Intro to Data Science
Cohort 2

Jonathan Joa

Scott McAllister

Table of Contents

1/ What is Data Science?

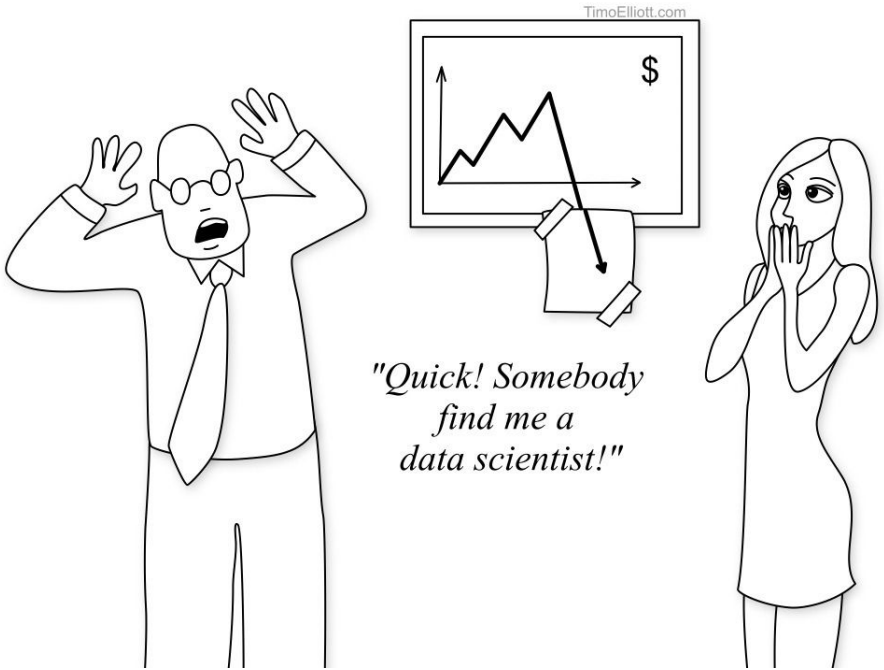
No, really. What *is* Data Science?

Tell us in your own words how you'd define **data science**.

Data Science

/ˈdadə,ˈdādə/ ˌˈsīəns/

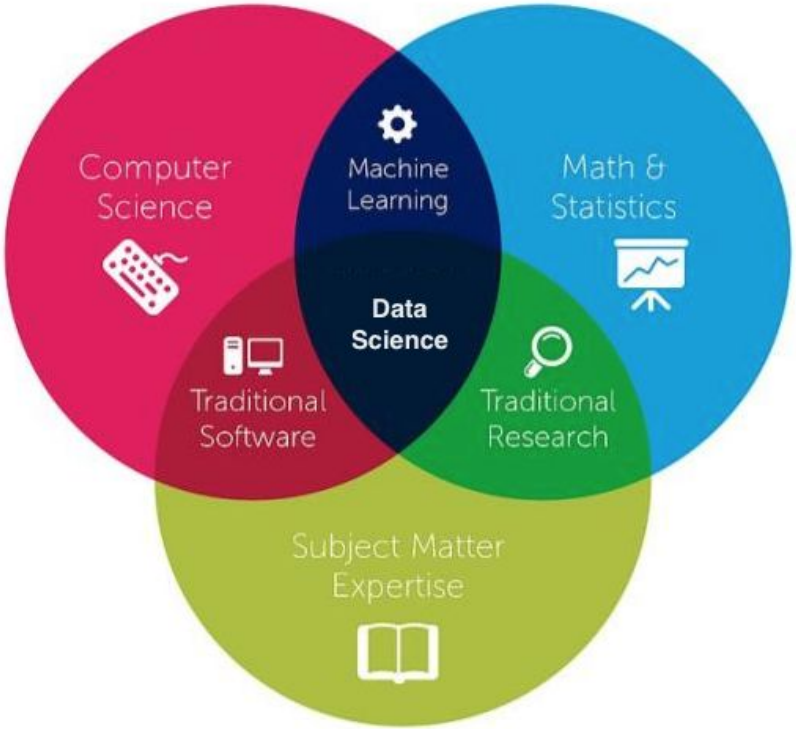
noun:
An interdisciplinary field focused on extracting knowledge from data in various forms.



Data Science

/ˈdadə,ˈdādə/ ˌˈsīəns/

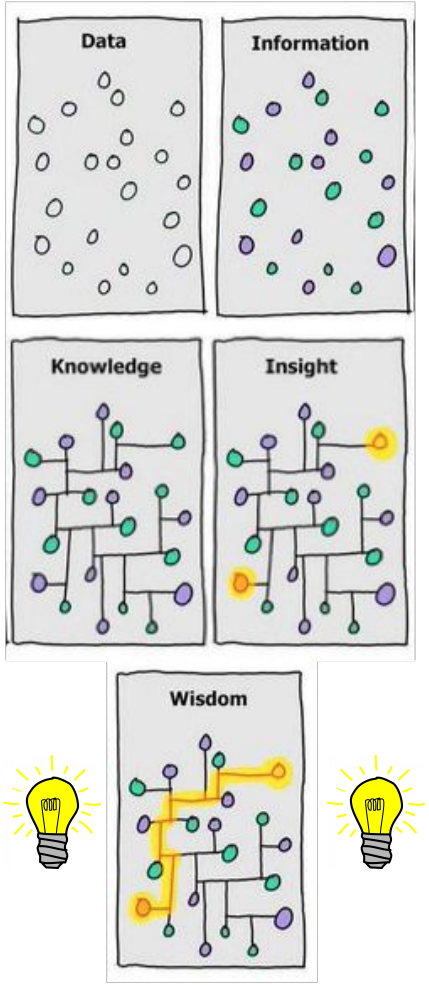
noun:
An interdisciplinary field focused on extracting knowledge from data in various forms.



Data Science

/ˈdadə,ˈdādə/ ˌˈsīəns/

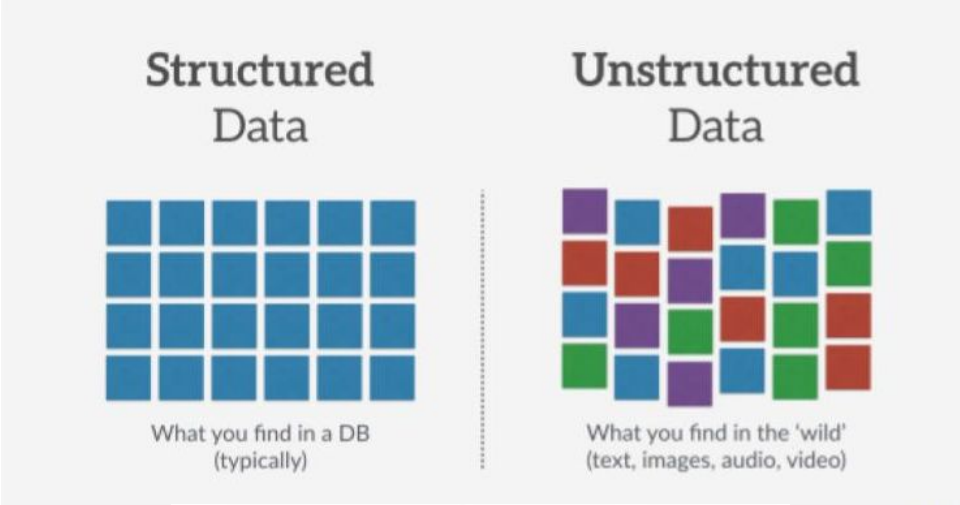
noun:
An interdisciplinary field focused on extracting knowledge from data in various forms.

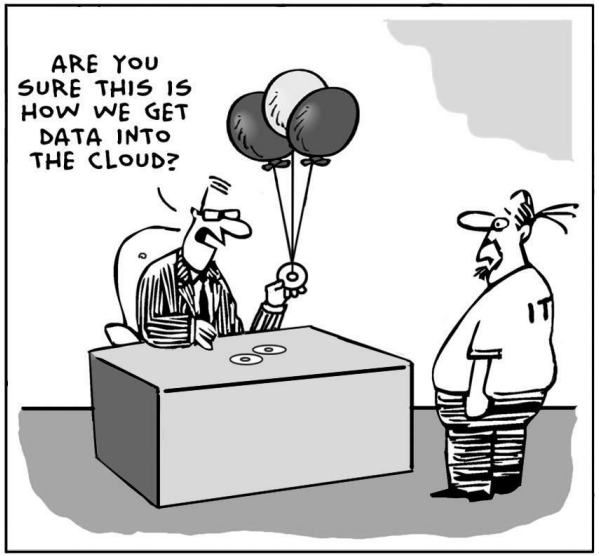


Data Science

/ˈdadə,ˈdādə/ ˌˈsīəns/

noun:
An interdisciplinary field focused on extracting knowledge from data in various forms.





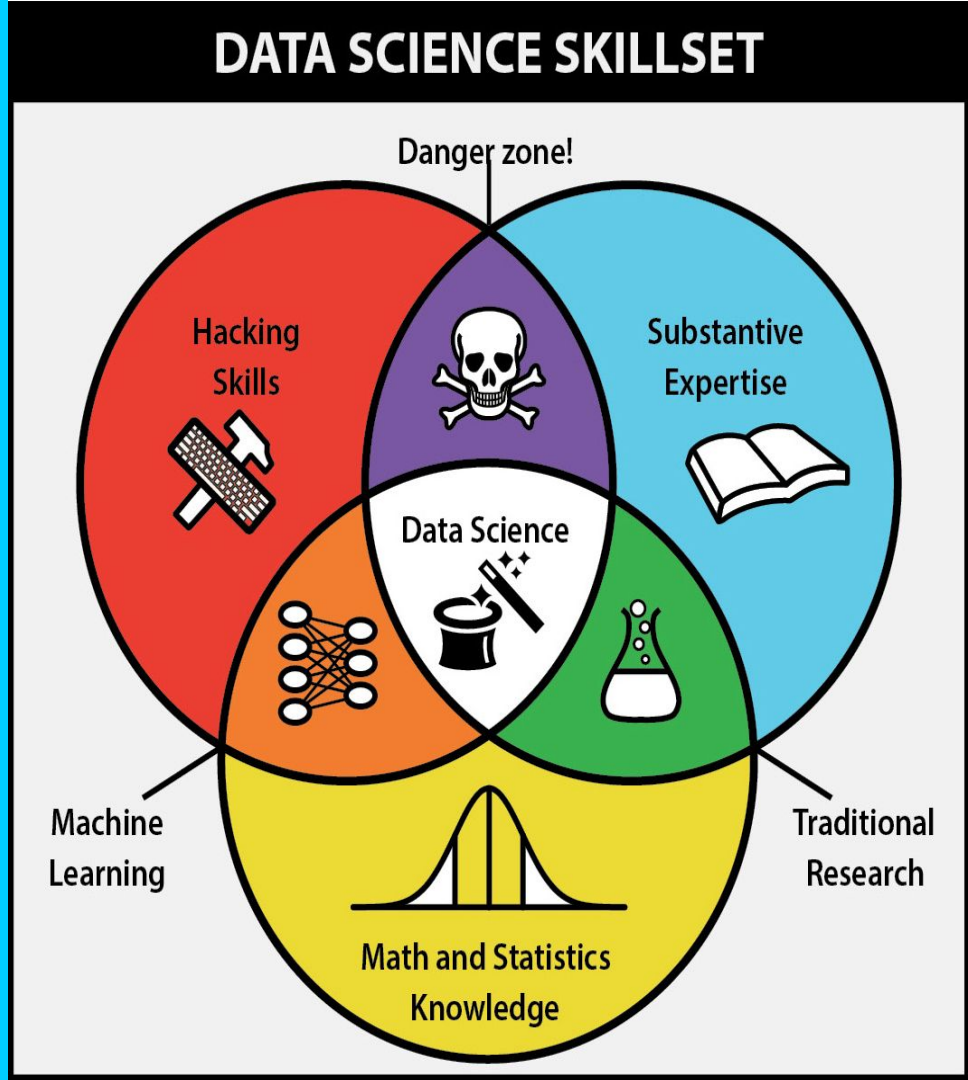
2/ Data Science Skills and Roles

Data Science Skills

- Data Science is the intersection of many fields.

- Note the danger zone. How does one end up there?

Click here



Data Scientist Skill Exercise

1. For your current position, count the number of competencies you currently utilize or would like to utilize for each skill set.

2. Sum your scores for each skill set.

3. Identify which of your skill sets have the highest scores and lowest scores.

4. Now that you've identified your strengths and weaknesses, how would you address them?

Hacking Skills
- Scripting language (e.g. R, Python)
- Database structures (e.g. SQL)
- Visualization tool knowledge

Math and Statistics
- Linear algebra
- Probability theory
- Statistical modeling

Substantive Experience
- Domain knowledge
- Influence
- Data curious

Traditional Research
- Experimental design
- User Experience
- Story-telling

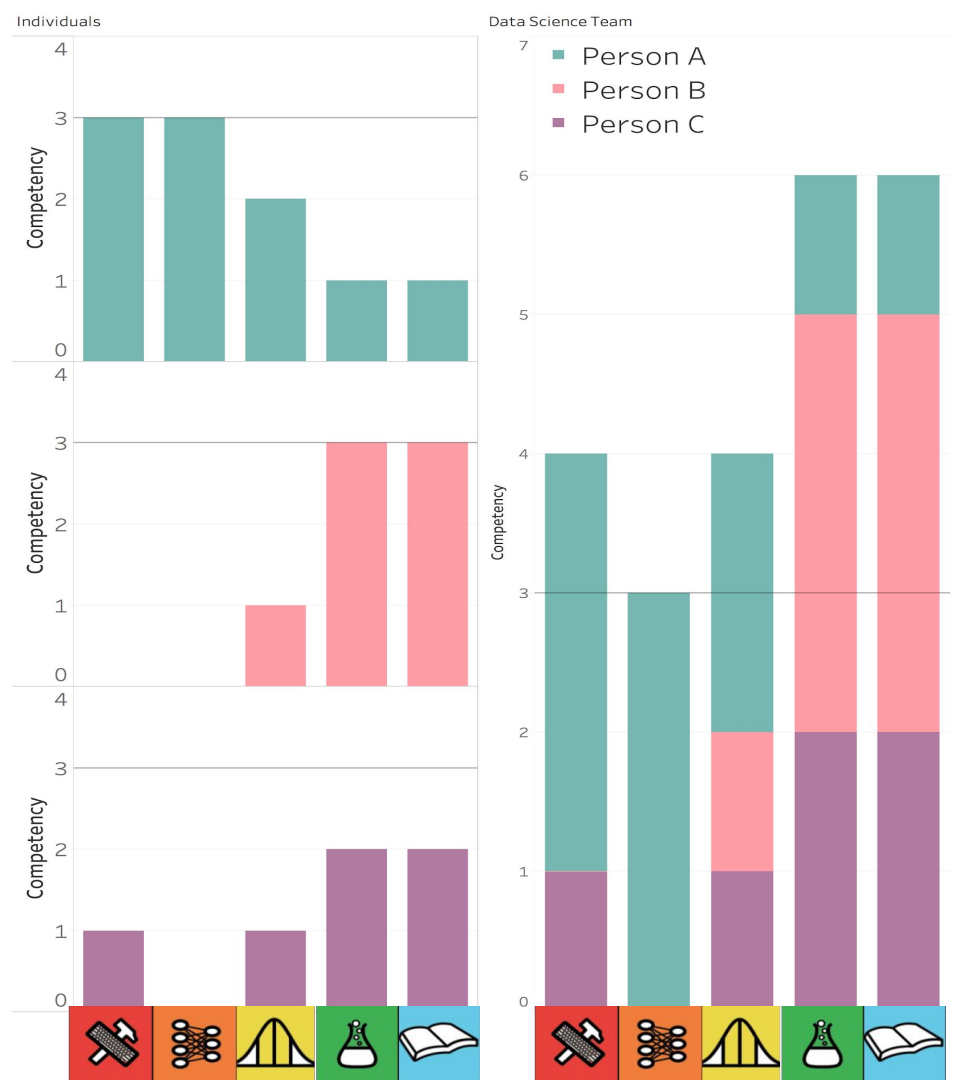
Machine Learning
- Supervised learning
- Unsupervised learning
- Cross validation

“The ideal data scientists aren’t just wunderkinds in advanced mathematics and statistics, they’re creative, non-linear thinkers with excellent communication skills….”

As much as you might want, you’ll never be that data science unicorn!

Why Data Science Teams?

- It’s tough to be an expert in everything.
- Diverse teams cover the spread.



Roles within a Data Science Team

- Skills match the job
- A successful team has members *outside your office*
- Cultivate relationships to bridge gaps

Role					
Database Admin	3	1	1		
Software Engineer	3	1	1	1	1
Analyst			1	3	2
Graphic Designer	2			2	2
Researcher	1		1	3	2
Systems Admin	3	1			2
Business Stakeholder			1	2	3
Data Scientist	3	3	3	1	2

Types of Data Scientists & What They Do

Data Analyst

- Export from SQL
- Excel or Tableau master
- Visualize data

Data Engineer

- Set up data infrastructures
- Clean, prepare and optimize data for consumption

Machine Engineer

- Applying formal mathematics & statistics
- Offering data-driven products

The Generalist

- A *little* of everything
- Automate mundane tasks
- Dashboarding

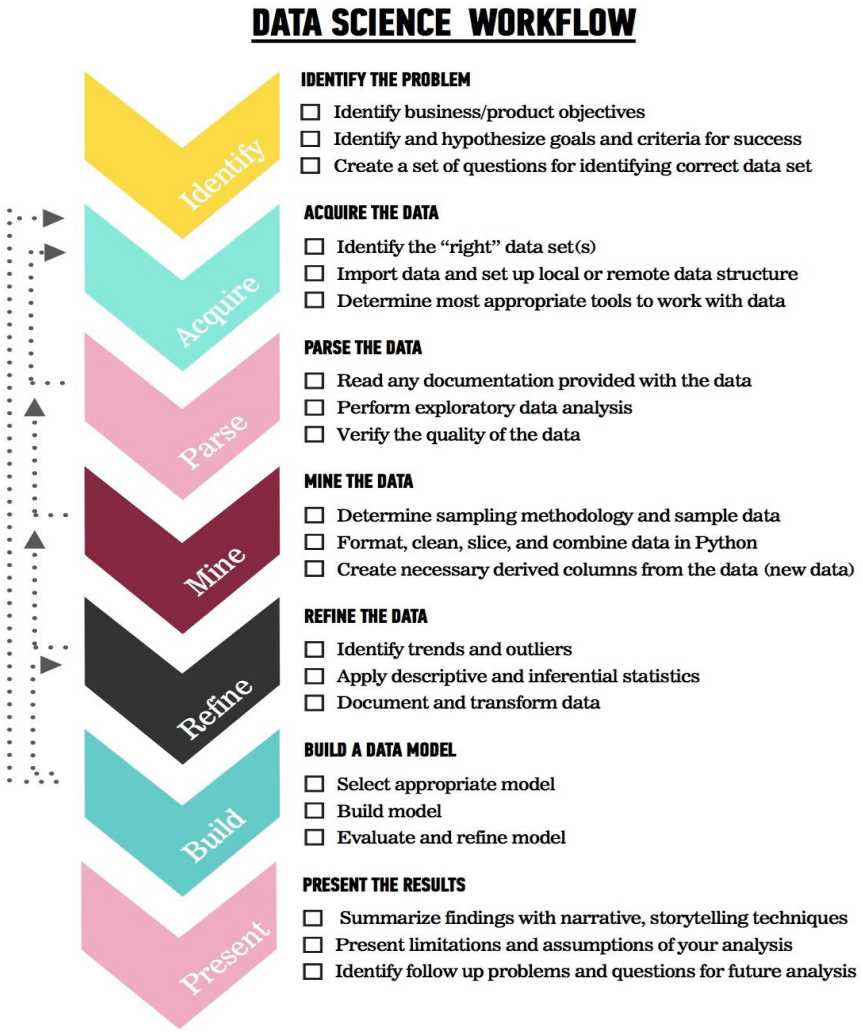
3/ The Data Science Workflow

The Workflow

There is no single template for solving a data science problem. The workflow changes with the dataset and the problem.

So here's a template...

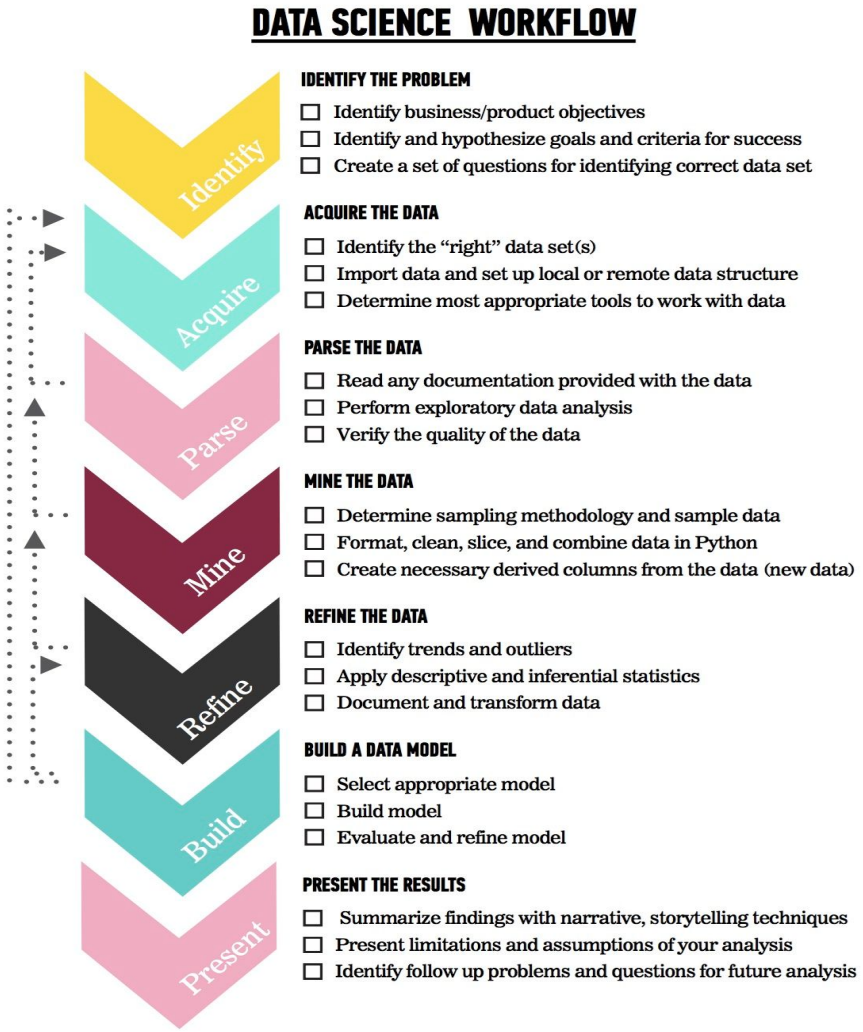
¯_(ツ)_/¯



The Workflow

Step 1: Identify the Problem

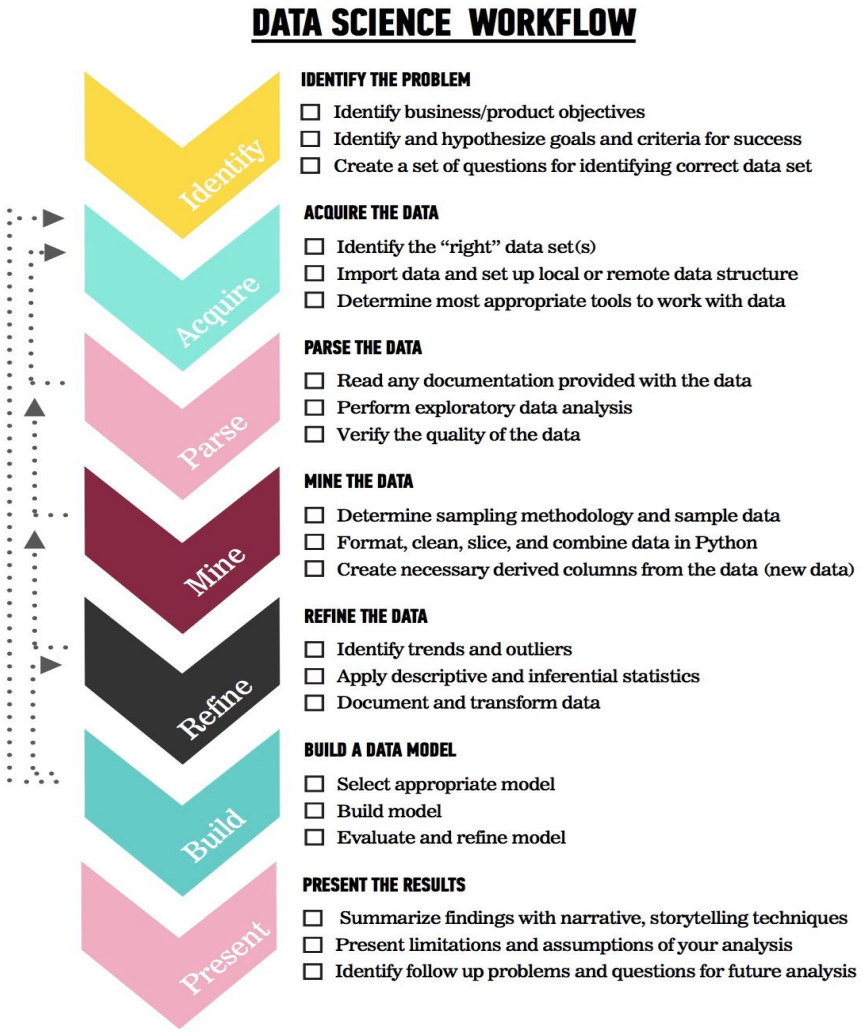
- Identify objectives
- Hypothesize criteria for success
- Create criteria for identifying the right data



The Workflow

Step 2: Acquire the Data

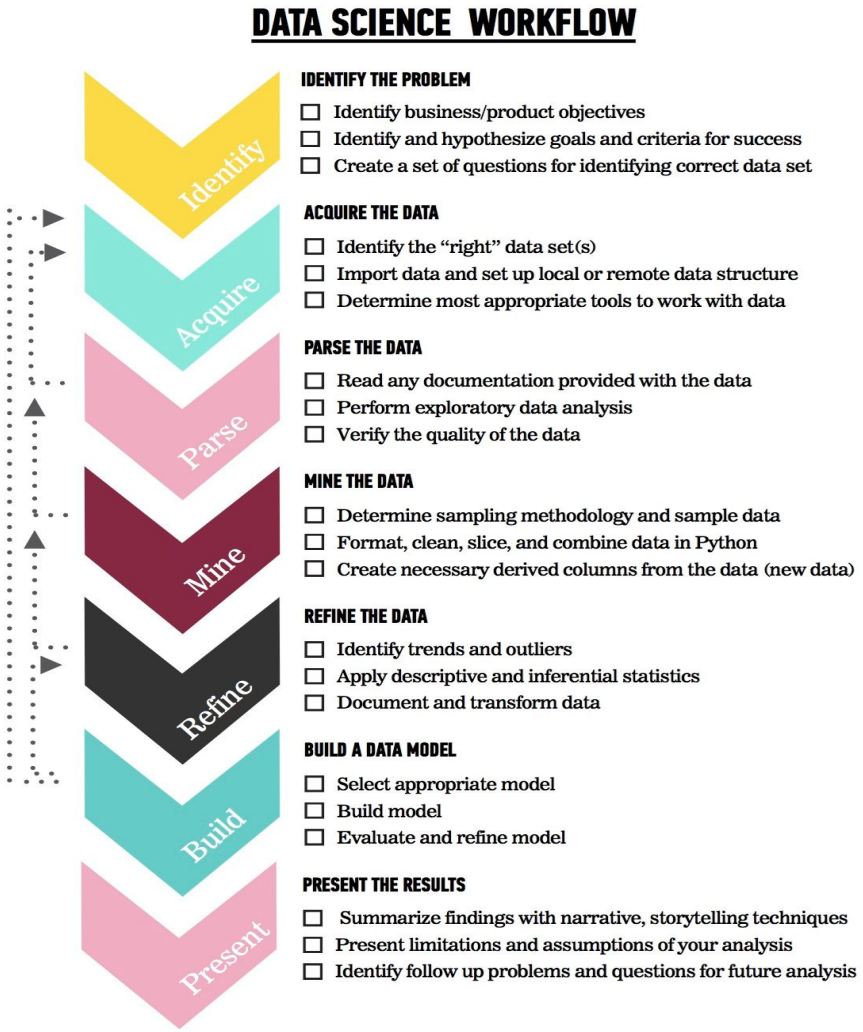
- Identify the “right” data set(s)
- Import the data and set up local/remote environment
- Determine the appropriate tools to work with the data given your role



The Workflow

Step 3: Parse the Data

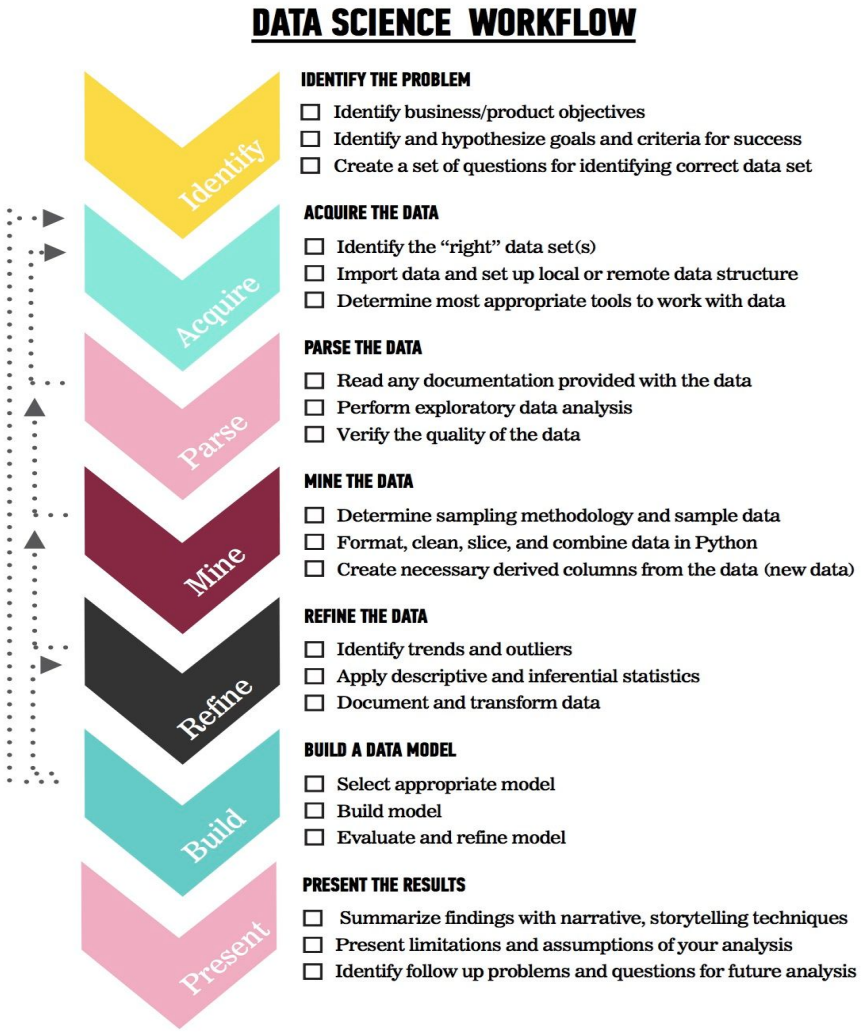
- Read the data documentation
- Perform exploratory data analysis
- Verify the quality of the data



The Workflow

Step 4: Mine the Data

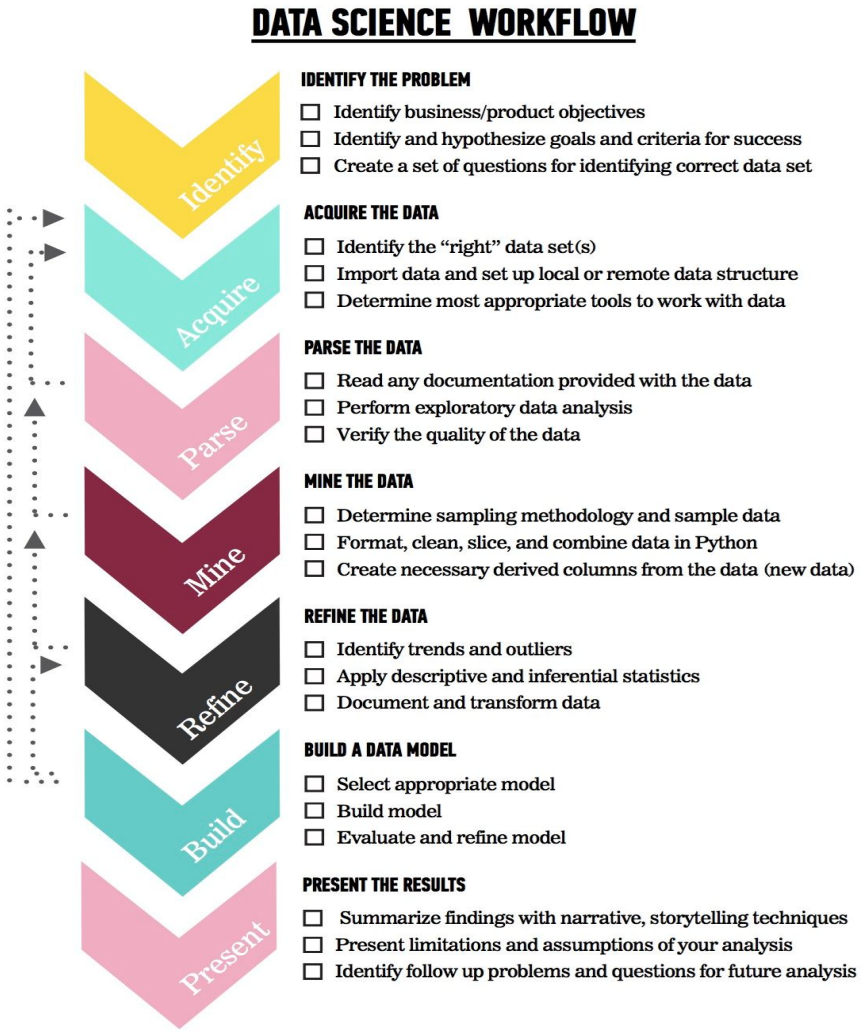
- Determine sampling methodology
- Munge the data
- Create derived columns



The Workflow

Step 5: Refine the Data

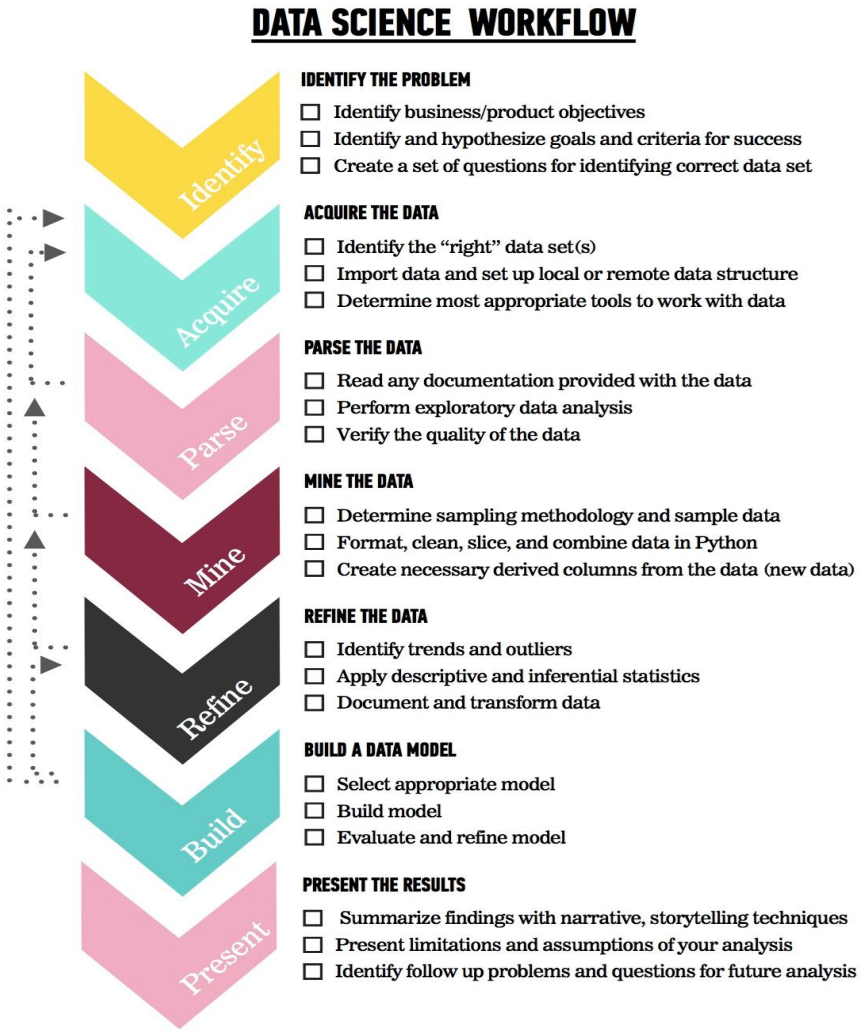
- Identify trends, outliers and missing values
- Apply descriptive and/or inferential statistics
- Document and transform the data



The Workflow

Step 6: Build a Model

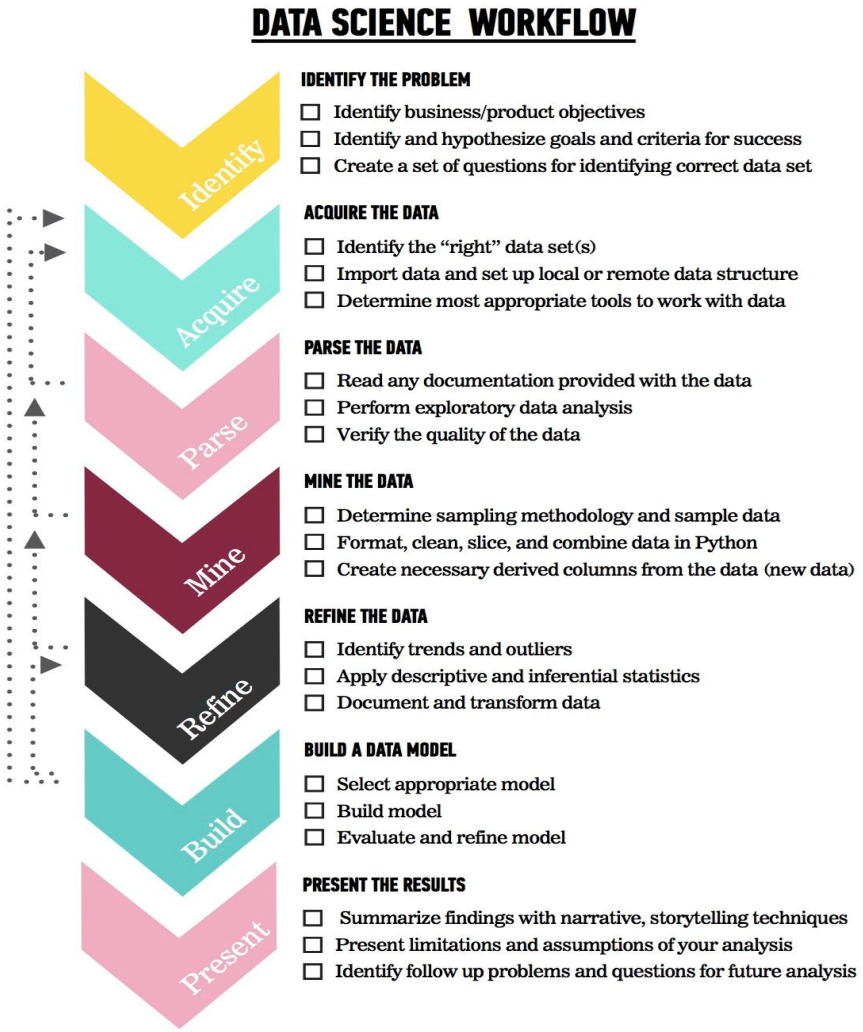
- Select appropriate model
- Build model
- Evaluate and refine the model



The Workflow

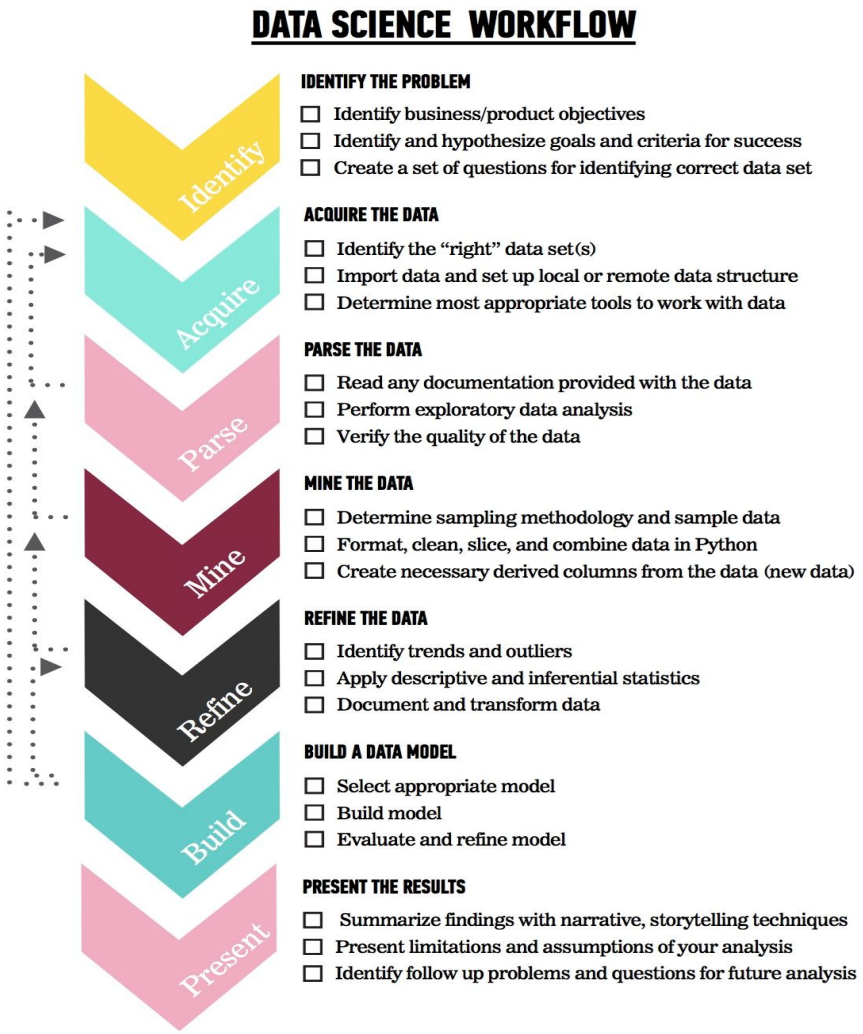
Step 7: Present the Results

- Summarize findings with narrative, storytelling techniques
- Present limitations and assumptions
- Chart future analyses



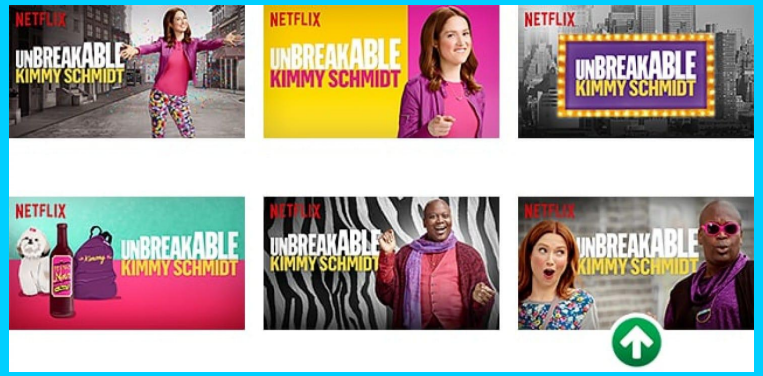
Workflow Exercise

- Do you see this workflow in your office/team?
- Where does *your* work fit into this workflow?
- Is there anything obstructing your movement through this workflow?



4/ Data Science Inside and Outside GSA

Netflix: Perfecting Promotional Artwork



The Data

- Session-level user information
- Subscriber and device details
- Historical Netflix user data
- Ad image data for “Unbreakable Kimmy Schmidt”

The Model

- A/B Testing, using recommendation algorithms (e.g. cosine and/or Jaccard similarity)

The Findings

- An image featuring a close-up of two characters showing silly expressions was the most popular.

The Application

- Nearly all promotional art goes through this process, sometimes increasing viewership by as much as 30%

OHRM: Modeling Bias in Performance Ratings

The Data

- Four years of performance ratings data
- Position-related data, such as grade, tenure, supervisory status, and job series
- Demographics, such as gender, age, race/national origin, and veteran status
- Derived data, such as such as the age difference between the rater and ratee

The Model

- Logistic regression, using the odds ratio to determine the likelihood of ratings

The Findings

- The analysis identified workforce attributes associated with higher performance ratings

The Application

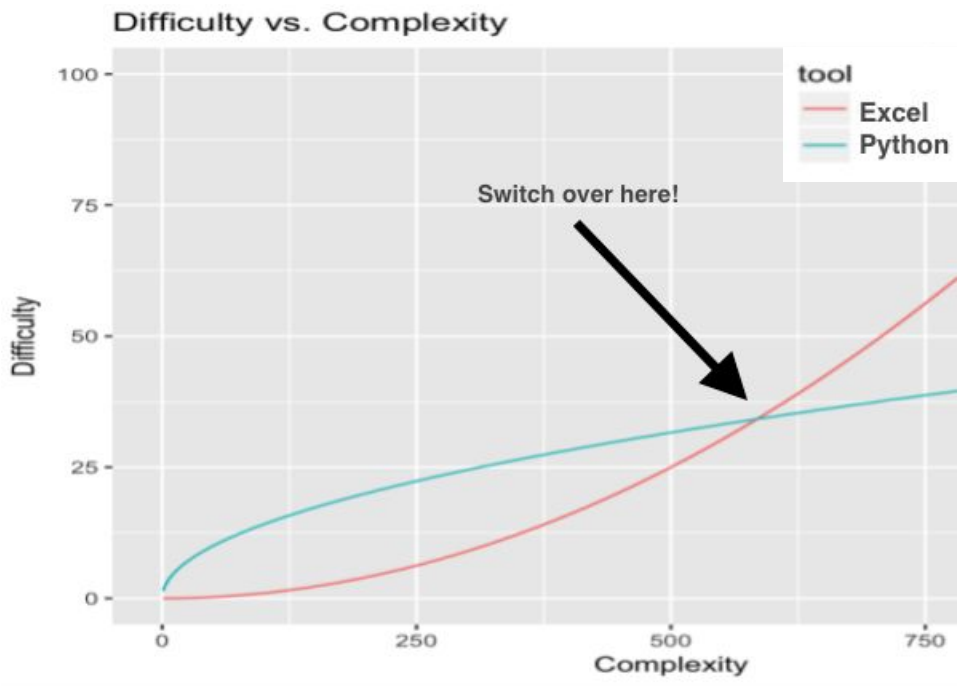
- The findings were incorporated into an unconscious bias training program for managers in partnership with the Office of Civil Rights

5/ Tools of the Trade

Python vs. Excel

- View these tools as complementary
  - Python actually has a package for working in both at the same time.

- When to make the switch:
  - Munging data
  - Automating tasks
  - Too much data
  - Machine learning



Python

[pahy-thon, -thuh n]

noun:
Python is an interpreted* high-level programming language for general-purpose programming. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

*Pretty much every Python implementation consists of an interpreter (rather than a compiler)

Python

[pahy-thon, -thuh n]

noun:
Python is an interpreted* high-level programming language for general-purpose programming. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

*Pretty much every Python implementation consists of an interpreter (rather than a compiler)

Compiled		Interpreted	
PROS	CONS	PROS	CONS
ready to run	**not** cross platform	cross-platform	interpreter required
often **faster**	inflexible	simpler to test	often **slower**
source code is **private**	extra step	easier to debug	source code is **public**

Python

[pahy-thon, -thuh n]

noun:
Python is an interpreted* high-level programming language for general-purpose programming. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

*Pretty much every Python implementation consists of an interpreter (rather than a compiler)

"Hello, World"

- C

```
#include <stdio.h>

int main(int argc, char ** argv)
{
   printf("Hello, World!\n");
}
```

- Java

```
public class Hello
{
   public static void main(String argv[])
   {
      System.out.println("Hello, World!");
   }
}
```

- now in Python

```
print "Hello, World!"
```

2

Monday, June 14, 2010

Python

[pahy-thon, -thuh n]

noun:
Python is an interpreted* high-level programming language for general-purpose programming. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

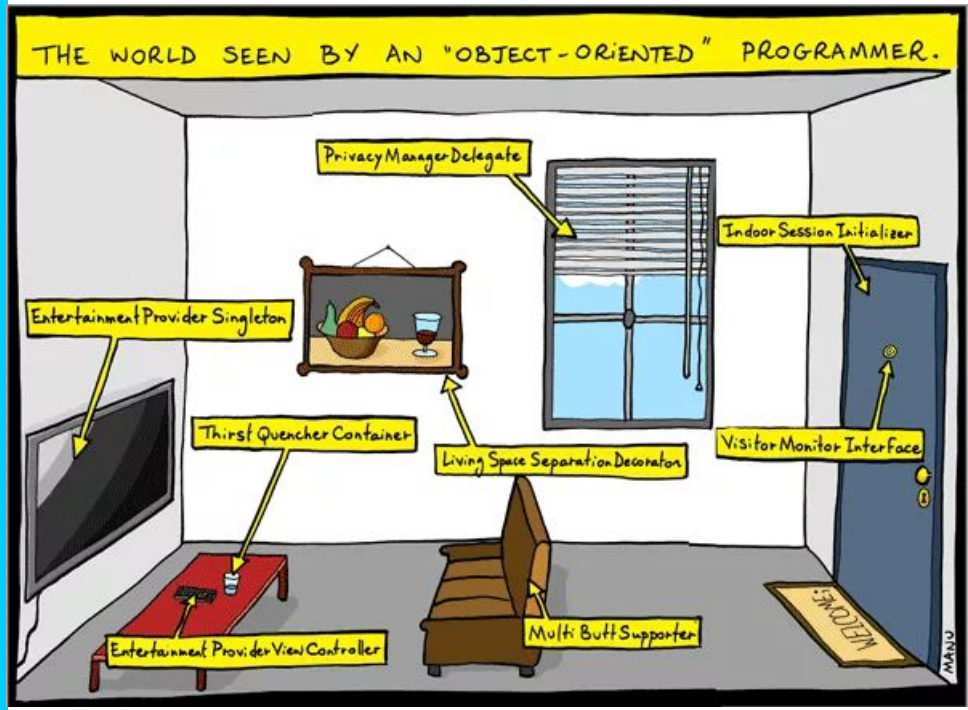
*Pretty much every Python implementation consists of an interpreter (rather than a compiler)

Python

[pahy-thon, -thuh n]

noun:
Python is an interpreted* high-level programming language for general-purpose programming. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

*Pretty much every Python implementation consists of an interpreter (rather than a compiler)

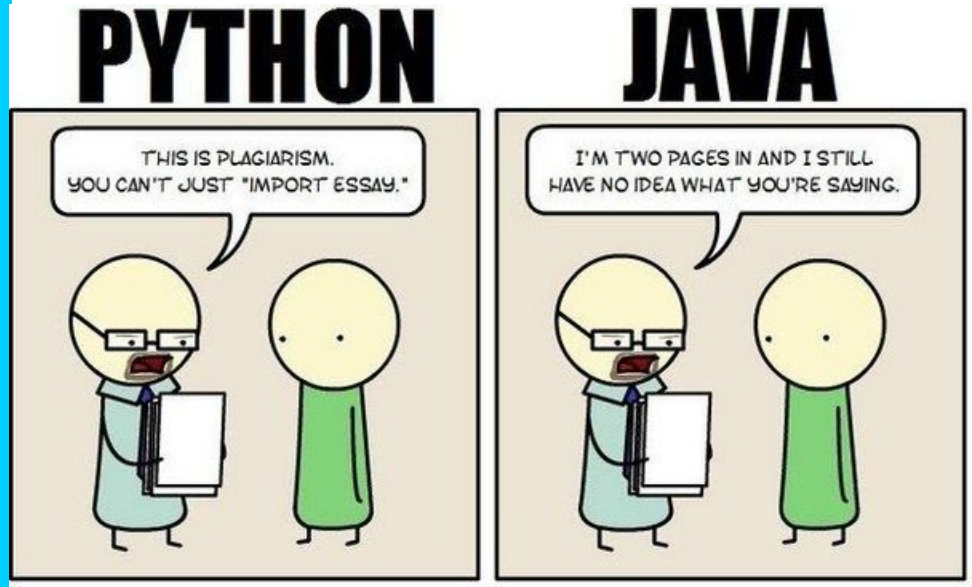


Python

[pahy-thon, -thuh n]

noun:
Python is an interpreted* high-level programming language for general-purpose programming. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

*Pretty much every Python implementation consists of an interpreter (rather than a compiler)



DSVD, Anaconda & Jupyter Notebook

DSVD (Data Science Virtual Desktop) is the virtual environment where you'll access all your data science tools.

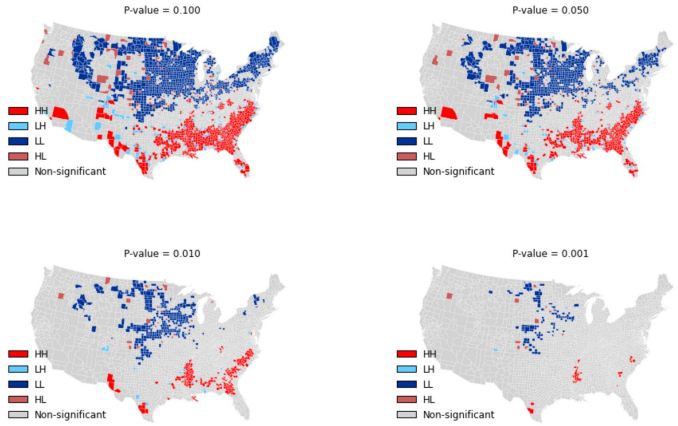
Anaconda is an open source Python distribution that simplifies module management.

Jupyter Notebook is an Interactive Development Environment that allows you to code as well as present.

jupyter nbviewer

JUPYTER FAQ

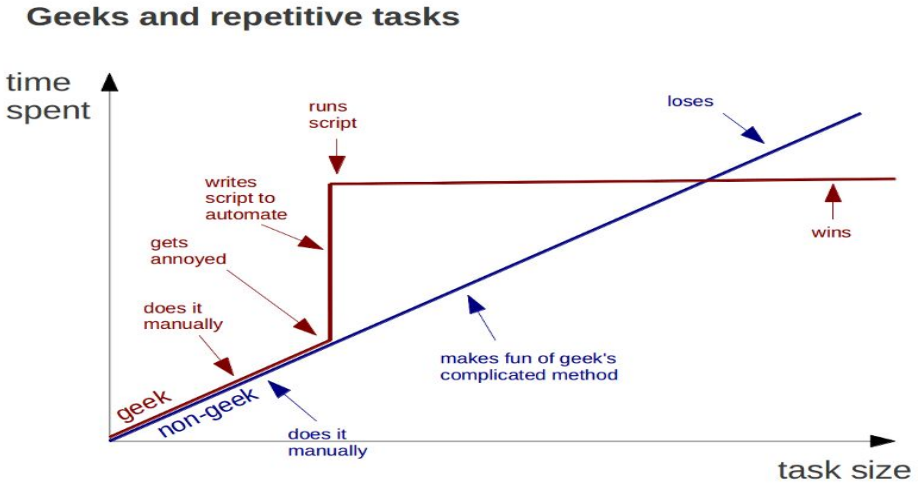
```
    ax.set_extent(extent, crs=ccrs.PlateCarree())
    ax.add_collection(polys)
    ax.outline_patch.set_visible(False)

    boxes, labels = maps.lisa_legend_components(lisa, p_thres=p_thres)
    plt.legend(boxes, labels, loc='lower left', frameon=False)
    ax.set_title('P-value = %.3f'%p_thres)

plt.show()
```



∞/ Unsolicited Advice

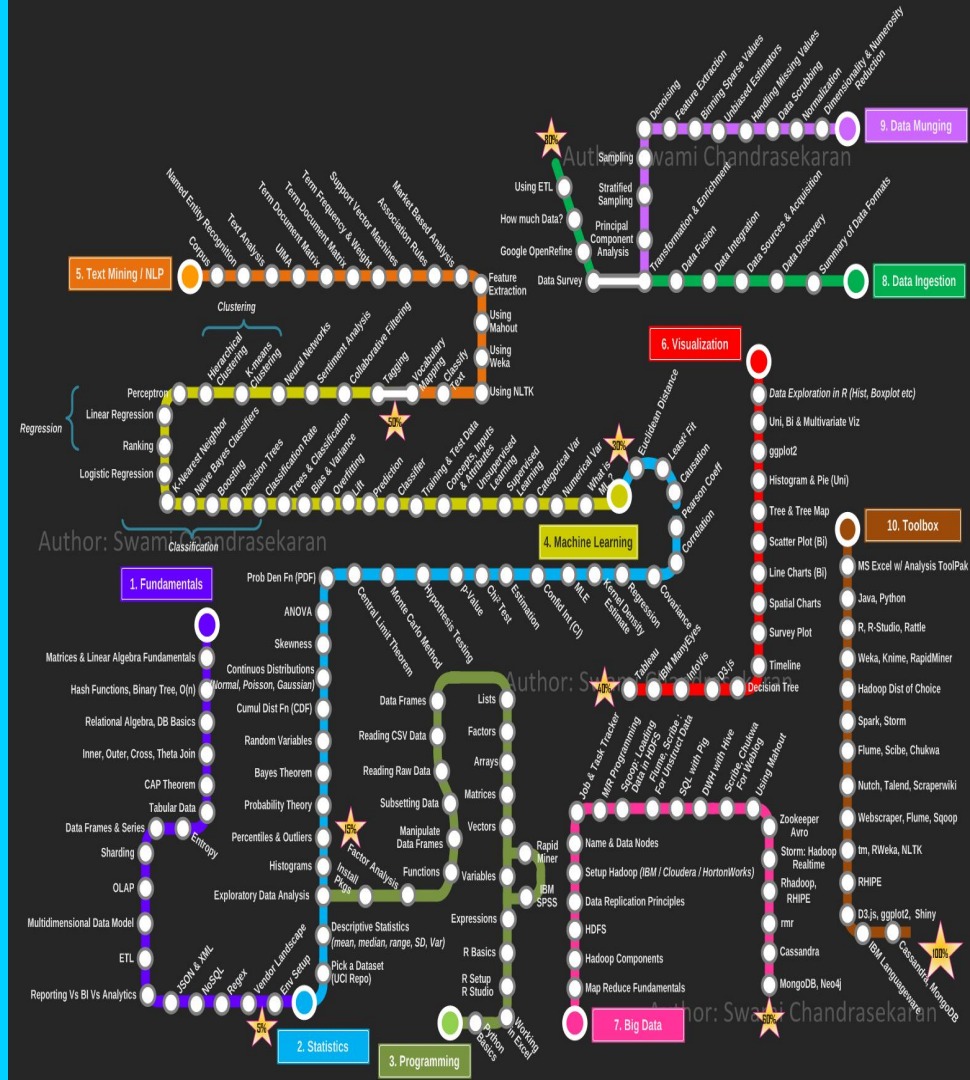
How to explain the benefits of automation

- When you start learning how to code, the ratio of reward to time invested (payout) is low.
- As the size/complexity of your task increases, the payout also increases.
- The key is to find and communicate these tradeoffs.



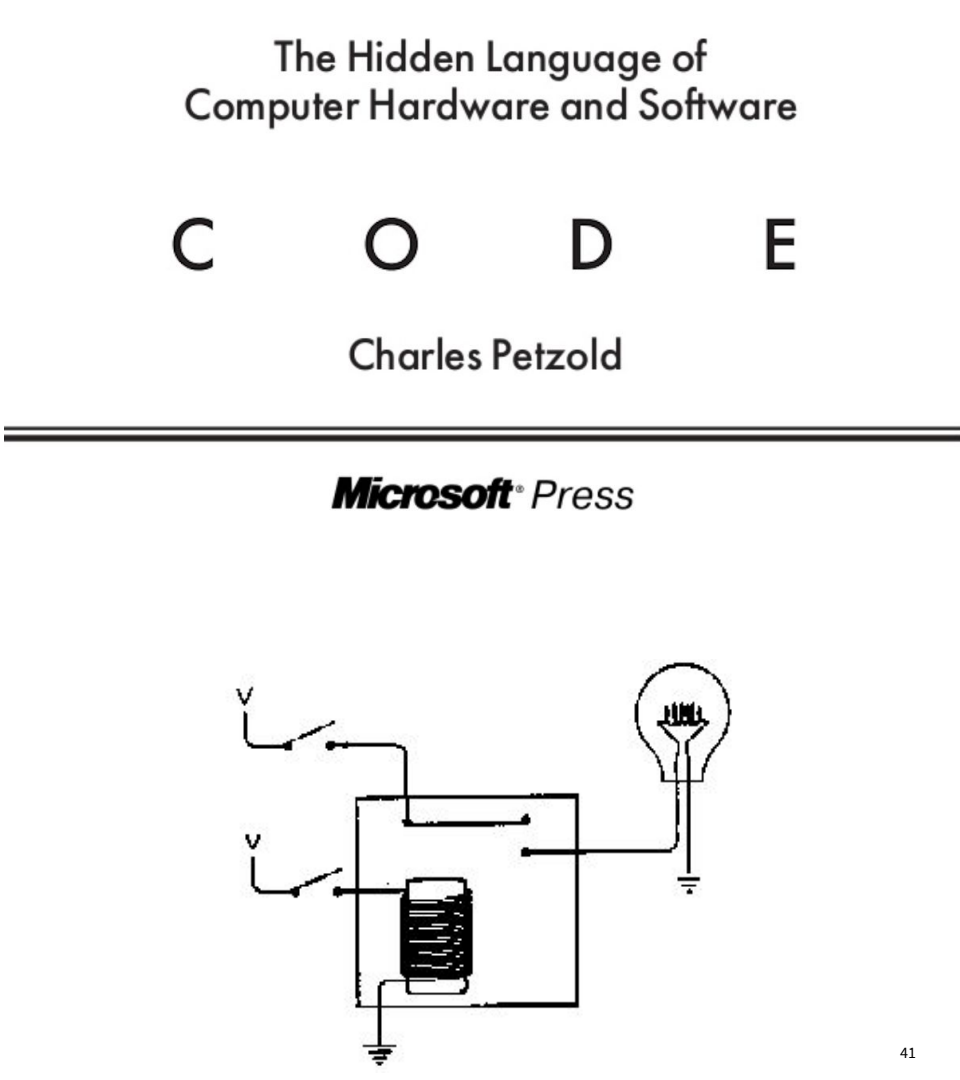
Only be confused by one thing at a time

- Learning data science is based on inductive chain learning: new concepts build upon mastery of old ones.
- Curiosity is the most important skill for a data scientist.
- When in doubt, google it (or go to stackoverflow).



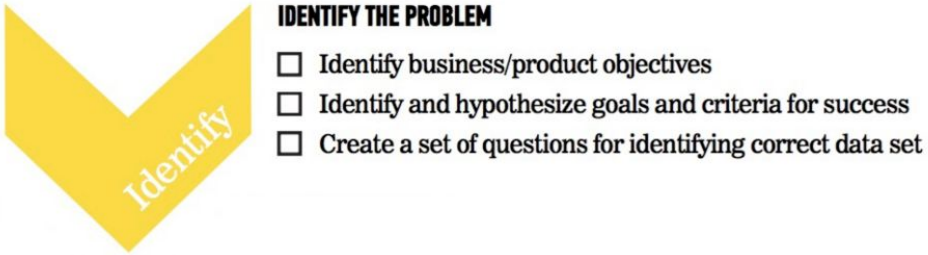
Read this book

- This book will demystify computers.
- This book will help you understand the language of computer science.
- This book is also very well written!



Problem Statement

- Homework!
- This curriculum will revolve around a capstone project within your office.
- Start with the first step of the Data Science Workflow.
- Be SMART! Specific, Measurable, Achievable, Realistic, and Timebound.



Thanks!

Contact

scott.mcallister@gsa.gov

jonathan.joa@gsa.gov

Work Together

github.com/GSA/training-pathway-data-practitioner