Primer to Data Science

Pri Oberoi, Star Ying

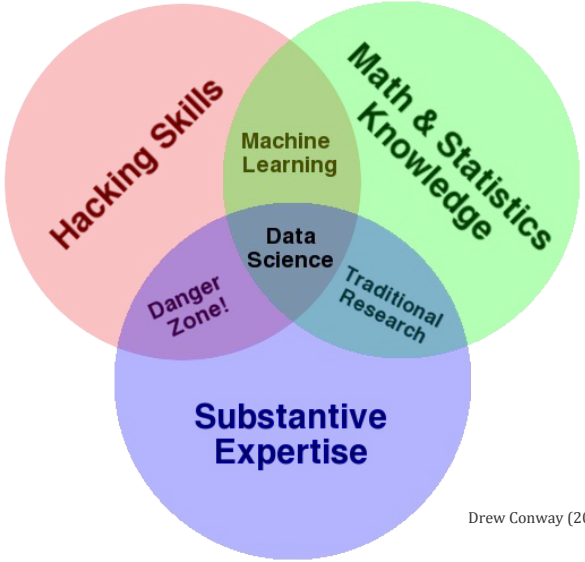
Get started:

https://github.com/StarCYing/open_data_day_dc

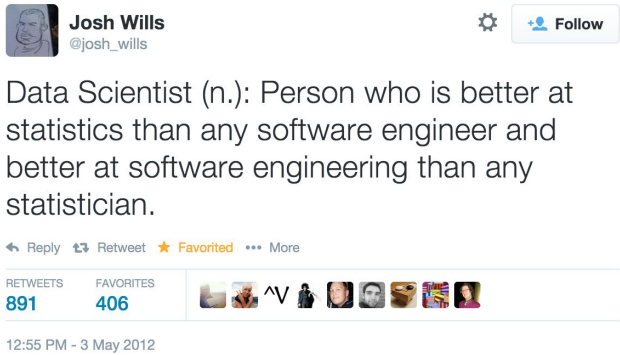
Follow the readme.md instructions to setup for the workshop

COMMERCE DATA SERVICE

The quick definition of data science



Drew Conway (2010)



Josh Wills
@josh_wills

Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

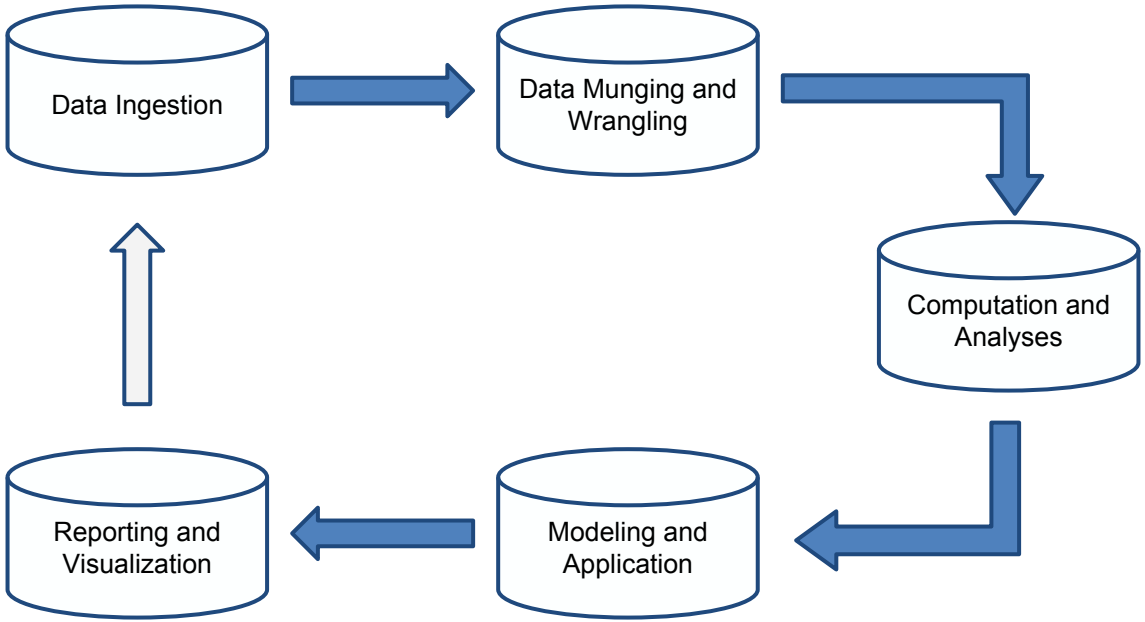
Reply Retweet Favorited More

RETWEETS 891 FAVORITES 406

12:55 PM - 3 May 2012

COMMERCE DATA SERVICE

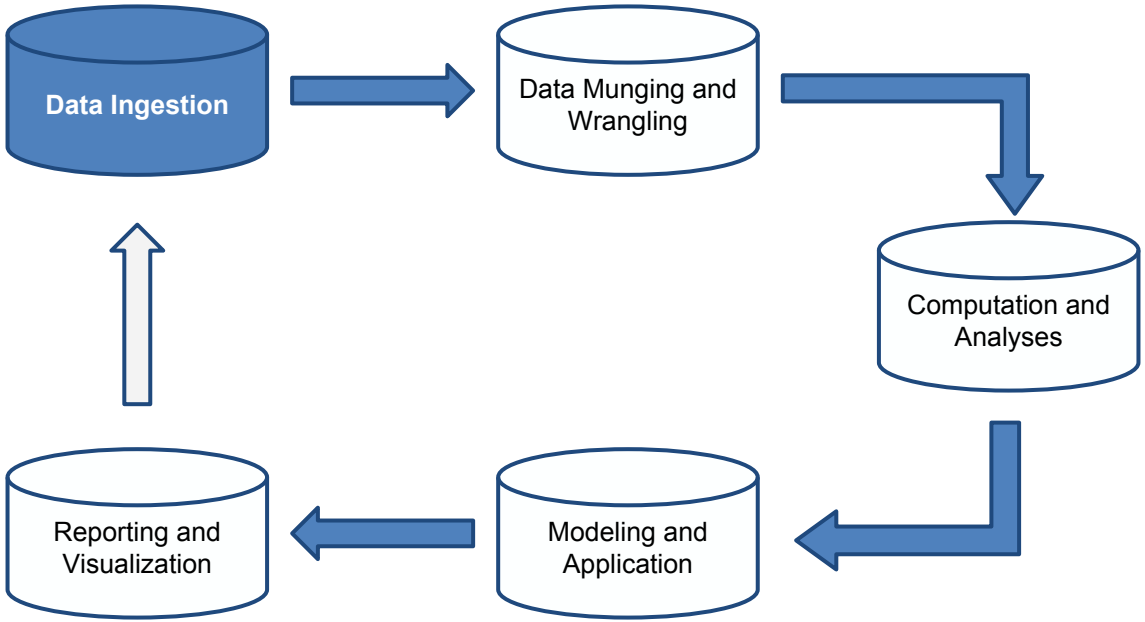
The quick definition of data science



COMMERCE DATA SERVICE

Data Ingestion

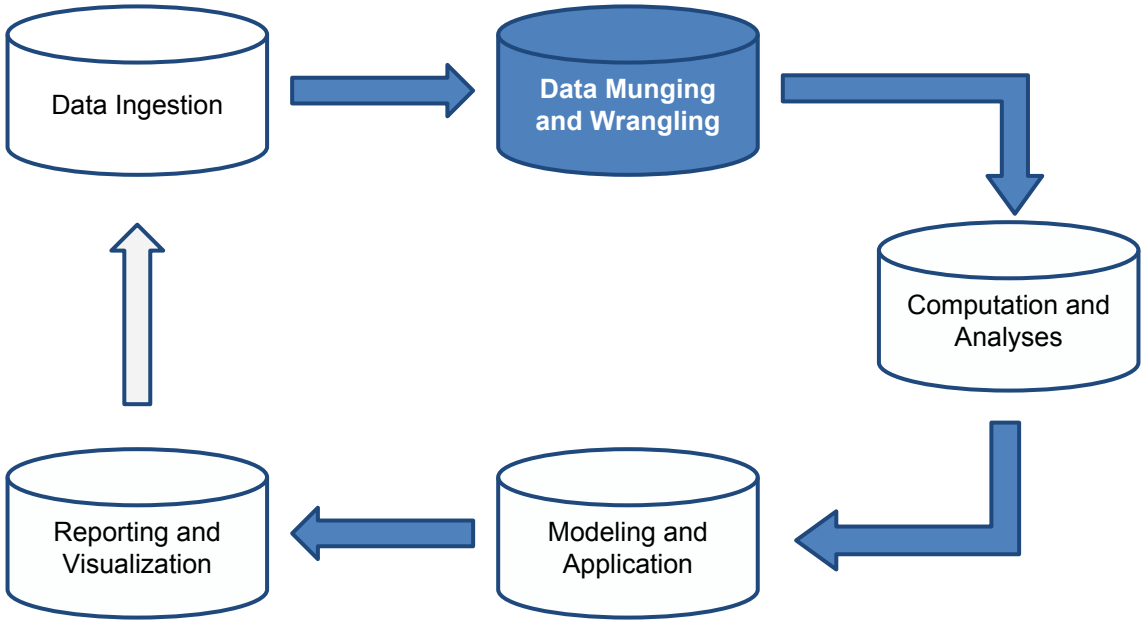
Means
Source
Question
Size
Velocity



COMMERCE DATA SERVICE

Data Munging and Wrangling

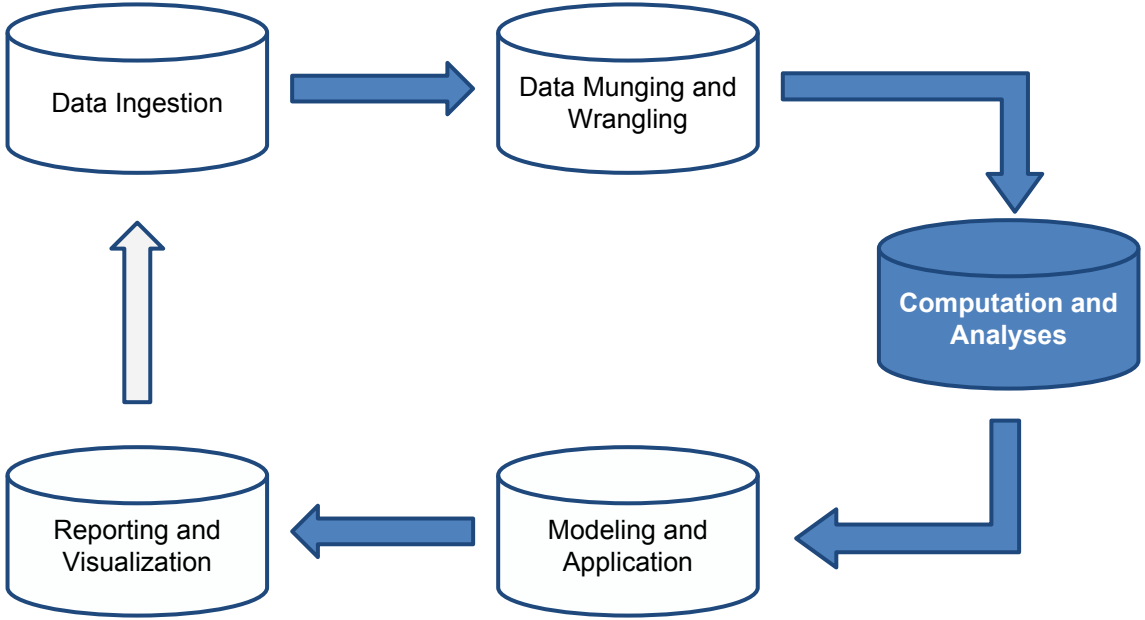
Warehouse
Extract
Transform
Filter
Aggregation
Training



COMMERCE DATA SERVICE

Computation and Analyses

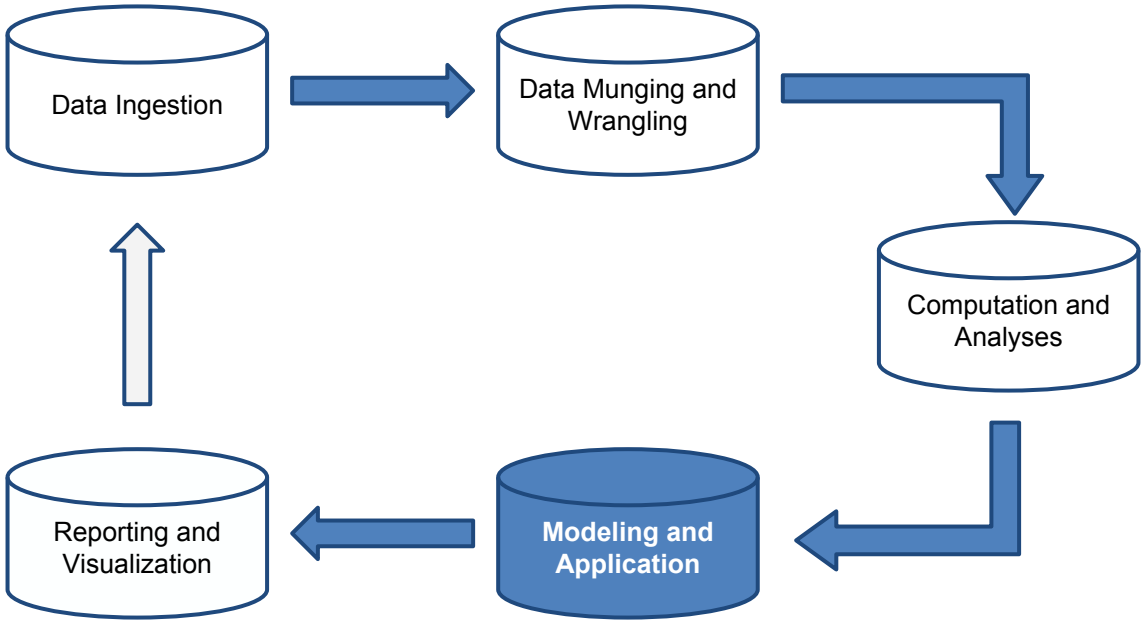
Hypothesis
Design
Method
Time



COMMERCE DATA SERVICE

Modeling and Application

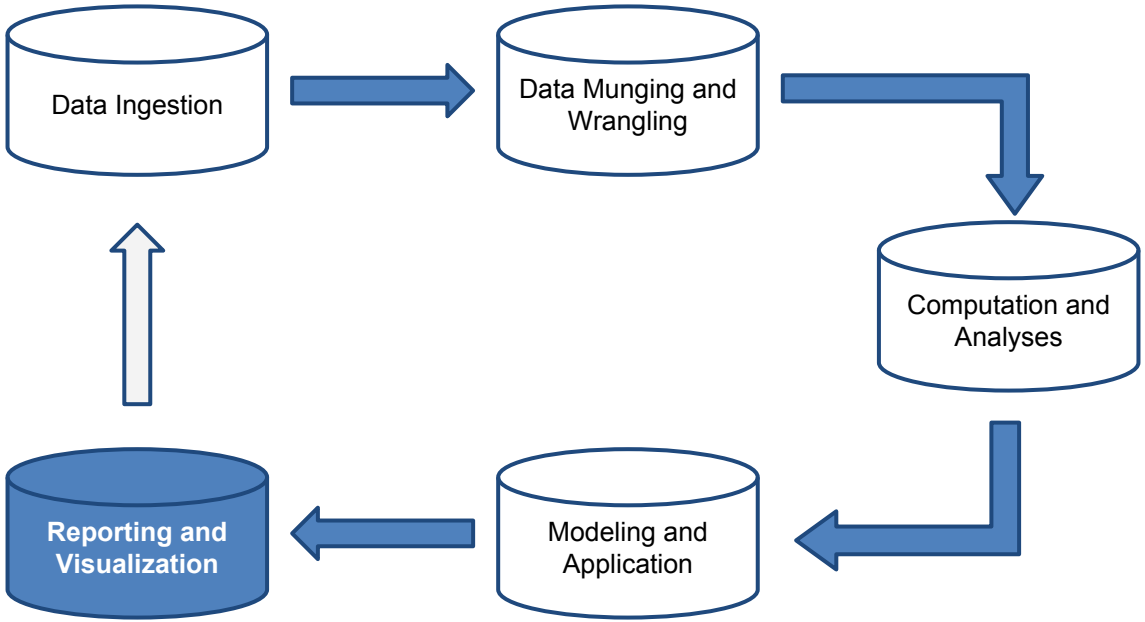
Supervised
Unsupervised
Regression
Classification
Clustering
Etc...



COMMERCE DATA SERVICE

Reporting and Visualization

Crucial
Active Learning
Error Detection
Mashups
Value



COMMERCE DATA SERVICE

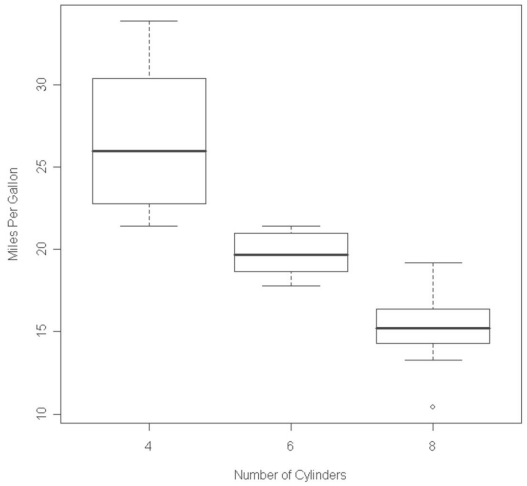
Modeling and Application

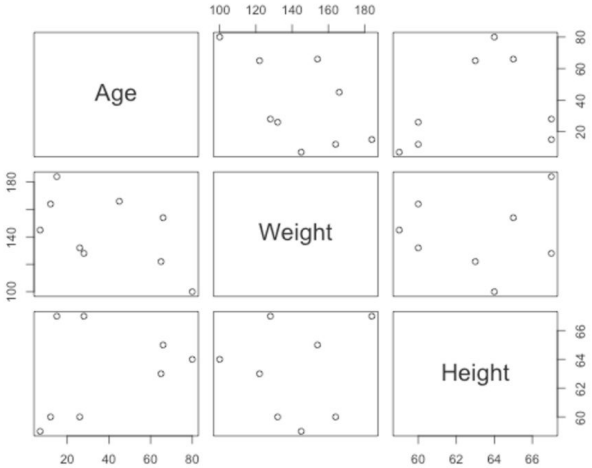
- Scope
  - What is the question/problem statement?
  - What data is available to answer this question?
- Explore
  - ‘Initial data touch’
  - Summary statistics
  - Simple visualizations
- Clean
- Model
  - choose a machine learning model (more on that in a minute)
- Evaluate
  - figure out if your model is any good

COMMERCE DATA SERVICE

Visualize

- Types
  - Boxplots
  - Scatter Plots
  - Tables
  - etc...
- Tools
  - R
  - Matplotlib

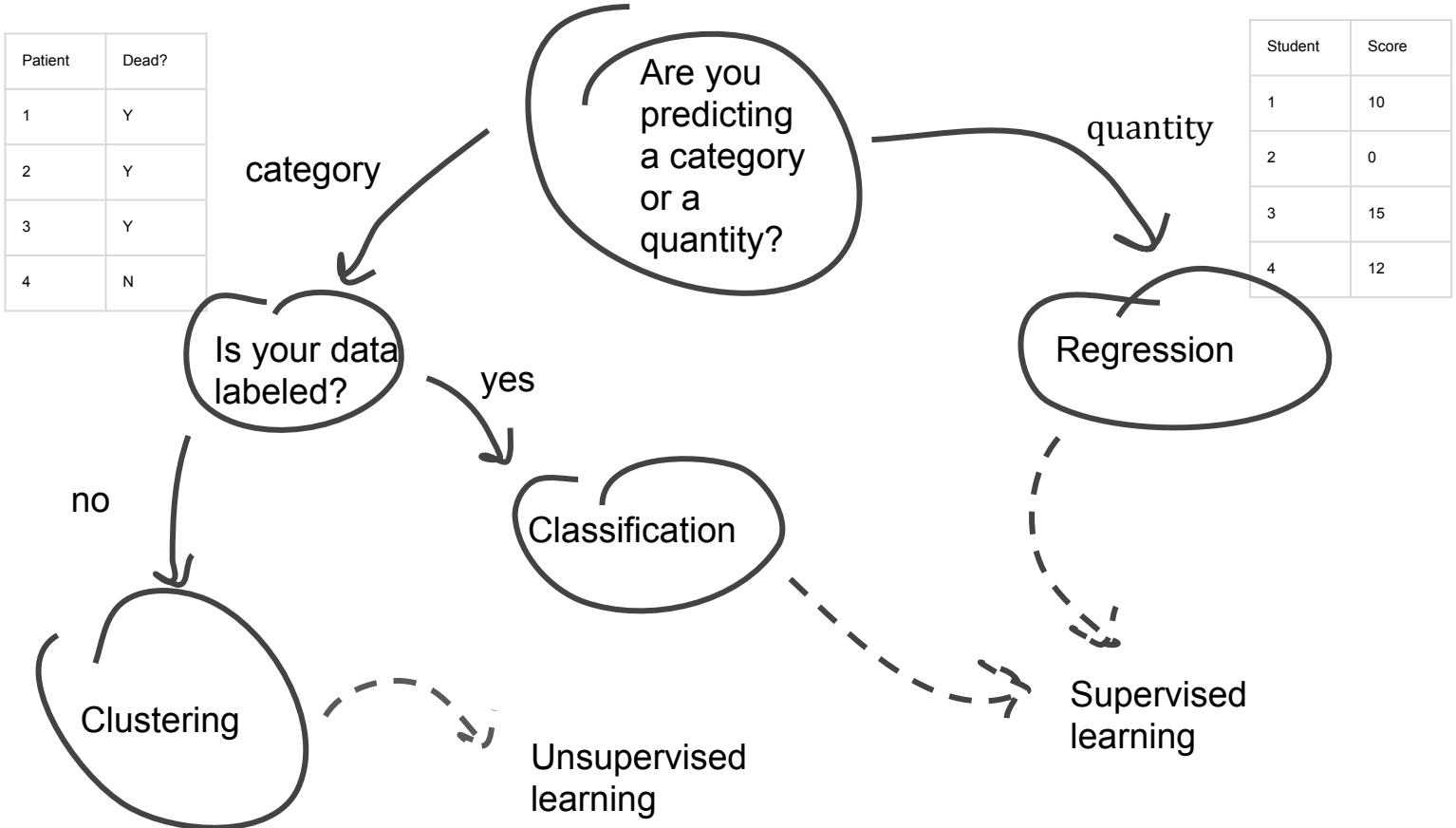




COMMERCE DATA SERVICE

Choose a machine learning method

Patient	Dead?
1	Y
2	Y
3	Y
4	N

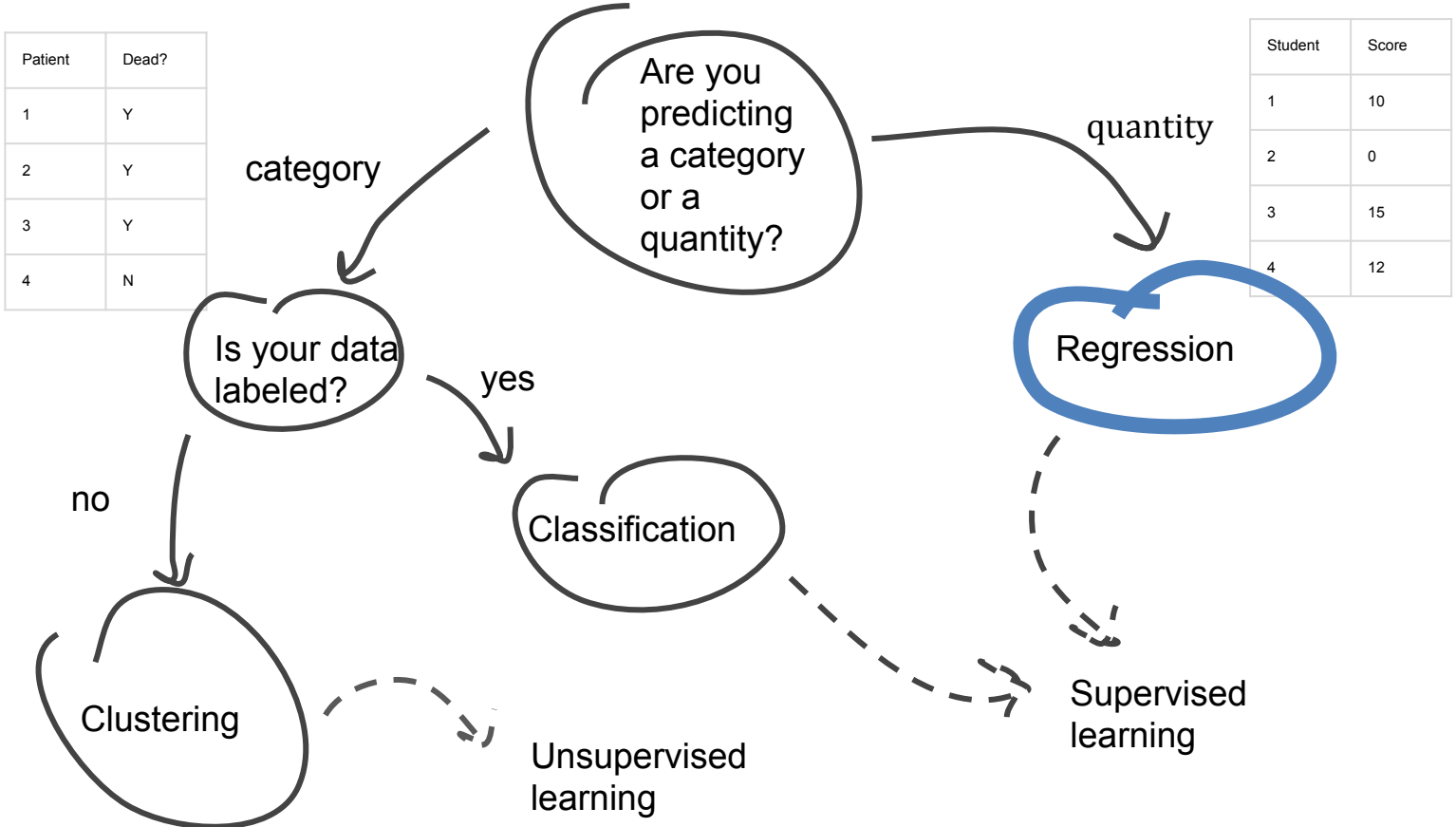


Student	Score
1	10
2	0
3	15
4	12

COMMERCE DATA SERVICE

Choose a machine learning method

Patient	Dead?
1	Y
2	Y
3	Y
4	N



Student	Score
1	10
2	0
3	15
4	12

Regression

Predict a numerical target using predictors (which can be numerical and categorical)

Example dataset: patient age, diagnosis, surgery date, discharge date, number of months survival post-discharge, where we are trying to predict how long a patient survives post-discharge.

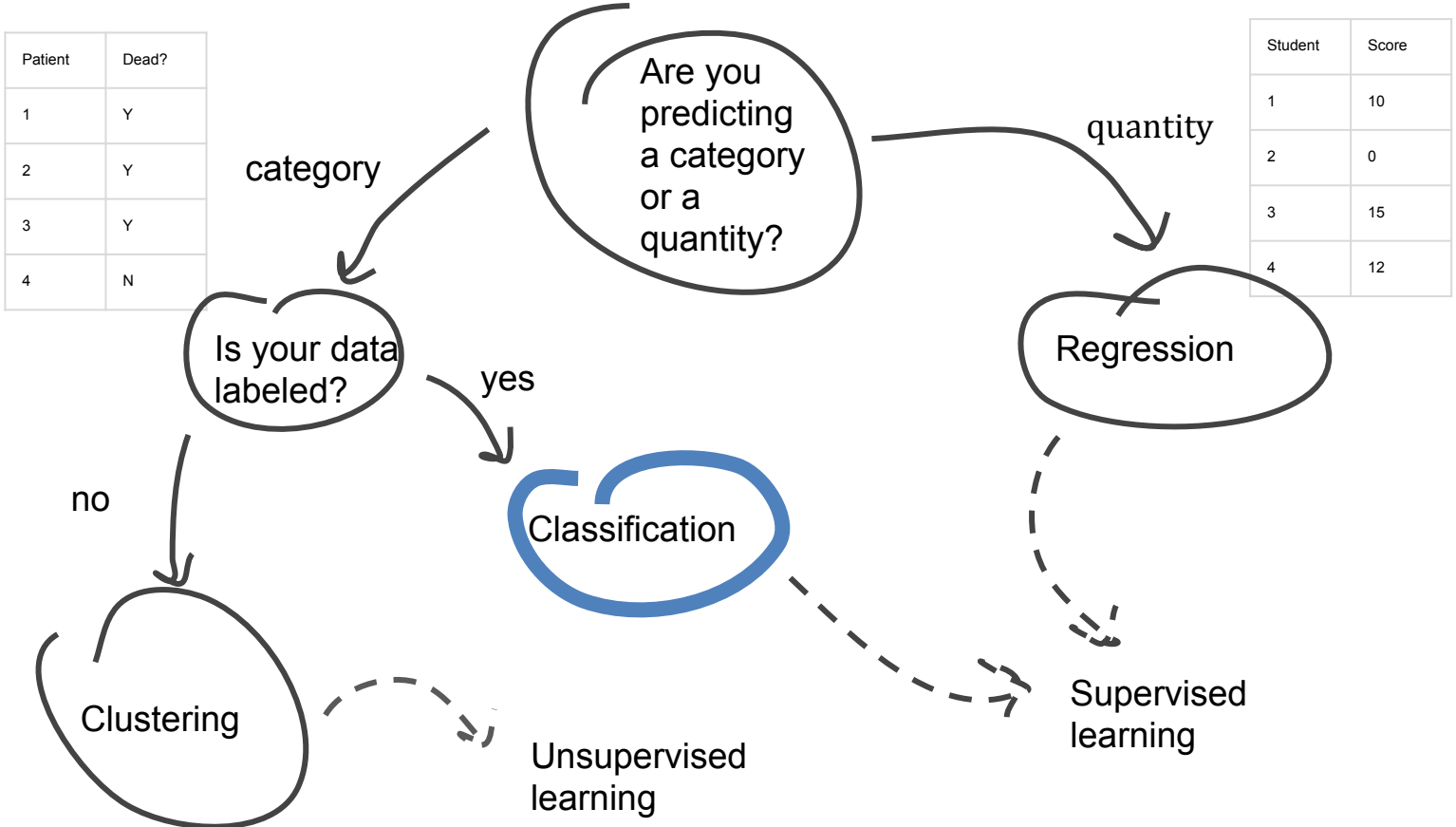
- Simple Linear Regression
- Multiple Linear Regression
- K-Nearest Neighbors Regression

COMMERCE DATA SERVICE

Choose a machine learning method

Patient	Dead?
1	Y
2	Y
3	Y
4	N

Student	Score
1	10
2	0
3	15
4	12



Classification

Identifying which category an object belongs to

Example dataset: patient age, diagnosis, surgery date, discharge date, label of whether the patient died, where we are trying to predict whether a patient will survive

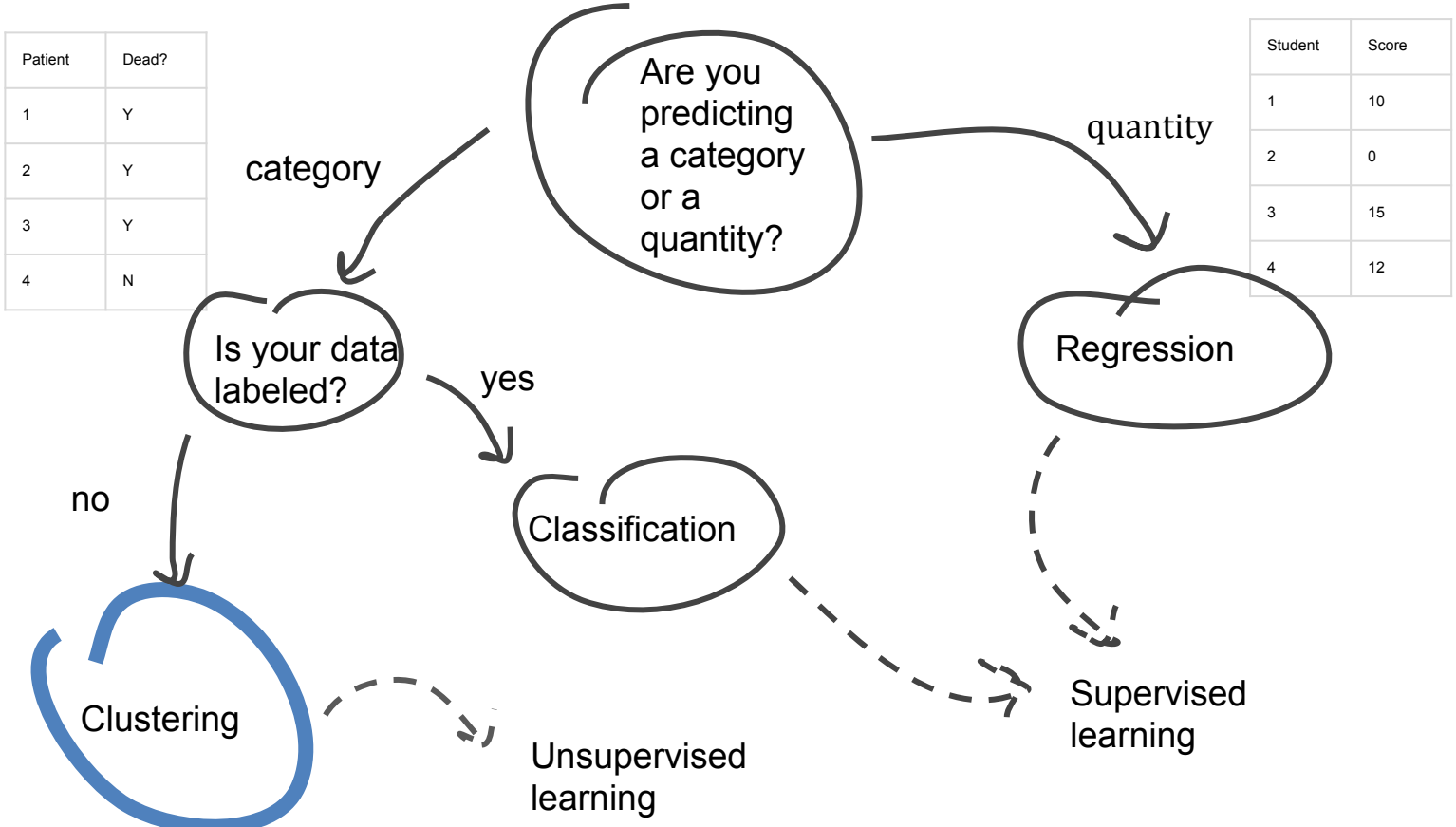
- Logistic Regression
- Multiple Logistic Regression
- Linear Discriminant Analysis
- K-Nearest Neighbors Classifier

COMMERCE DATA SERVICE

Choose a machine learning method

Patient	Dead?
1	Y
2	Y
3	Y
4	N

Student	Score
1	10
2	0
3	15
4	12



Clustering

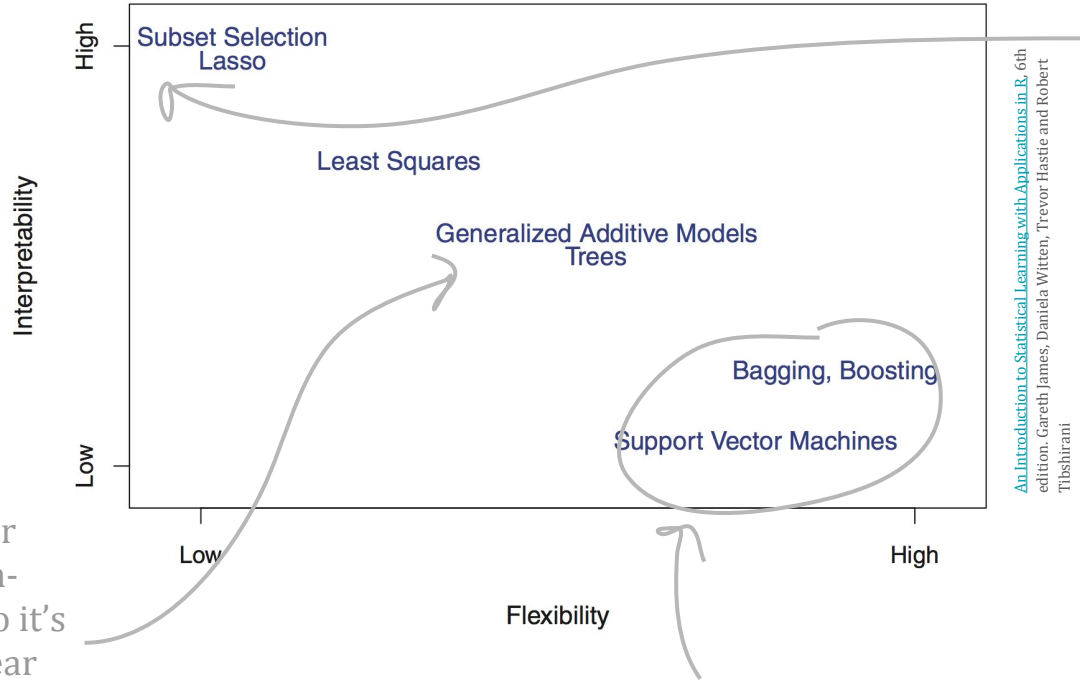
Grouping of similar objects into sets

Example datasets: clustering protein structures using amino acid sequence and secondary structures cluster structurally similar proteins together

- K-means
- Spectral Clustering
- Gaussian mixtures

COMMERCE DATA SERVICE

Prediction Accuracy versus Model Interpretability



Linear regression assumes a linear shape, which is pretty straightforward

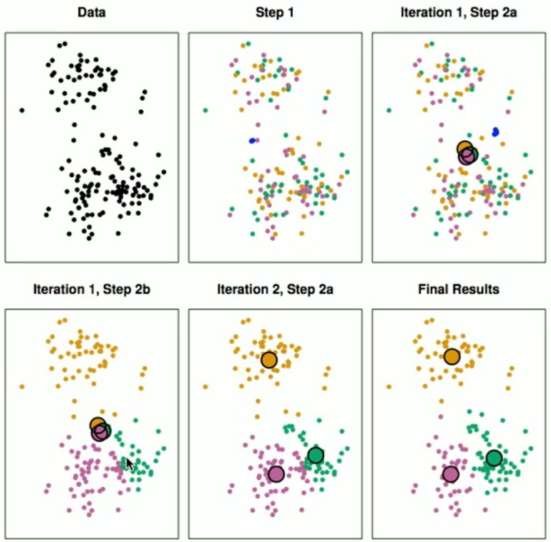
GAMs extend the linear model to allow for non-linear relationships. So it's more flexible than linear relationships but also less interpretable.

fully non-linear methods that are very flexible, but hard to interpret

COMMERCE DATA SERVICE

What news topics get published by NIST?

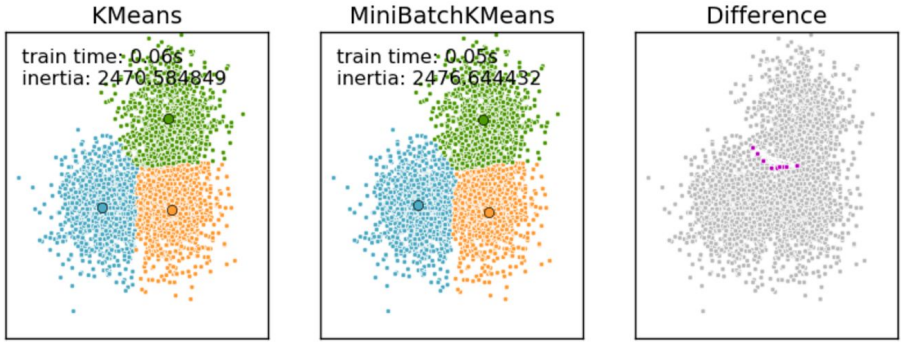
- K-Means Clustering on news published on NIST's newsfeed in 2014
- Their website doesn't indicate which subject area the article is about, so our data is unlabeled
- We know NIST publishes news on 15 subject areas, so we know k=15
- The goal is to find homogeneous clusters in your data, where we try to minimize the amount of variation within the cluster (Euclidean distance)
- Each iteration slightly improves the clustering



COMMERCE DATA SERVICE

Optimizing

- Change the number of clusters (k)
- Change the stop_words to include words common in your dataset (for example, National Institute of Standards and Technology)
- Mess with the init parameters sklearn.cluster.KMeans
- If you have a lot more data and the computing time is getting ridiculous, try MiniBatchKMeans.



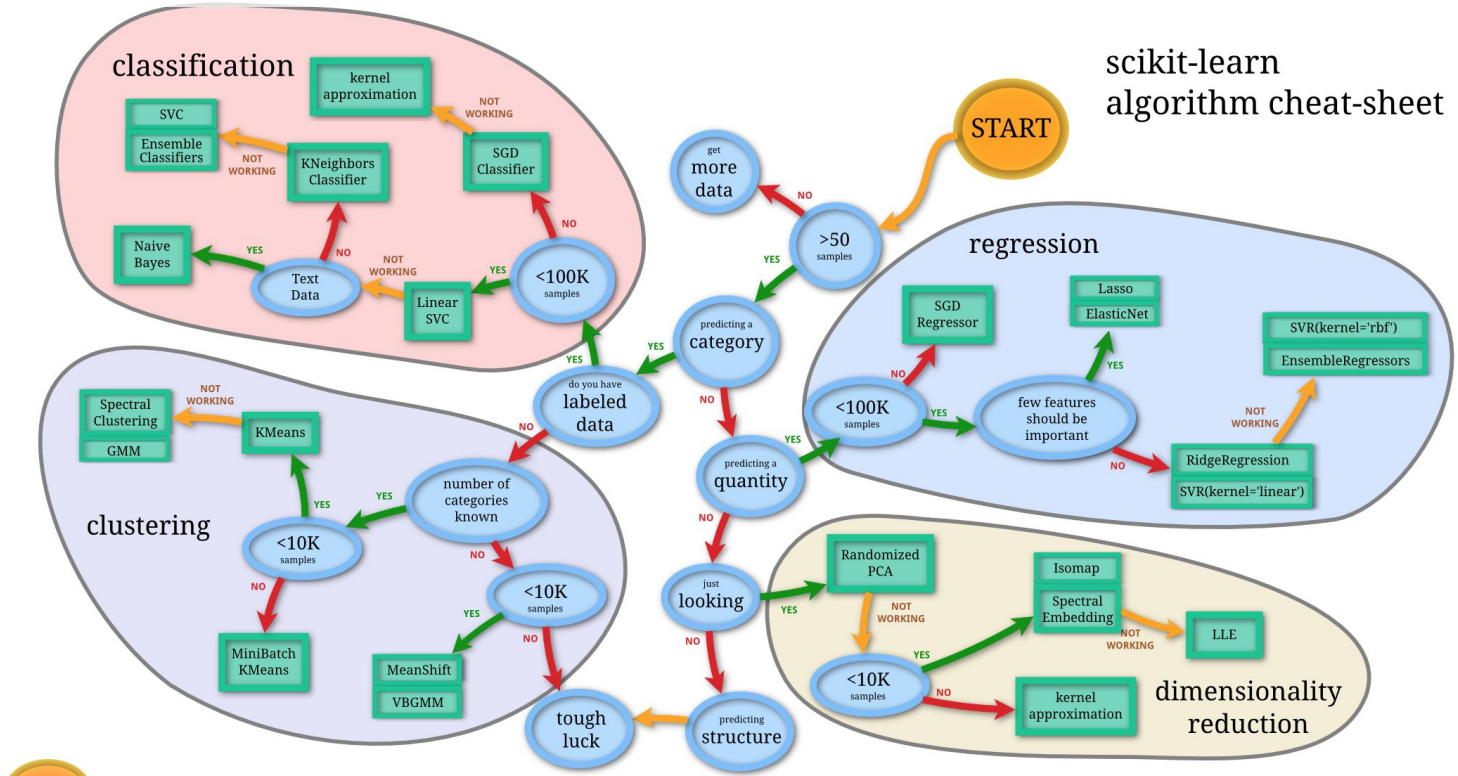
COMMERCE DATA SERVICE

Resources

- MIT OpenCourseware
  - Machine Learning
  - Statistics
  - Probability
  - Linear Algebra
  - Algorithms
  - Optimization
- An Introduction to Machine Learning with Python (Rebecca Bilbro)
- Machine Learning map for scikit learn and in general
- Binge watch machine learning
- Introduction to Statistical Learning in R

COMMERCE DATA SERVICE

scikit-learn algorithm cheat-sheet



Back

scikit learn

Question, Comments, Contact

Check out more of our open work at:

http://www.commerce.gov/datausability