Data Science and Big Data Analytics

Solomon Gizaw

Solomon Gizaw, PhD

Affiliations:

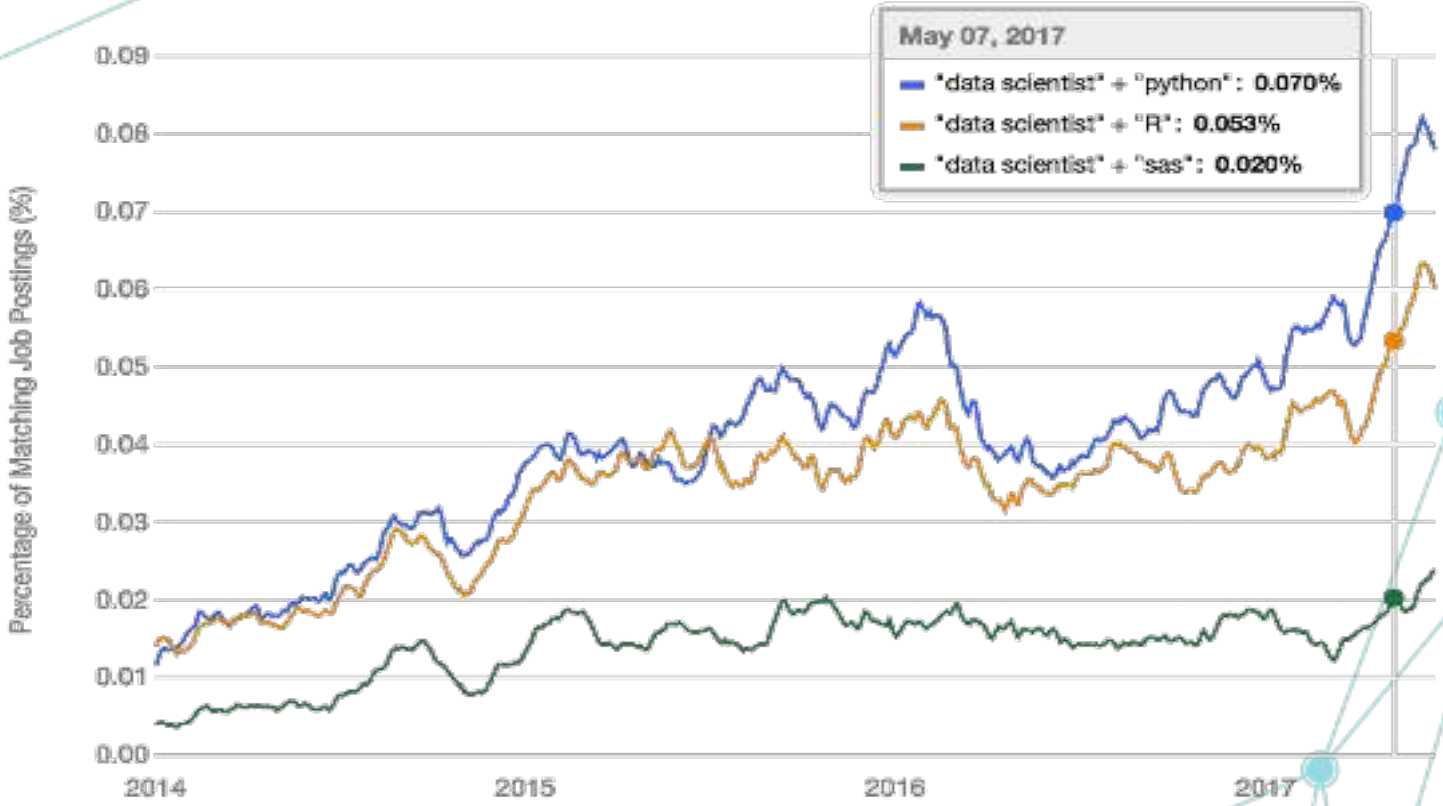
- Addis Ababa University: Computer Science
  - ❑ Assist Proff
- Addis Ababa University Computational Data Science graduate program
  - ❑ Director
- Pan African Information Communication Technology Professional Association (PAICTA)
  - ❑ Advisory Board
- Tiny ML Project
  - ❑ Coordinator, PI

Research Interests:

- Natural Language Processing,
- Localisation,
- Personalisation,
- Artificial Intelligence, Machine Learning

Email: solomong@aau.edu.et, Solomon@paicta.za, solohavi@gmail.com

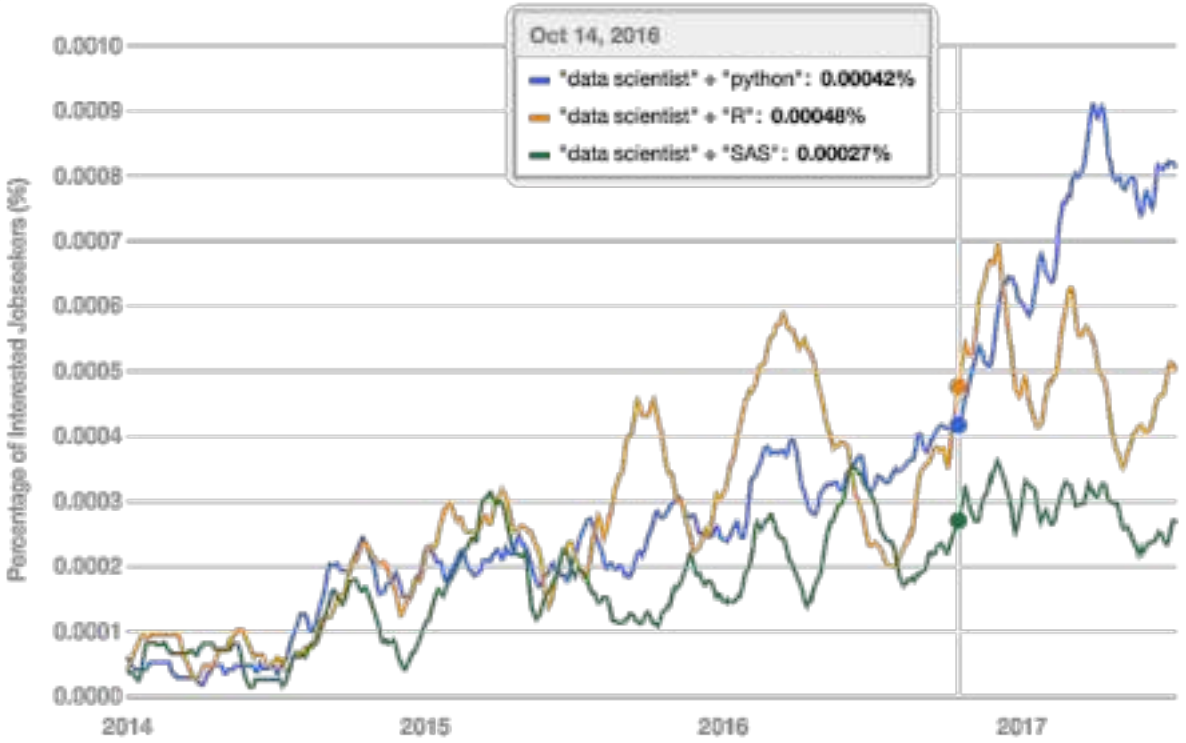
Data Scientist Job Demands



Job Postings

WeCloud DATA

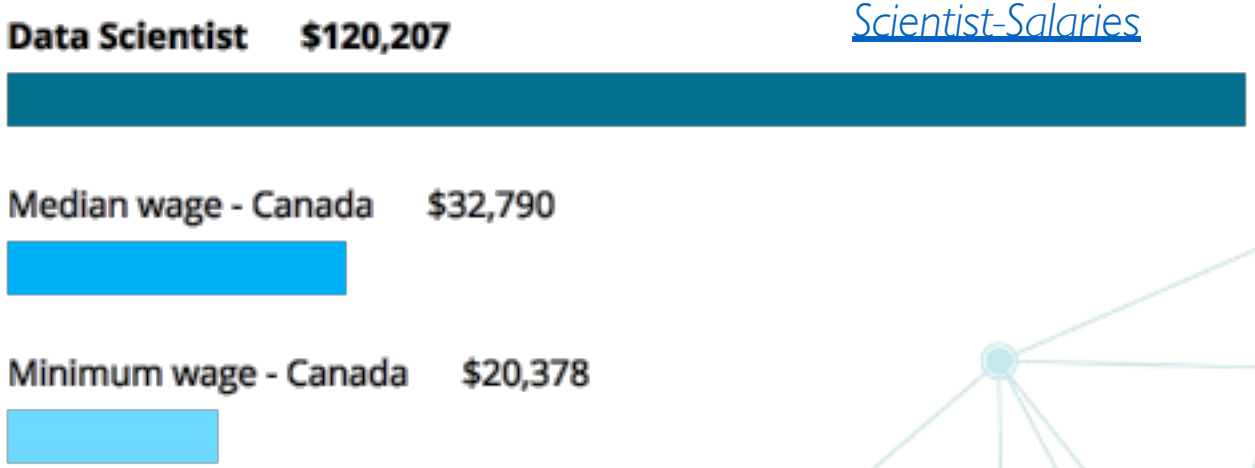
Data Scientist Job - Demand



Job Seeker Interest

Data Scientist Jobs - Salary

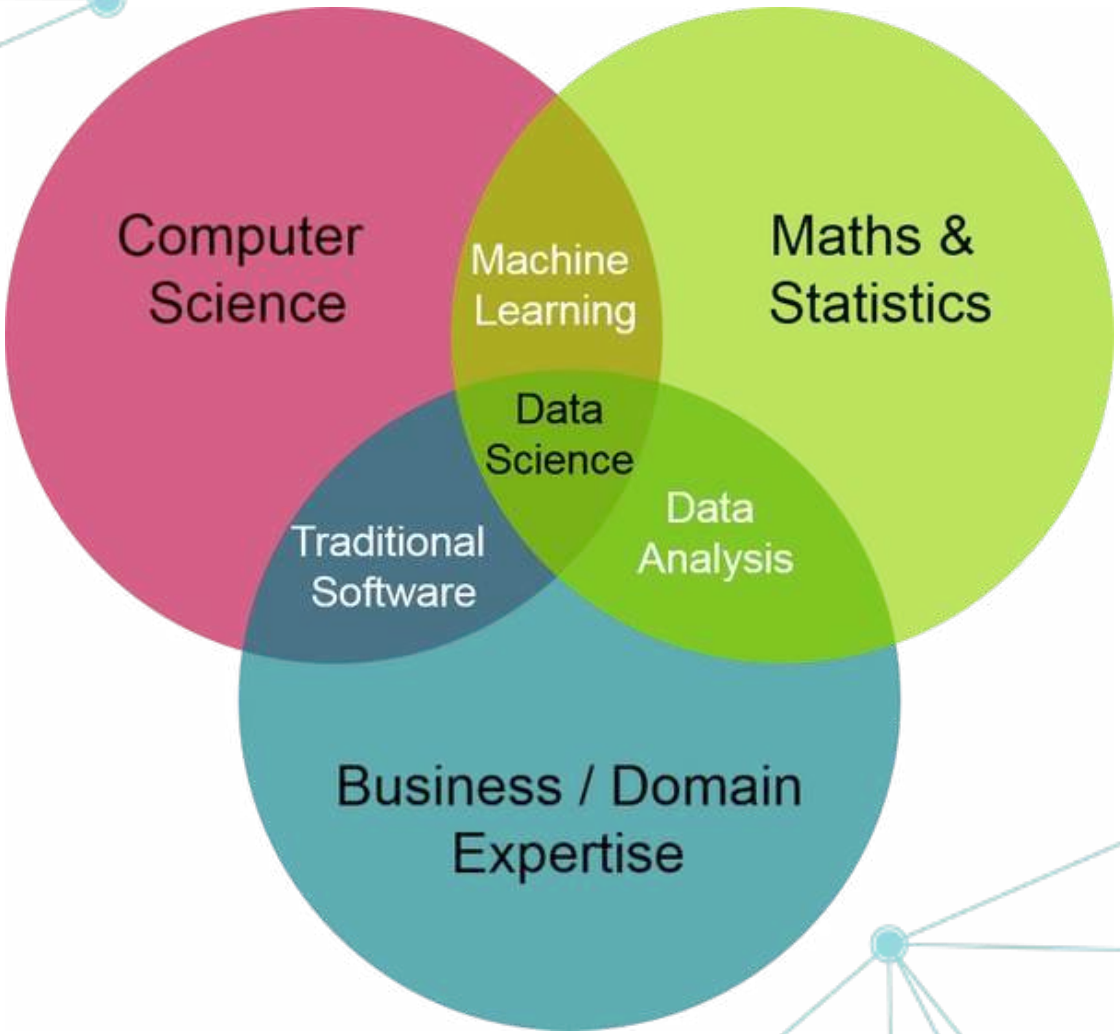
source: https://ca.indeed.com/Salaries/Data-Scientist-Salaries



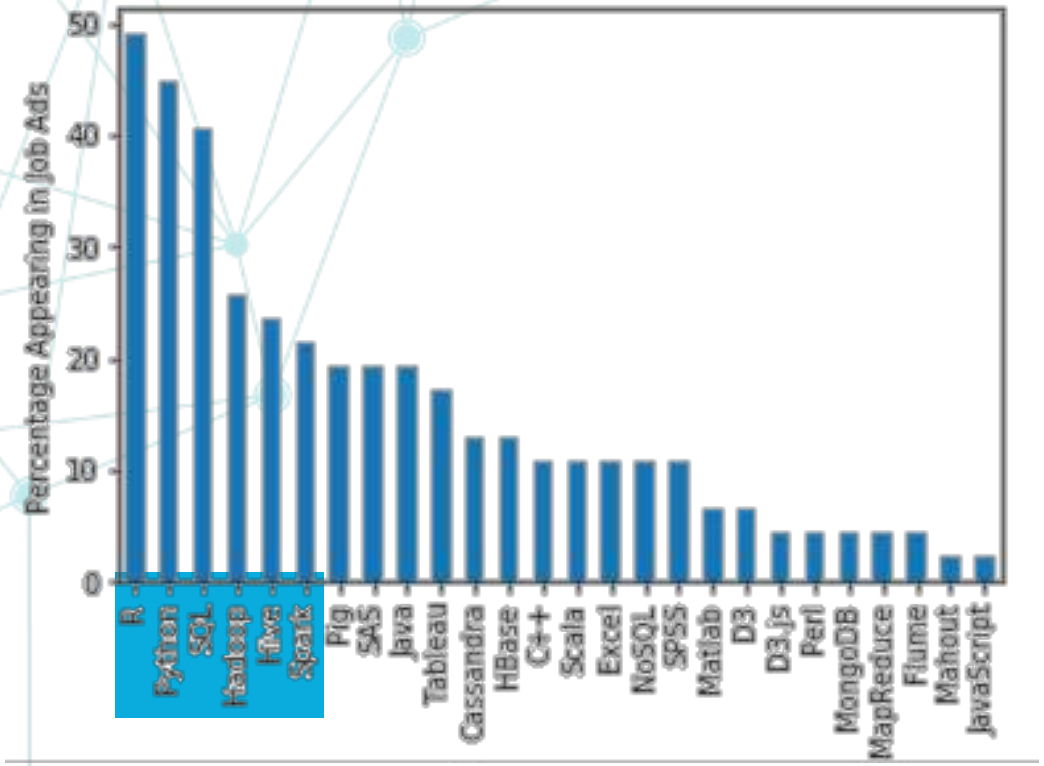
source: https://neuvoo.ca/salary/data-scientist/

neuvoo.ca/sal

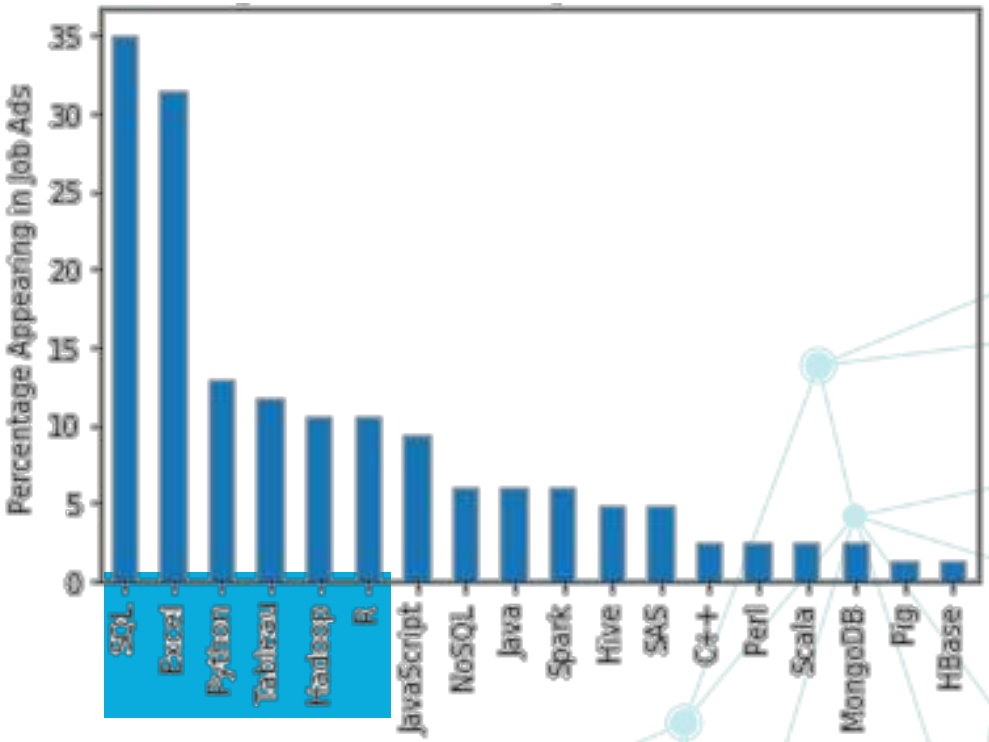
Data Science SkillsVenn Diagram



Skills required for Data Scientist |Analyst Jobs



Data Scientist



Data Analyst

Data Science Job Interview Questions

- Which data science frameworks and tools are you familiar with and use regularly?
- Here's my business problem and the data I collect. Tell me how you would approach it
- Why you should use NumPy arrays instead of nested Python lists?
- How do you handle missing values in Python?
- Explain the advantages and disadvantages of having more/fewer predictors in a model.

What is Data Science?

Why Even Data Science?

Research and Business collect data

- data in various forms
- data in big volumes
- data in fast speed

Make data-driven business decision

- data helps inform and answer business questions

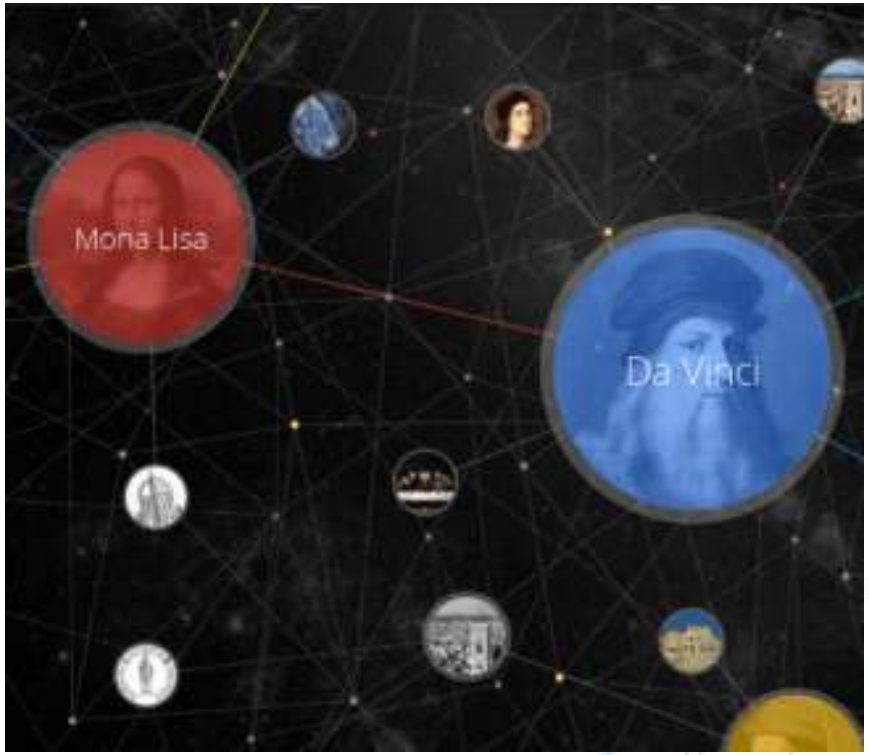
Build data products and turn data into gold

- personalized experience
- better risk management
- higher revenue
- design model

Data…

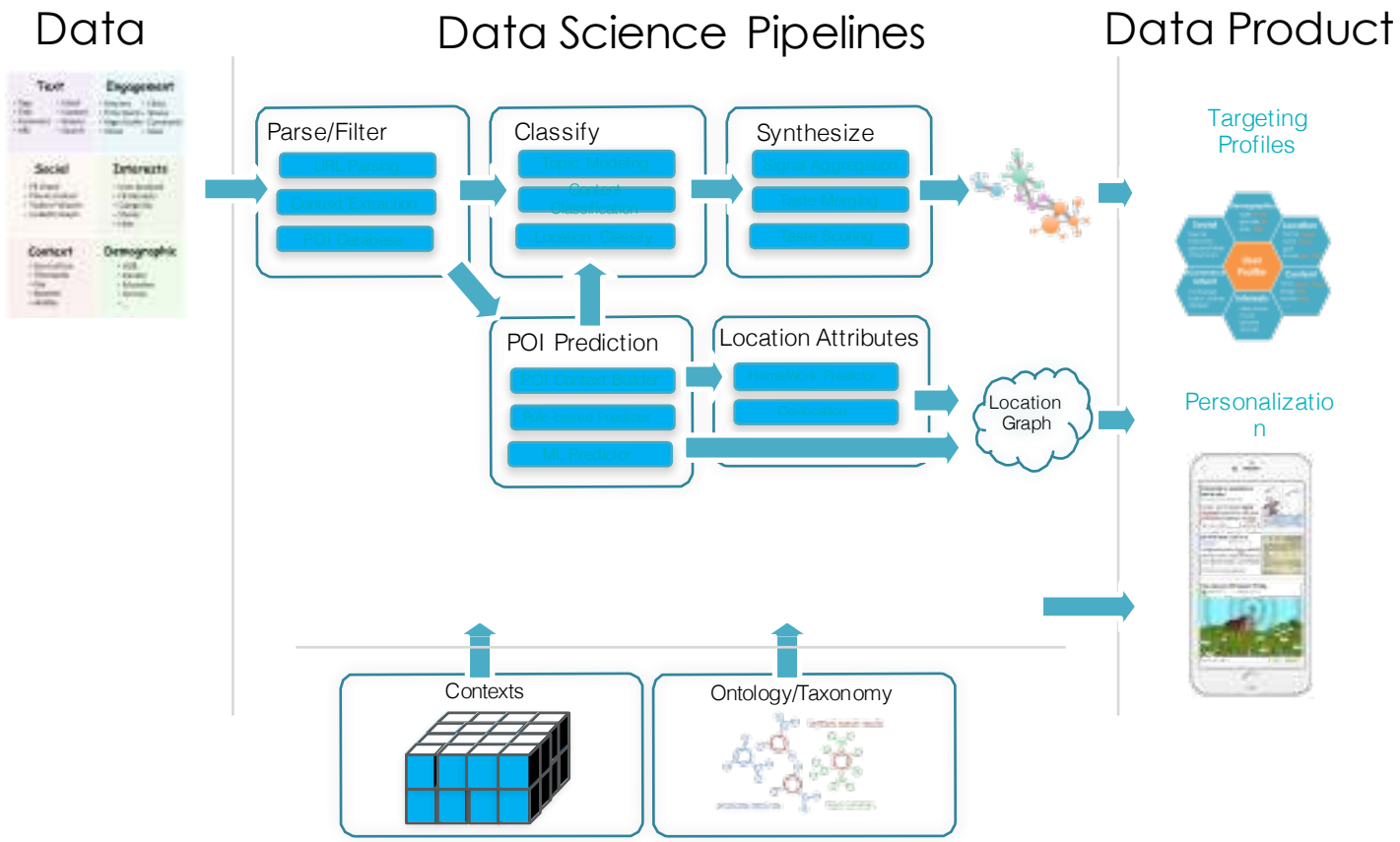
Data in many forms

- structured
  - tables
  - relational
- Unstructured
  - text
  - images
  - audio/video
- semi-structured
  - XML
  - JSON
  - Graph

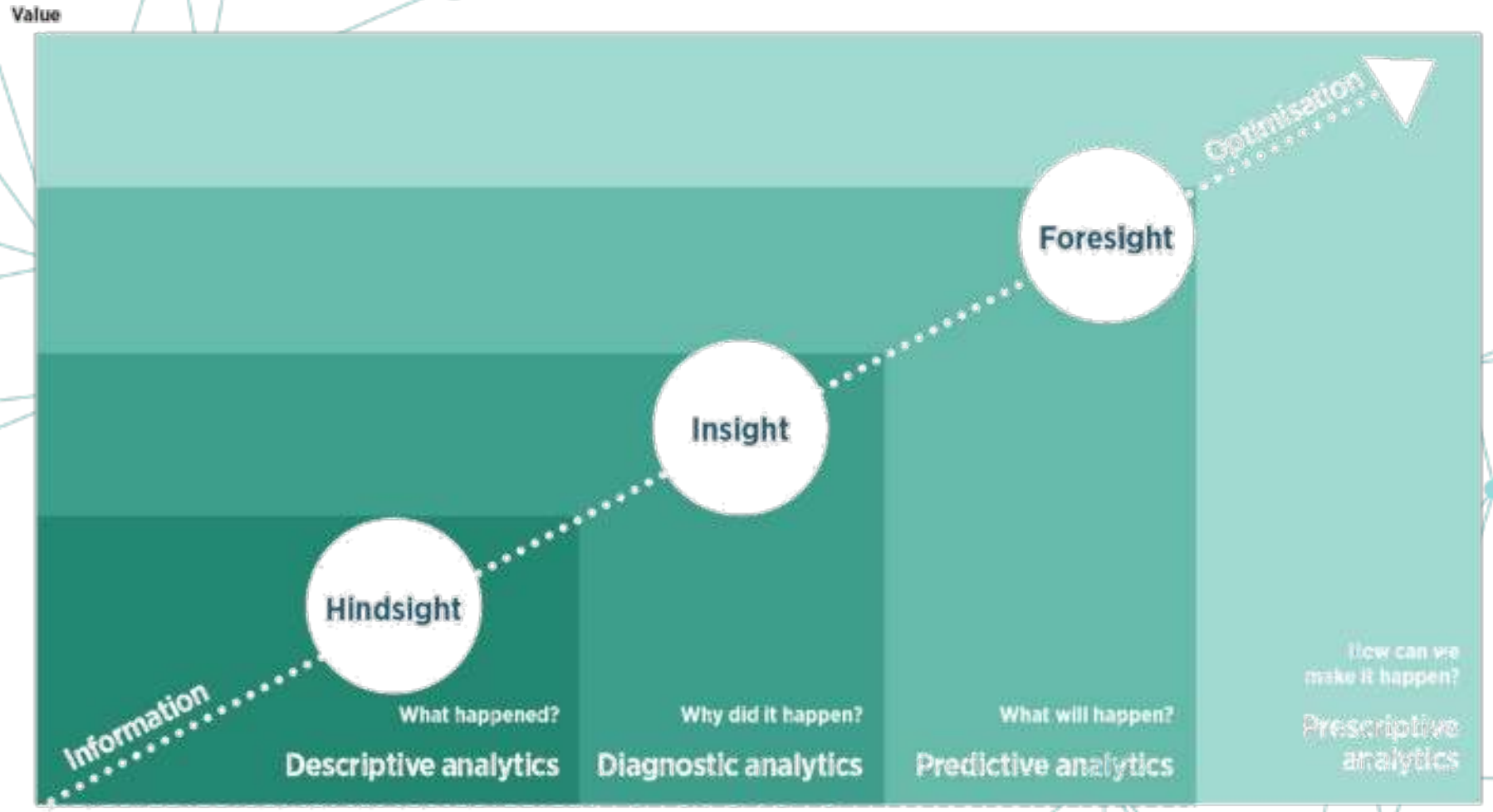


Science…

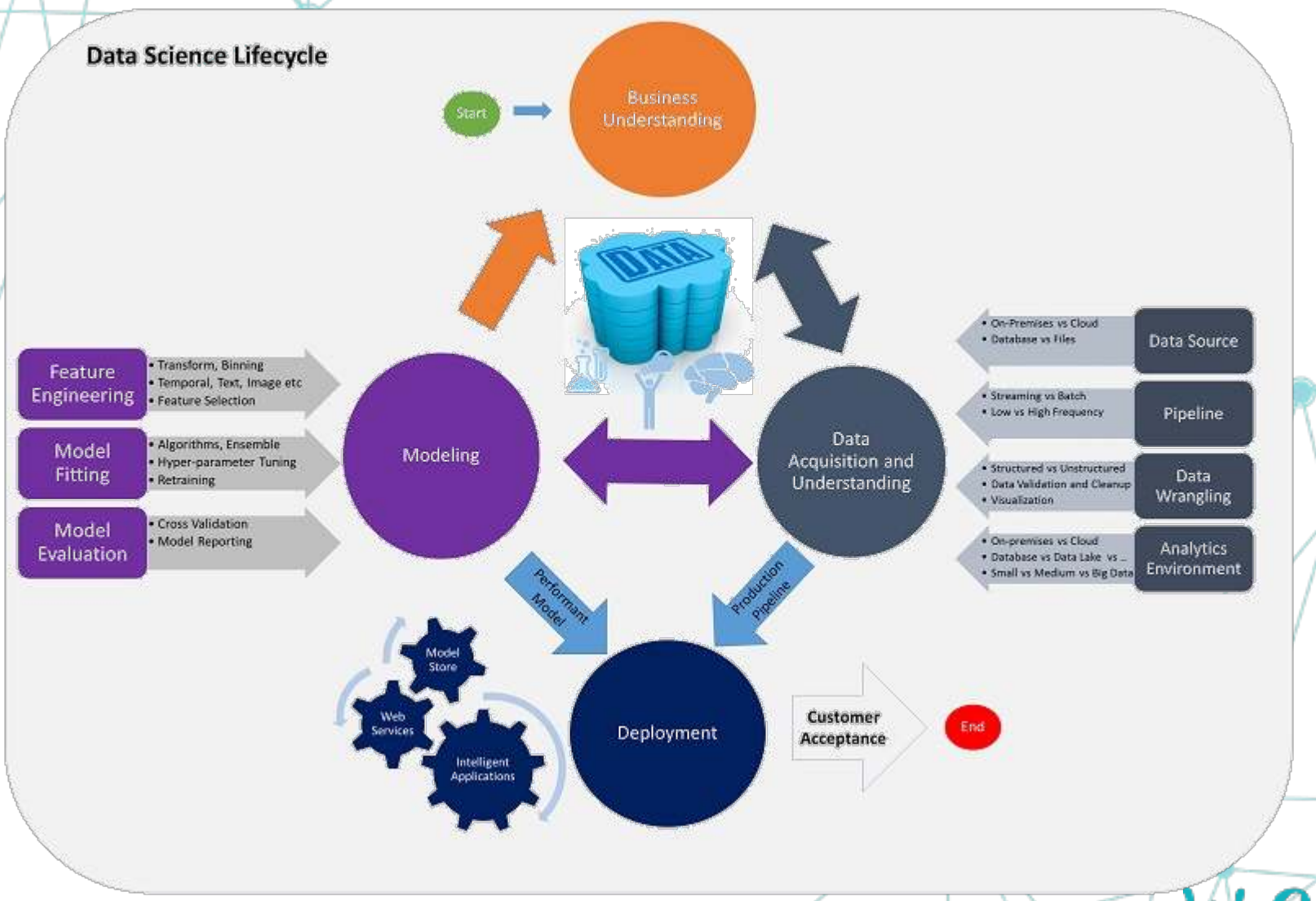
- Know something's wrong
- Get to know your customers
- Maximize customer value
- Understand the performance of your product



Why Data Science?



Data Science Lifecycle



Data Scientist

Chris Dixon
@cdixon

Follow

"A data scientist is a statistician who lives in San Francisco" via @smc90

Josh Wills
@josh_wills

Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Data Handling

Complex Analytics

Big Data

Storytelling

Thoughts of Data Science

- There's a lack of definitions around the most basic terminology.
  - What is "Big Data" anyway?
  - What does "data science" mean?
  - What is the relationship between Big Data and data science?
  - Is data science the science of Big Data?
  - Is data science only the stuff going on in companies like Google and Facebook and tech companies?
  - Why do many people refer to Big Data as crossing disciplines (astronomy, finance, tech, etc.) and to data science as only taking place in tech?
  - Just how *big* is big? Or is it just a relative term?
- These terms are so ambiguous, they're well-nigh meaningless.

Thoughts of Data Science

- Statisticians already feel that they are studying and working on the "Science of Data."
- Data science is *not* just a rebranding of statistics or machine learning but rather a field unto itself,

Thoughts of Data Science

- "Anything that has to call itself a science isn't."
- "data science" *itself* represents nothing, but of course what it represents may not be science but more of a craft.

Thoughts of Data Science

- There's is a difference between industry and academia.
  - But does it really have to be that way?
  - Why do many courses in school have to be so intrinsically out of touch with reality?
- The general experience of data scientists is
  - They have access to a *larger body of knowledge and methodology*, as well as a process,
  - Foundations in both statistics and computer science.

Why Data Science Now?

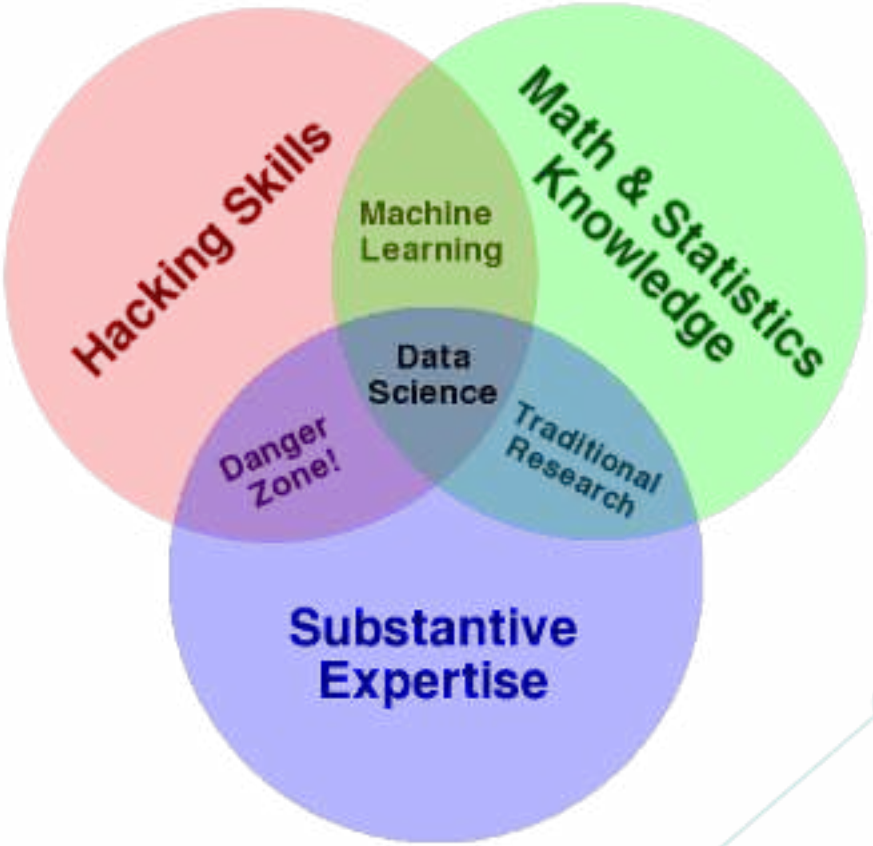
- Massive amounts of data about many aspects of our lives, and, simultaneously, an abundance of inexpensive computing power.

What is a Data Science?

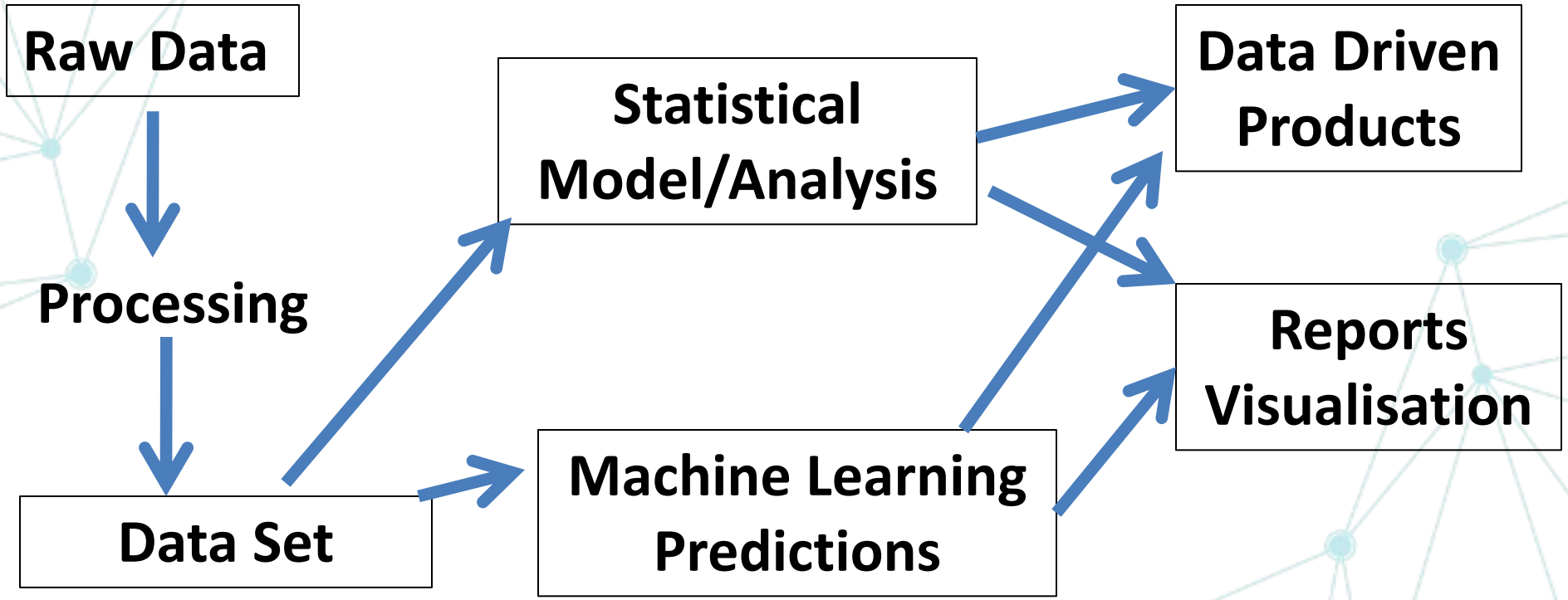
- Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.
- But data science is not merely hacking—because when hackers finish debugging their Bash one-liners and Pig scripts, few of them care about non-Euclidean distance metrics.
- Data science is not merely statistics, because when statisticians finish theorizing the perfect model, few could read a tab-delimited file into R if their job depended on it.
- Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible.

Introduction to Data Science

- What is a data Scientists?



What Does A Data Scientists do?



Basic Data Scientists Skill

- What does it mean for a data scientist to have "Substantive expertise" and why it is important?

  - Knows which questions to ask
  - Can interpret the data well
  - Understands structure of the data

Basic Data Scientists Skill

Background:

- In 1973, the University of California-Berkeley (UC-Berkeley) was sued for sex discrimination. Its admission data showed that men applying to graduate school at UC-Berkley were more likely to be admitted than women.
- The graduate schools had just accepted 44% of male applicants but only 35% of female applicants. The difference was so great that it was unlikely to be due to chance.
- By looking at the data more closely, you may realize that there is more to the story than meets the eye.

Basic Data Scientists Skill

This dataset contains information about the six most popular departments.

Feel free to examine and analyze the data with you favorite tool, like Excel, R, or even pen and paper.

Question for you

Now, do you agree that UC-Berkeley discriminated against women during the admission process?

Yes or No?

Admit	Gender	Dept,	Freq
Admitted	Male	A	512
Rejected	Male	A	313
Admitted	Female	A	89
Rejected	Female	A	19
Admitted	Male	B	353
Rejected	Male	B	207
Admitted	Female	B	17
Rejected	Female	B	8
Admitted	Male	C	120
Rejected	Male	C	205
Admitted	Female	C	202
Rejected	Female	C	391
Admitted	Male	D	138
Rejected	Male	D	279
Admitted	Female	D	131
Rejected	Female	D	244
Admitted	Male	E	53
Rejected	Male	E	138
Admitted	Female	E	94
Rejected	Female	E	299
Admitted	Male	F	22
Rejected	Male	F	351
Admitted	Female	F	24
Rejected	Female	F	317

How can we solve the world problems with Data Science?

Data Science can solve Problems you'd expect	Data Science can solve Problems you might not expect
• Netflix	• Bioinformatics
• Social Media	• Urban Planning
• Web Apps	• Astrophysics
	• Public Health

Definition of Data Science

1. Coding math, & statistics in applied settings
2. The analysis of diverse data
3. Inclusive analysis

Data Science Pathway

First.
Planning.

Second.
Data prep.

Third.
Modeling.

Fourth.
Follow up.

Planning

1. Define Goals
2. Organize Resources
3. Coordinate People
4. Schedule Project

Data Prep

1. Get data
2. Clean data
3. Explore data
4. Refine data

Modeling

1. Create Model
2. Validate model
3. Evaluate Model
4. Refine Model

Follow Up

1. Present Model
2. Deploy Model
3. Revisit Model
4. Archive assets

Roles in Data Science

Engineer

Focus on Back end, hardware. Software
Make DS Possible
Developer, Database Administrators

Big Data Specialists

Focus on Computer Science & Math
Machine Learning
Data Products

Roles in Data Science

Researchers

- Focus on Domain Specific research
- Physics , genetics
- Strong Statistics

Analyst

- Day-to-day tasks
- Web analytics, SQL
- Good for business

Roles in Data Science

Business

- Frames business relevant questions
- Manage Projects
- Must “Speak Data”

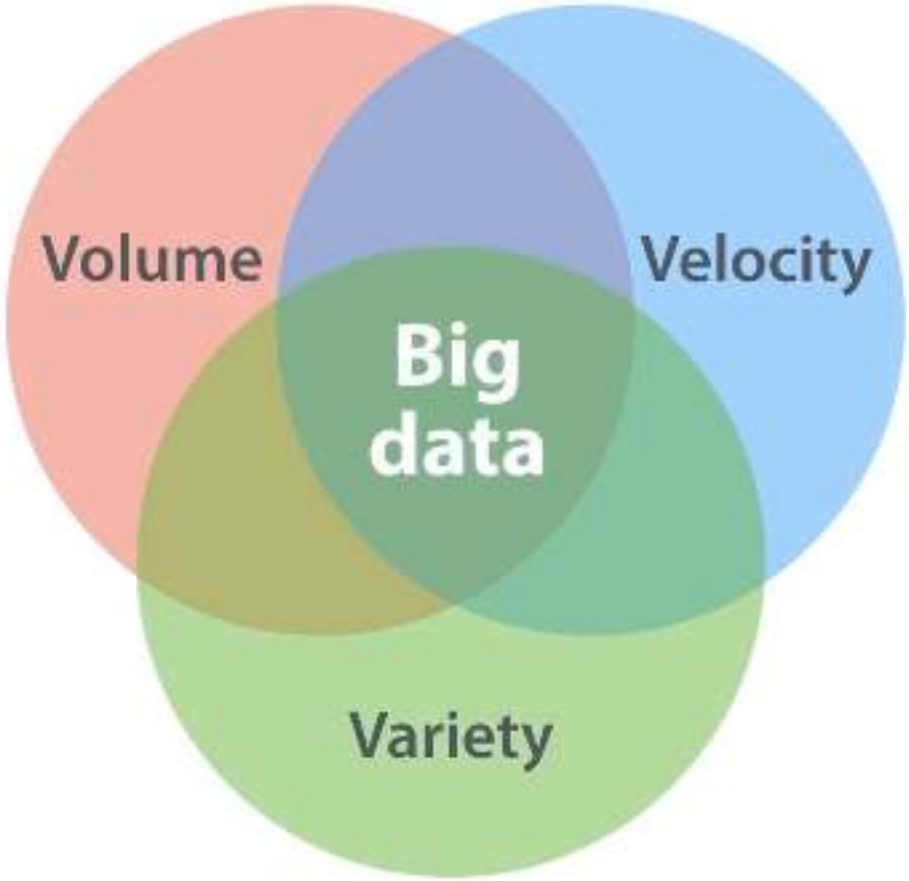
Entrepreneur

- Data Startup
- Needs data & business Skills
- Creative thoughts

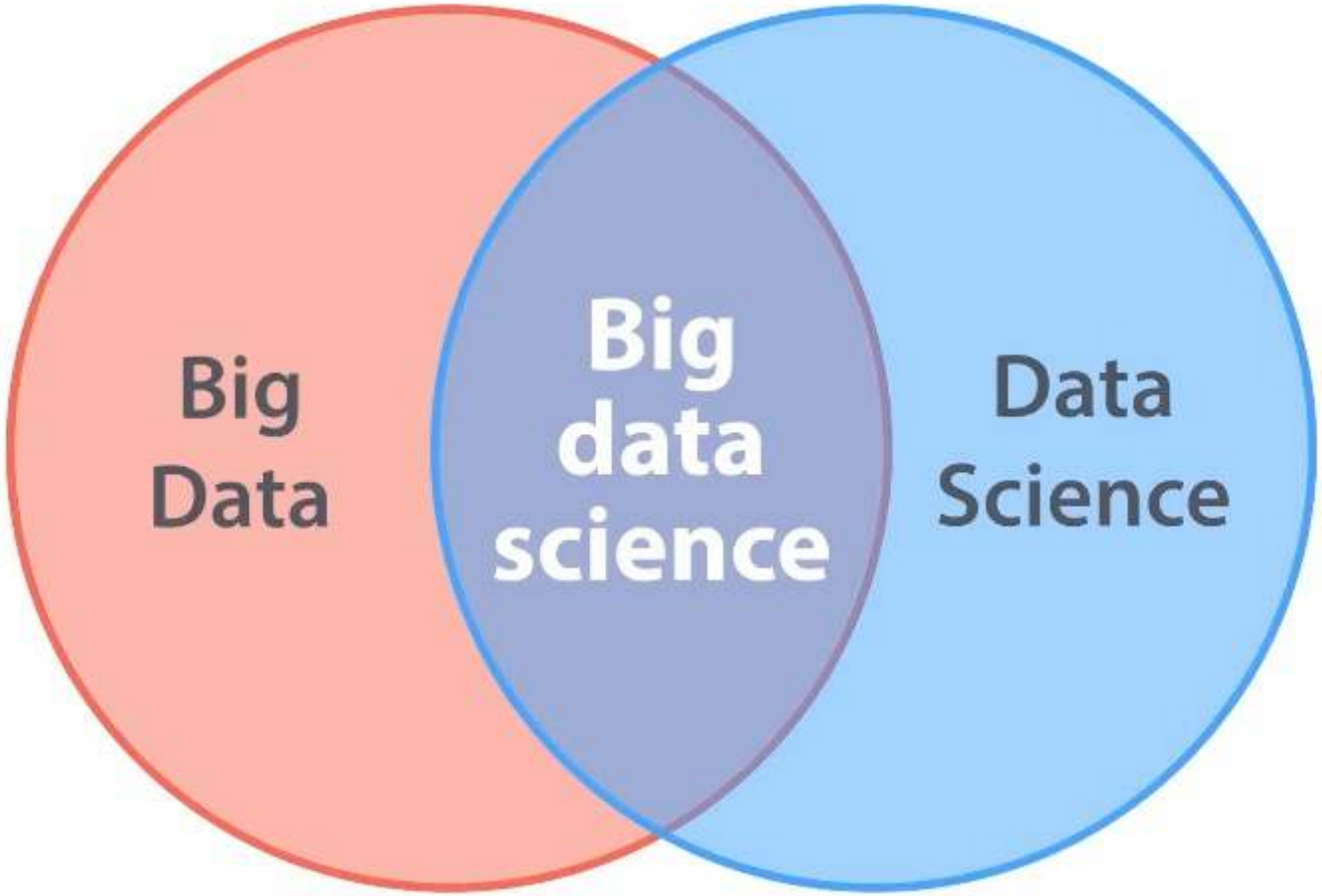
Roles in Data Science

- Unicorn
- A mythical creature with magical abilities
- A mythical data scientist with universal abilities

Big Data Vs Data Science



Big Data Vs Data Science



Big Data Vs Data Science

Big Data

Word Count, ML

Data Science

1Variable (Volume, Velocity or Variety)
Genetics data
Streaming sensor data
Facial Recognition

Big Data Science

Volume , Velocity, Variety

Python Packages

NumPy
& SciPy.

Matplotlib
& Seaborn.

Pandas.

scikit-learn.

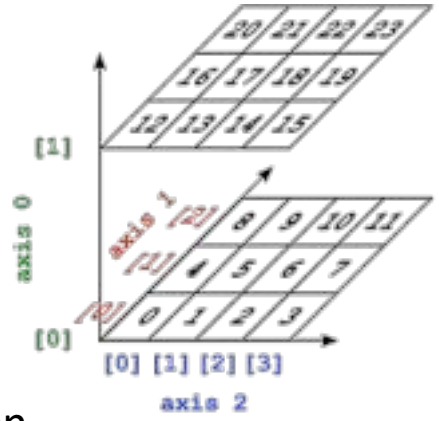
Data Science with Numpy

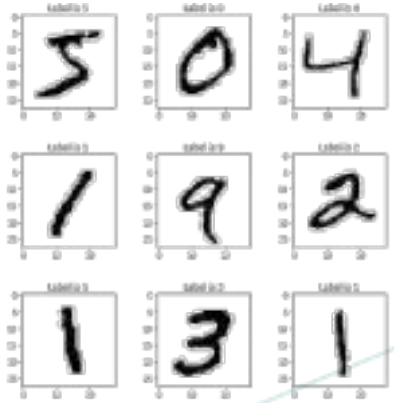
Learning

- data science lifecycle
- numpy arrays
  - array basics
  - array operations
  - array mathematics
  - array methods

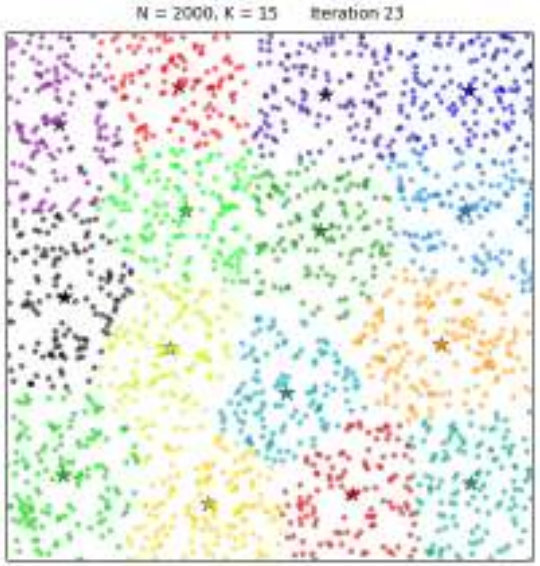
Outcome

- manipulating arrays and matrices
- k-means clustering





Python



Data Model and SQL

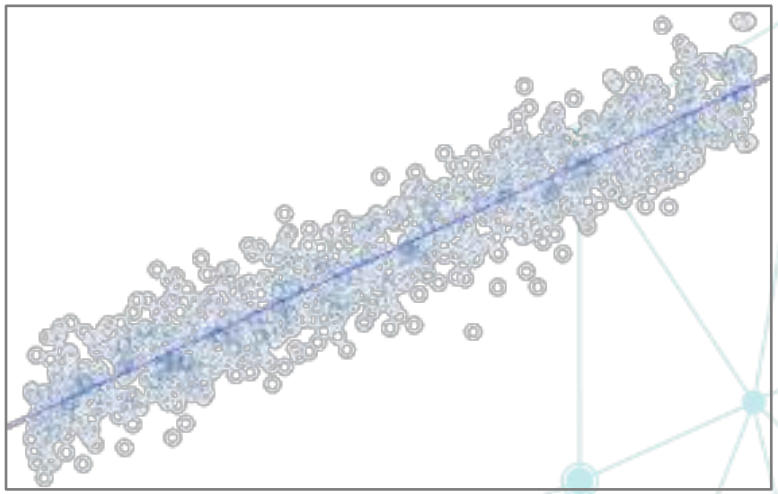
Learning

- Types of Data Model
- SQL
  - Create
  - Extract
  - Display

Outcome

- Analyse Data Models
- Manipulate Data using SQL

Statistics with Scipy



$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

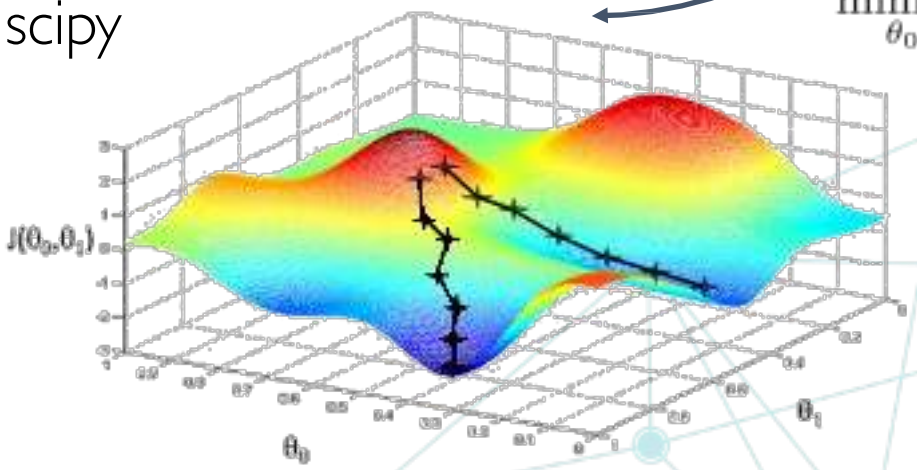
Learning

- scientific computing with Scipy
- statistics fundamentals with Scipy
- linear regression

Outcome

- hands-on statistics with scipy
- build a regression model with scipy

$$\underset{\theta_0,\theta_1}{\text{minimize}}\ J(\theta_0,\theta_1)$$



Data Collection via Web Scraping

Learning

- understanding various ways of data collection in data science
- build a simple web scraper to collect job insight

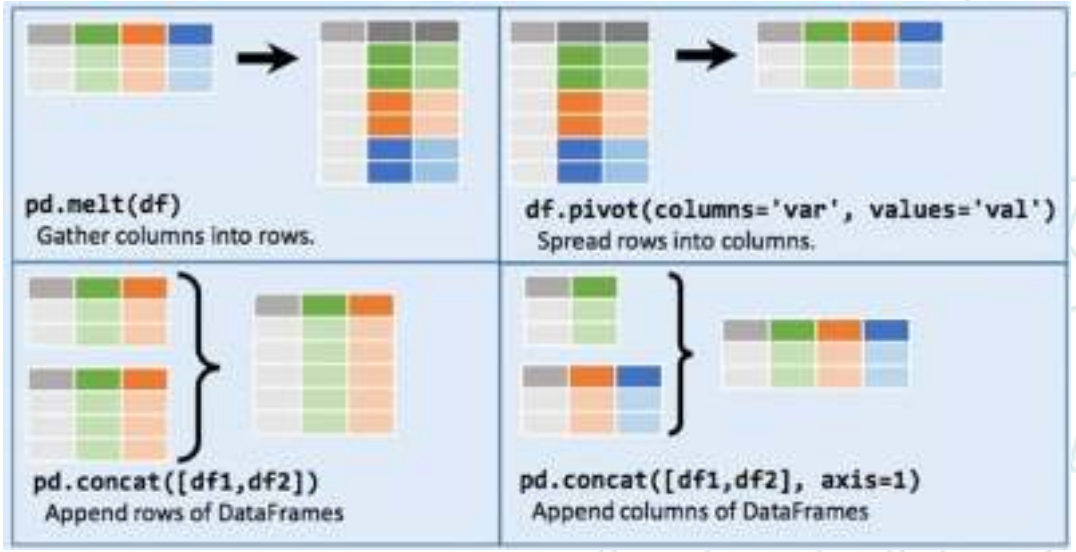
Outcome

- scrape websites for your next data project
- save hours on job search

Data Munging with Pandas

Learning

- Pandas Series
- Pandas DataFrames
- reading data from various data sources
- basic operations (indexing, slicing, sorting)
- summary statistics
- handling missing data, outliers and duplicates
- file I/O with pandas
- working directly with web data



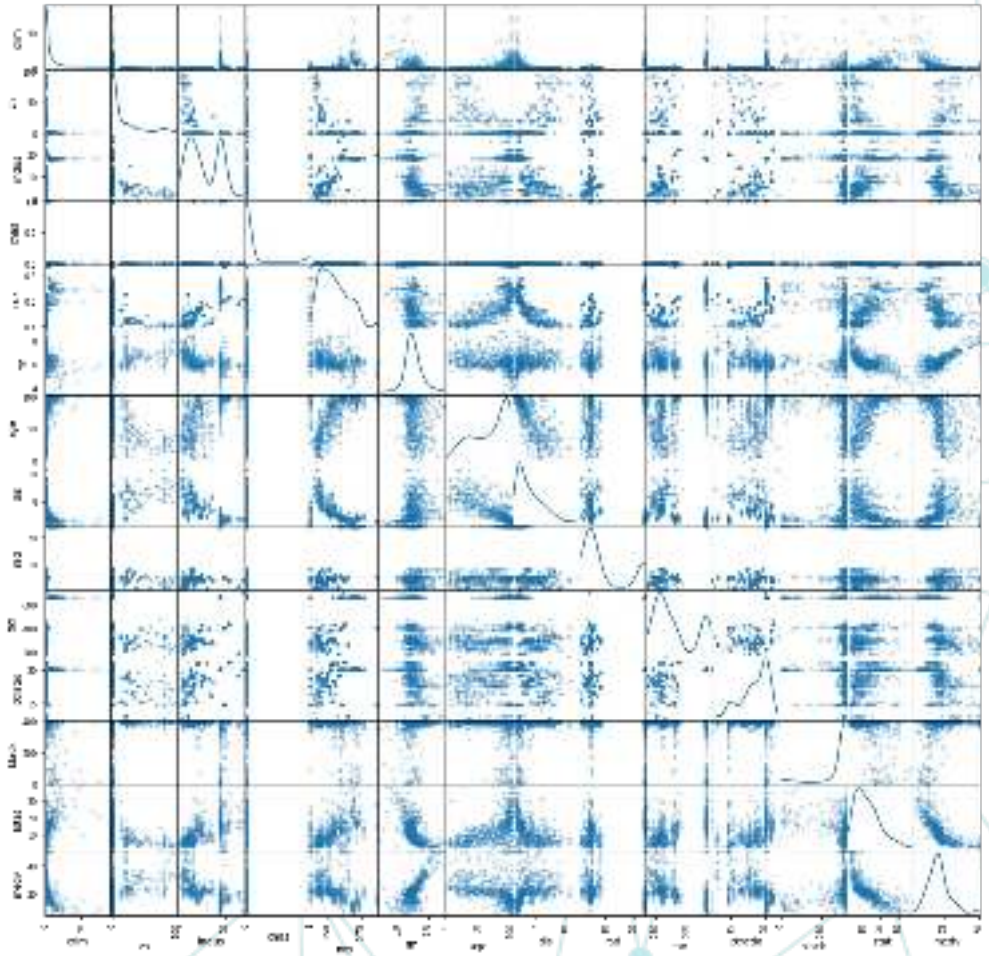
Outcome

- data manipulation with Python
- able to transform datasets in various ways

Data Analysis with Pandas and Data Visualization

Learning

- advanced Pandas operations
  - advanced indexing and slicing
  - aggregating data
  - shape operations
  - merging series and dataframes
  - pivoting
- Data visualization
  - plotting with pandas &matplotlib
  - plotting with Seaborn
  - plotting with plot.ly



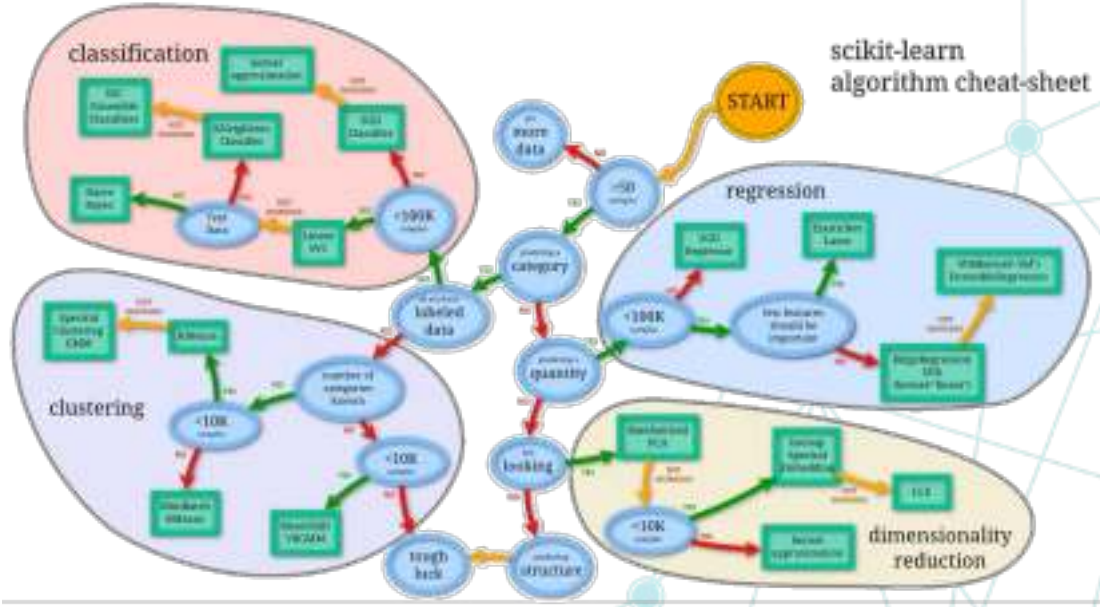
Outcome

- become a dataframe pro
- become adept at visualization

Predictive Modeling and Advanced Analytics

Learning

- classification vs regression models
- predictive modeling methods
- python's scikit-learn package for machine learning
  - data preprocessing
  - feature extraction
  - feature selection
  - model selection
  - result validation



Outcome

- understand machine learning basics
- build your first predictive model

Predictive Modeling and Advanced Analytics

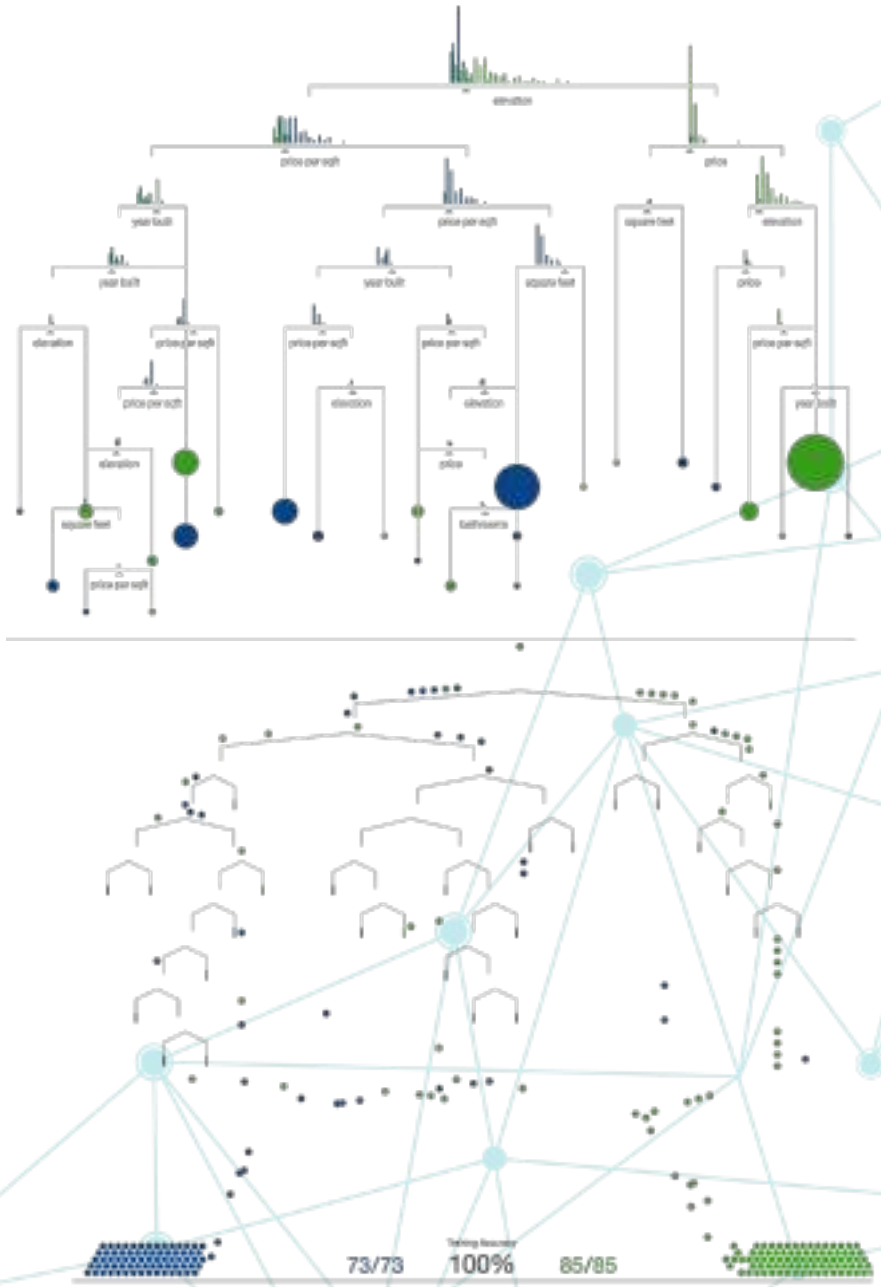
Machine Learning

Learning

- work on a hands-on machine learning project independently
- build your first classification and regression models

Outcome

- know what tool to use to solve data challenges
- able to perform customer segmentation model at work
- able to build a simple regression/classification model to help your business make better decision



Learning Outcome

Data Science Skills

- Get familiar with Data Science Lifecycle
- Be able to work on data problems independently
- Master essential python tools for data collection, munging, analysis and visualization