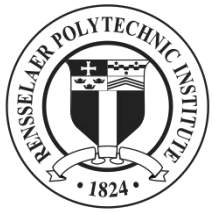
Introduction to Machine Learning Applications

Lecture-1

Jason Kuruzovich

kuruzj@rpi.edu



Rensselaer

Introductions

- Please post a 2-3 sentence introduction **and a picture** in the Webex Teams (or video if you like). Include something interesting about yourself.

https://introml.analyticsdojo.com/intro.html

Bookmark this course website.

Everything you need is there:

- Schedule
- Syllabus
- Lecture PPTs
- Assignments
- Notebooks
- Webex Teams
- Webex Meeting

Class details

Instructor: Jason Kuruzovich

When: Monday & Thursday 4:45 – 6:05 PM

Where: Online

Office Hours: Tuesday 2:00 AM – 4:00 PM

Website: https://introml.analyticsdojo.com/

Announcements: Webex Teams

Teaching Assistant

- Shailesh Divey diveys@rpi.edu
- PhD student

Agenda for today's lecture

- Me
- Data and society
- Why am I excited about Data Science?
- What does it mean to be a data scientist today?
- What will we cover in course?
- Assignment 1

Jason Kuruzovich

- Director of the Severino Center for Technological Entrepreneurship
- Associate Professor of Business Analytics
- Research on marketing, multichannel retailing, most recently entrepreneurship
- Increasingly entrepreneurship, using data to understand startup success

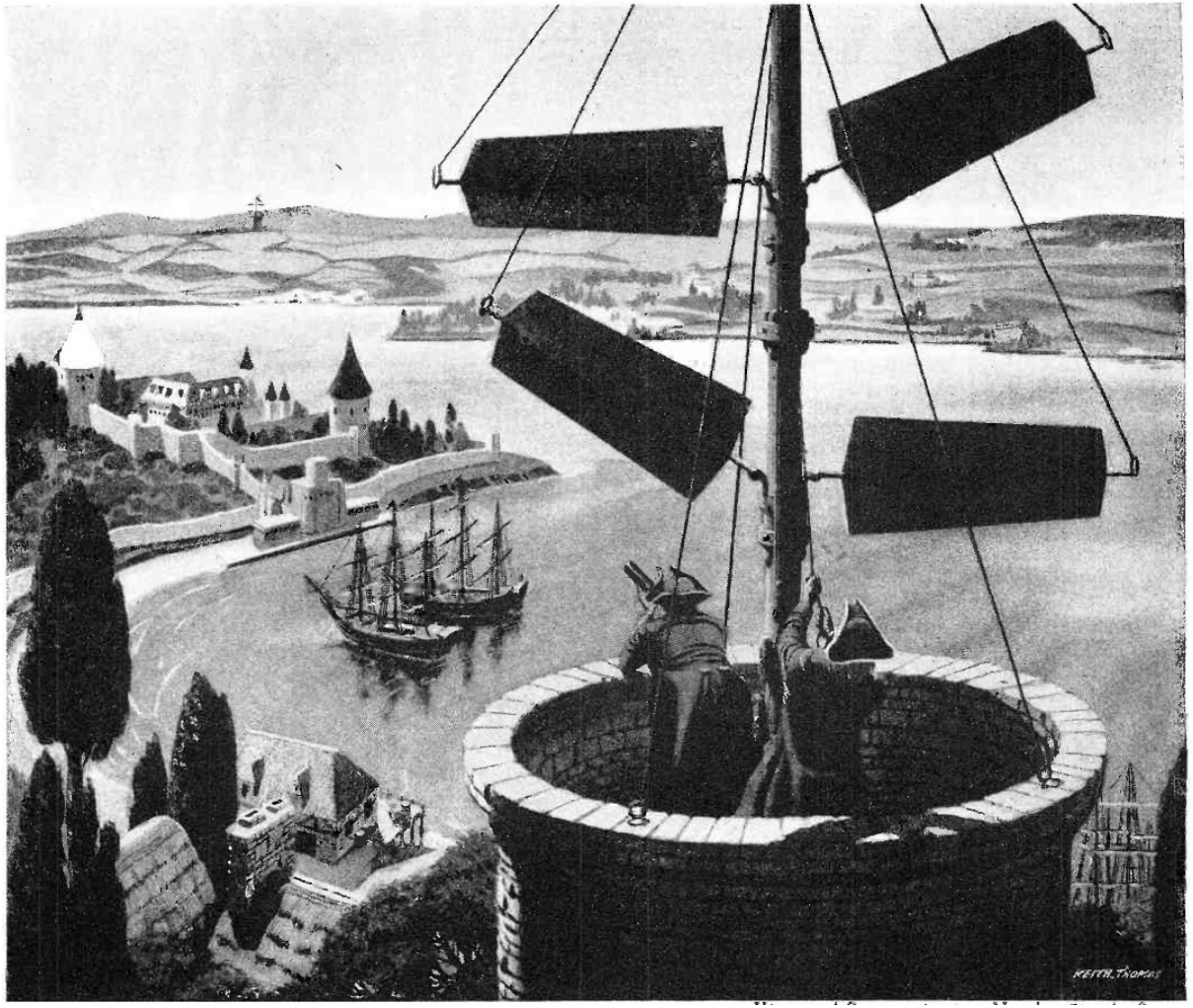
Data and Society

There has been profound progress in technology and data/information processes defining our society

Internet 0.1 Beta (18th Century)

Semaphore Telegraph

- Visual texting by position of the mechanical elements;



By The drawing is signed "Keith Thomas" in lower right corner [Public domain], via Wikimedia Commons

Internet 0.1 Beta (18th Century)

- Chappe system, as used simply for signaling letters and numbers.

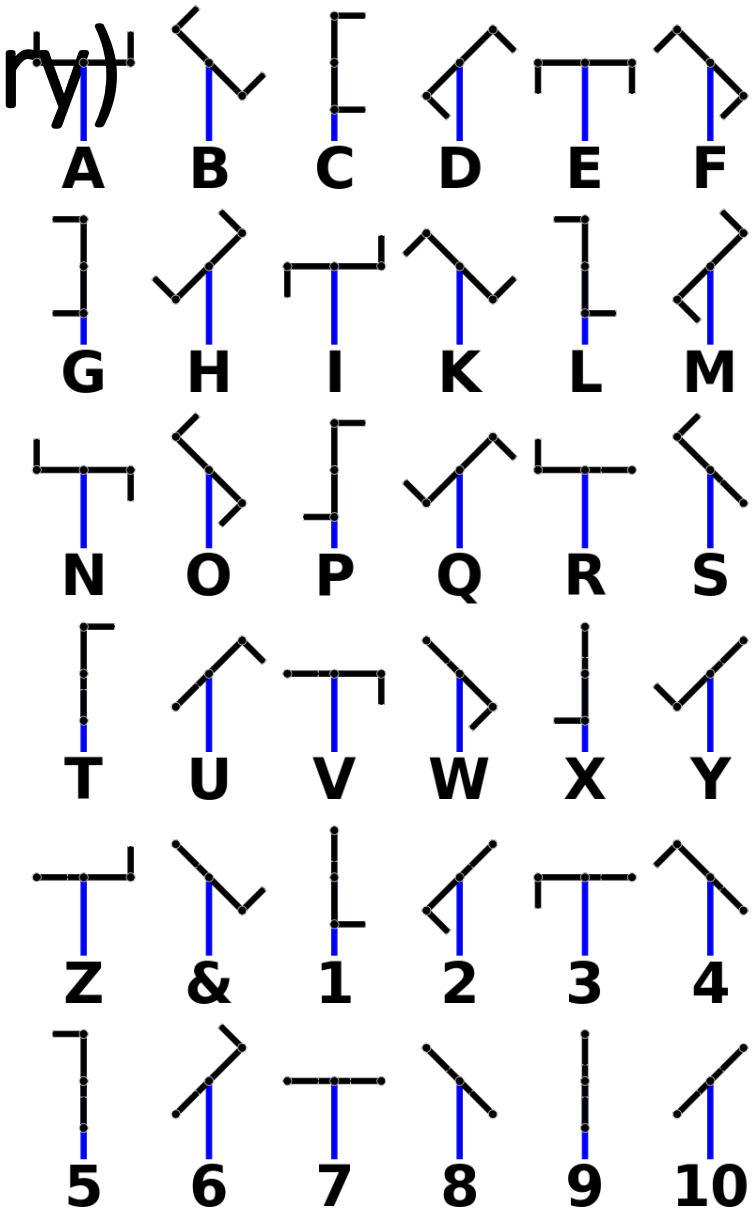
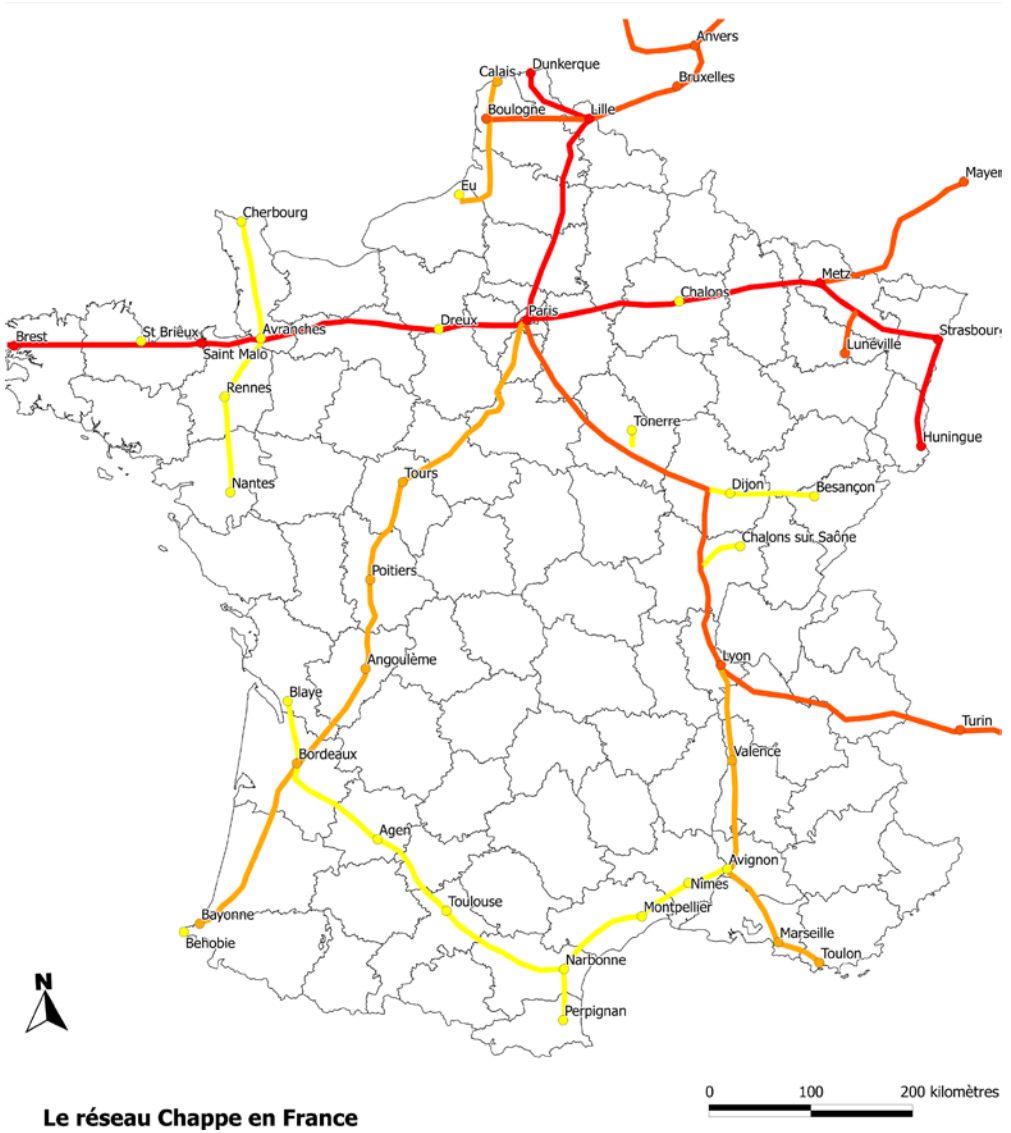


Image source: https://en.wikipedia.org/wiki/Semaphore_telegraph#/media/File:Chappe.svg

Internet 0.1 Beta (18th Century)

- Over 50 stations connecting France
- Shows the extent to which people will go to communicate

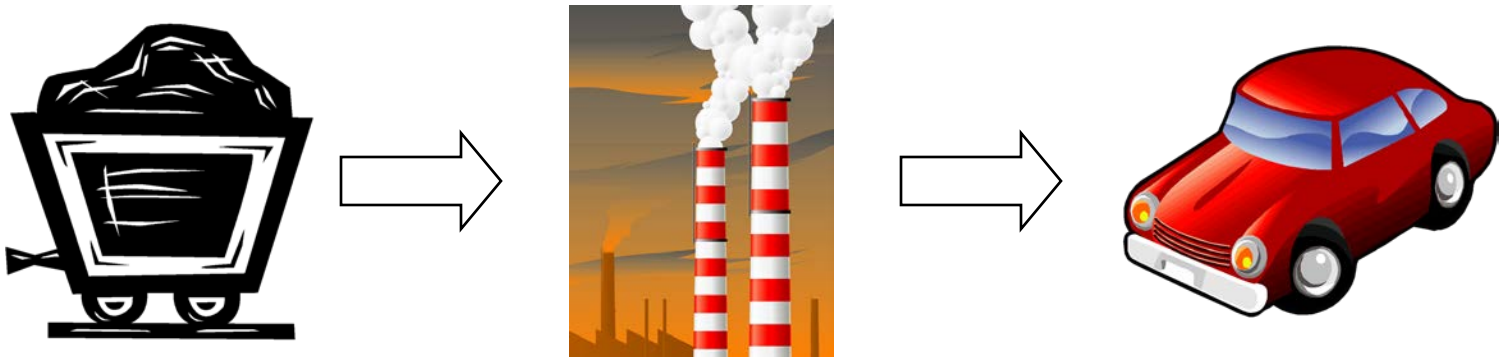


By Jeunamateur [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

“We create as much information in two days now as we did from the dawn of man through 2003.”

-Eric Schmidt, Former CEO of Google

Information Economy



INFORMATION-BASED BUSINESS PROCESS

INFORMATION TECHNOLOGY

Why am I excited about Data Science?

Data, Analytics, and AI are Changing the World

Martin Grandjean - Own work : http://www.martingrandjean.ch/wp-content/uploads/2013/10/Graphe3.png
•CC BY-SA 3.0

“Analytics is the discovery and communication of meaningful patterns in data.”

-Wikipedia

More data. More analytics.

The Internet, the Original Big Data Problem

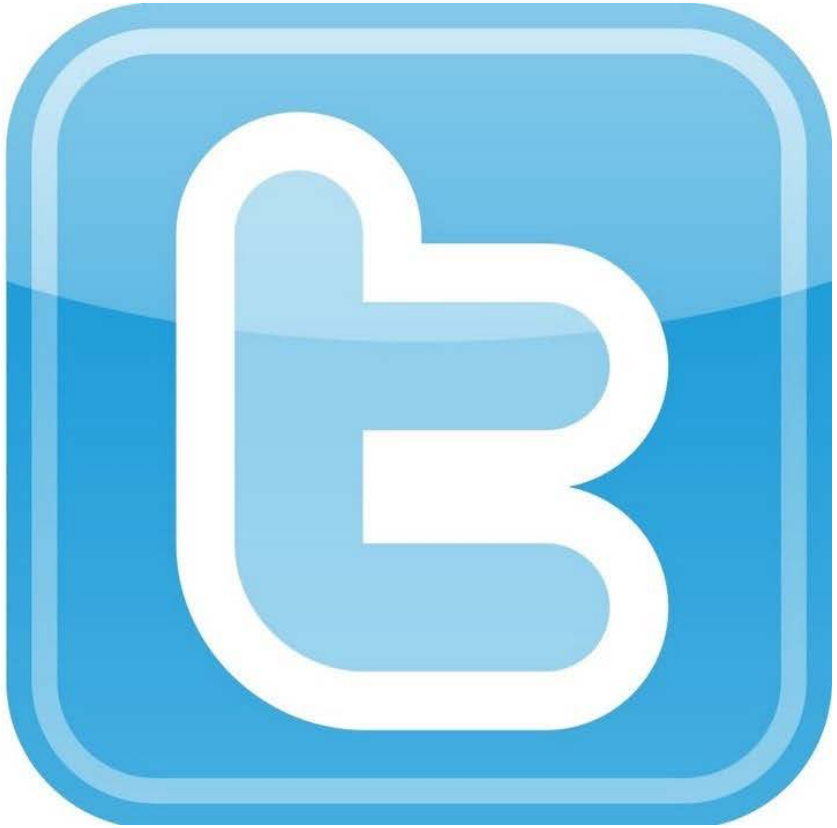
> “PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.”
>
> - From "Facts about Google and Competition” via Wikipedia [https://en.wikipedia.org/wiki/PageRank].

Internet of Things

“The internet of things (IoT) is the network of physical devices, vehicles, buildings and other items—embedded with electronics, software, sensors, actuators, and network connectivity that enable these objects to collect and exchange data.”

– Internet of Things Global Standards Initiative via Wikipedia.

Web 2.0 Social Networks



Disney at Open

ROLE OF DATA: How many tickets did we sell?

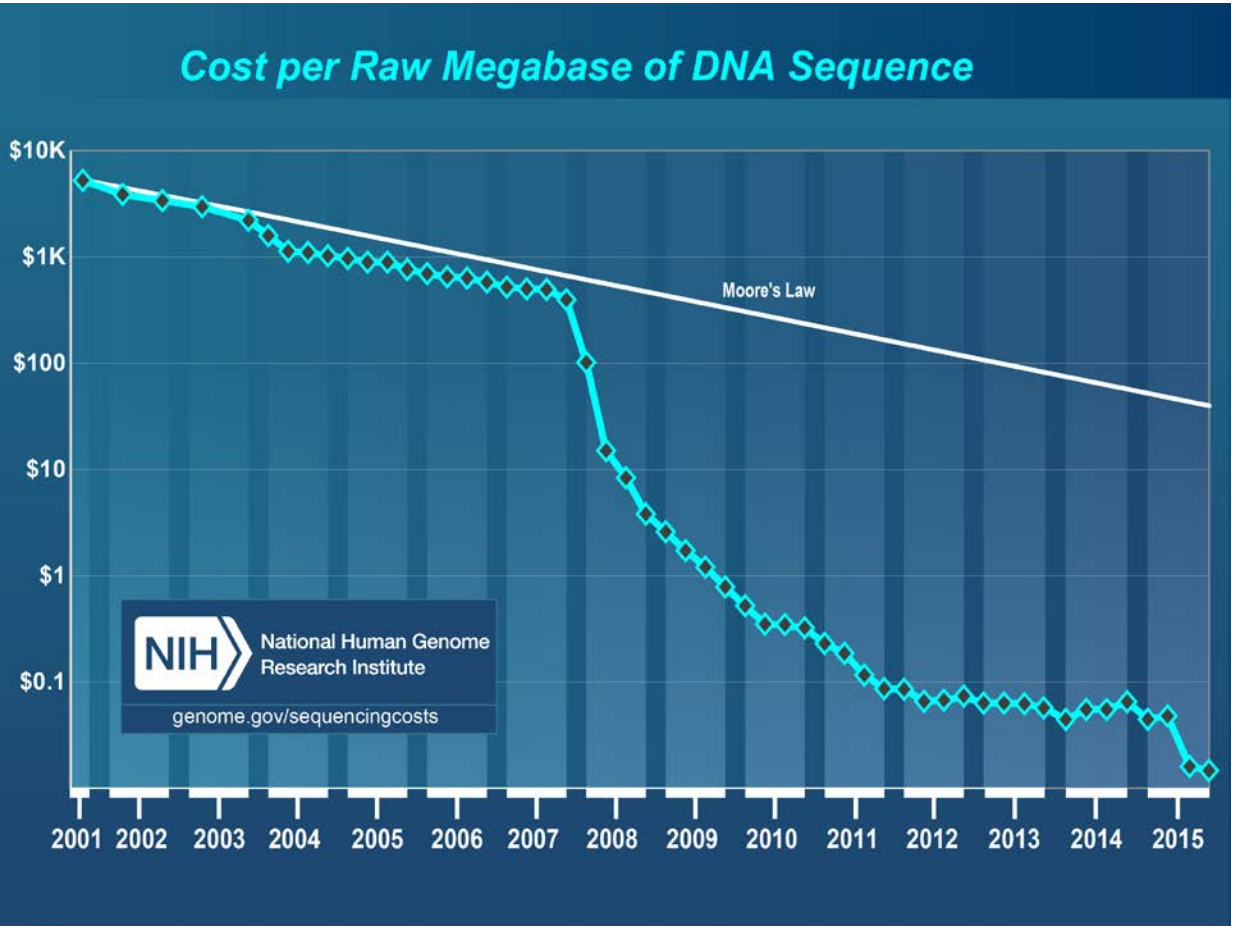
Disney – Data Warehouse Stage

ROLE OF DATA: How much did our customers spend? How can we understand different customer types?

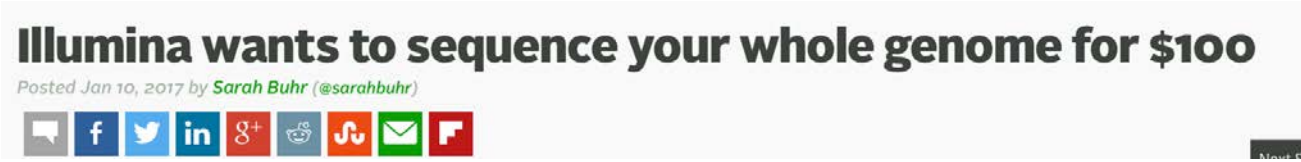
Disney – Big Data

ROLE OF DATA: What path did customers take through the park, when did they leave? How long did they stand in line? When did they spend money on souvenirs and where? How often did they go to the bathroom and did they have to wait? How long did they spend at dinner in the Mexican pavilion compared with the German pavilion? How does the speed of entry correlate with tipping behavior?

Big Data eScience



Tremendous drop in cost of sequencing DNA



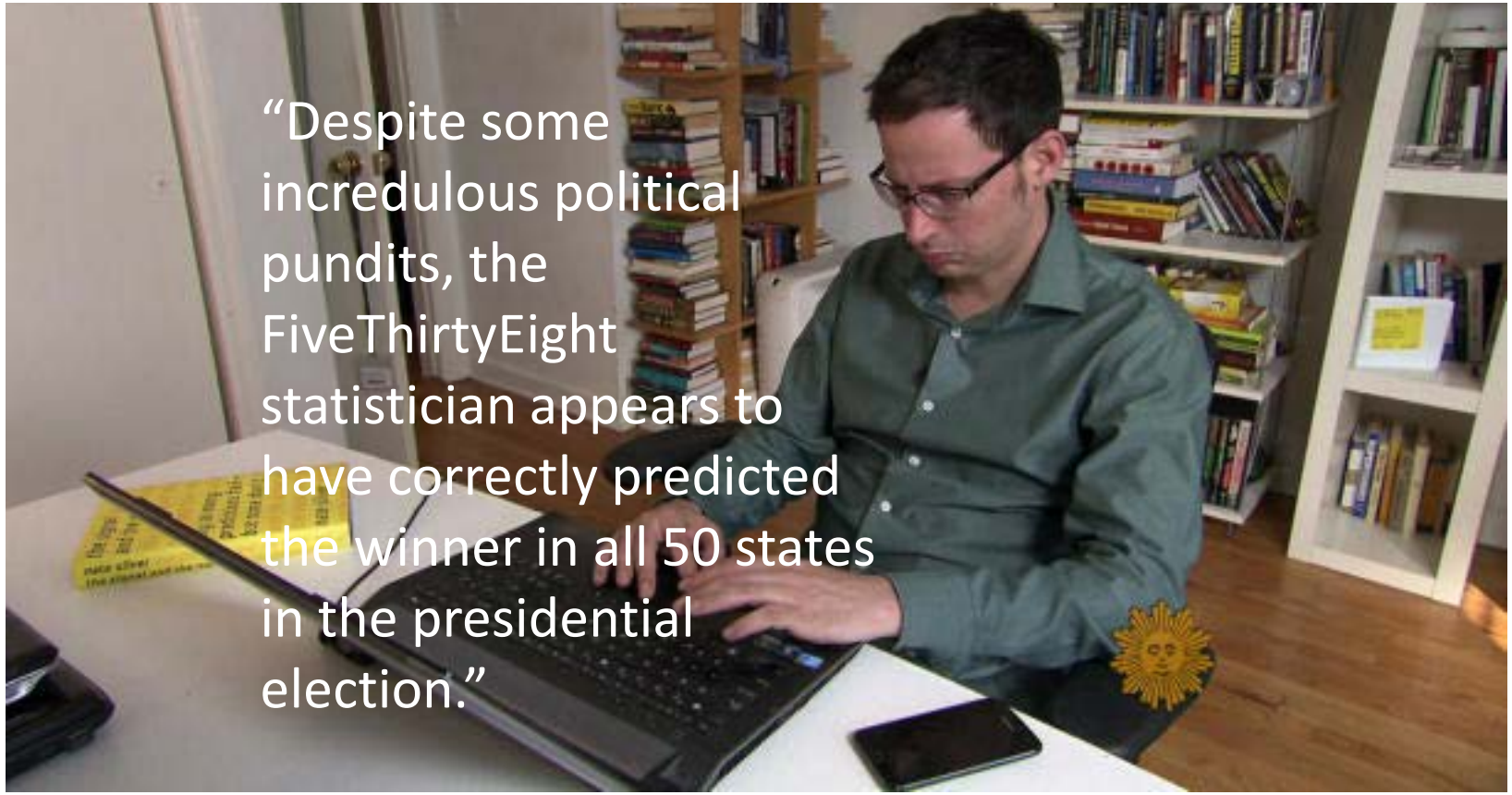
https://www.genome.gov/images/content/costpermb2015_4.jpg

Big Data and Astronomy

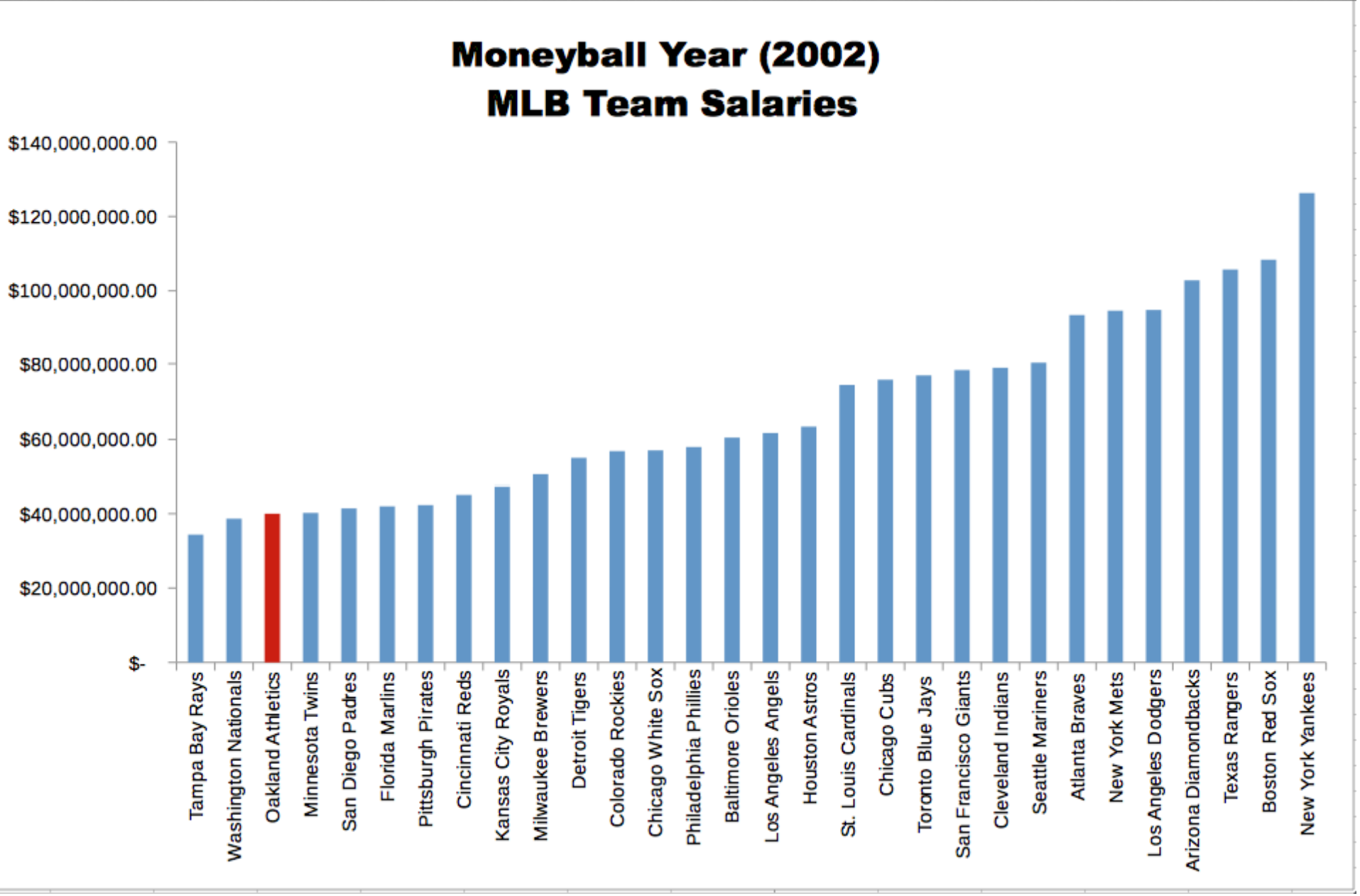
“To store the Big Data the MWA produces, you’d need almost three 1 TB hard drives every two hours.”

Source: http://www.skatelescope.org/news/pawsey-centre/

Obama's win a big vindication for Nate Silver, king of the quants



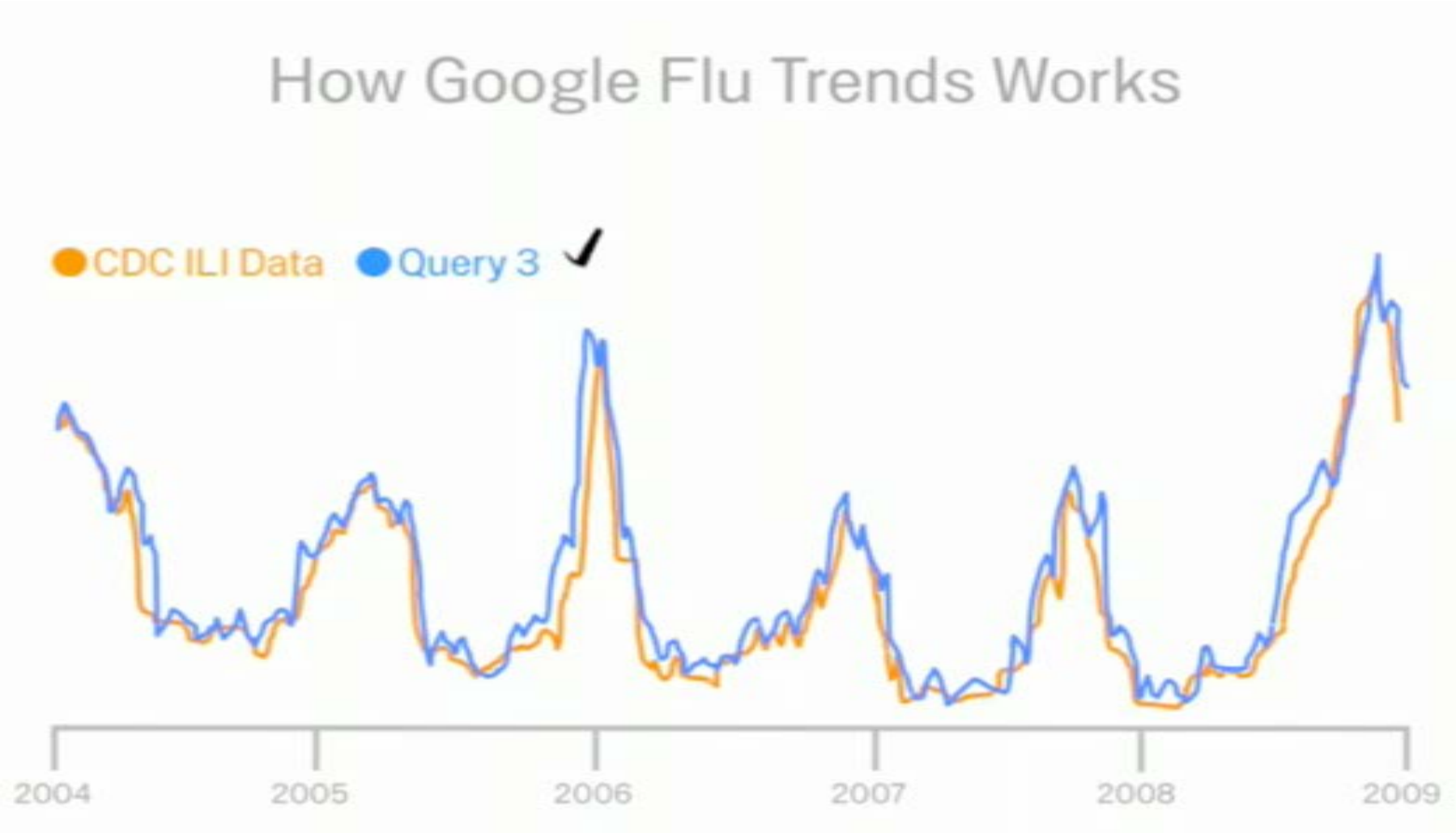
Source: http://news.cnet.com/8301-13510_3-57546161-21/obamas-win-a-big-vindication-for-nate-silver-king-of-the-quants/



http://www.youtube.com/watch?v=AiAHlZVgXjk

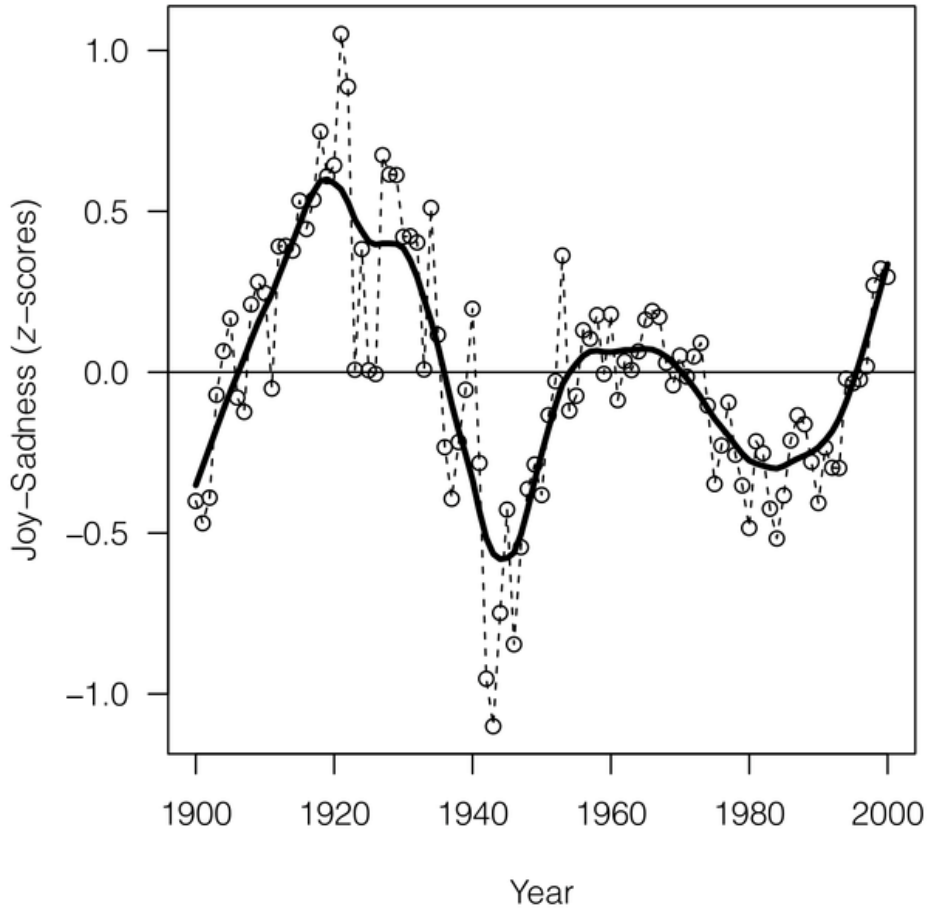
Darryl Leewood (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Google Flu Trends



http://www.google.org/flutrends/about/how.html

The Expression of Emotions in 20th Century Books



"using the data set provided by Google that includes word frequencies in roughly 4% of all books published up to the year 2008. We find evidence for distinct historical periods of positive and negative moods"

Source:
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0059030

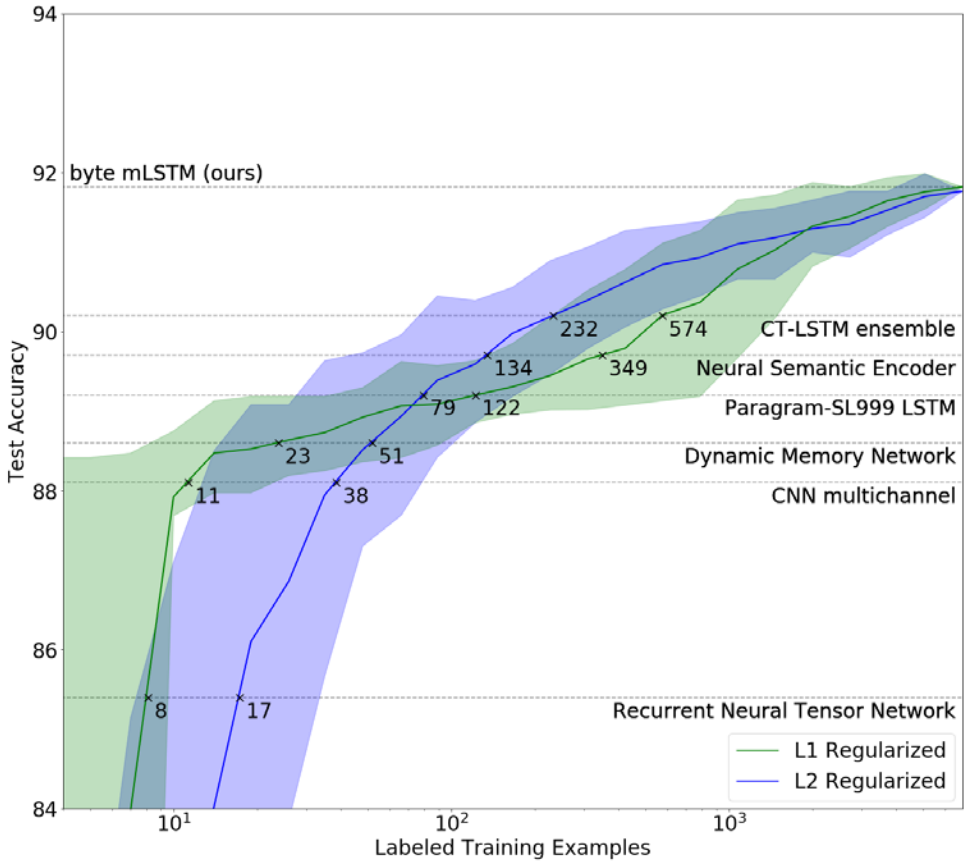

“AlphaGo is almost like the God of Go”
AlphaGo’s move 37 described as "So beautiful.”


Elon Musk
@elonmusk
Following
OpenAI first ever to defeat world's best players in competitive eSports. Vastly more complex than traditional board games like chess & Go.
1:15 AM - 12 Aug 2017
11,014 Retweets 37,113 Likes
1.1K
11K
37K
© Twitter / @elonmusk

OpenAI

“In 2012, Google had two deep-learning projects underway. Today, Google has more than 1,000 deep-learning projects in every major product category, including search, maps, translation, and self-driving cars.”

Unsupervised Sentiment Neuron Beat Supervised Sentiment Analysis



OpenAI developed an unsupervised system which learns an excellent representation of sentiment, despite being trained only to predict the next character in the text of Amazon reviews.

https://blog.openai.com/unsupervised-sentiment-neuron/

Startups are also Bringing AI to Market

gpt2

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

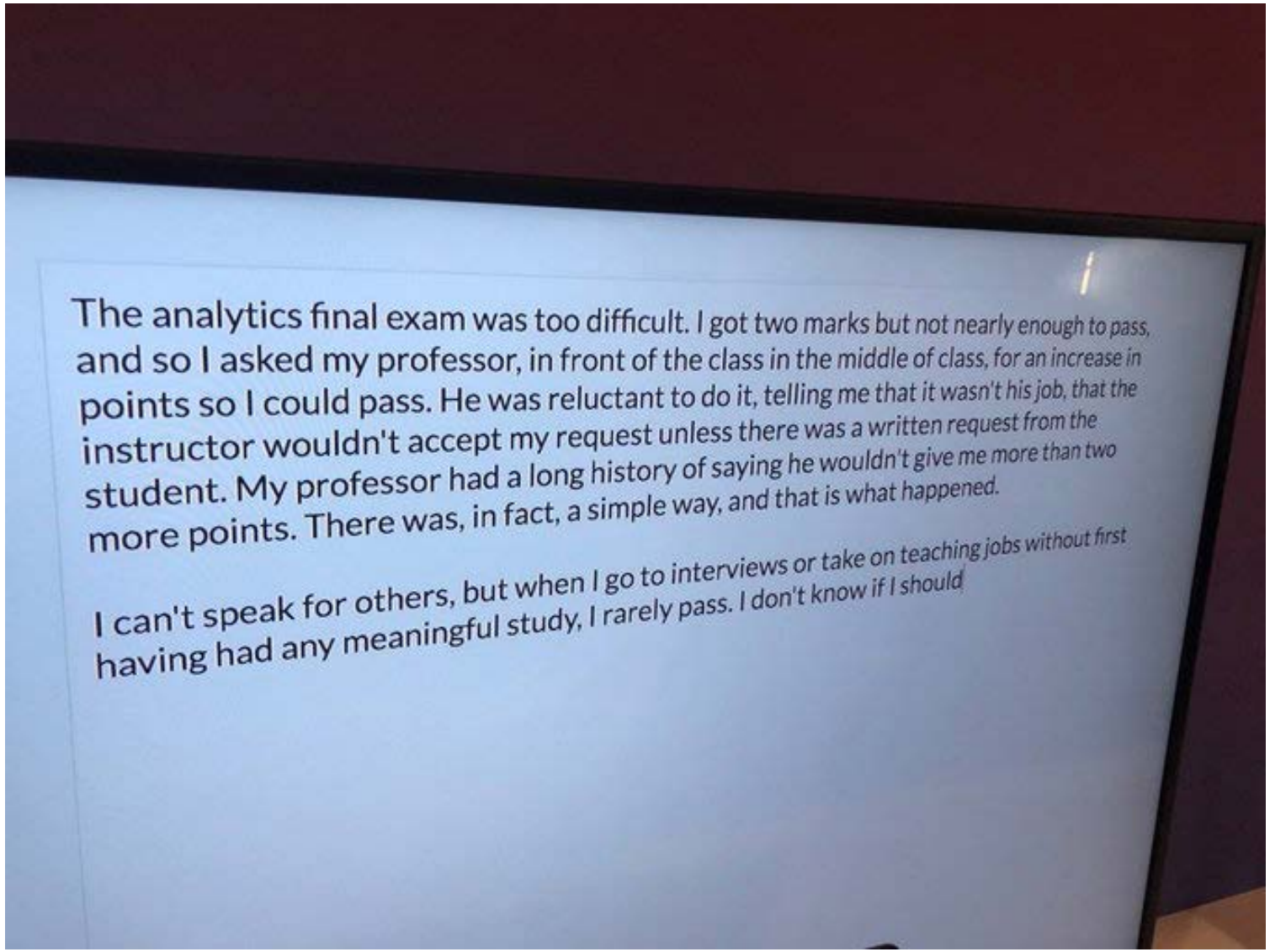
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.


The analytics final exam was too difficult. I got two marks but not nearly enough to pass, and so I asked my professor, in front of the class in the middle of class, for an increase in points so I could pass. He was reluctant to do it, telling me that it wasn't his job, that the instructor wouldn't accept my request unless there was a written request from the student. My professor had a long history of saying he wouldn't give me more than two more points. There was, in fact, a simple way, and that is what happened.
I can't speak for others, but when I go to interviews or take on teaching jobs without first having had any meaningful study, I rarely pass. I don't know if I should

What does it mean to be a data scientist today?

Let’s Look at Yesterday First

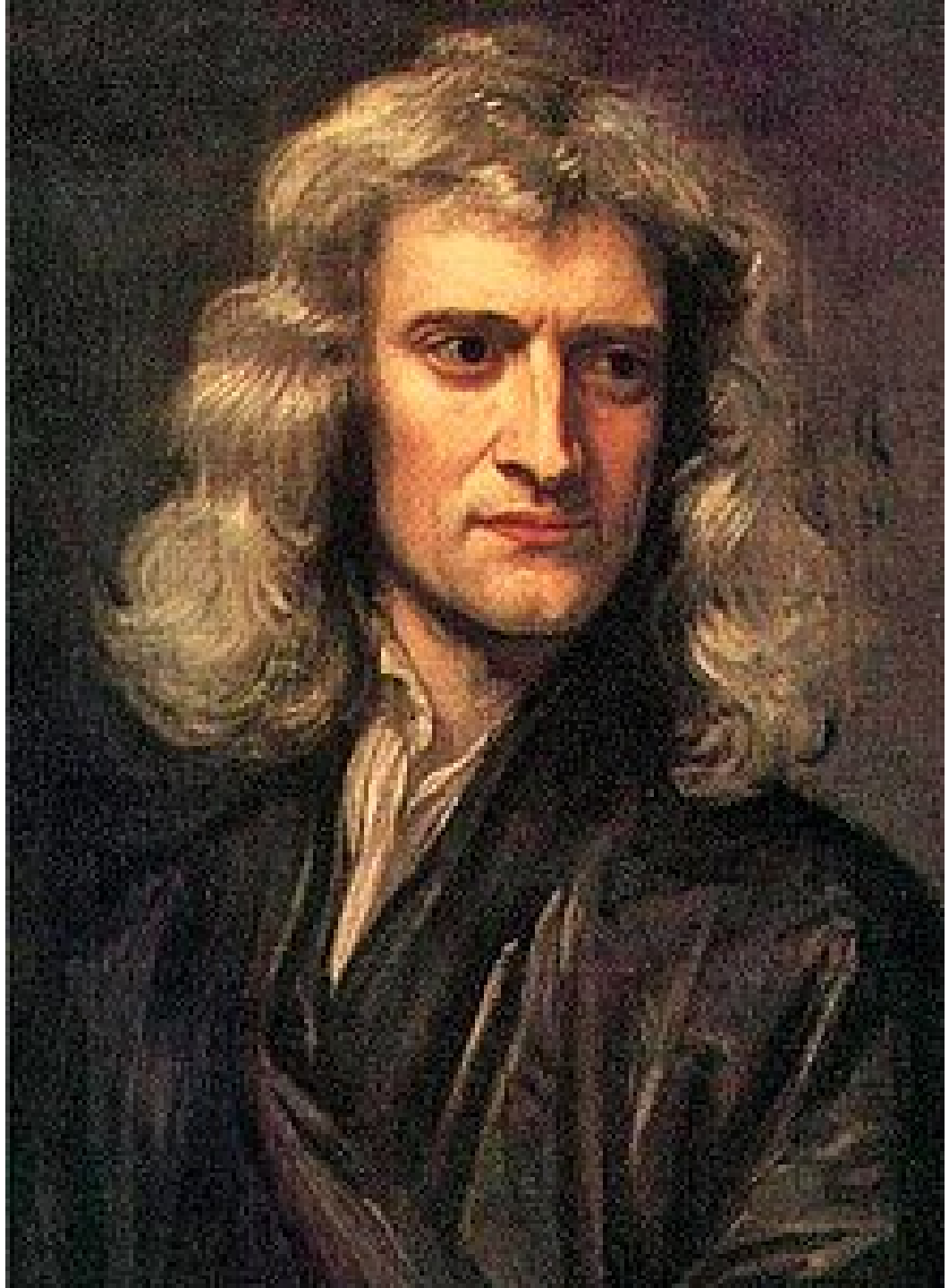
Tycho Brahe, 1546-1601
Observational Astronomer, 1572-1601

Data

Johannes Kepler, 1571-1630
Astronomer and Mathematician,
Laws of planetary motion,
1609-1619
Used Brahe's planetary observation data

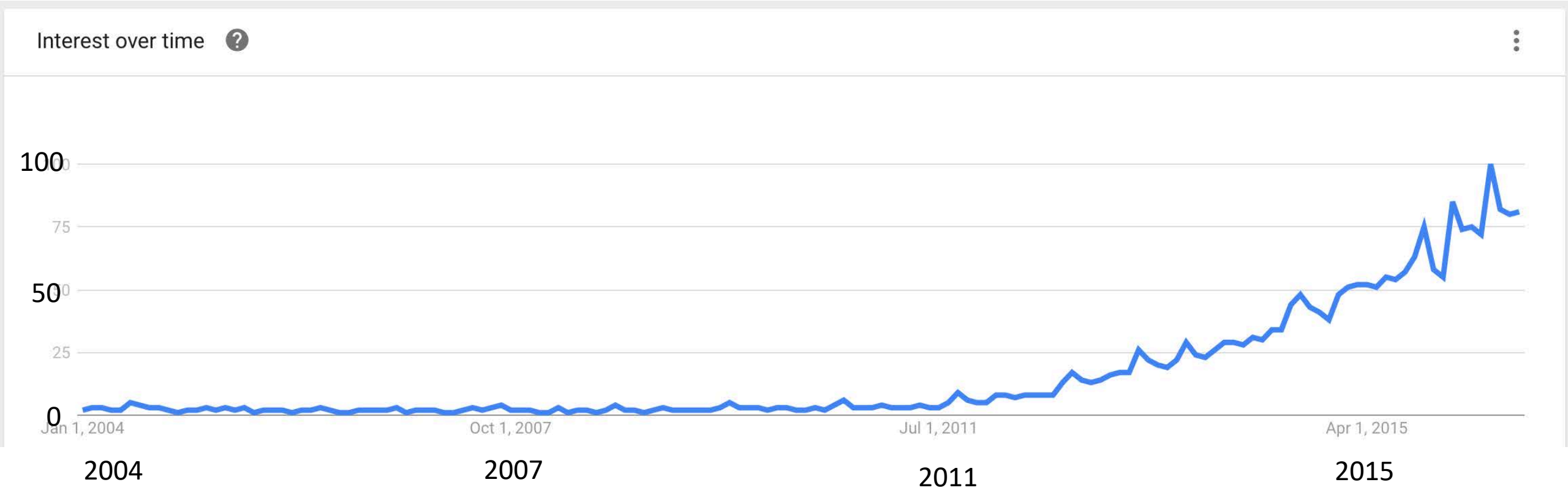
Correlation

Isaac Newton, 1643-1727
Mathematician,
Law of Gravity, 1687
Built on Kepler’s Laws



Causation

What is a “Data Scientist”?



By Special Collections, University of Houston Libraries [CC0], via Wikimedia Commons

The data scientist has been described as the sexiest job of the 21st century, and people with the broad range of skills to truly be a data scientist have been called unicorns.

“There’s a joke running around on Twitter that the definition of a data scientist is ‘a data analyst who lives in California.’”

— Malcolm Chisholm

http://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/

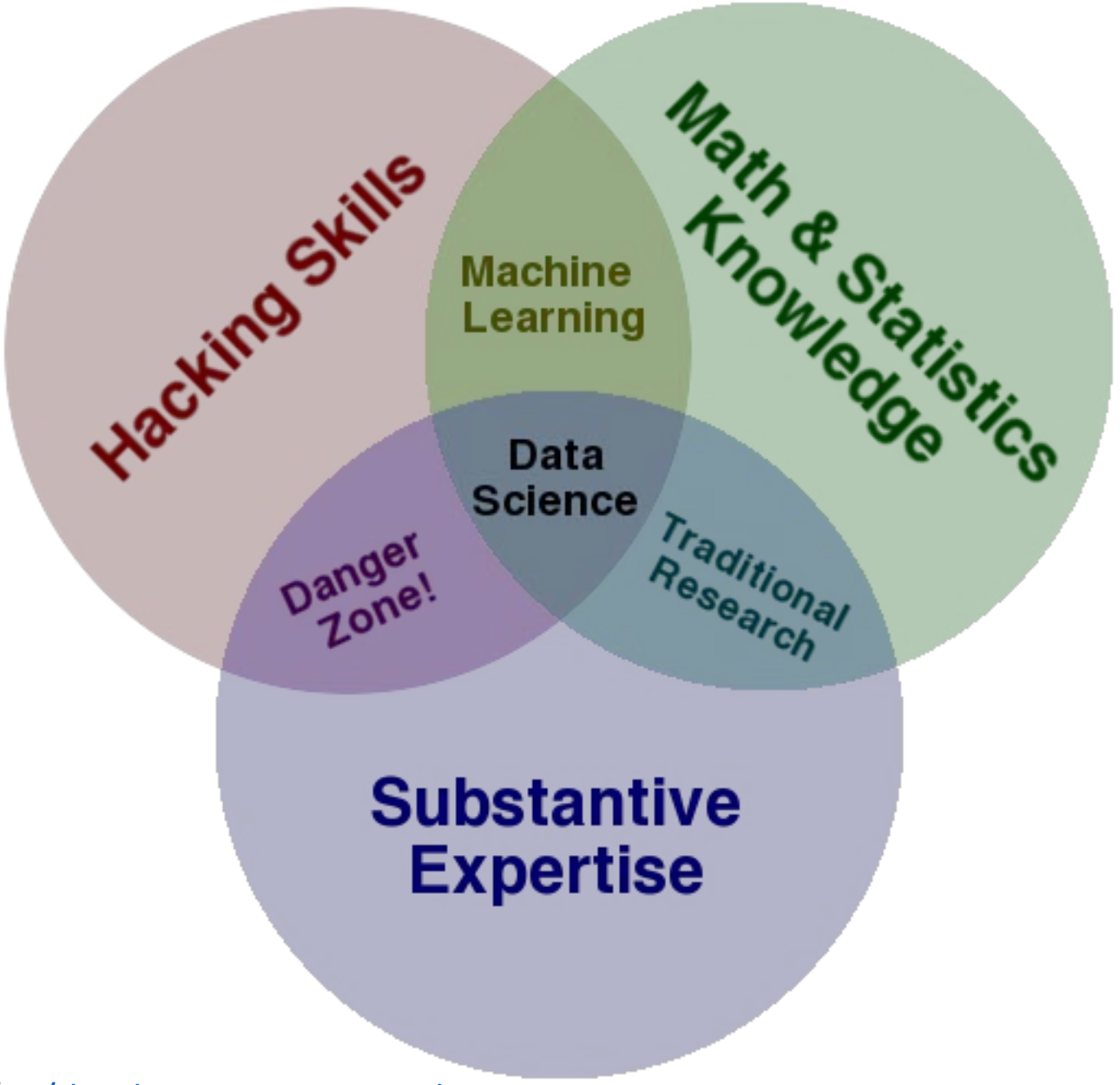
Data scientists are “analytically-minded, statistically and mathematically sophisticated data engineers who can infer insights into business and other complex systems out of large quantities of data.”

— Steve Hillion

http://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/

Data Science

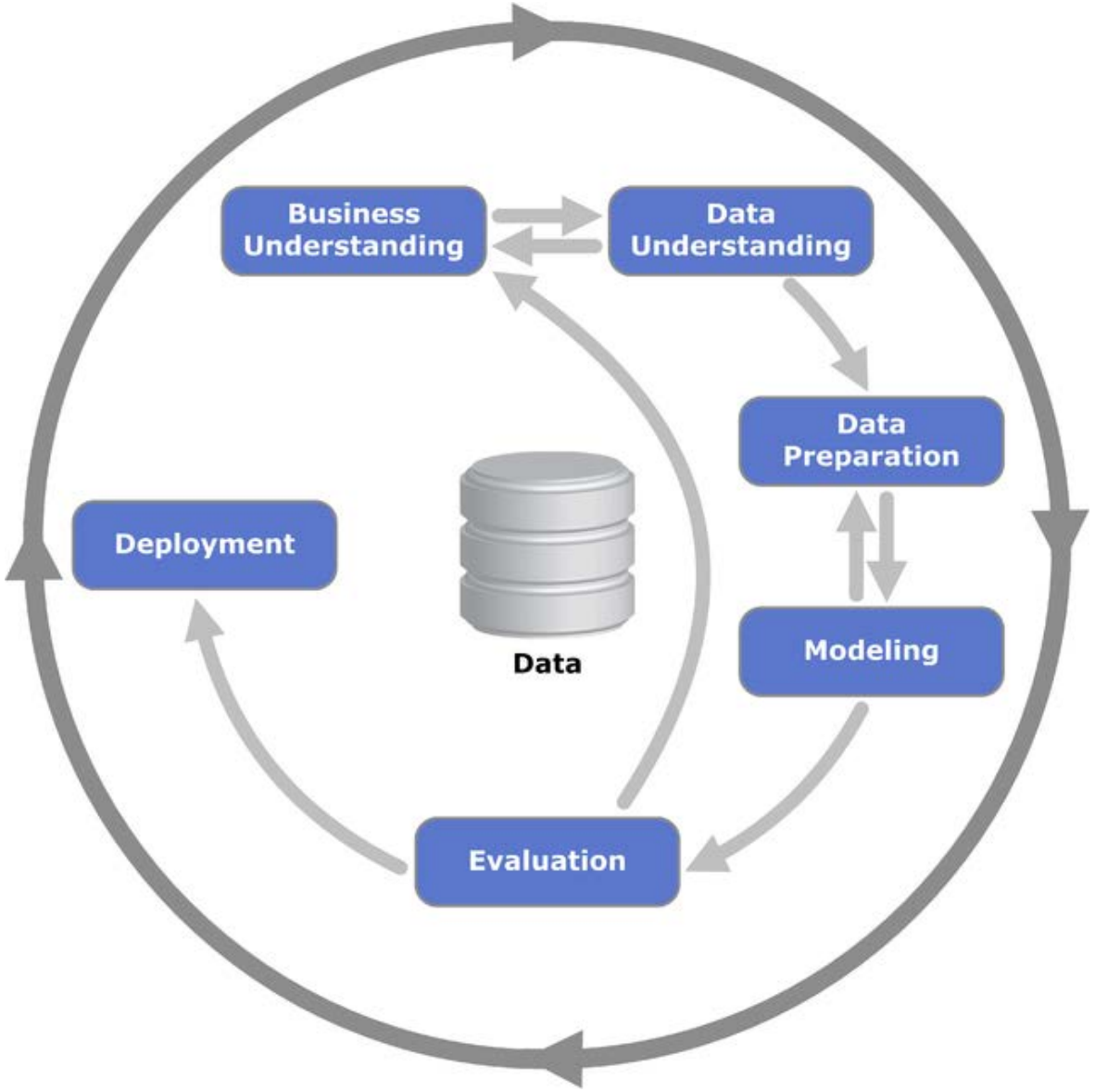
- Hacking Skills
- Math & Statistics Knowledge
- Substantive Expertise
- Ability to Learn and



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

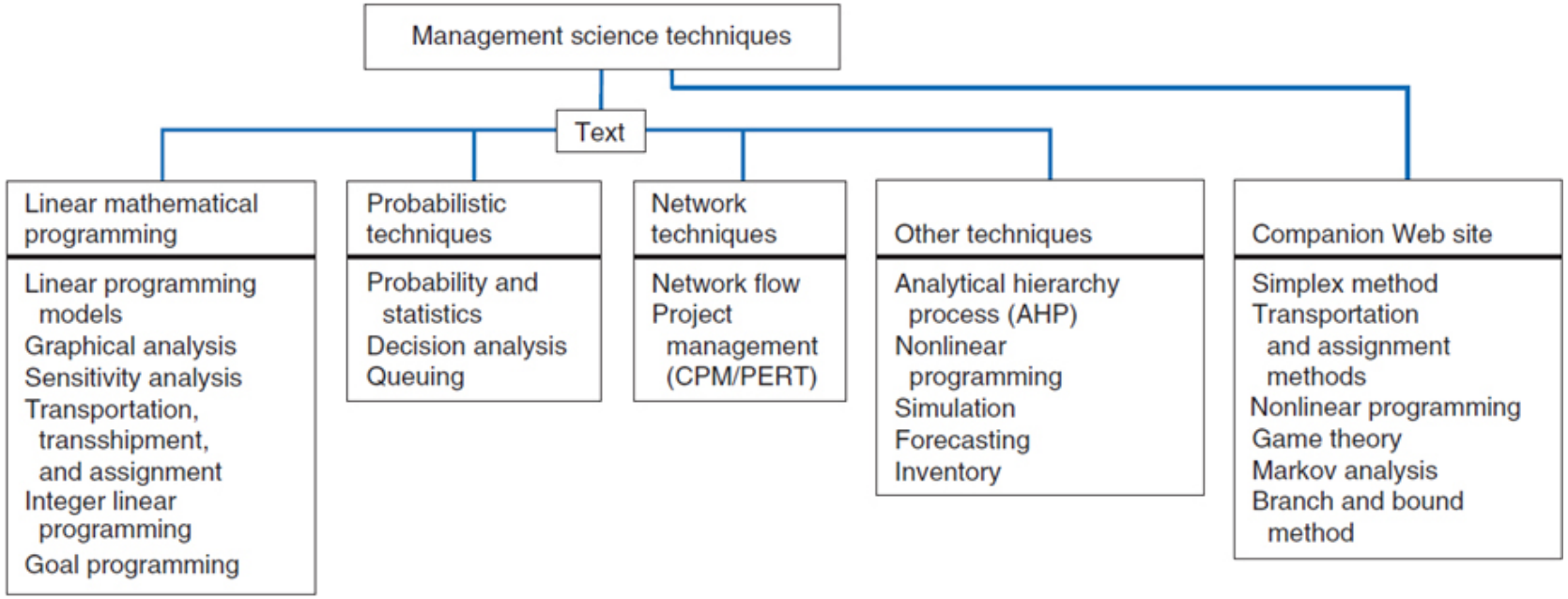
CRISP-DM

"Cross Industry Standard Process for Data Mining, commonly known by its acronym CRISP-DM, was a data mining process model that describes commonly used approaches that data mining experts use to tackle problems."
-Wikipedia



By Kenneth Jensen (Own work) [CC BY-SA 3.0], via Wikimedia Commons

Figure 1.6 Classification of Management Science Techniques



Quantitative Management
qm.analyticsdojo.com

Introduction to Machine Learning Applications
introml.analyticsdojo.com

Key Differences in Quant Business Issues

DESCRIPTIVE

- What is the average revenue per subscriber? What is the subscriber growth over the past quarter -> Dashboards

INFERENCE

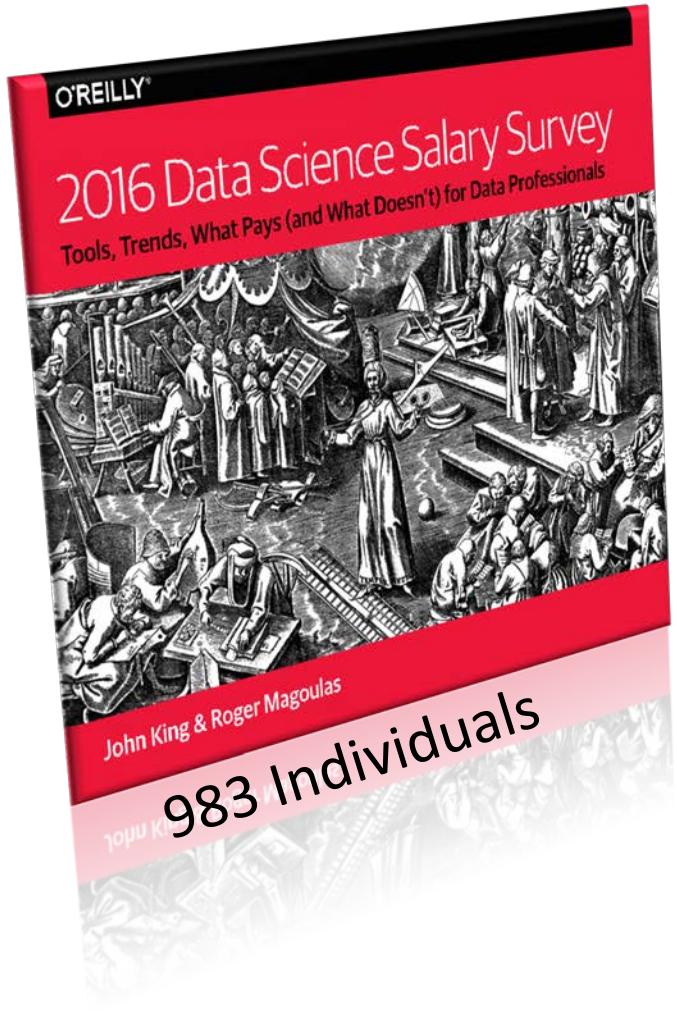
- Inference: Does X cause Y → Statistics/Economics (probabilistic techniques)

PREDICTION

- Can we predict Y given X (and a training dataset of the historical relationship) → Machine Learning (probabilistic techniques)

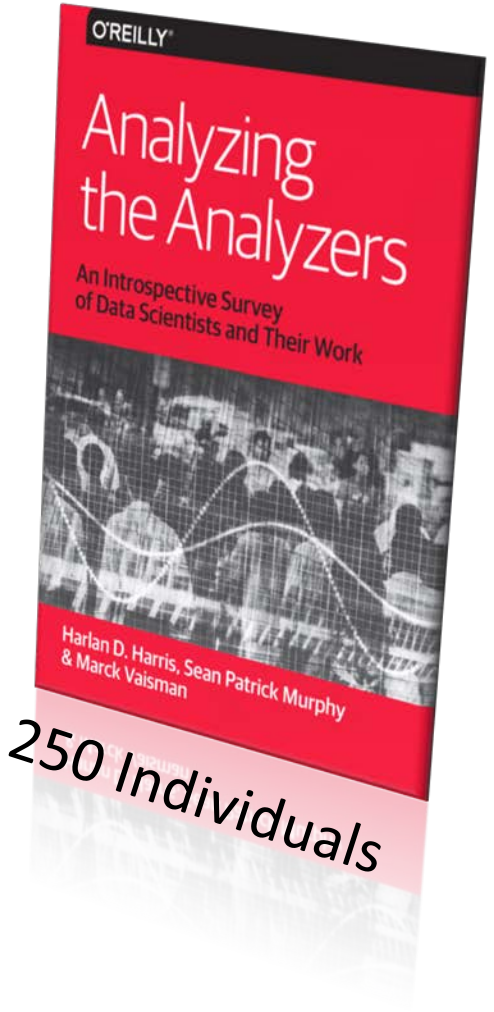
DETERMINISTIC SOLUTION

- Can we solve for X to optimize Y (and the constraints that X>5) → Linear Programming



O'REILLY Surveys

- Asked people involved in data science events to complete an online survey



Credit: DS100.org Class1 Presentation

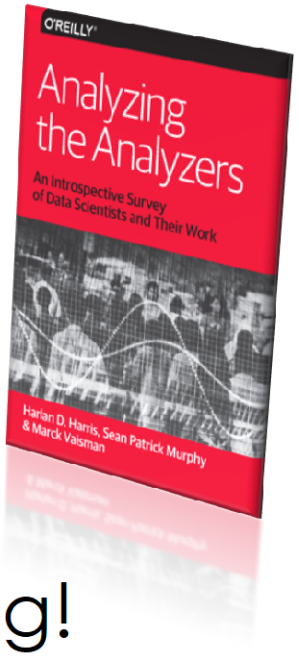
There is a lot of excitement around

Big Data

… how big is the data?

Credit: DS100.org Class1 Presentation

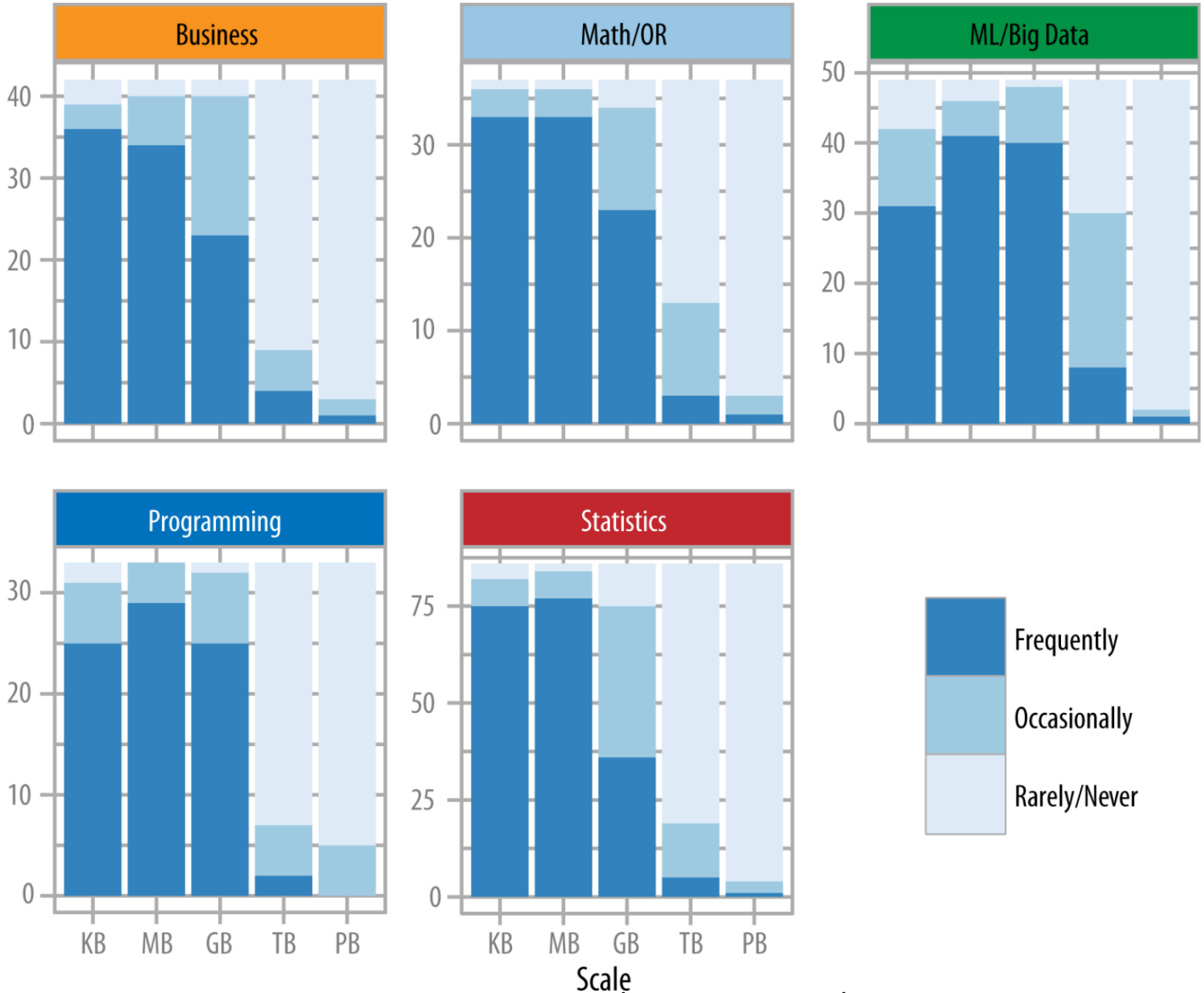
Scale of Data



Not usually big!

< 1GB

However some data scientists frequently process TB to PB data sets.



Credit: DS100.org Class1 Presentation

What do they do?

How involved are you in task ___:
(a) Major, (b) Minor, (c) None

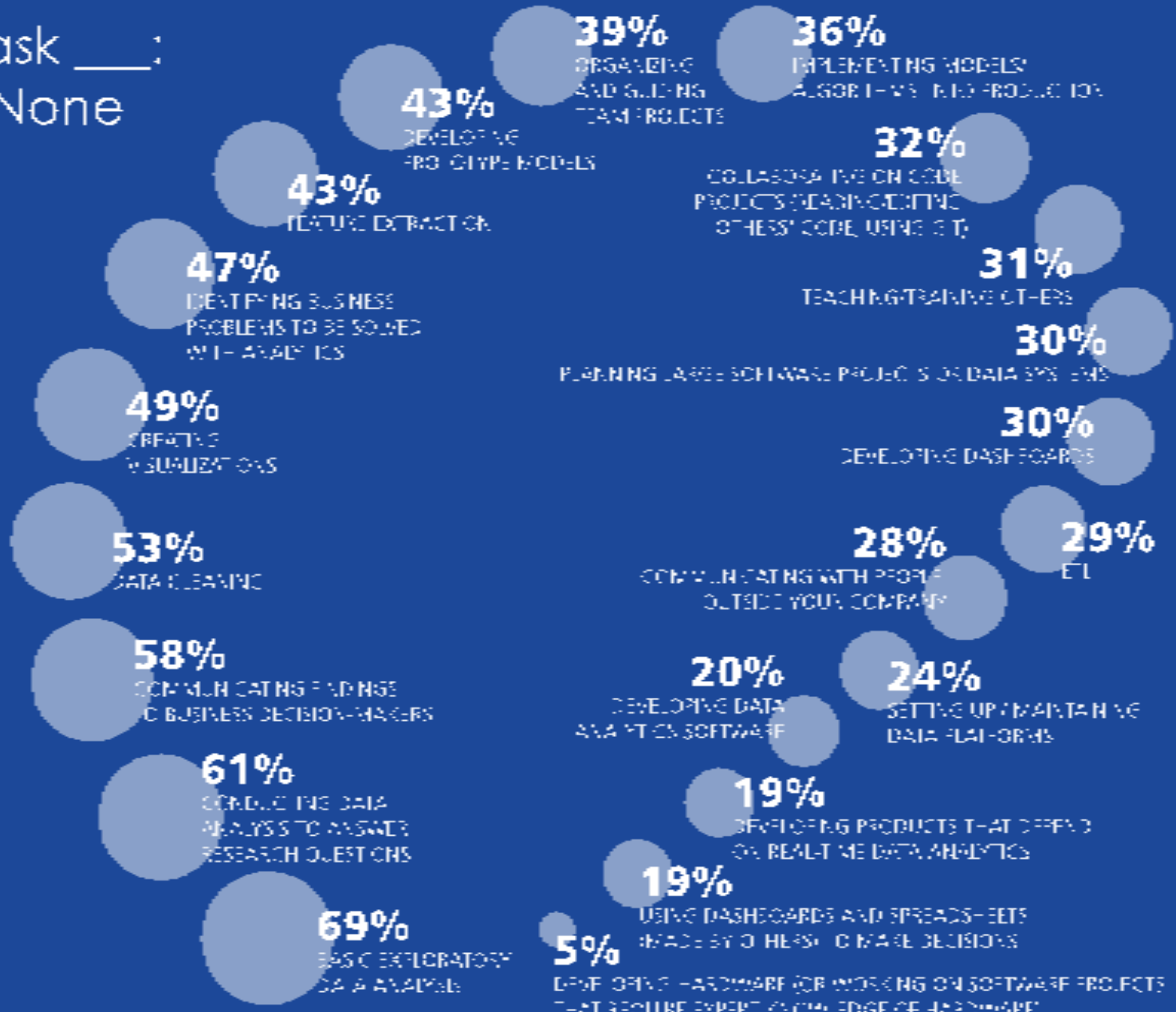
Developing Models
Implementing ML Algorithms
Visualization

Exploratory Data Analysis (EDA)
Researching Questions
Writing Reports,

…

Credit: DS100.org Class1 Presentation

How involved are you in task ___:
(a) Major, (b) Minor, (c) None



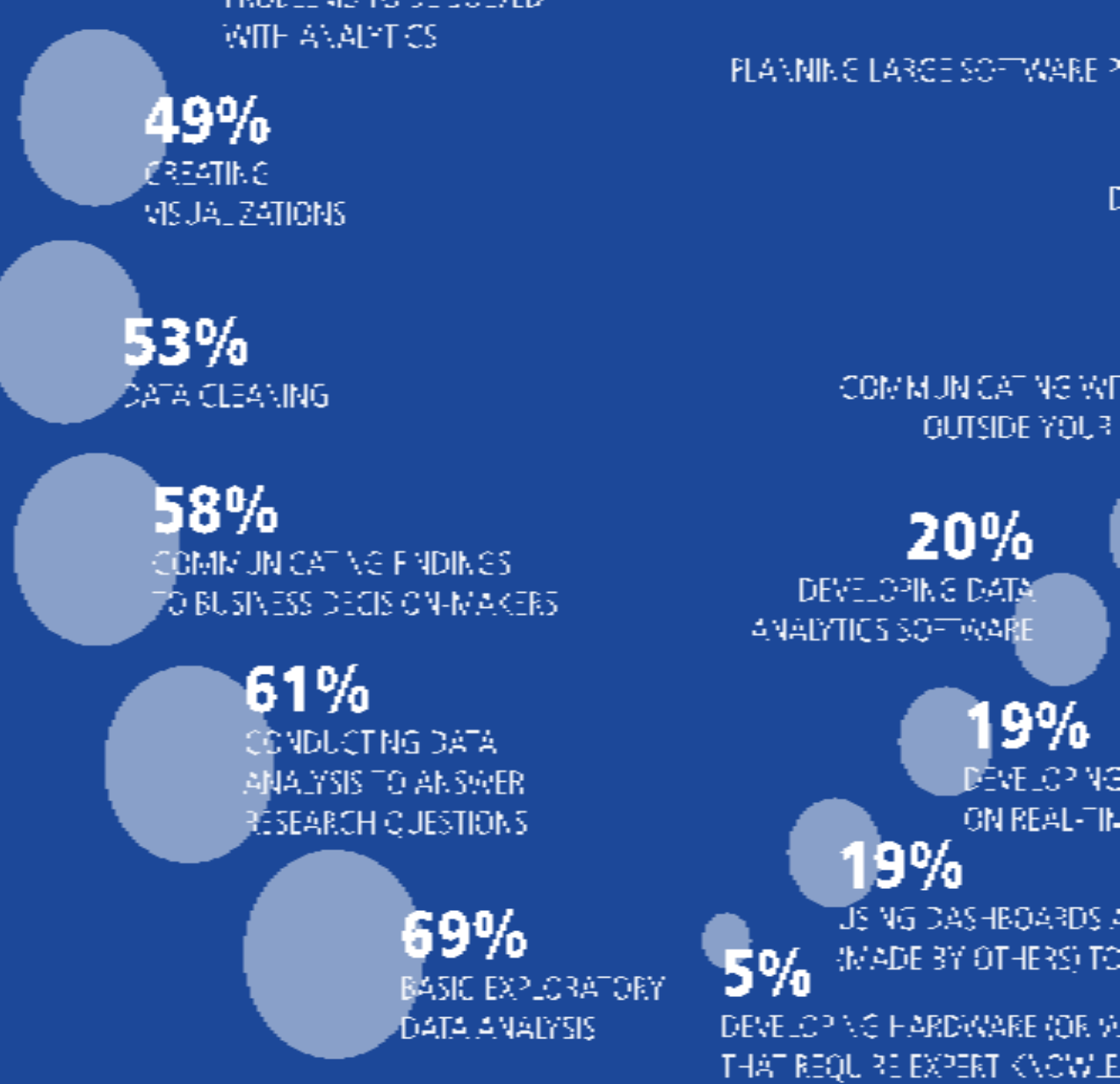
Credit: DS100.org Class1 Presentation

How involved are you in task ___:
(a) Major, (b) Minor, (c) None

Are the top items surprising?

Data Cleaning ☹

Where are Modeling / Prediction?

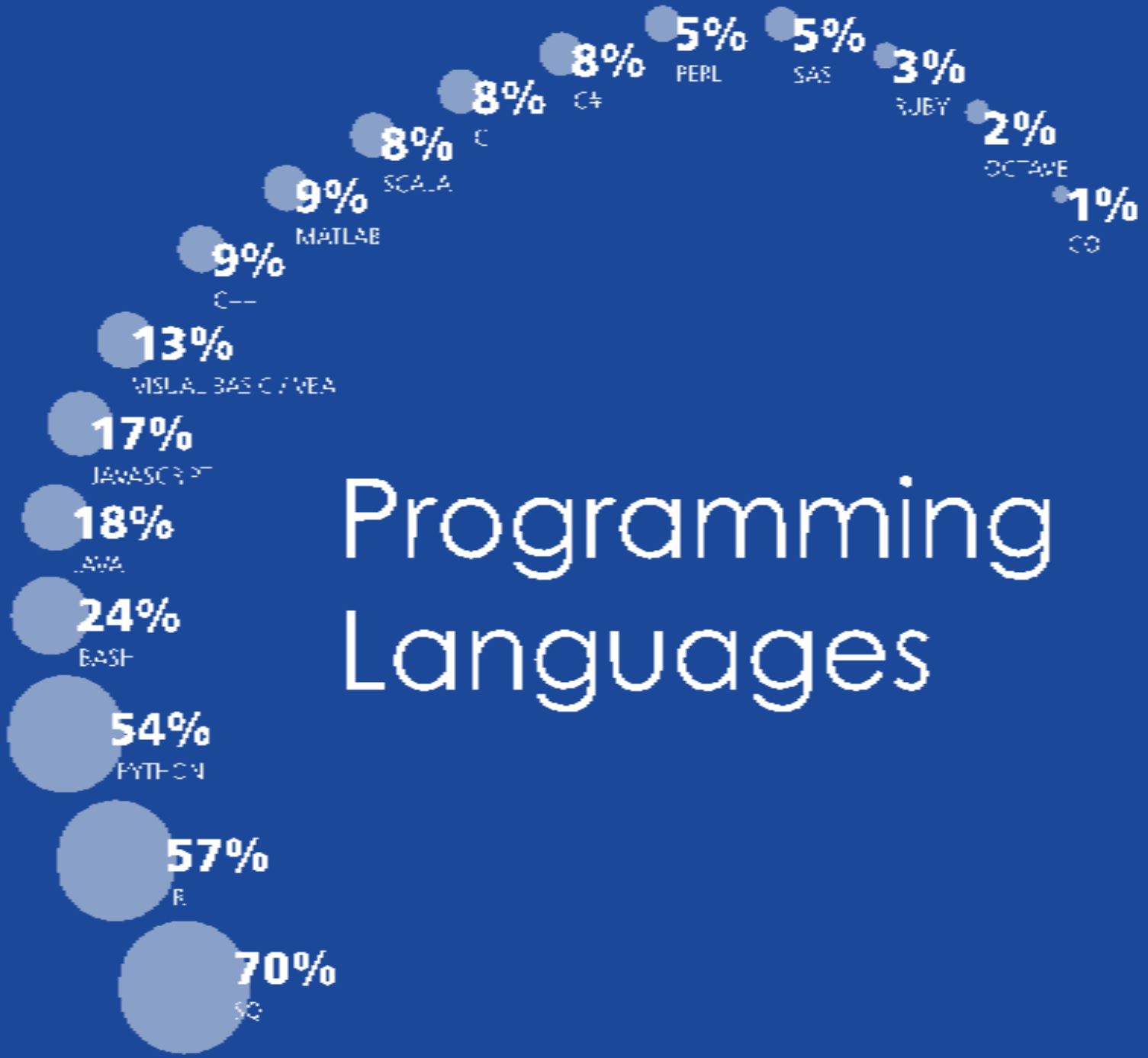


Credit: DS100.org Class1 Presentation

What tools do they use?

- Programming Languages
- Machine Learning
- Data Technologies

Credit: DS100.org Class1 Presentation



Programming Languages

SQL > R > Python

Cluster Analysis:

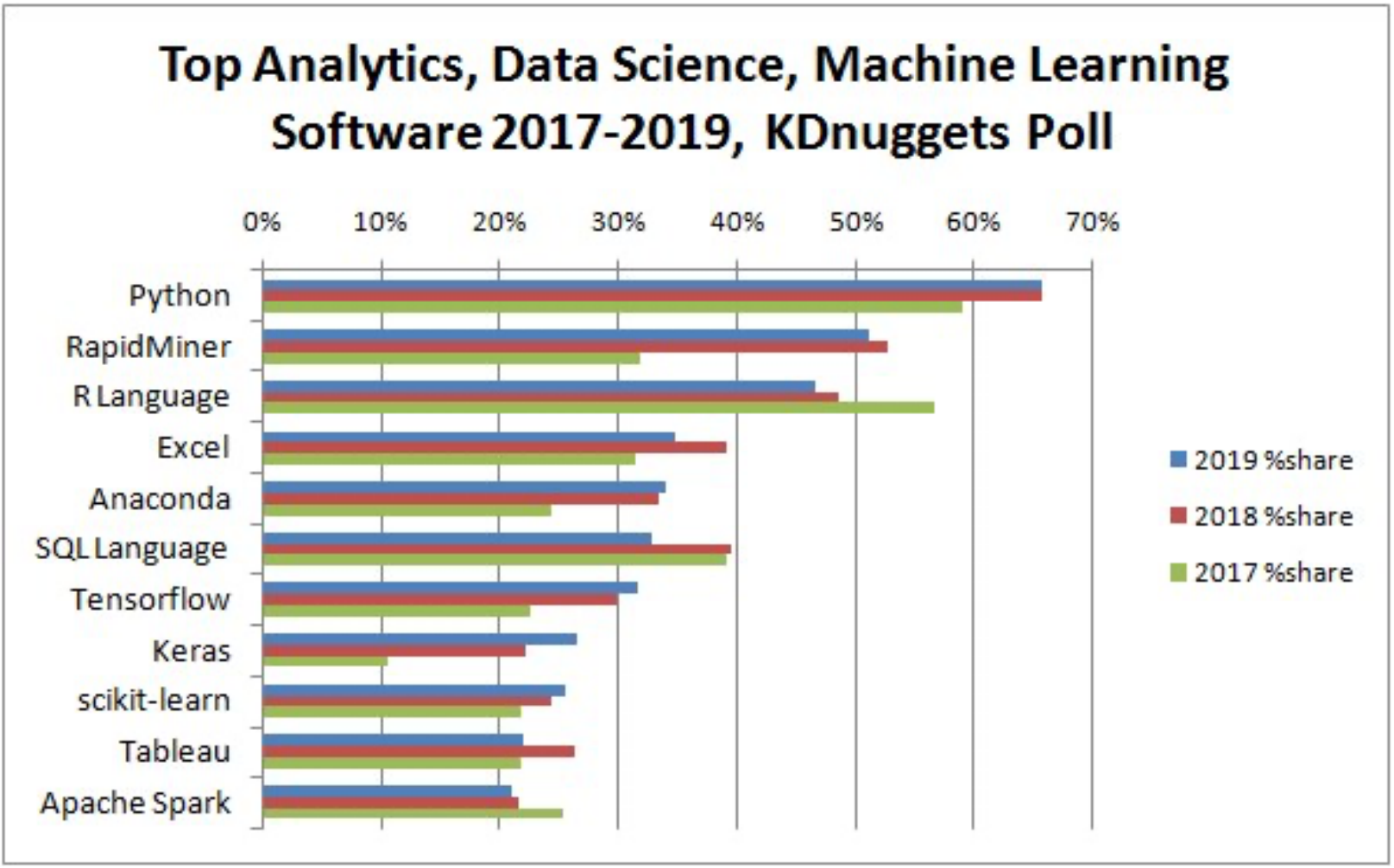
- Python > R: *data scientists*
- R > Python: *analysts*

Python users had higher salaries.

Highest Paid?

- Scala

Credit: DS100.org Class1 Presentation



Python has overtaking R.

Python is “glue” that holds machine learning ecosystems together.

What will we cover in course?

A Note on Philosophy….. If academics taught baseball…it might go something like this.

- Week 1: A history of bats, why their shape and materials.
- Week 2: The ball and stitching.
- Week 3: The bases.
- Week 4: Why 9 players?
- Week 5: When to steal bases.
- ….etc.

Doing Data Science=Playing Baseball

- Won’t include all the details
- Will provide an overall understanding of process and objectives.

Class Goals

- **Prepare** for advanced courses in analytics from across the RPI campus.
- **Enable** you to gain skills necessary to begin careers as data scientists.
- **Empower** you to apply analytics to solve real world problems.

What will we learn?

Class Overview

{JSON}

CSV

Spark

API

DATA MUNGING

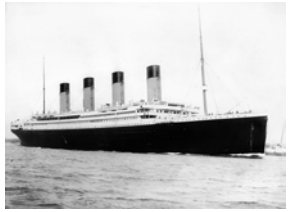
Retrieve-Filter-Missing Data-Data Cleaning-Aggregate-Merge-Missing Values-Feature Creation-Text Tools (Lemmatization-Stemming-Corpus-Bag of Words-TFIDF))-Sampling-K-Fold Cross Validation

DATA FUNDAMENTALS

Variable-Vector- Matrix-Dataframe-CSV-JSON-For Loop-if/else- Function

Modeling

kaggle



R

python

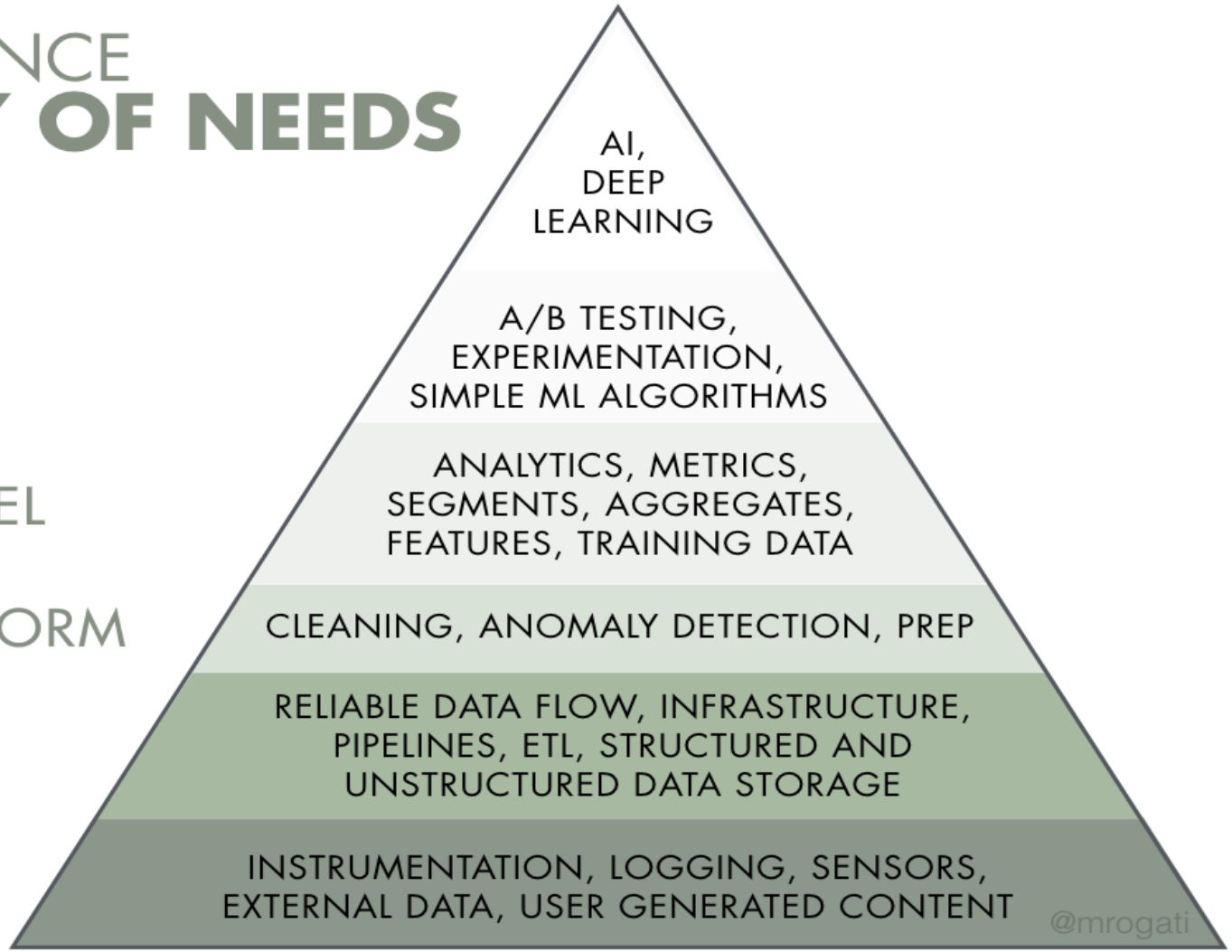
GitHub

Yummly

Basic Data Science Principles

- Defining the problem
- Data structures
- Missing data
- Exploratory data analysis
- Modelling
- Evaluation

THE DATA SCIENCE HIERARCHY OF NEEDS



LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

https://medium.com/@mrogati/the-ai-hierarchy-of-needs-18f111fcc007

Some of the topics we will focus

- Python Basics
- Introduction to R
- Unsupervised learning
- Classification
- Text and NLP
- Deep Learning
- …

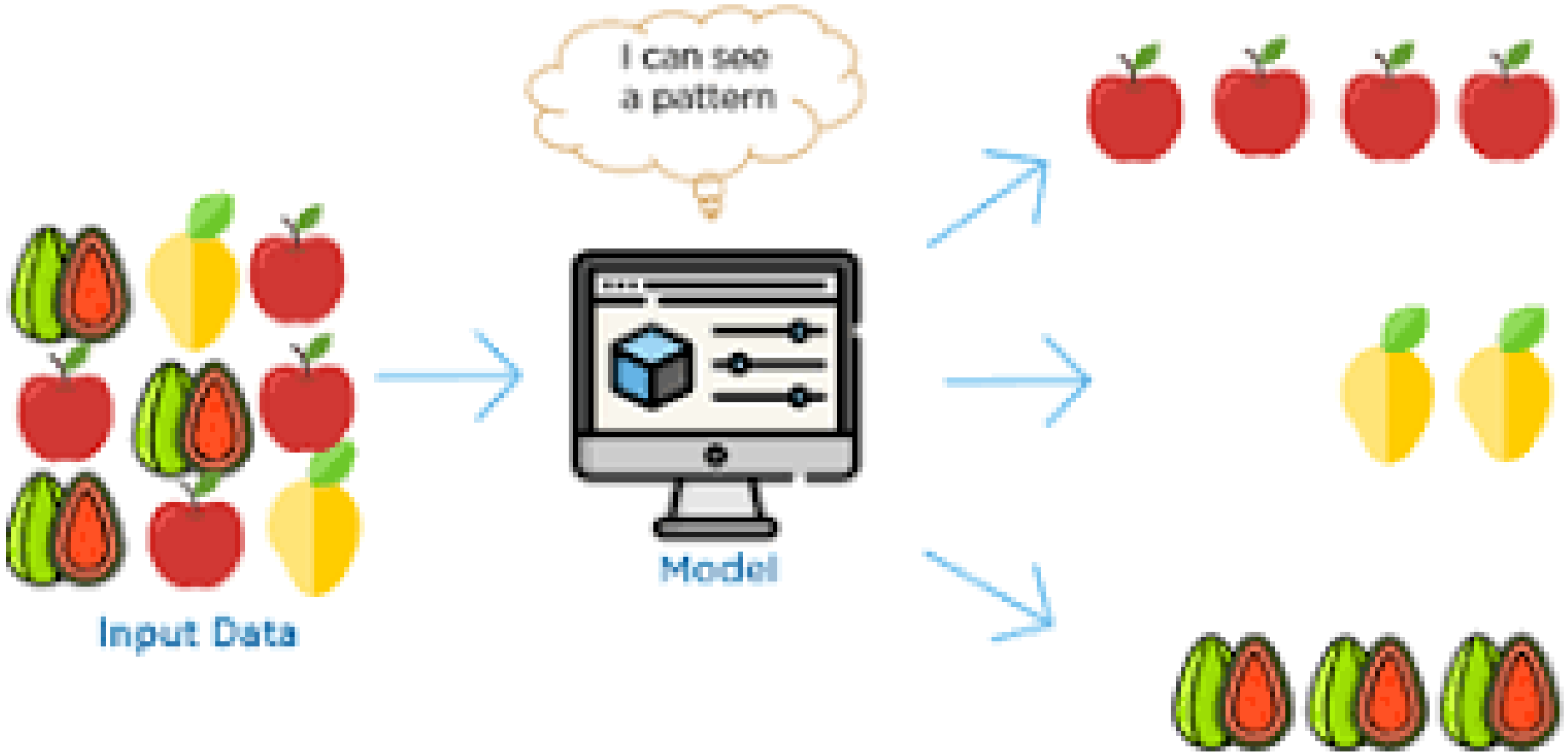
Computer Science Principles in Data Science

- Software development and version control
- Relational data models
- Issues surrounding parallelism and big data

Statistics (&CS) Principles in Data Science

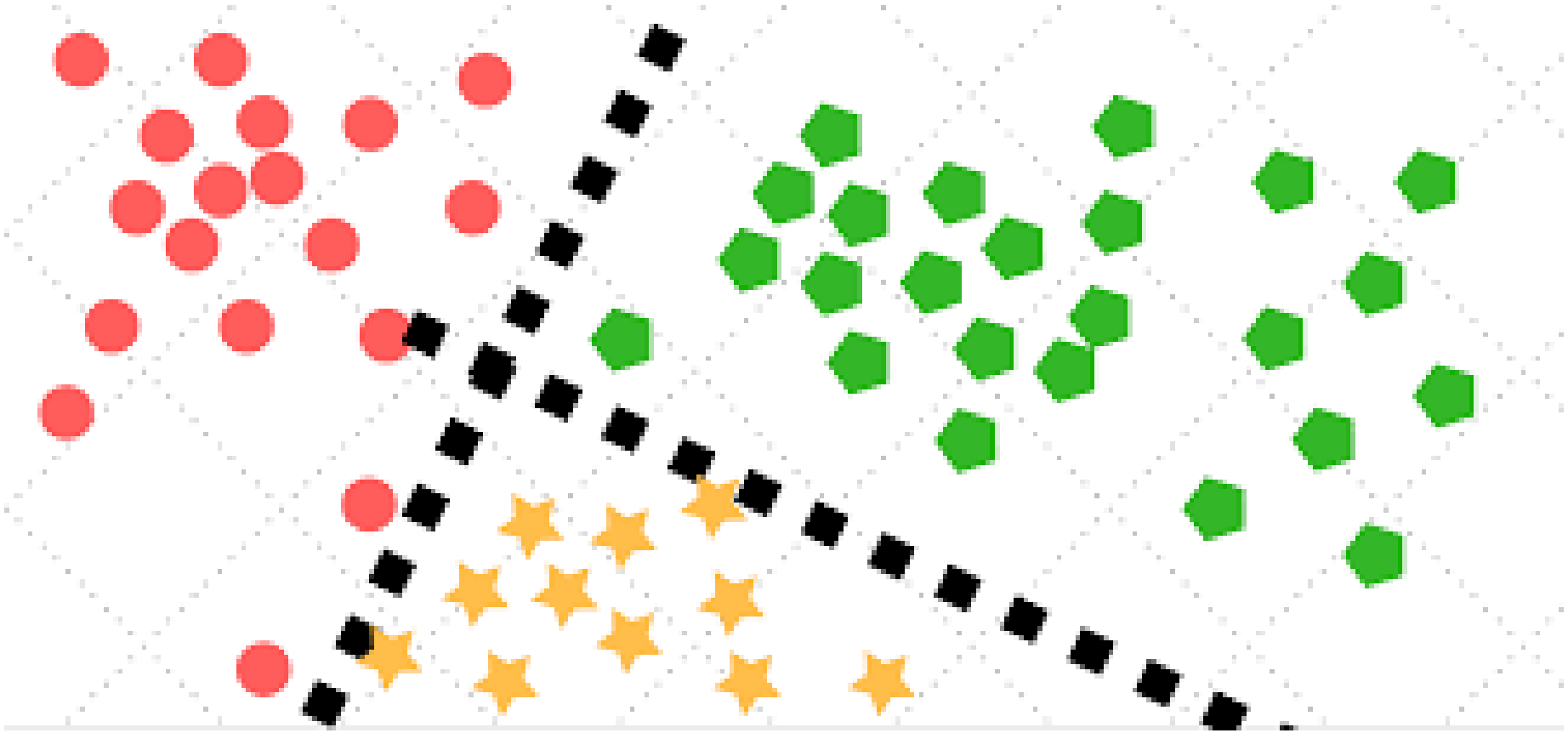
- Inference & prediction in modeling
- Model design and cross validation
- Feature extraction
- Processing of image and text data
- Issues surrounding parallelism and big data

Unsupervised Learning



Credits: Quora.com

Classification



Credits: nlpforhackers.io

Deep Learning

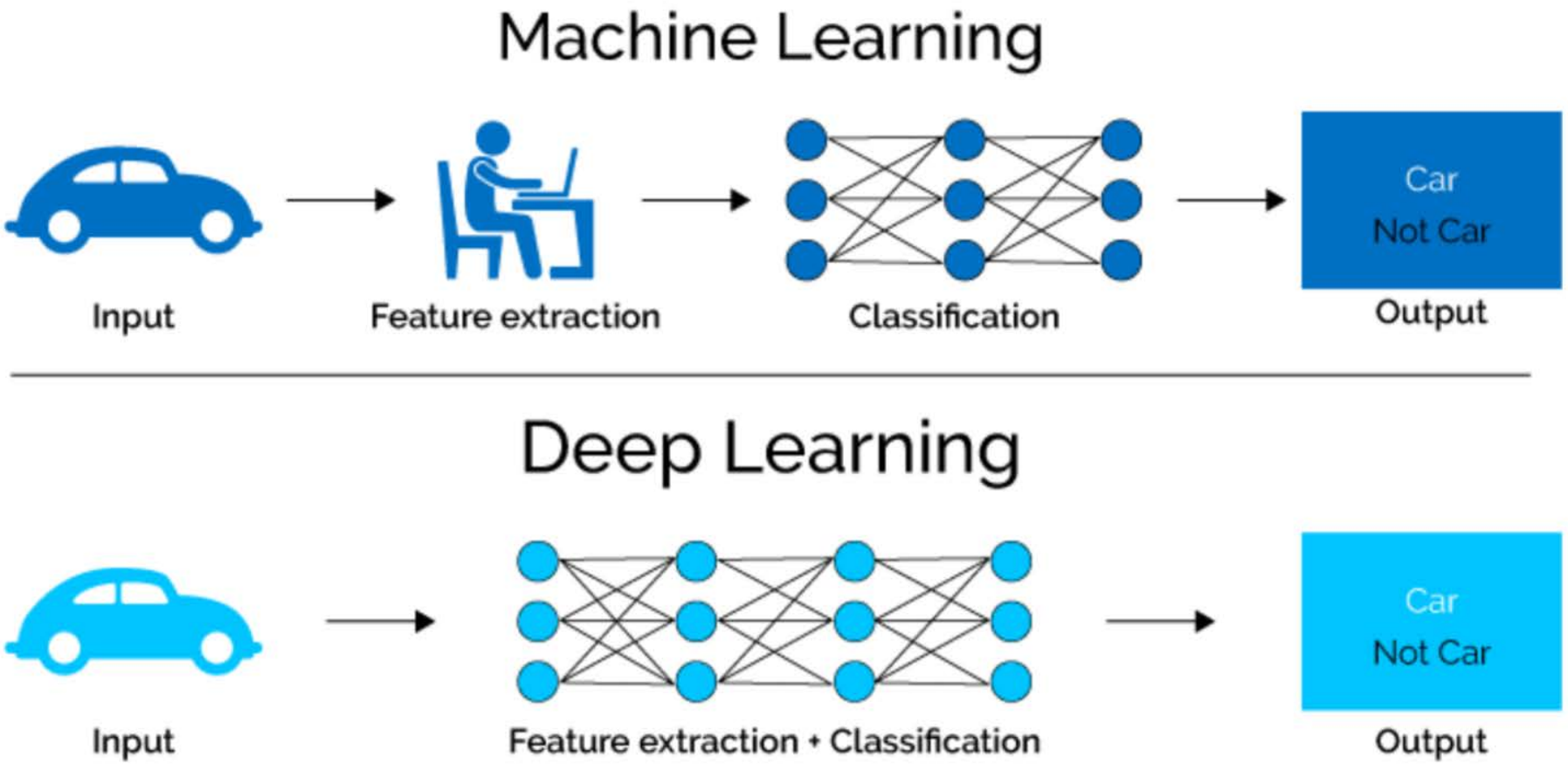


Figure 1: Machine Learning VS Deep Learning

Credits: towardsdatascience.com

Real Analytics on Real Data

Real data has issues

- Need to gain experience with issues like missing data, highly nested data
- 80% of the work a data scientist does is collecting, cleaning and organizing data

Real Tools for Analytics

- Both Python and R in Jupyter Notebooks
- Important packages (Pandas, Numpy, Seaborn)

Things will break:

- Learn how to troubleshoot code
- Get help through Slack

On Time Policy

- 3 days of “late” time for homework for sickness/deadline conflicts
- 20% per day for each late day
- Please stop by office hours virtually if having problems.

Quizzes

- Some quizzes through semester to incentivize watching videos on regular basis and not just before the test.
- There will be a window so you can take the quiz even if in a different time zone

Collaboration Policy

- It is OK to work in the same location as someone and ask questions. It is not OK to share code.
- You should produce everything that is submitted.

Communications & Homework Submissions

- Communication (Announcements etc.) will be done through Webex Teams.
  - Any questions related to the class should be posted on Webex Teams.
  - Use private option if you have a specific question that is related to your case.
  - Email communication with the professor/TA is only for emergencies.
  - Participation in the class could help in curving the grades at the end of the semester.
    - For example, 2 students received same number of cumulative points at the end, participation is considered to break the tie
- Homework submission to LMS.

Computing Environment

- Google Colab provides a computing environment for Python which is robust and free.

- Expected to eventually work on own laptop environment.

What did we learn?

Summary

- Our complex digital world world is generating tremendous amounts of data
- There a many tools to learn, but at the core a fundamental understanding of business, data, statistics, and programming is core to becoming a data science
- Descriptive, Inferential, Deterministic, Prediction for the semester are the main 4 types of problems for management sciences, machine learning focuses on prediction

1. Go to introml.analyticsdojo.com

2. Click on session 2.

3. Complete all items.