Shane Kimble

Beyond the Data Scientist: Overlooked Applications and Roles of Data Science

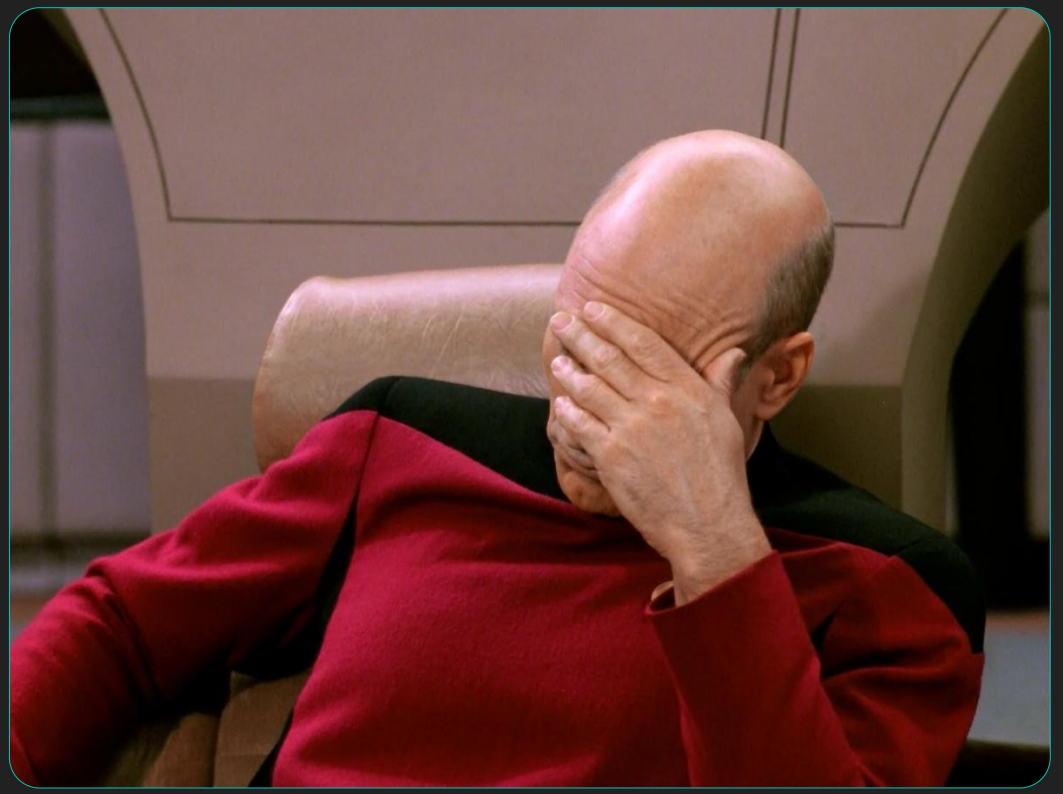
Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.

Overview

- Introduction
- Data Science
- Roles
- Applications

DISCLAIMER

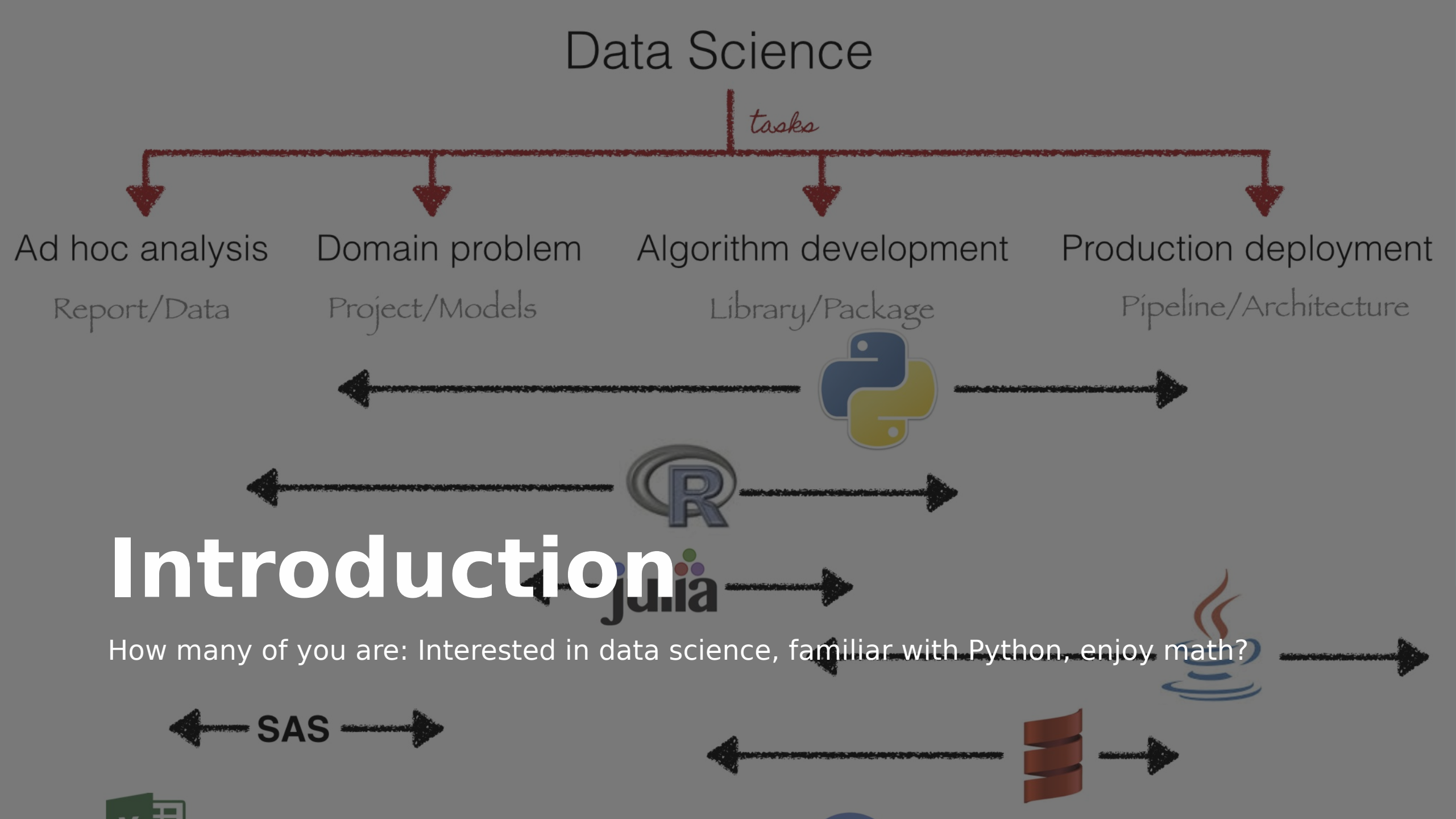
- This is NOT a presentation about data scientists!



Common Misconceptions

- The data scientist is the most important (King) role
- Data Science careers are hierarchical
- All data science and analytics involves coding
- Communication isn’t important in technical roles

PLEASE take data science career diagrams and flow charts with a grain of salt! They are NOT universal!

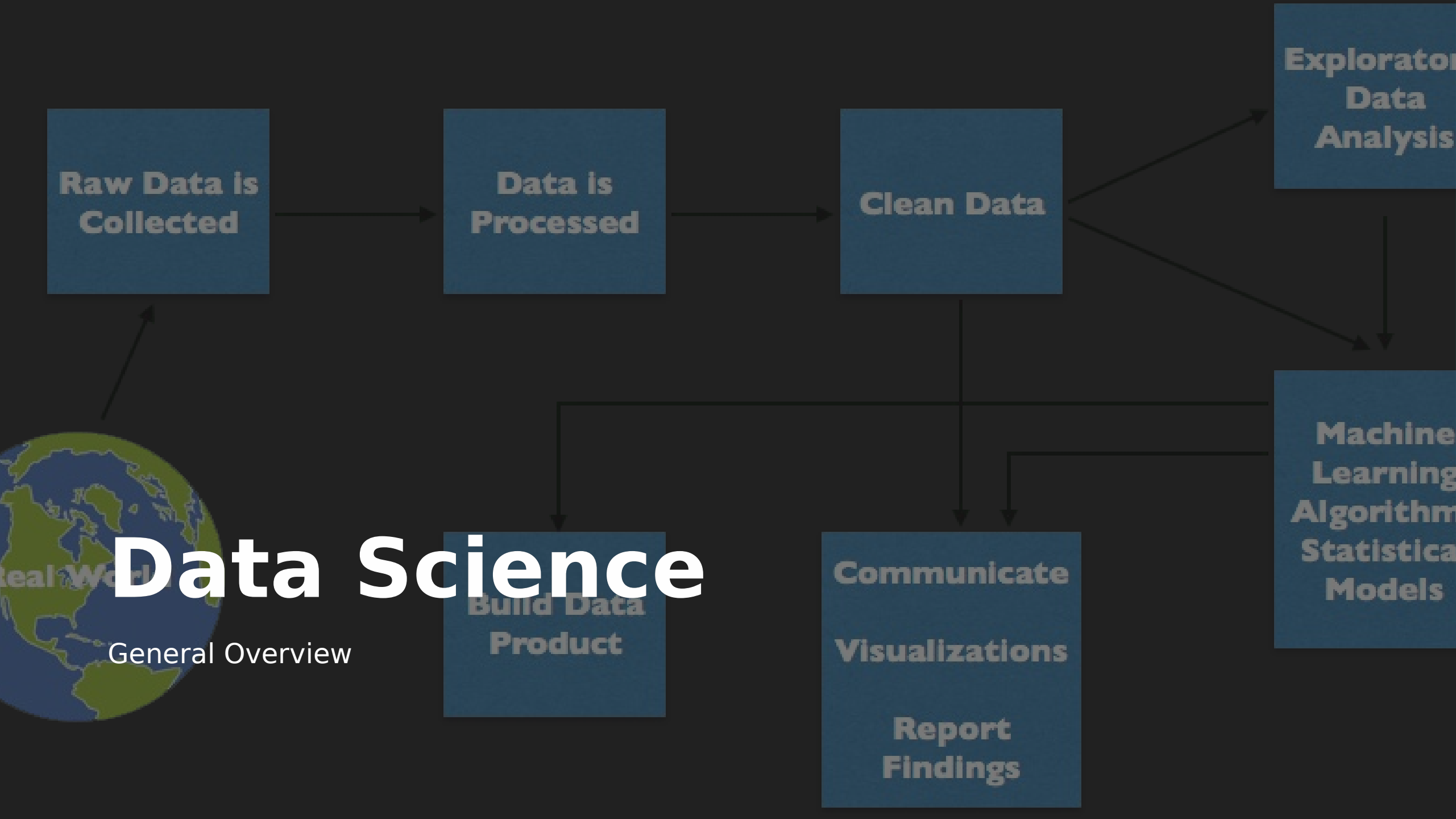


Introduction

How many of you are: Interested in data science, familiar with Python, enjoy math?

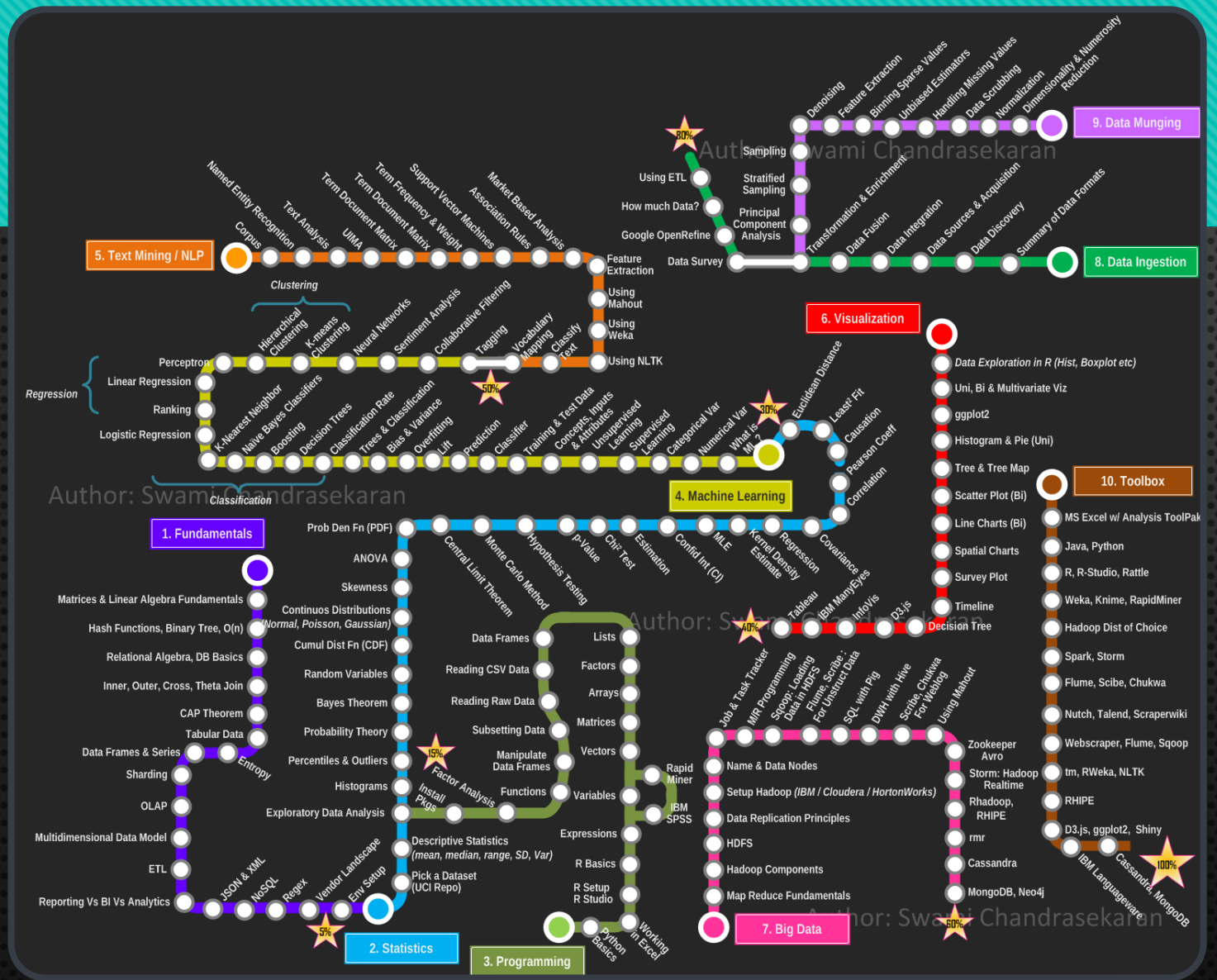
About Me

- Data Science Developer at EdjAnalytics
- Undergrad: Biology – Bellarmine University
- Grad: Computer Science – University of Louisville



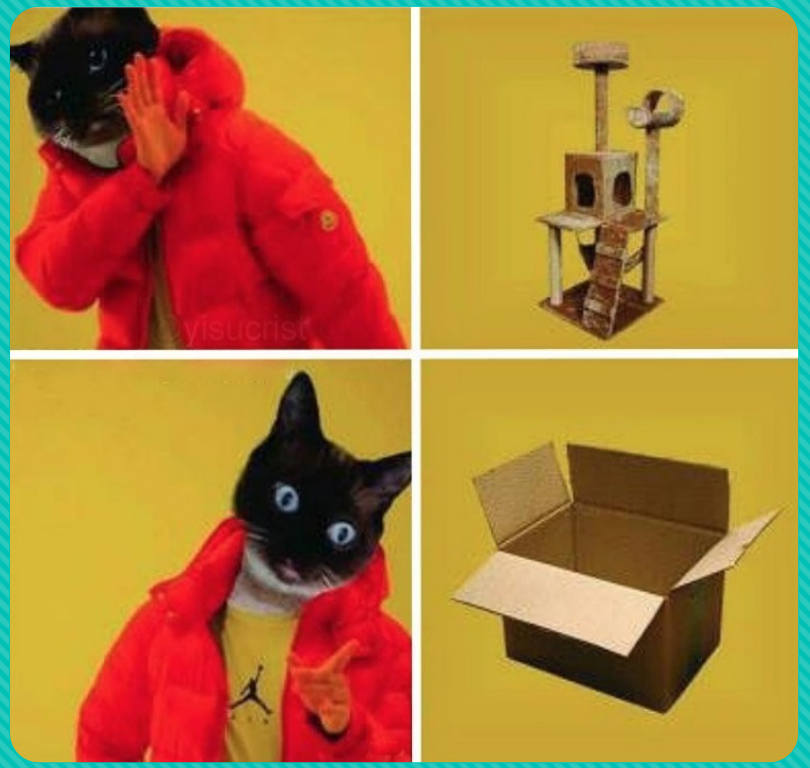
Data Science

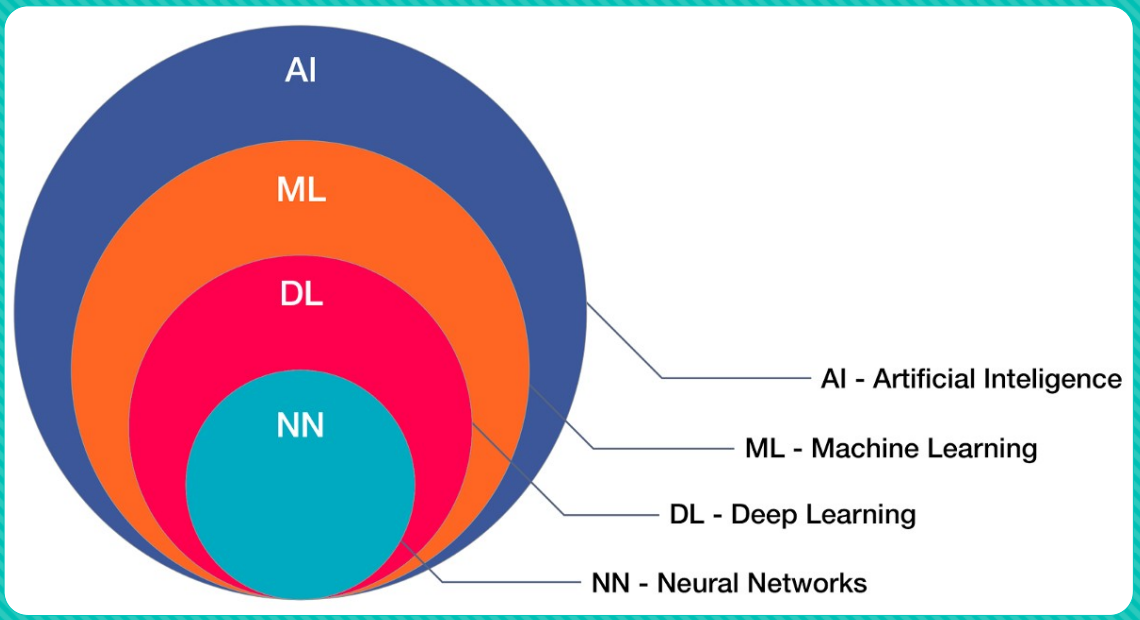
General Overview



Data Science doesn't have to be this terrifying

Don't get distracted





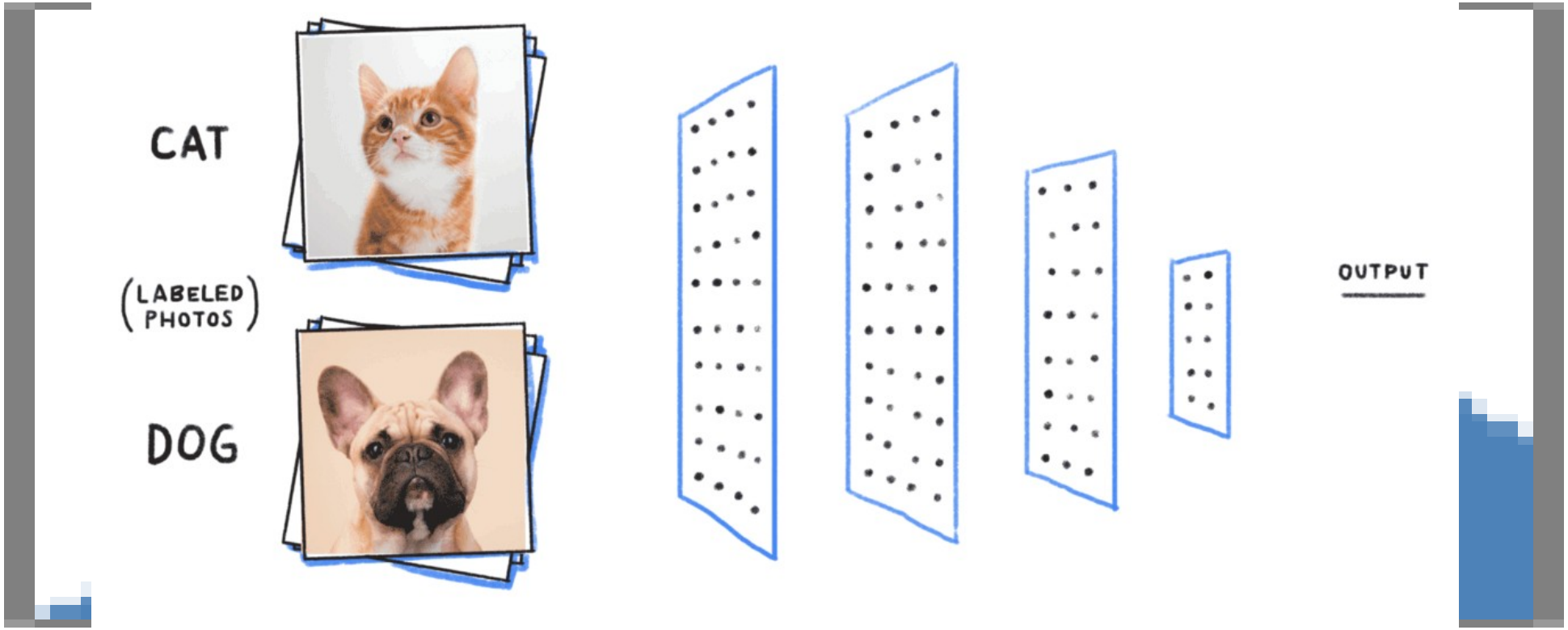
In a nutshell..

Data science is too broad. We'll focus on utility rather than intricacies and buzzwords.

Data Science Task

Examples include:

- Prediction – Generally Supervised
- Clustering – Generally Unsupervised
- (Anomaly) Detection
- (Dimensionality) Reduction
- (Feature) Selection
- (Reinforcement) Learning



Again, Utility

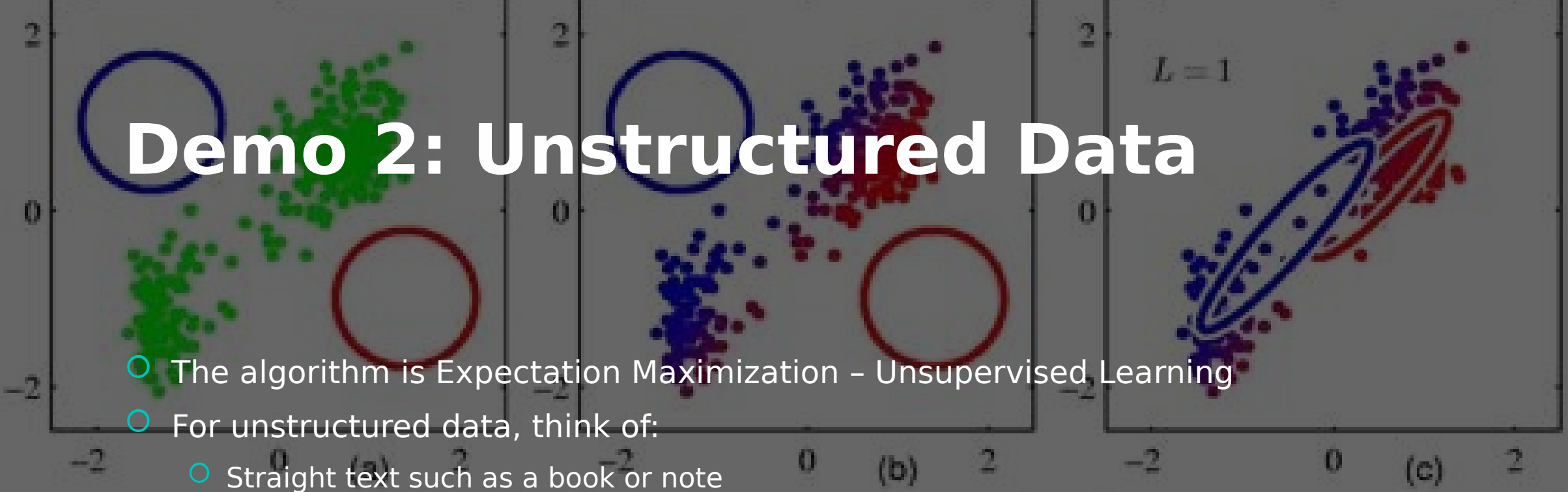
Does data science help your application or the problem you want to solve? It might not.

Demo 1: Structured Data

- The algorithm is Naïve Bayes – Supervised Learning
- For structured data, think of:
  - Excel spreadsheet
  - Feature = Variable = Column
  - Predict column y using columns X

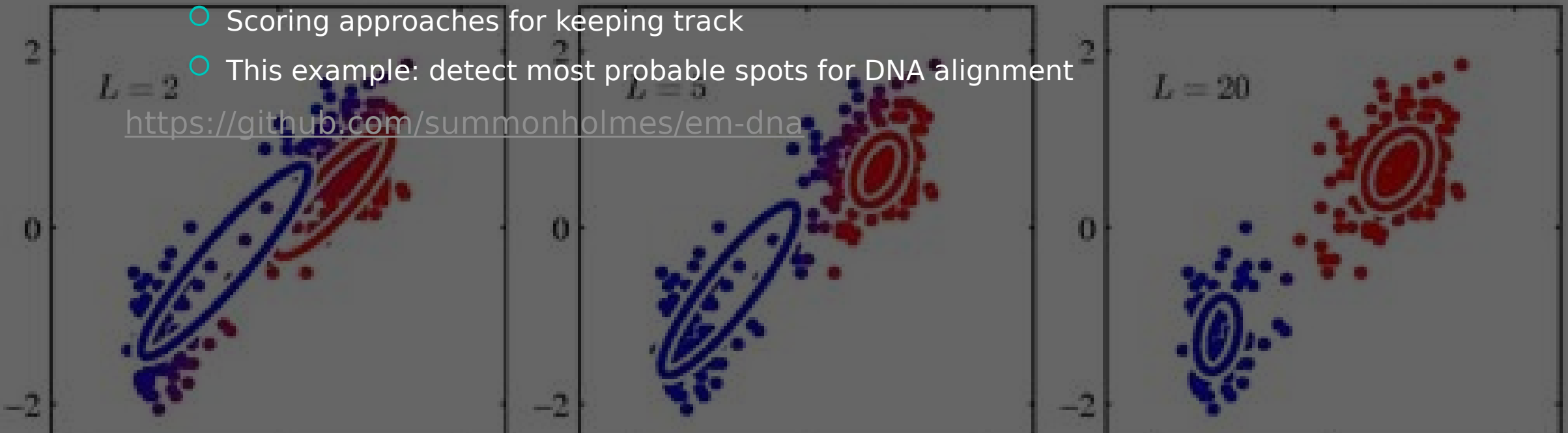
https://github.com/summonholmes/naive_bayes_pandas

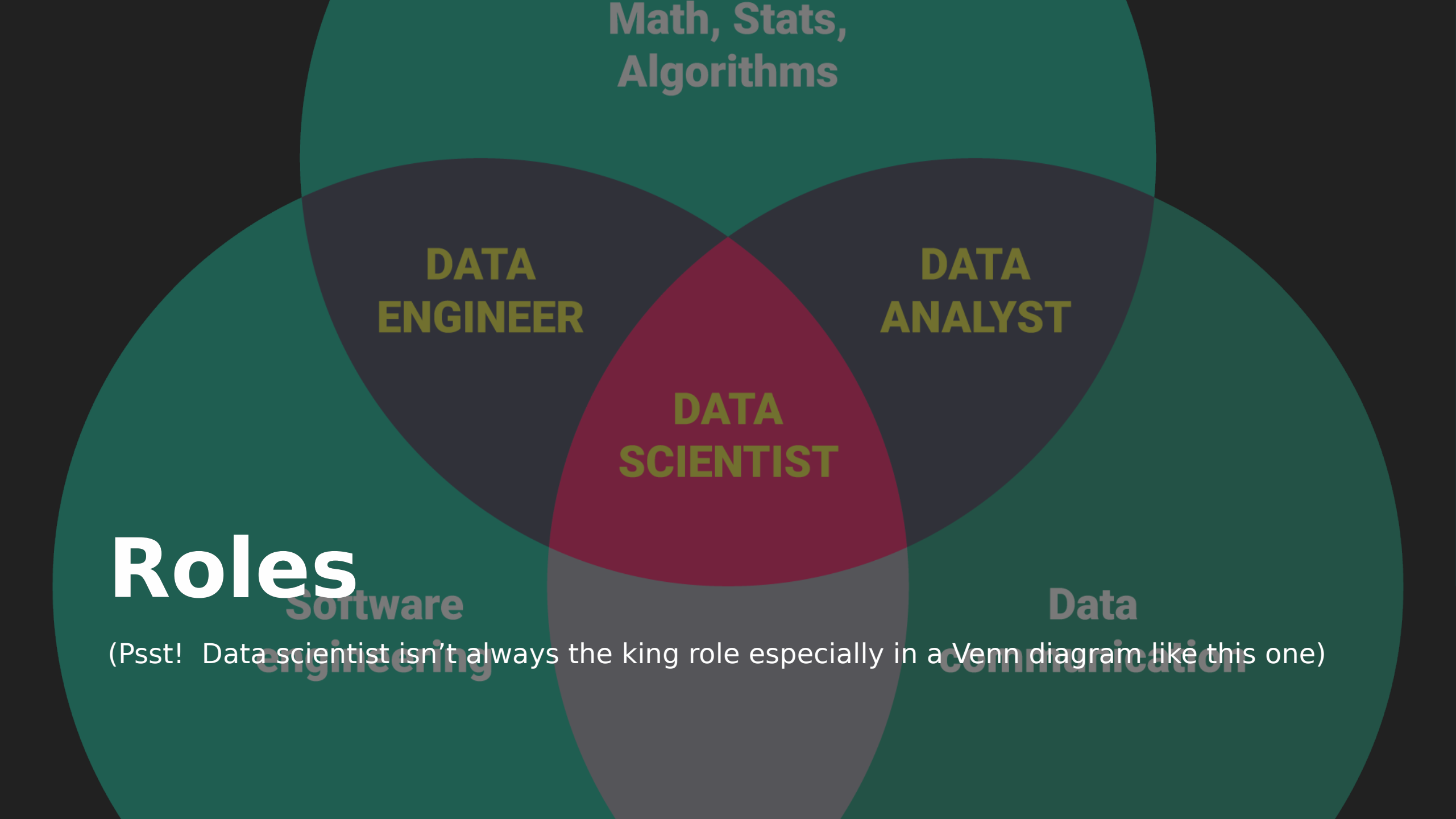
Demo 2: Unstructured Data



- The algorithm is Expectation Maximization – Unsupervised Learning
- For unstructured data, think of:
  - Straight text such as a book or note
  - Scoring approaches for keeping track
  - This example: detect most probable spots for DNA alignment

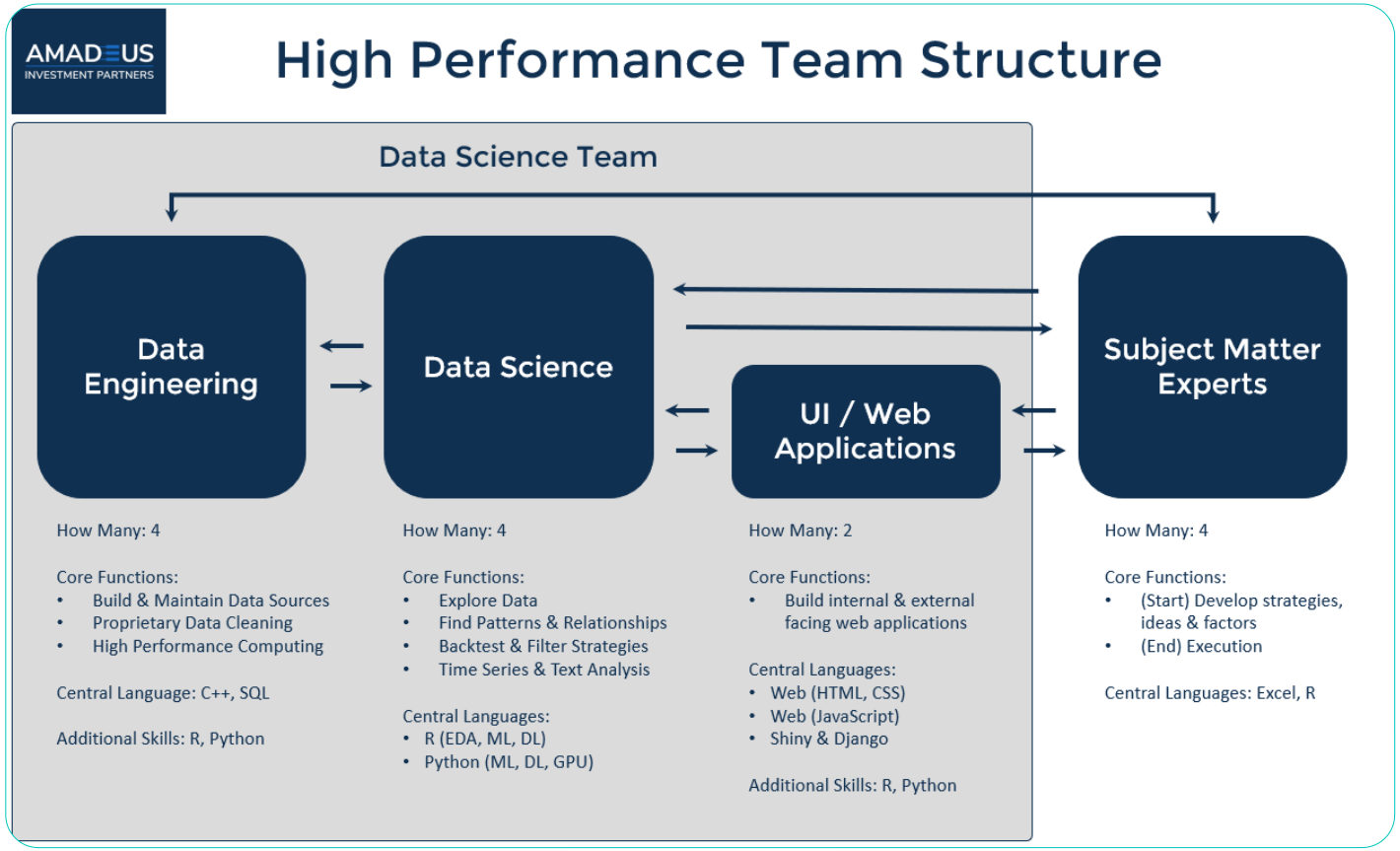
https://github.com/summonholmes/em-dna





Roles

(Psst! Data scientist isn’t always the king role especially in a Venn diagram like this one)

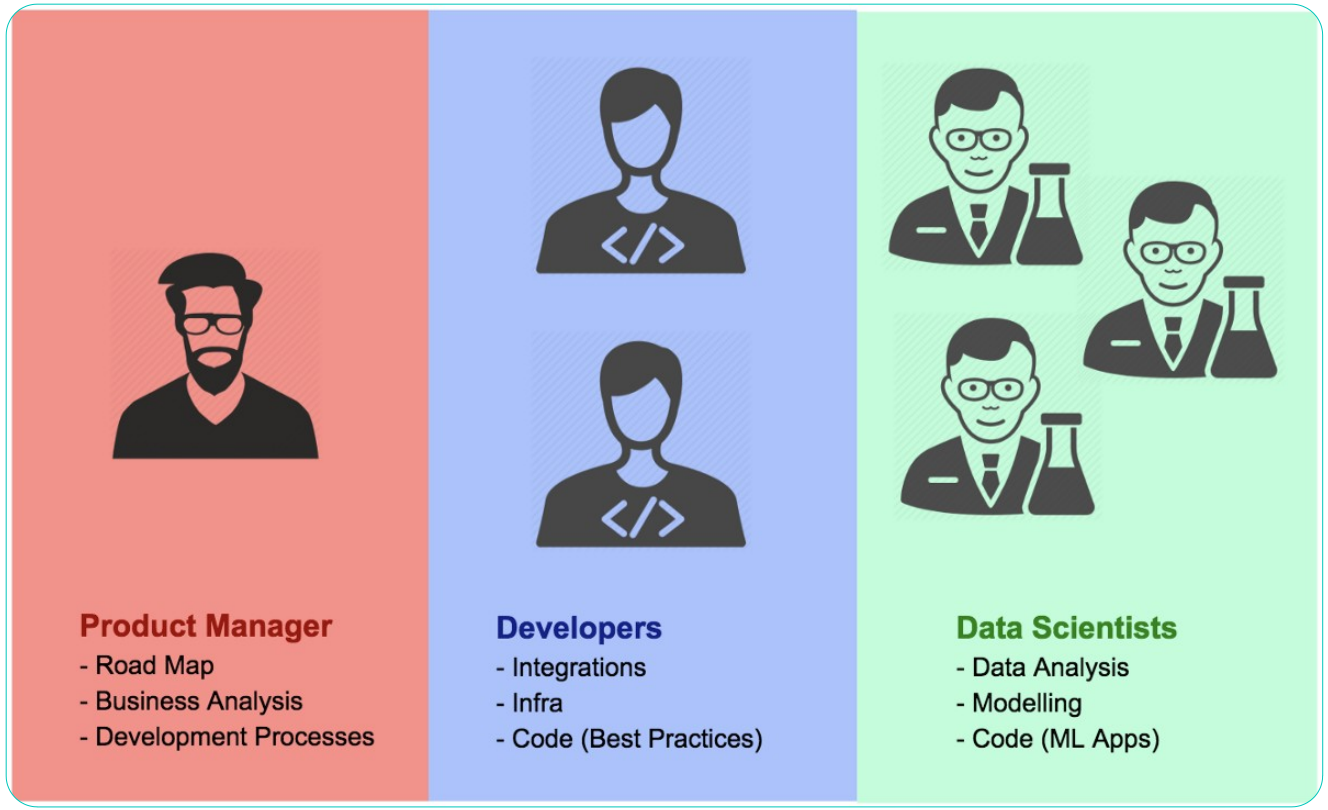


Data Science Project Overview

There’s a lot more to data science than data science!

Team Structure

- If a data science team doesn’t exist, IT may do data science by using enterprise software and contractors



Data Analyst

- Collect, process, and perform statistical analyses of data
- Sometimes thought of as a junior data scientist
- Very broad technical requirements:
  - May or may not know a combination of Excel, SQL, Python, R, Tableau, etc.

Data Scientist

MACHINE LEARNING

- Able to navigate through all phases of the data science process
- Well versed in machine learning and programming
- Standardized technical requirements:
  - SQL, Python, R, etc.

"What is a 'Data Scientist'? An analyst who lives in California."

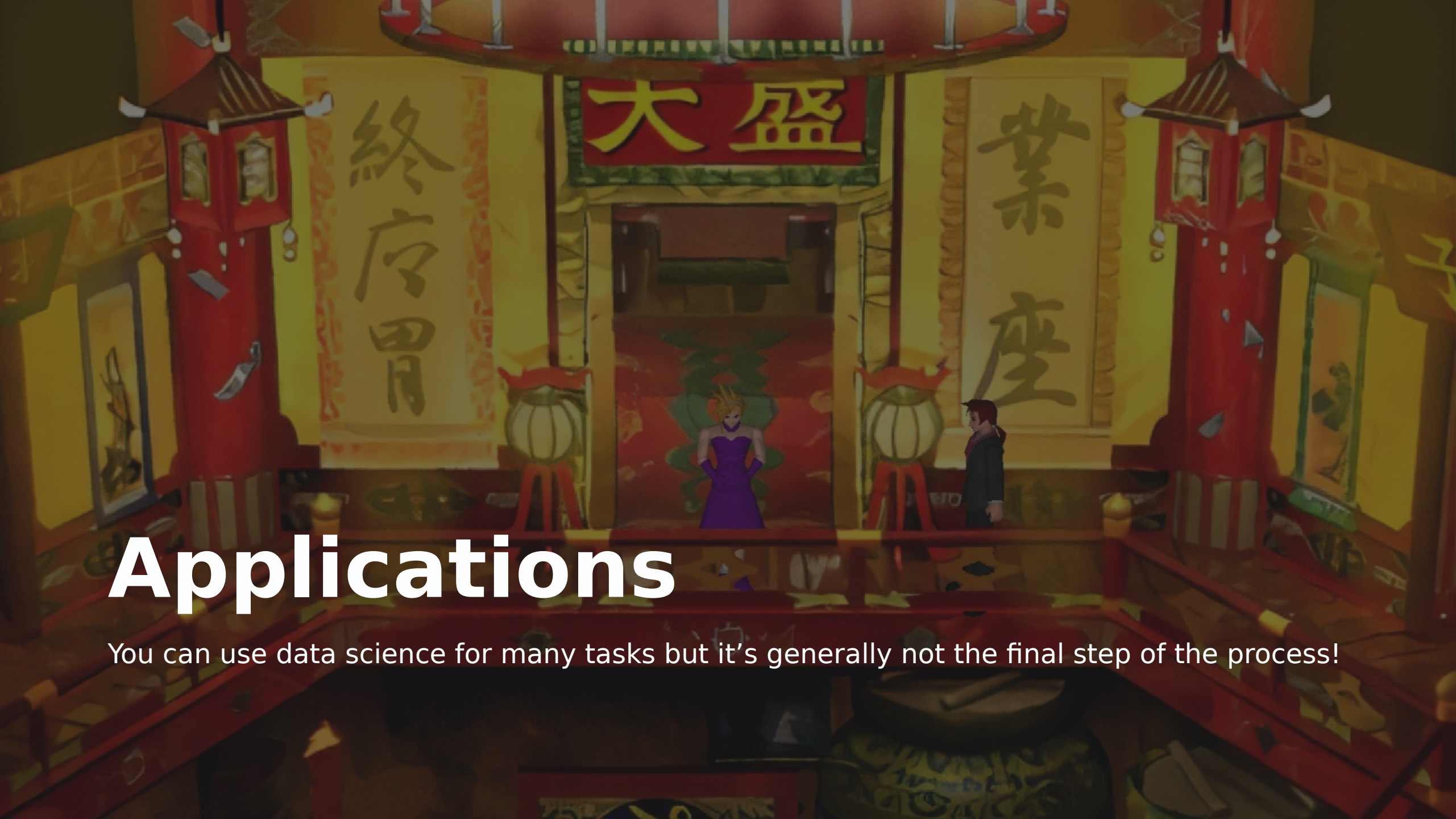
Product Manager

- Operates at the intersection of business, engineering, and user experience.
- Data strategy – the plan for how data is generated, collected, and consumed
- Always thinking about the data science process alignment with the needs of the end user

Other Roles

- Data Engineer – Maintain, implement, and test infrastructural components
- Data Architect – Warehousing and Performance
- Business Analyst – Convert business expectations into data analysis
- Statistician - Gather, review, analyze and draw conclusions from data
- Database Administrator – Logical and physical design of databases
- Machine Learning Engineer/Developer – Optimizing, training, monitoring, and maintaining models

We’re really good about putting people into boxes and roles. Rather than doing that, I think we should be focused on how we can use data to solve a problem. - DJ Patil



Applications

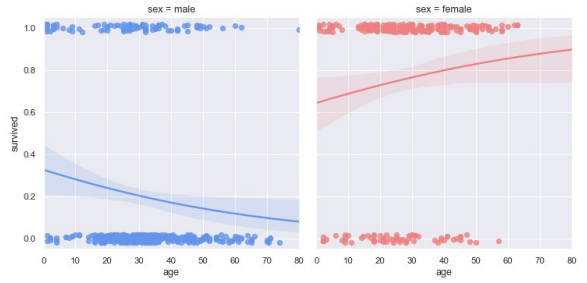
You can use data science for many tasks but it’s generally not the final step of the process!

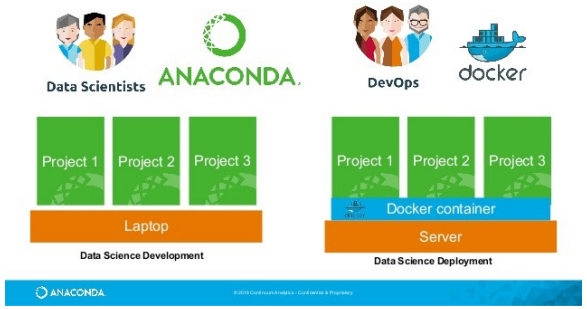
What now?

The data science code/results itself can’t be handed off to just anybody…

Results of the Data Science Process

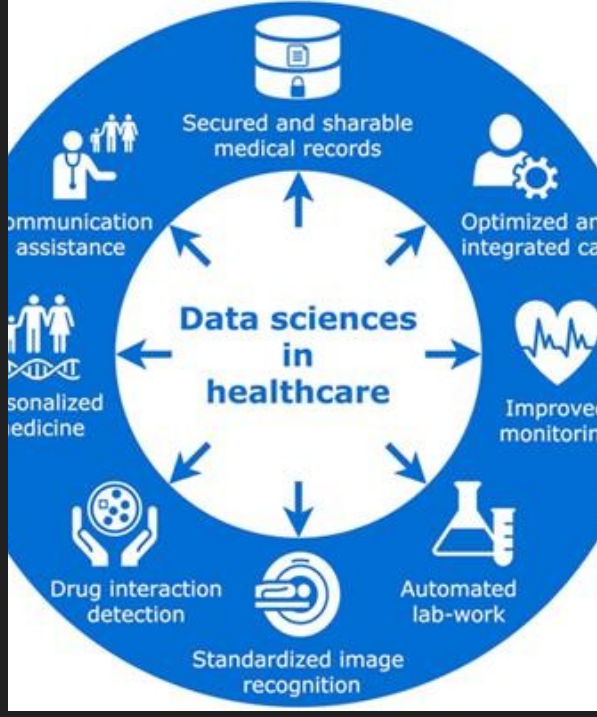
- Decision Making – The intent is to communicate results of an analysis and sometimes nothing more
- Deployment – The code is prepared to accommodate a native production environment
- User-facing Applications – The end-user experiences data science directly

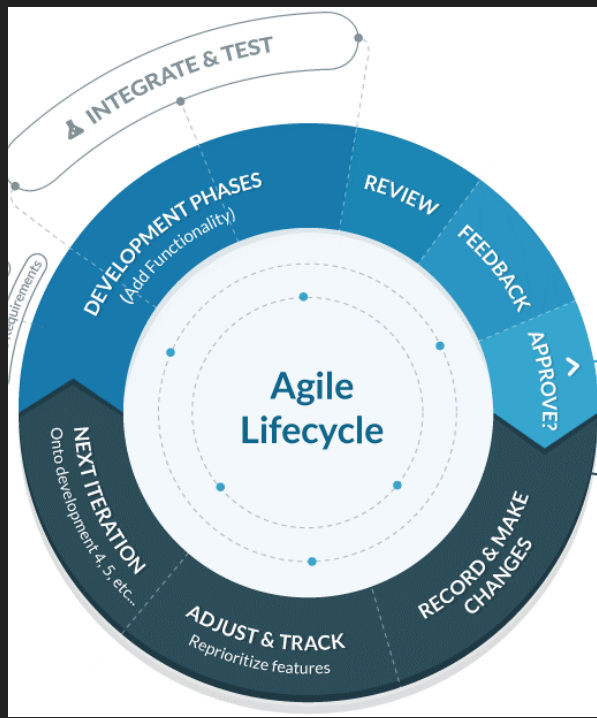


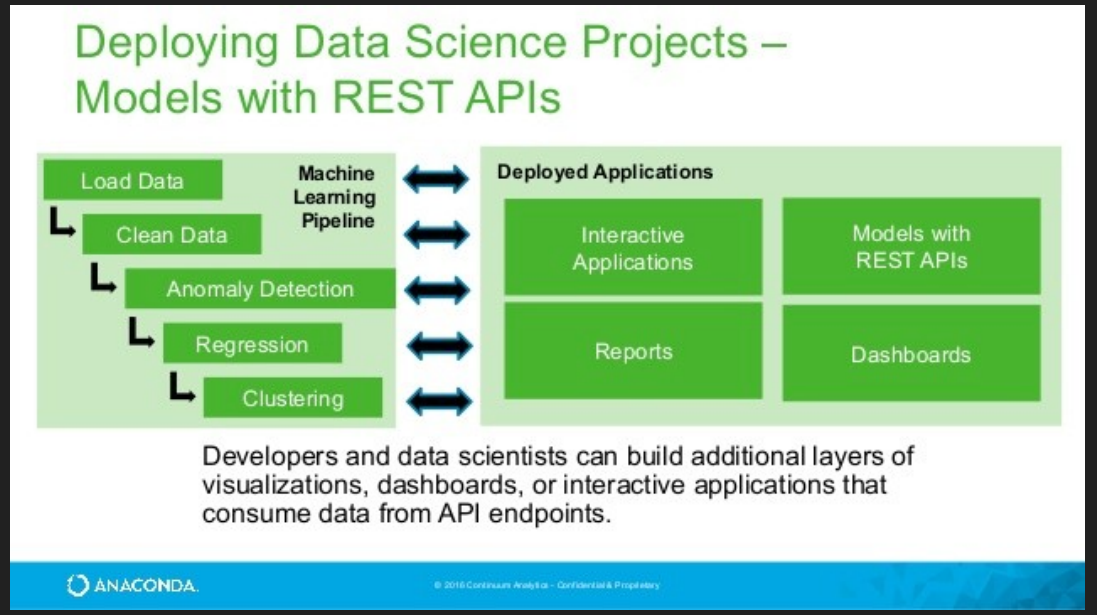


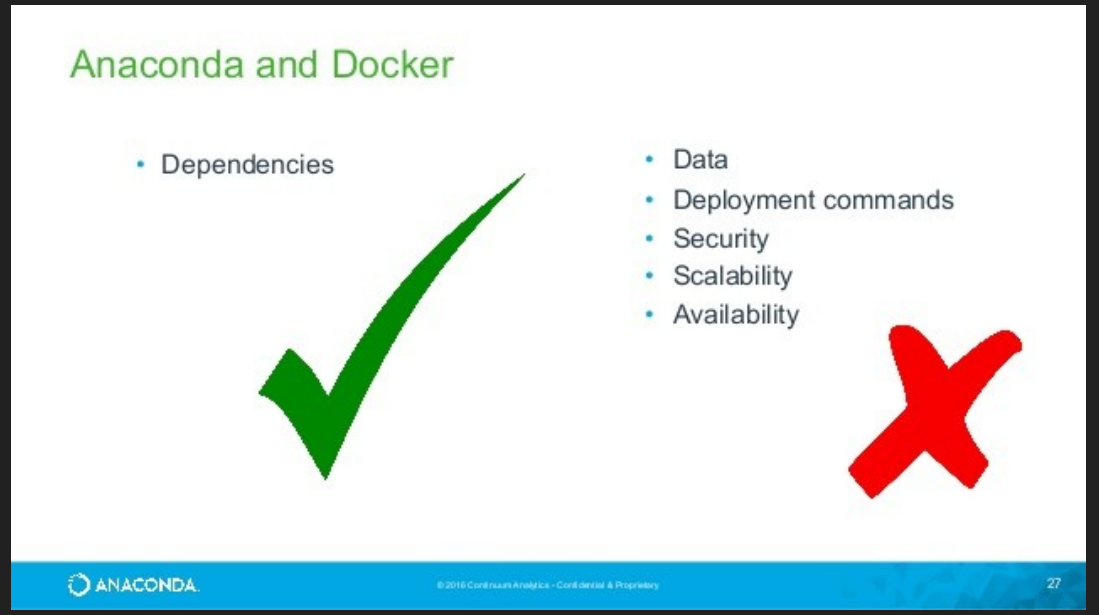
Decision Making

- Arguably the most common use of data science
- Often a prerequisite for deployments and user facing applications
  - Sometimes this IS the final step



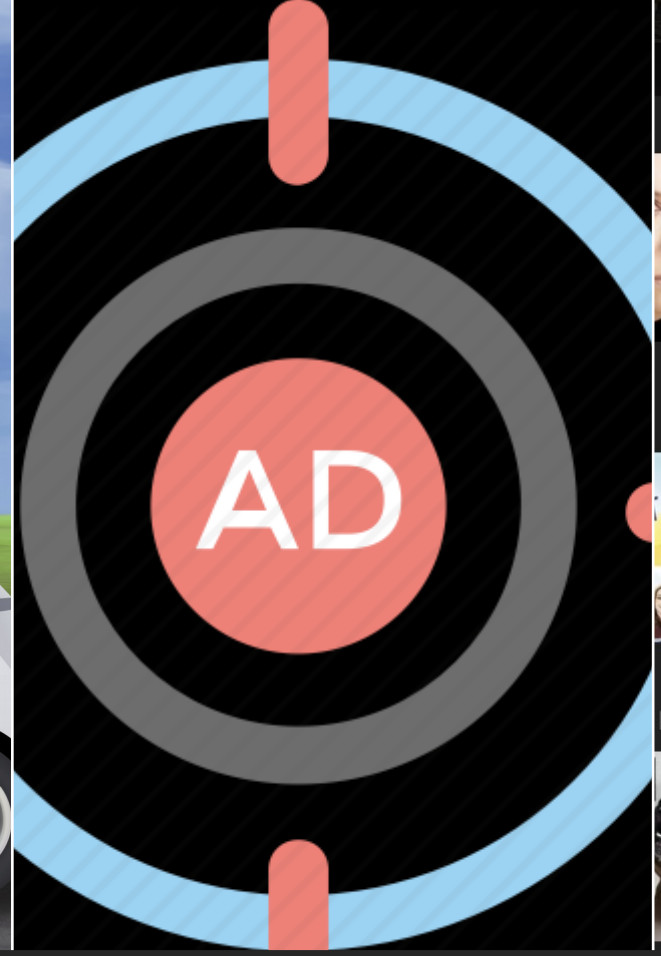


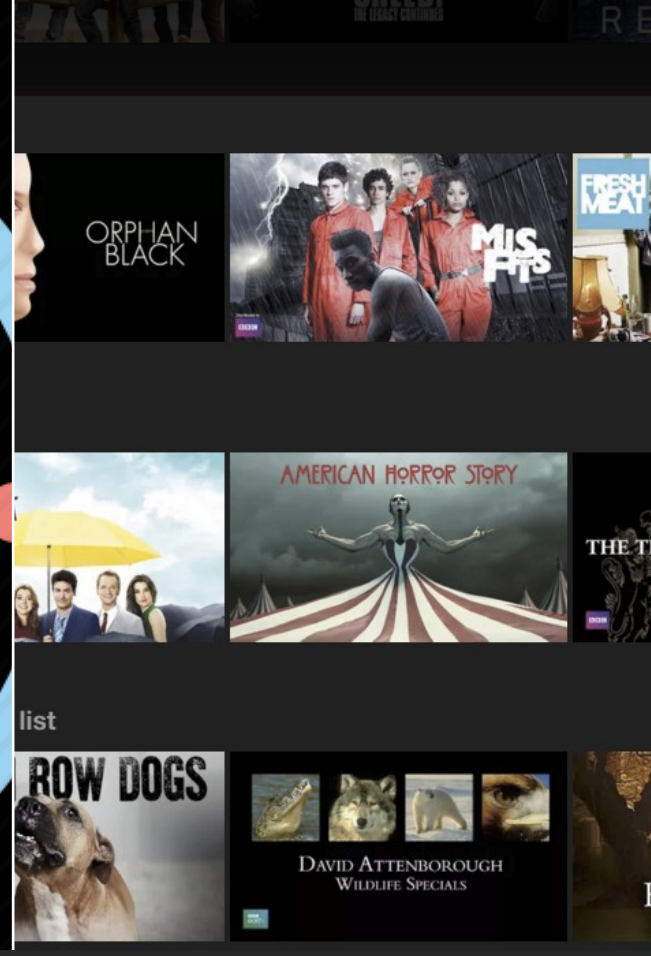




Deployment

For larger organizations and applications with lots of users





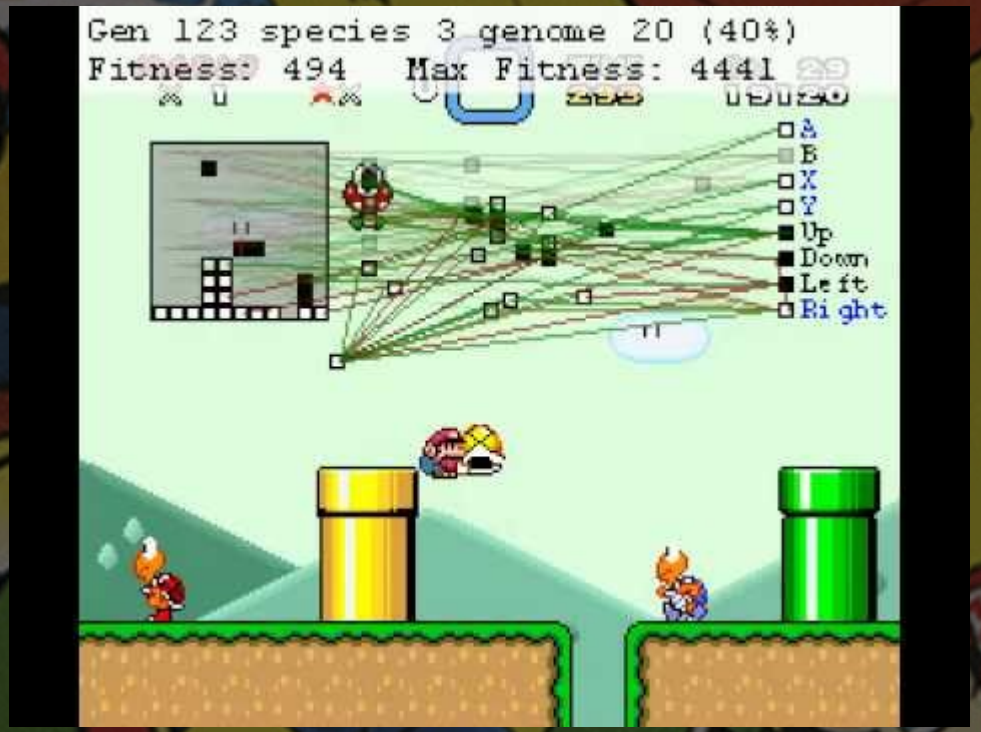
User-facing Applications

Less thought about career wise but ever present and rapidly evolving

Demo 3: Mar/IO

- Neural Networks and Genetic Algorithms
  - Reinforcement Learning
- LUA scripting within the Bizhawk emulator
- Super Mario World with Reinforcement Learning
  - Neural Networks and Genetic Algorithms

https://www.youtube.com/watch?v=qv6UVOQ0F44



Questions?

Feel free to reach out now or any time!

- Slack (Code Louisville): Shane Kimble
- Github: summonholmes
- LinkedIn: https://www.linkedin.com/in/shane-kimble-32991149/