Course Structure

Day 1 - The Basics

- Data science basics

Day 2 - The Algorithms

What *is* Data Science?

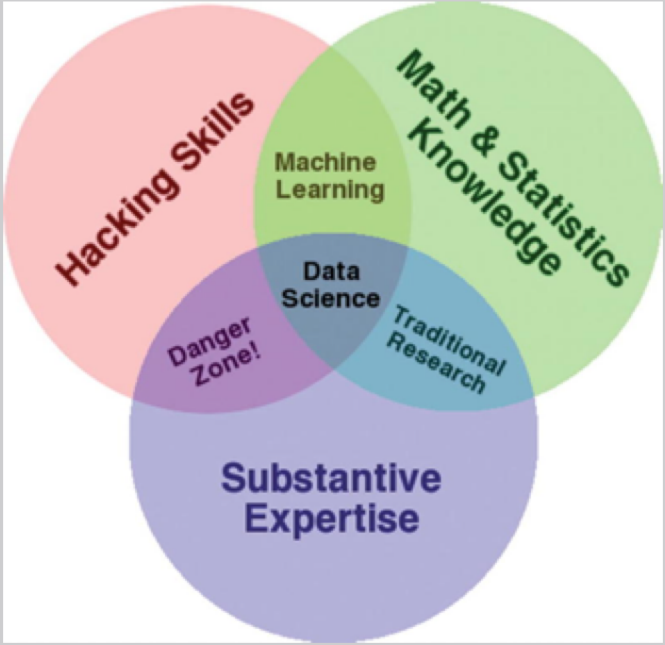
What is Data Science?

- The study of data (duh)

What is Data Science?

- The study of data (duh)
- Using data to help people and businesses

What is Data Science?



What is Data Science?

The technology is not the field

Data science vs. Big data vs. Data analytics vs….

Data science vs. Big data vs. Data analytics vs….

- Data science - using statistics, machine-learning, data gathering/cleaning strategies, etc. to gain insight from data

Data science vs. Big data vs. Data analytics vs….

- Data science - using statistics, machine-learning, data gathering/cleaning strategies, etc. to gain insight from data
- Big data - strategies for working with huge amounts of data effectively

Data science vs. Big data vs. Data analytics vs….

- Data science - using statistics, machine-learning, data gathering/cleaning strategies, etc. to gain insight from data
- Big data - strategies for working with huge amounts of data effectively
- Data analytics - automating processes for drawing conclusions from data

Data science vs. Big data vs. Data analytics vs….

Yes, these definitions are vague

Data science vs. Big data vs. Data analytics vs….

Most people who get hired for data science/big data/data analytics will end up doing all three

Data Science Examples

Recommender Systems - recommend new products/videos/songs based on previous behavior

Data Science Examples

Targeted Advertising - based on all the data we have about someone, what can we get them to buy?

Data Science Examples

Medical Diagnosis/Treatment - based on patient history, biological data, x-ray photos, analyzing DNA etc., we can recommend treatments, suggest diagnoses, and so on

Data Science Examples

Medical Diagnosis/Treatment - based on patient history, biological data, x-ray photos, analyzing DNA etc., we can recommend treatments, suggest diagnoses, and so on

Data Science Examples

Loan Default Prediction - based on data about a person, predict the probability that they'll be able to pay back a loan

Data Science Examples

Speech Recognition/Machine Vision/Etc. - allow computers to “sense” and interpret the surrounding world in a similar way as humans

Some Questions

How many of you…

Some Questions

How many of you…

- have a “rewards card” with a supermarket or retail store?

Some Questions

How many of you…

- have a “rewards card” with a supermarket or retail store?
- have a Facebook account?

Some Questions

How many of you…

- have a “rewards card” with a supermarket or retail store?
- have a Facebook account?
- use Google Maps?

Some Questions

How many of you…

- have a “rewards card” with a supermarket or retail store?
- have a Facebook account?
- use Google Maps?
- get “discounted” medical insurance premiums by handing over some medical data?

Some Questions

How many of you…

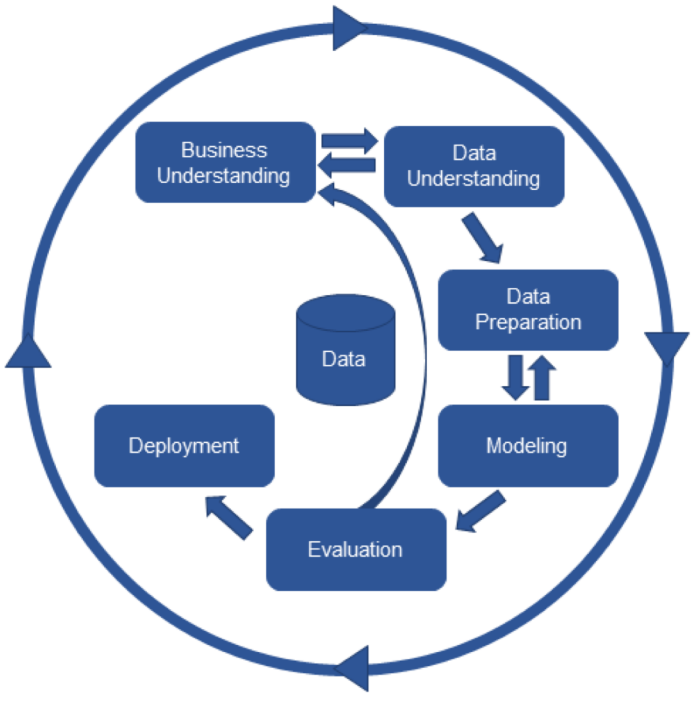
- have a “rewards card” with a supermarket or retail store?
- have a Facebook account?
- use Google Maps?
- get “discounted” medical insurance premiums by handing over some medical data?
- have a “Snapshot” car sensor from your car insurance company?

Data Science Examples

Data is a business asset

CRISP-DM: The Data Science Cycle

CRISP-DM: The Data Science Cycle



CRISP-DM: The Data Science Cycle

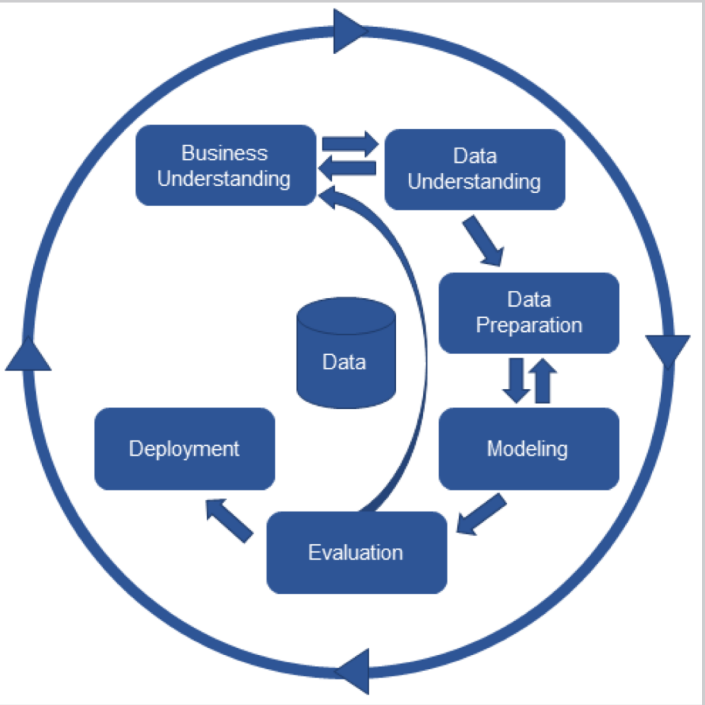
Business Understanding - The business needs to know what it actually wants

CRISP-DM: The Data Science Cycle

Business Understanding - The business needs to know what it actually wants

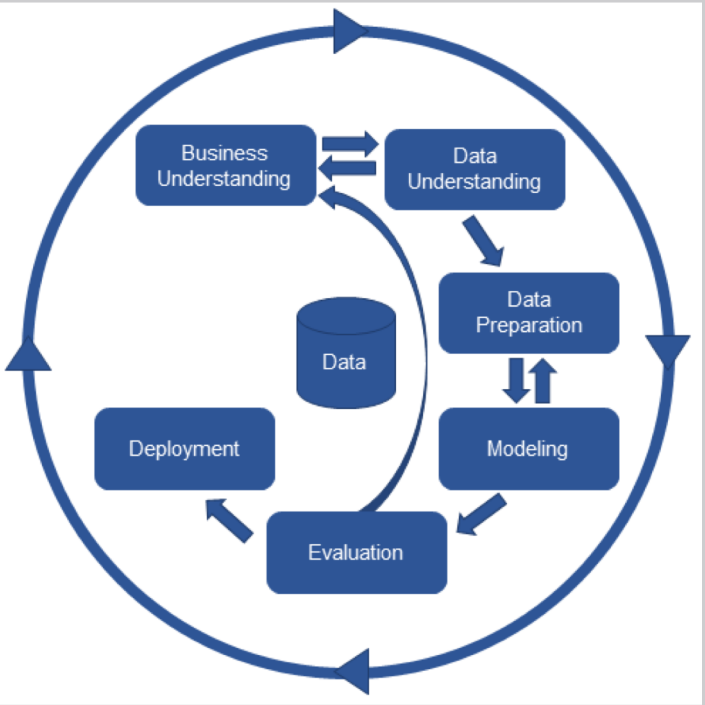
Data Understanding - are the business' wants actually possible with the current dataset?

CRISP-DM: The Data Science Cycle



Data Preparation - transform raw data into the proper form for the problem at hand

CRISP-DM: The Data Science Cycle



Data Preparation - transform raw data into the proper form for the problem at hand

Modeling - the "fun part", use stats, machine learning, etc. to create a model that can be used on new data

CRISP-DM: The Data Science Cycle

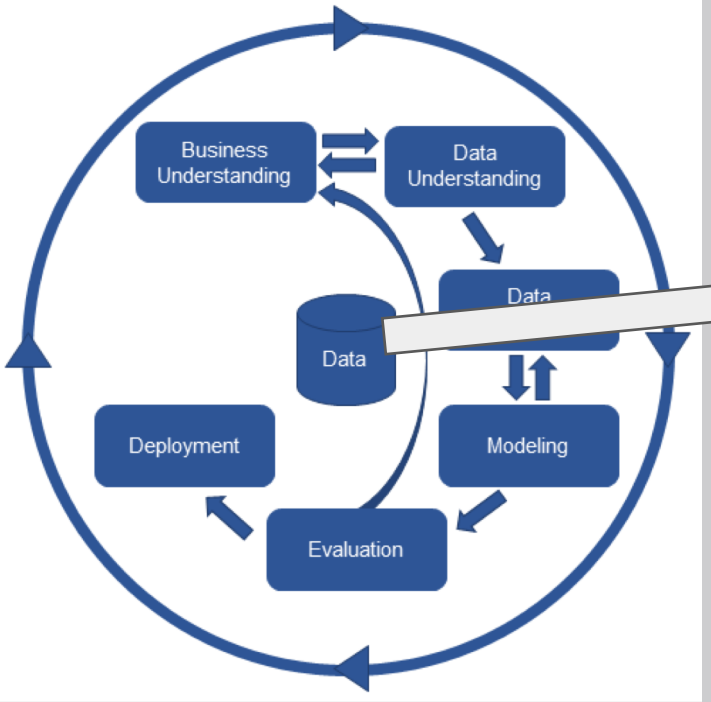
Evaluation - do the patterns we've discovered in our data hold up in real-life?

CRISP-DM: The Data Science Cycle

Evaluation - do the patterns we've discovered in our data hold up in real-life?

Deployment - send our models out into the real-world

CRISP-DM: The Data Science Cycle



And let's not forget the data here!

Types of Problems in Data Science

There are many algorithms in Data Science, but relatively few types of problems

Types of Problems in Data Science

1. Classification - sort entities into actionable groups

Types of Problems in Data Science

1. Classification - sort entities into actionable groups
2. Regression - predict the value of some variable based on the values of some attributes

Types of Problems in Data Science

1. Classification - sort entities into actionable groups
2. Regression - predict the value of some variable based on the values of some attributes
3. Similarity matching - identify similar items in a dataset and use that to find other similar items outside that dataset

Types of Problems in Data Science

1. Classification - sort entities into actionable groups
2. Regression - predict the value of some variable based on the values of some attributes
3. Similarity matching - identify similar items in a dataset and use that to find other similar items outside that dataset
4. Clustering - find naturally occurring groups of individuals in a dataset

Types of Problems in Data Science

1. Classification - sort entities into actionable groups
2. Regression - predict the value of some variable based on the values of some attributes
3. Similarity matching - identify similar items in a dataset and use that to find other similar items outside that dataset
4. Clustering - find naturally occurring groups of individuals in a dataset
5. Co-Occurrence Grouping - find items that occur together frequently in transactions

Types of Problems in Data Science

1. Classification - sort entities into actionable groups
2. Regression - predict the value of some variable based on the values of some attributes
3. Similarity matching - identify similar items in a dataset and use that to find other similar items outside that dataset
4. Clustering - find naturally occurring groups of individuals in a dataset
5. Co-Occurrence Grouping - find items that occur together frequently in transactions
6. Profiling - define a “typical behavior” of a group/individual

Types of Problems in Data Science

1. Classification - sort entities into actionable groups
2. Regression - predict the value of some variable based on the values of some attributes
3. Similarity matching - identify similar items in a dataset and use that to find other similar items outside that dataset
4. Clustering - find naturally occurring groups of individuals in a dataset
5. Co-Occurrence Grouping - find items that occur together frequently in transactions
6. Profiling - define a “typical behavior” of a group/individual
7. Link Prediction - predict connections and the strength of those connections between data items

Types of Problems in Data Science

1. Classification - sort entities into actionable groups
2. Regression - predict the value of some variable based on the values of some attributes
3. Similarity matching - identify similar items in a dataset and use that to find other similar items outside that dataset
4. Clustering - find naturally occurring groups of individuals in a dataset
5. Co-Occurrence Grouping - find items that occur together frequently in transactions
6. Profiling - define a “typical behavior” of a group/individual
7. Link Prediction - predict connections and the strength of those connections between data items
8. Data Reduction - reduce the size of large data, without losing any of its “important aspects”

Types of Problems in Data Science

1. Classification - sort entities into actionable groups
2. Regression - predict the value of some variable based on the values of some attributes
3. Similarity matching - identify similar items in a dataset and use that to find other similar items outside that dataset
4. Clustering - find naturally occurring groups of individuals in a dataset
5. Co-Occurrence Grouping - find items that occur together frequently in transactions
6. Profiling - define a “typical behavior” of a group/individual
7. Link Prediction - predict connections and the strength of those connections between data items
8. Data Reduction - reduce the size of large data, without losing any of its “important aspects”
9. Causal Modeling - identify cause and effect in datasets

Data Difficulties

- Null values - `null, NULL, N/A…`

Data Difficulties

- Null values - `null, NULL, N/A…`
- Blank spaces - lots of them

Data Difficulties

- Null values - `null, NULL, N/A…`
- Blank spaces - lots of them
- Parse errors - we need to handle them

Data Difficulties

- Null values - `null, NULL, N/A…`
- Blank spaces - lots of them
- Parse errors - we need to handle them
- Outliers - “funky” values

Building Models in Data Science

Models are simplified representations of reality

Supervised vs. Unsupervised Learning

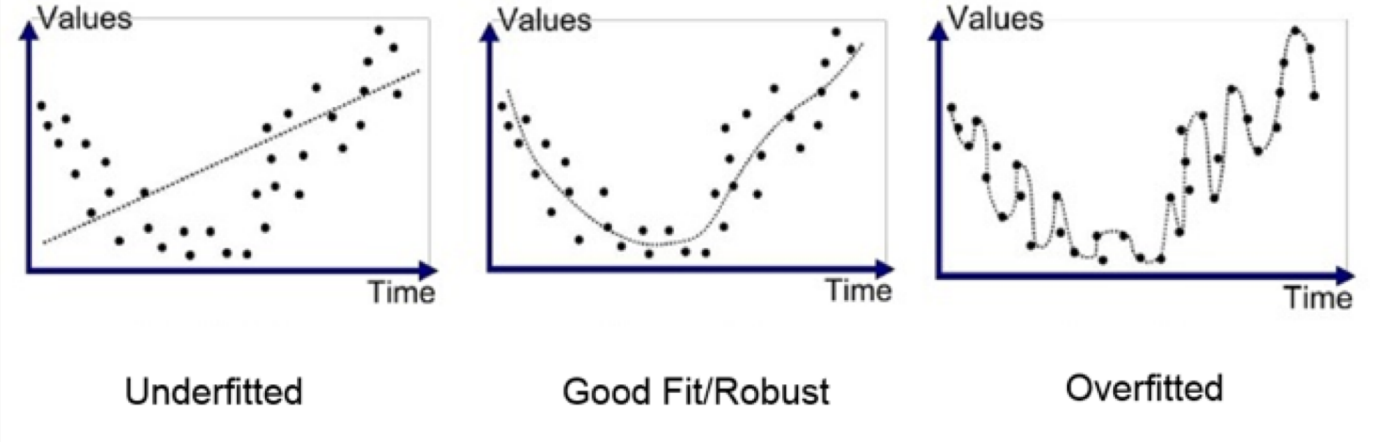
Supervised vs. Unsupervised Learning

- **Unsupervised learning** - “Do our customers naturally fall into different groups?”

Supervised vs. Unsupervised Learning

- **Unsupervised learning** - “Do our customers naturally fall into different groups?”
- **Supervised learning** - “Can we find groups of customers that are very likely to buy X?”

Overfitting



Overfitting

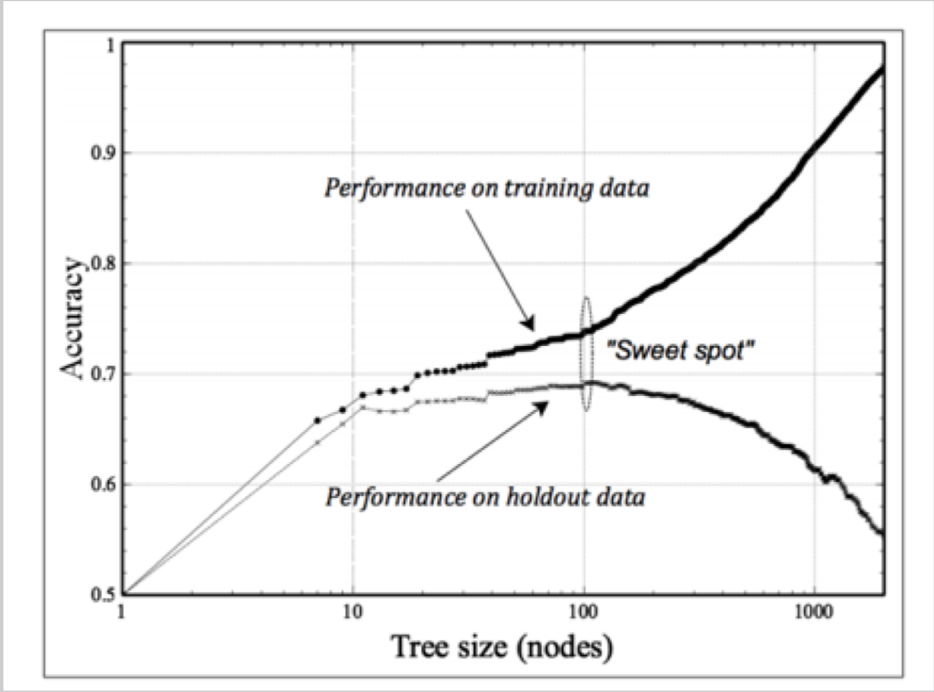
"If you stare too long at a dataset, you will find a pattern"

Overfitting

Base rate - the accuracy a given model would have if it just guessed all the time

Overfitting

Overfitting



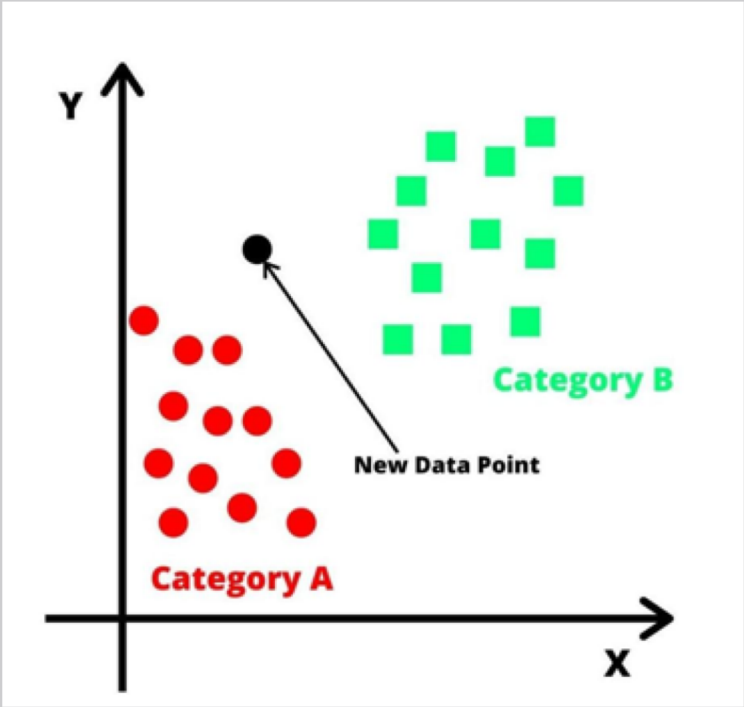
Avoiding Overfitting

- **Holdout data** - keep some data to the side for validation
- **Cross validation** - build the model several times, each time with a different subset as holdout data
- **Nested holdout testing** - two “levels” of holdout testing

Avoiding Overfitting

How representative is your training data of the actual population?

K-Nearest Neighbor Algorithm



Naive Bayes Classifier

$$P(A \mid B) = (P(B \mid A) * P(A)) / P(B)$$

Linear Regression