Data Science #1 - Scraping and collecting data

Tal Ne'eman
Developer Advocate
IBM Alpha Zone

talne@il.ibm.com

IBM **Developer**

IBM

Hi, I’m Tal

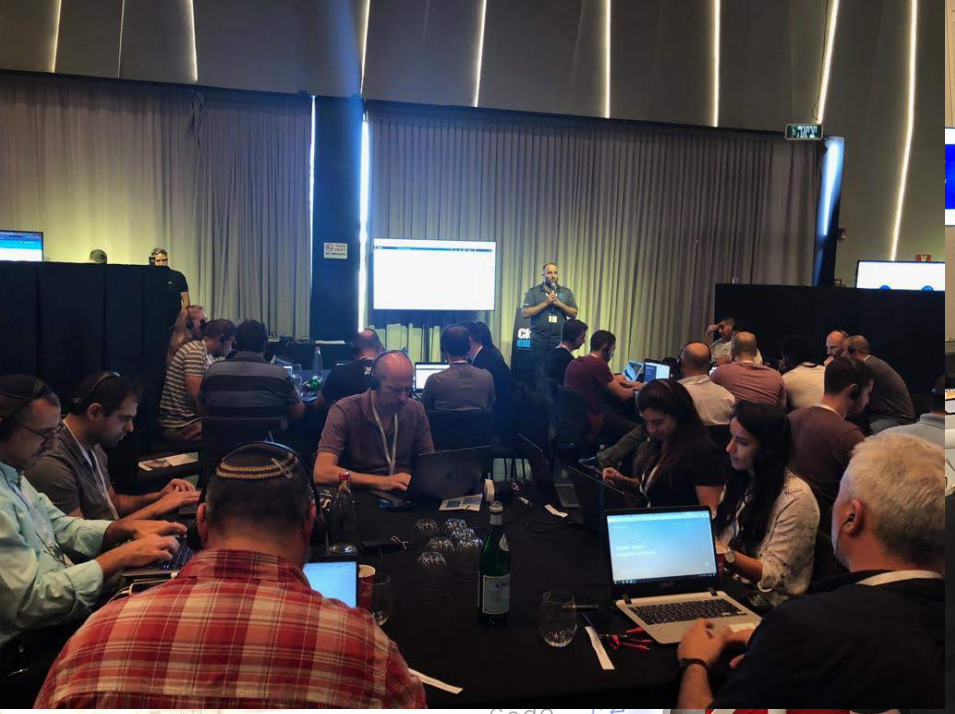
I’m a developer advocate for IBM.

I participate in meetups, hackathons, webinars and write articles about technology for IBM and other organizations.

IBM Startup & Developer - Tel Aviv

http://ibm.biz/ibmtlvmeetup

Warning: I am a lowly developer



developer.ibm.com

2020 CALL FOR CODE®

Global Challenge

COVID-19 has revealed limits of the systems we take for granted, compromising our health, our planet, and our survival.

We've expanded the 2020 Call for Code Global Challenge to take on COVID-19, while we remain steadfastly committed to combating climate change.

Take on the most pressing issues of our time

Build solutions that fight back.

- Build & deploy solutions to **help seize & reduce** the impact of **COVID-19**
- Build & deploy solutions to help **halt & reverse** the impact of **Climate Change**

In 2020, Call for Code is aligned to the UN 75th anniversary global conversation theme of **climate change**, with a focus on:

- **water sustainability**
- **energy sustainability**
- **disaster resiliency**

The 2020 Call for Code Global Challenge is being expanded to address the **COVID-19 pandemic**, with a focus on:

- **crisis communication**
- **remote education**
- **community cooperation**

Join a **movement** of:

- **210,000+** problem solvers
- **165+** nations
- **8,000+** applications built

Have the chance to win:

- **$200,000** USD
- Open Source support from **The Linux Foundation**
- Meetings with **mentors** & potential **investors**
- **Implementation** support through **Code and Response™**

Call for Code Founding Partner

Call for Code Creator

Call for Code Charitable Partner

Call for Code Affiliate

Challenge Timeline

- Climate change
- COVID-19

February 26
Challenge kickoff

March 20
COVID-19 track added

April 27
Deadline for initial COVID-19 submissions

May 5
Top 3 COVID-19 solutions announced at IBM Think

May 15
Initial deployment support starts for top 3 COVID-19 solutions

July 31
Final submission deadline for both tracks

October
Winners of the 2020 Call for Code Global Challenge announced at award ceremony

Start coding

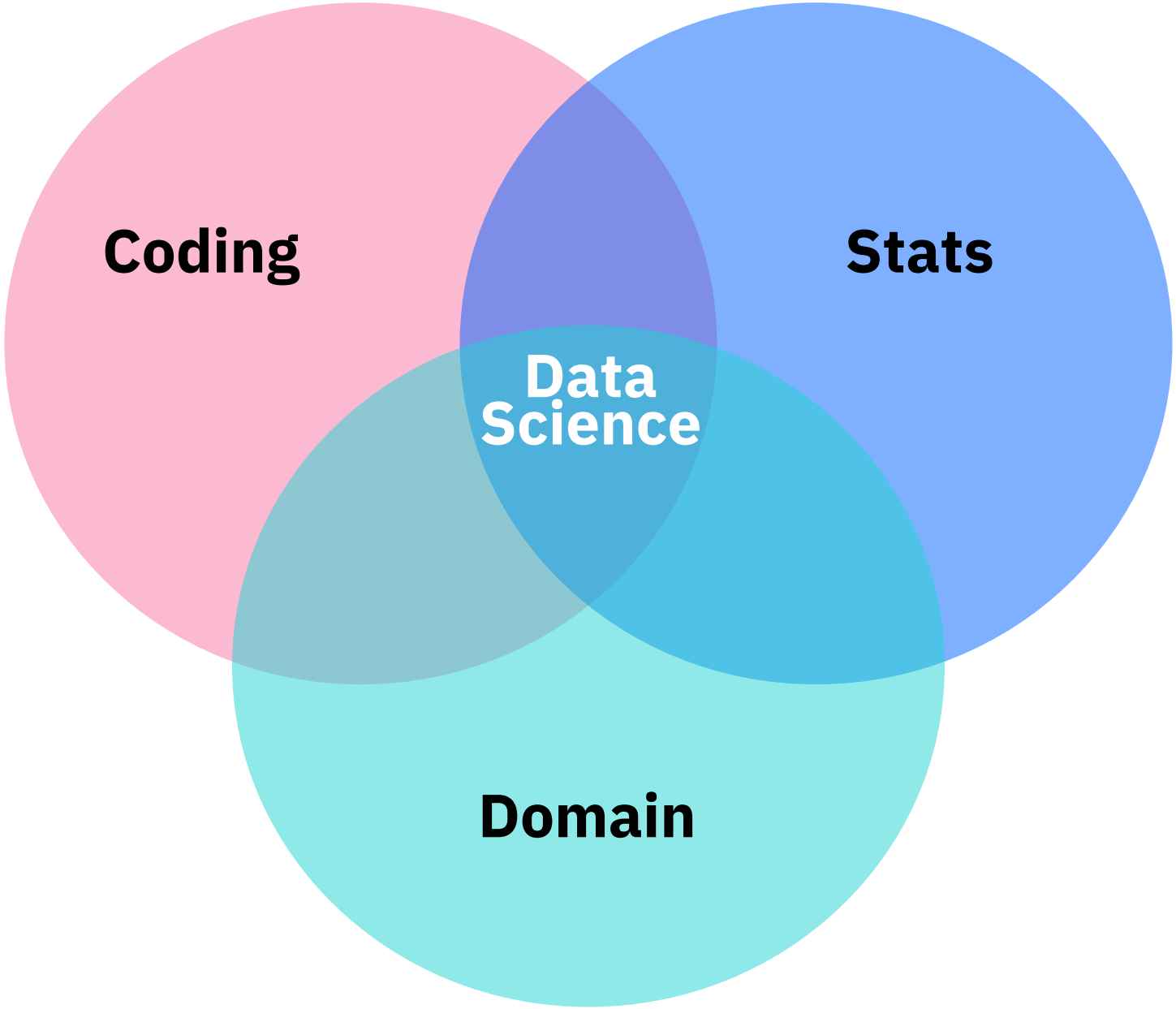
Build with open tech

Find your squad

Submit your idea

http://ibm.biz/Bdqe65

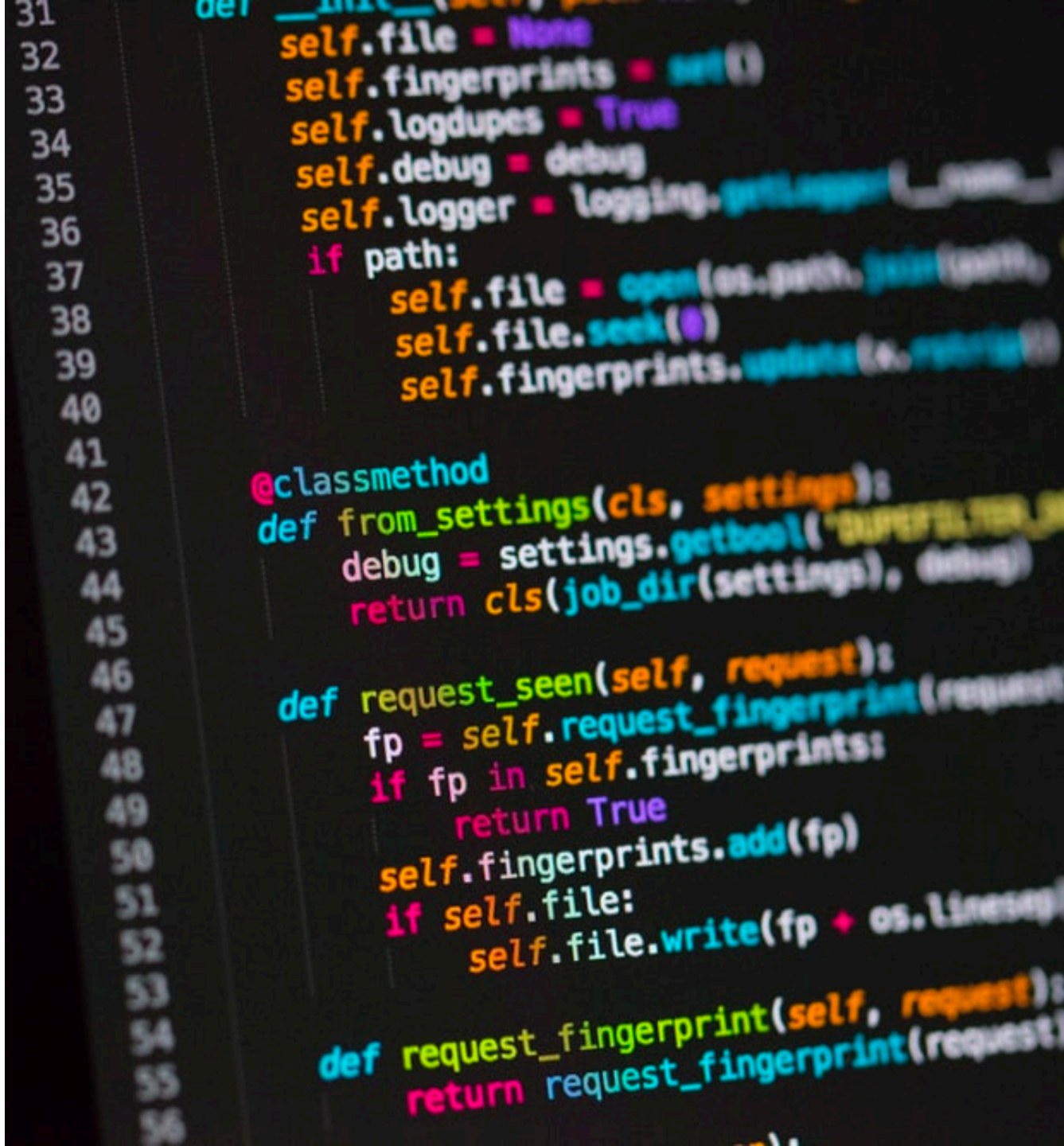
Drew Conway's data science Venn diagram



http://j.mp/ds-venn

Coding :

- Gather, prepare data
- Different types of sources
- Unusual formats
- Requires creativity



Stats

R & Python

Database

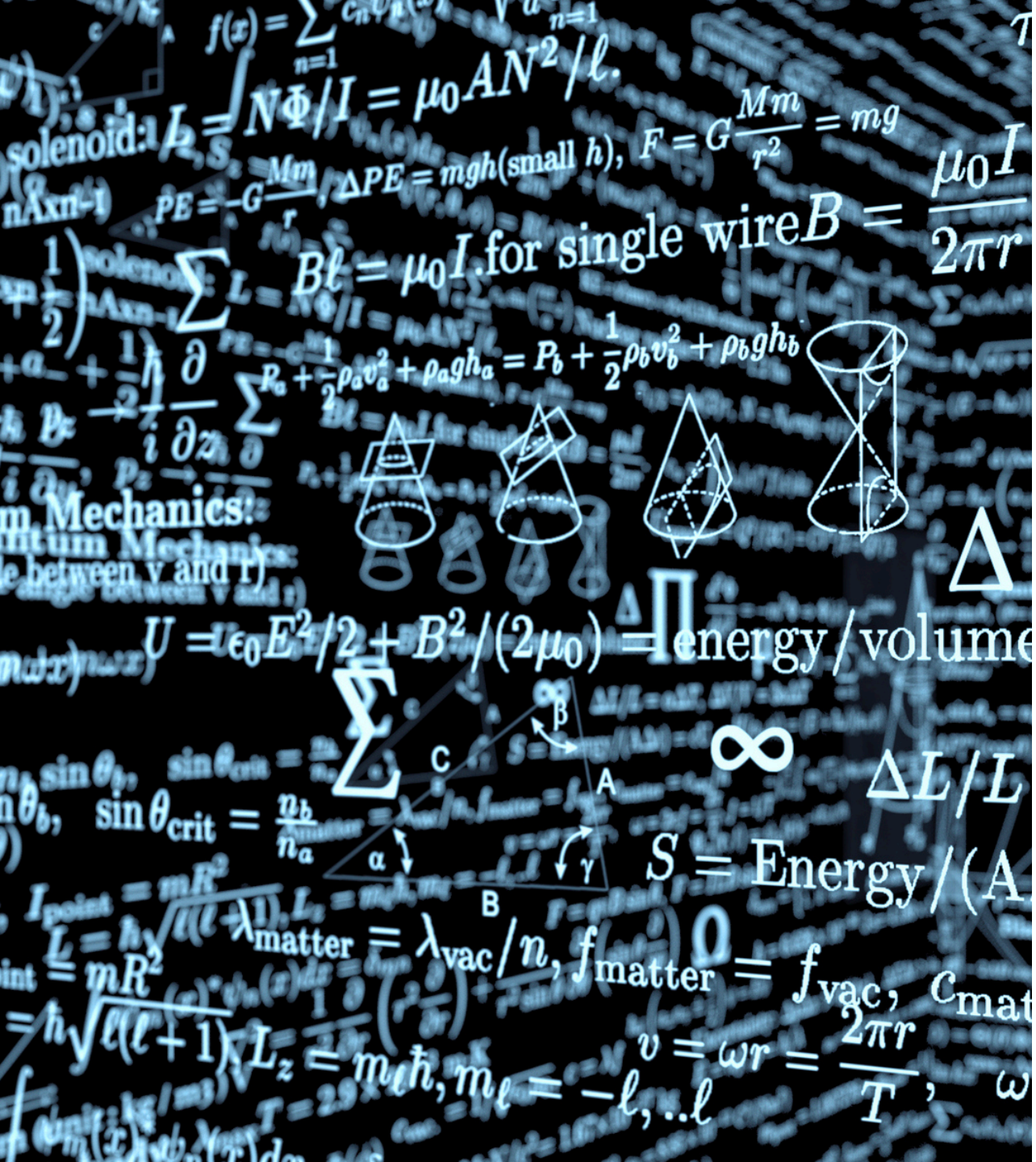
SQL

Command line

Bash

Search

Regex



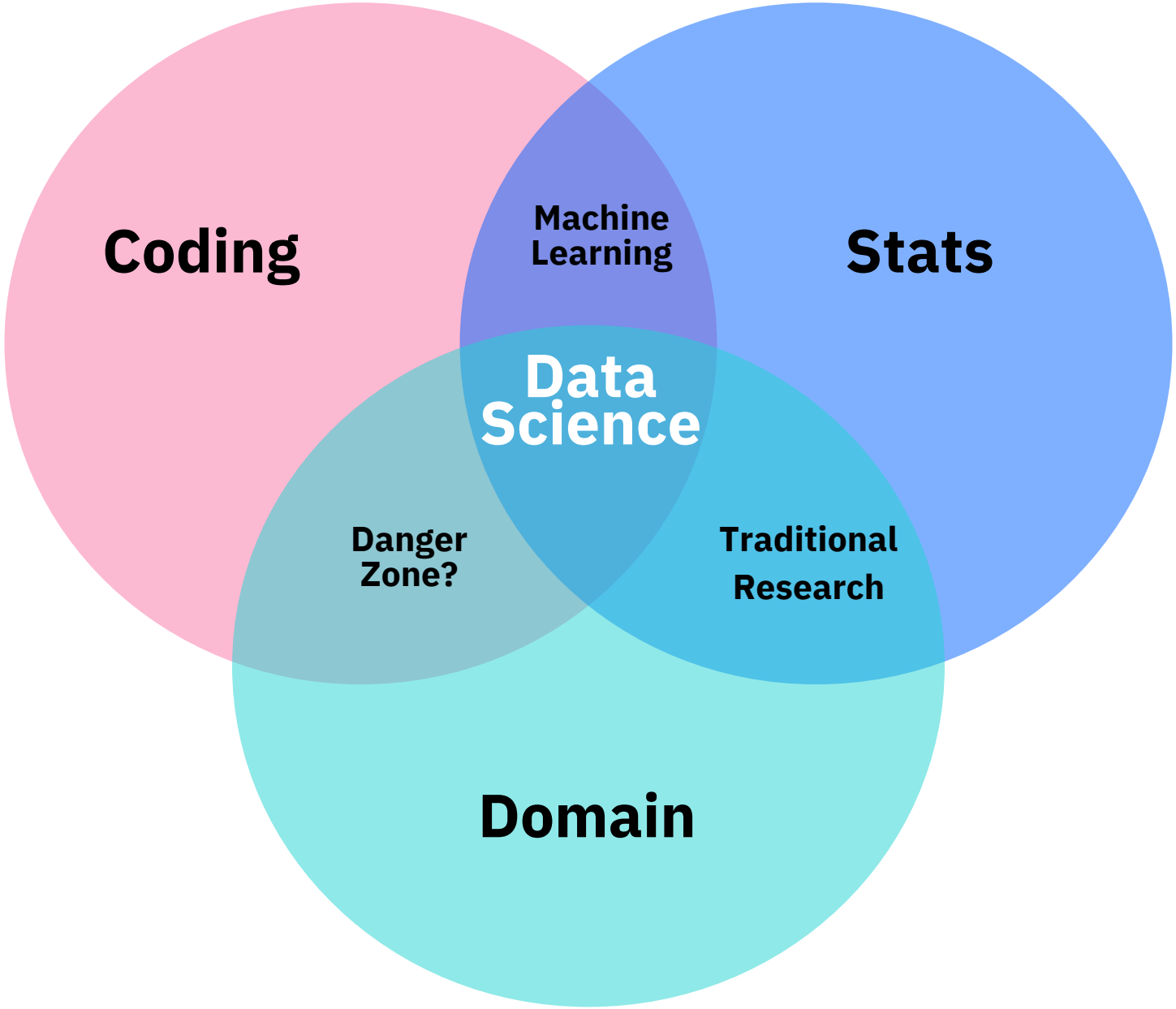
Math & stats :

- Probability, algebra, regression, etc
- Choose procedures
- Diagnose problems

Domain :

- Expertise in field
- Goals, methods & constrains
- Can implement well

Drew Conway's data science Venn diagram



http://j.mp/ds-venn

ML :

- Machine Learning
- Coding & math

  without domain
- “Black Box” models

Research :

- Math & domain without coding
- Data is structured
- Effort is in method & interpretation

Danger Zone :

- Coding & domain without math
- Unlikely to happen
- Words counts, maps

Data scientists types

Code	Stats	Domain
Coders who can do math, stats & business	Statisticians who can code & do business	Business people who can code & do numbers
Most common	Less common	Least common

Steps in Data Science world

First

Planning

Second

Data prep

Third

Modeling

Fourth

Follow up

Planning :

- Define goals
- Organizer resources
- Coordinate people
- Schedule project

Data prep :

- Get data
- Clean data
- Explore data
- Refine data

Modeling :

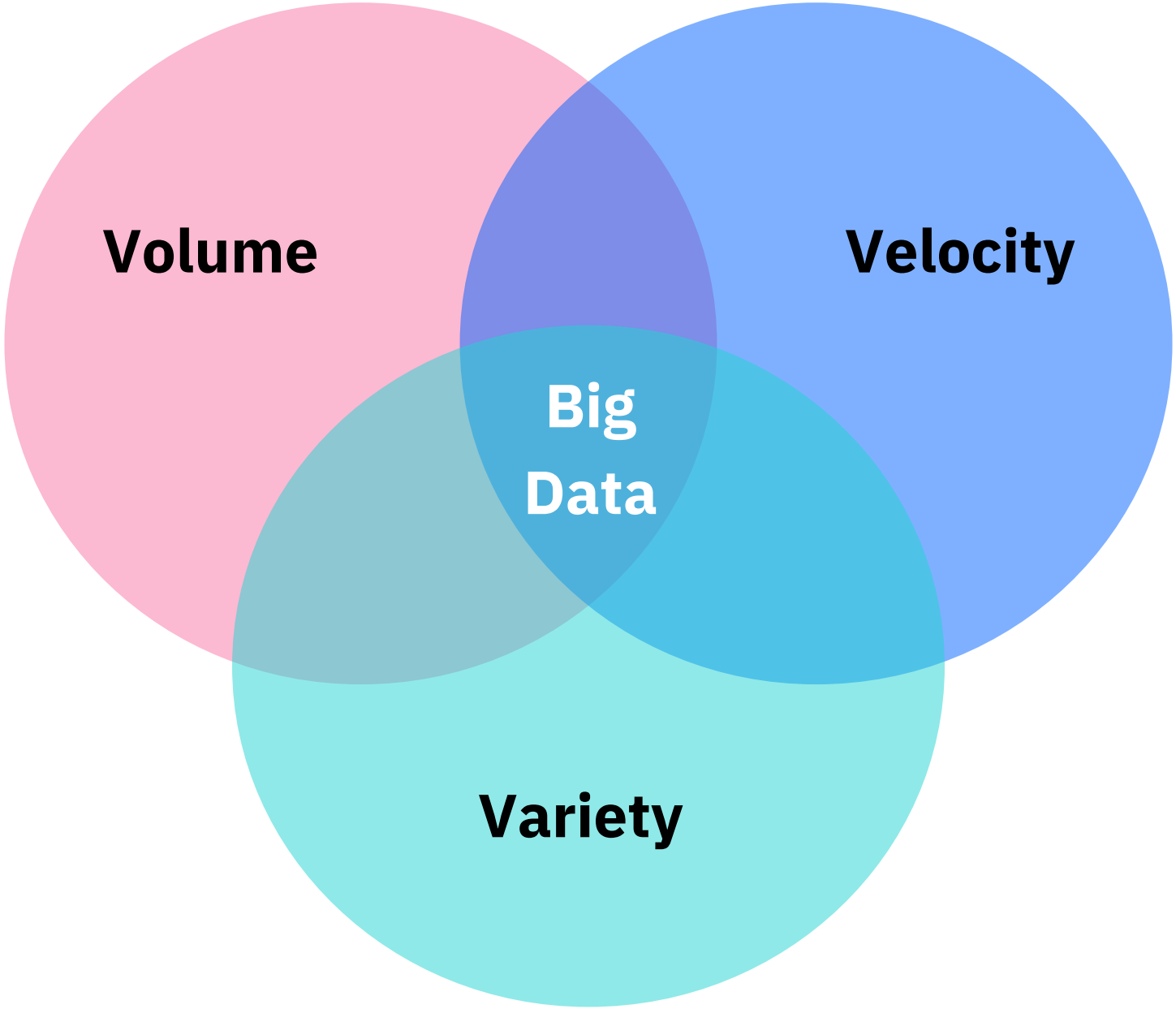
- Create model
- Validate model
- Evaluate model
- Refine model

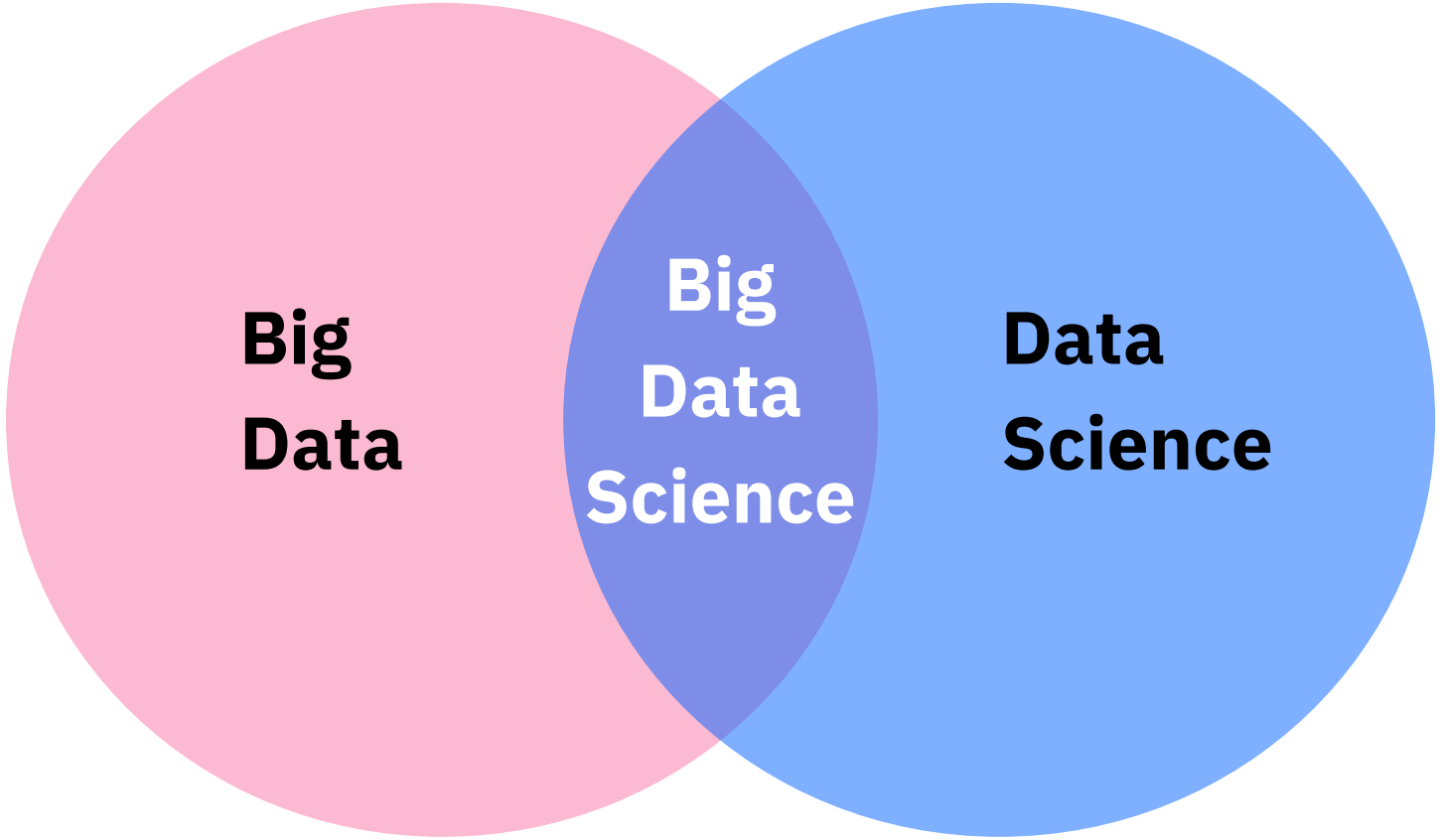
Follow up

- Present model
- Deploy model
- Revisit model
- Archive assets

A word about Big Data

Big Data Venn diagram





Big data != Data Science

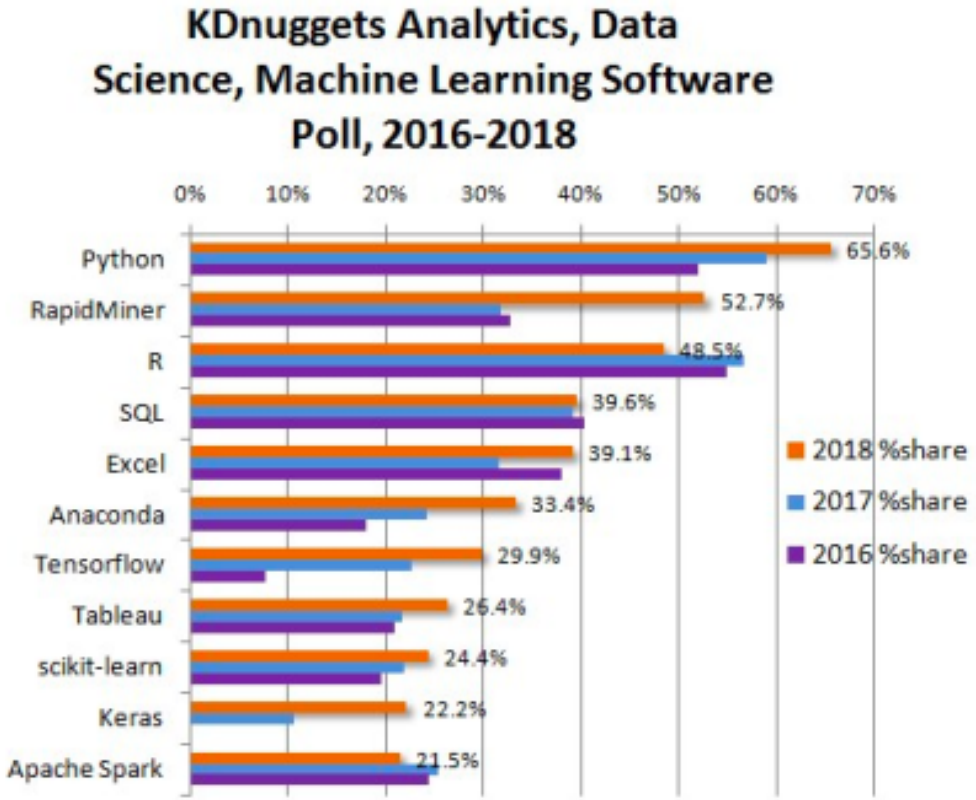
Some common ground

Coding and Data Science tools

Tools for coding

Language Rank	Types	Spectrum Ranking
1. Python		100.0
2. C++		99.7
3. Java		97.5
4. C		96.7
5. C#		89.4
6. PHP		84.9
7. R		82.9
8. JavaScript		82.6
9. Go		76.4
10. Assembly		74.1

Tools for Data science



-Data science != Coding

-Share tools & practices

-But statistics is critical

Methods for Data Science

Sourcing methods

- Existing (company record / open data sets / third party)
- Data APIs
- Scrapping (HTML/PDF using apps & code)
- Make data (GIGO)

Company records

- Potentially quick, easy & free
- Standardized ?
- Original team ?
- Identifiers for specific insights
- Quality ?
- Documentations ?

Open data

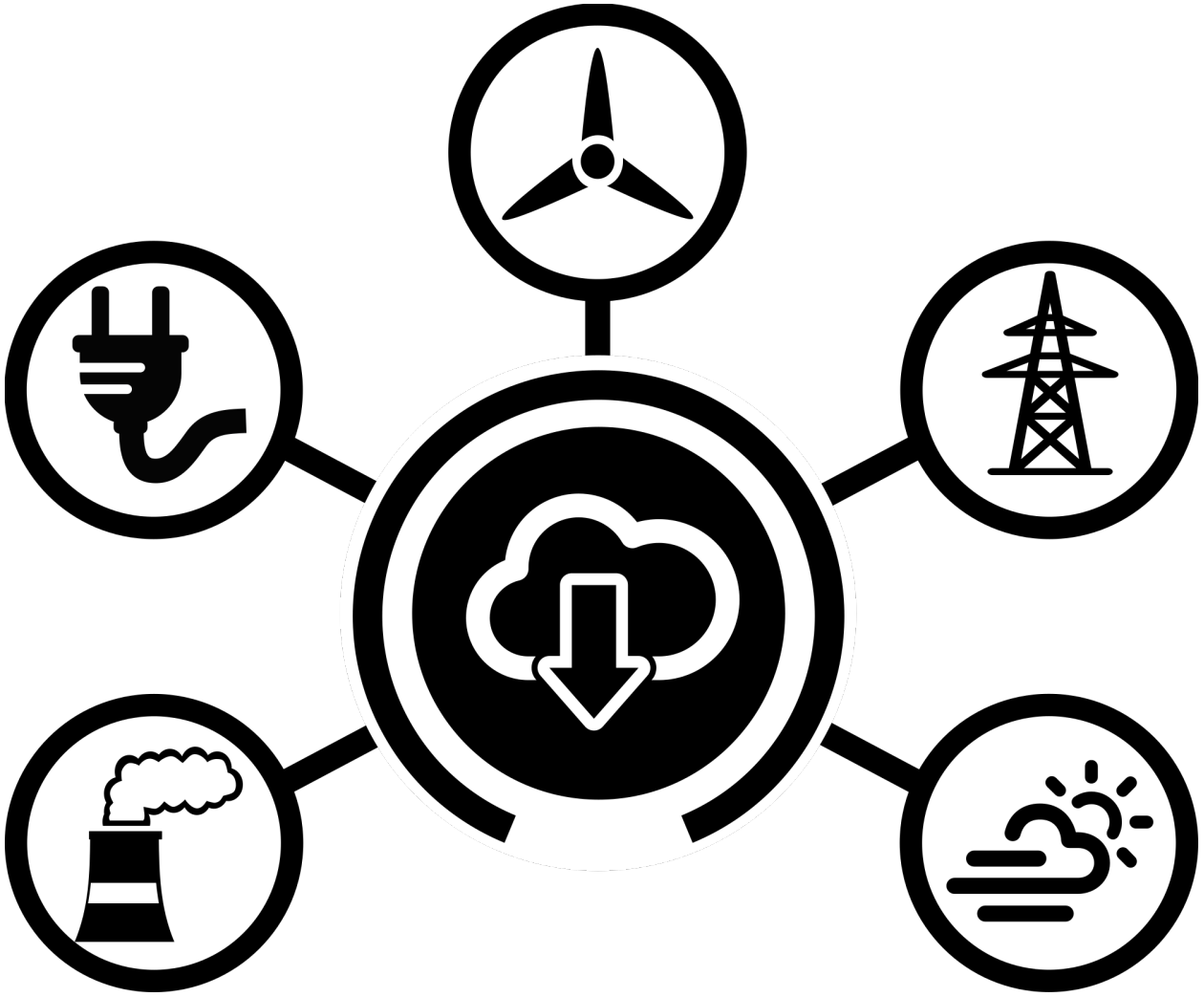
- Valuable
- Range of topics, times, groups, etc.
- Well-formatted & well-documented
- Biased samples?
- Sharing?
- Privacy & confidentiality

Third party

- Save time & effort
- Individual level
- Can get summaries
- Can be expensive
- Still need to validate
- Distasteful to many people

Open data sources :

- Data.gov
- Open-data.Europa.eu
- Uniced.org/statistics
- Who.int/gho
- Developer.nytimes.com
- Google.com/publicdata
- Aws.amazon.com/datasets



Social APIs

- Facebook
- Twitter
- Google Talk
- FourSquare
- SoundCloud

Visual APIs

- Google Maps
- YouTube
- AccuWeather
- Pinterest
- Flickr

Scrapping apps

- Import.io
- ScrapperWiki
- Tabula
- GoogleSheets
- Excel

Scrapping Code

- Use any program language
- HTML (look for tags you need <h1>, <p>, <table>)
- PDF (text elements)
- Media (images / videos)

Make your own data

- Interviews
- Surveys
- Card sorting
- Experiments

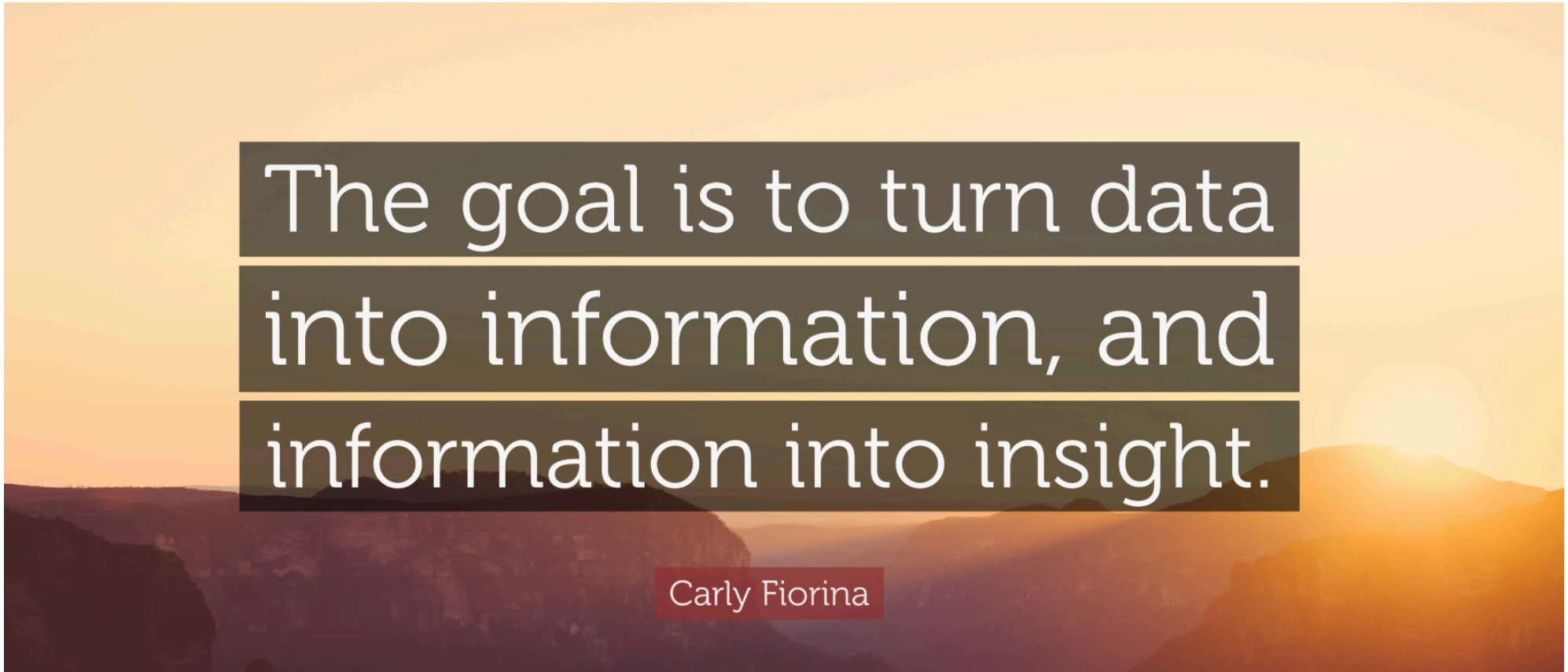
Sourcing sum :

- Get the raw materials
- Many possible methods
- Check quality of data

Remember – No Data, No Data Science

The goal is insight

Tools are just tools ! Remember your goal and choose the tools that will help you to reach it.



Use Open source tools

- Jupyter

- Rstudio

https://cloud.ibm.com/

https://raw.githubusercontent.com/tal2k4xj/Watson-Q-Learning/master/ScrapingBeautifulSoup.ipynb

https://www.dataquest.io/blog/web-scraping-tutorial-python/