COSC 3337 : Data Science I

N. Rizk

College of Natural and Applied Sciences

Department of Computer Science

University of Houston

Terminology

- Analyzing data has a long history!
- There have been many terms that have been used to describe such endeavors:
  - Statistics
  - Artificial Intelligence
  - Machine learning
  - Data analytics

The Good

Experiments, observations, and numerical simulations in many areas of science and business are currently generating terabytes of data, and in some cases are on the verge of generating petabytes and beyond. Analyses of the information contained in these data sets have already led to major breakthroughs in fields ranging from genomics to astronomy and high-energy physics and to the development of new information-based industries.

- Frontiers in Massive Data Analysis, National Research Council of the National Academies

The Bad

Given a large mass of data, we can by judicious selection construct perfectly plausible unassailable theories—all of which, some of which, or none of which may be right.

- Paul Arnold Srere

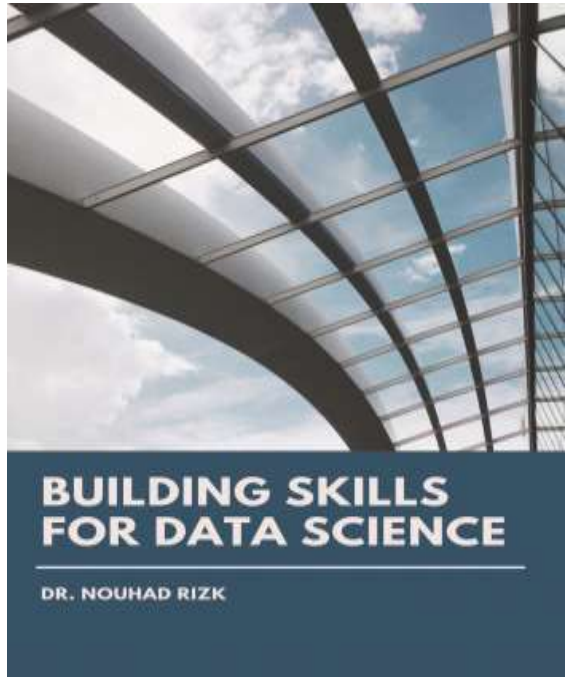
The Hopeful

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—that'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. So the complimentary scarce factor is the ability to understand that data and extract value from it.

- Hal Varian, Google's Chief Economist, http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers

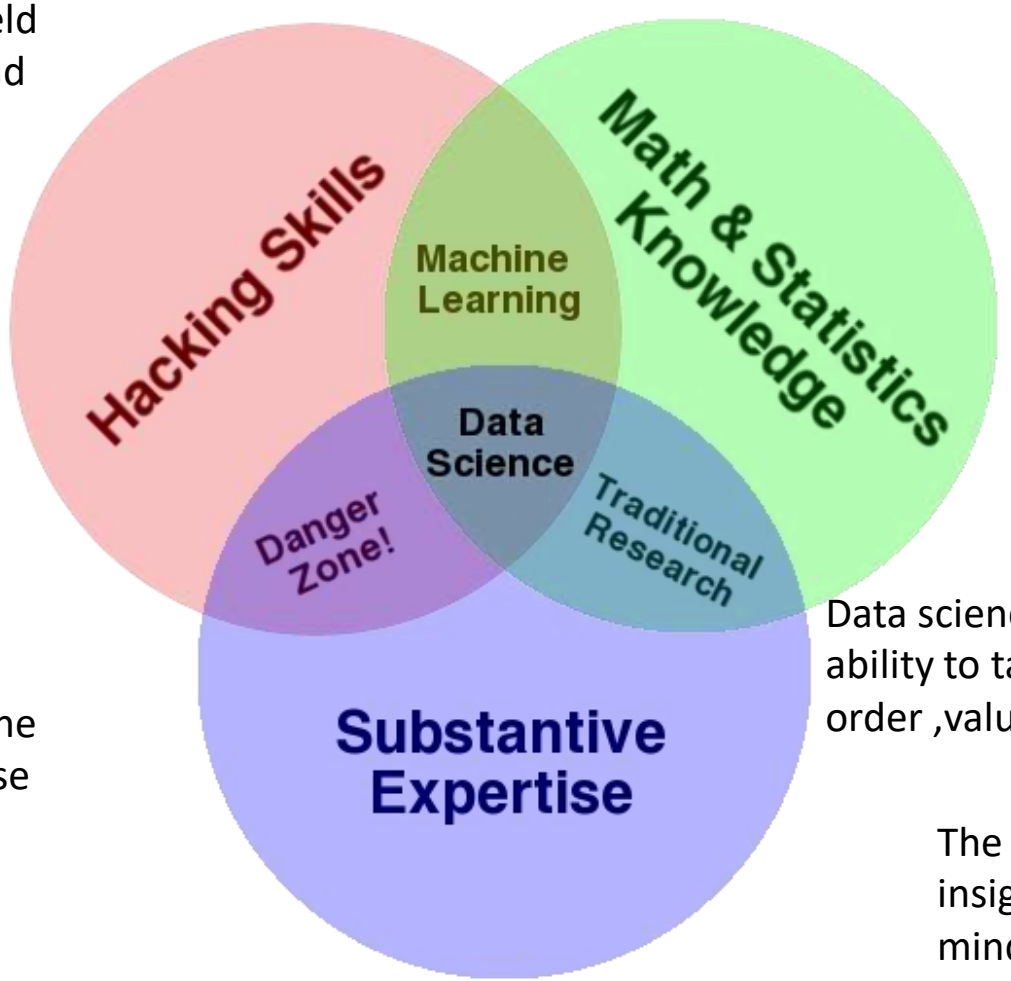
My personal goal: Getting students to be able to think critically about data.

Human beings are really good at **pattern detection...**



Skills for Data Science

Data science is the field that joins statistics and programming skills in applied settings



Data science is the analysis of diverse data

Data science qualities
ability to take unstructured data and find order ,value and meaning

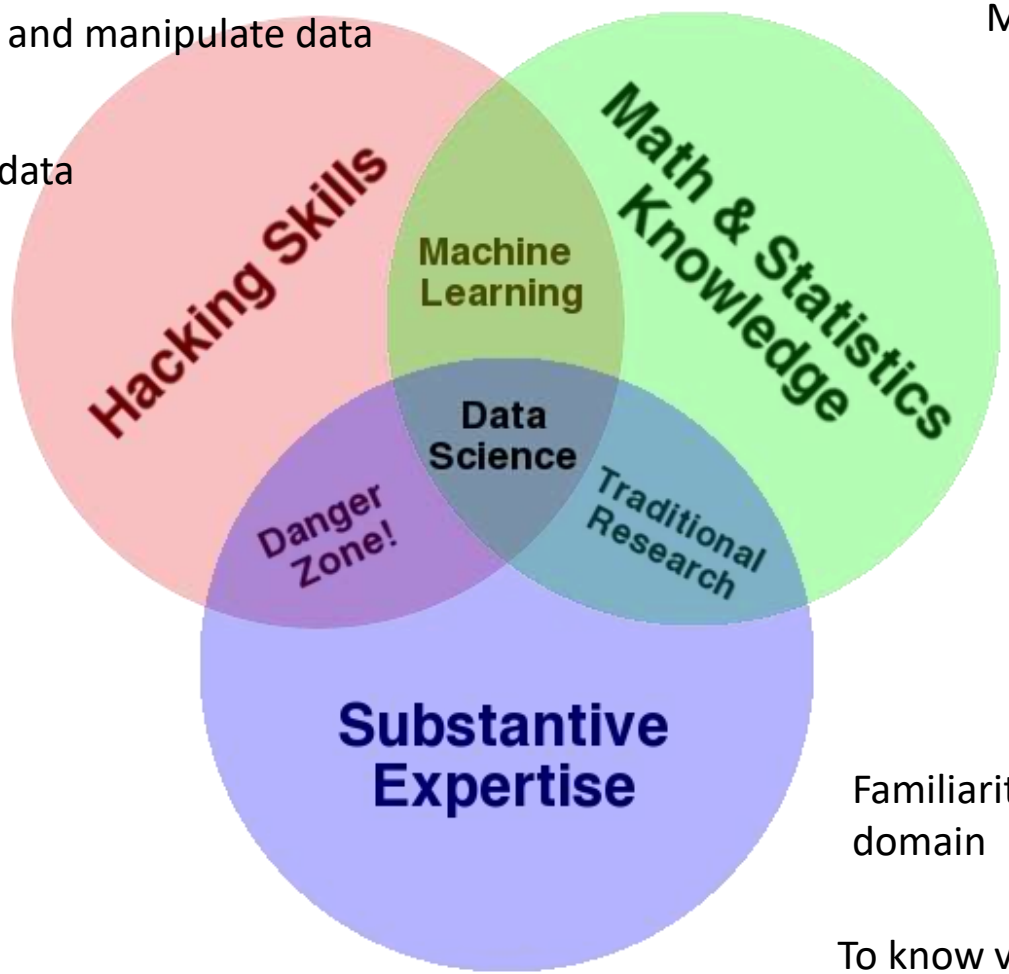
The demand is because they give insights in what is going on in people minds

Which is most important?

Coding + retrieve and manipulate data

Make sense of data

Gather and prepare data
Unusual data format
Creativity skills

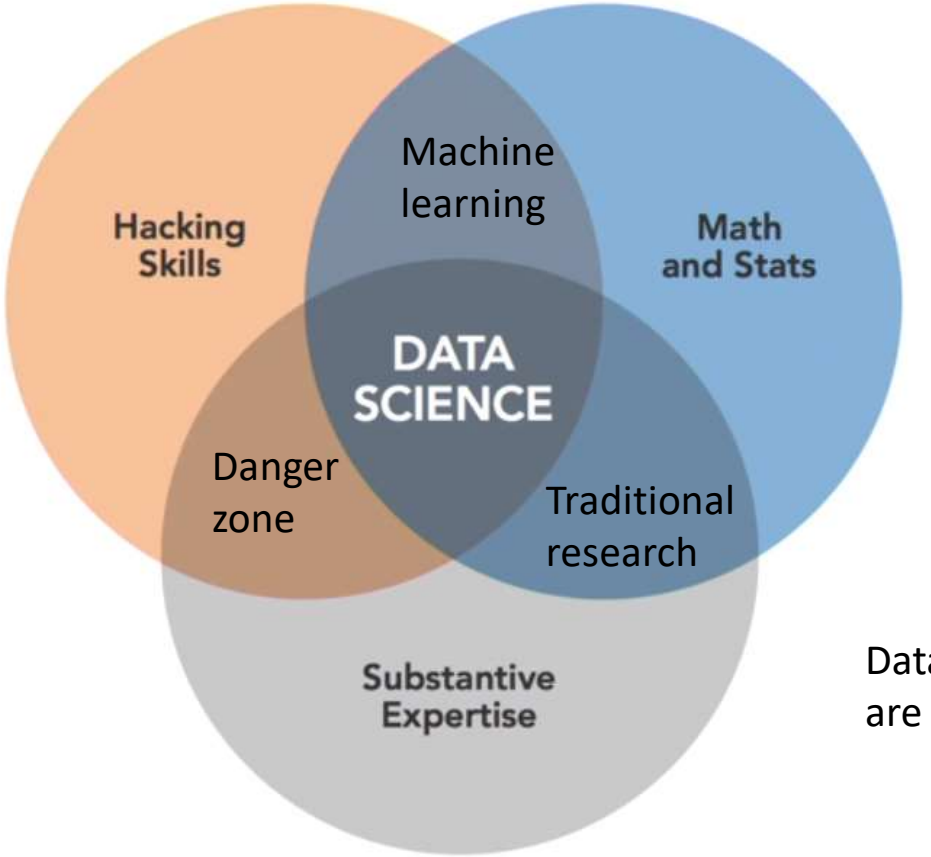


Choose procedure to
Answer question
How to diagnose problem
How to develop and improve procedures
to deal with new data challenges

Familiarity with applied domain

To know value , goals and restrictions of a specific domain

Black box predictive model



Unlikely to happen as this person can develop math and statistics expertise

Data sets and analysis are structured

Role of a data scientist?diverse skills ?Background ?Emphasis

How to do **collaborative** project in data science?

1- planning

2-data preparation

3-modeling

4-follow up

1-planning tasks

- a. Deifne goals
- b. Organize resources (+time)
- c. Coordinate people
- d. Schedule the project

2-data preparation tasks

- a. Get data
- b. Clean data
- c. Explore data
- d. Refine data (choice of cases obsverved)

3- modeling tasks (data analysis)

- a. Create model
- b. Validate model (accurate ?generalize well)
- c. Evaluate model (how accurate and informative)
- d. Refine data (?tweaks to make more informative and implementable)

4- follow up tasks

- a. present model
- b. deploy model
- c. Revisit model
- d. Archive assets (all the steps)

Data Science A Collaboration

Technicians who do data science??

Mathematics

Computer Science

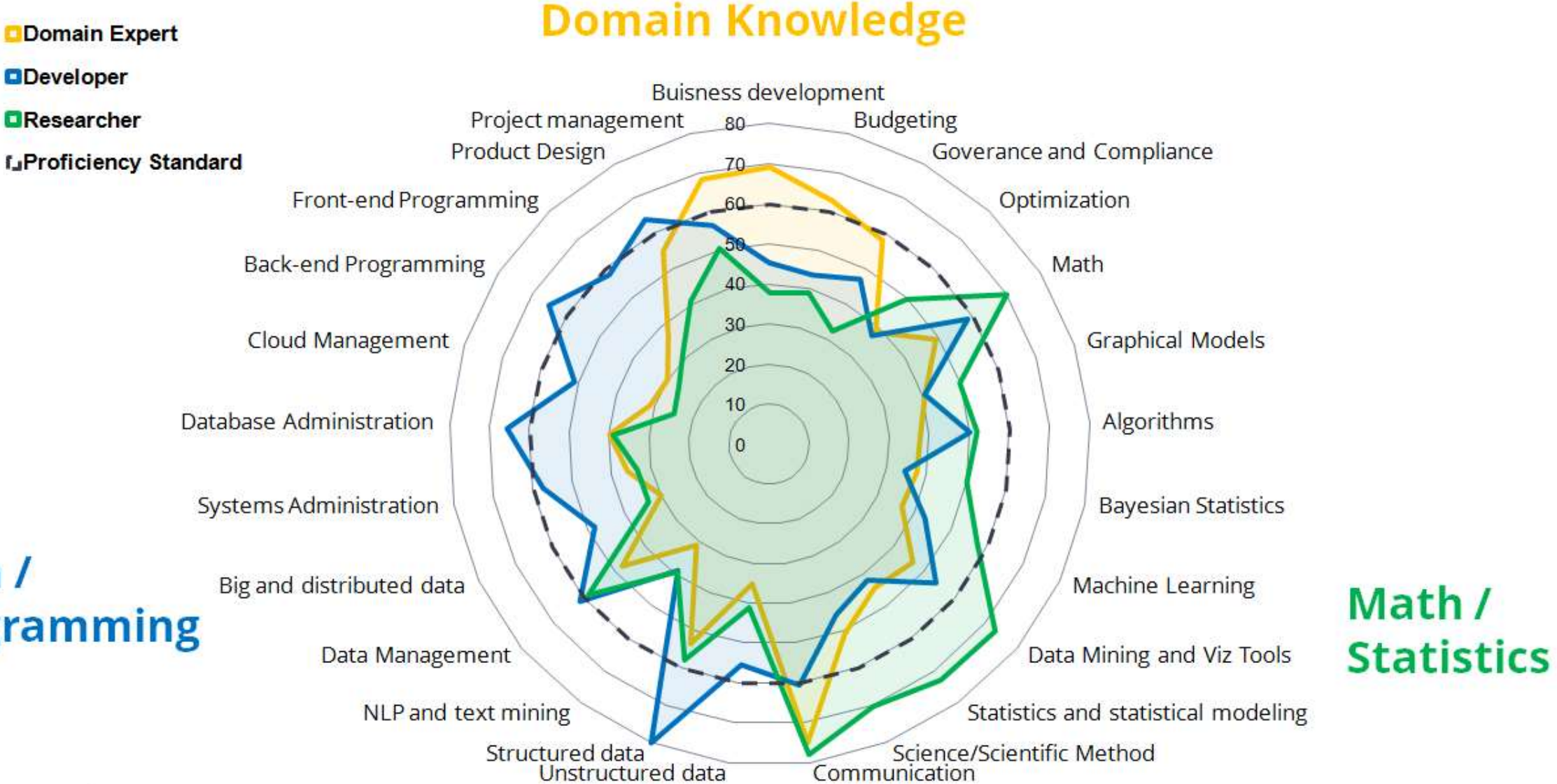
Business

Contextually oriented managers who puts results into practice

Conclusion what is data science?

- Data science is not technical
- Contextual skills are critical
- Data science fosters diversity

Skill Proficiency Varies by Data Science Role

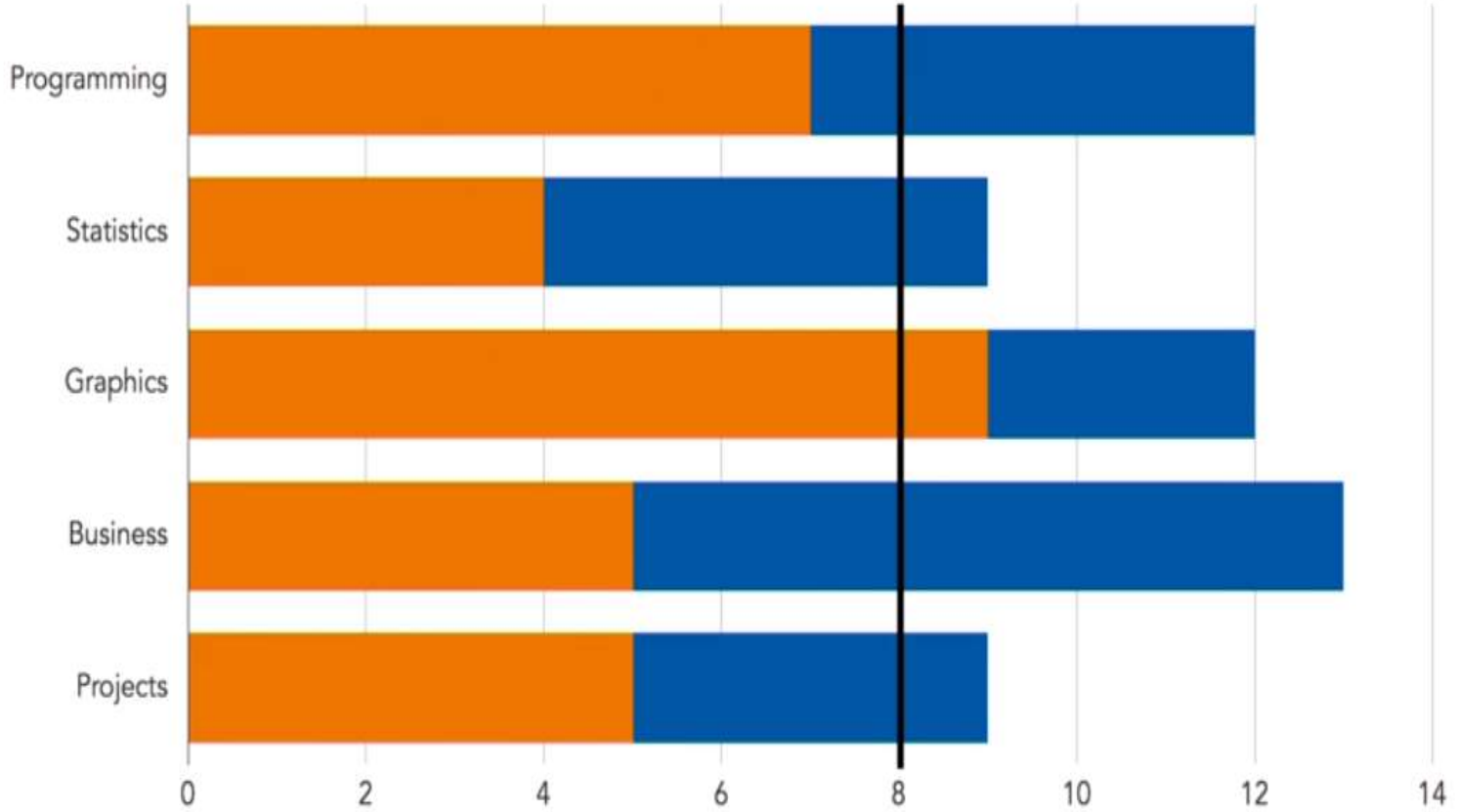


Data are based on responses to AnalyticsWeek and Business Over Broadway Data Science Survey. From September 2015.

Diverse roles, different goals & skills, different contexts

- Engineer/ developer : hardware software
- (data engineer , database administrator)
- Big data specialist (cs+math =>machine learning)
- Researcher (+statistics expertise)
- Analyst (day-to-day web analytics, sql, visualizations)
- Executive business :manage project o implement solutions
- Entrepreneur data based startups,planning , solutions
- Full stack data scientist ALL EXPERTISE (Unicorn)

Combined Skills



What is Big Data?

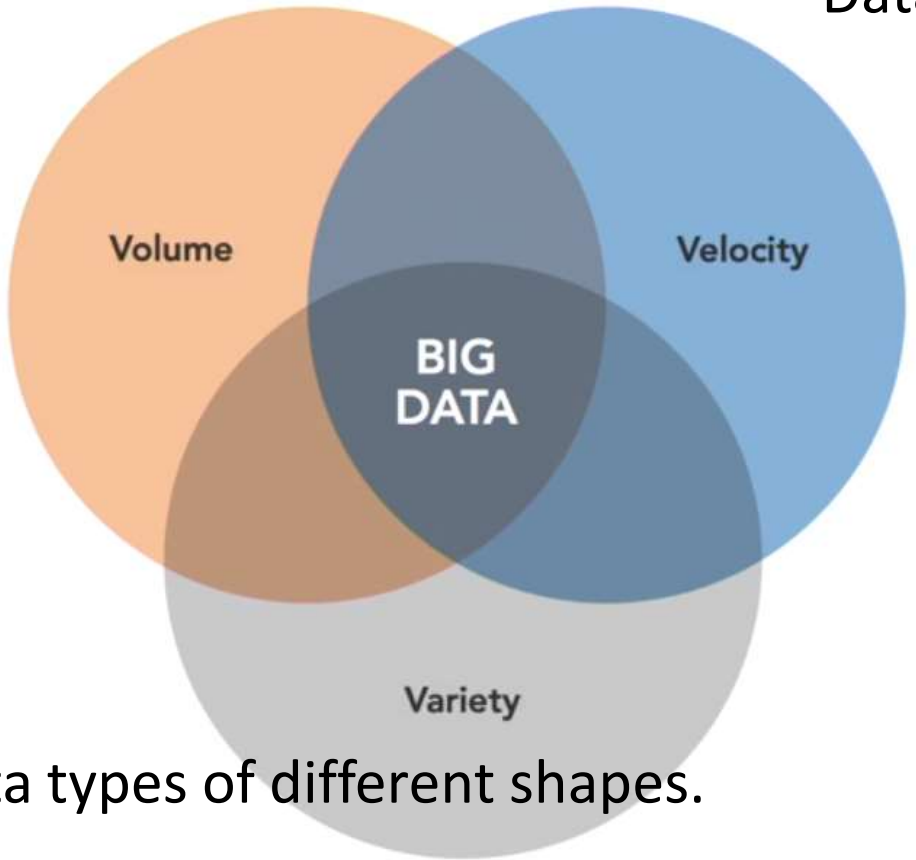
- The are many examples of "data", but what makes some of it “big”? The classic definition revolves around the **three Vs**.
- **Volume, velocity, and variety.**
  - **Volume:** There is a just a lot of it being generated all the time. Things get interesting and “big”, when you can’t fit it all on one computer anymore. Why? There are many ideas here such as MapReduce, Hadoop, etc. that all revolve around being able to process data that goes from Terabytes, to Petabytes, to Exabytes.
  - **Velocity:** Data is being generated very quickly. Can you even store? If not, then what do you get rid of and what do you keep? it all
  - **Variety:** The data types you mention all take different shapes. What does it mean to store them so that you can play with or compare them?

http://pl.wikipedia.org/wiki/Green_Giant#mediaviewer/Plik:Jolly_green_giant.jpg

3 V's

Use MapReduce,
Hadoop,

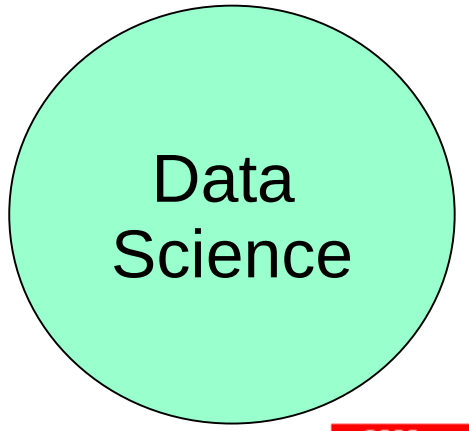
Data is being generated very quickly



data types of different shapes.

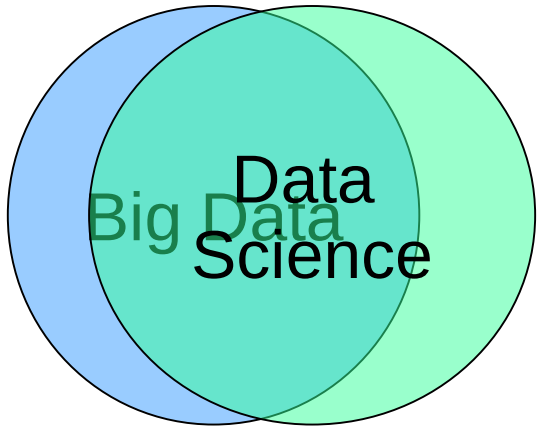
Is Big Data the same as Data Science?

- Are Big Data and Data Science the same thing?
  - Data Science can be done on small data sets.
  - And not everything done using Big Data would necessarily be called Data Science.



Is Big Data the same as Data Science?

- Are Big Data and Data Science the same thing?
  - Data Science can be done on small data sets.
  - And not everything done using Big Data would necessarily be called Data Science.
  - But there certainly is a substantial overlap!



Differ but share same goals and techniques

Big data without data science

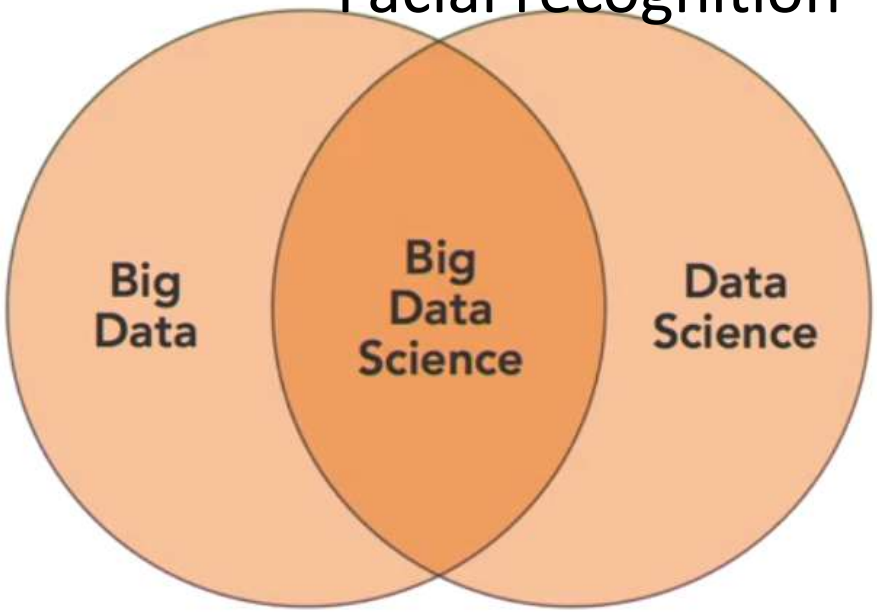
➔machine learning & words count (+volume or velocity –Math's skills)

Data science without Big data

Genetic data sets (huge data sets but consistent

Streaming sensor data (large +structured)

Facial recognition (variety +small amount phot



Big data science

Volume +velocity +variety

➔Unicorn needed of course

Programming for data (word counts)

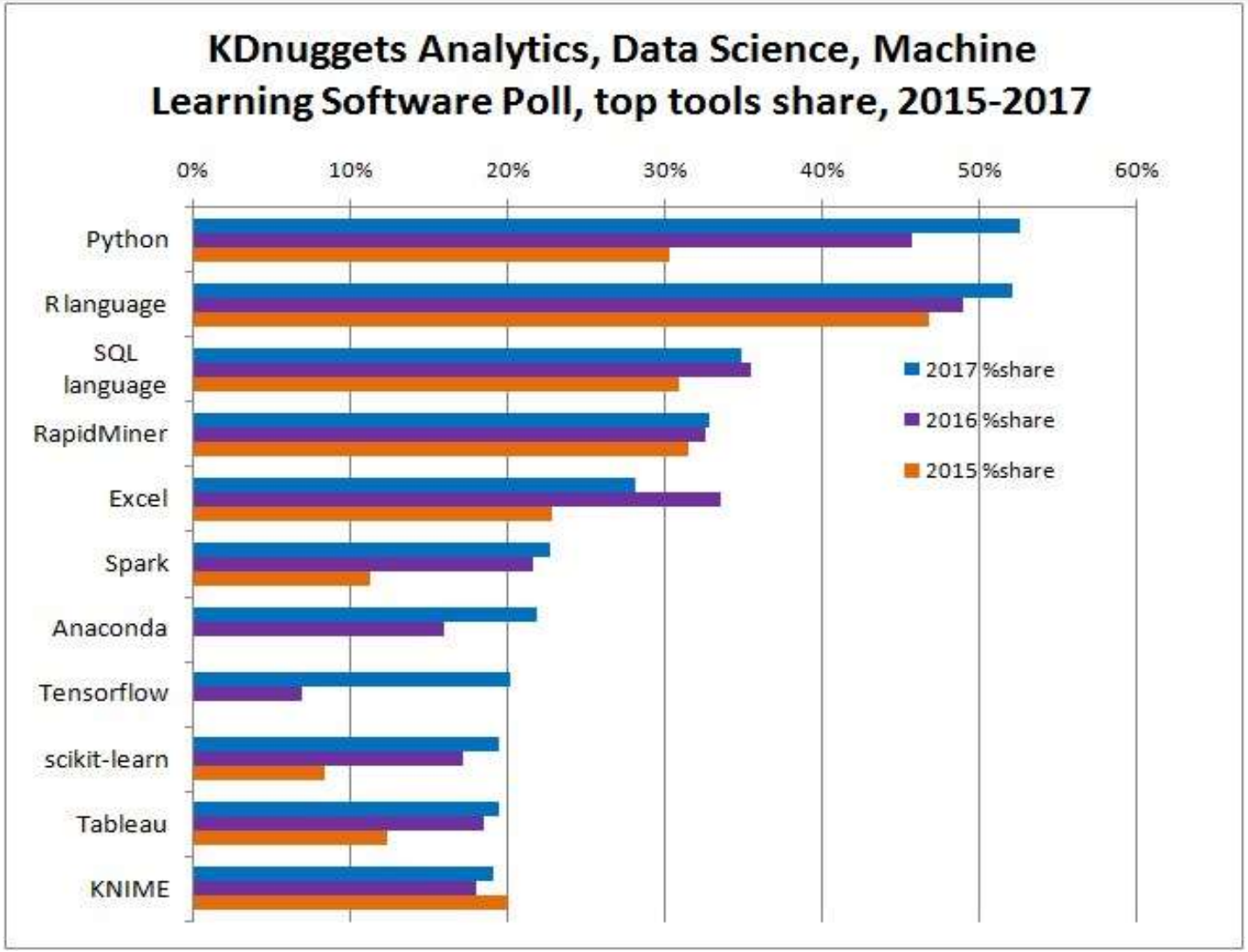
```
class MAPPER
    method Map(docid a, doc d)
        for all term t ∈ doc d do
            EMIT (term t, count 1)
```

Frequency of each set of words?

But data with uncertainty and variability needs Data Science &stats

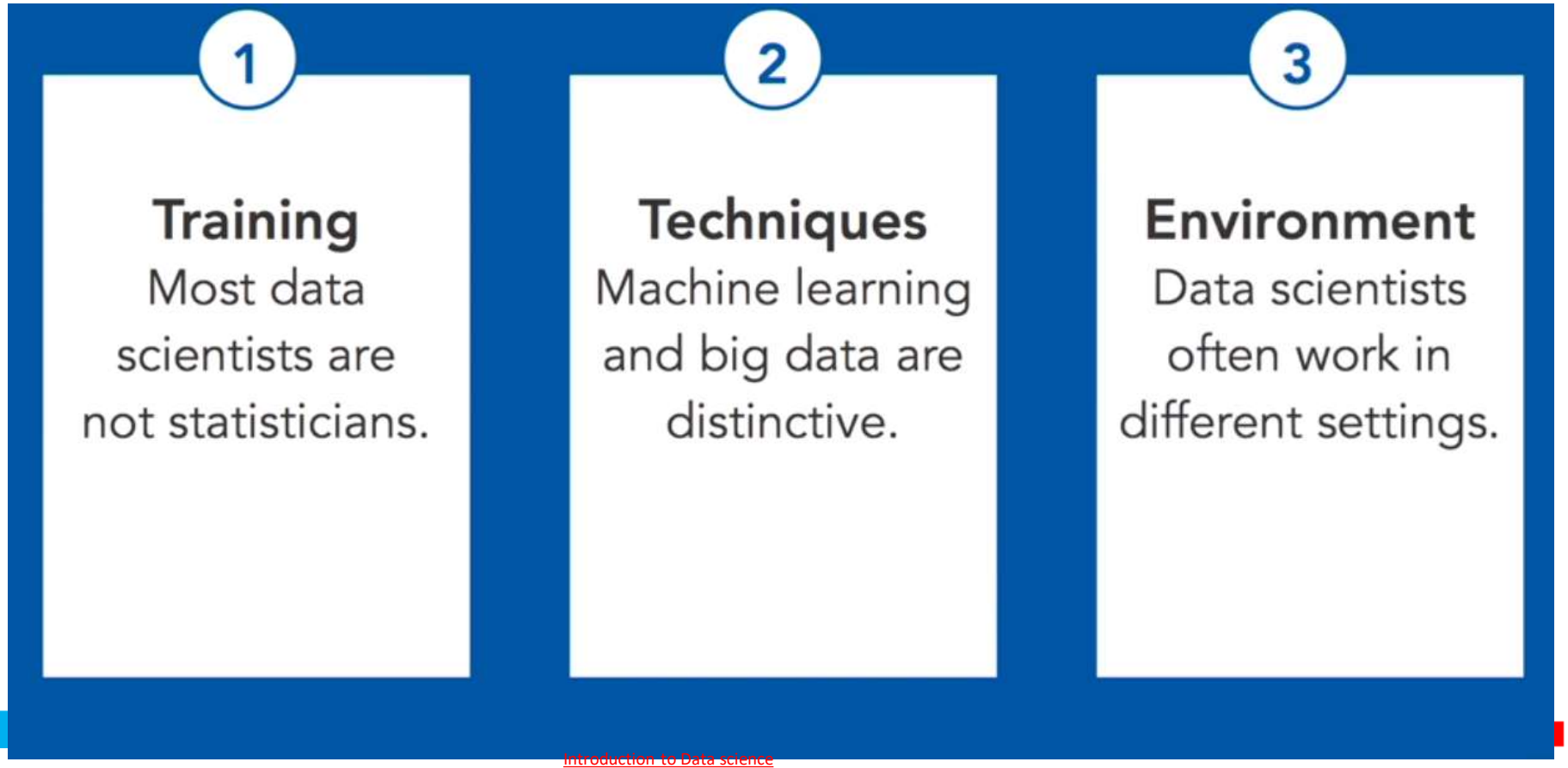
Programming tools

Tools for data science



Data science vs statistics

- Data science requires statistics but is not a subset
- Both fields use data but have different motivation and goals; different background and different context



Measure !!

- Measurement boosts awareness
- Awareness contribute to quality
- Measure thoughtfully and sensitively

Data science metrics

- Key performance Indicators KPIs
- (nonfinancial, timely, simple, significant impact)
- SMART goals (specific, measurable, assignable, realistic, and time-bound)
- Classification accuracy (sensitivity (event exists?),specifity (avoid false negative),positive predictive value, negative predictive value

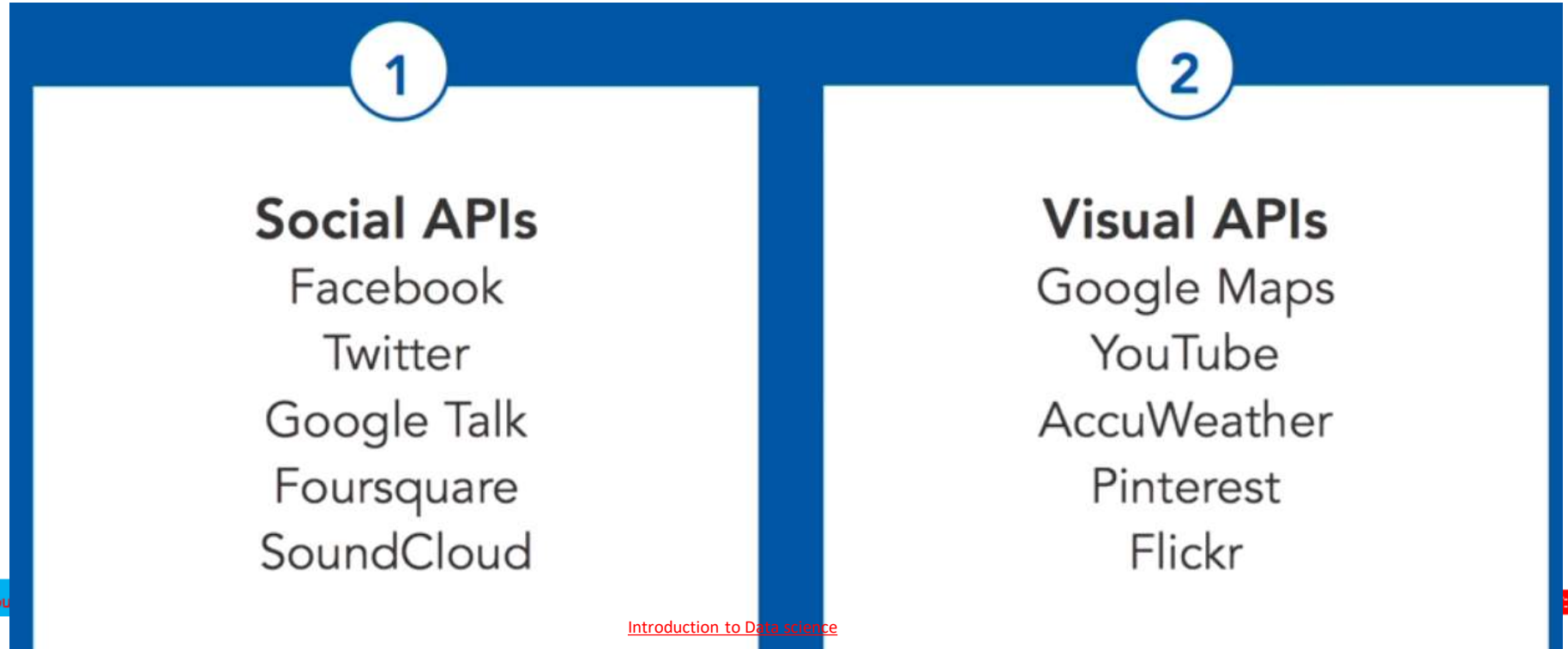
	Event present	Event absent	Total
Test positive	True positives	False positives	Total positives
Test negative	False negatives	True negatives	Total negatives
	Total present	Total absent	Total

Existing data sets: care while interpreting data

- In house data
- Quick easy and free , formatted, identifiers might be available
- Cons: non existent data , inadequate documentation, biases answers
- Open data: enormous data, formatted, documented, biases or unclear data,
- Third party data :Data as a service DaaS,data brokers, processed data , individual level data, can be very expensive , requires validation

Gather data (API)

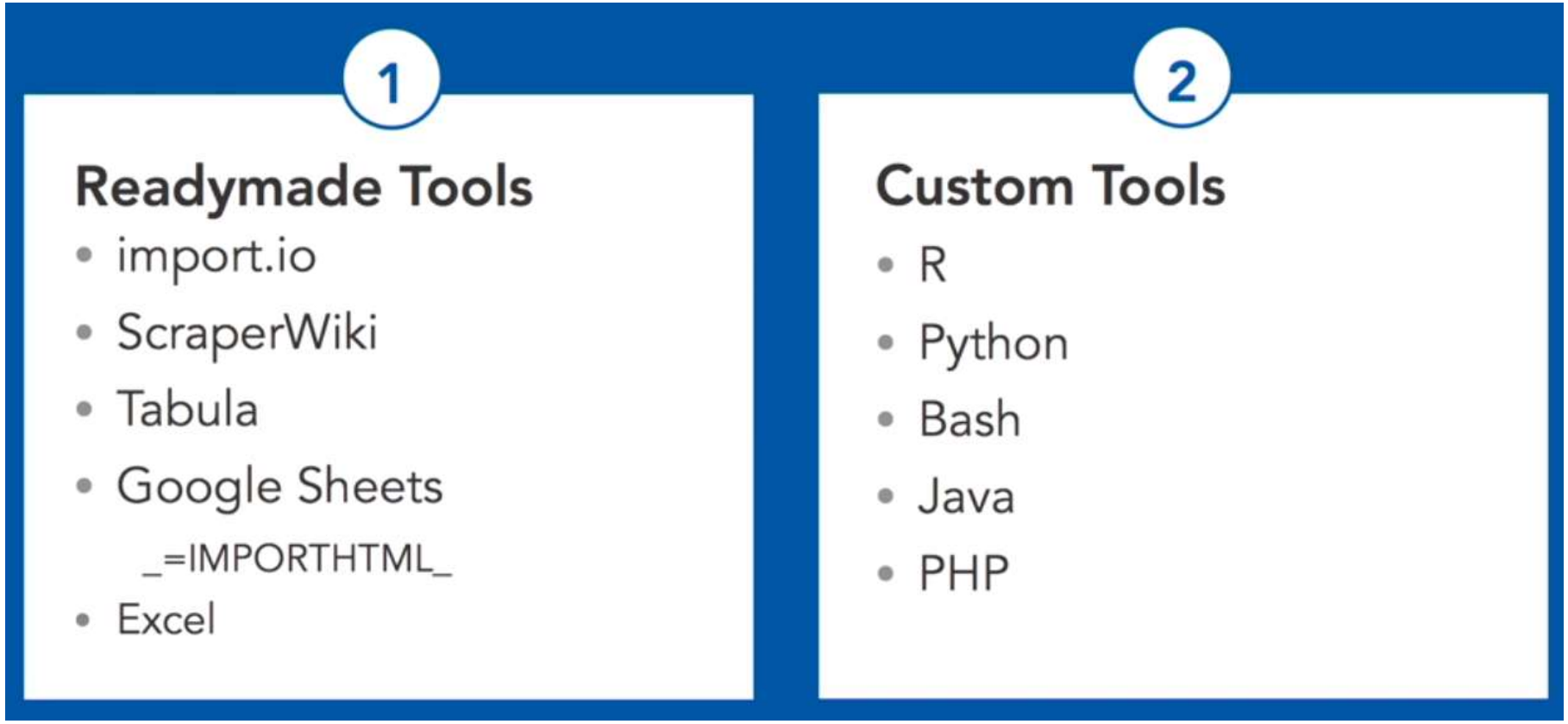
- APIs:Application programing interface
- Rest API: Representation state Transfer API(software architecture style of Wo web):send via HTTp ; send directly from web to program(javascript,JSON); lar agnostic



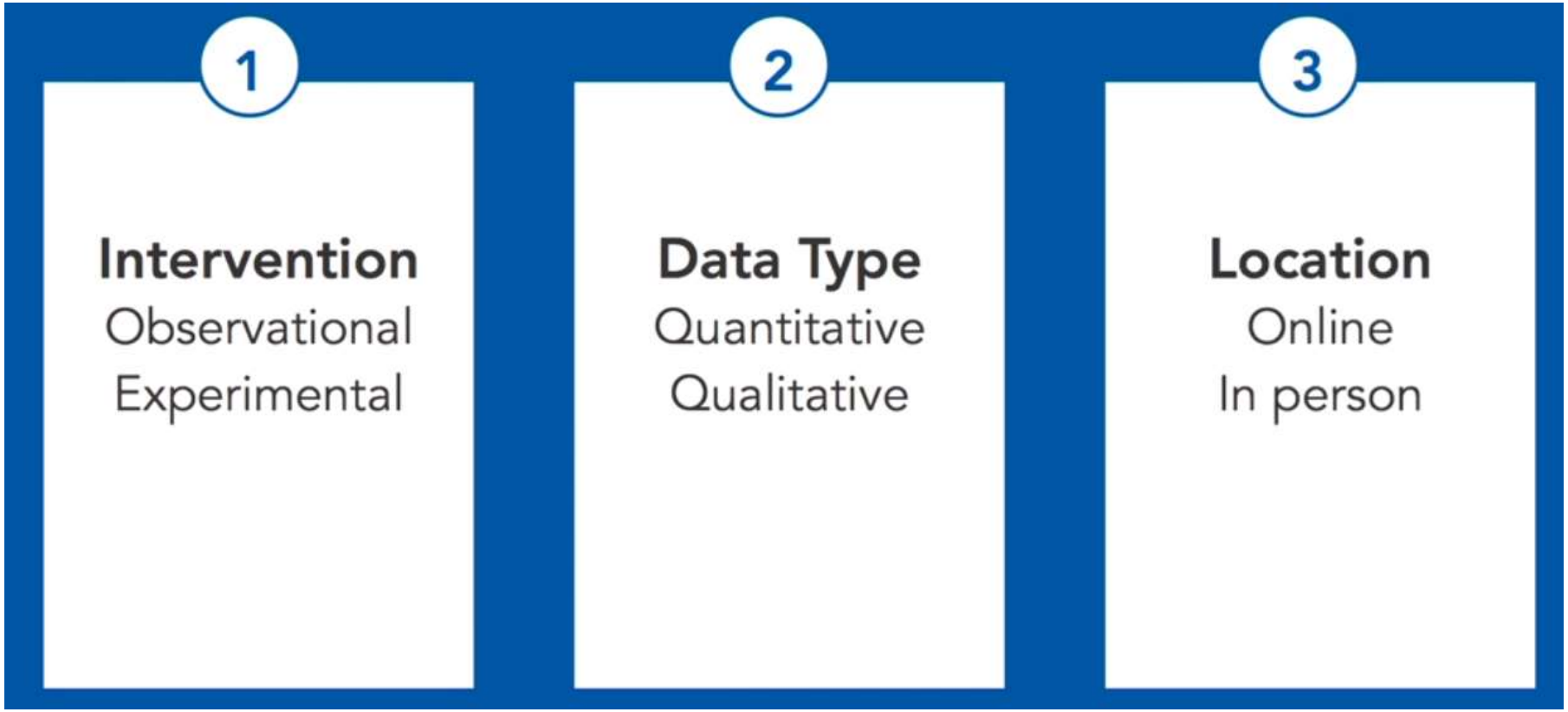
Access data from the web and feed it into Python

```
#Import the modules
import requests
import json
# Get the feed
r =requests.get("http://gdata.youtube.com/feeds/api/standardfeeds/top_rated?v=2&alt=jsonc")
r.text
# Convert it to a Python dictionary
data = json.loads(r.text)
# Loop through the result.
for item in data['data']['items']:
    print "Video Title: %s" % (item['title'])
    print "Video Category: %s" % (item['category'])
    print "Video ID: %s" % (item['id'])
    print "Video Rating: %f" % (item['rating'])
    print "Embed URL: %s" % (item['player']['default'])
```

Gather data (scraping data when API does not exist)



Creating data



Interviews (time consuming), surveys (easy), card sorting, experiments,

Training

Exploratory graphs

- Review data
- Check assumptions
- Check anomalies
- Data suggestion
- Quickly check shape, gap, outliers
- ==========
- Bar chart for categories
- Box plots for quantitative variable
- Histograms (overlay shapes for comparison)
- Scatter plot matrices

Exploration: a critical first step in analysis

- Single distributions
- Joint distributions (associations)
- Unusual cases
- Error in the data
- Missing data

Exploration using

- Coding (R, python, JavaScript)
- Applications (tableau, Qlik , excel)
- By hand (as John Tukey: Exploratory Data Analysis)

Exploratory statistics

- Robust statistics:(stable, less affected by outliers, many choices: median, avg..; not easy)
- Resample data: empirical estimate of sampling variability
- (jackknife:sample without replacement; bootstrap samples with replacement;permutation:shuffle across different group,cross-validation:test/training)
- Transform data (functions while preserving the order ..such as Tukey's ladder of power used when we expect outliers

Programing

- Excel flexible
- R language (7000 packages CRAN)
- Python (Jupyter Interface)
- SQL (RDBS)
- Web formats (html,xml,Json (javascript object notation),javascript (d3.js library that generates dynamic visualization that can be shared on the web)

What do you need to be ready?

- Anaconda - Scientific Python package with IPython Notebook.
  https://www.continuum.io/downloads