What is Data Science

Peyman Hesami

Data Science Foundations - General Assembly

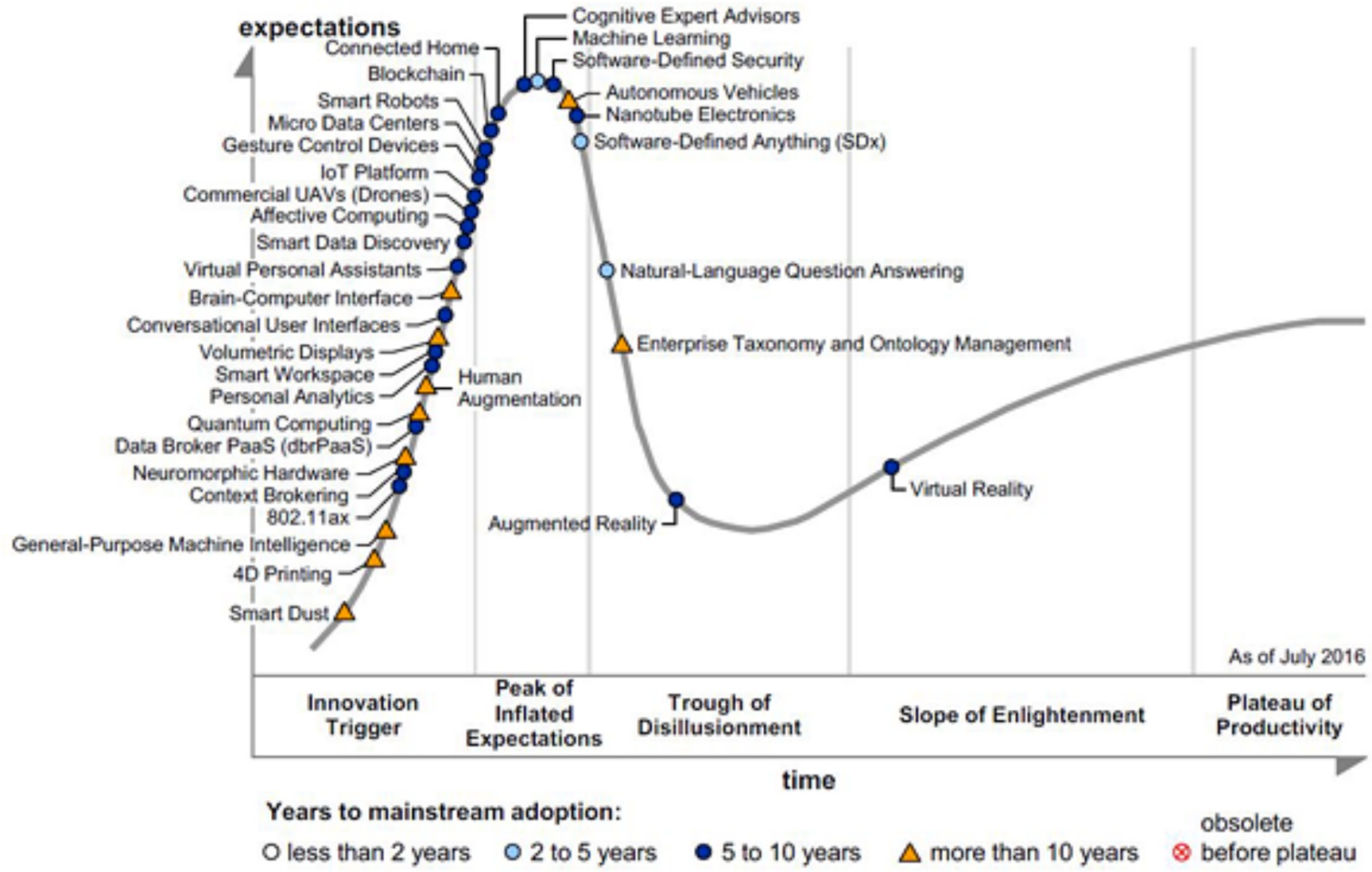
August 1st, 2017

Learning Objectives

After this lesson, you will be able to:

- Describe the roles and components of a successful development environment.
- Define data science and the data science workflow.
- Apply the data science workflow to solve a task.
- Discuss common data science terminology and processes.

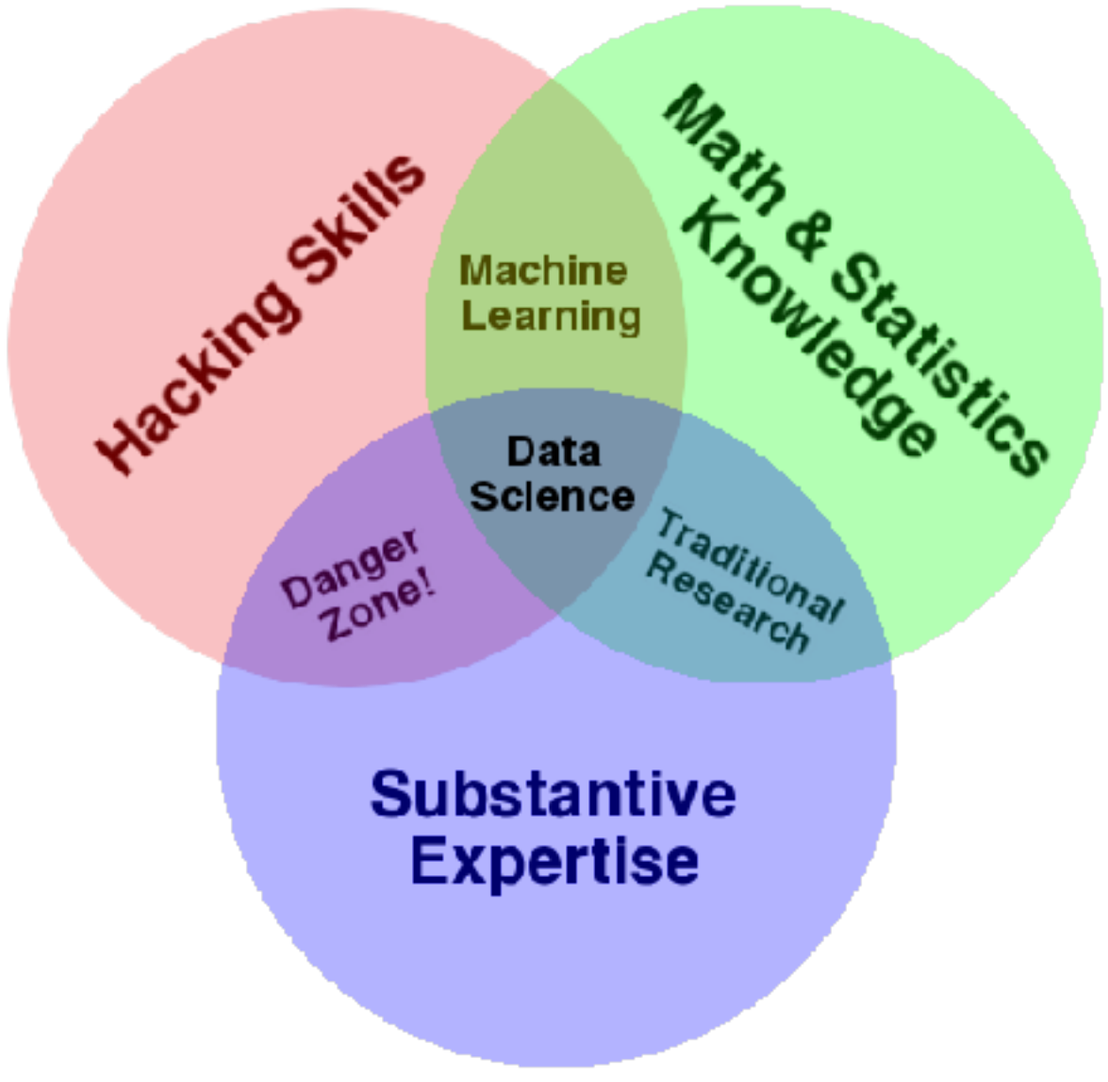
2016 Gartner Hype Cycle



Source: Gartner (July 2016)

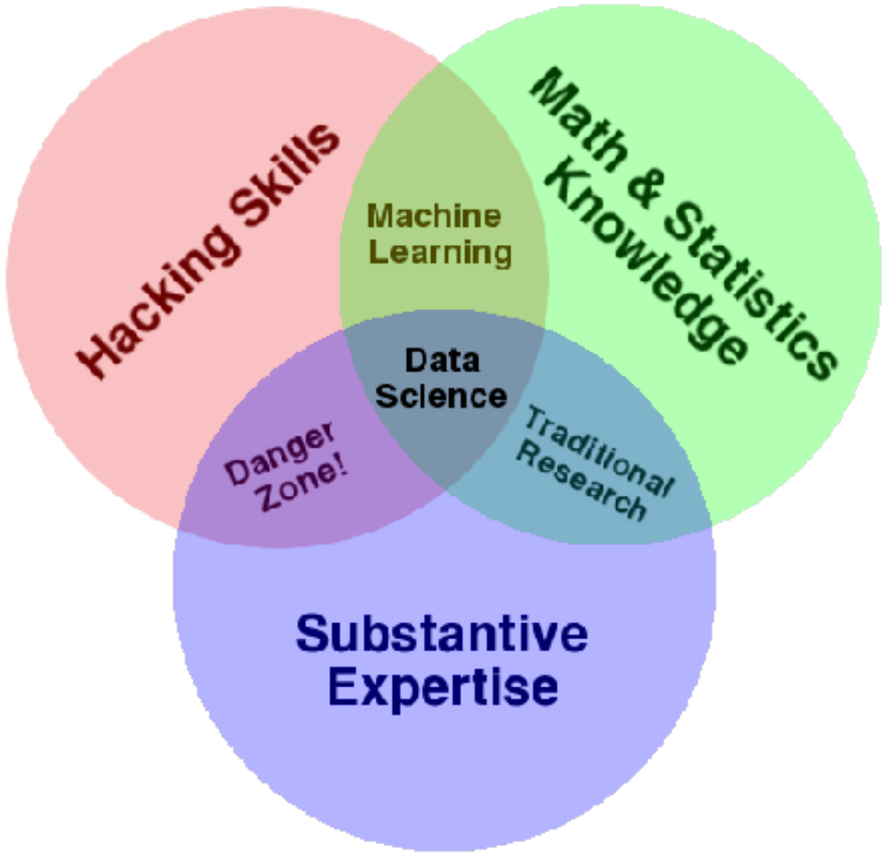
2015 Gartner Hype Cycle

What is a Data Scientist?



- The ***ART*** of transforming ***DATA*** into ***INSIGHTS!***

Who is Data Scientist?



+

=

- Data scientists need to learn hacking and math skills

- What do we do different?! Our goal is to teach you some of the data science secret sauces as well!

What is Data?

- "Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation" — *Merriam-Webster.*

- *Structured vs. semi-structured vs. unstructured*

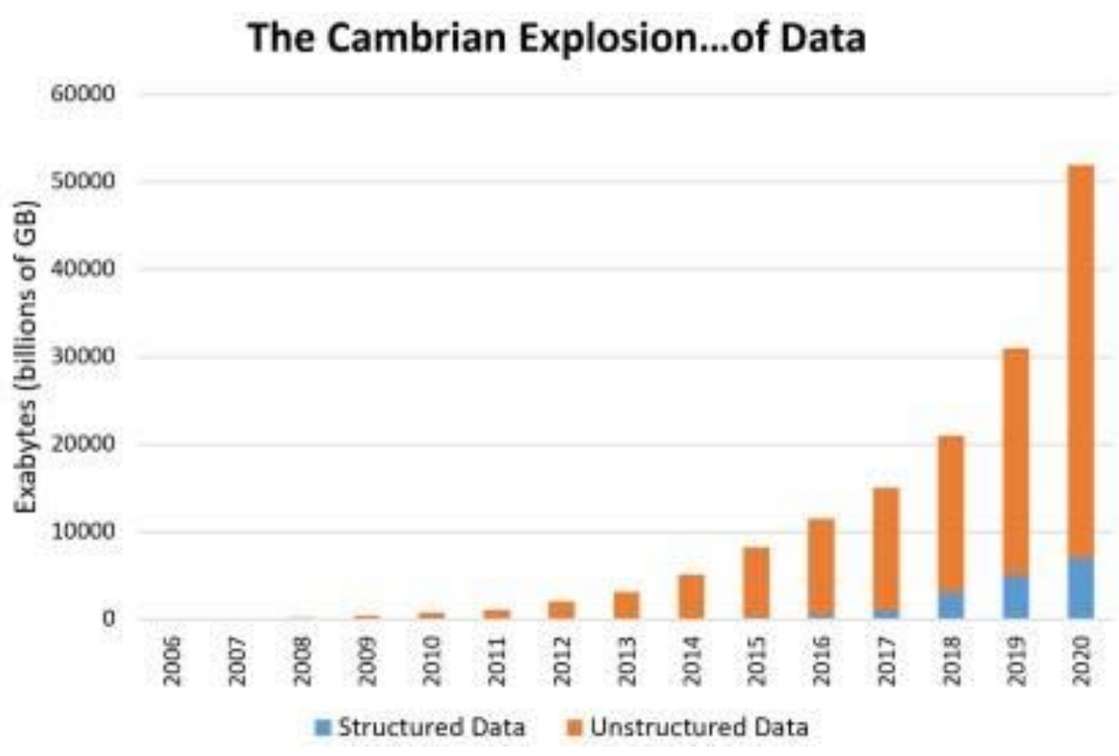
- *80-90% of all potentially usable business information may originate in unstructured form- 1*998, Merrill Lynch rule of thumb

- *Yelp reviews data type?!*

What is Data?

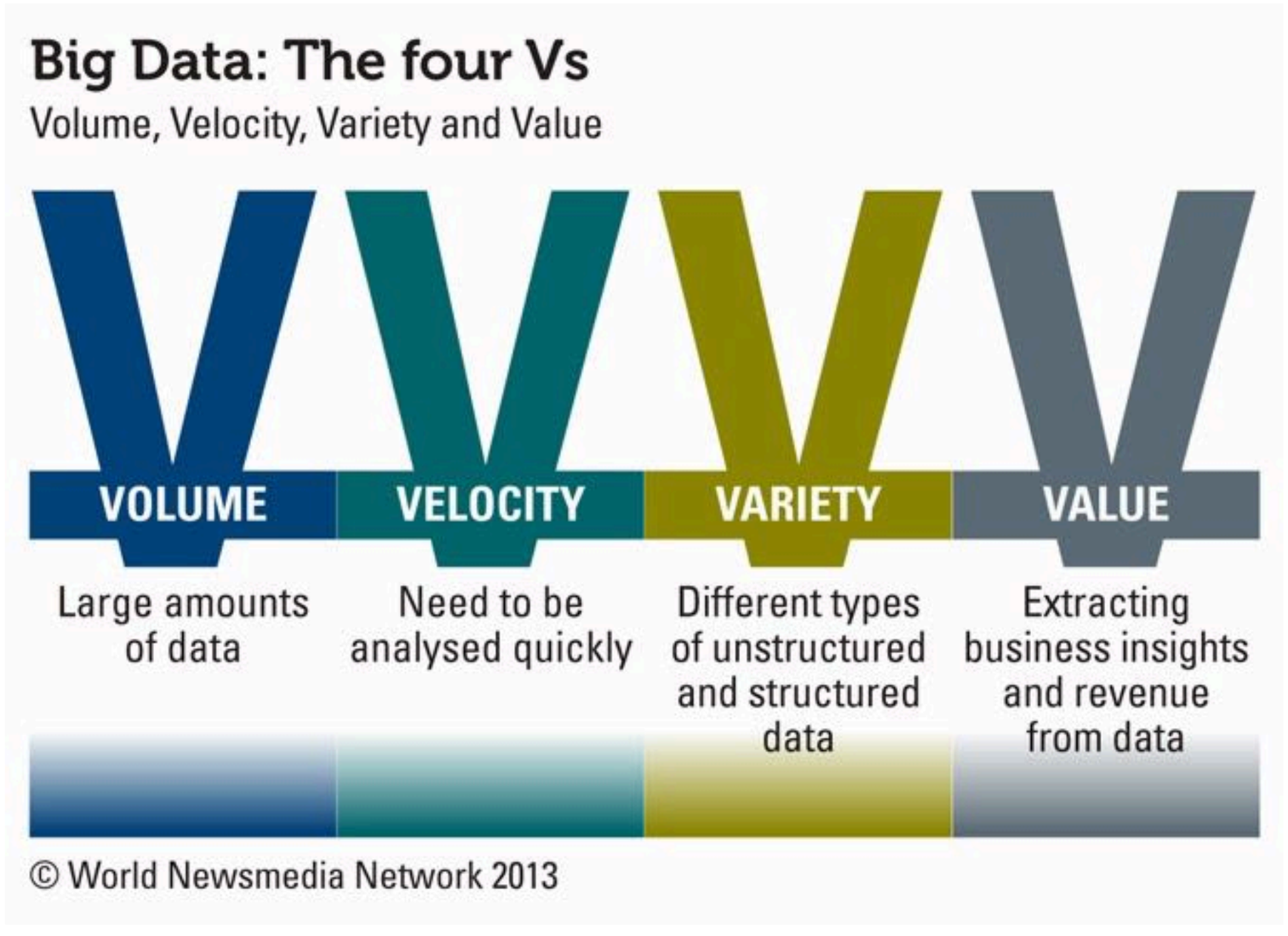
- IDC and EMC project that data will grow to 40 zettabytes by 2020 (from 4 ZB in 2016)

- 1 ZB = 1000^7bytes = 1 billion terabytes = 1 trillion gigabytes



Source: EEtimes.com

4 V’s of Big Data



Who Uses Data Science?

- Targeted Marketing @ Target!

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill, FORBES STAFF
Welcome to The Not-So Private Parts where technology & privacy collide

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

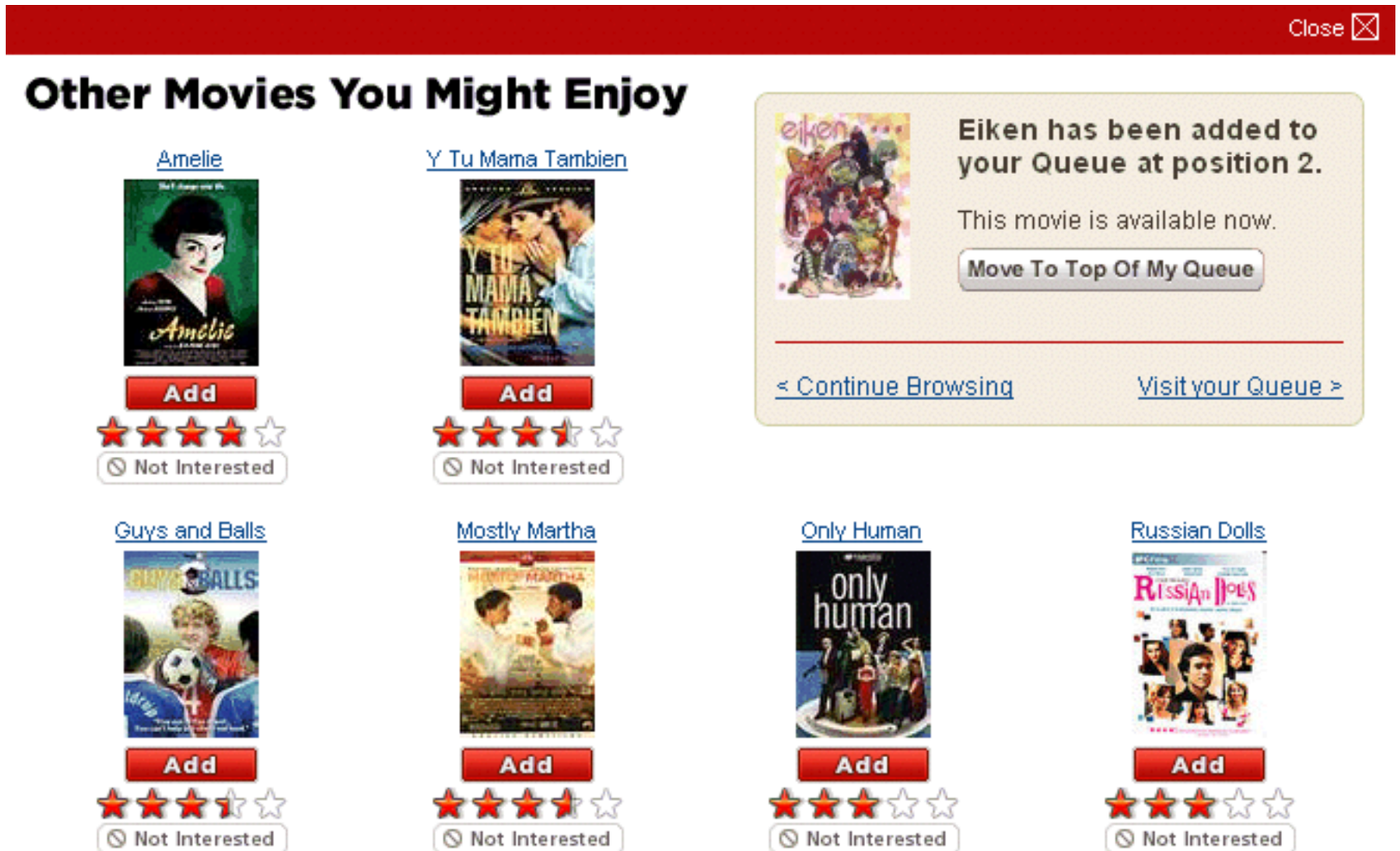
TARGET

Target has got you in its aim

"My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

Who Uses Data Science?

- Providing movie recommendations on Netflix.

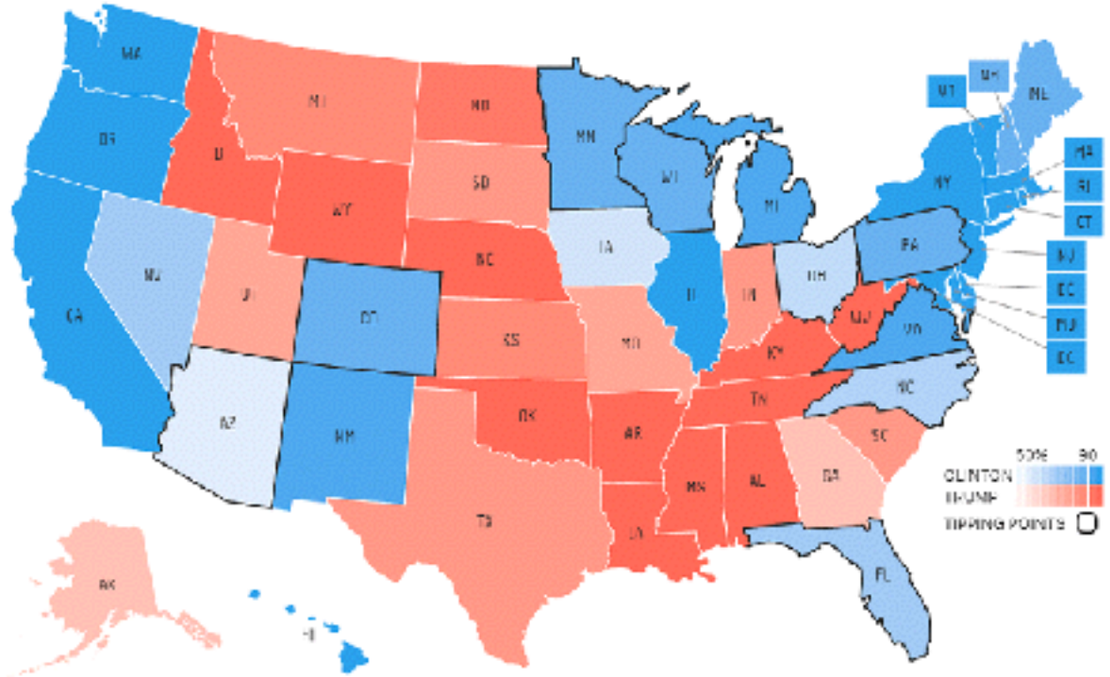


Who Uses Data Science?

- Making product suggestions on Amazon!

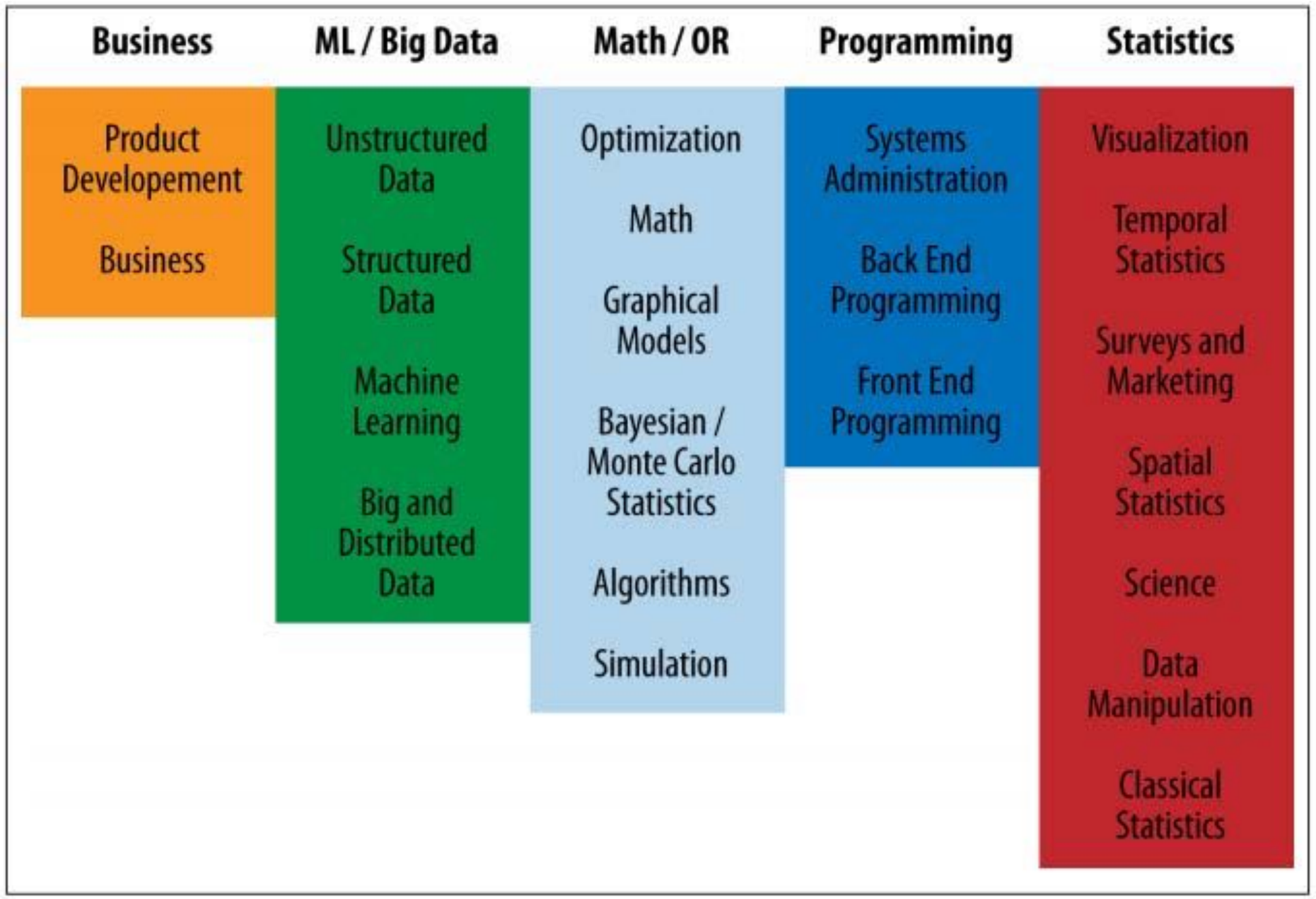
Who Uses Data Science?

- Offering election and sports coverage on the stats site FiveThirtyEight.

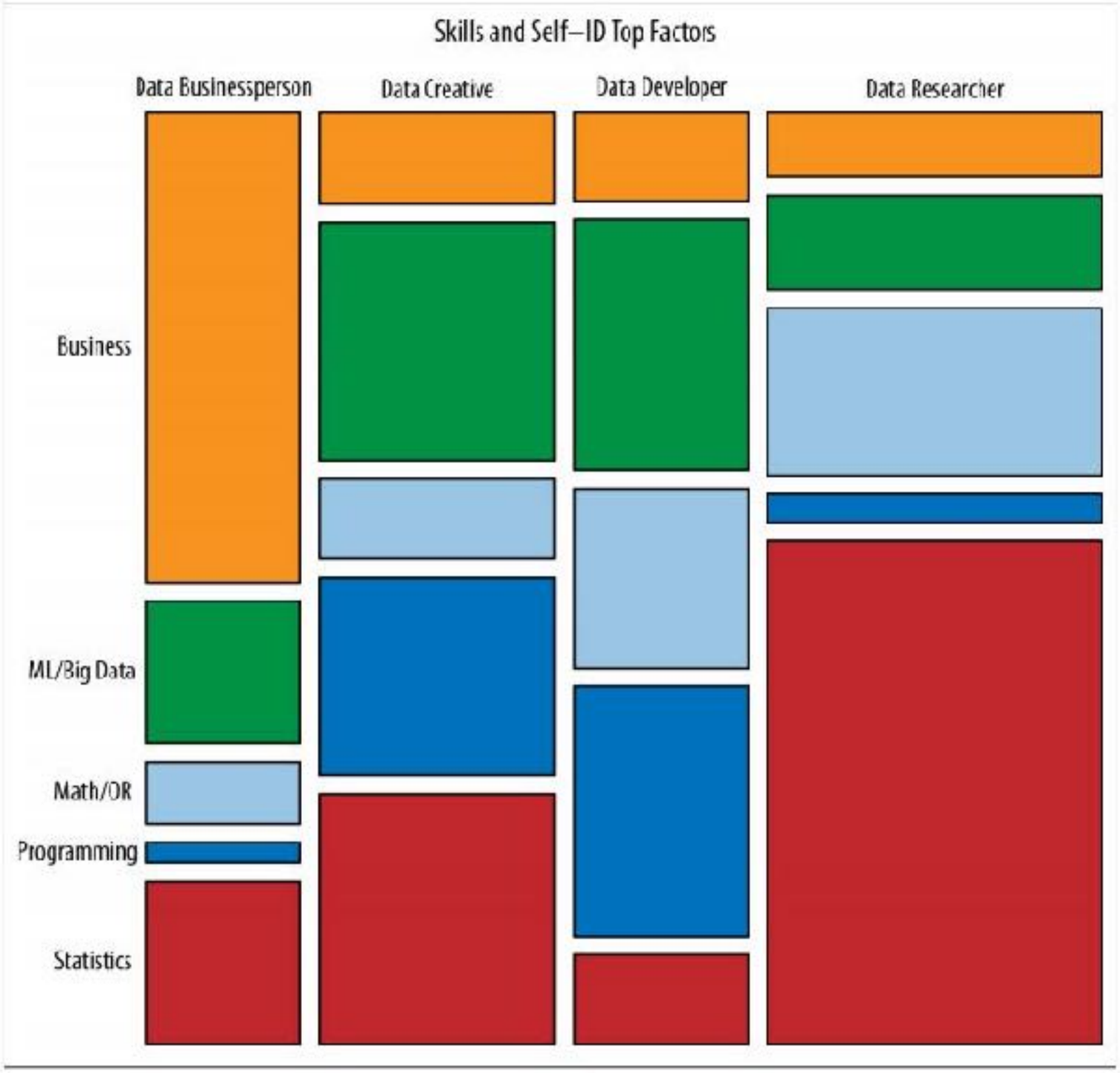


- Any other example? “how did it know that?”

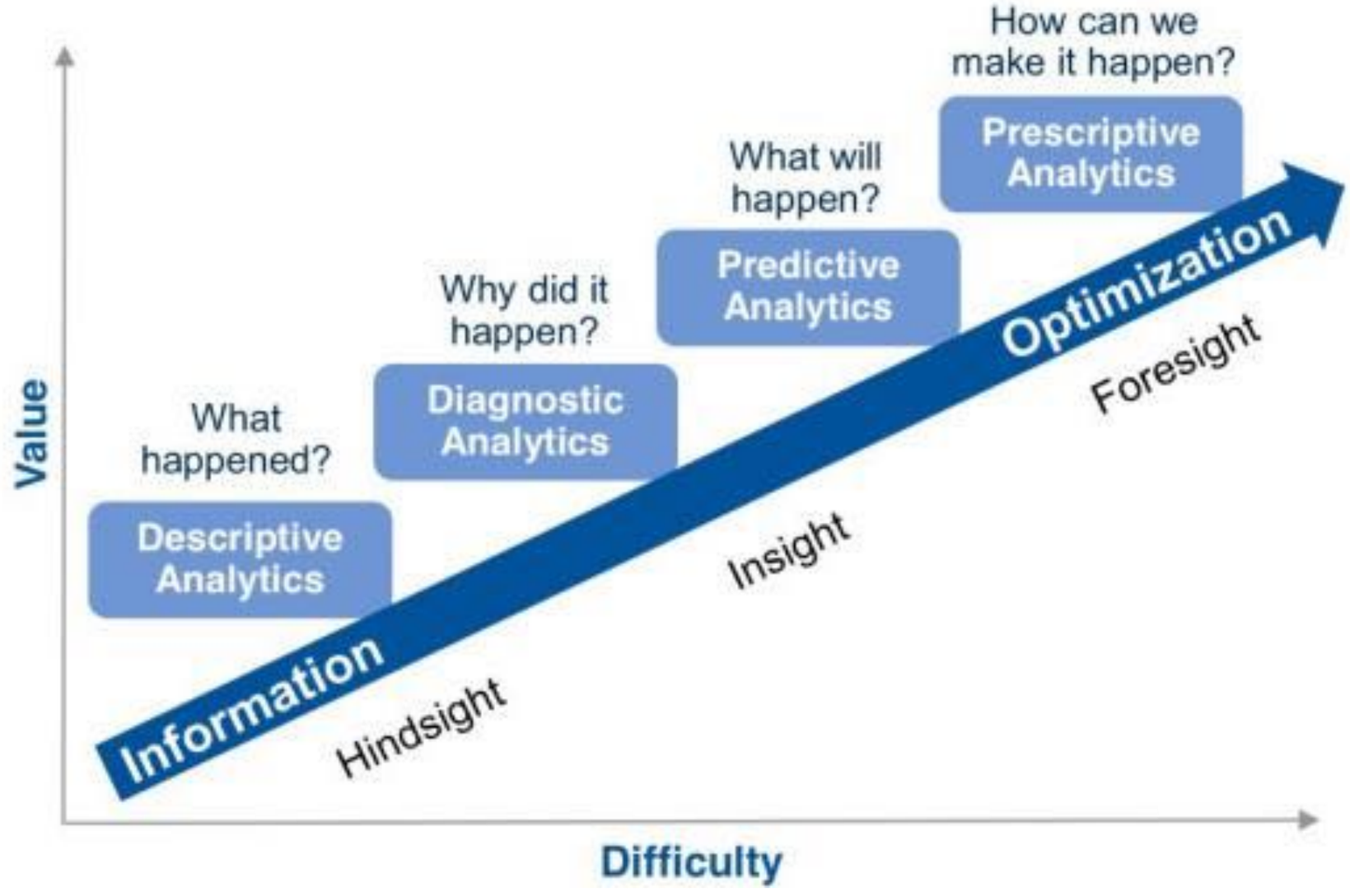
Data Scientist Skills



Data Scientist Skills



Type of Data Science Projects



- Source: http://www.rosebt.com/blog/descriptive-diagnostic-predictive-prescriptive-analytics

Data Science Outcomes

- **Reliable**: Accurate findings

- **Reproducible**: Others can follow your steps and achieve the same results

Data Science Workflow

Develop a hypothesis-driven approach to your analysis

Select, import, explore, and clean your data

Structure, visualize, and complete your analysis

Make recommendations and business decisions from your data

Present insights from your data to different audiences

Data Science Workflow-Frame

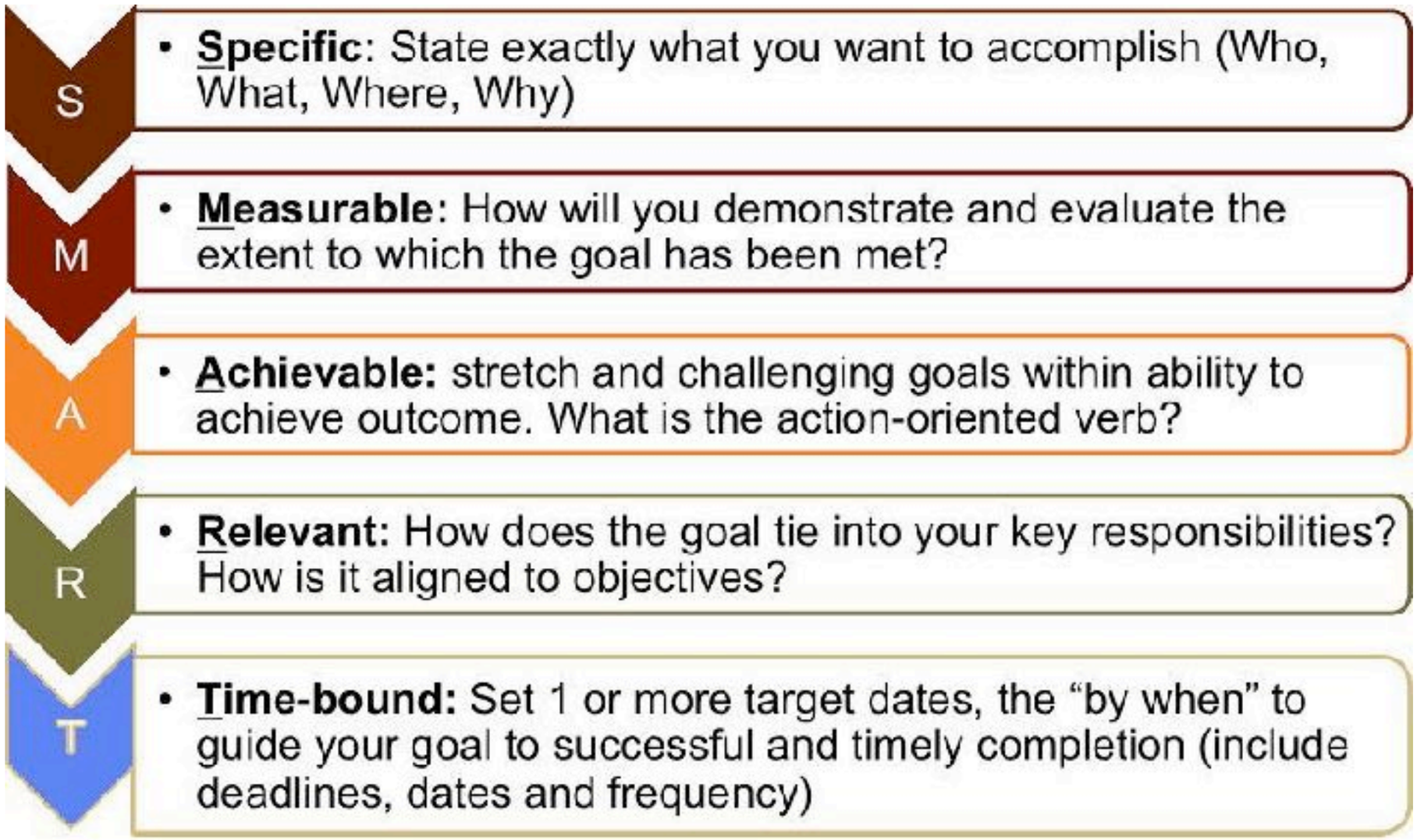
- *As a senior data scientist/manager/CIO/CTO/CEO, you need to identify the business objective.*

- *Why do we need a good question?*

- *“A problem well stated is half solved.” — Charles Kettering*

- ***A good question:***
  - *Sets you up for success as you begin analysis.*
  - *Establishes the basis for reproducibility.*
  - *Enables collaboration through clear goals.*

Characteristics of Good Question



- Source: https://greaterphoenix.score.org/blog/smart-goals-–-recipe-success

Common Questions Asked in Data Science

- **From a business perspective, we can ask:**
  - What is the likelihood that a customer will buy this product?
  - How much demand will there be for my service tomorrow?
  - Is this the cheapest way to deliver my goods?

- **From a data science perspective, these will translate to**
  - Does X predict Y? (Where X is a set of data and y is an outcome.)
  - Are there any distinct groups in our data?
  - Is one of our observations “weird”?

Project: Futurama

- For those of you not familiar with *Futurama*, here are some quick notes about the show:
  - It's an animated comedy series set in the year 3000
  - It focuses on the adventures of the space delivery company Planet Express

Project: Futurama

- Using Planet Express' customer data from January 3001-3005, determine how likely previous customers are to request a repeat delivery using demographic information (e.g., profession, company size, location) and previous delivery data (e.g., days since last delivery, number of total deliveries, etc.).
    - Identify the business/product objectives.
        - ✓How likely are previous customers to request a repeat delivery?
    - Identify and hypothesize goals and criteria for success.
        - ✓What factors are likely to influence a customer's decision to reuse Planet Express for delivery?
- These types of questions will help you identify the correct data set!

Data Science Workflow

Develop a hypothesis-driven approach to your analysis

Select, import, explore, and clean your data

Structure, visualize, and complete your analysis

Make recommendations and business decisions from your data

Present insights from your data to different audiences

Data Preparation:

Understand, Structure, and Clean the Data¶

- Ideal Data vs. Data That is Available

Data Preparation

- The data could be incomplete, non-existant, or unable to meet the criteria necessary to answer your question

- Different sources of data (CSV, XML, TXT, JSON) might need to be integrated

Data Preparations- Understand the Data

- Building data dictionaries and source documentation to understand the data

Variable	Description	Type of Variable
Profession	Title of the Account Owner	Categorical
Company Size	1- small, 2- medium, 3- large	Categorical
Location	Planet of the Company	Categorical
Days Since Last Delivery	Integer	Continuous
Number of Deliveries	Integer	Continuous

- Performing exploratory surface analysis via filtering, sorting, and simple visualizations.

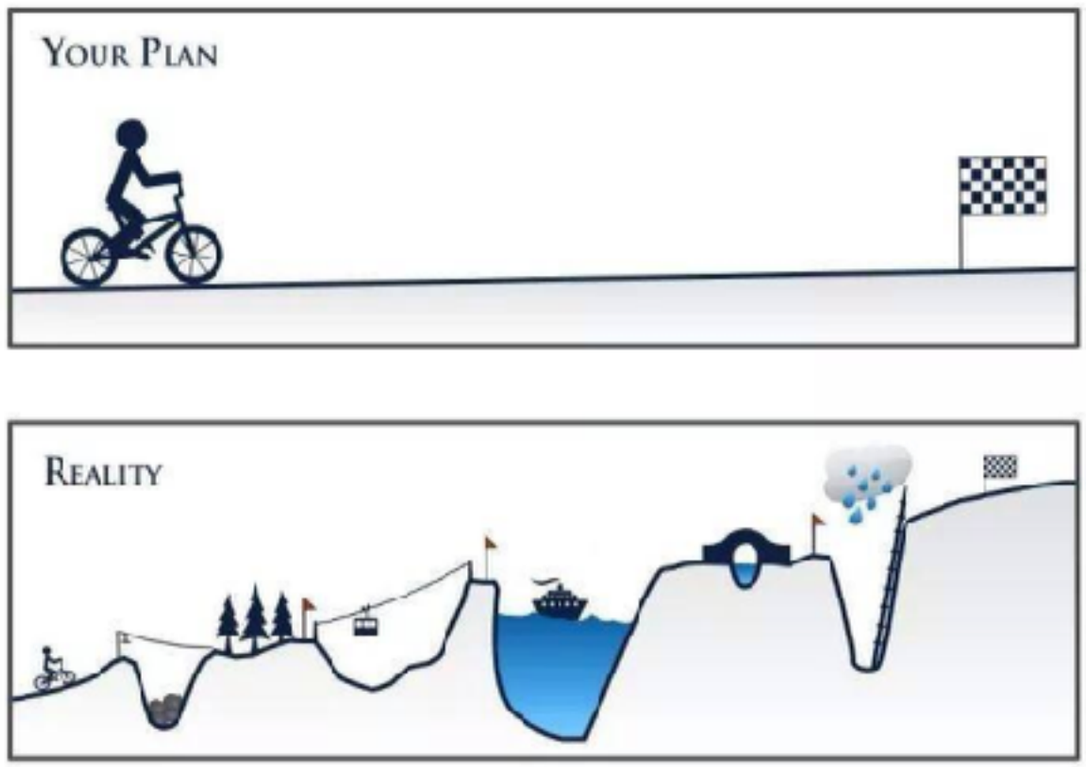
- Assessing preliminary outliers and trends.

Data Preparation- Clean the Data

- Sampling the data and determining sampling methodology.
- Formatting and cleaning data (e.g., dates, number signs, formatting).
- Defining how to appropriately address missing values
- Categorizing, manipulating, slicing, formatting, and integrating data.
- Determining most appropriate methods for aggregating, cleaning, etc

Expectations vs. Reality

- Expect to spend 80% of your time on data cleaning!!!



Source: KDNuggets

- Insightful data scientists learn their secret sauces in data preparation!!

Data Science Workflow

Develop a hypothesis-driven approach to your analysis

Select, import, explore, and clean your data

Structure, visualize, and complete your analysis

Make recommendations and business decisions from your data

Present insights from your data to different audiences

Analyze: Exploratory Data Analysis

- Descriptive and diagnostic analyses:

Variable	Mean (STD) or Frequency (%)
Number of Deliveries	50.0 (10)
Earth	50 (10%)
Amphibios 9	100 (20%)
Bogad	100 (20%)
Colgate 8	100 (20%)
Other	150 (30%)

- These descriptive stats allow us to:
  - Identify trends and outliers.
  - Choose visualization techniques for different data types.
  - Transform data

Analyze: Feature Engineering

- Creating new/transformed features based on the data
- One of the secret sauces of a successful data scientist!
  - Simple: **Mass/Size = Density**
  - Complex: **image to spectrogram**

Analyze: Create a Model

- Predictive data analysis: Create a model to predict the outcome we are interested in

- "*We completed a logistic regression. We calculated the probability of a customer placing another order with Planet Express.*"

- The steps for model building are:
    - Selecting the appropriate model
    - Building the model
    - Evaluating and refining the model
    - Predicting outcomes and action items

Data Science Workflow

Develop a hypothesis-driven approach to your analysis

Select, import, explore, and clean your data

Structure, visualize, and complete your analysis

Make recommendations and business decisions from your data

Present insights from your data to different audiences

Interpret The Results

- Interpreting the results with subject matter experts
- Were you questions answered?

  ➡What do you need to do to answers the ones that weren't answered?

- Do your findings support any business recommendations, actions, or decisions?

  ➡How does your data support these recommendations?

Interpret The Results- Futurama

- Conclusion:

"Customers from large companies were twice as likely to place another order with Planet Express than customers from small companies."

- Recommendation:

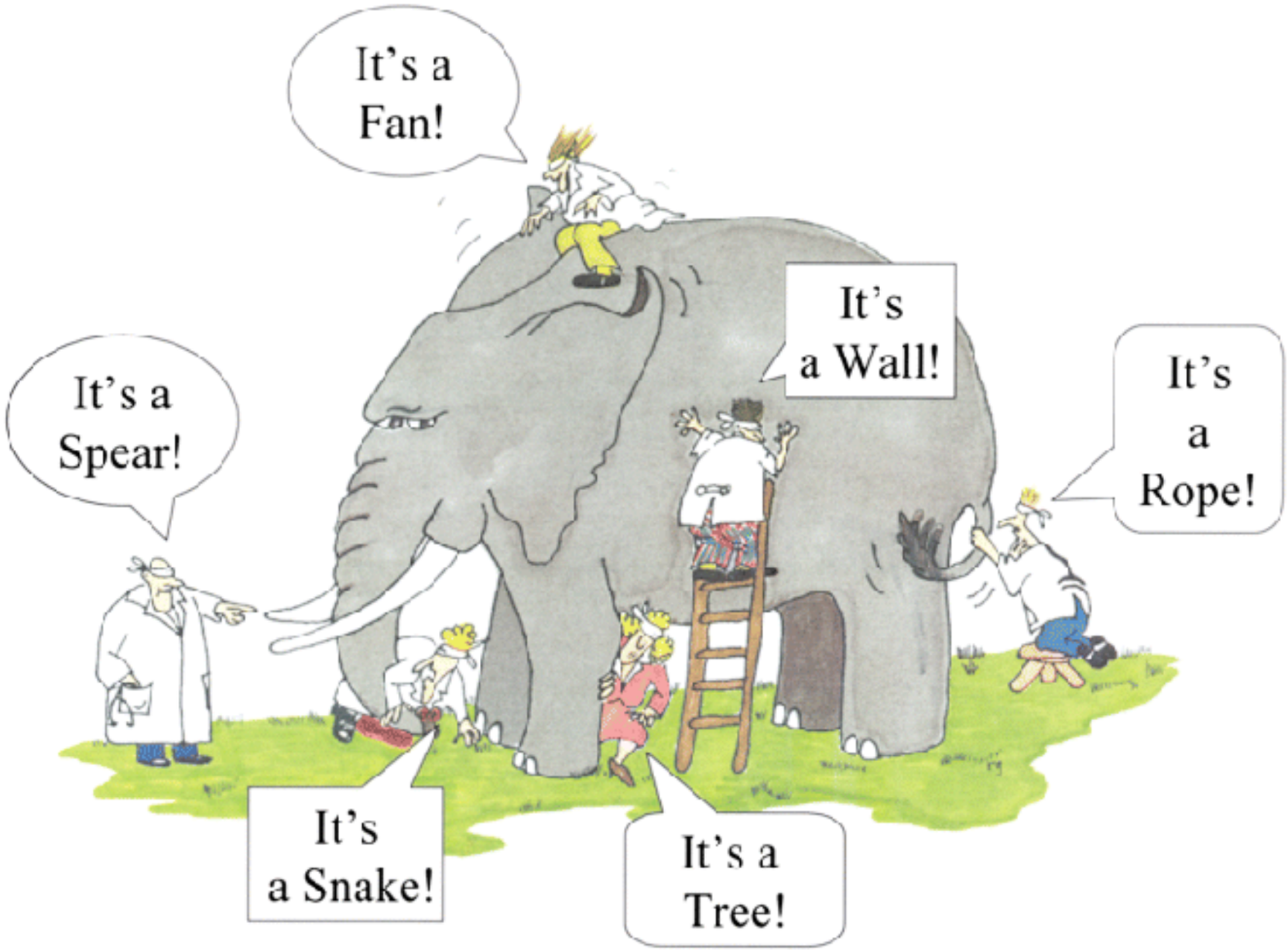
“We should target more large companies to use our delivery service."

Communicate the Results

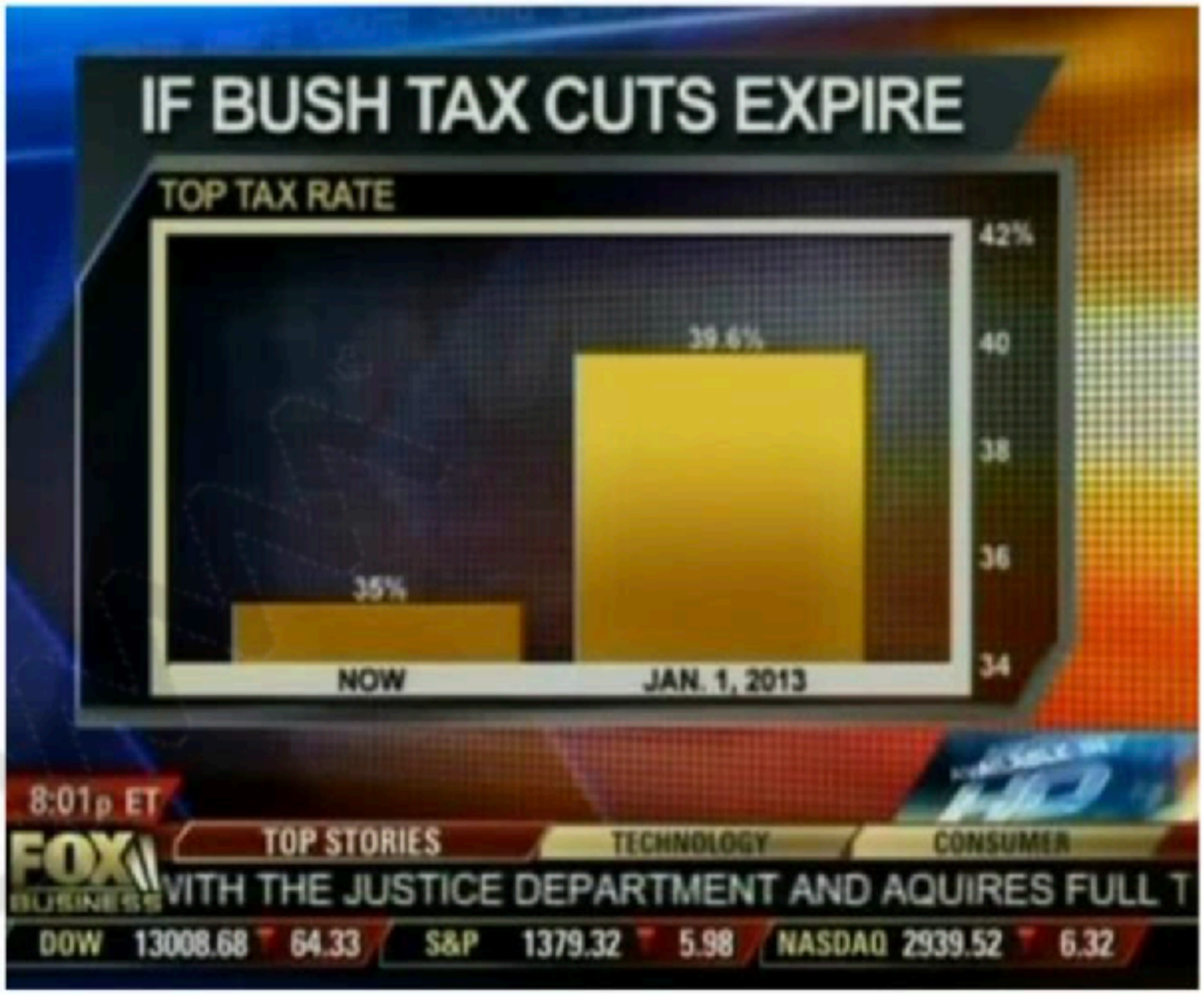
- Audience focused! CEO vs. CTO!
- No matter how brilliant your model is, if you are not able to effectively communicate your results, they will not be used
- Communicate through simple words and powerful visualizations
    - *“Customers from large companies had twice the odds for placing another order with Planet Express compared to customers from small companies.”*

Ethics of Data Science

- Cherry picking the results!

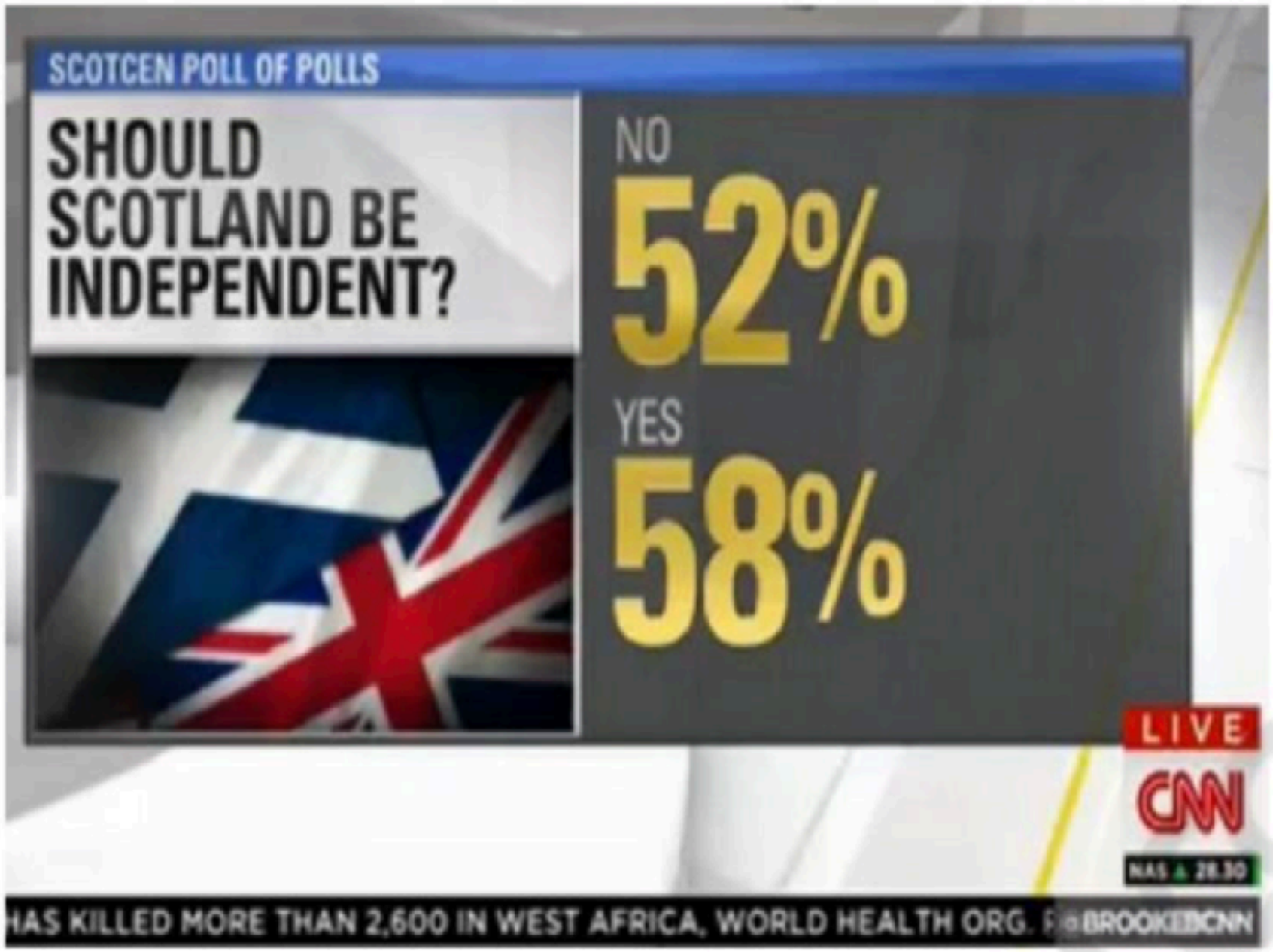


Ethics of Data Science

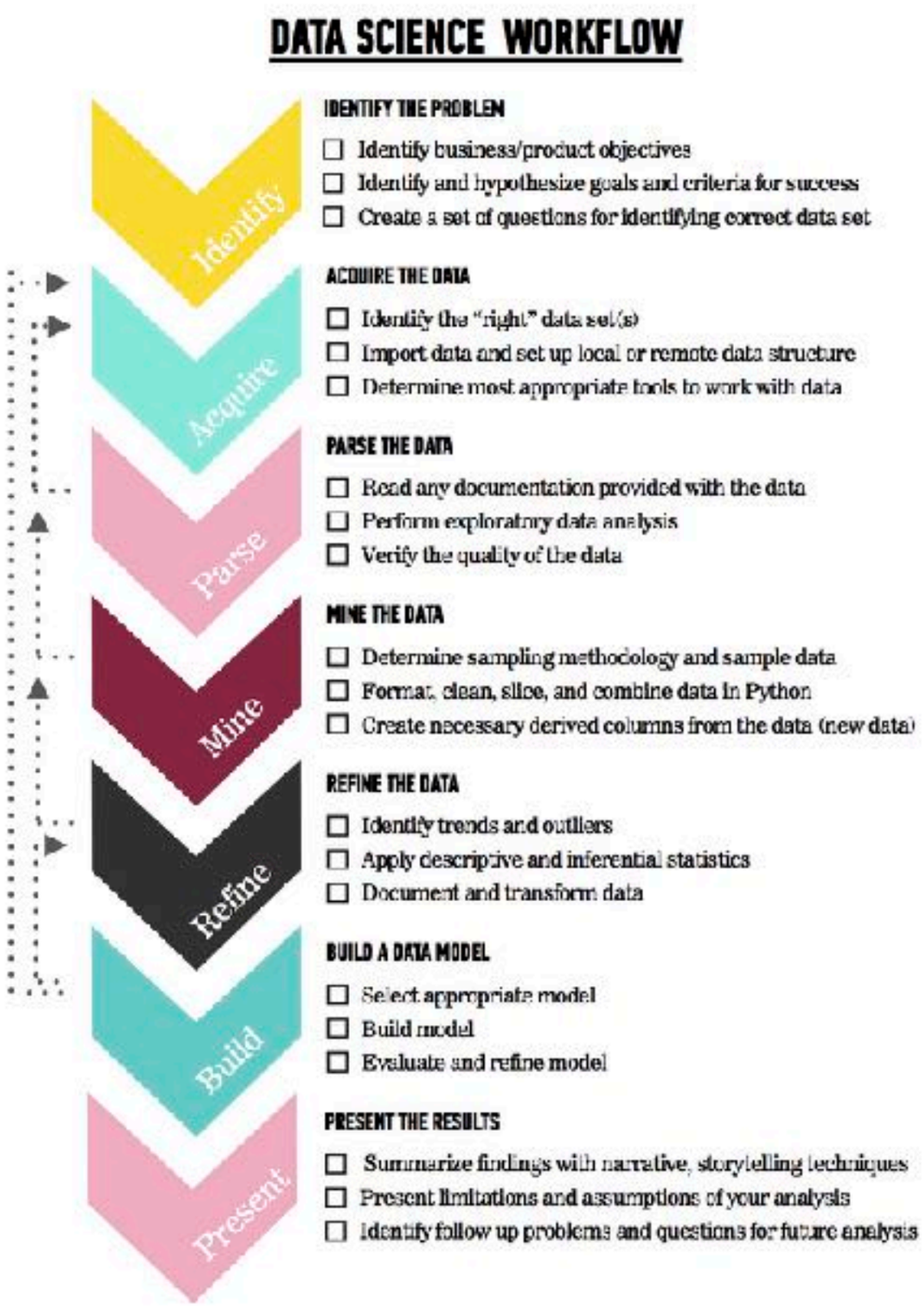


Ethics of Data Science

Ethics of Data Science



Iterations in Data Science



Group Practice:
The Data Science Workflow

- Use four of the steps from the Data Science Workflow (Frame, Prepare, Analyze, Communicate) to analyze the following subject:

Number of hours of sleep

Group Practice:
Number of Hours of Sleep

- **FRAME: Understand the Problems:** In groups of 5, develop one research question about number of hours of sleep and form a hypothesis. Examples:
    - Does age/ethnicity/work schedule/favorite cuisine impact the number of hours of sleep? How?
    - Does month/season impact the number of hours of sleep? How?
    - Does work schedule/favorite cuisine impact the number of hours of sleep? How?
- **PREPARE & ANALYZE**: Obtain the Data and Examine It¶
    - Rotate through the members of your group to "collect the data" and record the raw data in a table
    - Manually analyze the data and build a dummy predictor
- **PRESENT:** Communicate the Results of Your Analysis¶
    - Summarize your findings in a narrative/visualizations

Common Data Science Terminologies

150 observations

(n = 150)

Feature matrix “X” has n rows and p columns

Response “y” is a vector with length n

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

response

4 features (p = 4)

Data Types

- **Numerical:** Simply enough, numerical data is typically data represented by numbers.
    - Examples: sales, height, passenger count, etc.
- **Categorical:** Categorical data, on the other hand, is any data *not* represented by numbers.
    - Examples: color, name, organism type, etc.
- What is image data type?!

Dataset Types

Datasets can comprise both categorical and numerical data, and data set types can offer additional information about the data as a whole.

- **Cross-sectional :** All information is determined at the same time; all data come from the same time period

- **Time series:** The information is collected over a period of time for a single group

- **Longitudinal/Panel:** The information is collected over a period of time for several groups

Why Do Data Types Matter?¶

- It has high impact on choosing the right analysis algorithm

- Different data types have different limitations and strengths

- Certain types of analyses aren’t possible with certain data types

Machine Learning

- "A field of study that gives computers the ability to learn without being explicitly programmed." — Arthur Samuel, AI pioneer

- Another definition states that, “Machine learning is the semi-automatic extraction of knowledge from data.”

- **Knowledge from data:** The process starts with a question that might be answerable using data

- **Automatic extraction:** A computer provides the insight

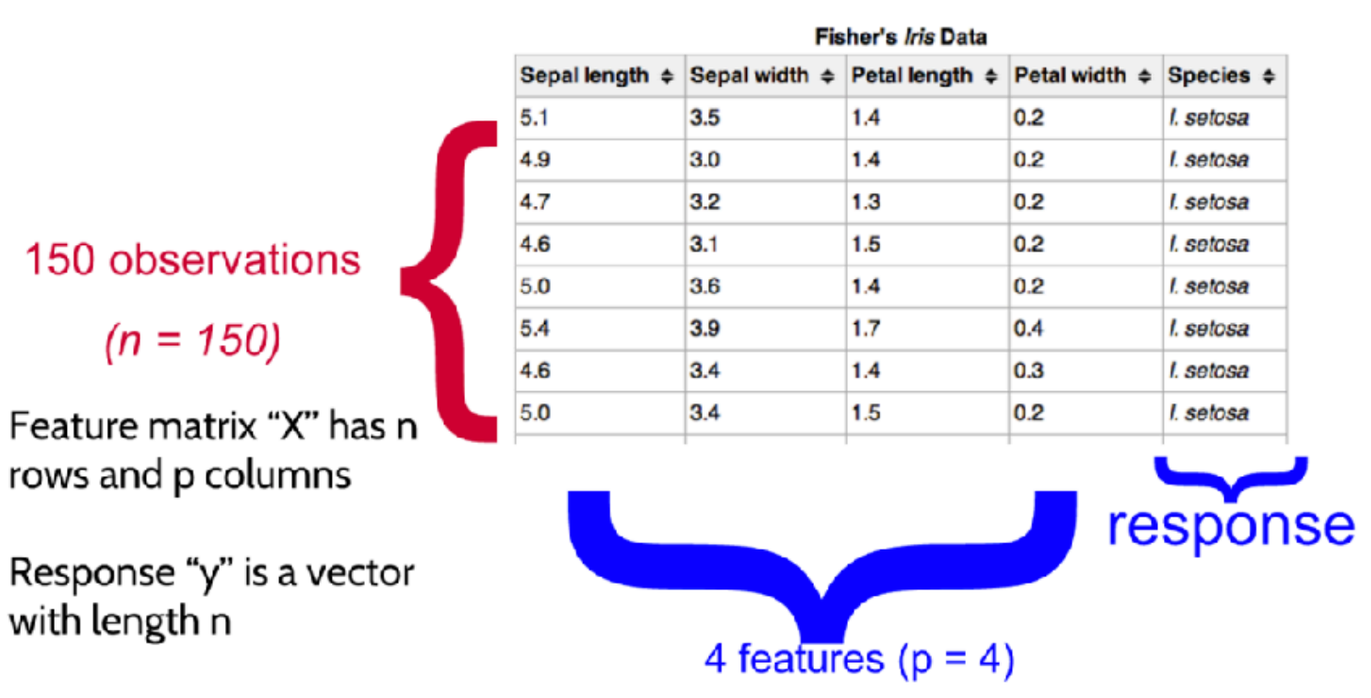
- **Semi-automatic:** It still requires many smart decisions by a human

Machine Learning

- Teaching a dog parody!

Types of Machine Learning Problems

- **Supervised learning (a.k.a., "predictive modeling"):** The data is labeled. Machine is learning the label.



Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

- **Unsupervised learning:** The data is unlabeled. Machine is learning for hidden patterns/insights

Supervised Learning

1. We train a **machine learning model** using **labeled data**
   - The “machine learning model” learns the relationship between the features and the response.

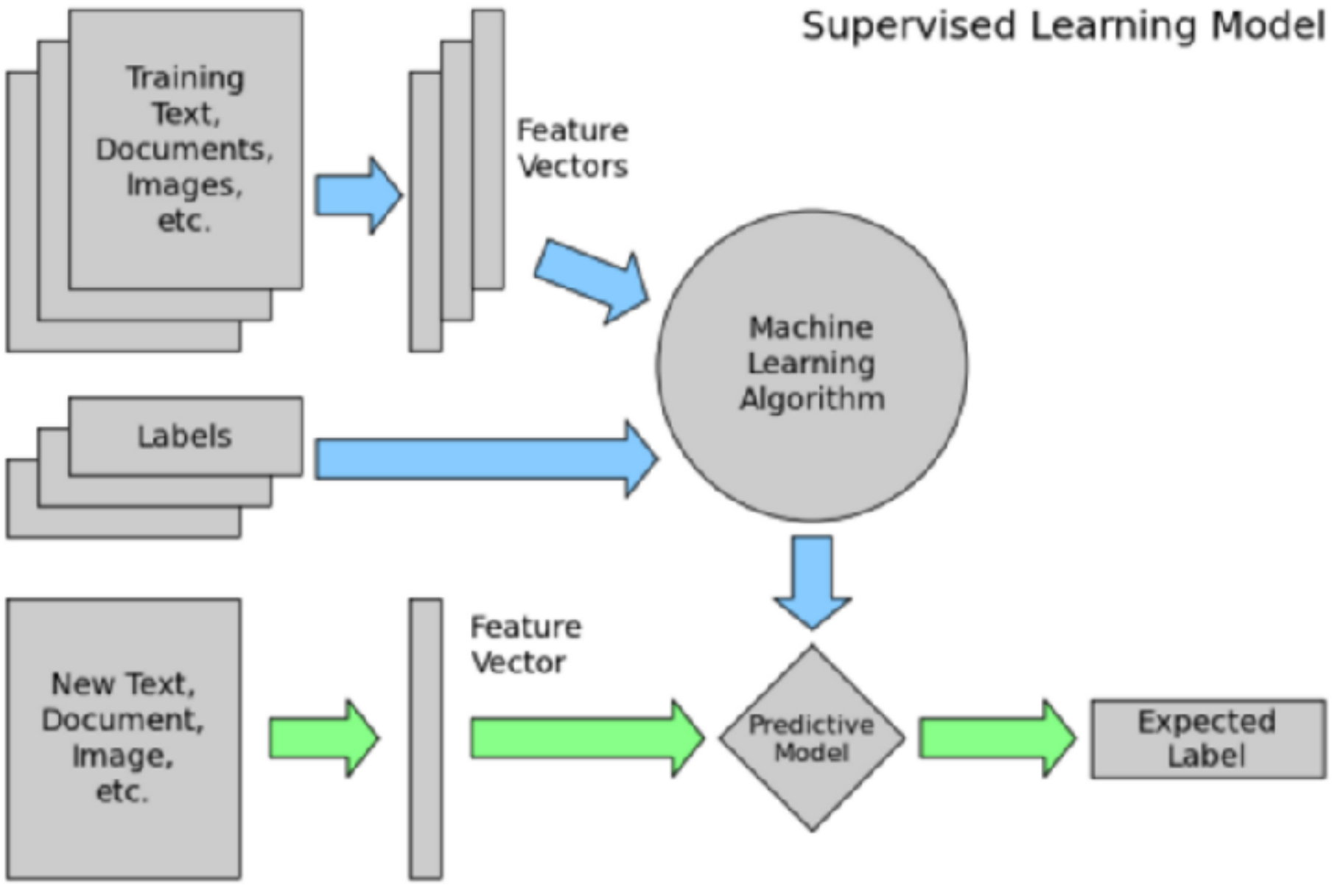
2. We make predictions on **new data** for which the response is unknown.

The primary goal of supervised learning is to build a model that “generalizes” — i.e., accurately predicts the **future** rather than the **past**!

Supervised Learning Toy Example

	Location	Weather	Fun? (Label)
Sample1	Prison	Cold	No
Sample2	Beach	Cold	No
Sample3	Beach	Warm	Yes
New Data	Prison	Hot	?

Supervised Learning



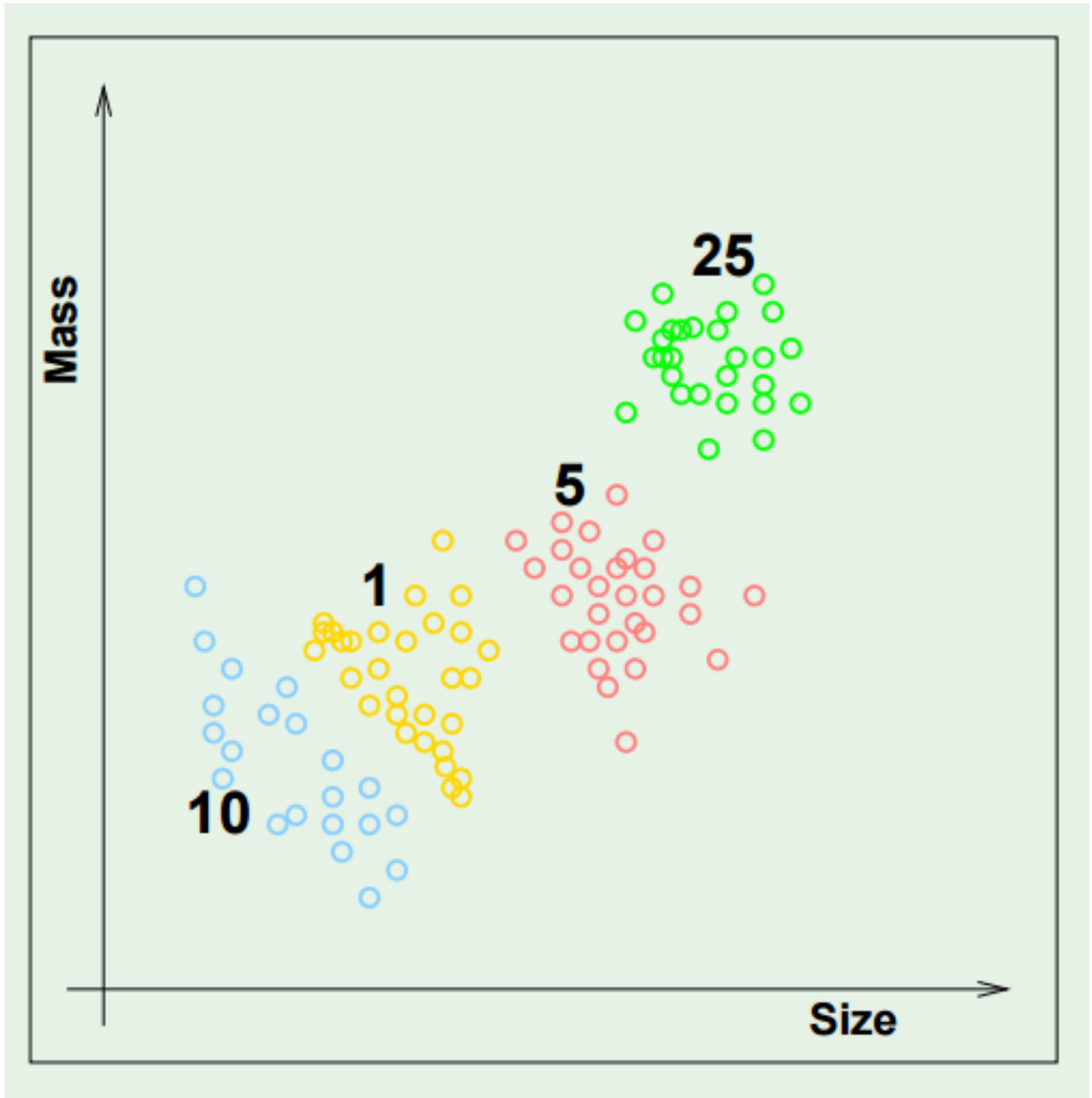
Types of Supervised Learning

- **Regression**
    - The outcome/label we are trying to predict is continuous.
    - Examples: price, blood pressure, etc.
- **Classification**
    - The outcome/label we are trying to predict is categorical
    - Examples: Spam/Ham, cancer class of tissue sample, etc.

The type of supervised learning problem has nothing to do with the features; only the response matters!

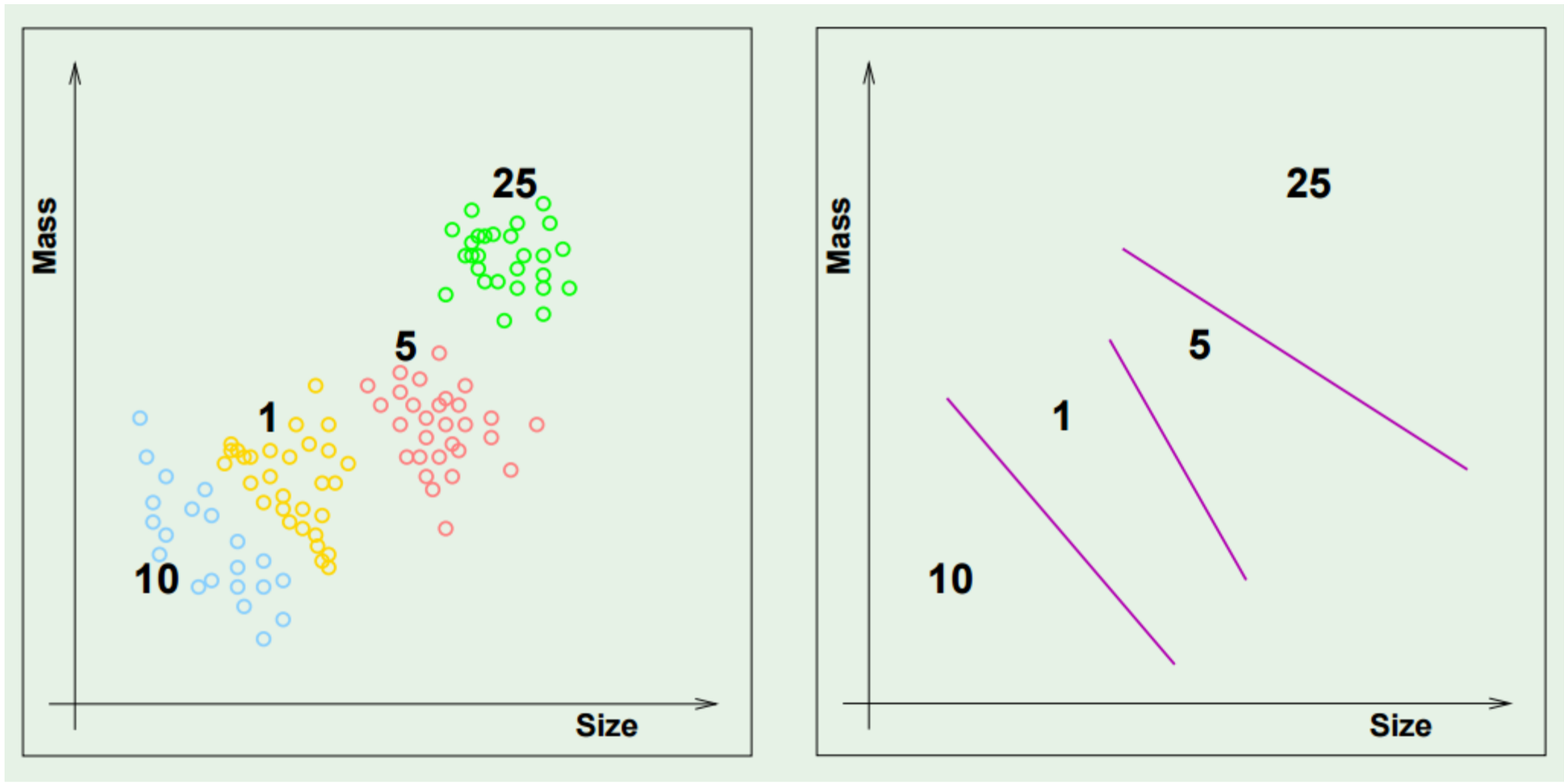
Supervised Learning Example: Coin Classifier

- **Observations:** Coins
- **Features:** Size and mass.
- **Response:** Hand-labeled coin type



Supervised Learning Example: Coin Classifier

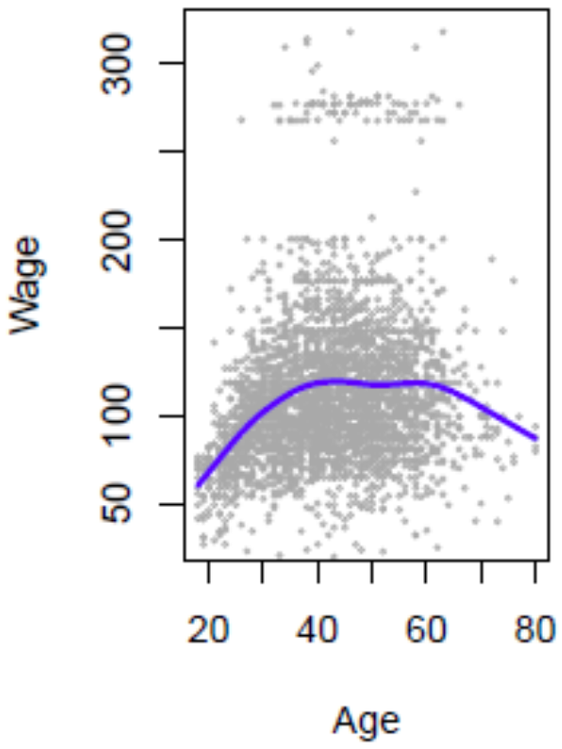
- Train a machine learning model using labeled data.
    - The model learns the relationship between the features and the coin type.
- Make predictions on new data for which the response is unknown.
    - Give the model a new coin, and it will predict the coin type automatically.

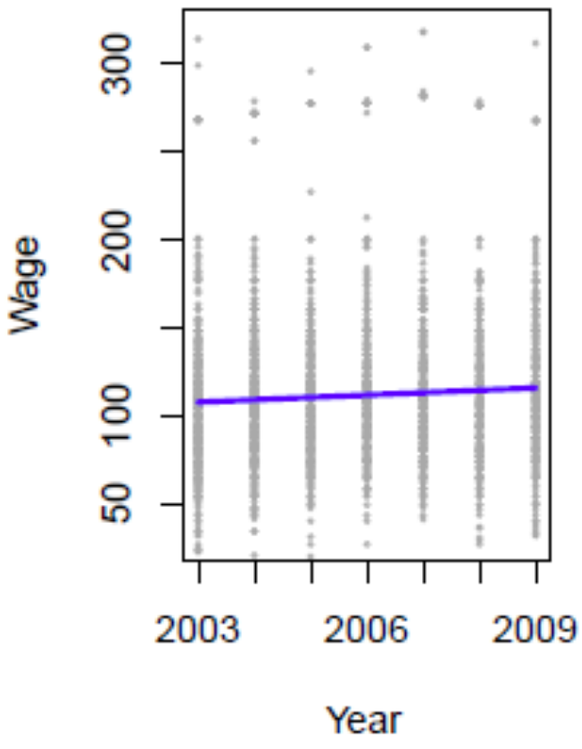


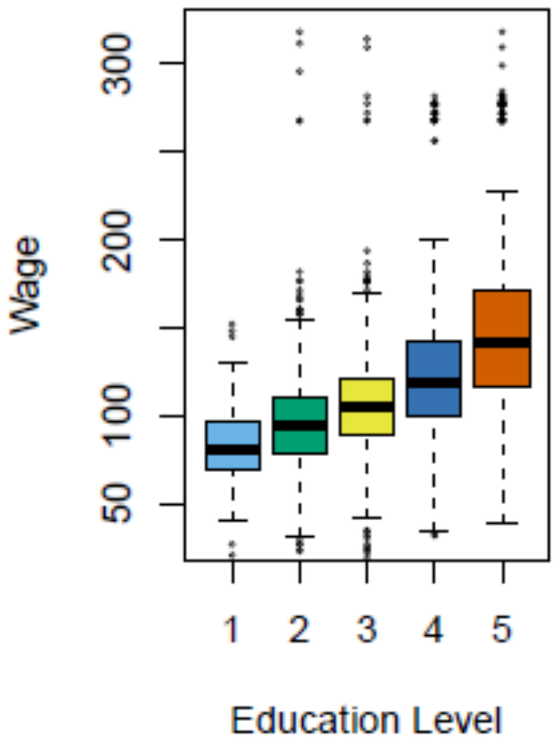
Practice: Regression or Classification?

- With a nearby partner (or two), decide if the problems below are classification, regression, or both.

1. Predict salary using demographic data:

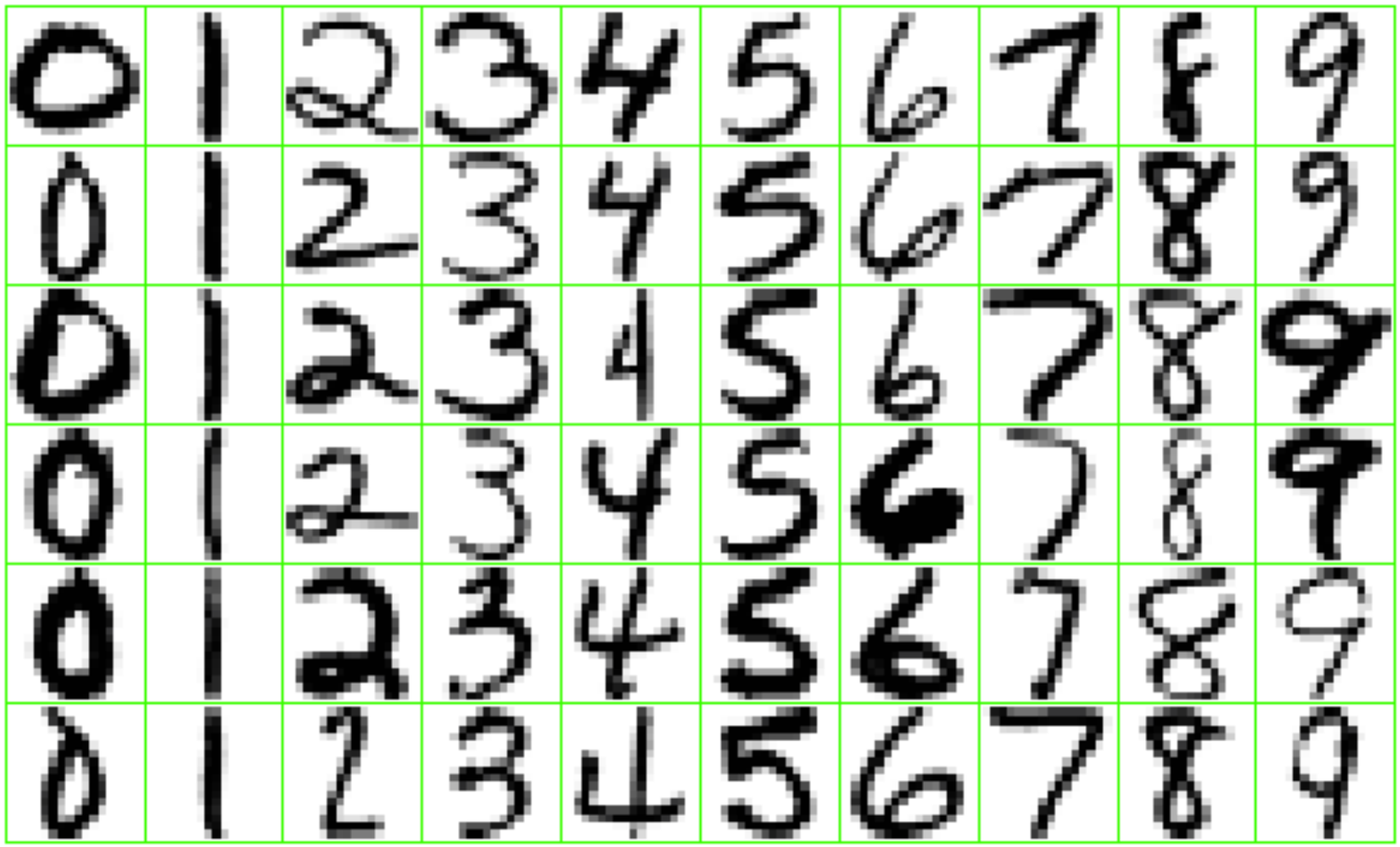






Practice: Regression or Classification?

2. Identify the numbers in a handwritten zip code:



Practice: Regression or Classification?

3. Consider the following problem (http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695):

- **Problem:** Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick.

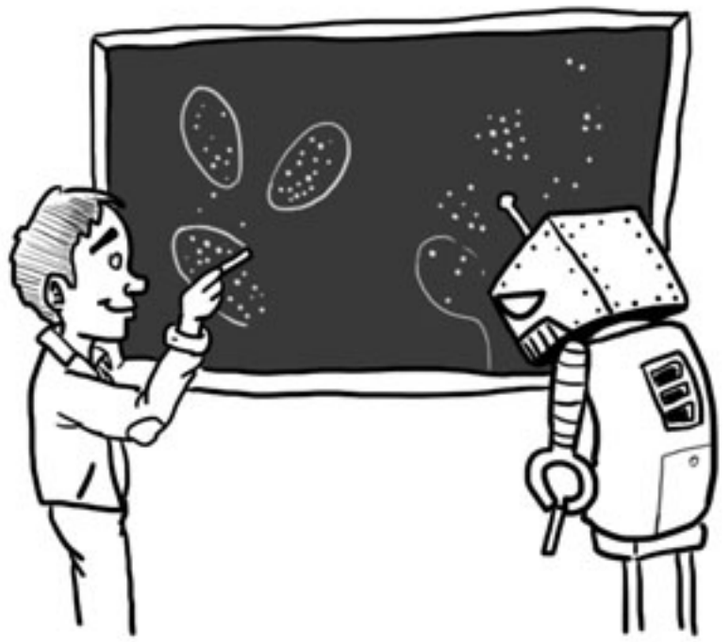
- **Goal:** Detect subtle patterns in the data that predict infection before it occurs.

- **Data:** 16 vital signs such as heart rate, respiration rate, blood pressure, etc.

- **Impact:** The model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear.

Unsupervised Learning

- Unsupervised learning has some clear differences from supervised learning. With unsupervised learning:
    - Data is not labeled. No response variable, only observations with features
    - There is no “right answer”
    - Extracts structure from data
    - Its goal is “representation”



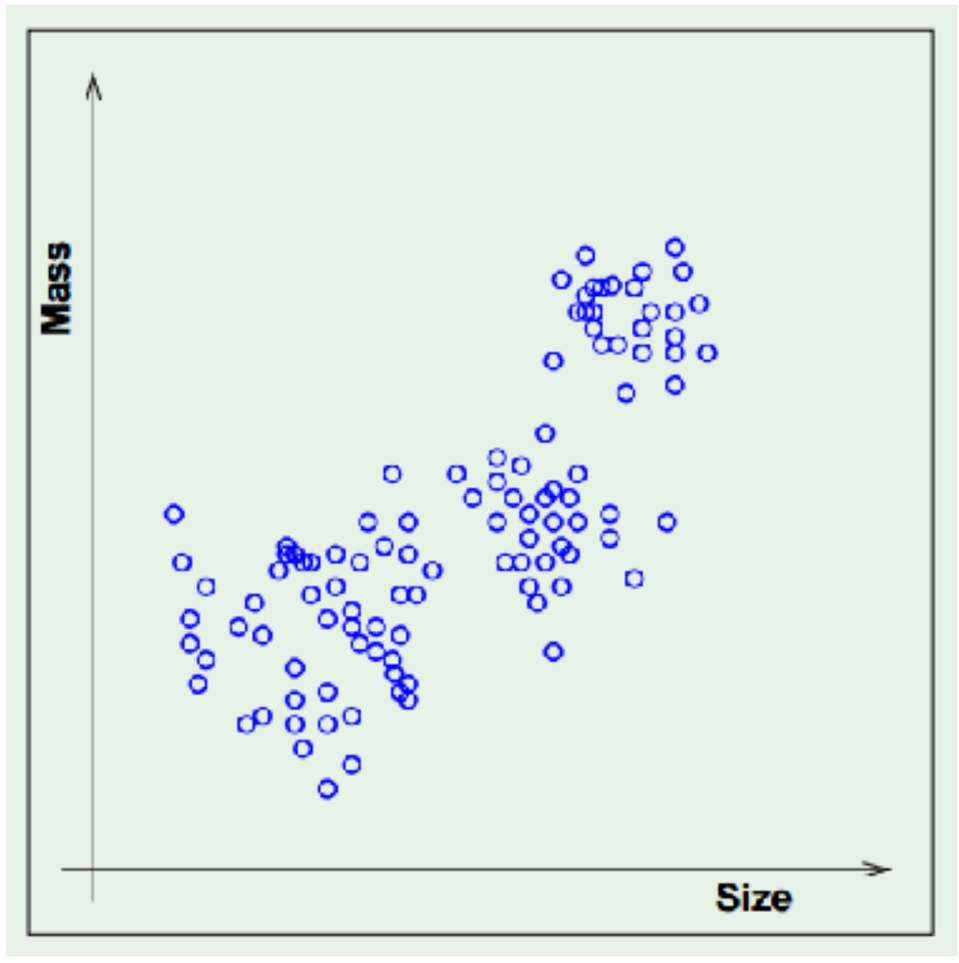
- Example: Segmenting grocery store shoppers into “clusters”/ groups that exhibit similar behaviors

Types of Unsupervised Learning

- **Clustering:** Groups “similar” data points together
- **Dimensionality Reduction:** Reduce the dimensionality of a data set by extracting features that capture most of the variance in the data

Unsupervised Learning: Clustering Example

- Considering our coin example:
    - Cluster the coins based on “similarity.”
    - Inspect the grouping that the algorithm found
- Hopefully this would put the coins into four separate groups.

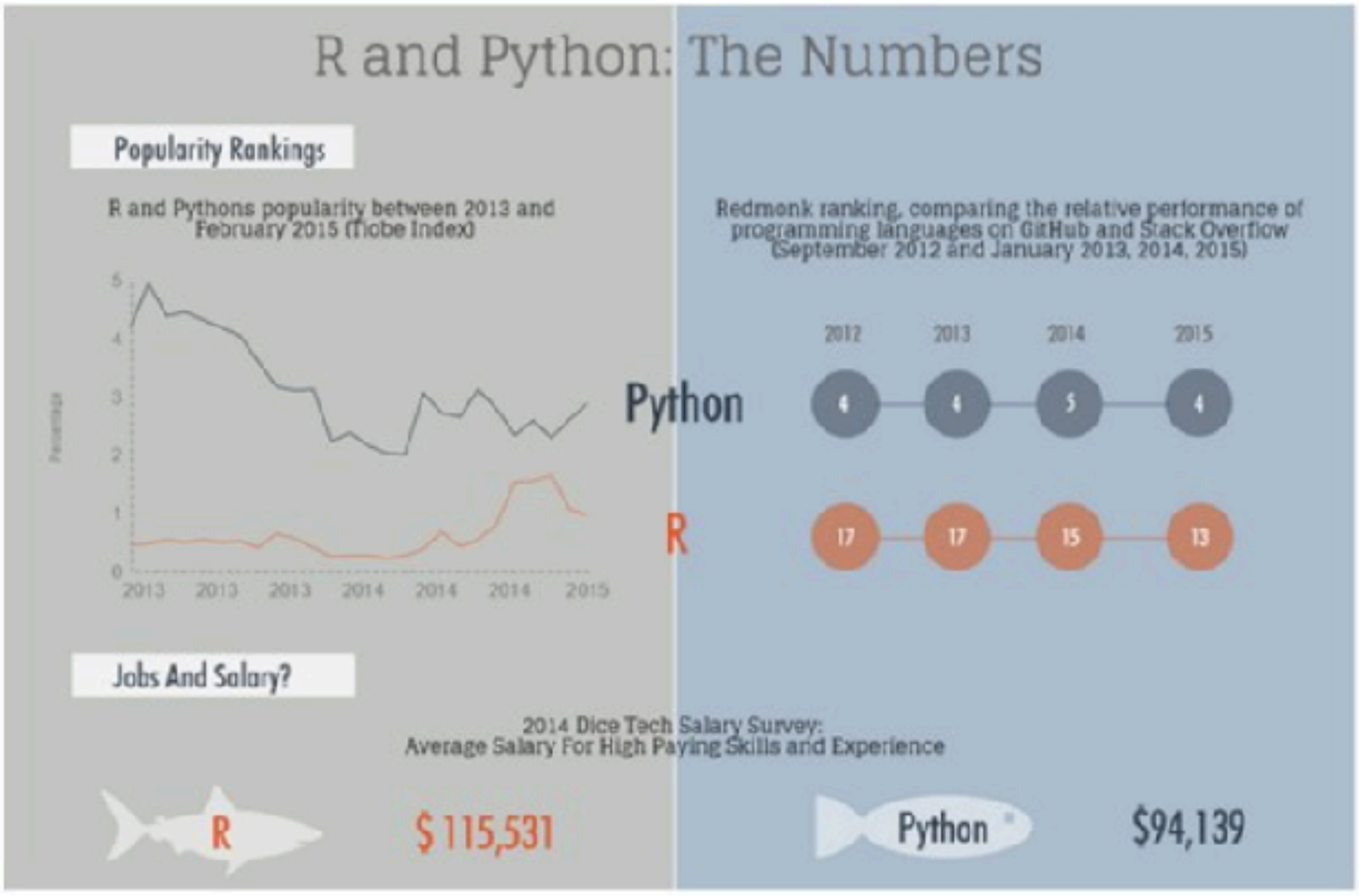


Supervised vs. Unsupervised Learning: Summary

Criteria	Supervised	Unsupervised
Prior Knowledge of label	Yes	No
Use Case	Classify new samples into known classes	Suggest groups based on the patterns in the data
Algorithms	Regression/Classifications	Clustering
Complexity	Often low to medium	Often high

Tools

- **Python versus R:**
  - No clear winner; both have their pros and cons and equally popular
  - Python is more popular between programmer community while R is more popular among statistician



Tools

- **Reasons for Choosing Python:**
  - It was created for simplicity and readability
  - It allows for rapid prototyping and ease of production
  - Strong support for machine learning libraries (scikit-learn)

Tools

- **scikit-learn (Author: David Cournapeau)**
  - Typically, we won't be implementing machine learning algorithms from scratch
  - Scikit-learn, referred to as "sklearn," is a popular machine learning library in Python.
  - Its benefits include ease of use and great documentation

```
from sklearn import datasets
from sklearn.model_selection import cross_val_predict
from sklearn import linear_model
import matplotlib.pyplot as plt

lr = linear_model.LinearRegression()
boston = datasets.load_boston()
y = boston.target

# cross_val_predict returns an array of the same size as `y` where each entry
# is a prediction obtained by cross validation:
predicted = cross_val_predict(lr, boston.data, y, cv=10)
```

Summary

- By now, you should be able to answer the following questions easily:
  - What is data science?
  - What is the data science workflow?
  - What is Machine Learning and its place in data science workflow

Demo

- A real small world data science project!

Extras

- Let's say we are a real estate agent looking to price a house using only its square feet. We know there are other features that can highly influence this outcome, but we are just focusing on the square footage for now. We know that, as square footage increases, so does price. At this point you may be thinking that a simple algebra equation could be useful; one that helps us price the house by its square footage.

- Recently we sold a house whose square footage was 2,500 for about $285,000. If we apply this information to a normal linear equation (Y = mx + b) we can create a simple algorithm to help us price a house.

285,000 = 2,500x + b —> x = 114, b = 0

- **Final Algorithm:**

Price = 114x

- Typically, our models will be more complex, and we'll consider a greater amount of prior data to help us develop a final algorithm

Extras

- **Algorithm Training :** In our example, we used previously known information to find our coefficients. This action is also known as "Training." But let's make something clear:

- Model building would be the task of constructing an actual algorithm.

    - This is the linear model of Y = mx + b

- Training involves figuring out the coefficient and the Y intercept the model uses for _our intended purpose_.

    - The coefficients uncovered via training were m= 114 and b=0