Big Data, what is it?

Big Data, what is it?

traditional
computer science

data that will not fit in main memory.

Big Data, what is it?

traditional
computer science

data that will not fit in main memory.

For example...

busy web server access logs

graph of the entire Web

all of Wikipedia

daily satellite imagery over a year

Big Data, what is it?

traditional
computer science

data that will not fit
in main memory.

data with a *large*
number of observations
and/or features.

statistics

Big Data, what is it?

Tall data:

edge list of a large graph

rgb values per pixel location in large images

data with a *large* number of observations and/or features.

statistics

Wide data: *mobile app usage statistics of 100 people*

Big Data, what is it?

traditional
computer science

data that will not fit
in main memory.

data with a *large*
number of observations
and/or features.

statistics

Big Data, what is it?

traditional computer science

data that will not fit in main memory.

data with a *large* number of observations and/or features.

statistics

other fields

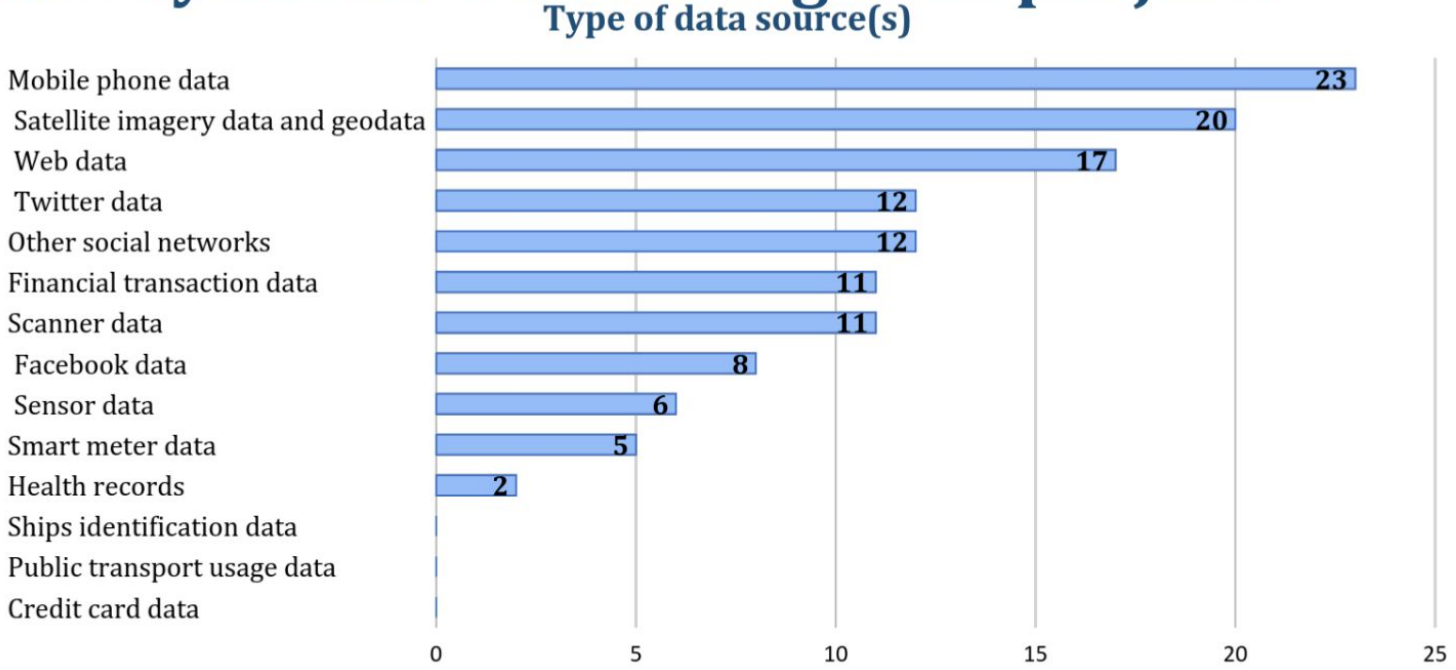
non-traditional sample size (i.e. > 300 subjects); can't analyze in stats tools (Excel).

Big Data, what is it? *Government View*

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP (2016)

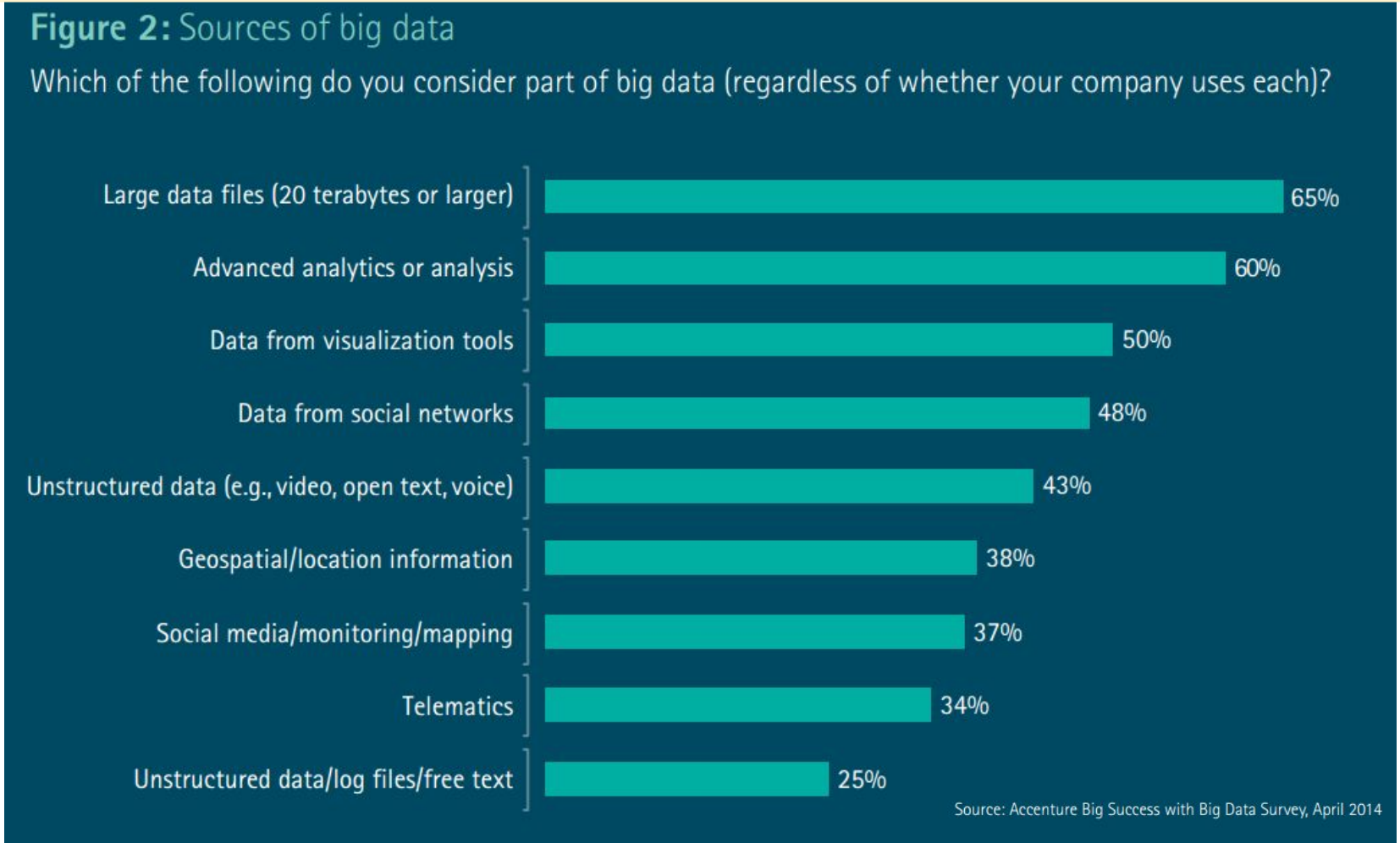
BigData
UN Global Working Group

1. Survey of SDG-related Big Data projects



- Mobile (23), Satellite imagery (20) and social media (12+12+8) are the most prominent sources

Big Data, what is it? *Industry View*



Big Data, what is it?

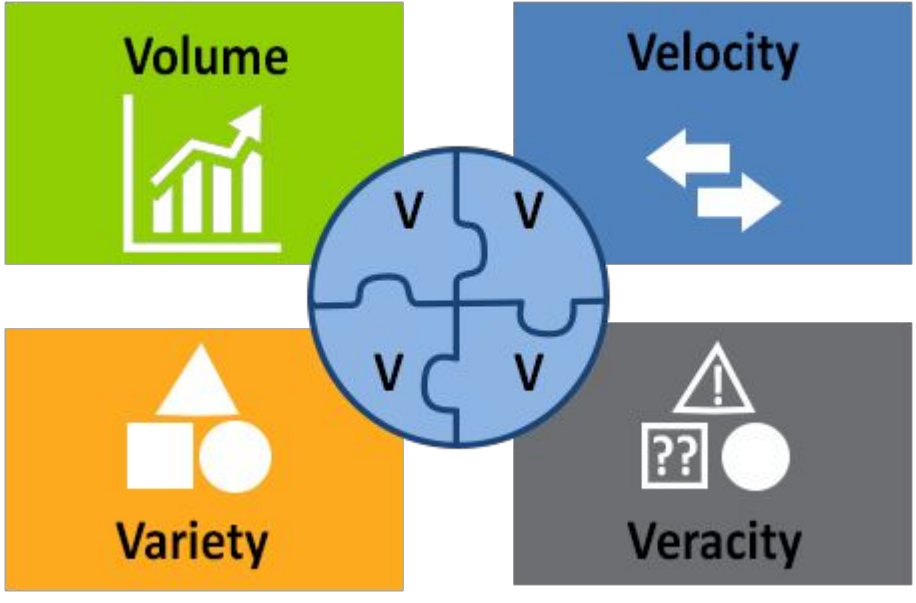
Analyses which can handle the 3 Vs and do it with quality (veracity):

(Laney, 2001: META Group)

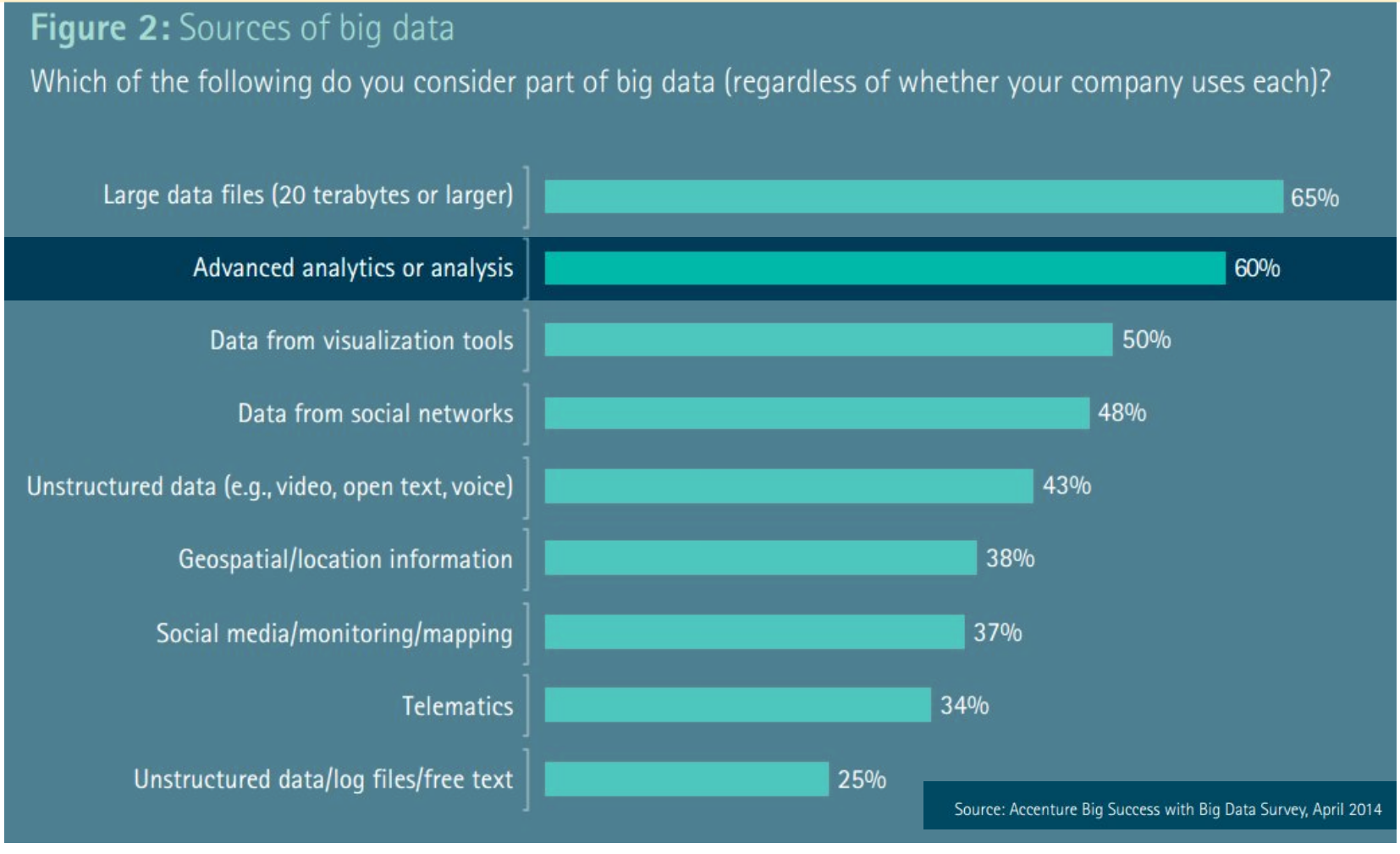
1. *large quantity*

2. *arriving quickly*

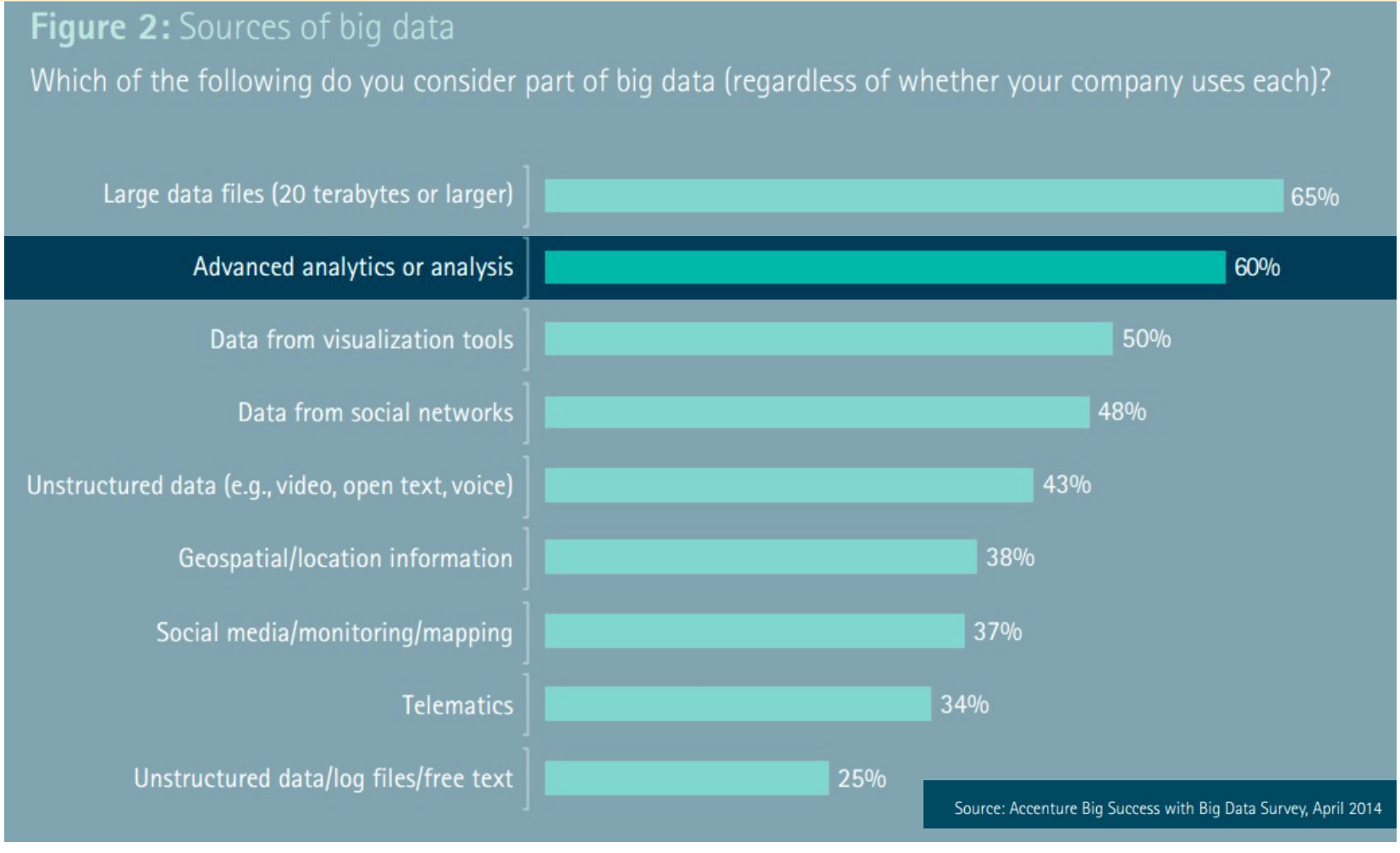
3. *[un]structed, multi-modal*



Big Data, what is it? *Industry View*



Big Data, a type of analytics



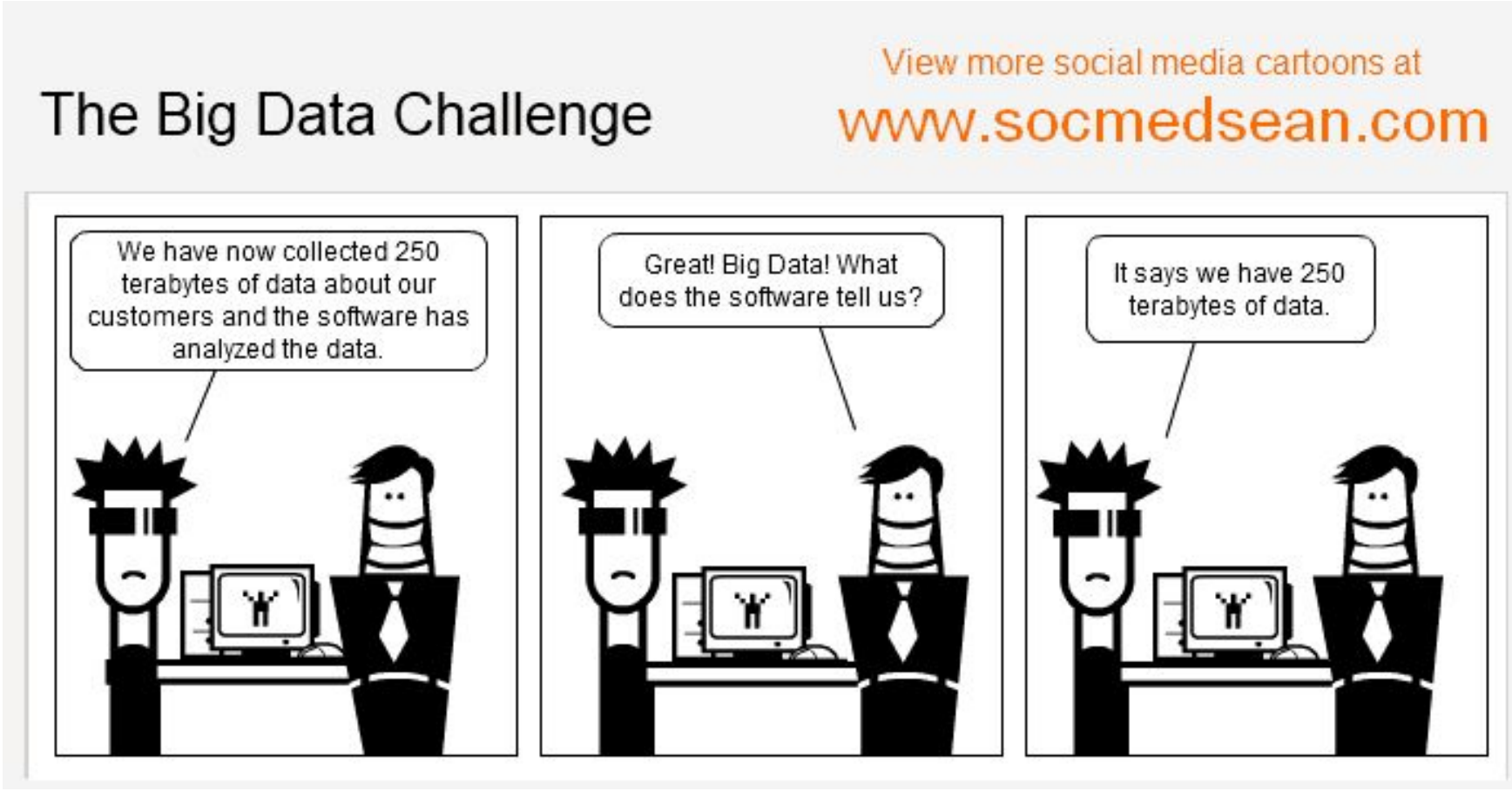
Big Data, a type of analytics

?

Big Data, a type of analytics

Big Data, a type of analytics

Big Data, a type of analytics



Big Data, what is it?

(ChatGPT, January 2023)

?

Big Data, what is it?

(ChatGPT, January 2023)

Big data refers to the large and complex sets of data that are generated and collected by organizations and individuals on a daily basis. These data sets can come from a variety of sources, such as social media, online transactions, and sensor data, and can be structured or unstructured.

One of the main challenges of big data is the ability to store, process, and analyze it effectively. Traditional data processing methods and technologies are often not able to handle the volume, velocity, and variety of big data. As a result, new technologies and approaches, such as distributed computing and machine learning, have been developed to help organizations make sense of their big data.

…

Big Data, what is it?

(ChatGPT, January 2023)

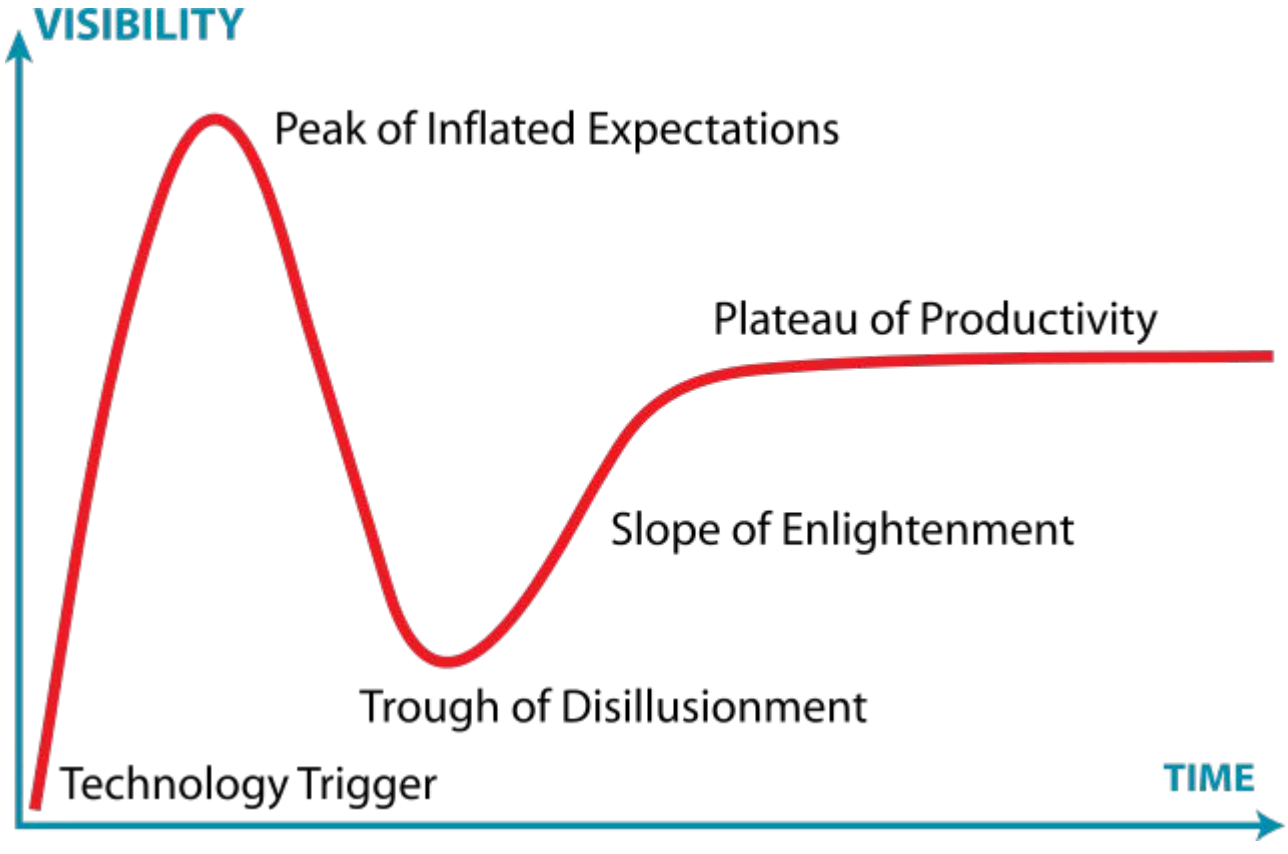
Big data refers to the large and complex sets of data that are generated and collected by organizations and individuals on a daily basis. These data sets can come from a variety of sources, such as social media, online transactions, and sensor data, and can be structured or unstructured.

One of the main challenges of big data is the ability to store, process, and analyze it effectively. Traditional data processing methods and technologies are often not able to handle the volume, velocity, and variety of big data. As a result, new technologies and approaches, such as distributed computing and machine learning, have been developed to help organizations make sense of their big data.

Big data can have a wide range of applications, from improving business operations and customer service to enabling new scientific discoveries and advancements in healthcare. For example, in business, big data can be used to gain insights into customer behavior, identify new market opportunities, and optimize supply chain operations. In healthcare, big data can be used to improve patient outcomes and develop personalized treatment plans.

Overall, big data is a rapidly growing field with many potential benefits for organizations and individuals, but also has the potential for privacy and security concerns. Therefore, it is important for organizations to have a robust data governance framework and for individuals to understand the implications of data collection and use.

Big Data, a buzz word?



(Gartner Hype Cycle)

Big Data, a buzz word?

2008

2011

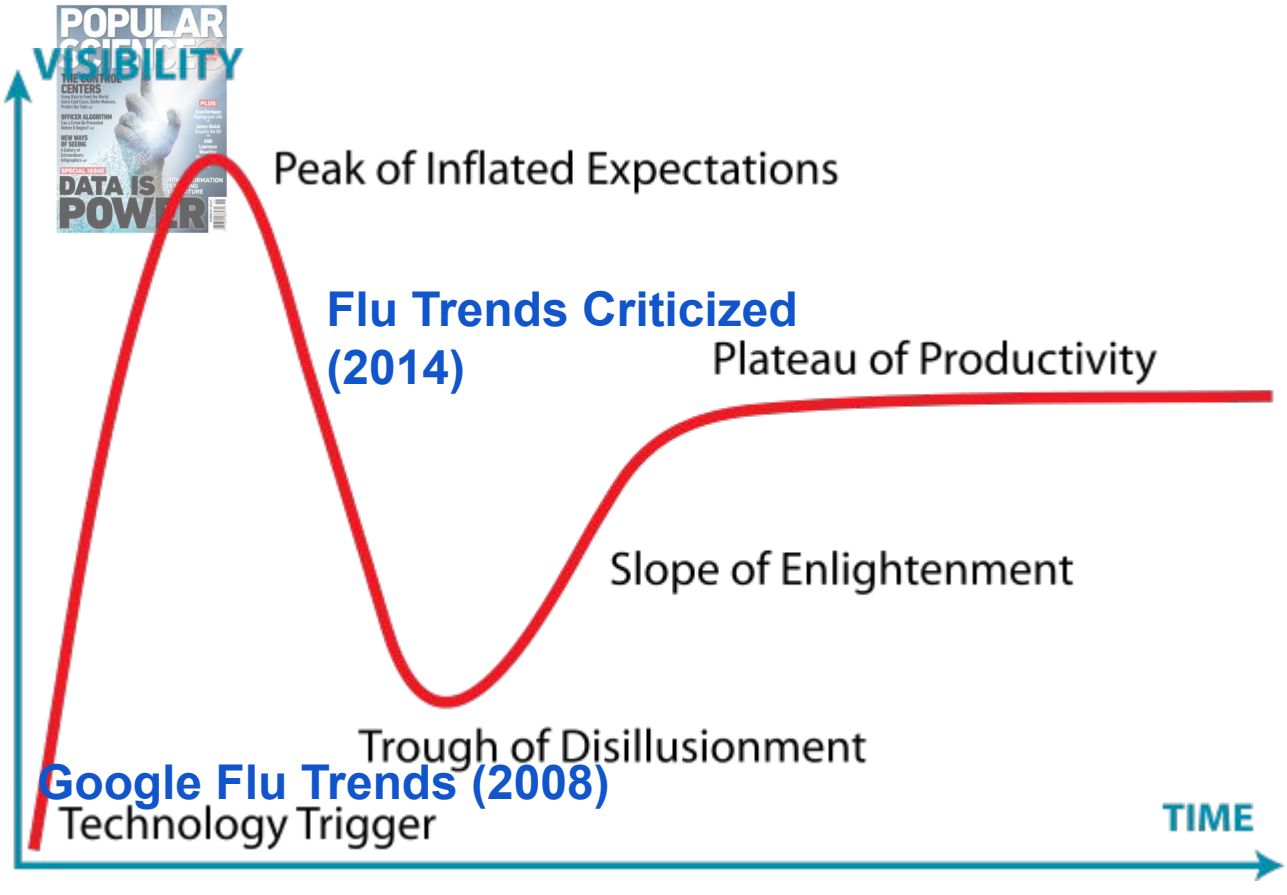
2012

2010

2011

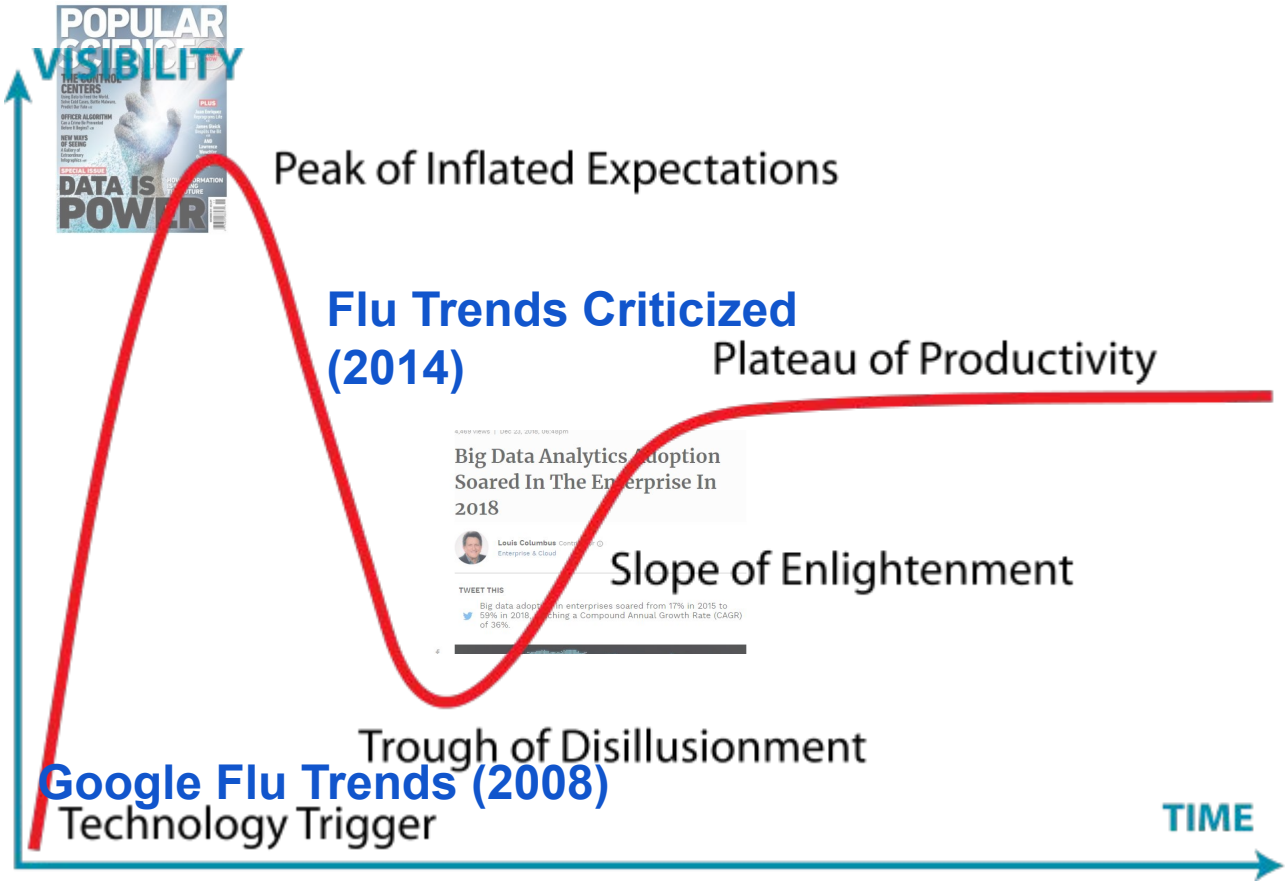
2018

Big Data, a buzz word?



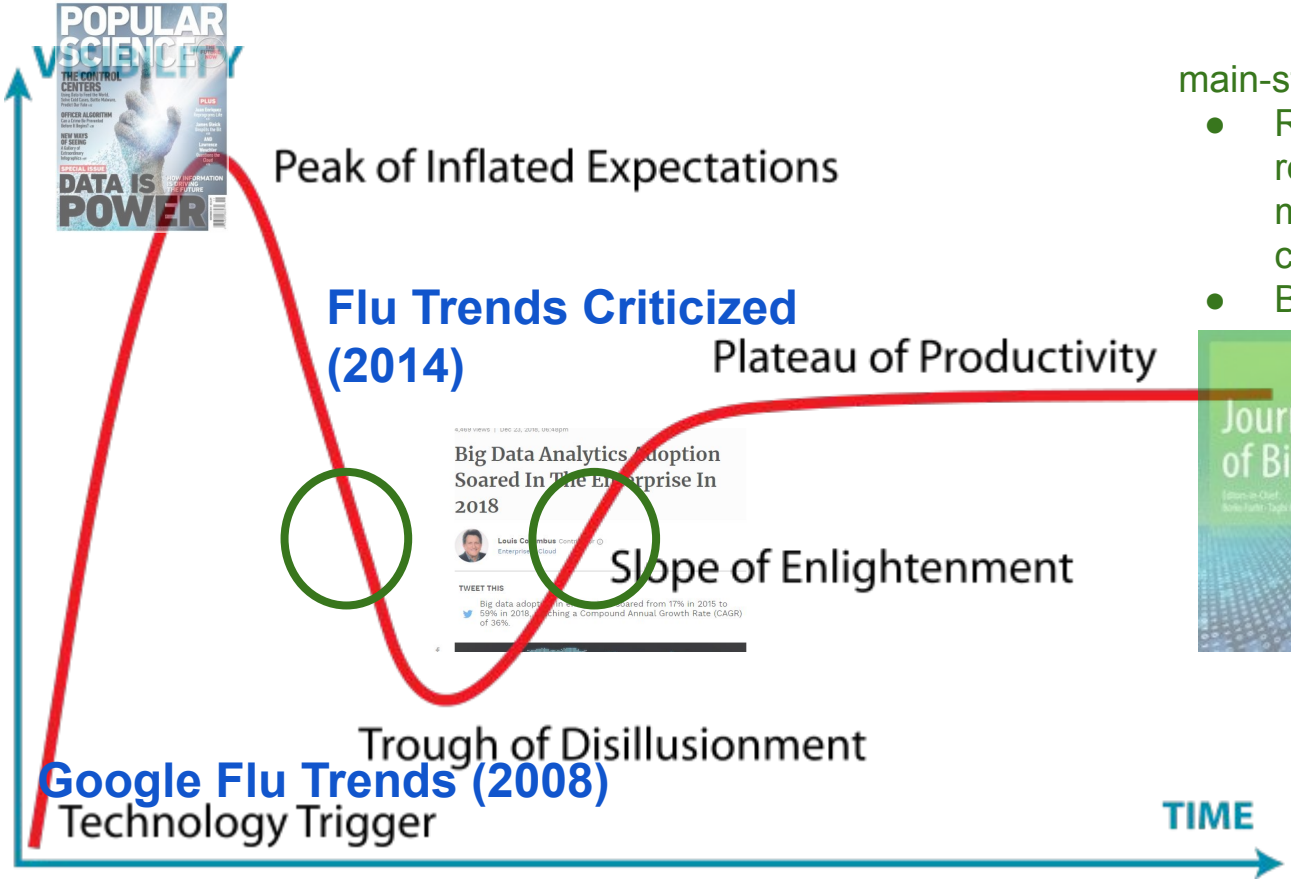
(Gartner Hype Cycle)

Big Data, a buzz word?



(Gartner Hype Cycle)

Big Data, a buzz word?



main-stream study being established

- Realization of what subfields are really doing “big data” (i.e. data mining, ML, Statistics, computational social sciences).
- Best practices being established.

(Gartner Hype Cycle)

Big Data, a buzz word?

2008

2011

2012

2022

2010

2011

2018

Big Data, a buzz word?

Google Scholar

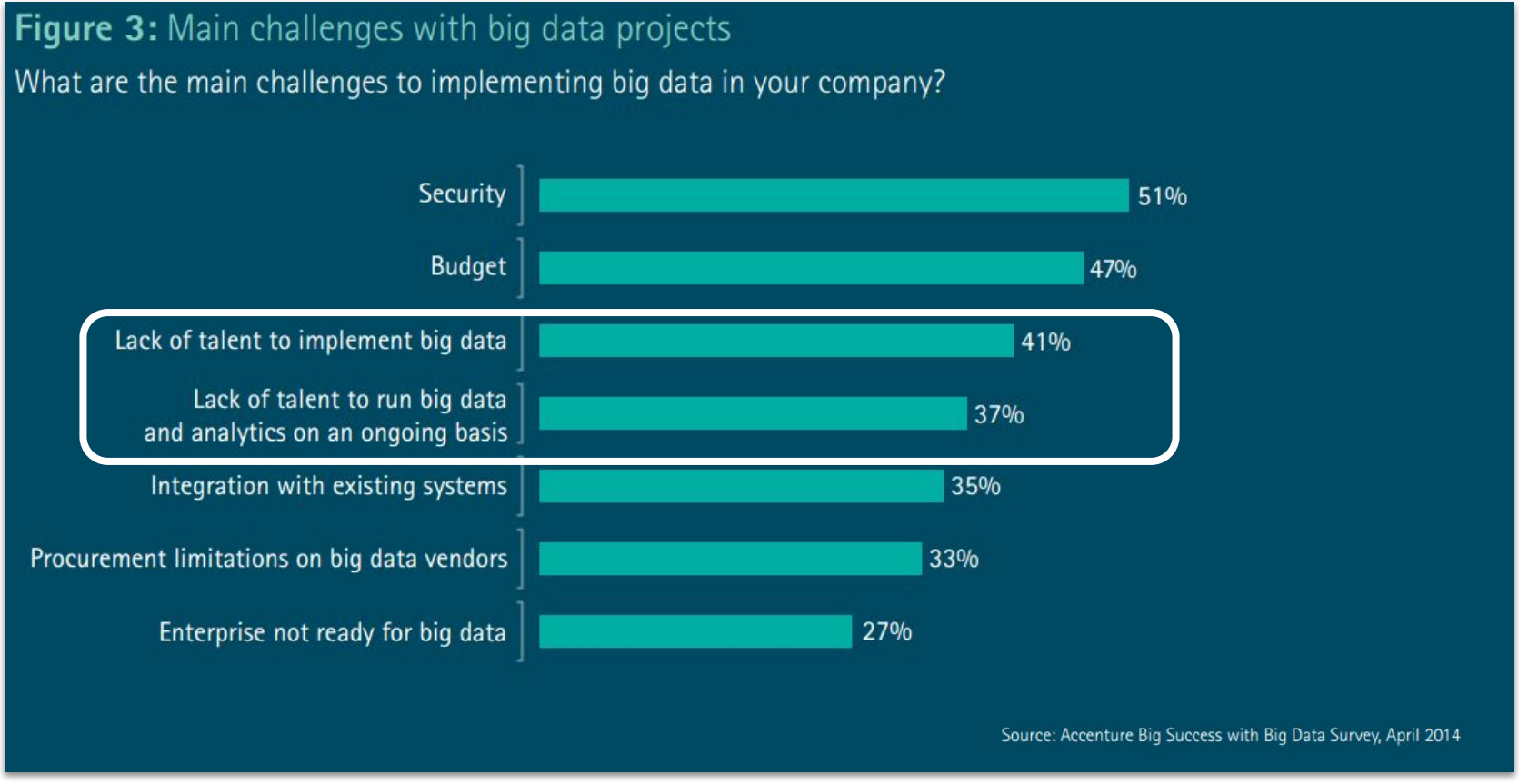
Top publications

Categories > Engineering & Computer Science > **Data Mining & Analysis**

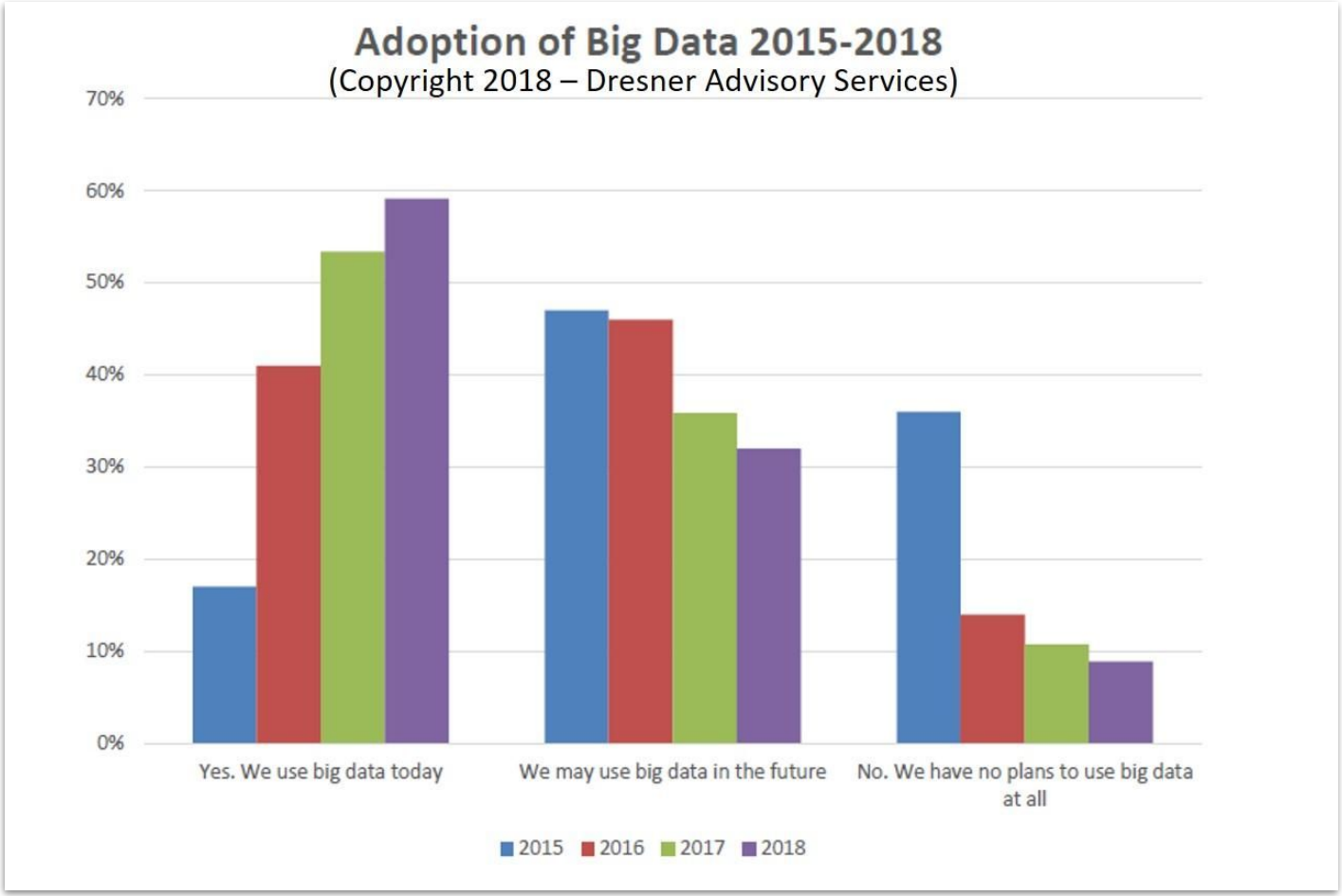
	Publication	h5-index	h5-me
1.	ACM SIGKDD International Conference on Knowledge Discovery & Data Mining	104	183
2.	IEEE Transactions on Knowledge and Data Engineering	87	132
3.	International Conference on Artificial Intelligence and Statistics	68	101
4.	ACM International Conference on Web Search and Data Mining	61	120
5.	IEEE International Conference on Data Mining	54	90
6.	ACM Conference on Recommender Systems	50	84
7.	Knowledge and Information Systems	46	64
8.	IEEE International Conference on Big Data	45	66
9.	Journal of Big Data	42	74
10.	ACM Transactions on Intelligent Systems and Technology (TIST)	40	62
11.	Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery	38	77
12.	Data Mining and Knowledge Discovery	38	68

Big Data, in demand?

Big Data, in demand?



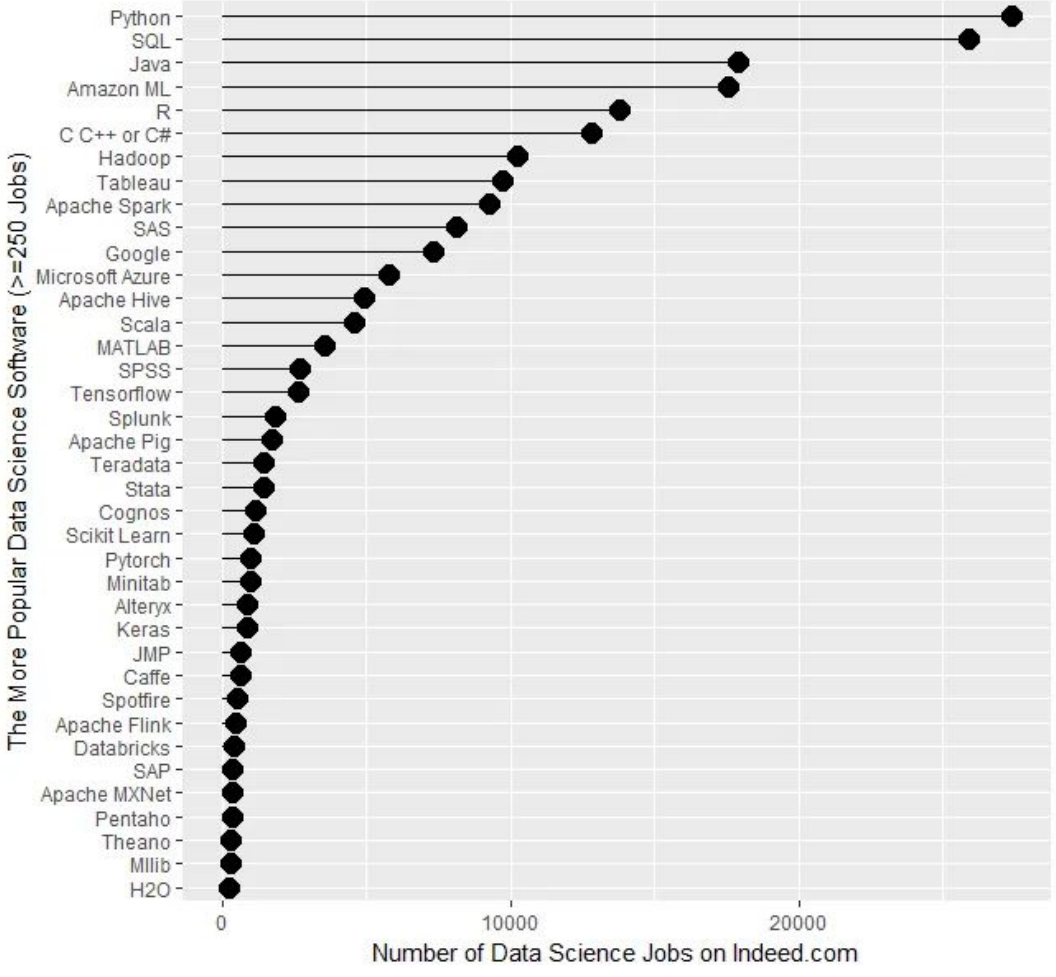
Big Data, in demand?

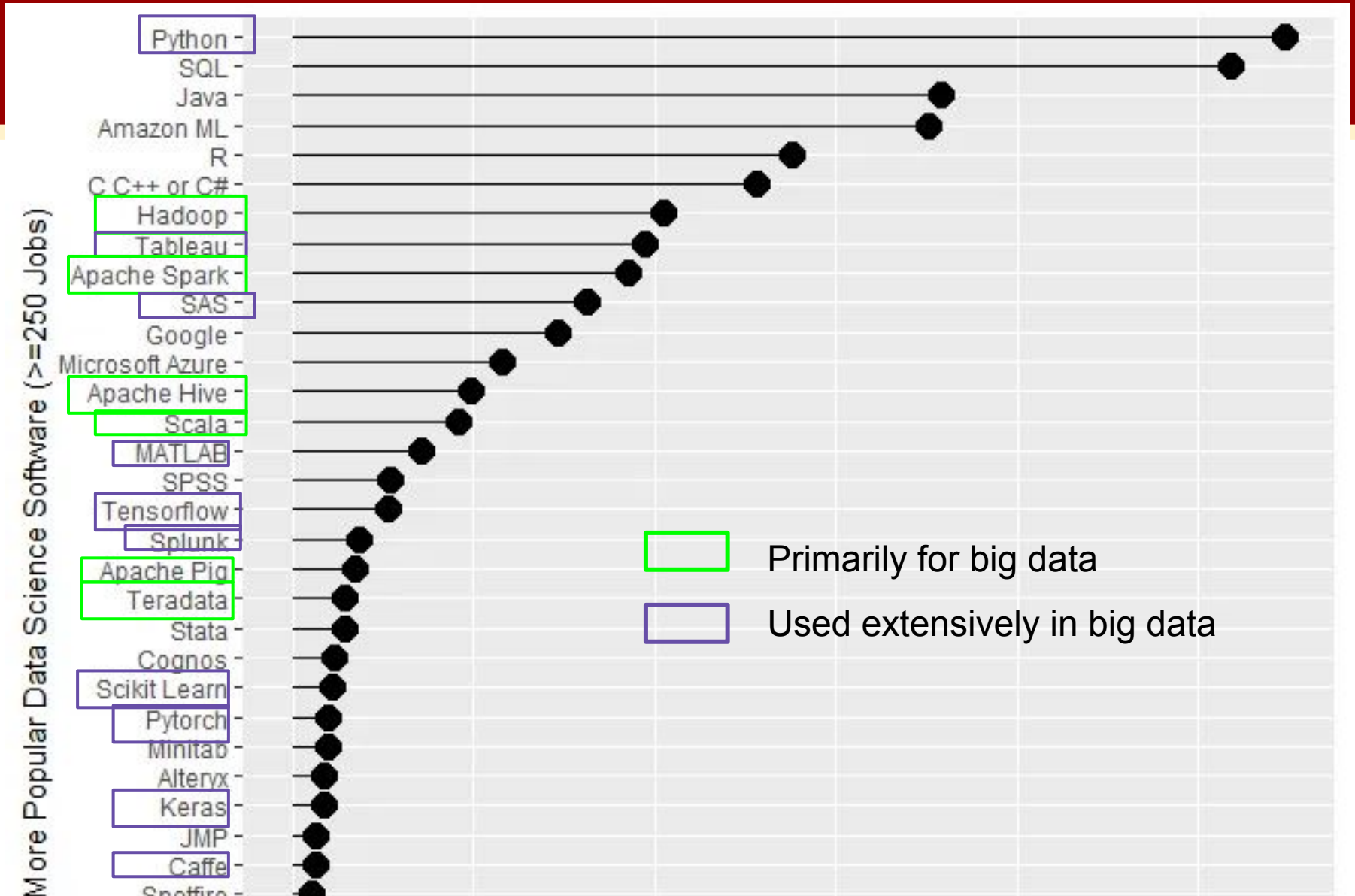


https://www.forbes.com/sites/louiscolumbus/2018/12/23/big-data-analytics-adoption-soared-in-the-enterprise-in-2018/

Big Data, in demand?

By the requirements in job ads. (Muenchen,2019)




More Popular Data Science Software (>=250 Jobs)
Python
SQL
Java
Amazon ML
R
C C++ or C#
Hadoop
Tableau
Apache Spark
SAS
Google
Microsoft Azure
Apache Hive
Scala
MATLAB
SPSS
Tensorflow
Splunk
Apache Pig
Teradata
Stata
Cognos
Scikit Learn
Pytorch
Minitab
Alteryx
Keras
JMP
Caffe
Spotfire
Primarily for big data
Used extensively in big data

Big Data, What is it?

Top big data trends in 2021

Edge computing

Explosive growth in data generated from cloud systems, sensors, smart devices and video streaming is driving adoption of edge computing. Data processing is done on the periphery of the network as close to the originating source as possible.

Cloud and hybrid cloud computing

Cloud computing enables organizations to process nearly limitless amounts of data. Hybrid cloud approaches are being developed to enable companies in regulated industries to take advantage of cloud's economic and technical advantages.

Data lakes

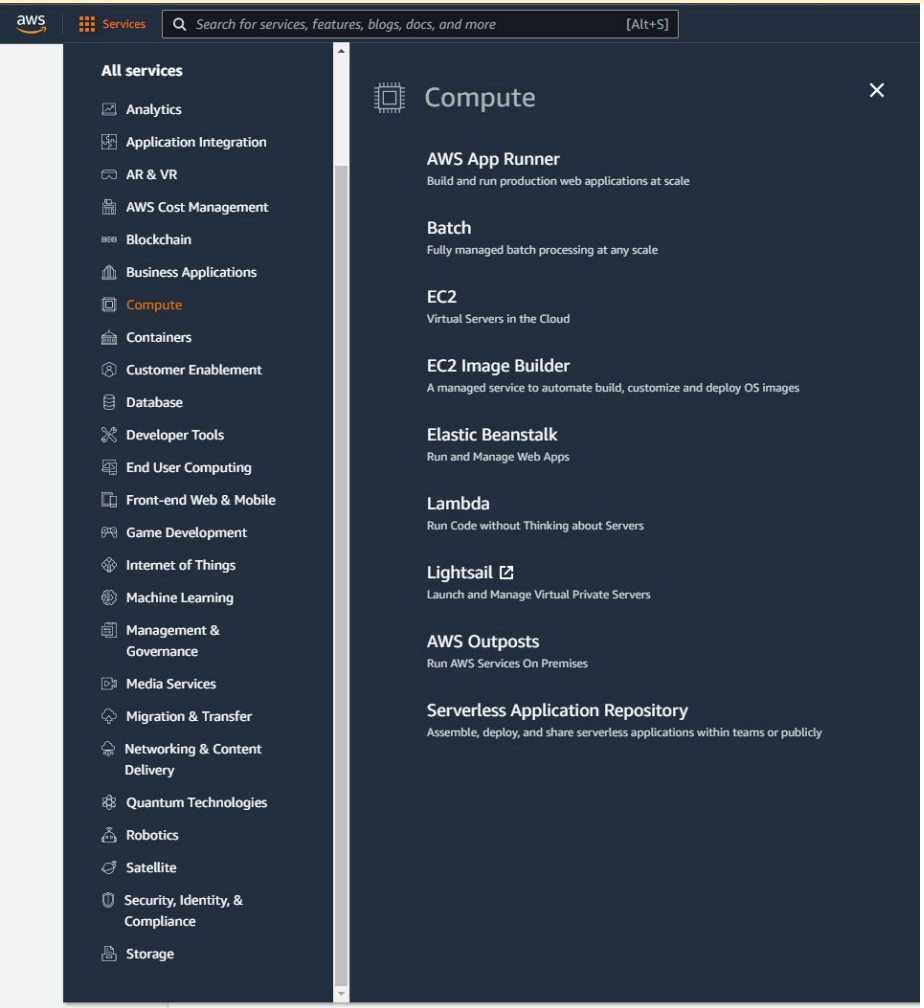
These large repositories store structured and unstructured data in its native format. Data scientists often extract just what's needed for a project, eliminating costly ETL processes required of centralized data warehouses.

Machine learning and AI technologies

Machine learning and other AI technologies are revolutionizing big data analytics. AI's ability to ingest and analyze massive amounts of structured and unstructured data is being used by companies to optimize and improve business operations.

SOURCE: COGNILYTICA; DESIGN: LINDA KOURY; ICONS: LIMEART/GETTY IMAGES

Big Data, What is it?



Libraries, tools and architectures for working with large datasets quickly.

Big Data, What is it?

Short Answer:

Big Data ≈ Data Mining ≈ Predictive Analytics ≈ Data Science

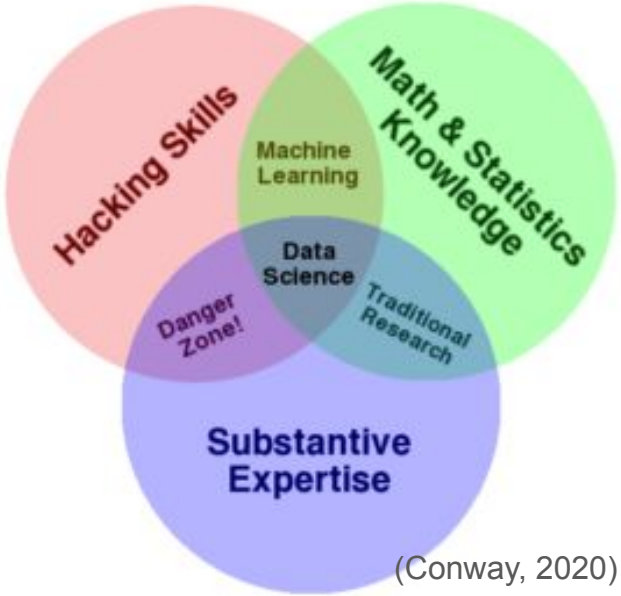
(Leskovec et al., 2017)

Big Data, What is it?

Short Answer:

Big Data ≈ Data Mining ≈ Predictive Analytics ≈ Data Science

(Leskovec et al., 2017)



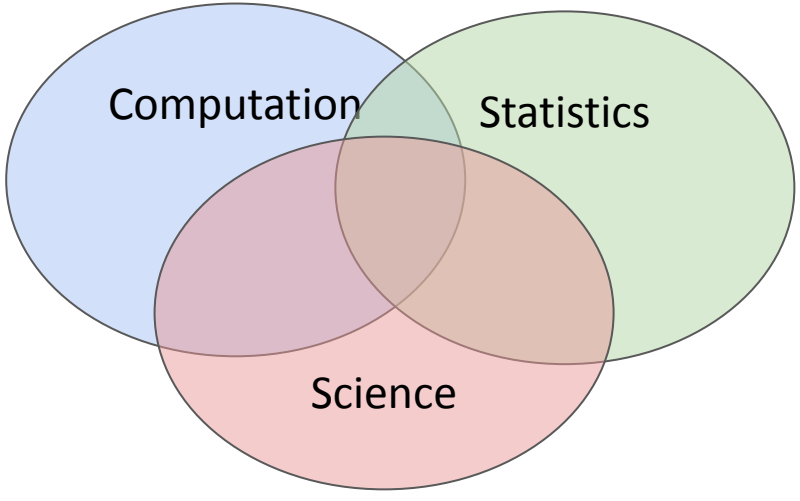
(Conway, 2020)

Big Data, What is it?

Short Answer:

Big Data ≈ Data Mining ≈ Predictive Analytics ≈ Data Science

(Leskovec et al., 2017)



Big Data, What is it?

Goal: Generalizations
A *model* or *summarization* of the data.

E.g.

- Google's PageRank: *summarizes* web pages by a single number.
- Twitter financial market predictions: *Models* the stock market according to shifts in sentiment in Twitter.
- Distinguish tissue type in medical images: *Summarizes* millions of pixels into clusters.
- Mental health diagnosis in social media: *Models* presence of diagnosis as a distribution (a summary) of linguistic patterns.
- Frequent co-occurring purchases: *Summarize* billions of purchases as items that frequently are bought together.

Big Data, What is it?

Goal: Generalizations
A *model* or *summarization* of the data.

1. Descriptive analytics
   Describe (*generalizes*) the data itself

2. Predictive analytics
   Create something *generalizeable* to new data

Big Data Analytics, The Class

Goal: Generalizations
A *model* or *summarization* of the data.

Data/Workflow Frameworks

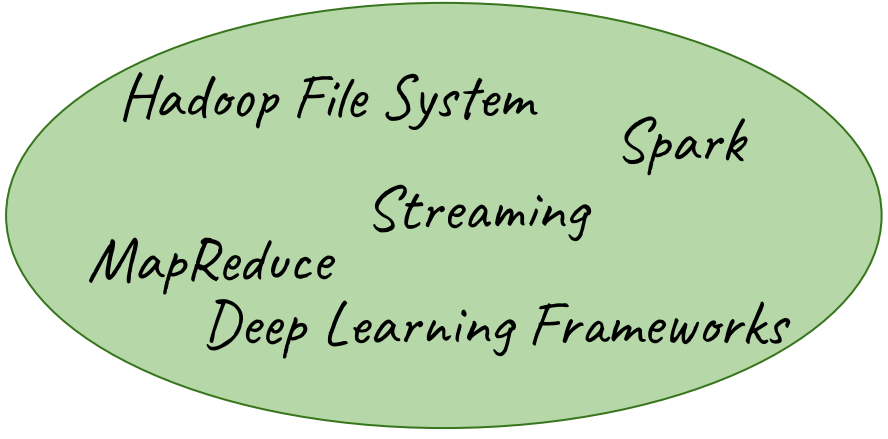
Analytics and Algorithms

Big Data Analytics, The Class

Goal: Generalizations
A *model* or *summarization* of the data.

Data/Workflow Frameworks

Analytics and Algorithms

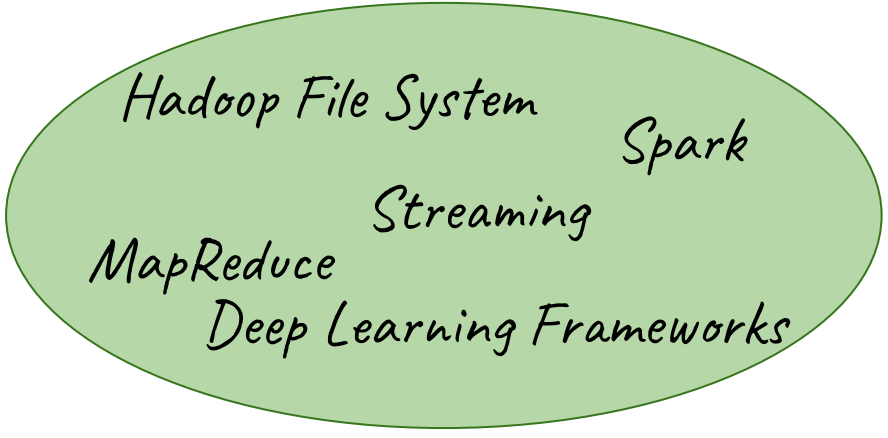


Big Data Analytics, The Class

Goal: Generalizations
A *model* or *summarization* of the data.

Data/Workflow Frameworks

Analytics and Algorithms



Any Questions ??