Data Science
CS 575

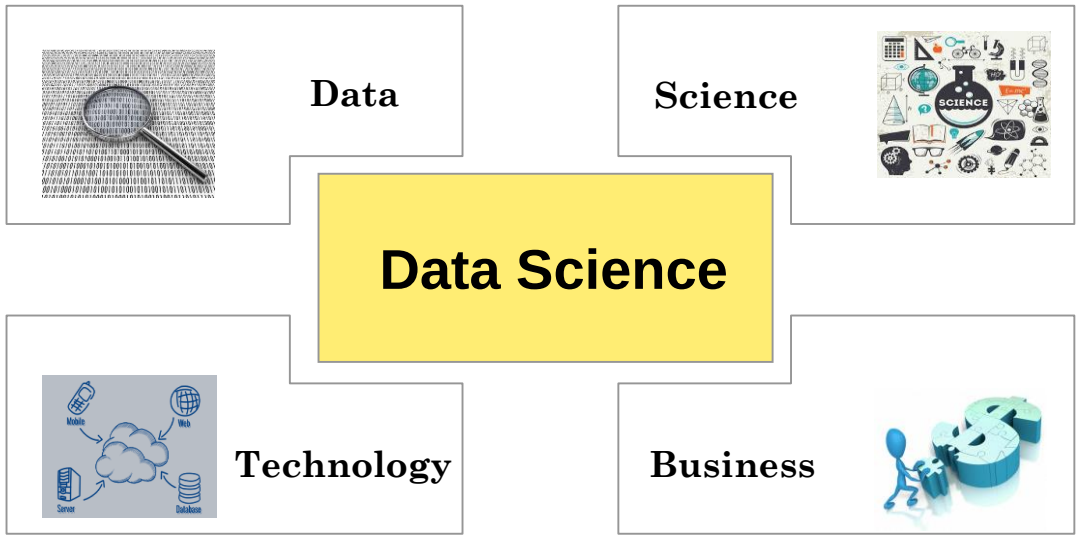
Introduction

Outline

- What ?
- Why ?
- How ?

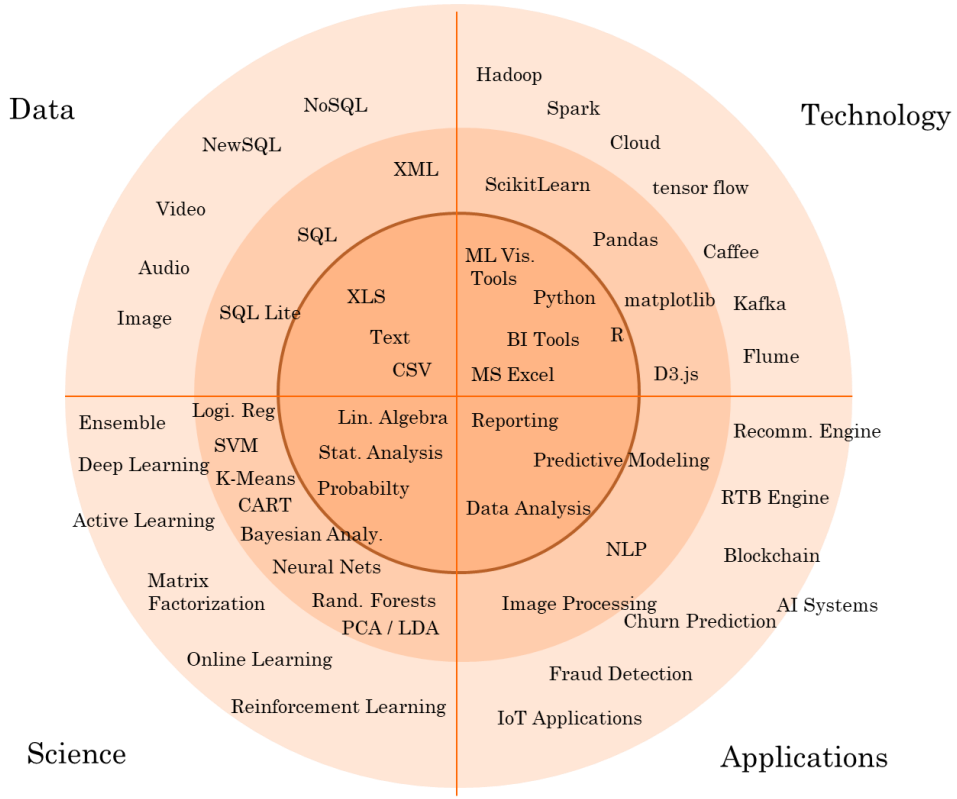
- What?

What is Data Science?

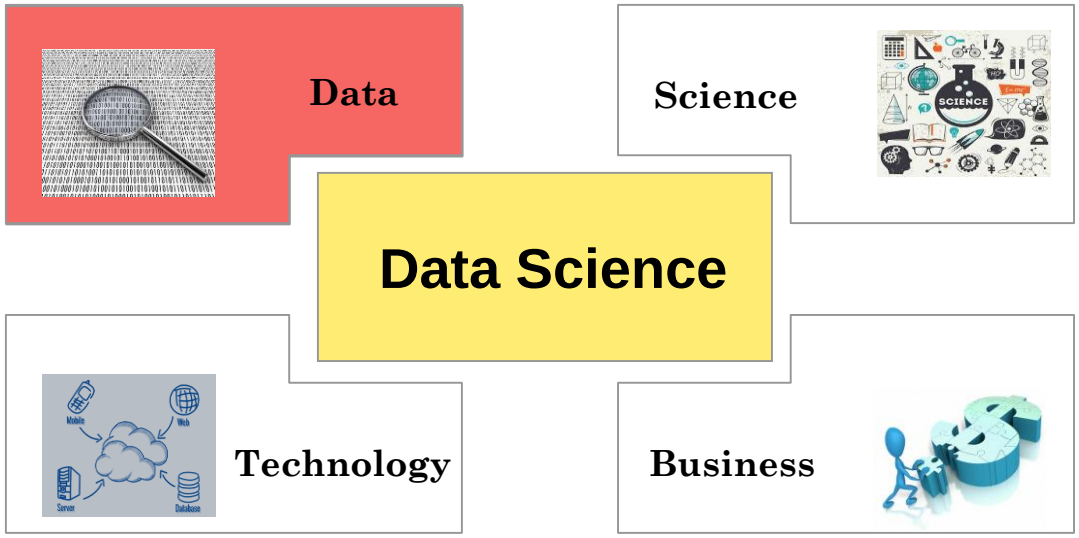


An interdisciplinary field that employs sophisticated tools and techniques to extract knowledge and actionable insights from structured or unstructured data in order to optimize business objectives.

Data Science – The Ingridients

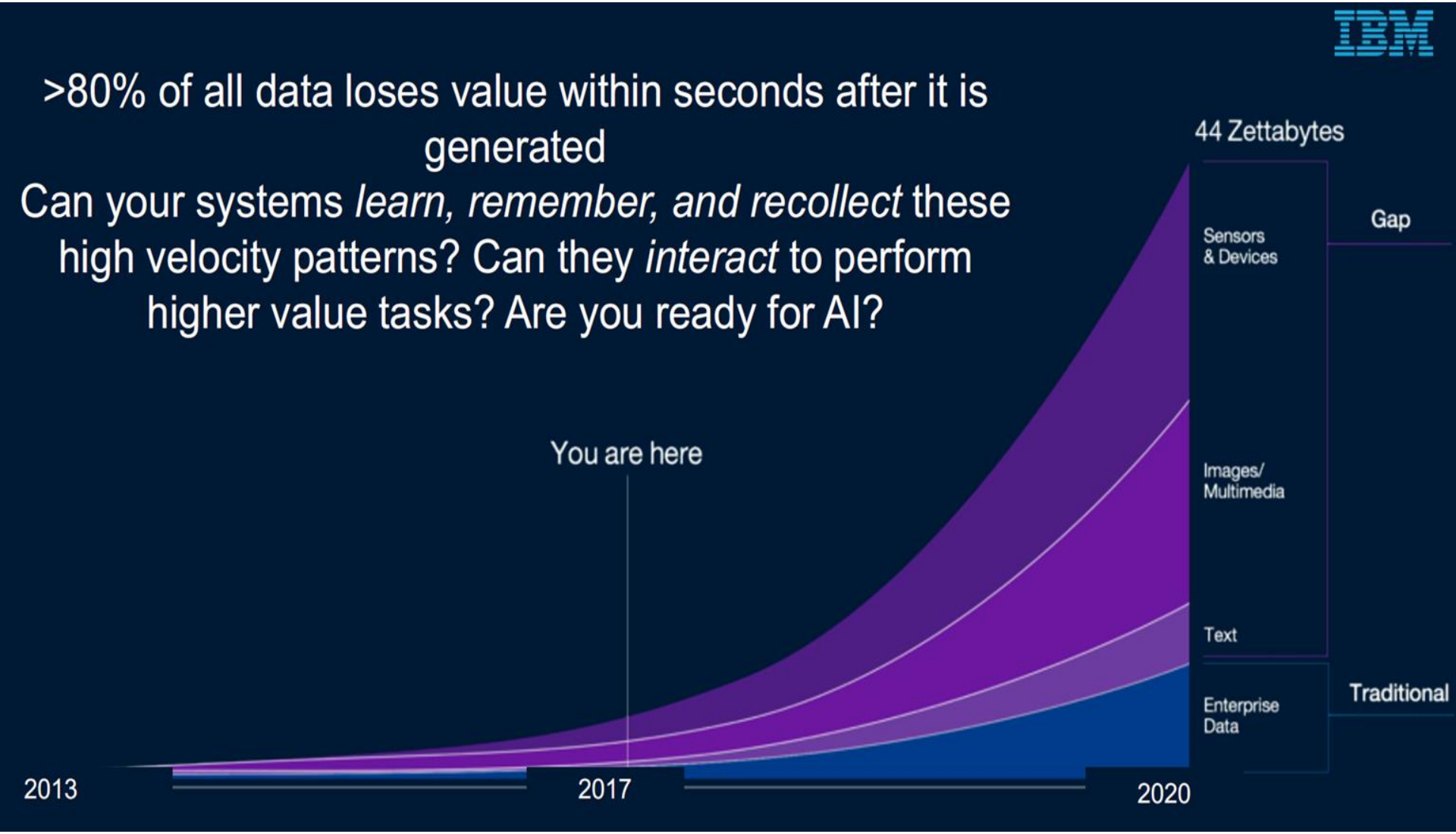


Data Science – The Data



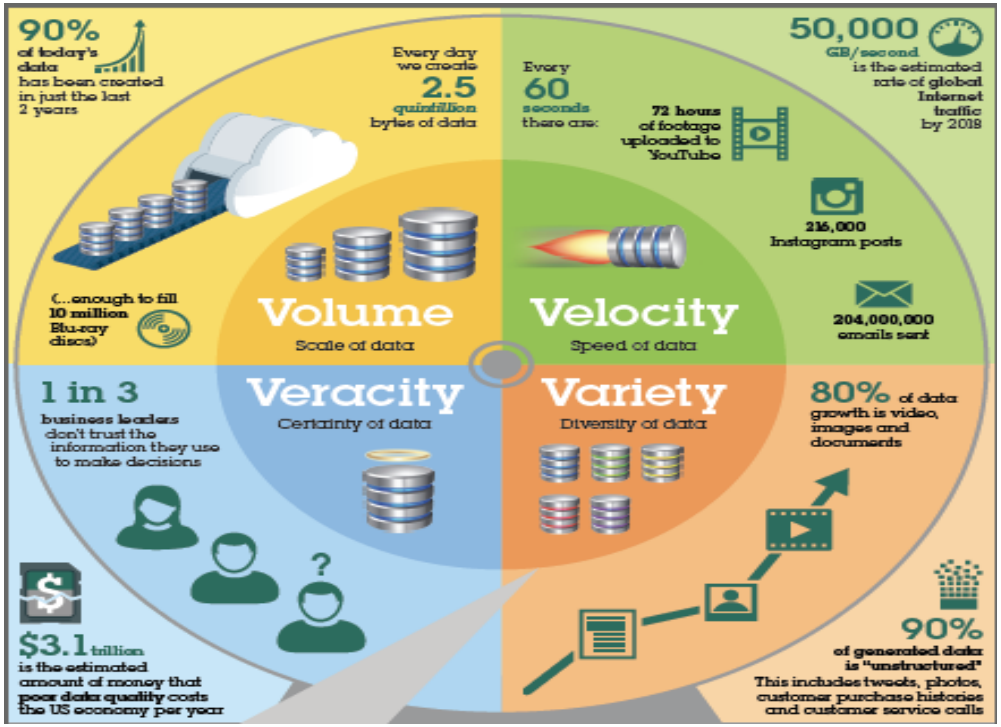
>80% of all data loses value within seconds after it is generated

Can your systems *learn, remember, and recollect* these high velocity patterns? Can they *interact* to perform higher value tasks? Are you ready for AI?



Data – The 4 V's of Big Data

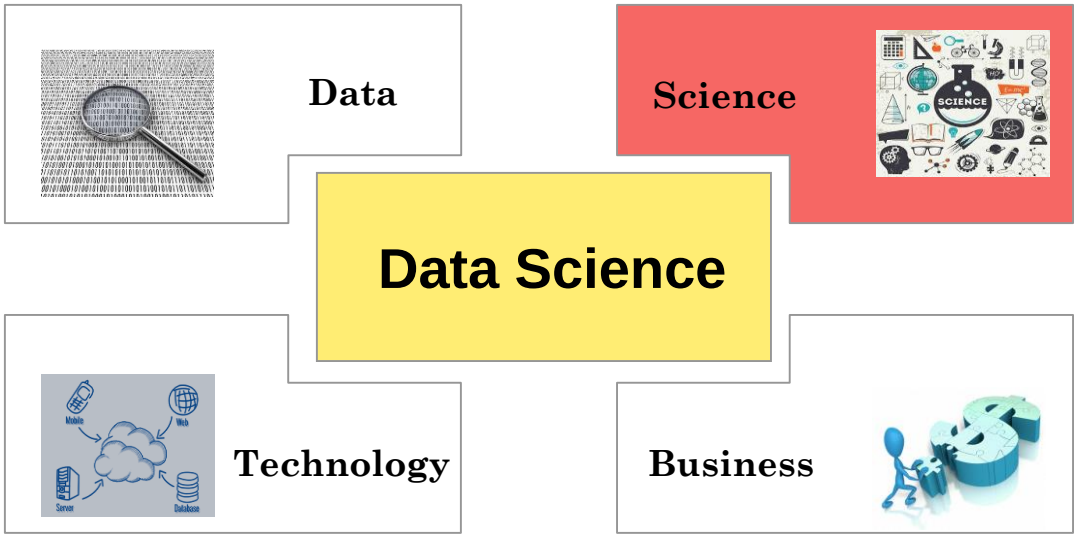
Data is the new oil - crude oil ☺



Source: IBM

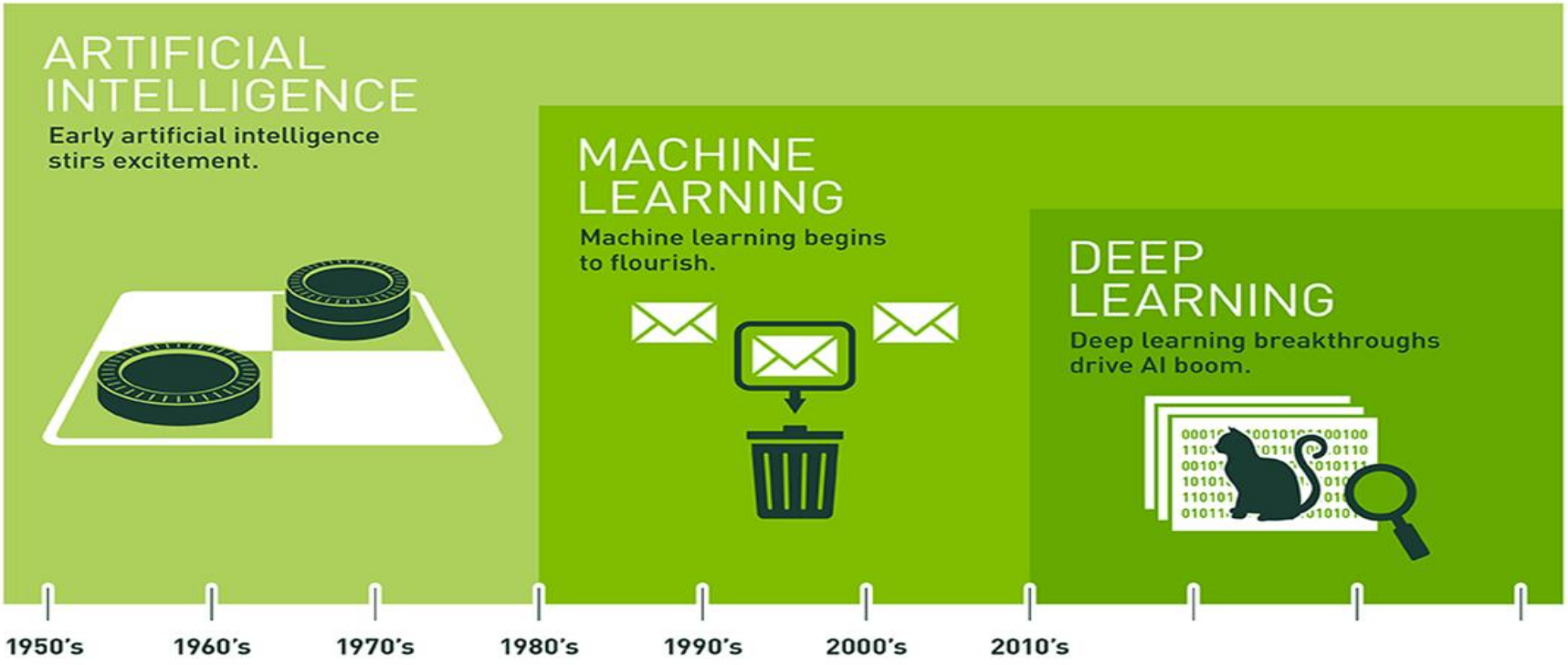

Value

Data Science – The Science

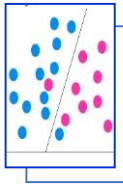


The Science – Artificial Intelligence

Data Science is the practical discipline which implements AI

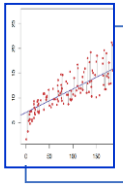


Machine Learning Methods and Business Application



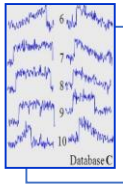
Classification

- Among all the customers which are likely to given response?



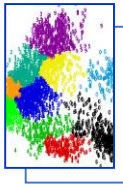
Regression

- How much will a given customer use the service?



Similarity Matching

- Attempts to identify similar individuals based on known about them.

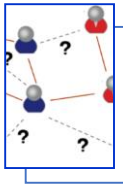


Clustering

- Do our customers form natural groups or segments?

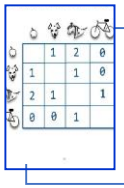
Profiling

- What is the typical cell phone usage of this customer segment?



Link Prediction

- Recommending movies to customers on the basis of watched and rated movies.



Co-occurrence

- Which items are commonly purchased together?

Answering Business Questions

Who are the most profitable customers?

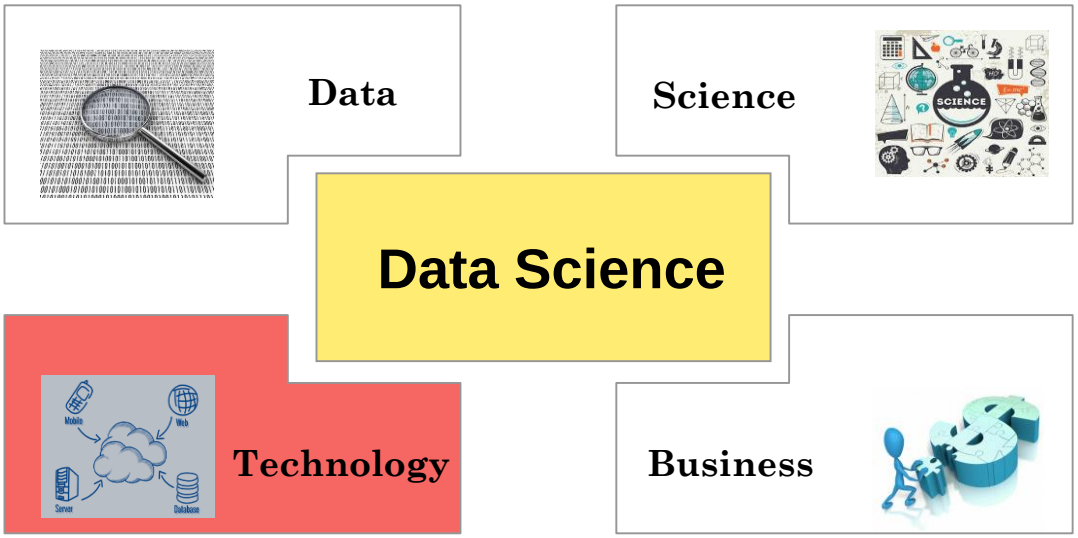
Is there really a difference between the profitable customers and the average customer?

But who really are these customers? Can I characterize them?

Will some particular new customer be profitable? How much revenue should I expect this customer to generate?

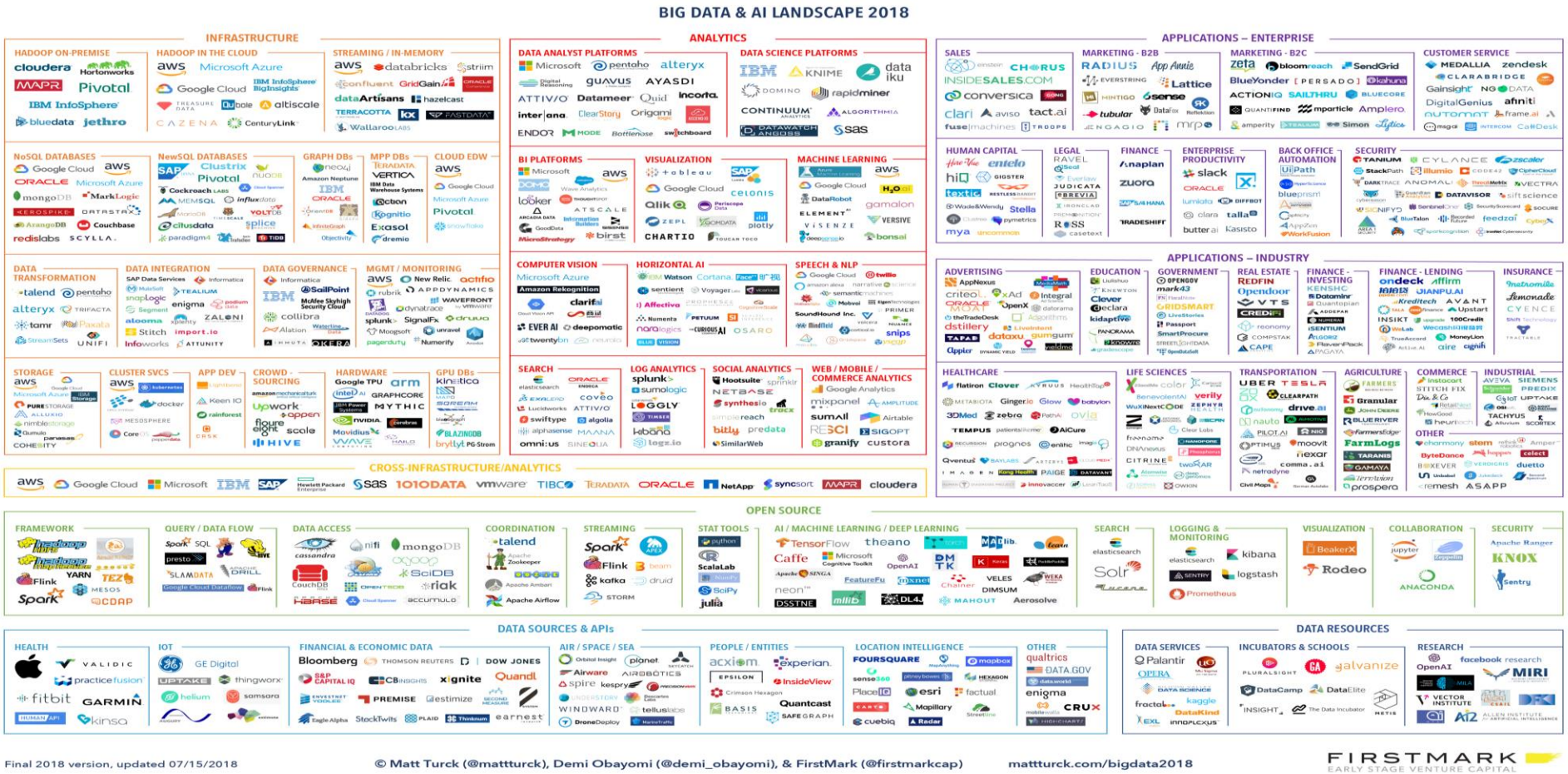
Reference: *Data Science for business by foster provost & tom fawcett*

The Technology

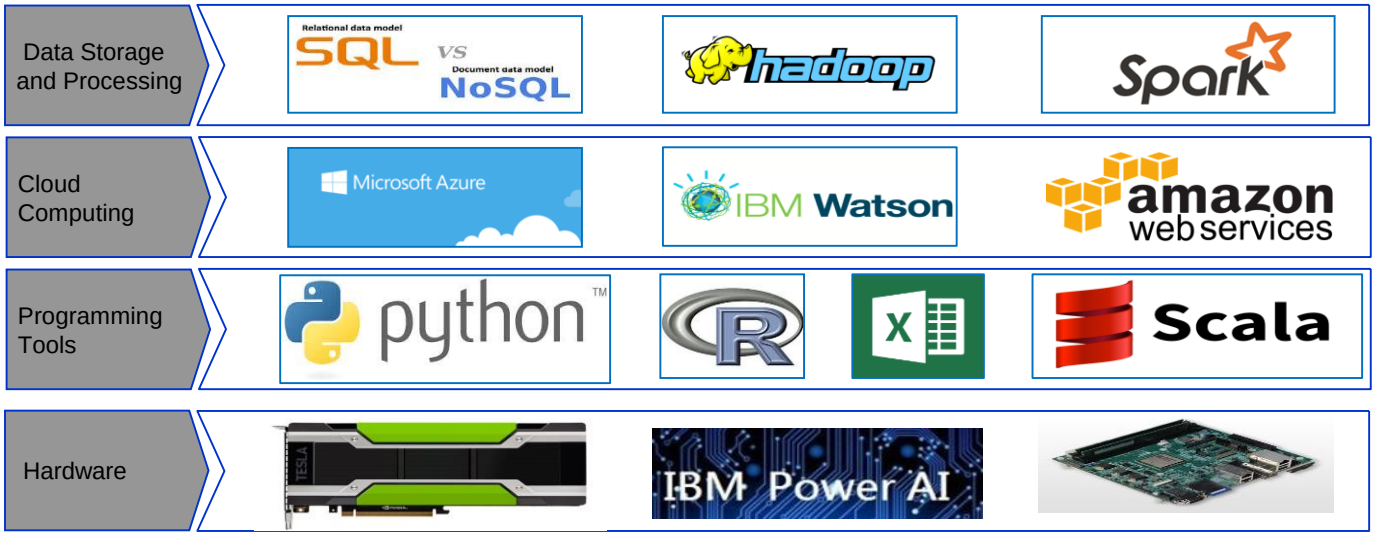


- The Tools that put life into a logical plan.

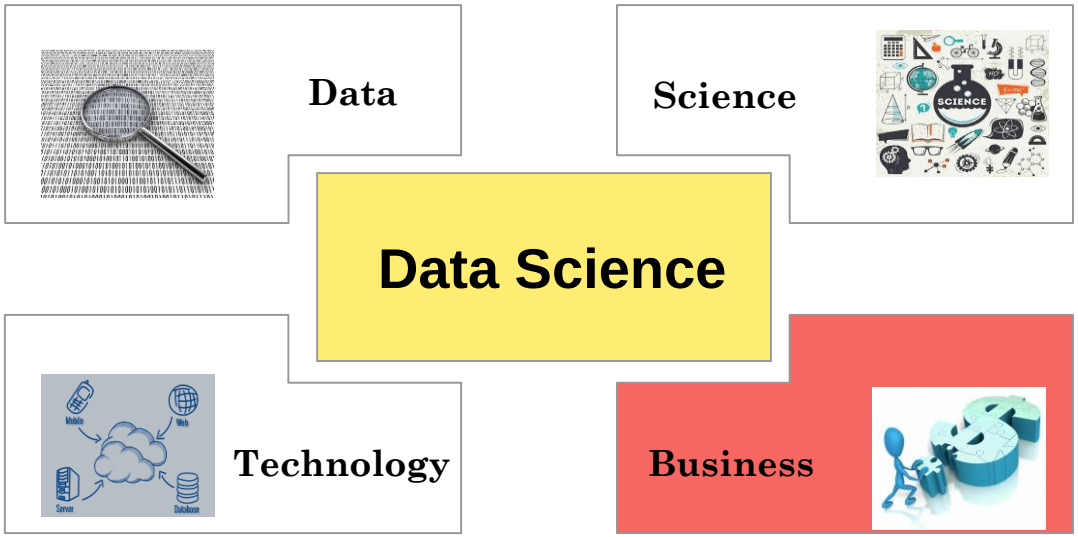
The Technology - ~~is a mess~~



The Technology – A Subset



Applications/Business – What matters at the end



The Applications – Some Areas


Automotive


Electronics


Defense


Logistics

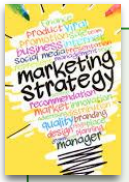

Banking

Insurance

E-commerce


Retail


Energy



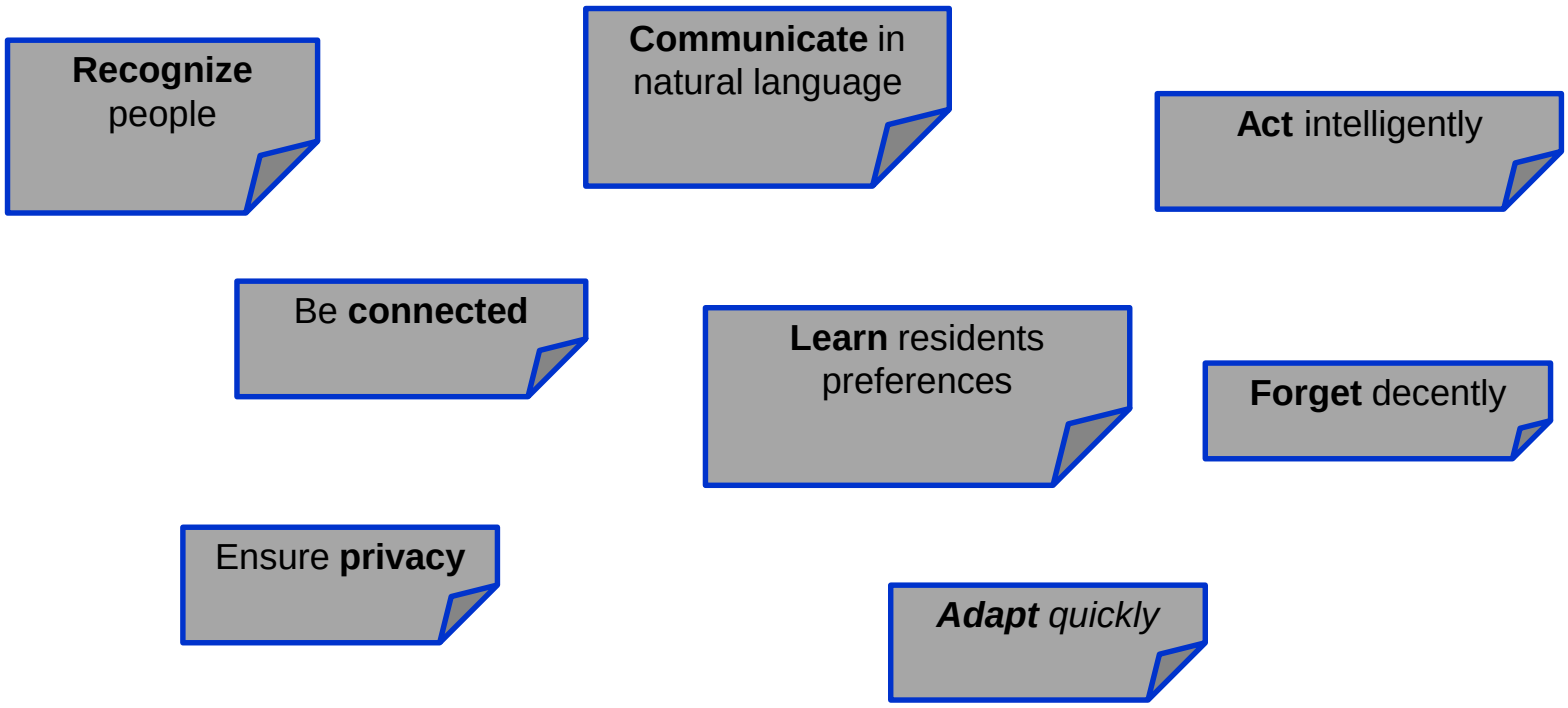
Marketing


Manufacturing


Healthcare


HUF HAUS
Menü
Design of an AI System
Self Learning Home
https://www.youtube.com/watch?v=U-Nty35p370

Self-Learning Home - Expectations



- Why ?

The Age of big data

Big Data

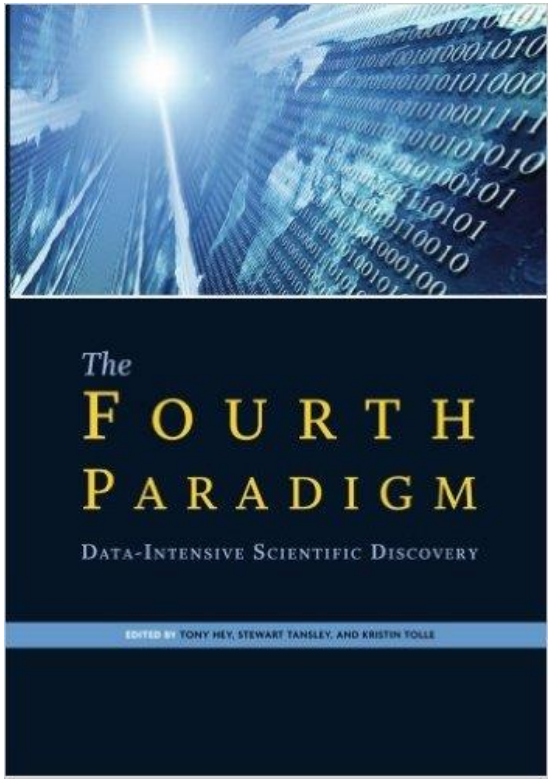
- Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.”
- Eric Schmidt, Google

Science Paradigms

- Thousand year ago:
  - Science was empirical
    - Describing natural phenomena
- Last few hundred years:
  - Theoretical branch
    - Using models, generalizations
- Last few decades:
  - Computational branch
    - Simulating complex phenomena

Science Paradigms

- Today: data exploration (eScience)
  - Unify theory, experiments, and simulation
    - Data captured by instruments or generated by simulator
    - Processed by software
    - Information/knowledge stored in computer
    - Scientist analyzes database/files using data management and statistics.



Jim Gray, Microsoft

#1 Job (2016/2017)

2016

Data Scientist (#1), Tax Manager (#2) and Solutions Architect (#3) stand out as the three Best Jobs in America for 2016. But which other jobs made the cut?

https://www.glassdoor.com/blog/25-jobs-america-2016/

2017

1 Data Scientist

4.8 / 5
Job Score

4.4 / 5
Job Satisfaction

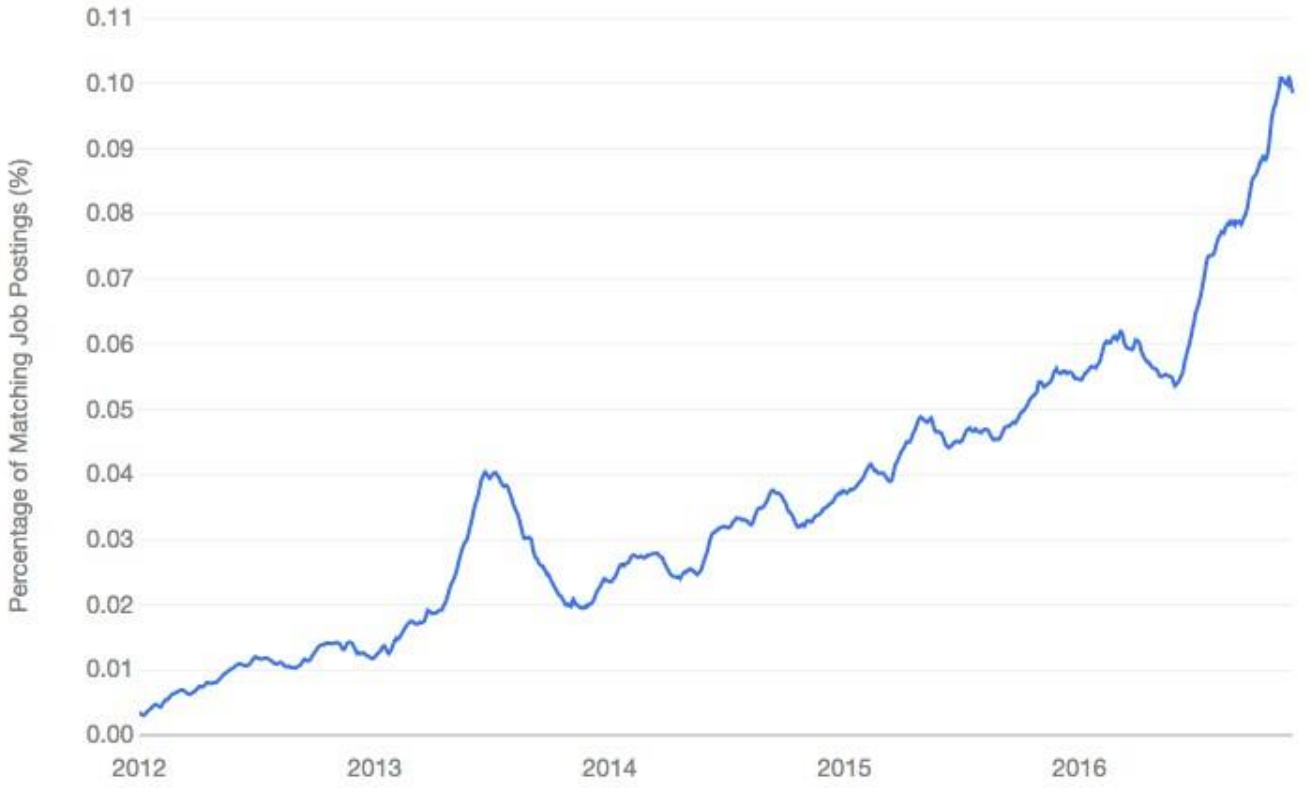
$110,000
Median Base Salary

4,184
Job Openings

View Jobs

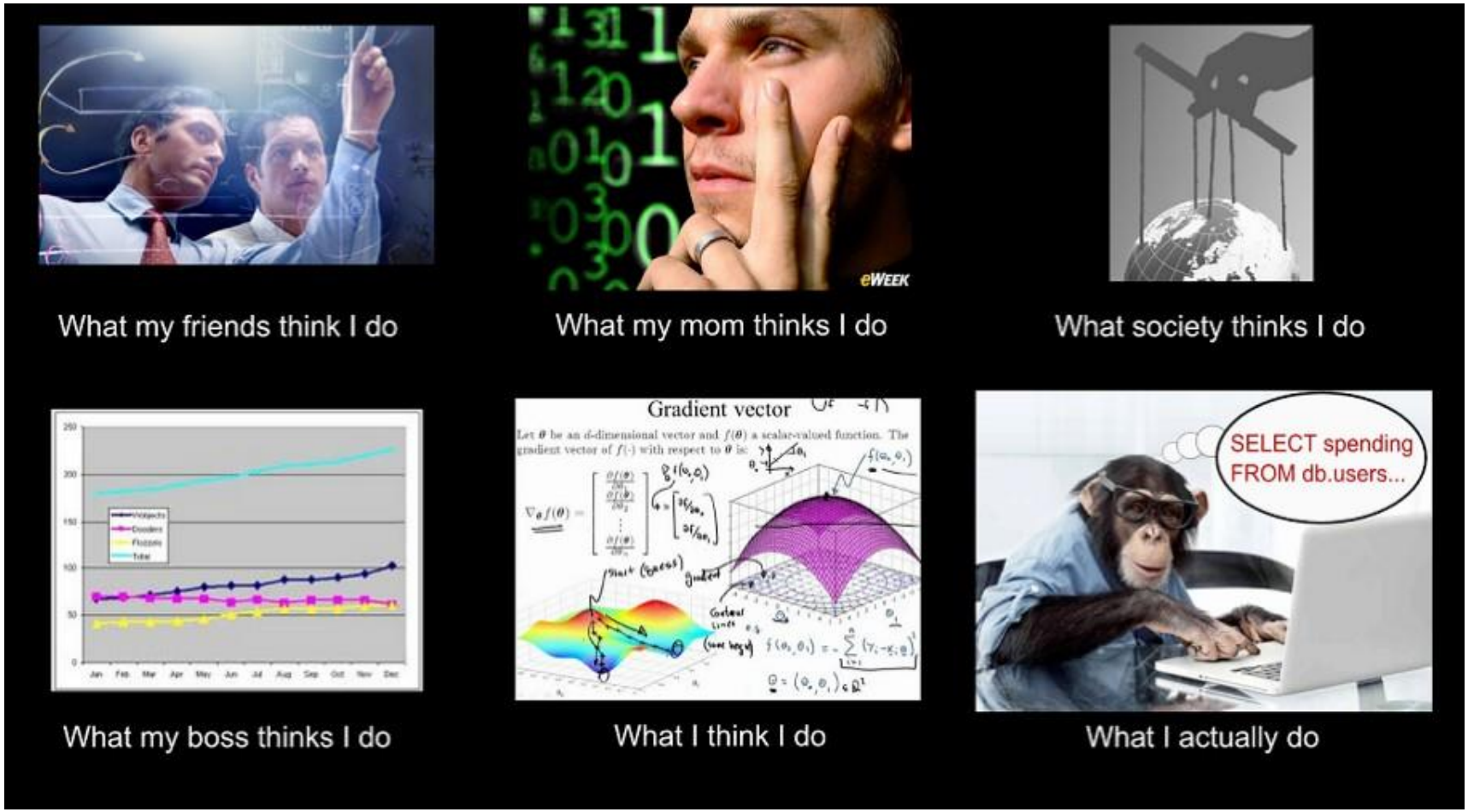
https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

INDEED JOB TRENDS FOR "DATA SCIENTIST"



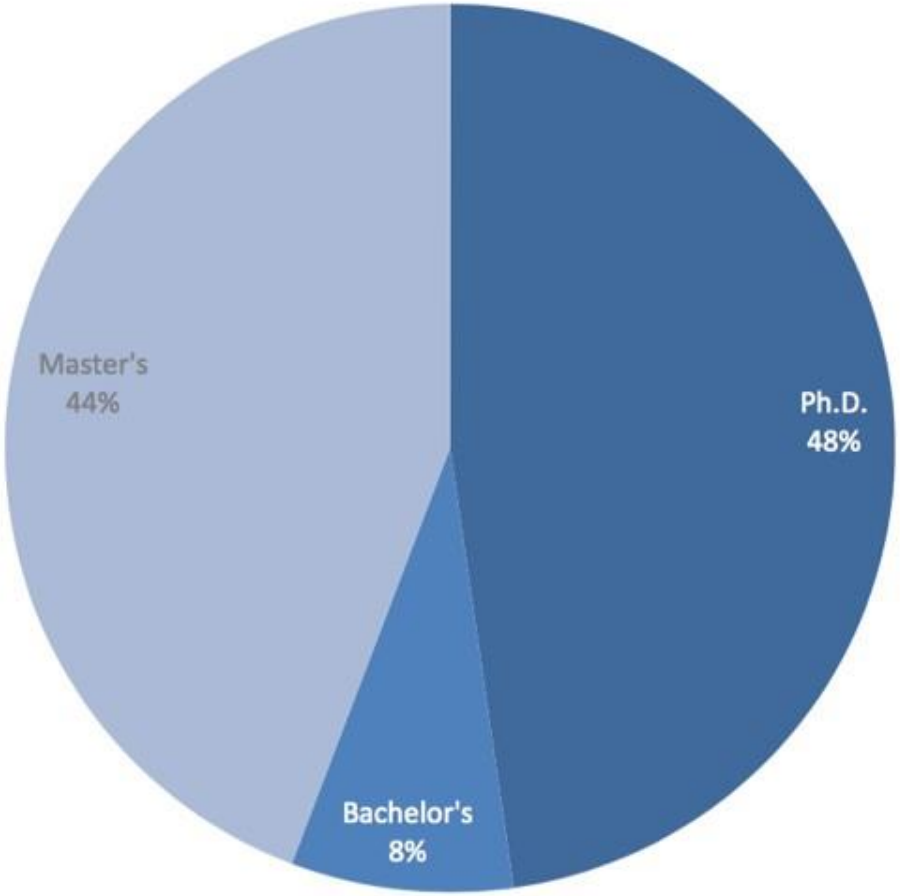
https://www.indeed.com/jobtrends/q-%22Data-Scientist%22.html

TRAINING TO BE A DATA SCIENTIST



http://www.sintetia.com/wp-content/uploads/2014/05/Data-Scientist-What-I-really-do.png

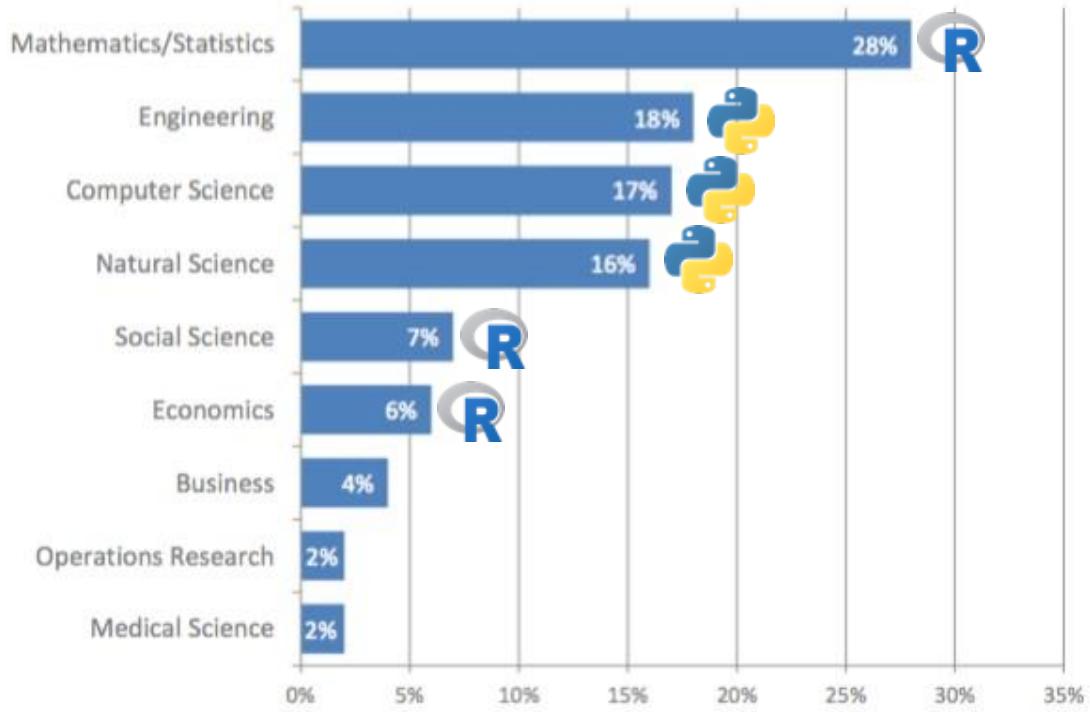
DO YOU NEED A PH.D.?



Burtch Works 2016 Study, Data Scientist Education Levels

http://www.burtchworks.com/files/2016/04/Burtch-Works-Study_DS-2016-final.pdf

POPULAR DATA SCIENCE BACKGROUNDS

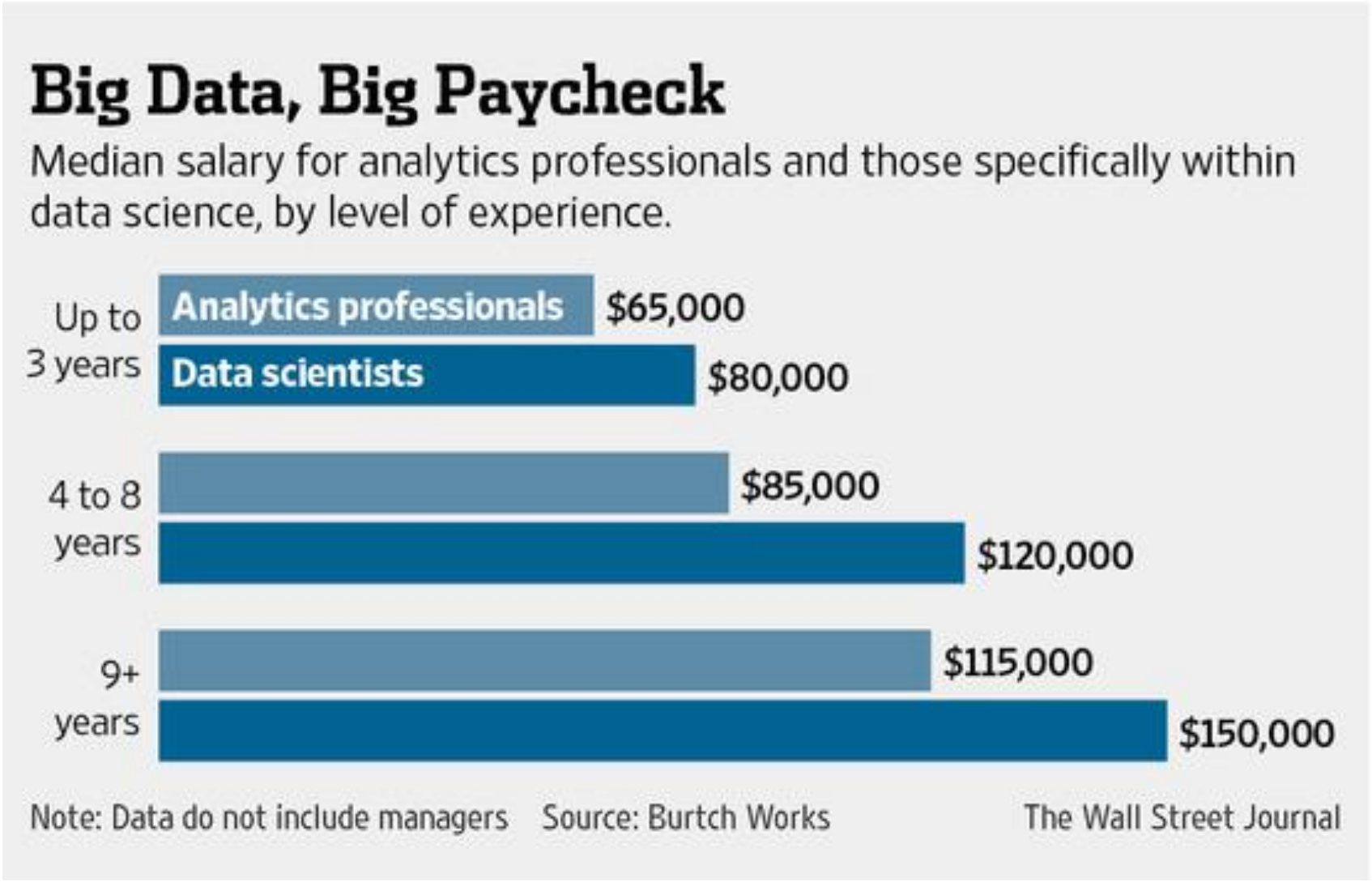


Burtch Works 2016 Study, Data Scientist Backgrounds

http://www.burtchworks.com/files/2016/04/Burtch-Works-Study_DS-2016-final.pdf

= *R more common*

= *Python more common*


Big Data, Big Paycheck
Median salary for analytics professionals and those specifically within data science, by level of experience.
Up to 3 years
Analytics professionals
$65,000
Data scientists
$80,000
4 to 8 years
$85,000
$120,000
9+ years
$115,000
$150,000
Note: Data do not include managers
Source: Burtch Works
The Wall Street Journal

Data Scientist

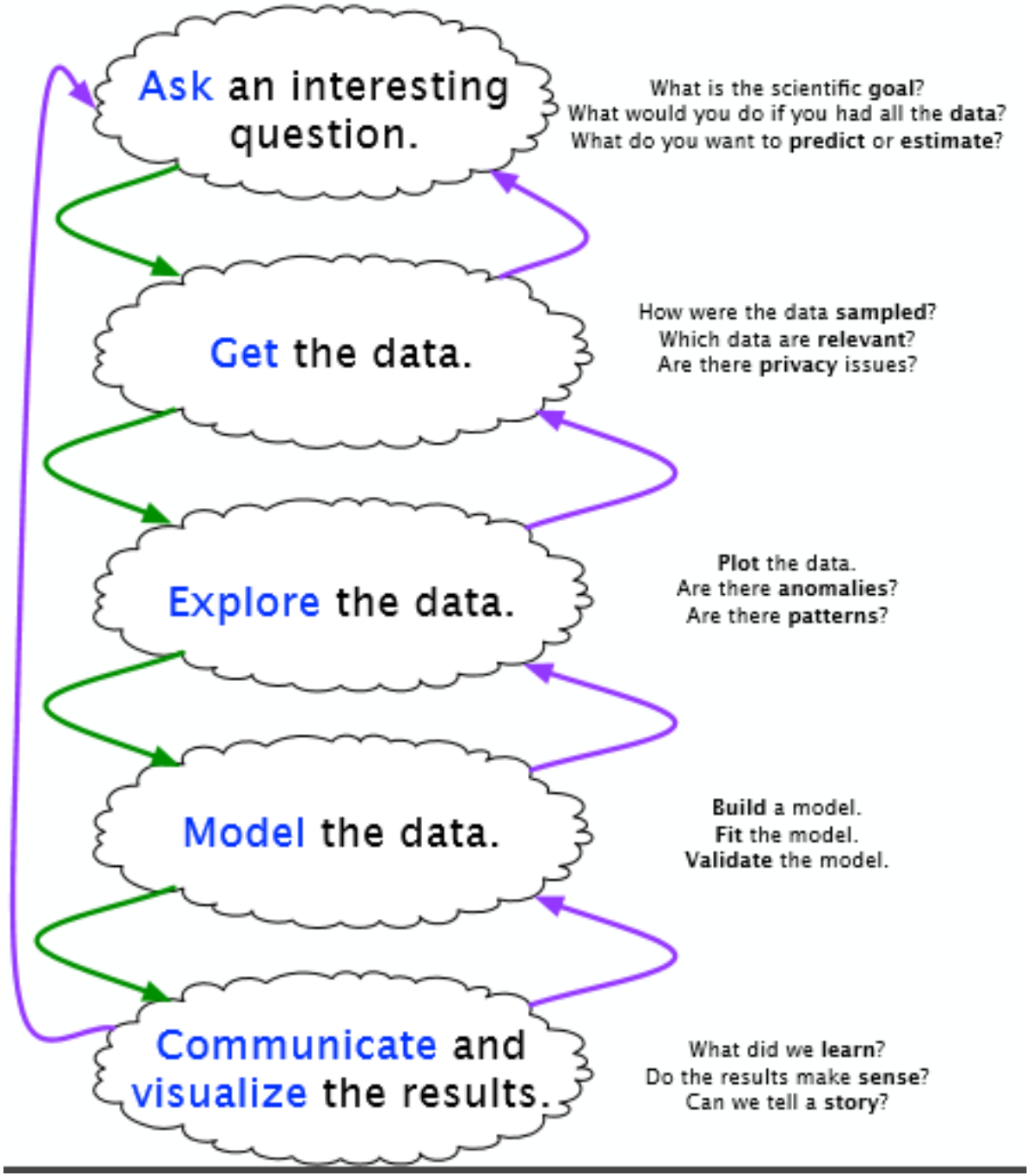
- “By 2018, the US could face a shortage of upto 190,000 workers with analytical skills”

  McKinsey Global Institute

- “Data is the new science. Big data holds the answers.” – **Pat Gelsinger, CEO, EMC**, Big Bets on Big Data

- The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data.”

– Hal Varian Prof. Emeritus UC Berkeley Chief Economist, Google


Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

How ?

- Course outline

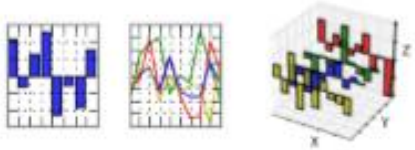
Tools

Programming

IP[y]: IPython
Interactive Computing

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



scikits learn
machine learning in Python

NumPy

SciPy.org ENTHOUGHT

matplotlib