Big Data Analytics: What is Big Data?

H. Andrew Schwartz

CSE545
Spring 2022

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

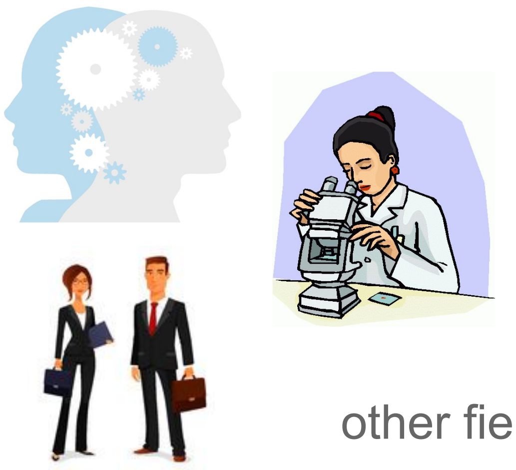
Big Data, what is it?

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. > 100 subjects); can't analyze in stats tools (Excel).

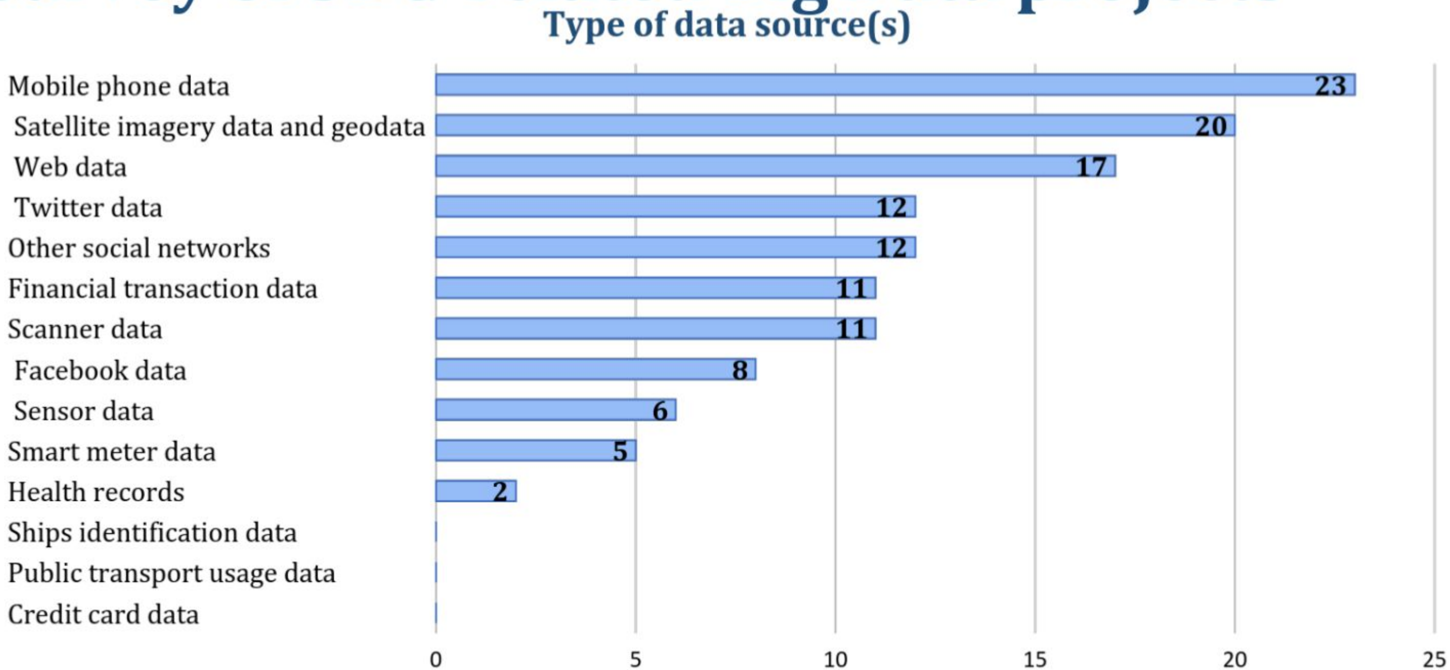
Big Data, what is it? *Government View*

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

(2016)

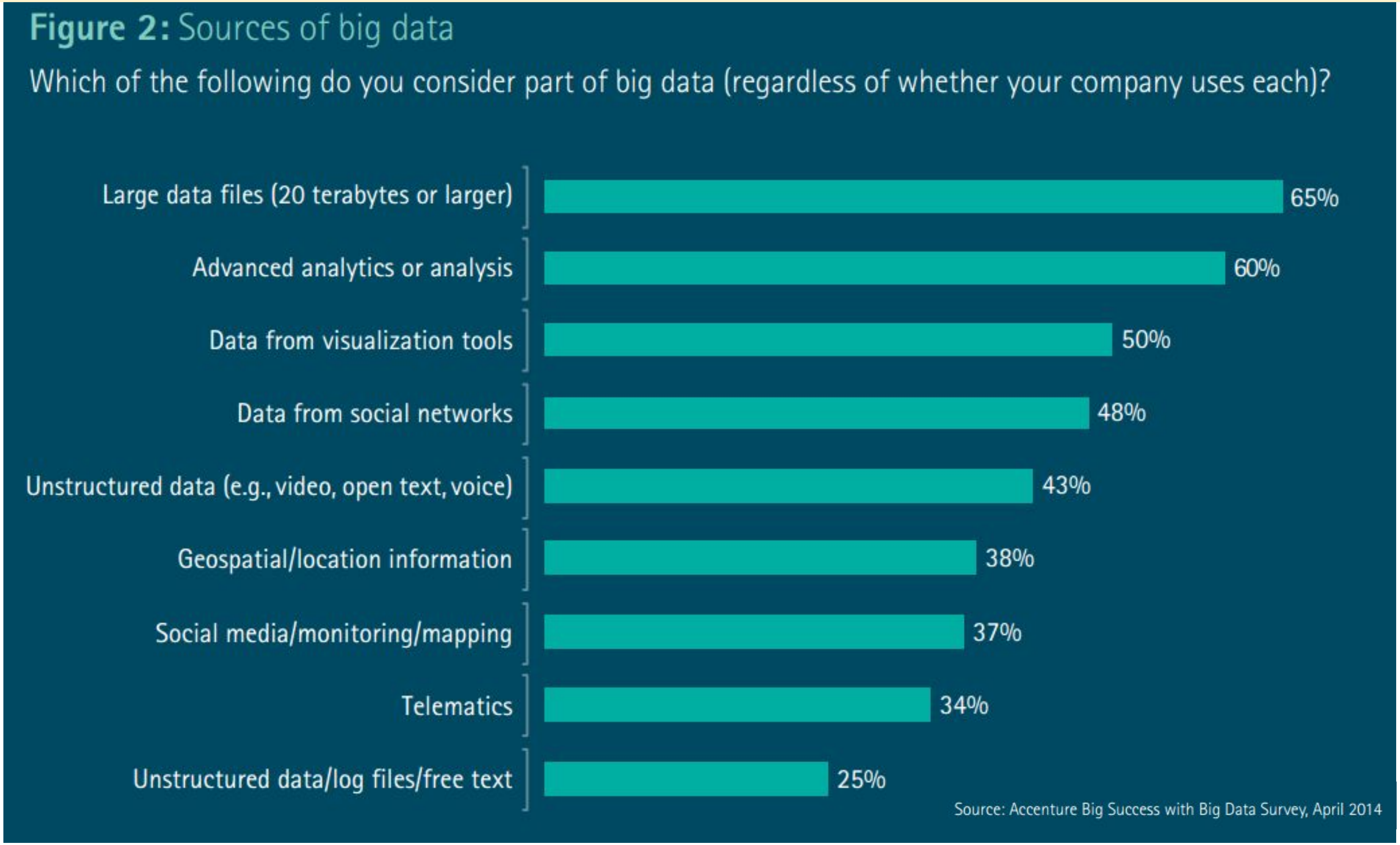
BigData
UN Global Working Group

1. Survey of SDG-related Big Data projects



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

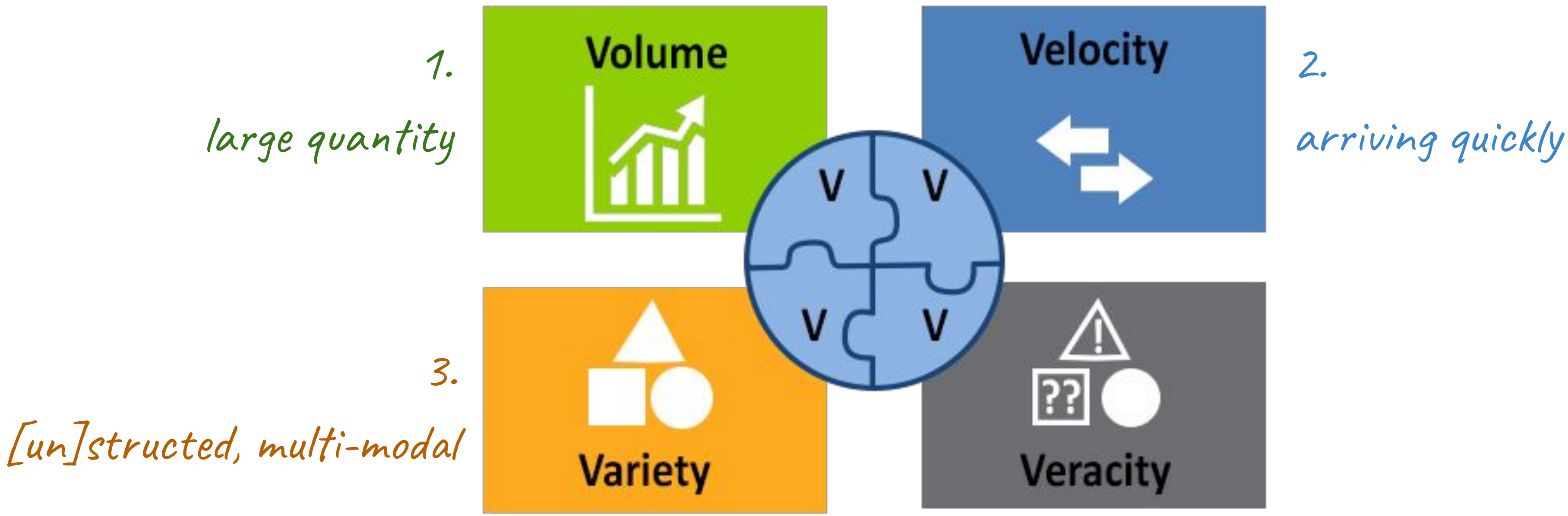
Big Data, what is it? *Industry View*



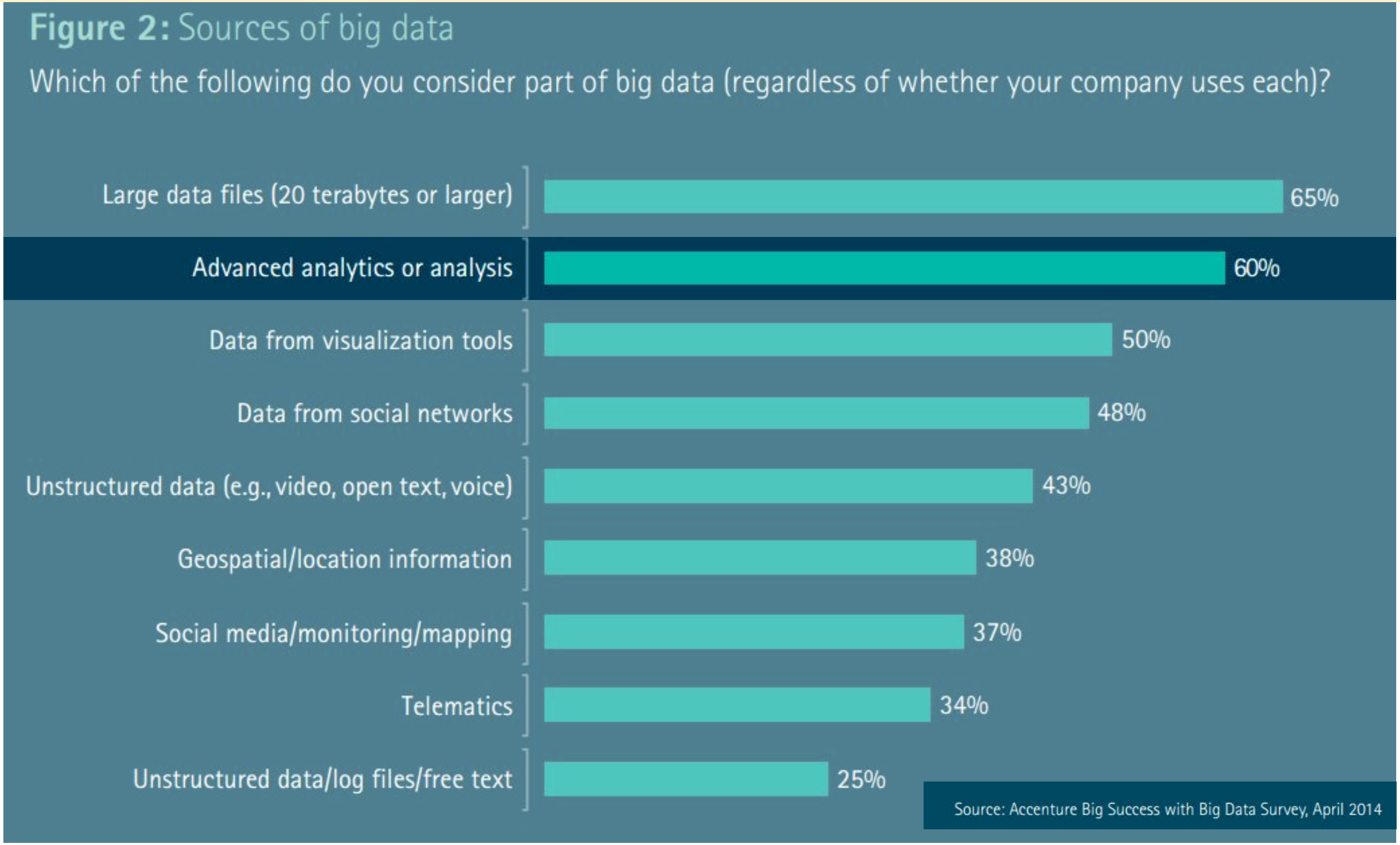
Big Data, what is it?

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

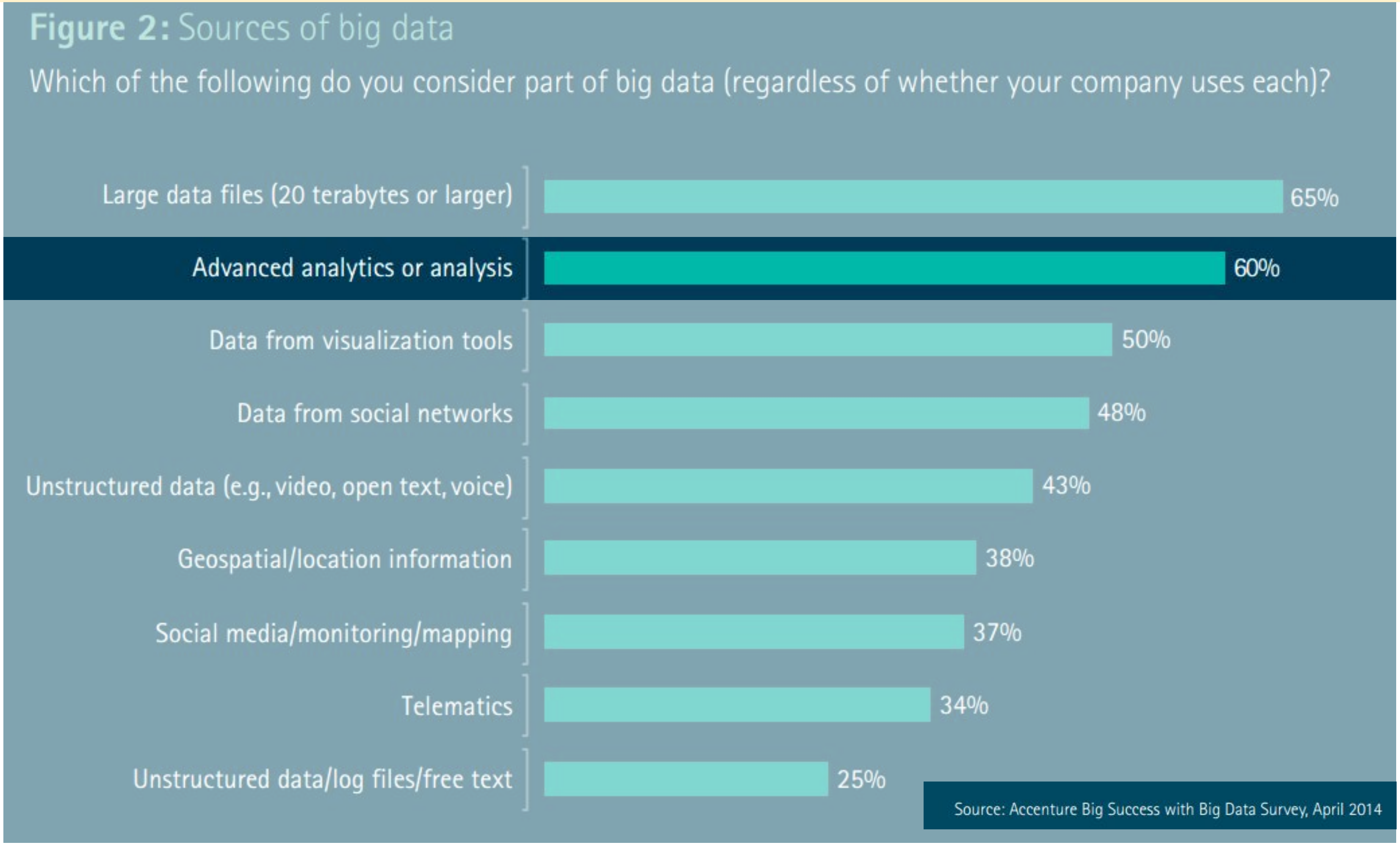
(Laney, 2001: META Group)



Big Data, what is it? *Industry View*



Big Data, a type of analytics



Big Data, a type of analytics

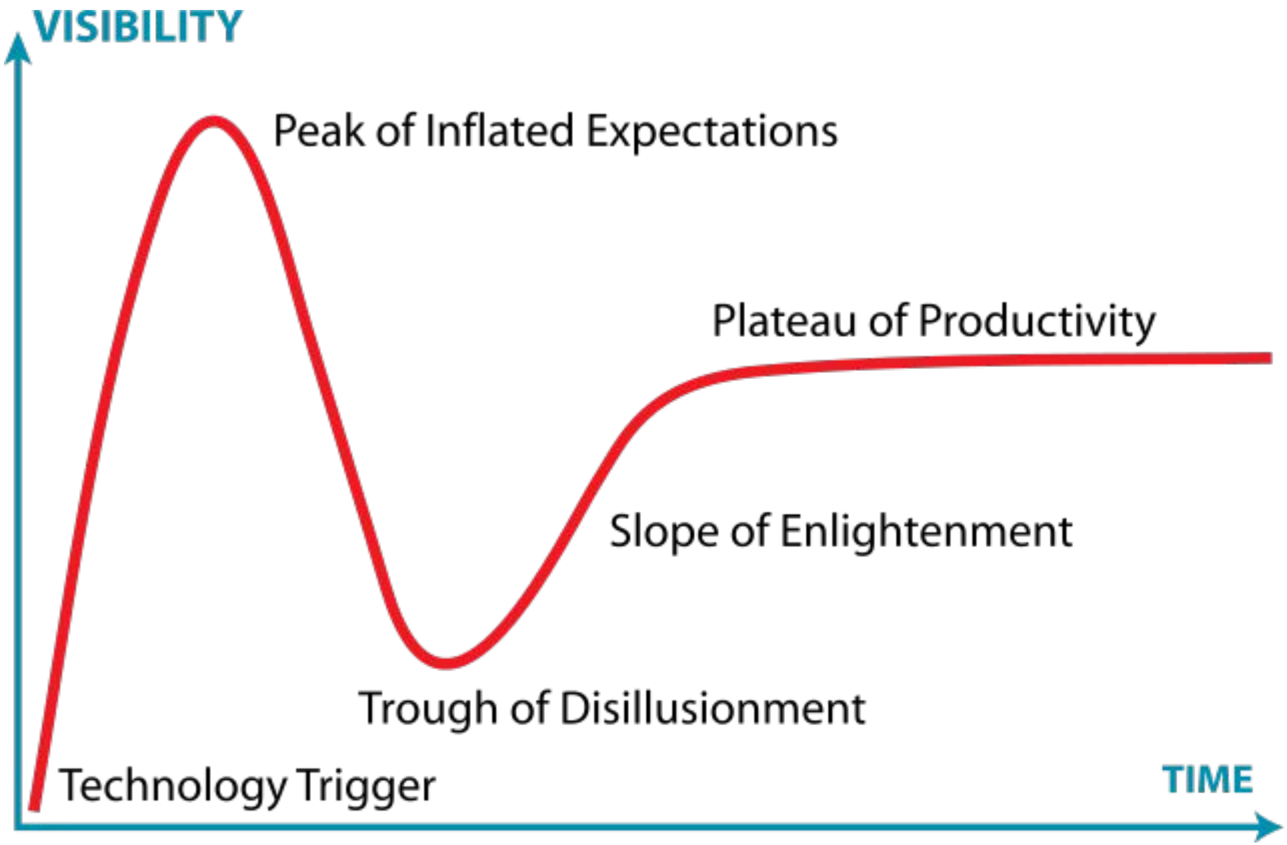
?

Big Data, a type of analytics

Big Data, a type of analytics

Big Data, a type of analytics

Big Data, a buzz word?



(Gartner Hype Cycle)

Big Data, a buzz word?

2008

2011

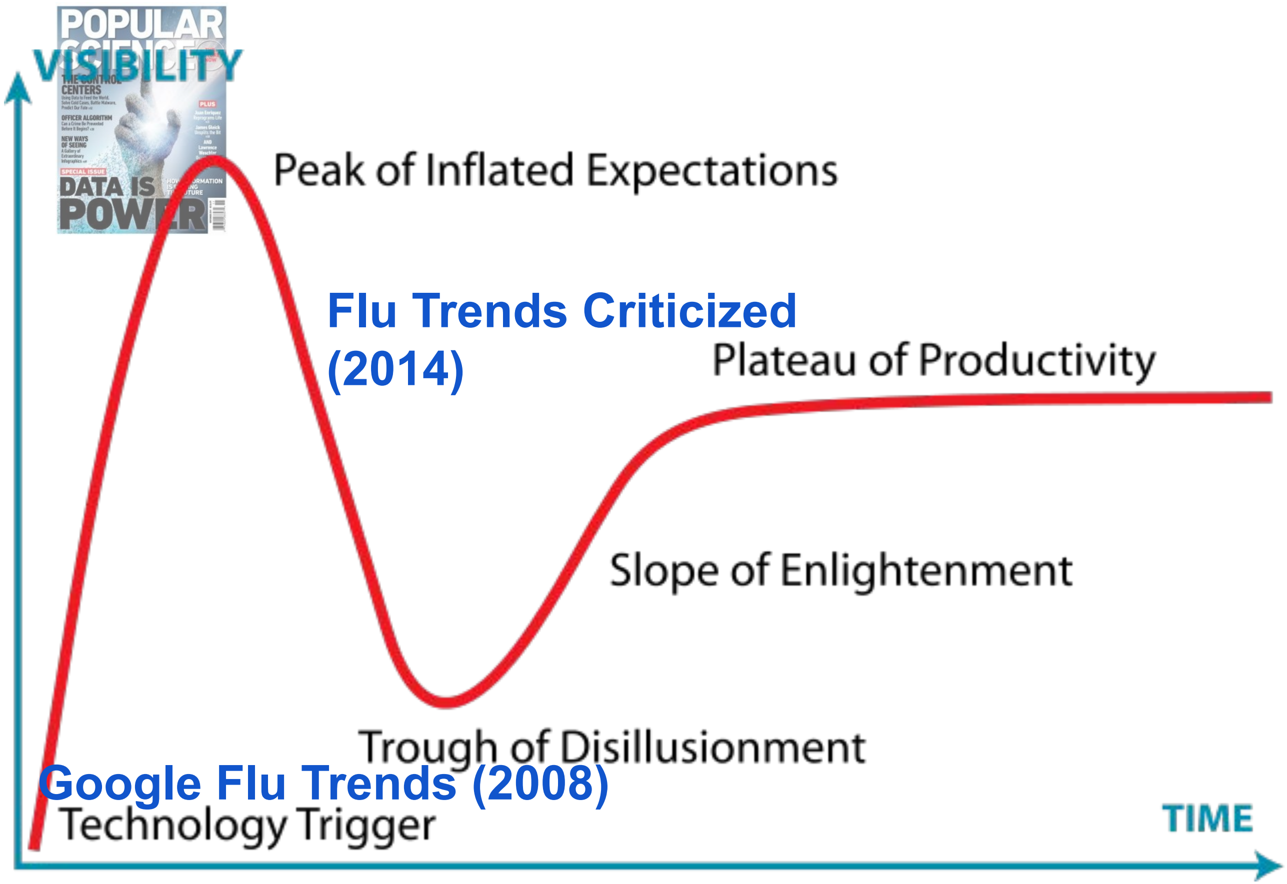
2012

2010

2011

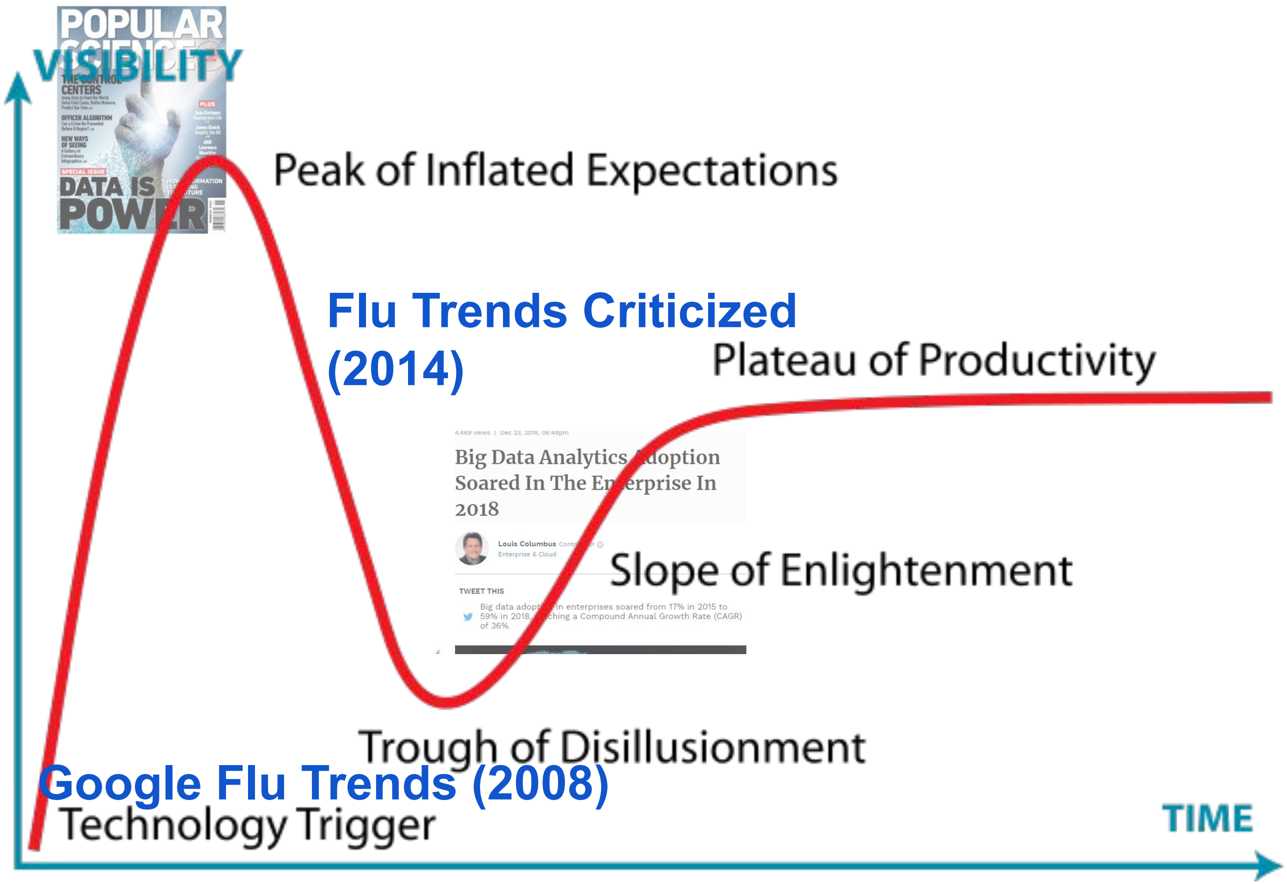
2018

Big Data, a buzz word?



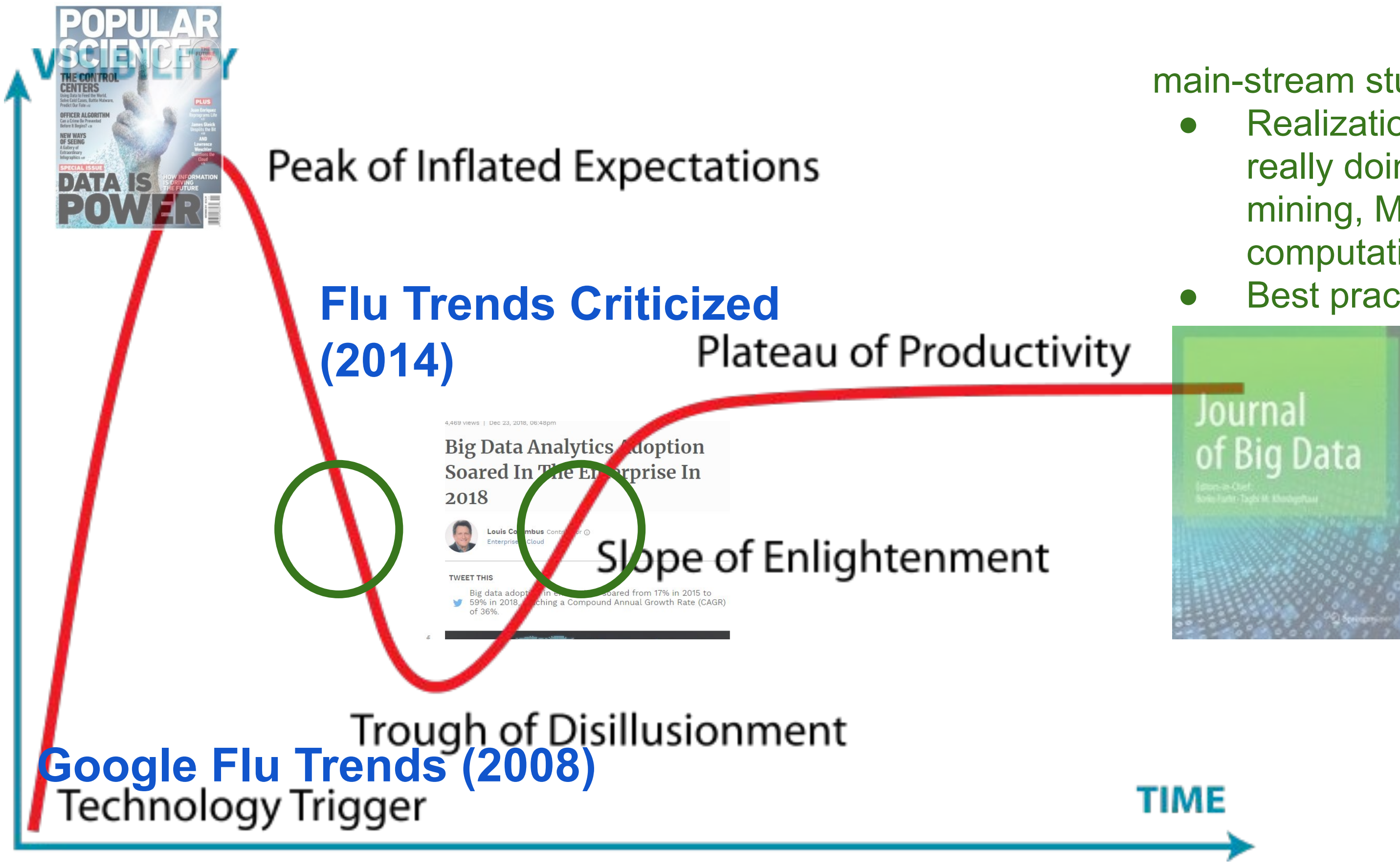
(Gartner Hype Cycle)

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

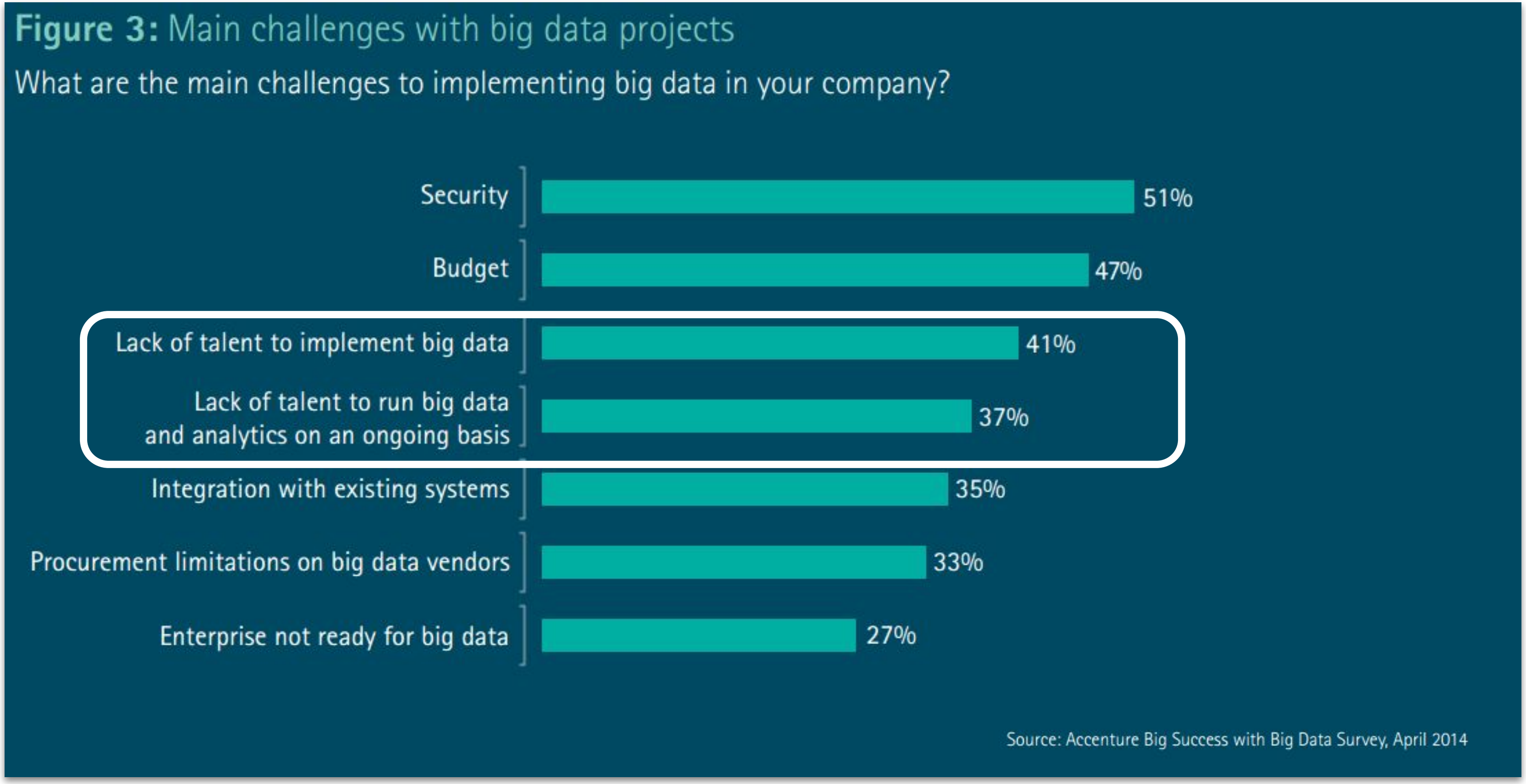
Top publications

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

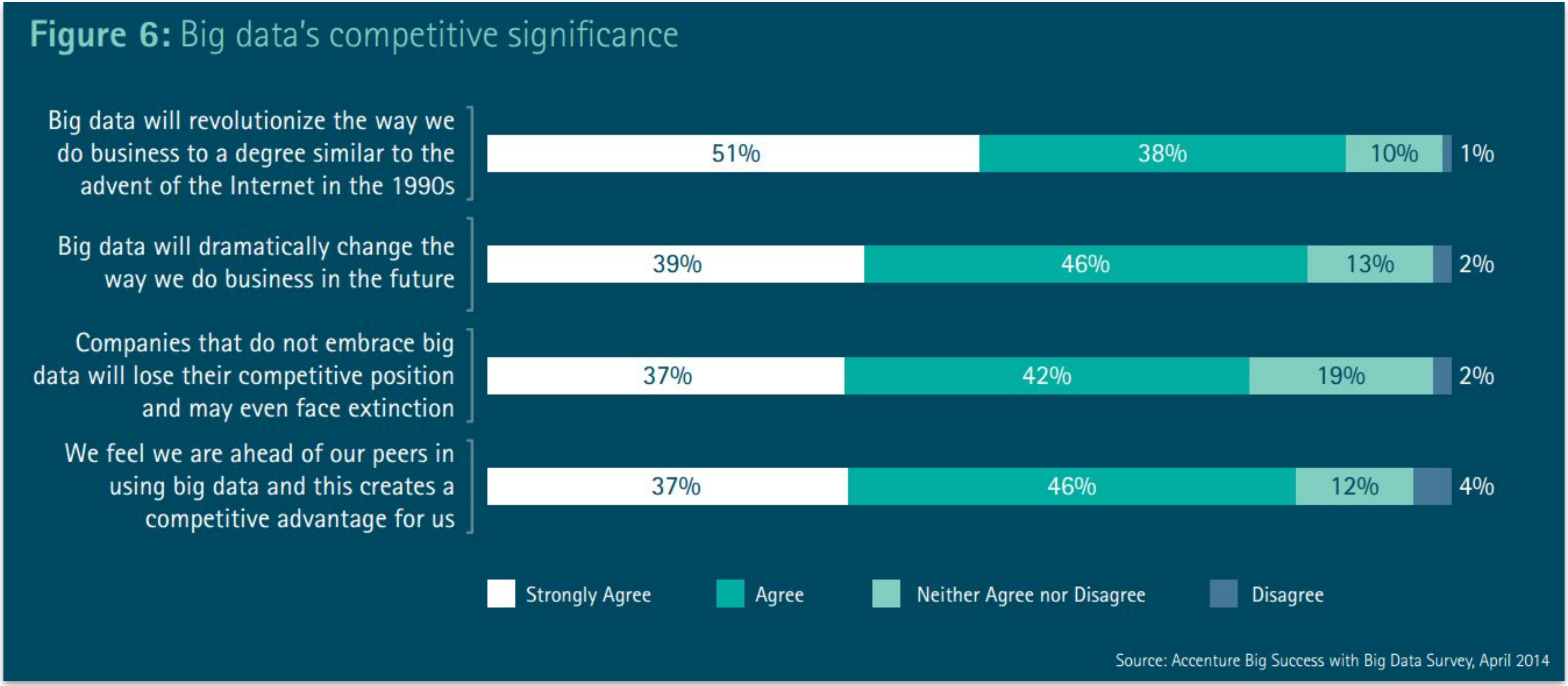
	Publication	h5-index	h5-med
1.	ACM SIGKDD International Conference on Knowledge Discovery & Data Mining	104	183
2.	IEEE Transactions on Knowledge and Data Engineering	87	132
3.	International Conference on Artificial Intelligence and Statistics	68	101
4.	ACM International Conference on Web Search and 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?

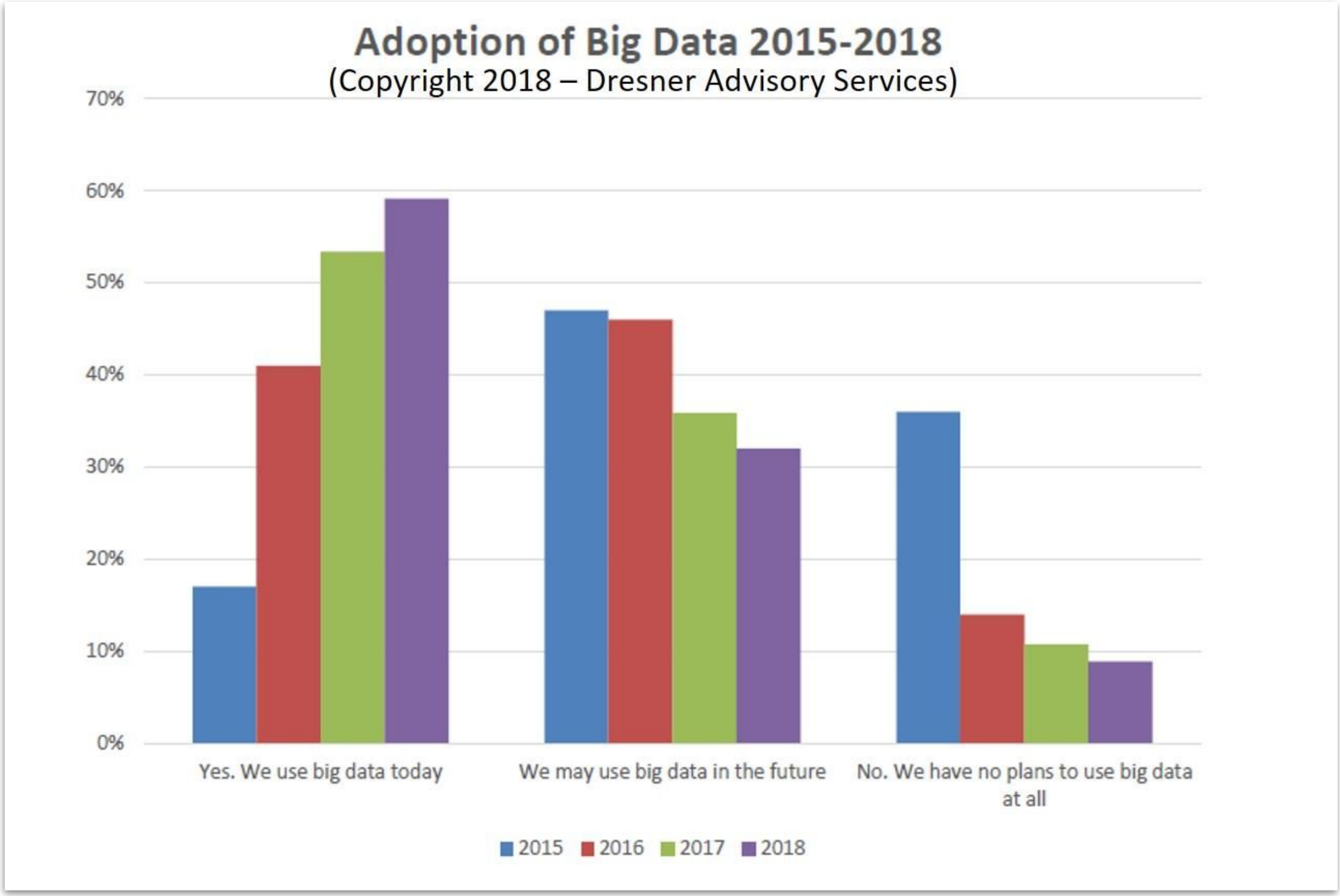
Big Data, in demand?



Big Data, in demand?



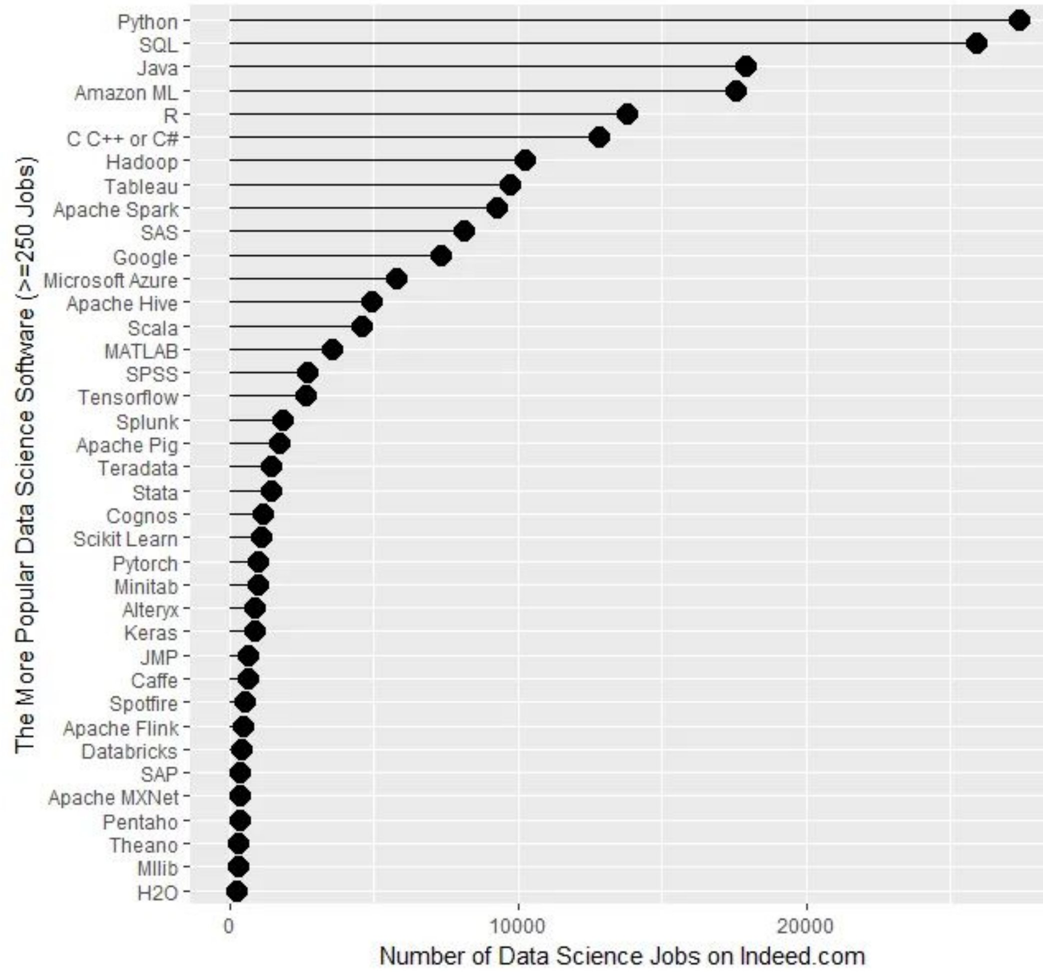
Big Data, in demand?

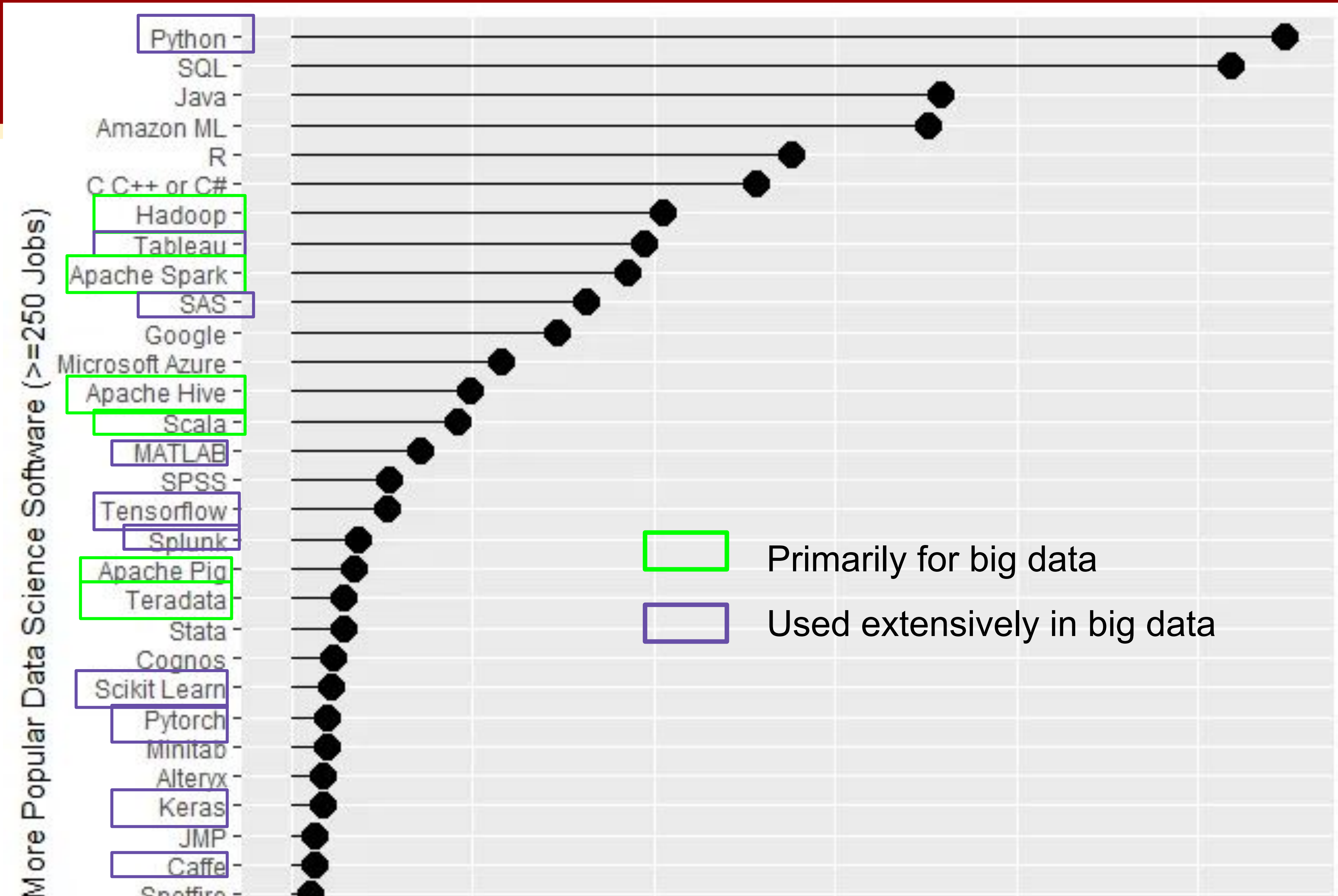


https://www.forbes.com/sites/louiscolumbus/2018/12/23/big-data-analytics-adoption-soared-in-the-enterprise-in-2018/

Big Data, in demand?

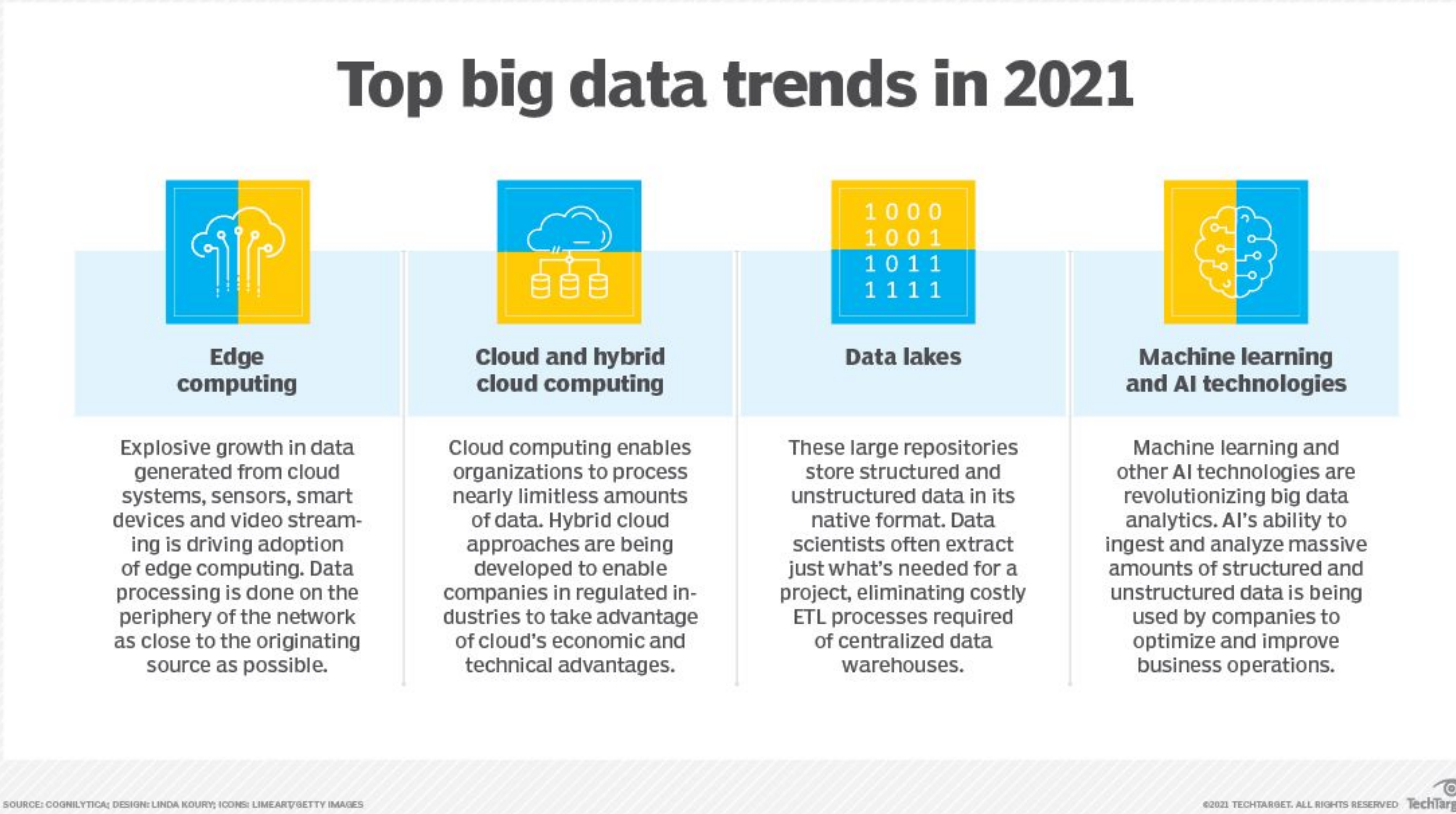
By the requirements in job ads. (Muenchen,2019)



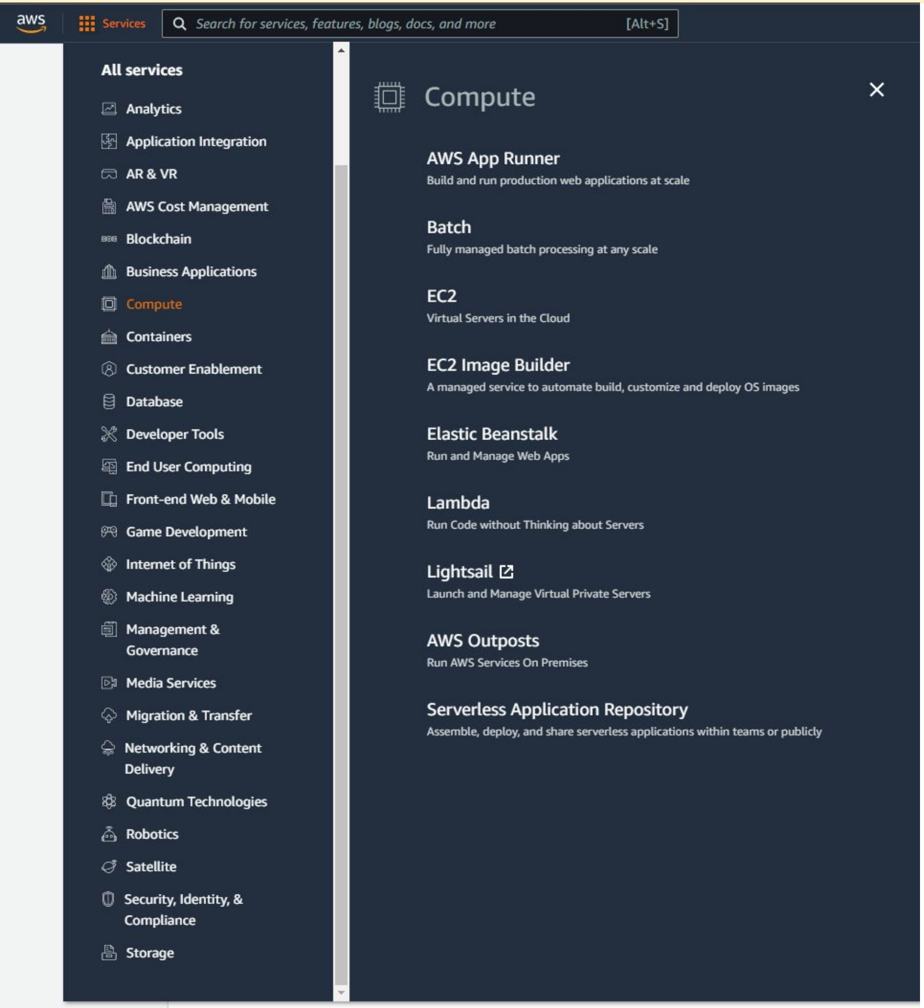


More Popular Data Science Software (>=250 Jobs)
Python
SQL
Java
Amazon ML
R
C C++ or C#
Hadoop
Tableau
Apache Spark
SAS
Google
Microsoft Azure
Apache Hive
Scala
MATLAB
SPSS
Tensorflow
Splunk
Apache Pig
Teradata
Stata
Cognos
Scikit Learn
Pytorch
Minitab
Alteryx
Keras
JMP
Caffe
Primarily for big data
Used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

Big Data, What is it?



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

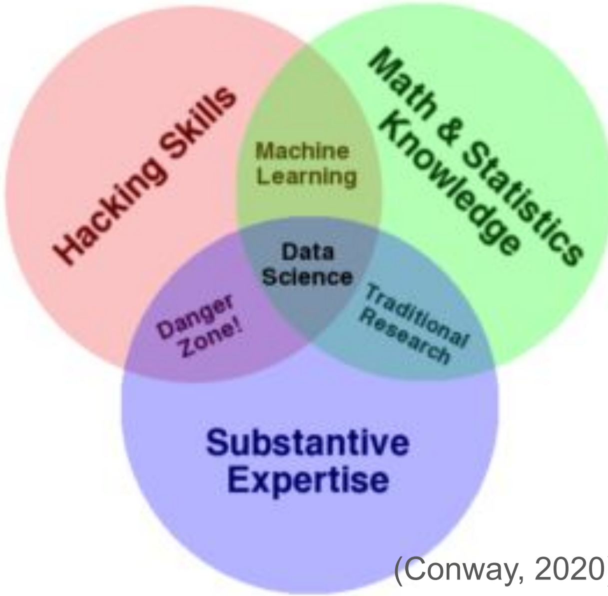
(Leskovec et al., 2017)

Big Data, What is it?

Short Answer:

Big Data ≈ Data Mining ≈ Predictive Analytics ≈ Data Science

(Leskovec et al., 2017)



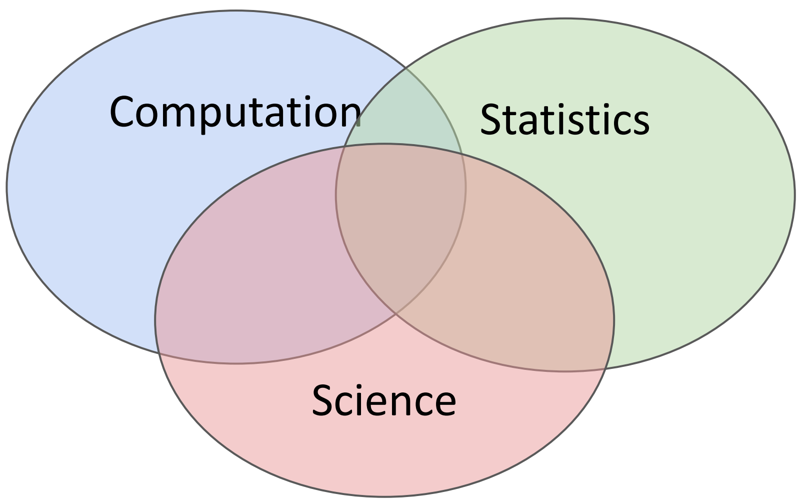
(Conway, 2020)

Big Data, What is it?

Short Answer:

Big Data ≈ Data Mining ≈ Predictive Analytics ≈ Data Science

(Leskovec et al., 2017)



Big Data, What is it?

Short Answer:

Big Data ≈ Data Mining ≈ Predictive Analytics ≈ Data Science

(Leskovec et al., 2017)

CSE545 focuses on:

How to analyze data that is mostly too large for main memory.

Analyses only possible with a *large* number of observations or features.

Big Data, What is it?

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

How to analyze data that is mostly too large for main memory.

Analyses only possible with a *large* number of observations or features.

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

Core Data Science Courses

CSE 519: Data Science Fundamentals

CSE 544: Prob/Stat for Data Scientists

CSE 545: Big Data Analytics

CSE 512: Machine Learning

CSE 537: Artificial Intelligence

CSE 548: Analysis of Algorithms

CSE 564: Visualization

Applications of Data Science

CSE 527:
Computer Vision

CSE 538:
Natural Language Processing

CSE 549:
Computational Biology

...

Big Data Analytics, The Class

Core Data Science Courses

CSE 519: Data Science Fundamentals

CSE 544: Prob/Stat for Data Scientists

CSE 545: Big Data Analytics

CSE 512: Machine Learning

CSE 537: Artificial Intelligence

CSE 548: Analysis of Algorithms

CSE 564: Visualization

Applications of Data Science

CSE 527:
Computer Vision

CSE 538:
Natural Language Processing

CSE 549:
Computational Biology

...

Key Distinction:
??

Big Data Analytics, The Class

Core Data Science Courses

CSE 519: Data Science Fundamentals

CSE 544: Prob/Stat for Data Scientists

CSE 545: Big Data Analytics

CSE 512: Machine Learning

CSE 537: Artificial Intelligence

CSE 548: Analysis of Algorithms

CSE 564: Visualization

Applications of Data Science

CSE 527:
Computer Vision

CSE 538:
Natural Language Processing

CSE 549:
Computational Biology

...

Key Distinction:
Focus on scalability and algorithms/analyses not possible without large data.

Big Data Analytics, The Class

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

Data/Workflow Frameworks

Analyses and Algorithms

Big Data Analytics, The Class

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

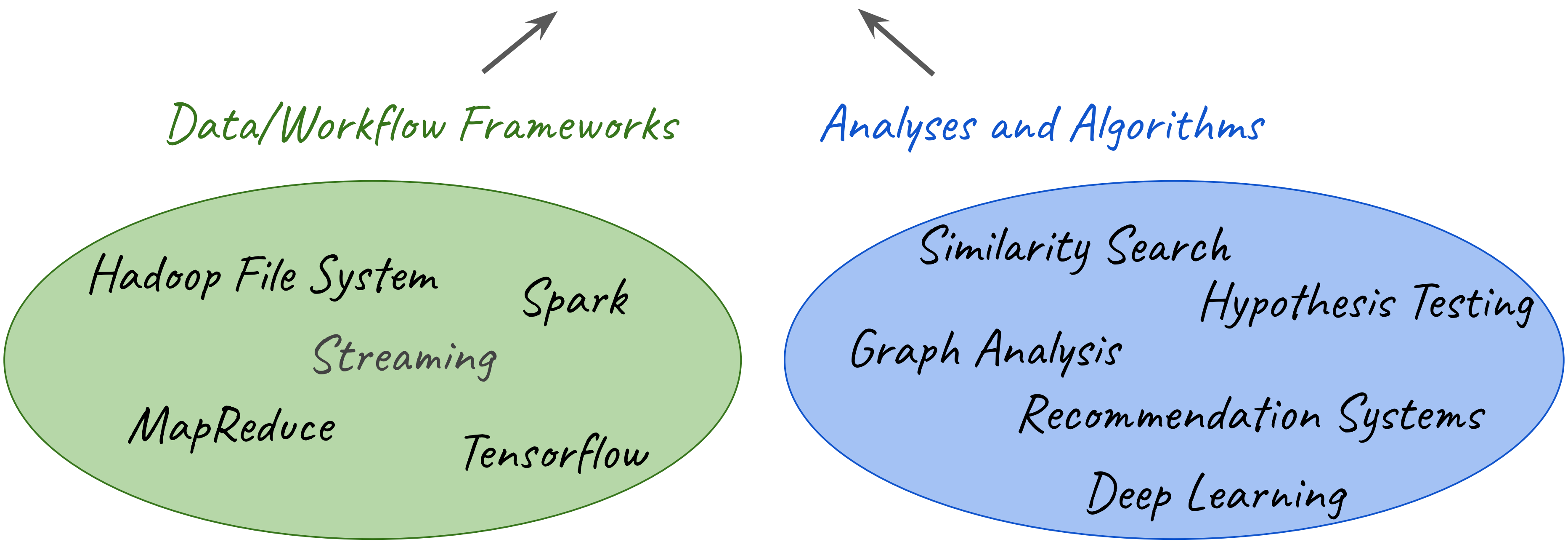
Data/Workflow Frameworks

Hadoop File System
Spark
Streaming
MapReduce
Tensorflow

Analyses and Algorithms

Big Data Analytics, The Class

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



Big Data Analytics, The Class

http://www3.cs.stonybrook.edu/~has/CSE545/

Big Data Analytics, The Class

How to succeed:

1. Do the weekly readings see [syllabus]
2. Take notes associated with the lectures. If needed:
   a. consult lecture recordings in Blackboard.
   b. watch recordings from MMDS website
3. Practice exercises in the back of each reading.
4. Attend class and actively participate.
5. Begin assignments early and seek help if trouble (e.g. office hours).

Preliminaries

Ideas and methods that will repeatedly appear:

- Normalization (TF.IDF)
- Power Laws
- Hash functions
- IO Boundedness (Secondary Storage)
- Unstructured Data
- Probability Theory
- **Bonferroni's Principle**

Normalization

Count data often need *normalizing* -- putting the numbers on the same “scale”.

Prototypical example: TF.IDF

Normalization

Count data often need *normalizing* -- putting the numbers on the same "scale".

Prototypical example: TF.IDF of word *i* in document *j:*

Term Frequency:

$$tf_{ij} = \frac{count_{ij}}{\max_k count_{kj}}$$

$$tf.idf_{ij} = tf_{ij} \times idf_i$$

Inverse Document Frequency:

$$idf_i = log_2(\frac{docs_*}{docs_i}) \propto \frac{1}{\frac{docs_i}{docs_*}}$$

where docs is the number of documents containing word *i*.

Normalization

Count data often need *normalizing* -- putting the numbers on the same “scale”.

Prototypical example: TF.IDF of word *i* in document *j:*

Term Frequency:

$$tf_{ij} = \frac{count_{ij}}{\boxed{\max_k count_{kj}}}$$

$$tf.idf_{ij} = tf_{ij} \times idf_i$$

Inverse Document Frequency:

$$idf_i = \boxed{log_2(}\frac{docs_*}{docs_i}) \propto \frac{1}{\boxed{\frac{docs_i}{docs_*}}}$$

where docs is the number of documents containing word *i*.

Normalization

Standardize: puts different sets of data (typically vectors or random variables) on the same scale with the same center.

- Subtract the mean (i.e. “mean center”)
- Divide by standard deviation

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

Power Law

Characterized many frequency patterns when ordered from most to least:

County Populations [r-bloggers.com]

links into webpages [Broader et al., 2000]

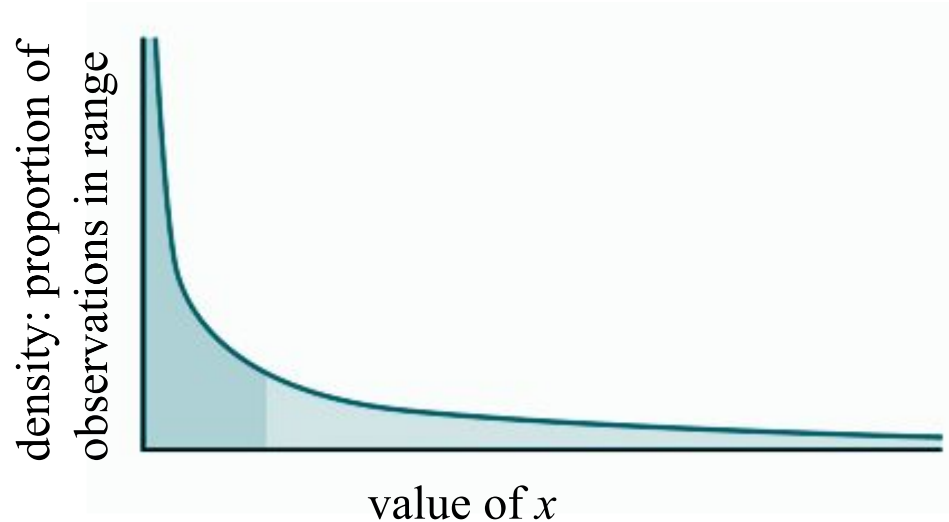
Sales of products [see book]

Frequency of words [Wikipedia, “Zipf’s Law”]

(“popularity” based statistics, especially without limits)

Power Law

$$\log y = b + a \log x$$



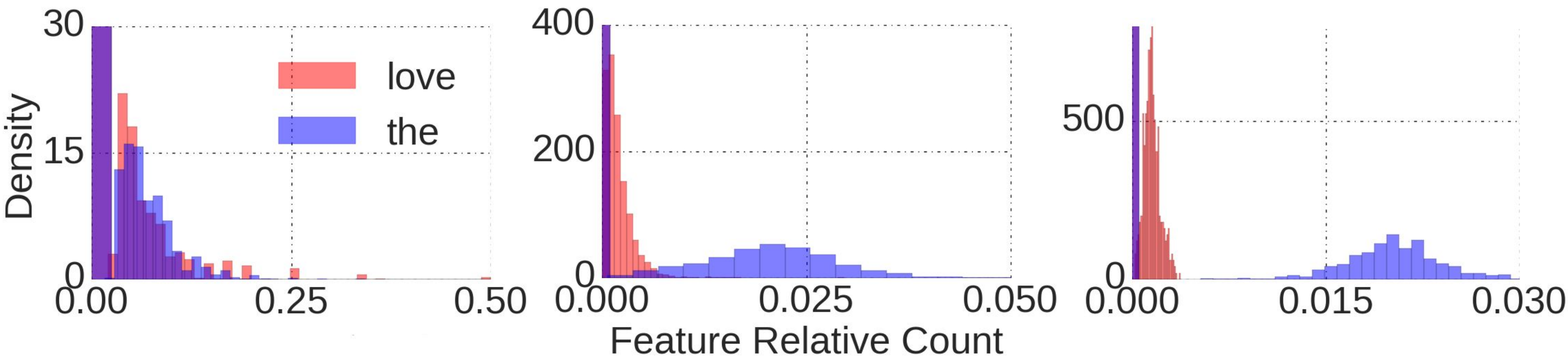
raising to the natural log:

$$y = e^b e^{a \log x} = e^b x^a = cx^a$$

where c is just a constant

Characterizes “the Matthew Effect” -- the rich get richer

Power Law



message-level user-level county-level

Almodaresi, F., Ungar, L., Kulkarni, V., Zakeri, M., Giorgi, S., & Schwartz, H. A. (2017). On the Distribution of Lexical Features at Multiple Levels of Analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 79-84).

Hash Functions and Indexes

Review:

h: *hash-key -> bucket-number*

Objective: uniformly distribute hash-keys across buckets.

Example: storing word counts.

Hash Functions and Indexes

Review:

h: *hash-key -> bucket-number*

Objective: uniformly distribute hash-keys across buckets.

Example: storing word counts.

$$h(word) = \left(\sum_{char \in word} ascii(char) \right) \% \, \#buckets$$

Hash Functions and Indexes

Review:

h: *hash-key -> bucket-number*

Objective: uniformly distribute hash-keys across buckets.

Example: storing word counts.

$$h(word) = \left(\sum_{char \in word} ascii(char) \right) \% \#buckets$$

Data structures utilizing hash-tables (i.e. O(1) lookup; dictionaries, sets in python) are a friend of big data algorithms! Review further if needed.

Hash Functions and Indexes

Review:

h: *hash-key -> bucket-number*

Objective: uniformly distribute hash-keys across buckets.

Example: storing word counts.

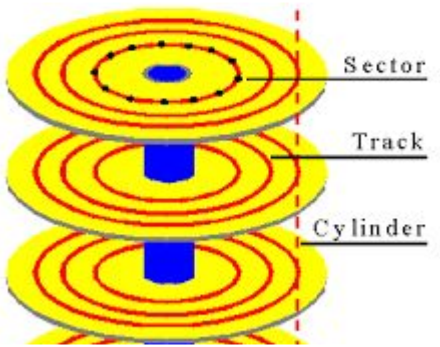
Database Indexes: Retrieve all records with a given *value.* (also review if unfamiliar / forgot)

Data structures utilizing hash-tables (i.e. O(1) lookup; dictionaries, sets in python) are a friend of big data algorithms! Review further if needed.

IO Bounded

Reading a word from disk versus main memory: 10^5 slower!

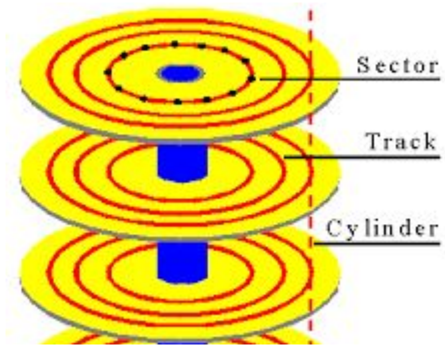
Reading many contiguously stored words is faster per word, but fast modern disks still only reach 150MB/s for sequential reads.



IO Bounded

Reading a word from disk versus main memory: 10^5 slower!

Reading many contiguously stored words is faster per word, but fast modern disks still only reach 150MB/s for sequential reads.



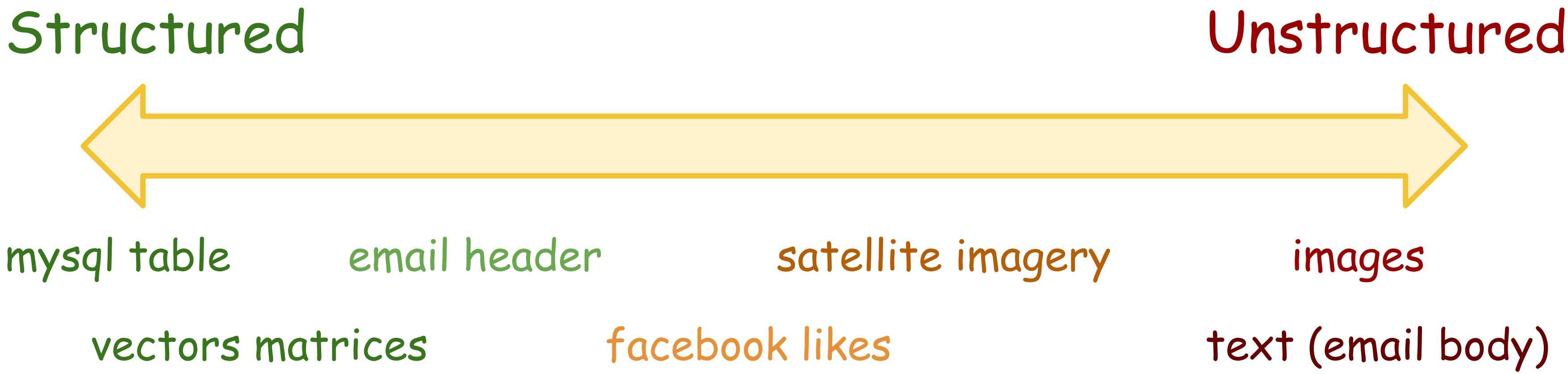
IO Bound: biggest performance bottleneck is reading / writing to disk.

(starts around 100 GBs; ~10 minutes just to read).

Unstructured Data Continuum

- Unstructured ≈ requires processing to get what is of interest
- Feature extraction used to turn unstructured into structured
- Near infinite amounts of potential features in unstructured data

Unstructured Data Continuum



- Unstructured ≈ requires processing to get what is of interest
- Feature extraction used to turn unstructured into structured
- Near infinite amounts of potential features in unstructured data

Bonferroni's Principle

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

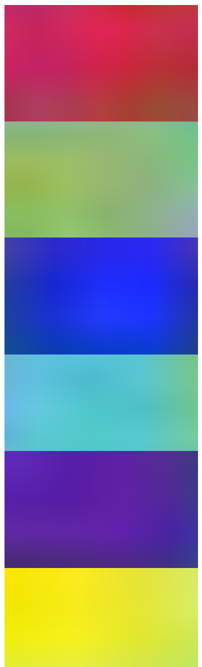
Generalize: Find patterns that didn't just happen by chance.

Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:

- Red
- Green
- Blue
- Teal
- Purple
- Yellow

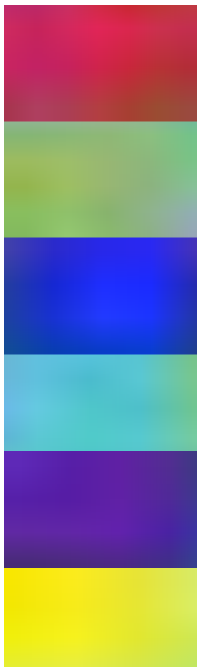


Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:

Red
Green
Blue
Teal
Purple
Yellow



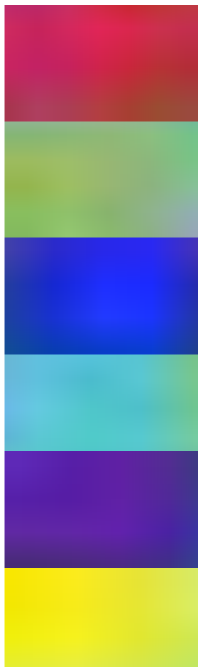
Is a color not selling?

Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:

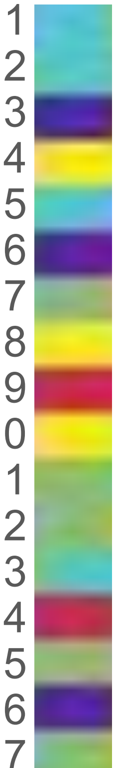
Red
Green
Blue
Teal
Purple
Yellow



first day, 17 sales:

Is a color not selling?

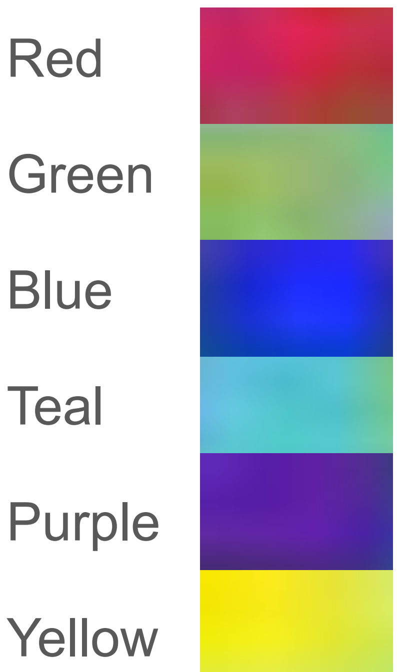
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17



Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:



first day, 17 sales:

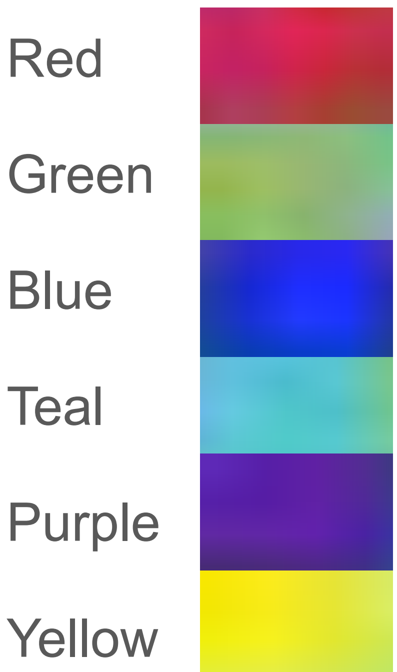
Is a color not selling?

How to define "not selling" so as not to make a mistake?

Bonforroni’s Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:



first day, 17 sales:

Is a color not selling?

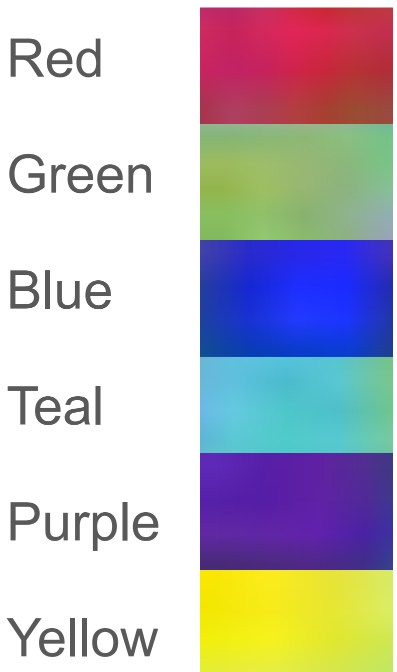
How to define "not selling" so as not to make a mistake?

Counterfactual argument:
Let’s assume the color is as likely to sell as any other. Then what is the probability we observe this many sales?

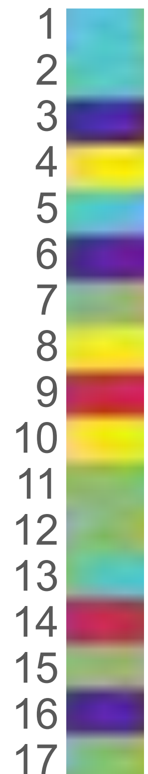
Bonforroni’s Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:



first day, 17 sales:



Is a color not selling?

Let’s assume the color is as likely to sell as any other. Then what is the probability we observe this many sales?

Bonforroni’s Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:

Red

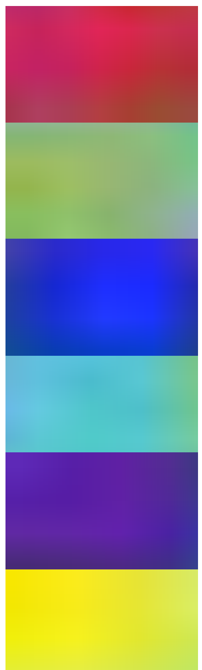
Green

Blue

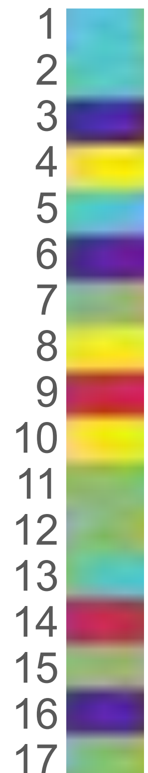
Teal

Purple

Yellow



first day, 17 sales:



Is a color not selling?

Let’s assume the color is as likely to sell as any other. Then what is the probability we observe this many sales?

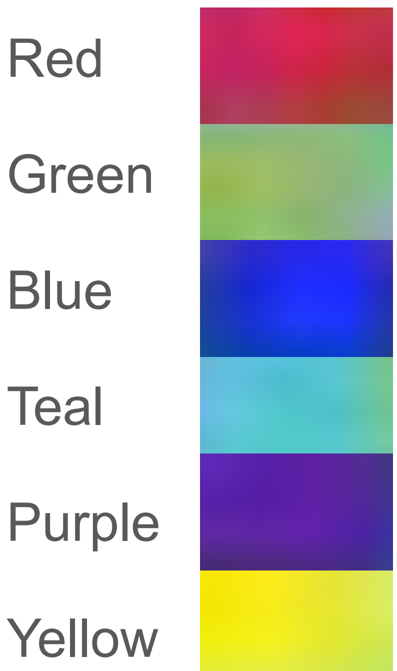
(|blue| == 0) =

??

Bonforroni's Principle; Task Example

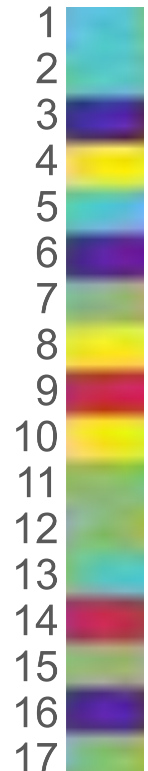
snazzyphones.com wants to know which case to eliminate.

6 total cases:



first day, 17 sales:

Is a color not selling?



Let's assume the color is as likely to sell as any other. Then what is the probability we observe this many sales?

$(|blue| == 0) =$

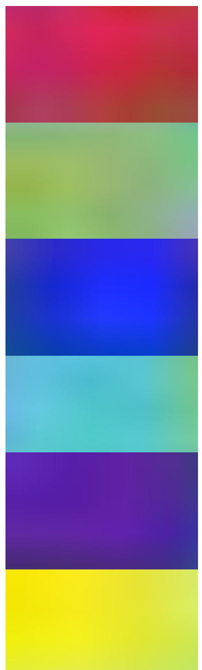
$(5/6)^{\wedge}17 =$ 4.5 % chance

Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

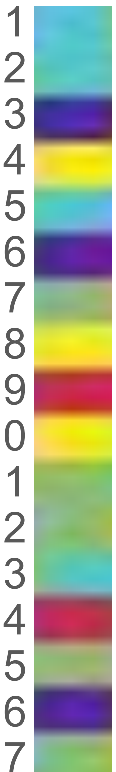
6 total cases:

Red
Green
Blue
Teal
Purple
Yellow



first day, 17 sales:

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17



Is a color not selling?

Let's assume the color is as likely to sell as any other. Then what is the probability we observe this many sales?

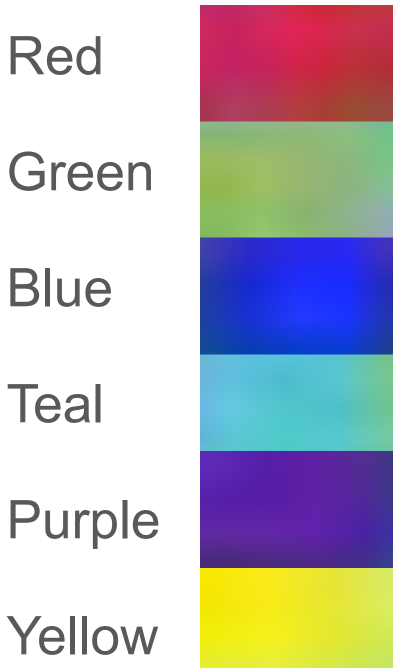
$(|blue| == 0) =$

$(5/6)^{17} \cancel{=}$ 4.5 % chance

Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:



first day, 17 sales:

Is a color not selling?

Let's assume the color is as likely to sell as any other. Then what is the probability we observe this many sales?

$(|blue| == 0) =$

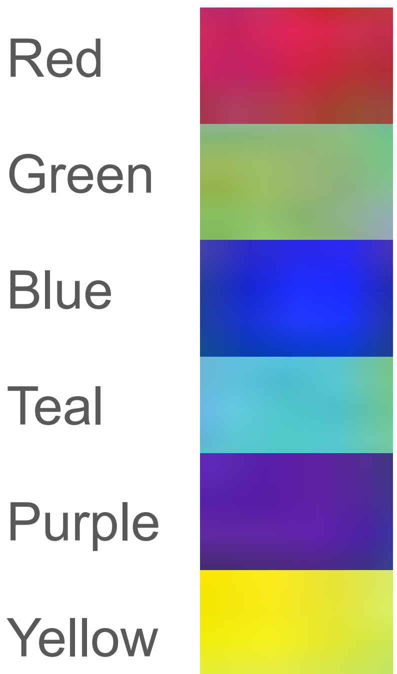
$(5/6)^{17} \cancel{=}$ 4.5 % chance

$p\ (|*| == 0) =\ ??$

Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:



first day, 17 sales:

<u>*Is a color not selling?*</u>

Let's assume the color is as likely to sell as any other. Then what is the probability we observe this many sales?

$(|blue| == 0) =$

$(5/6)^{\wedge}17 \cancel{=}$ 4.5 % chance

$p\,(|*| == 0) =$ **27.0 % chance**

any single color doesn't appear

Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:

first day, 17 sales:

Is a color not selling?

27% is roughly a 1 in 4 chance!
In other words, just due to chance, we would expect 1 out of every 4 times that there are 17 sales that at least one color does not appear at all.

Would you trust eliminating a color is a good data-informed decision to make with these odds?
< 5% or 1 in 20 odds is typical standard for science.

e the color is as likely to
ther. Then what is the
we observe this many

) =

7 ~~X~~ = 4.5 % chance

Yellow

15
16
17

p (1^1 == 0) = 27.0 % chance

any single color doesn't appear

Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

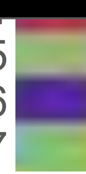
6 total cases: first day, 17 sales: *Is a color not selling?*

27% is roughly a 1 in 4 chance!
In other words, just due to chance, we would expect 1 out of every 4 times that there are 17 sales that at least one color does not appear at all.

Would you trust eliminating a color is a good data-informed decision to make with these odds?
< 5% or 1 in 20 odds is typical standard for science.

Once we look at the data and see a particular pattern, it's easy to think in terms of chance for that specific pattern and forget one started with a broader question.

Yellow 

15
16
17


p (|^| == 0) = 27.0 % chance

any single color doesn't appear

Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:

first day, 17 sales:

Is a color not selling?

27% is roughly a 1 in
In other words, just o
out of every 4 times
least one color does

Would you trust elim
data-informed decisi
< 5% or 1 in 20 odds

This is often mentioned as one of the main reasons for a so-called "replication crisis" in many sciences: In some fields, it has been suggested that over 50% of findings fail to replicate. (https://en.wikipedia.org/wiki/Replication_crisis#Tackling_publication_bias_with_pre-registration_of_studies)

e the
ther
we ob

Once we look at the data and see a particular pattern, it's easy to think in terms of chance for that specific pattern and forget one started with a broader question.

Yellow

15
16
17

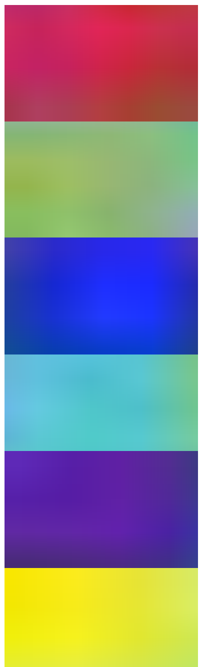
p ((^(== 0) =

any single color doesn't appear

Bonforroni's Principle; Task Example

snazzyphones.com wants to know which case to eliminate.

6 total cases:



first day, 17 sales:

<u>*Is a color not selling?*</u>

Let's assume the color is as likely to sell as any other. Then what is the probability we observe this many sales?

$(|blue| == 0) =$

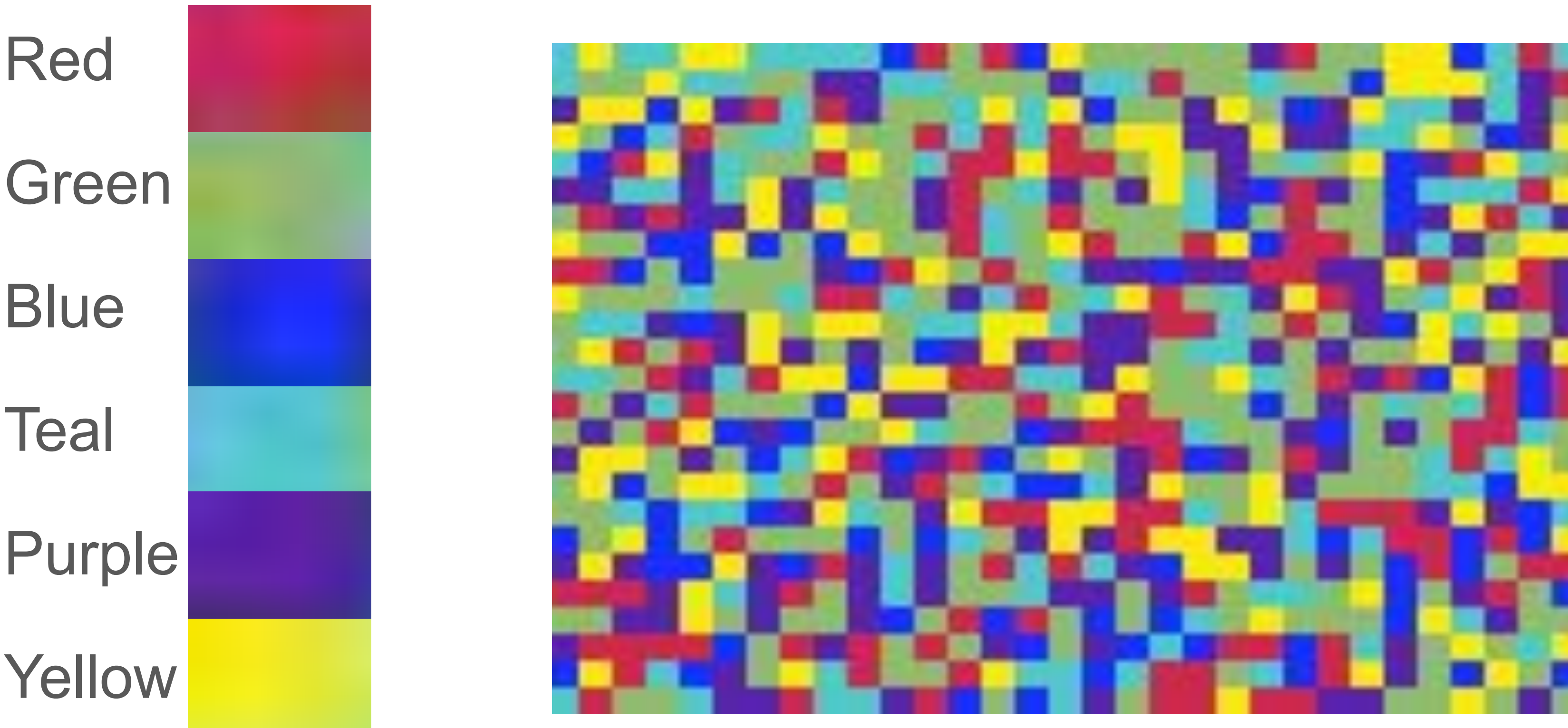
$(5/6)^{17} \cancel{=}$ 4.5 % chance

$p(|*| == 0) =$ 27.0 % chance

any single color doesn't appear

Statistical Limits

Bonferroni's Principle



Bonferroni's Principle

Roughly, calculating the probability of any of n *findings* being true requires n times the probability as testing for 1 finding.

https://xkcd.com/882/

In brief, one can only look for so many patterns (i.e. features) in the data before one finds something just by chance (i.e. finding something that does **not** generalize).

"Data mining" is a bad word in some communities!

Bonferroni's Principle

Note: *Bonferroni's principle* is simply an abstract idea inspired by a precisely defined method of hypothesis testing called "Bonferroni correction".

We will go over this <u>correction method</u> later. The ***<u>principle</u>*** is the more important idea to understand as a big data practitioner.

In brief, one can only look for so many patterns (i.e. features) in the data before one finds something just by chance (i.e. finding something that does **not** generalize).

"Data mining" is a bad word in some communities!

Bonferroni's Principle

The Many Faces of the Bonferroni Principle

Domain	**Concept**	**Mitigation Techniques**
Machine Learning	*Overfitting*	*Regularization;* Out-of-Sample Testing *(Cross-Validation)*
Scientific Process	*P-Hacking*	*Multi-test Correction*
Cognitive Bias	*Confirmation Bias*	Awareness* Turn to Science and Empirical Evidence.
Layman Terms	Falsely believing: "It's not just a coincidence"	Rationality*: Turn to Science / Empirical Evidence.

Preliminaries

Ideas and methods that will repeatedly appear:

- Normalization (TF.IDF)
- Power Laws
- Hash functions
- IO Boundedness (Secondary Storage)
- Unstructured Data
- Probability Theory
- **Bonferroni's Principle**