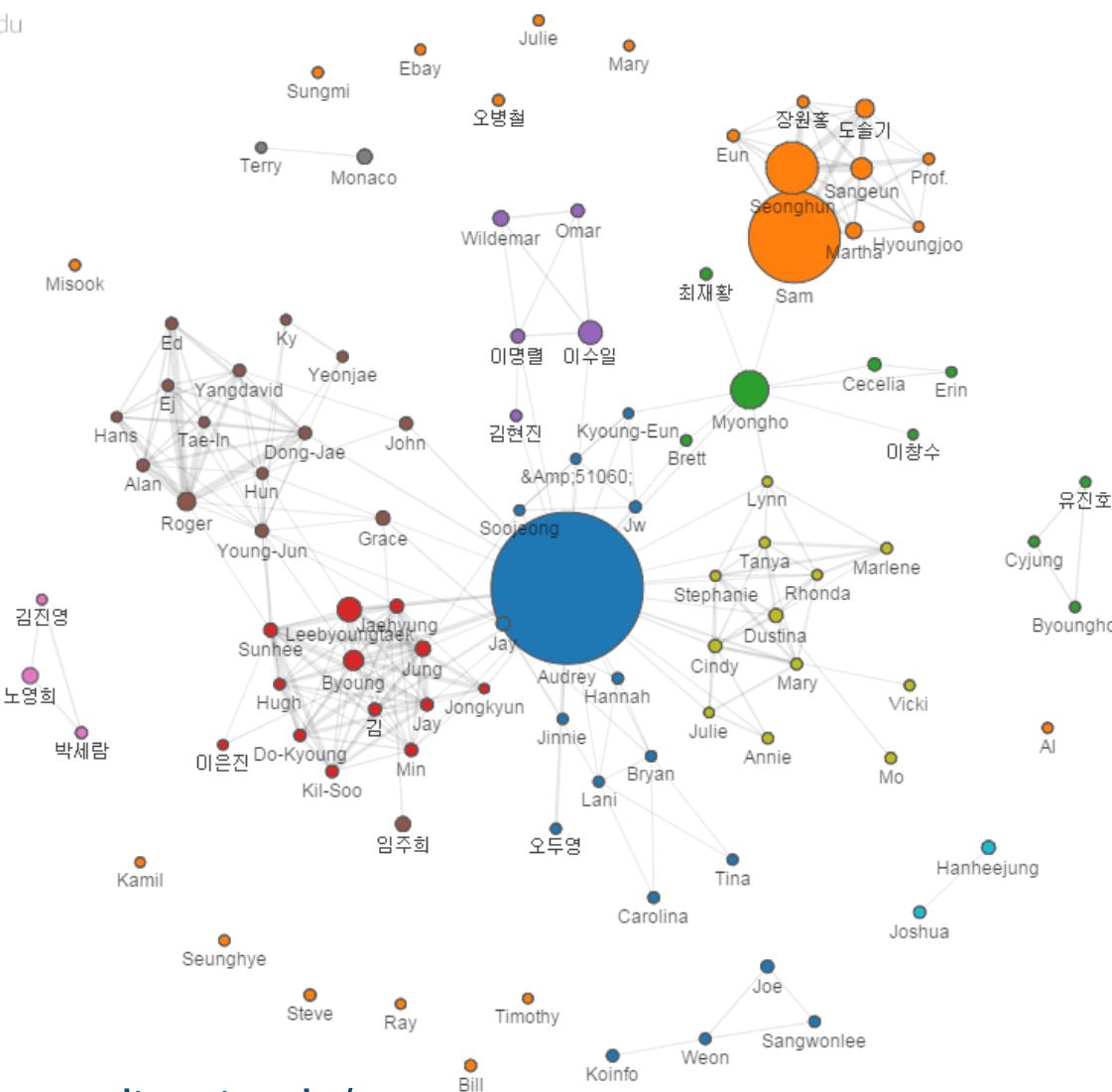


355,332,734

Tweets sent **today**

Personal Email (8 years) Analysis

immersion.media.mit.edu

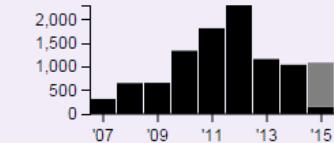


Joseph Lee

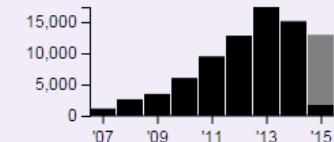
467 collaborators
79,949 emails

My Stats [Top Collaborators](#)

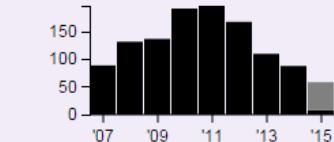
Emails Sent



Emails Received



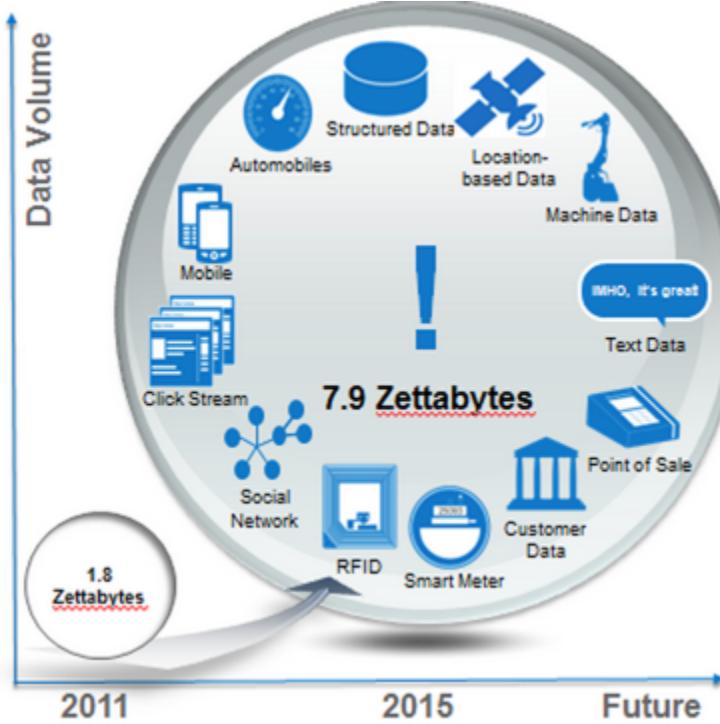
New Collaborators



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

8.0 years
23 Feb 2007 - 19 Feb 2015

Big Data - Variety

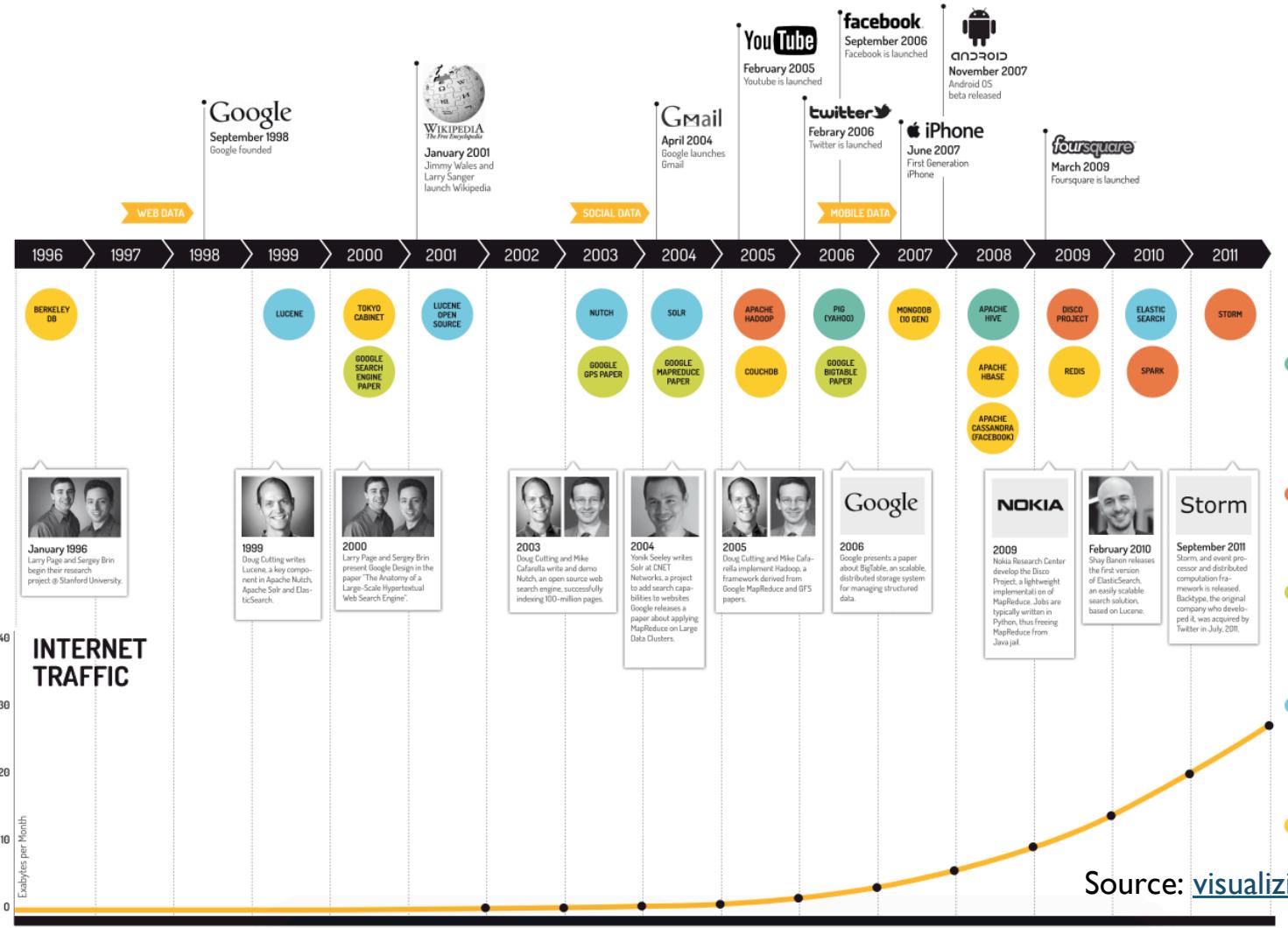


BY 2015
90%
of data will be
UNSTRUCTURED

GROWTH OF THE WORLD'S "DIGITAL UNIVERSE"



History of Big Data



History of Big Data (by TED)



Big Data Sets Example (#1 in Cloud Market)



The screenshot shows the AWS Public Data Sets homepage. At the top, there's a navigation bar with links for 'Sign Up', 'My Account / Console', and '한국어'. Below the navigation is a search bar and a main title 'AWS Public Data Sets'.

The left sidebar has a 'Browse By Category' section with links to various scientific fields: Astronomy, Biology, Chemistry, Climate, Economics, Encyclopedic, Geographic, and Mathematics. It also includes a 'Developer Resources' section with links to Amazon Machine Images (AMIs), Articles & Tutorials, Customer Apps, Developer Tools, Documentation, Release Notes, and Sample Code & Libraries.

The main content area features a heading 'Public Data Sets' with a sub-section titled 'Featured Public Data Sets'. It lists three datasets: 'Google Books Ngrams', 'Common Crawl Corpus', and '1000 Genomes Project'.

Public Data Sets

Public Data Sets on AWS provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications. AWS is hosting the public data sets at no charge for the community, and like all AWS services, users pay only for the compute and storage they use for their own applications. Learn more about [Public Data Sets on AWS](#) and visit the [Public Data Sets forum](#).

Featured Public Data Sets

Google Books Ngrams

A data set containing Google Books n-gram corpora. This data set is freely available on Amazon S3 in a Hadoop friendly file format and is licensed under a Creative Commons Attribution 3.0 Unported License. The original dataset is available from <http://books.google.com/ngrams/>.

Common Crawl Corpus

A corpus of web crawl data composed of over 5 billion web pages. This data set is freely available on Amazon S3 and is released under the Common Crawl Terms of Use.

1000 Genomes Project

The 1000 Genomes Project, initiated in 2008, is an international public-private consortium that aims to build the most detailed map of human genetic variation available.

Showing 1-25 of 57 results.

Sort by: Date - newest first 

Google Books Ngrams

A data set containing Google Books n-gram corpora. This data set is freely available on Amazon S3 in a Hadoop

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

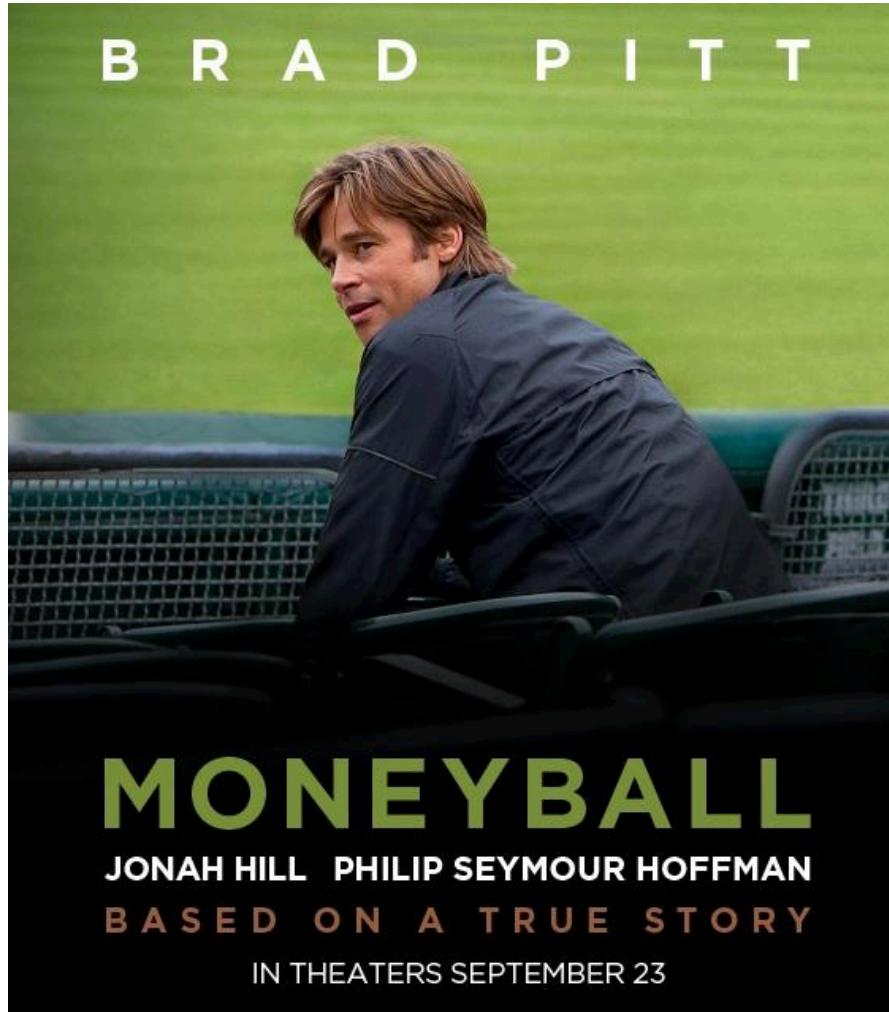
FLU TRENDS



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



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



Limited View



Snippet View



No Preview Available



[View the entire book](#)

[View a limited number of pages](#)

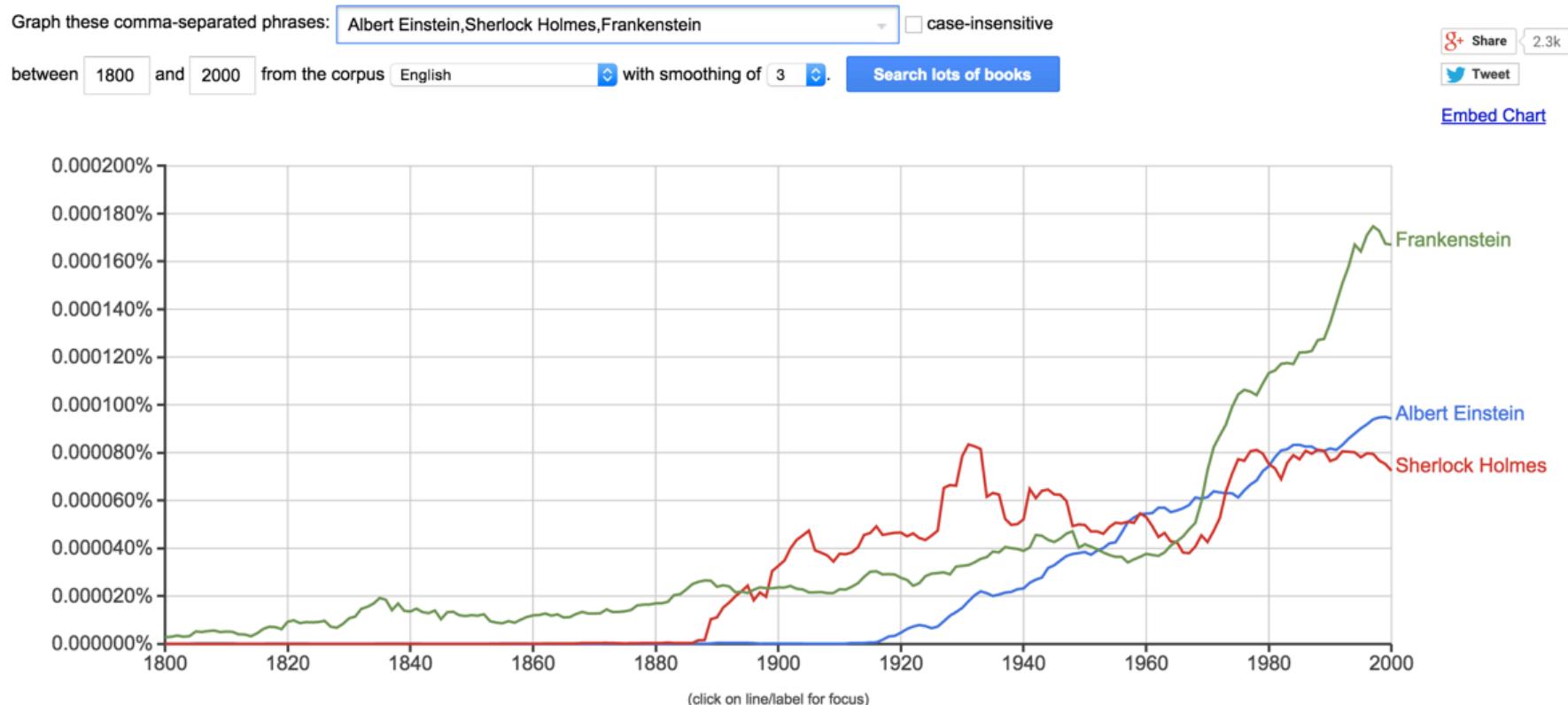
[View a few sentences surrounding the search term](#)

[View basic information about the book](#)

Big Data Use Cases

Google Ngram Viewer

Google books Ngram Viewer



Big Data Use Cases

By World Future Forum



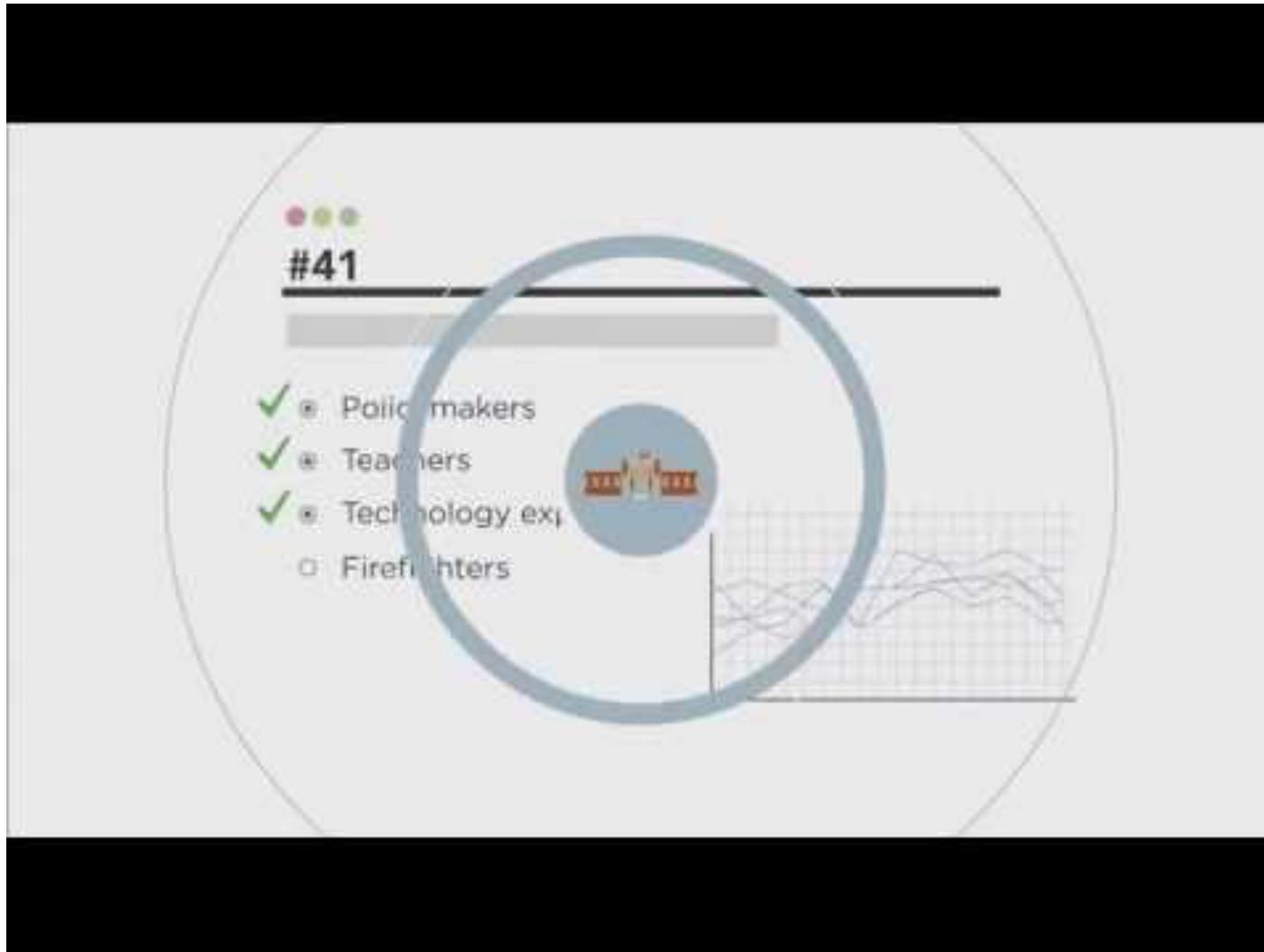
Big Data Use Cases

Daily Life Related

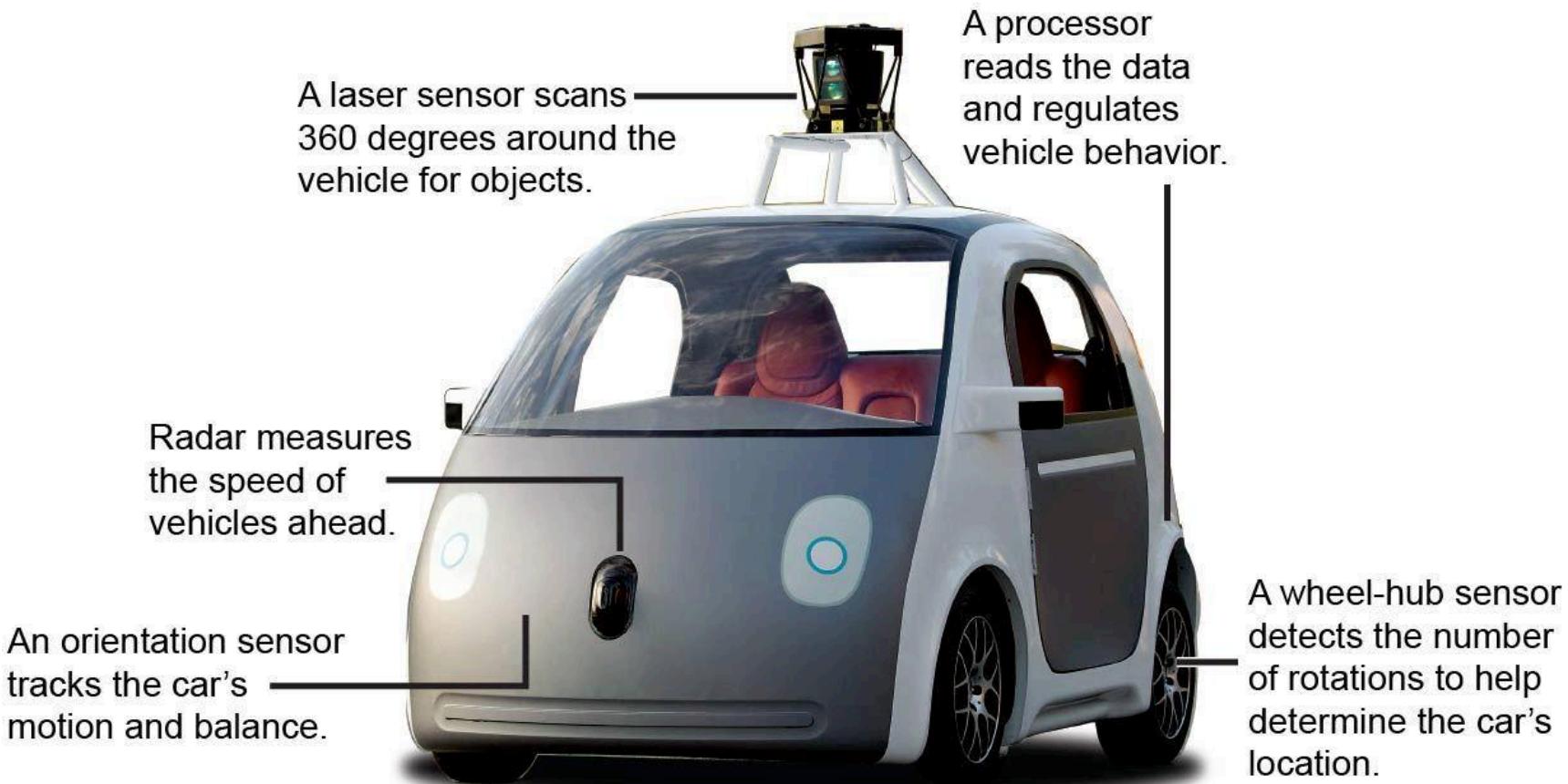


Big Data Use Cases

Education Related



Big Data Use Cases



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

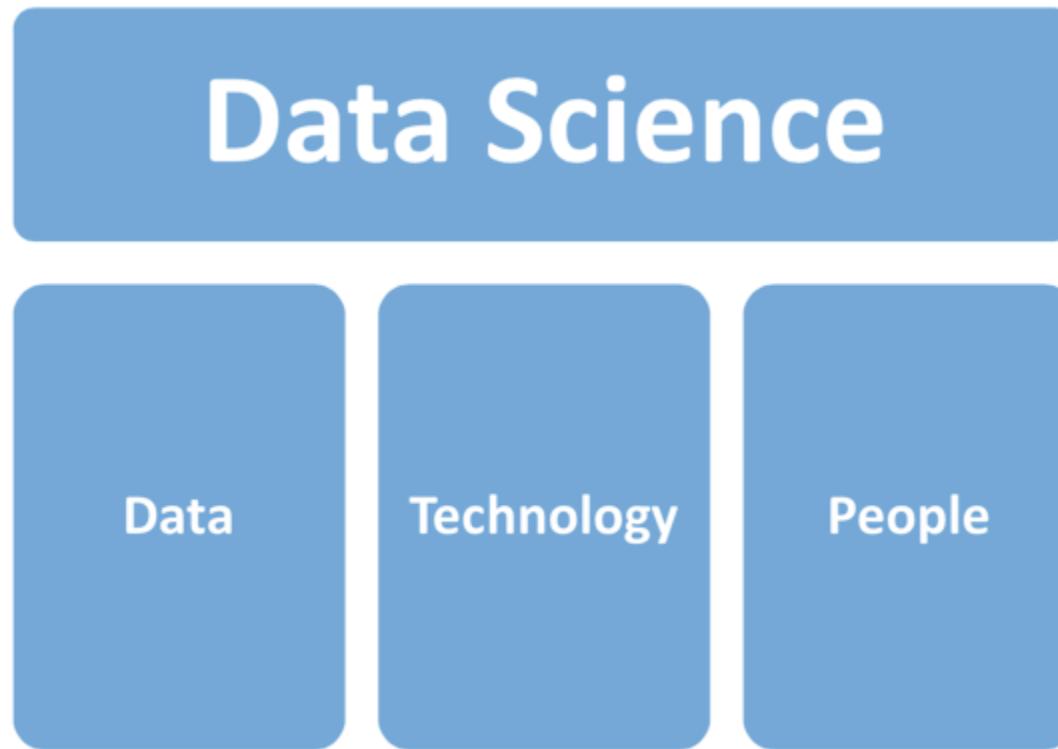
(Vasant Dhar, 2013)

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

(Provost & Fawcett, 2013)

What is Data Science?

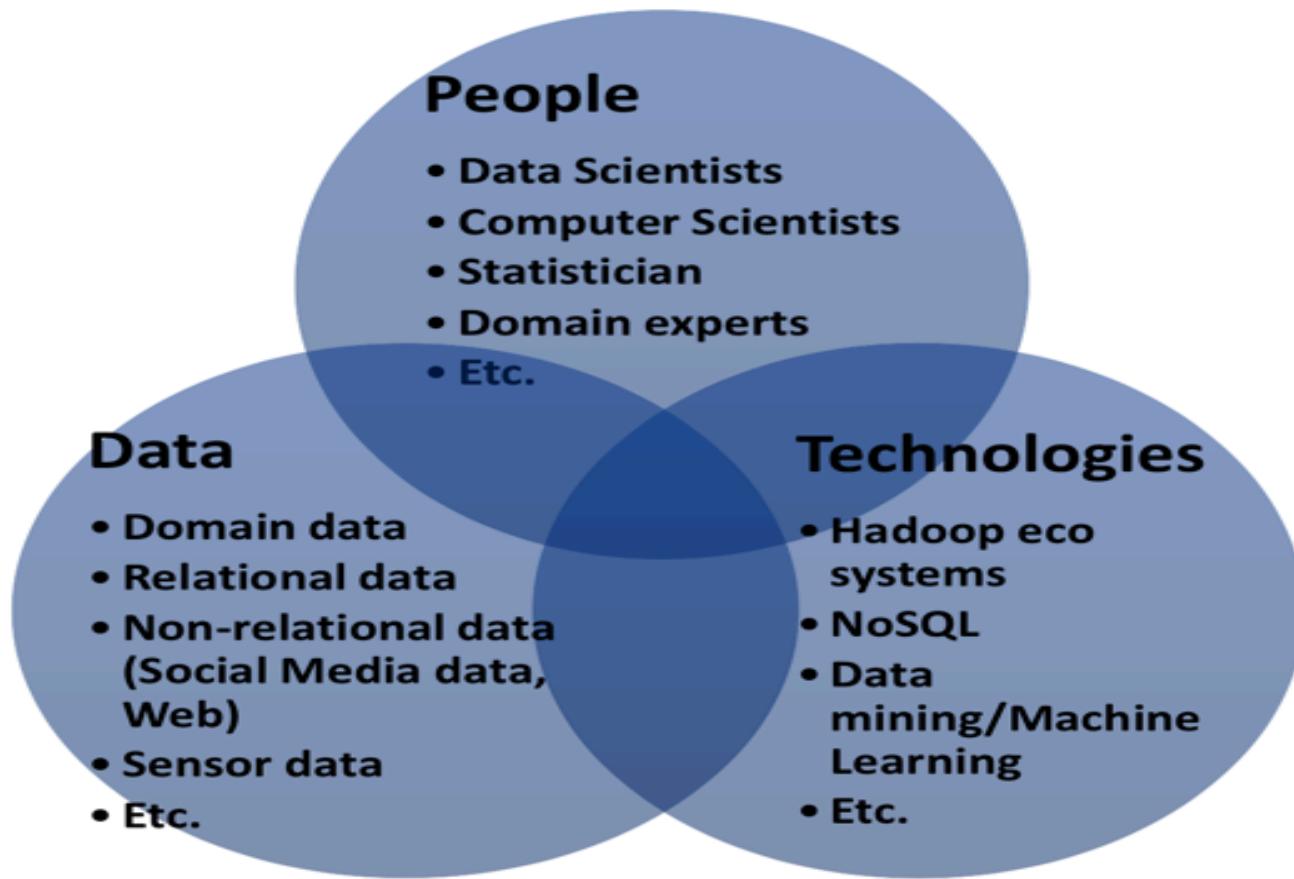
- Three pillars of Data Science



(Source: Il-Yoel Song, 2014)

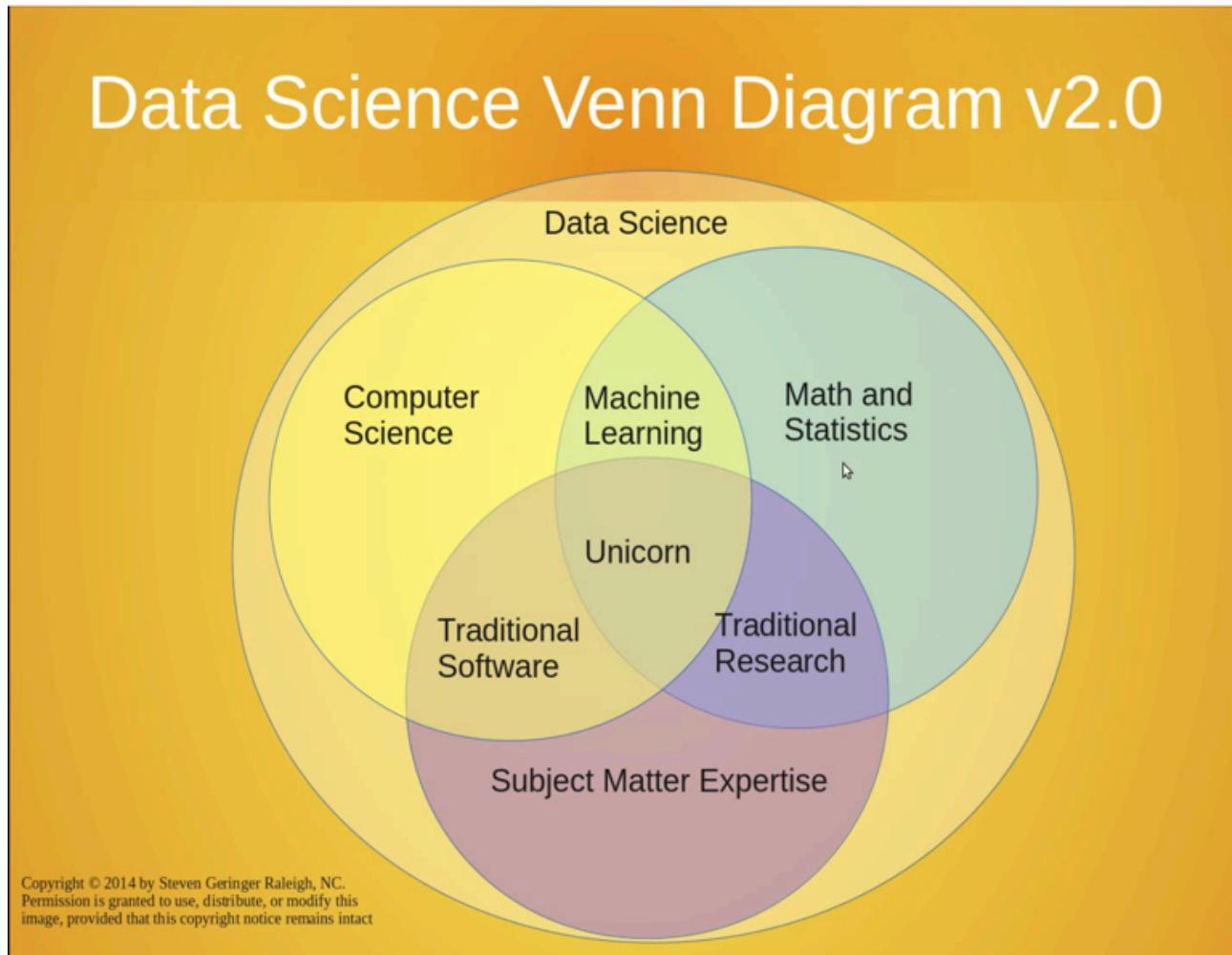
What is Data Science?

- Three pillars of Data Science



(Source: Il-Yoel Song, 2014)

What is Data Science?

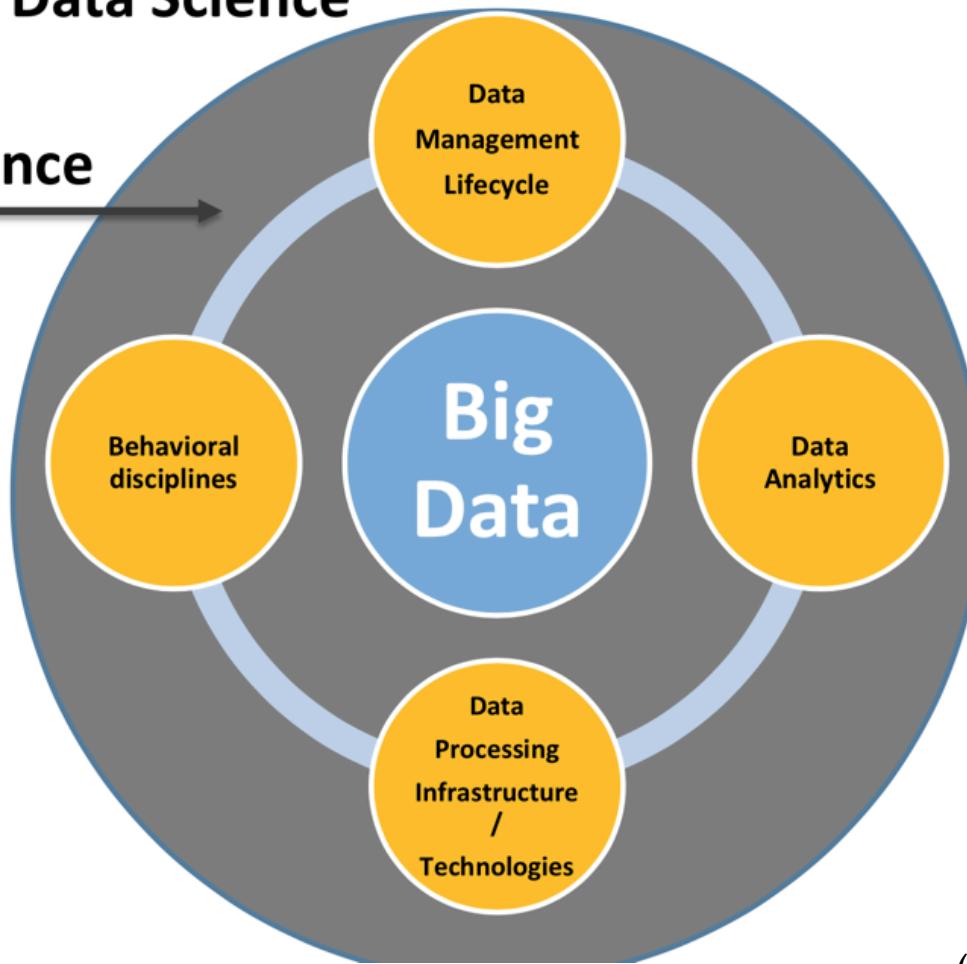


(Source: Steve Geringer, 2014)

Data Science vs Big Data

Big Data vs. Data Science

Data Science →



(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

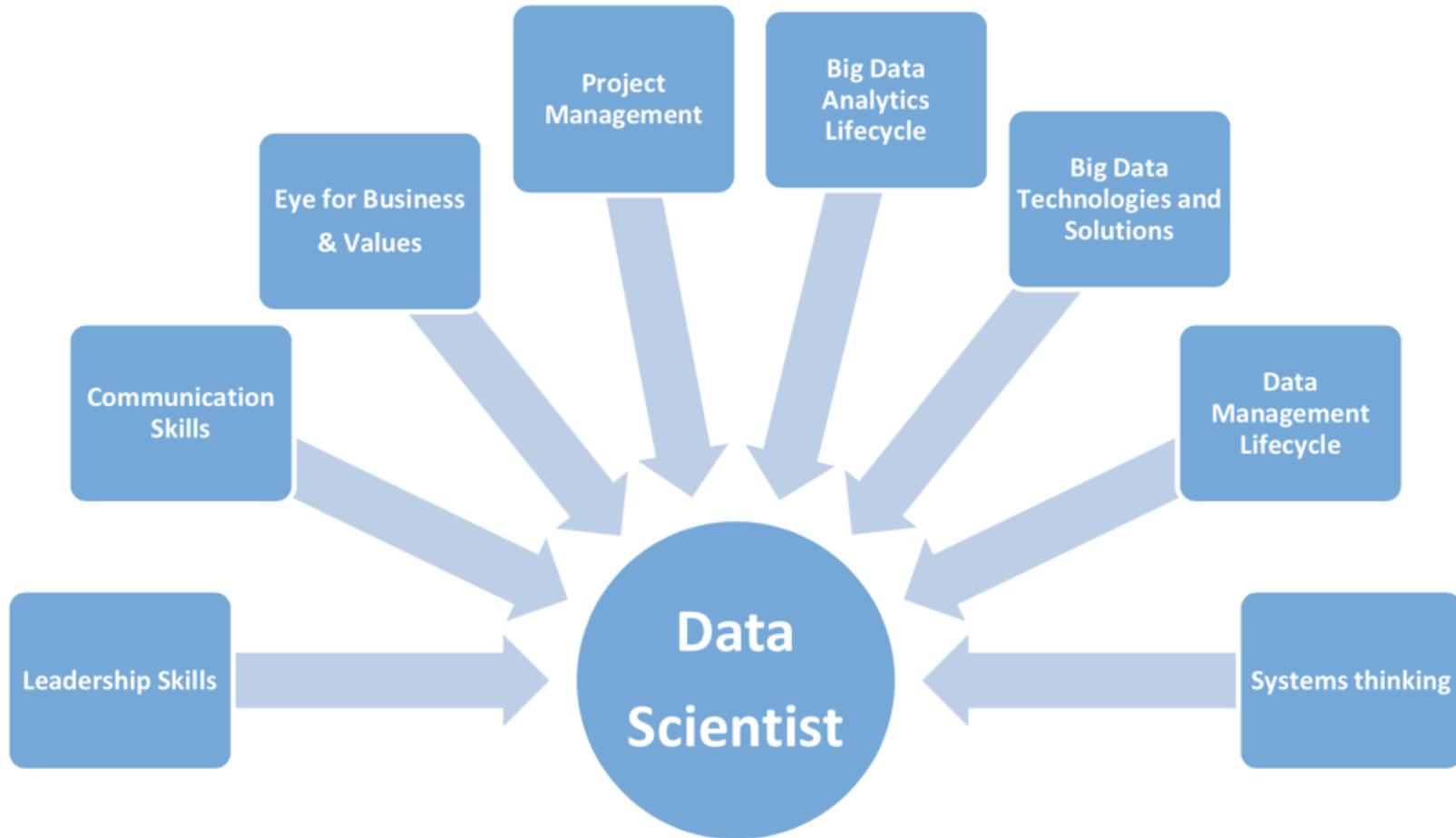
What Data Scientists Do?

- 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

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**

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, Tableau, Flare, Shiny, ggplot2...
- SAS, SPSS, Matlab...
- NoSQL, MongoDB, Couchbase, Cassandra...
- And more...

Data Scientist Toolkit



Data Scientist Toolkit

Source Data



Store Data



Convert & ETL



Transform Data



Exploratory Analysis



Model Build & Generate Insights



Visualisation



Model Execution in Production

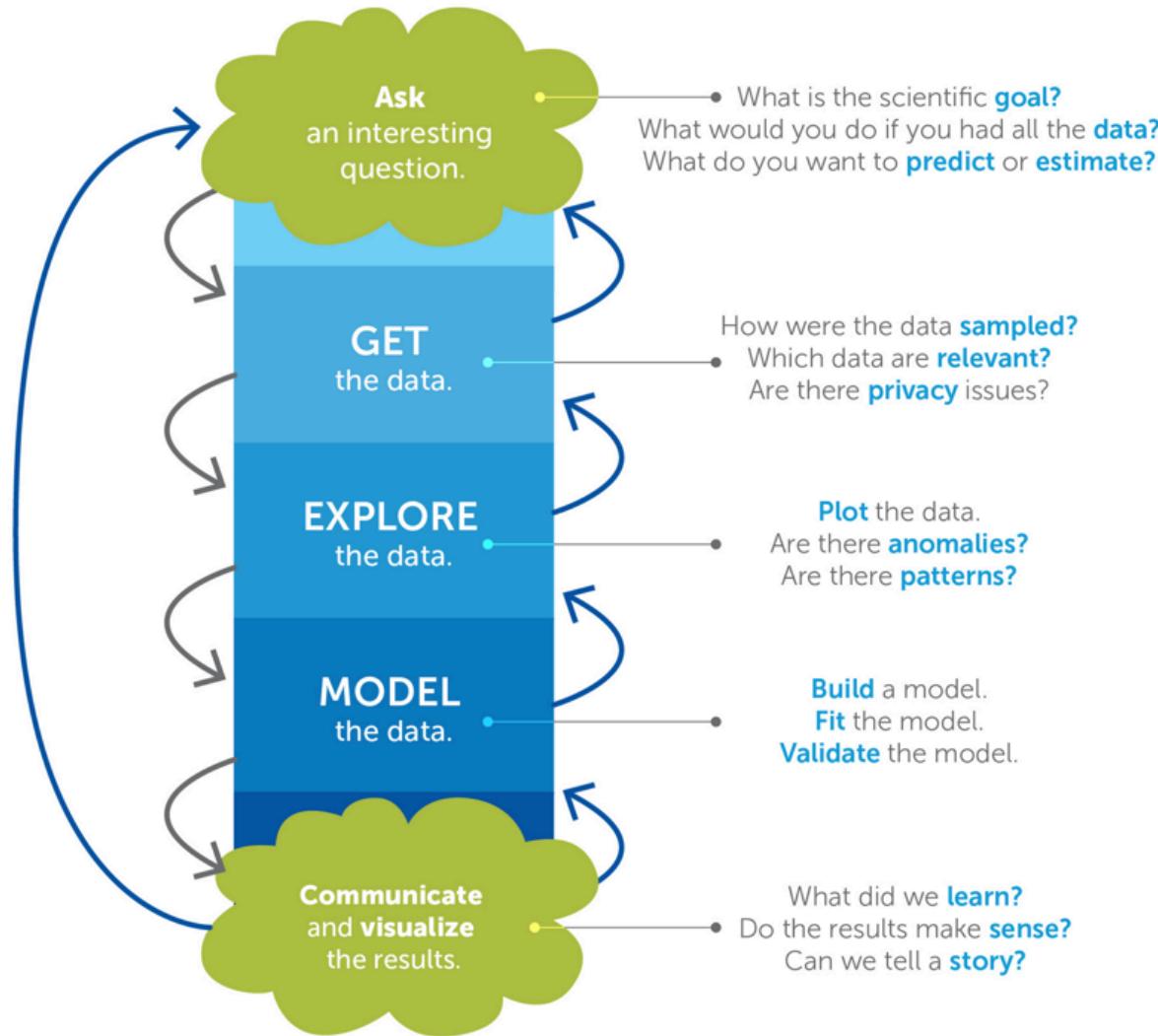


Data Scientist Interview

Data Scientist

EMC Academic Alliance

Data Science Process



(Source: [Opera solutions](#))

What is Data Product ?

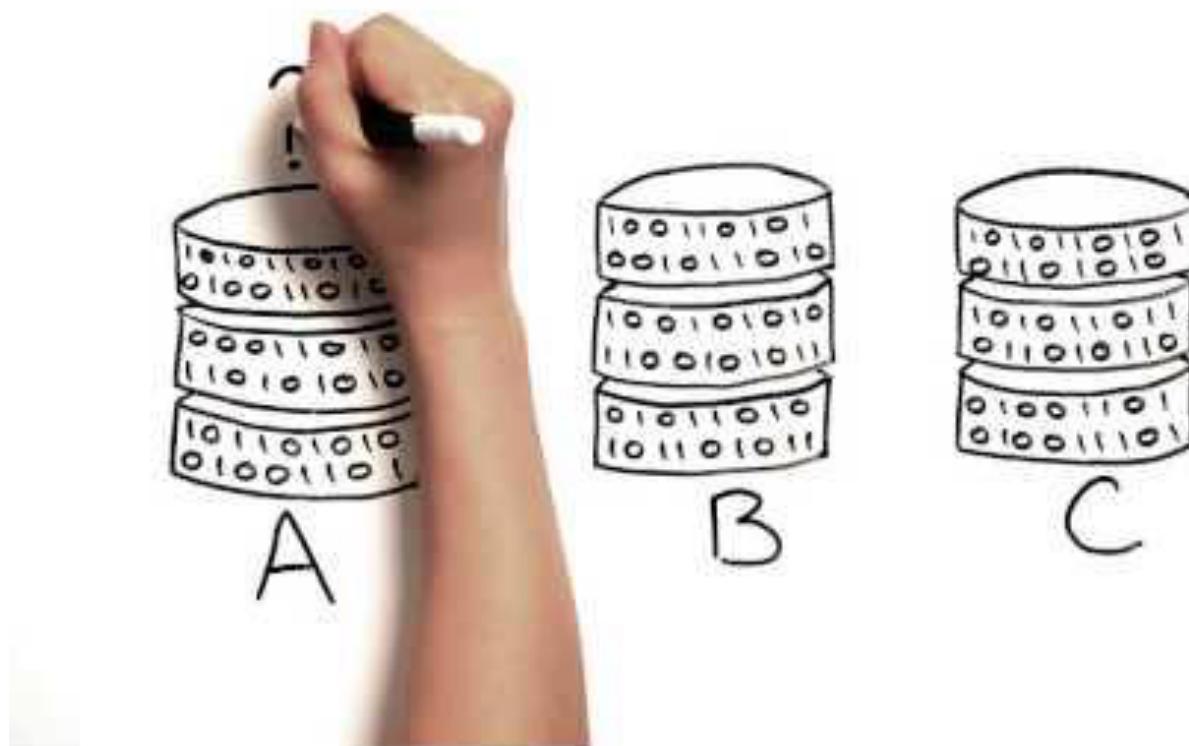
A data product is the production output from a statistical analysis. Data products automate complex analysis tasks or use technology to expand the utility of a data informed model, algorithm or inference

(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)