Data Science and the Data Scientist Toolkit

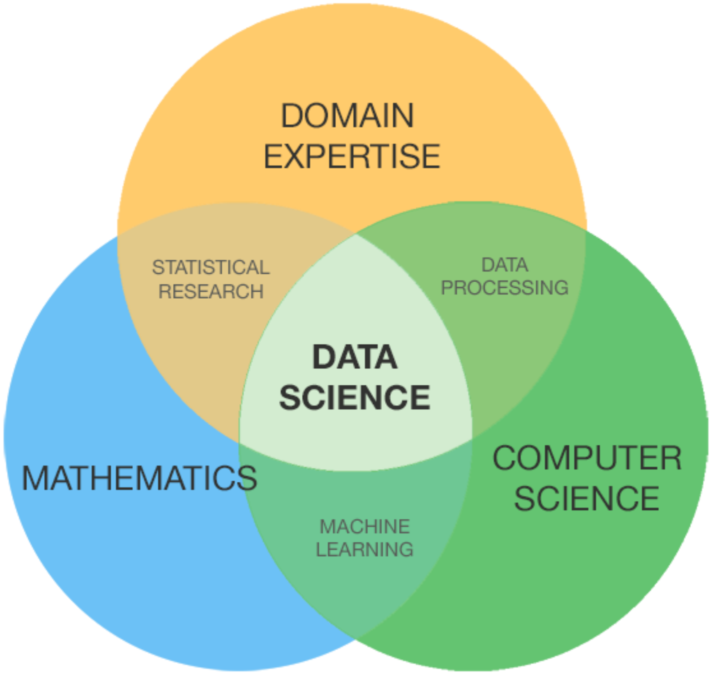
// FLATIRON SCHOOL

Agenda

- What is Data Science?
  - Roles and Responsibilities
  - The Process
- The Data Science Toolkit (Phase 1)

So:
What is
Data Science?

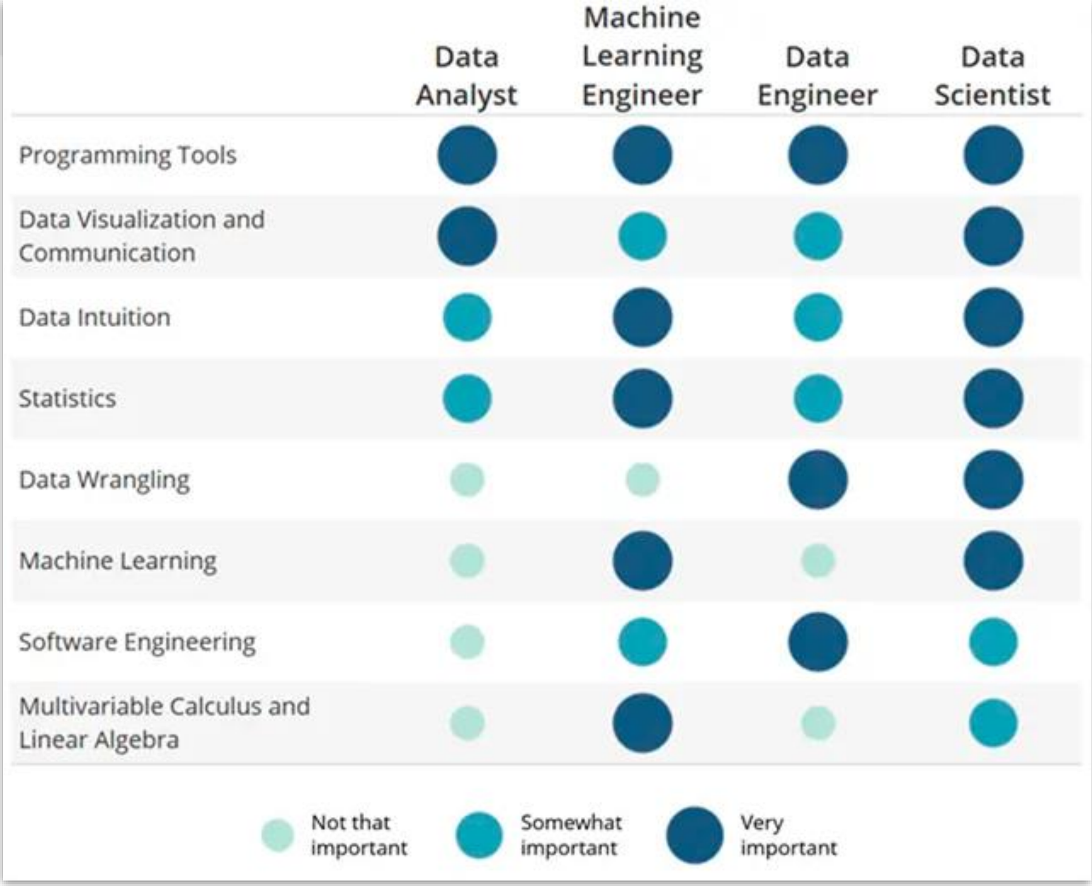
The Data Science Venn Diagram



Source: Palmer, Shelly. Data Science for the C-Suite.
New York: Digital Living Press, 2015. Print.

A data scientist is responsible for **collecting, analyzing and interpreting** data on various scales. **Offshoot of several traditional technical roles**, including mathematician, scientist, statistician and computer professional.

Common Roles & Responsibilities

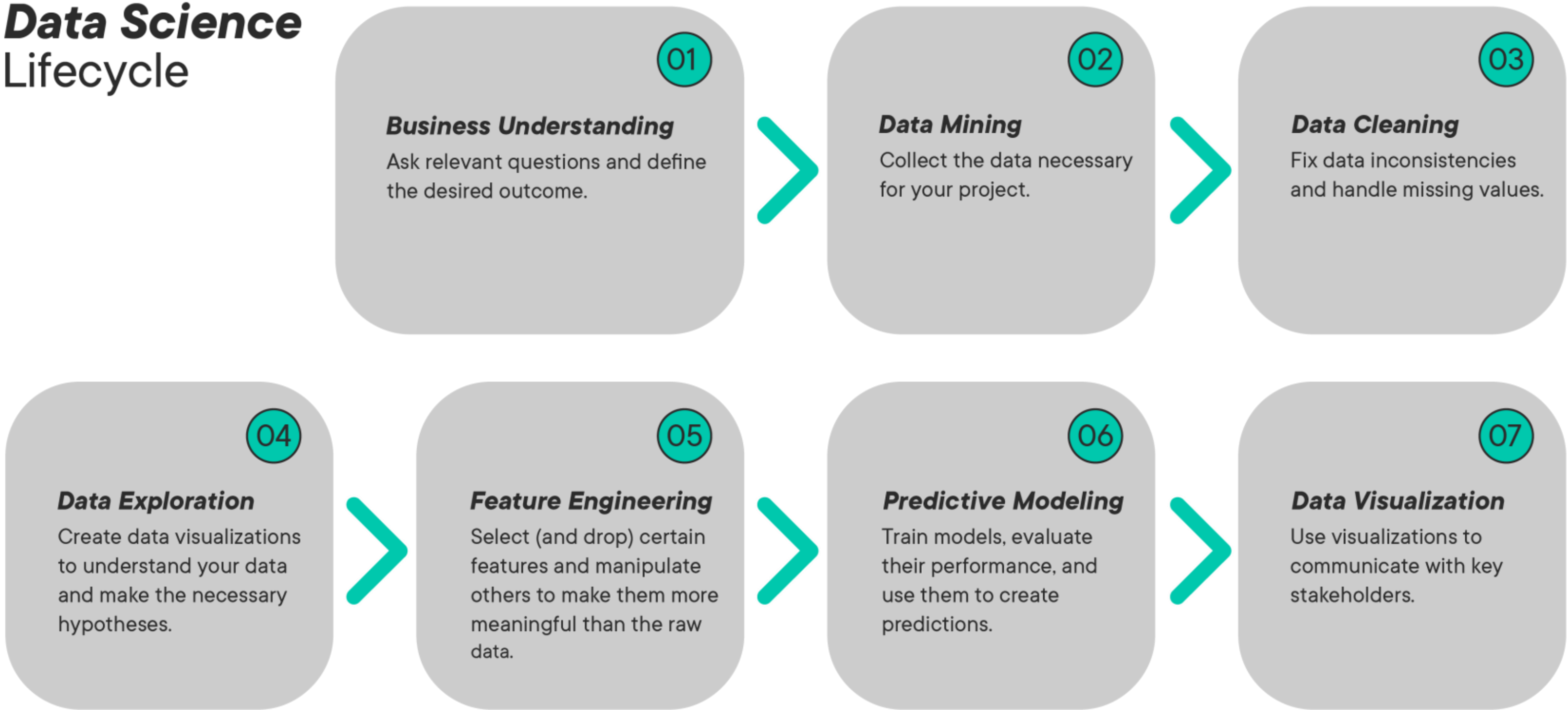


	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	●	●	●	●
Data Visualization and Communication	●	◉	◉	●
Data Intuition	◉	●	◉	●
Statistics	◉	●	◉	●
Data Wrangling	○	○	●	●
Machine Learning	○	●	○	●
Software Engineering	○	◉	●	◉
Multivariable Calculus and Linear Algebra	○	●	○	◉

○ Not that important ◉ Somewhat important ● Very important

The Data Science Process

Data Science Lifecycle



01 ***Business Understanding***

Ask relevant questions and define the desired outcome.

02 ***Data Mining***

Collect the data necessary for your project.

03 ***Data Cleaning***

Fix data inconsistencies and handle missing values.

04 ***Data Exploration***

Create data visualizations to understand your data and make the necessary hypotheses.

05 ***Feature Engineering***

Select (and drop) certain features and manipulate others to make them more meaningful than the raw data.

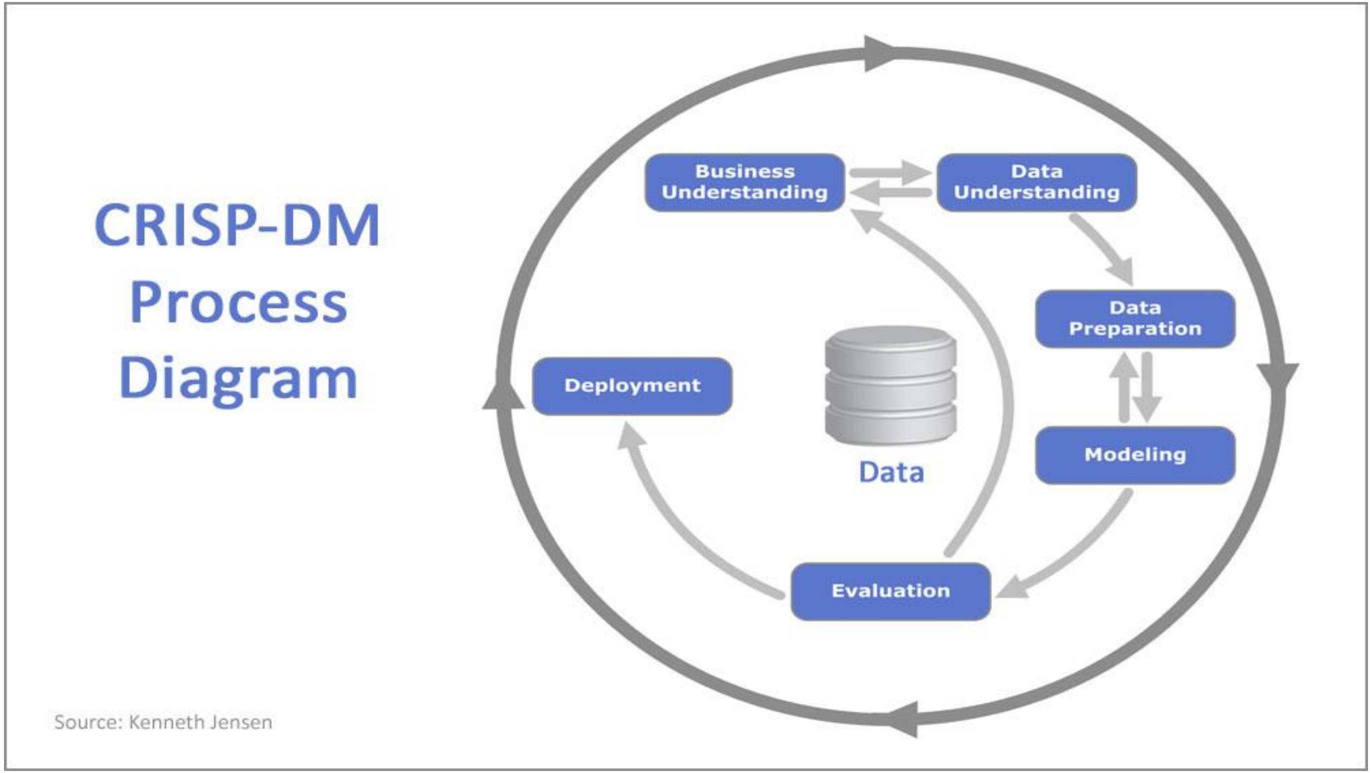
06 ***Predictive Modeling***

Train models, evaluate their performance, and use them to create predictions.

07 ***Data Visualization***

Use visualizations to communicate with key stakeholders.

But it’s actually an iterative process…



Source: Kenneth Jensen

Asking the right questions

An irrelevant question + data/machine learning/stats = an irrelevant answer

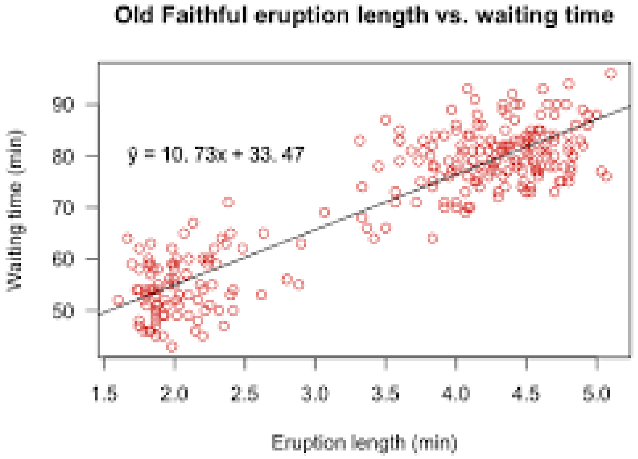
Problem Formulation

Transformation:

Question into data science problem.

Some typical Data Science problems

Regression:



Old Faithful

Predict time between eruptions based on previous eruption duration.

Some typical Data Science problems

Classification:

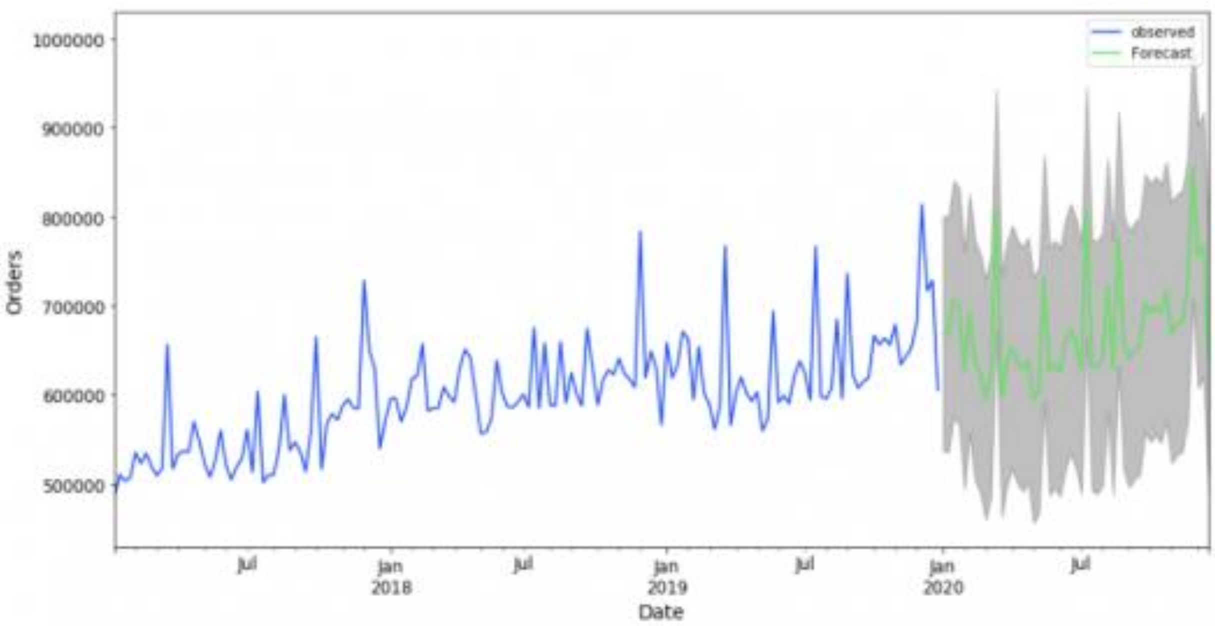
or

Koala

Red Panda?

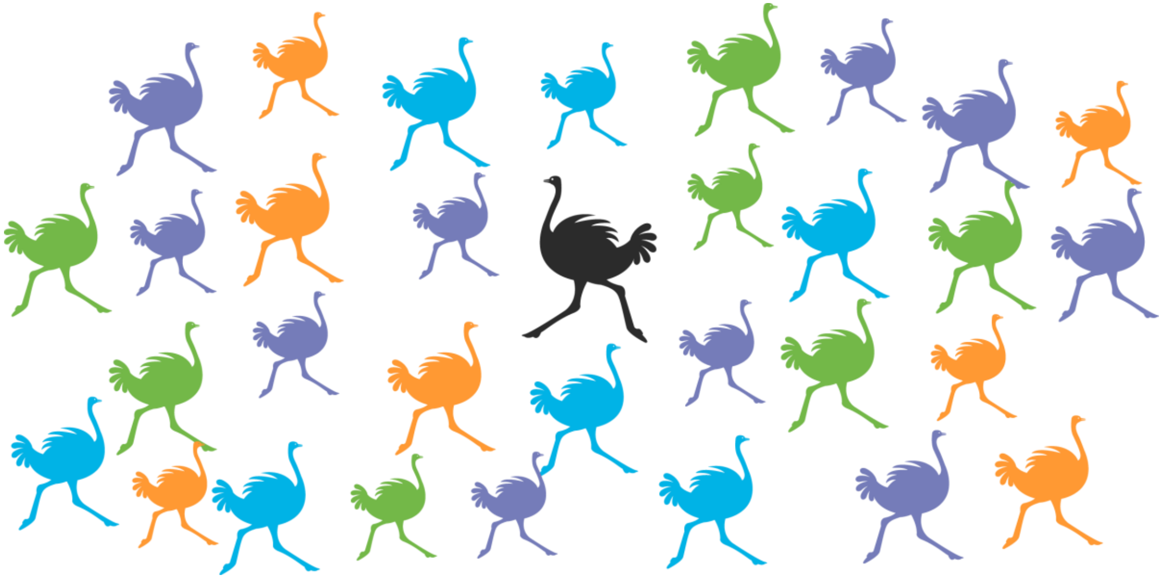
Some typical Data Science problems

Time series forecasting:

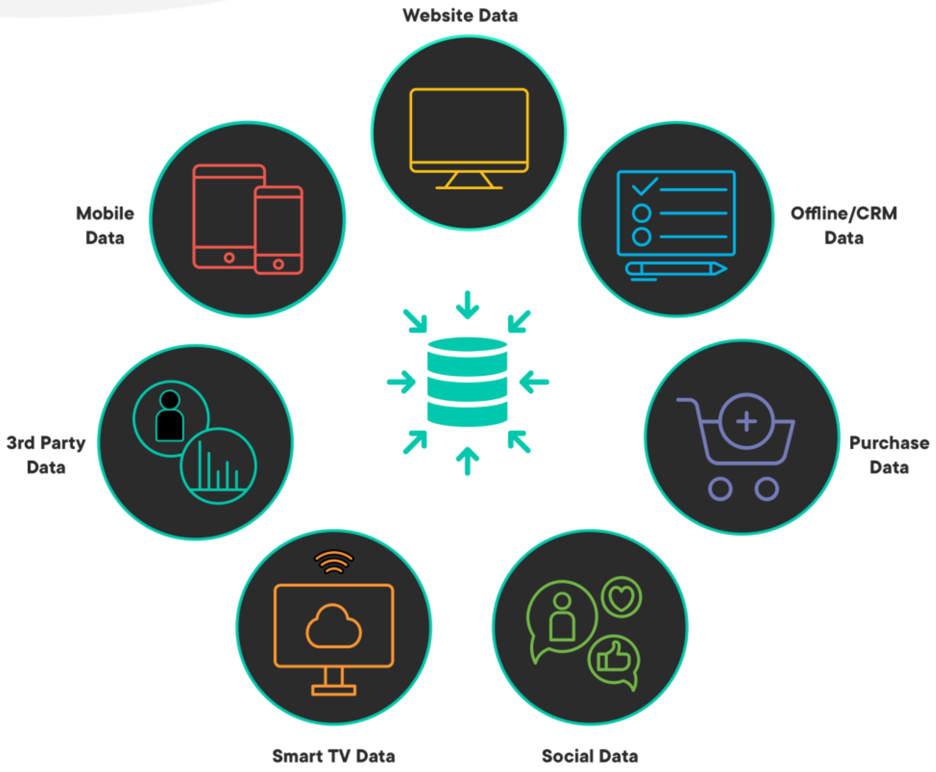


Some typical Data Science problems

Anomaly detection:



Data you might encounter



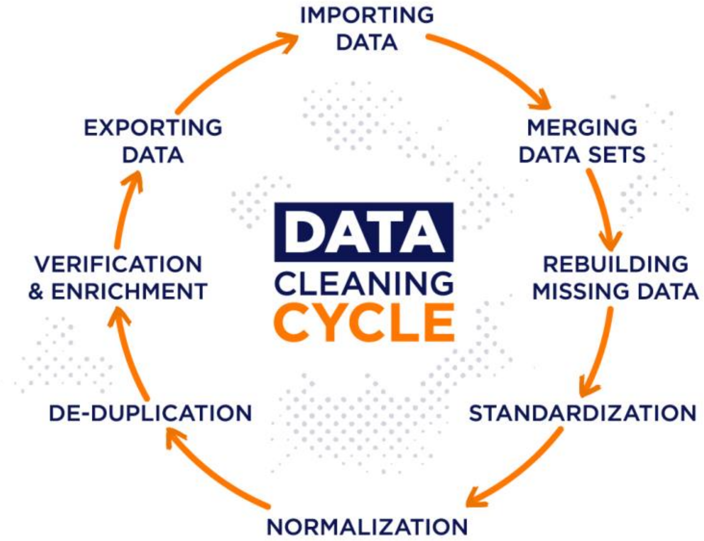
Be the data sculptor

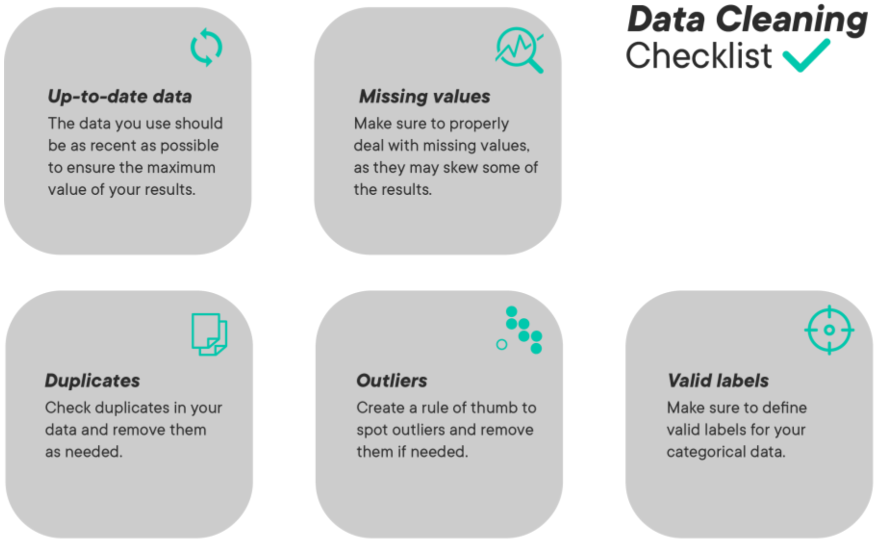
Reshape the data:

- Clean and transform the data to your will.
- Data in useful form: modeling, answering your question.

An art form.

Where data scientists spend most of their time…





Data Cleaning Checklist ✓

Up-to-date data
The data you use should be as recent as possible to ensure the maximum value of your results.

Missing values
Make sure to properly deal with missing values, as they may skew some of the results.

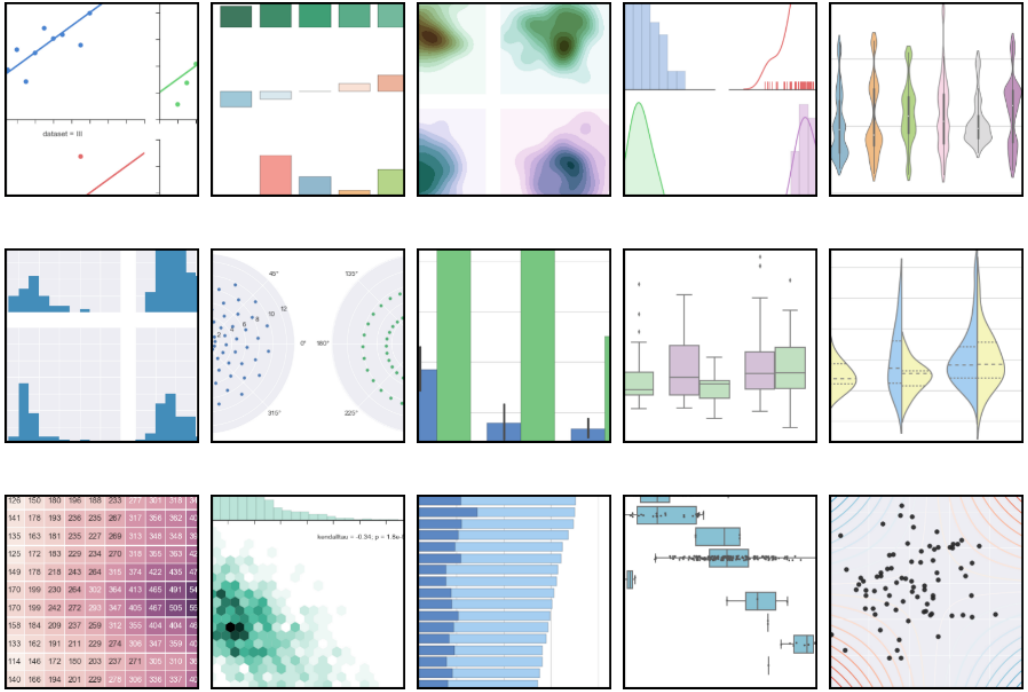
Duplicates
Check duplicates in your data and remove them as needed.

Outliers
Create a rule of thumb to spot outliers and remove them if needed.

Valid labels
Make sure to define valid labels for your categorical data.

Exploratory Data Analysis

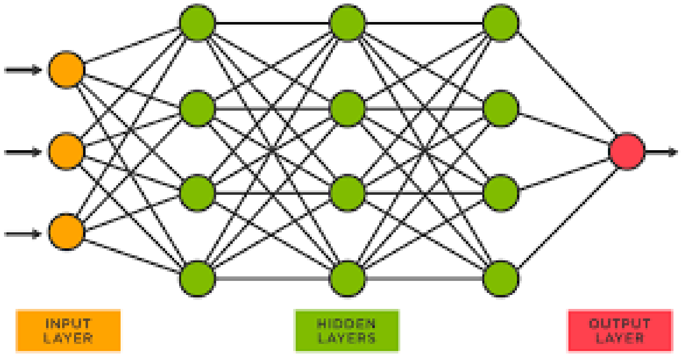
Visualize and understand your data to transform into useful form.



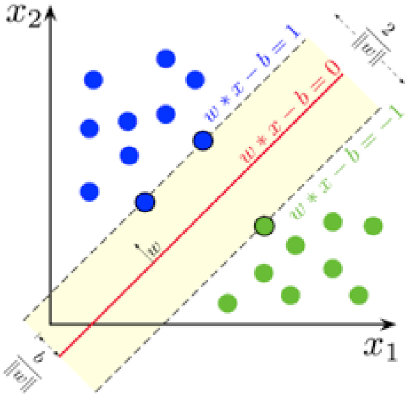
Feature Engineering

Transform raw data into meaningful features that directly address the problem you are trying to solve.

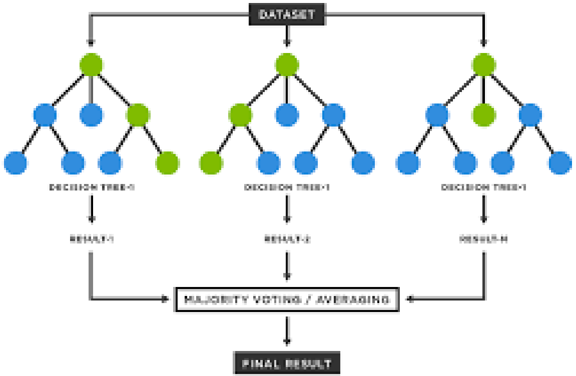
Modeling



Neural Network.



Support Vector Machine



Random Forests

And more…try different models, tune, see what works best.

Presenting/visualizing results

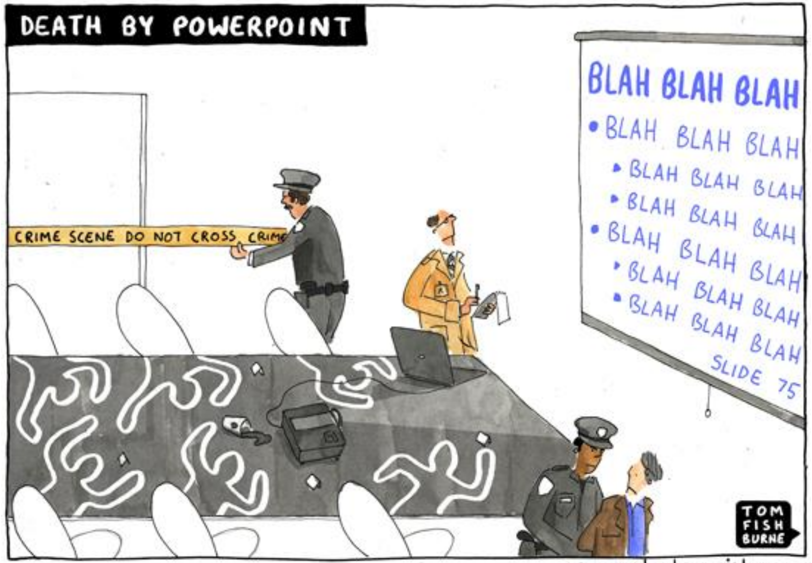
This is key to a data scientist: Presentations/Reports

- Know your audience

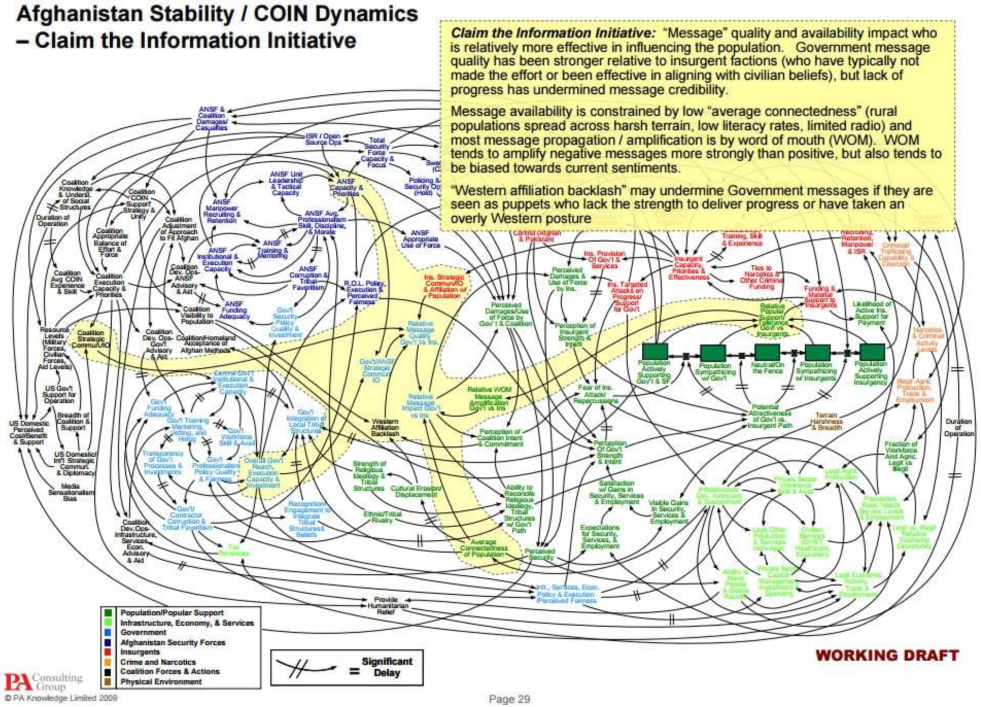
- State the problem clearly.
- How did you go about solving the problem?
- Key factors
- Visualizations of data and model
- Making recommendations.

Bad Sign

Avoid



© marketoonist.com



Afghanistan Stability / COIN Dynamics – Claim the Information Initiative

Claim the Information Initiative: "Message" quality and availability impact who is relatively more effective in influencing the population. Government message quality has been stronger relative to insurgent factions (who have typically not made the effort or been effective in aligning with civilian beliefs), but lack of progress has undermined message credibility.

Message availability is constrained by low "average connectedness" (rural populations spread across harsh terrain, low literacy rates, limited radio) and most message propagation / amplification is by word of mouth (WOM). WOM tends to amplify negative messages more strongly than positive, but also tends to be biased towards current sentiments.

"Western affiliation backlash" may undermine Government messages if they are seen as puppets who lack the strength to deliver progress or have taken an overly Western posture

- 🟩 Population/Popular Support
- 🟩 Infrastructure, Economy, & Services
- 🟦 Government
- 🟦 Afghanistan Security Forces
- 🟥 Insurgents
- 🟧 Crime and Narcotics
- ■ Coalition Forces & Actions
- 🟫 Physical Environment

—//→ = Significant Delay

ANSF & Coalition Damages/ Casualties
ISR / Open Source Ops
Total Security Force Capacity & Focus
Coalition Knowledge & Underst. of Social Structures
Coalition COIN Support Strategy & Unity
Duration of Operation
Western Affiliation Backlash
Relative WOM Message Amplification Gov't vs Ins
Population Actively Supporting Gov't & SF
Population Sympathizing w/ Gov't
Neutral/On the Fence
Population Sympathizing w/ Insurgents
Population Actively Supporting Insurgency
Average Connectedness of Population
Provide Humanitarian Relief
Perceived Security

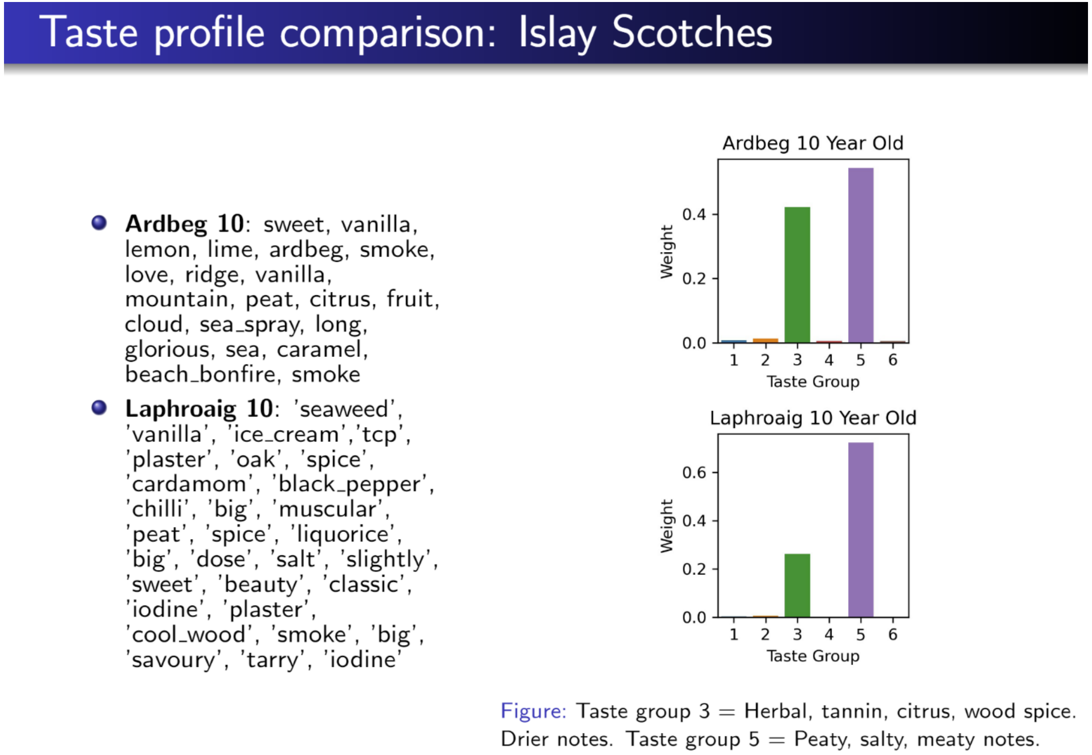
WORKING DRAFT

PA Consulting Group

© PA Knowledge Limited 2009

Page 29

Yeah, OK.

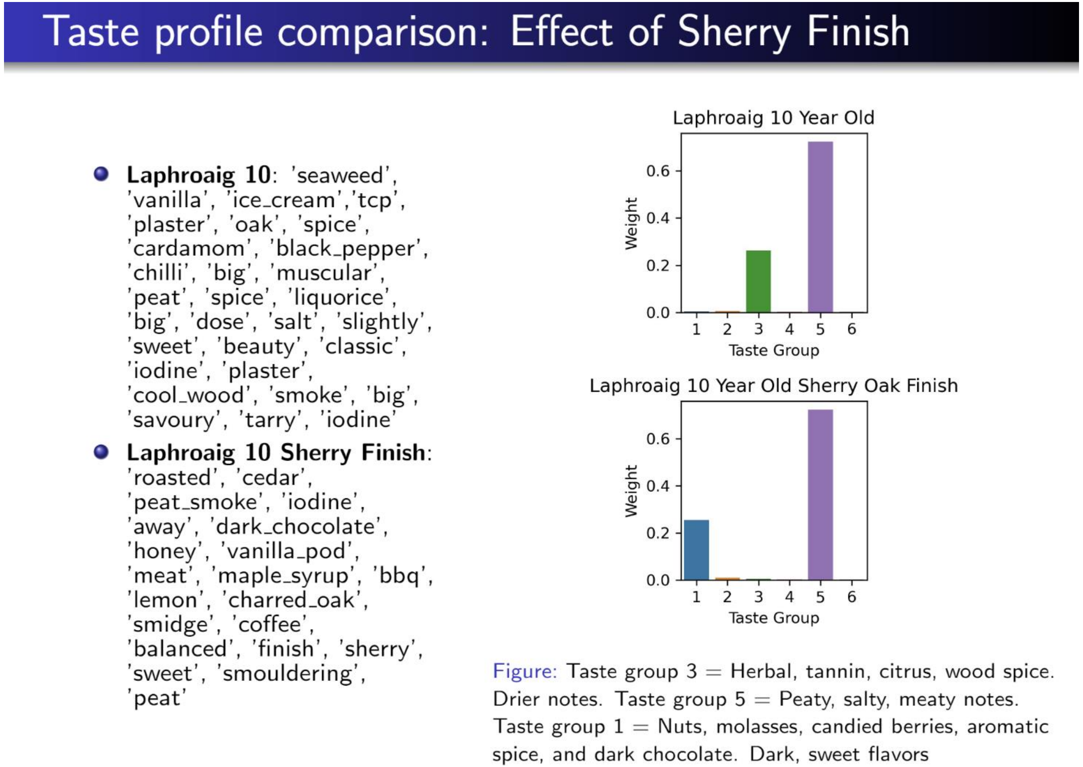


Taste profile comparison: Islay Scotches

- **Ardbeg 10**: sweet, vanilla, lemon, lime, ardbeg, smoke, love, ridge, vanilla, mountain, peat, citrus, fruit, cloud, sea_spray, long, glorious, sea, caramel, beach_bonfire, smoke
- **Laphroaig 10**: 'seaweed', 'vanilla', 'ice_cream','tcp', 'plaster', 'oak', 'spice', 'cardamom', 'black_pepper', 'chilli', 'big', 'muscular', 'peat', 'spice', 'liquorice', 'big', 'dose', 'salt', 'slightly', 'sweet', 'beauty', 'classic', 'iodine', 'plaster', 'cool_wood', 'smoke', 'big', 'savoury', 'tarry', 'iodine'

Figure: Taste group 3 = Herbal, tannin, citrus, wood spice. Drier notes. Taste group 5 = Peaty, salty, meaty notes.

Yeah, OK.



Taste profile comparison: Effect of Sherry Finish

- **Laphroaig 10**: ’seaweed’, ’vanilla’, ’ice_cream’,’tcp’, ’plaster’, ’oak’, ’spice’, ’cardamom’, ’black_pepper’, ’chilli’, ’big’, ’muscular’, ’peat’, ’spice’, ’liquorice’, ’big’, ’dose’, ’salt’, ’slightly’, ’sweet’, ’beauty’, ’classic’, ’iodine’, ’plaster’, ’cool_wood’, ’smoke’, ’big’, ’savoury’, ’tarry’, ’iodine’
- **Laphroaig 10 Sherry Finish**: ’roasted’, ’cedar’, ’peat_smoke’, ’iodine’, ’away’, ’dark_chocolate’, ’honey’, ’vanilla_pod’, ’meat’, ’maple_syrup’, ’bbq’, ’lemon’, ’charred_oak’, ’smidge’, ’coffee’, ’balanced’, ’finish’, ’sherry’, ’sweet’, ’smouldering’, ’peat’

Figure: Taste group 3 = Herbal, tannin, citrus, wood spice. Drier notes. Taste group 5 = Peaty, salty, meaty notes. Taste group 1 = Nuts, molasses, candied berries, aromatic spice, and dark chocolate. Dark, sweet flavors

The Data Science Toolkit

Data Science Toolkit - Phase 1

Languages

python™

SQL

Interfaces

jupyter

illumiDesk

Version Control

git

GitHub

Versioning

ANACONDA.

PyPi

Languages

Python

- Free, open source, versatile, powerful
- Not just for data science!
- Object-oriented (everything is an ‘object’)
- The Zen of Python

Structured Query Language (SQL)

- Connect to, change, and retrieve data from relational databases
- Developed in the 1970s, still going strong
- Many flavors

Interfaces

Jupyter Notebooks

- Streamlined document-centric interface for running and sharing code

IllumiDesk

- Hosts Jupyter Notebooks in the cloud

Code-Focused Text Editor

- Write text files in a code-native format
- **VS Code** is one of many that would work

Version Control

Git

- Distributed version tracking on any files
- Folder → “Repository”

GitHub

- Hosts Git repositories
- Collaborate and share code with others
- Backbone of the open source community
- Your Data Science portfolio!

Versioning

Anaconda

- Package management and deployment
- Designed with Data Science in mind
- Create and share environments

Python Package Index (PyPi)

- Database of public Python libraries
- Package installer (pip)
- Not everything is on Anaconda

Now:
Time to Get Started!