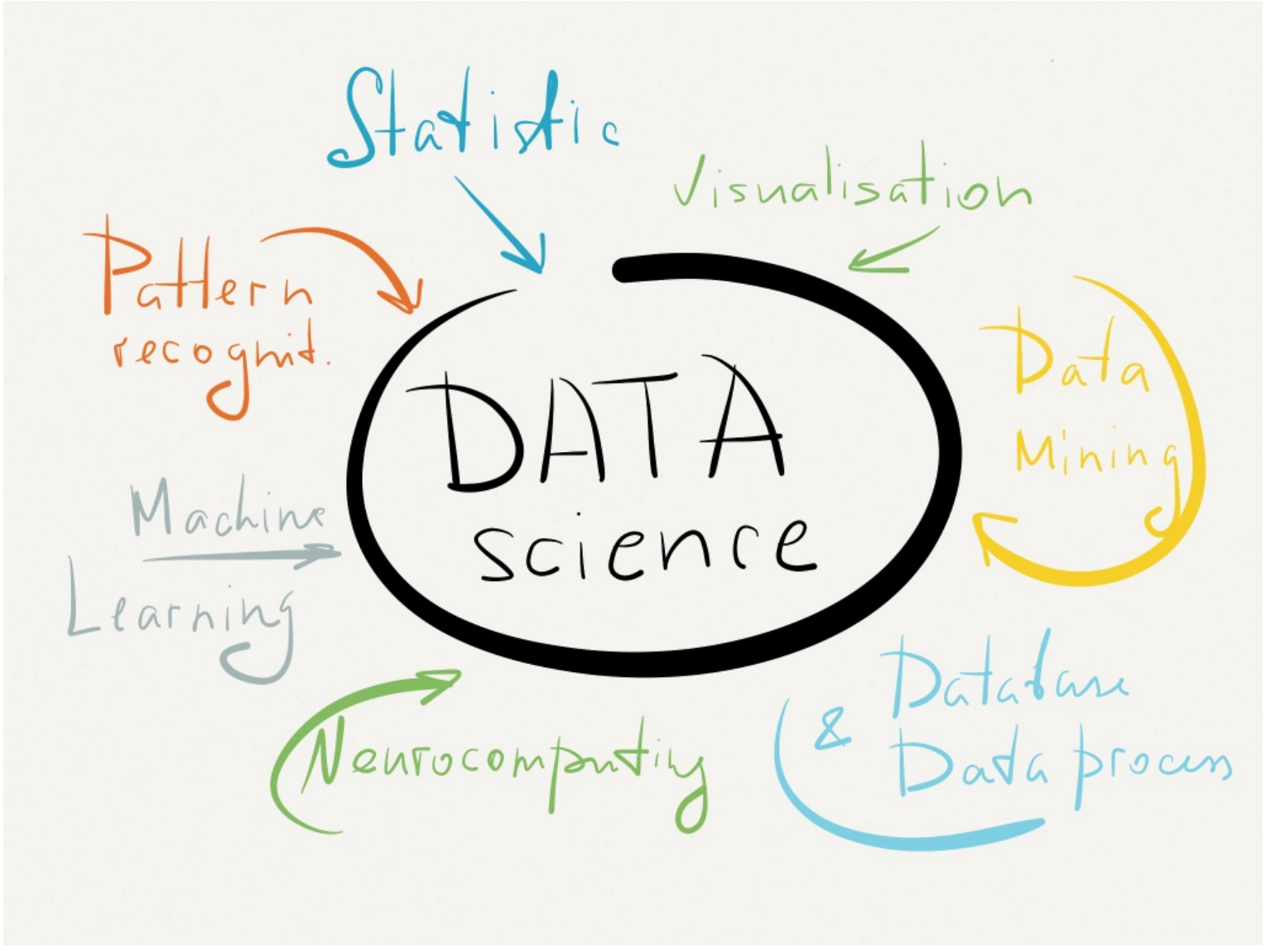
Introduction to Data Science using Python

Issam Laradji

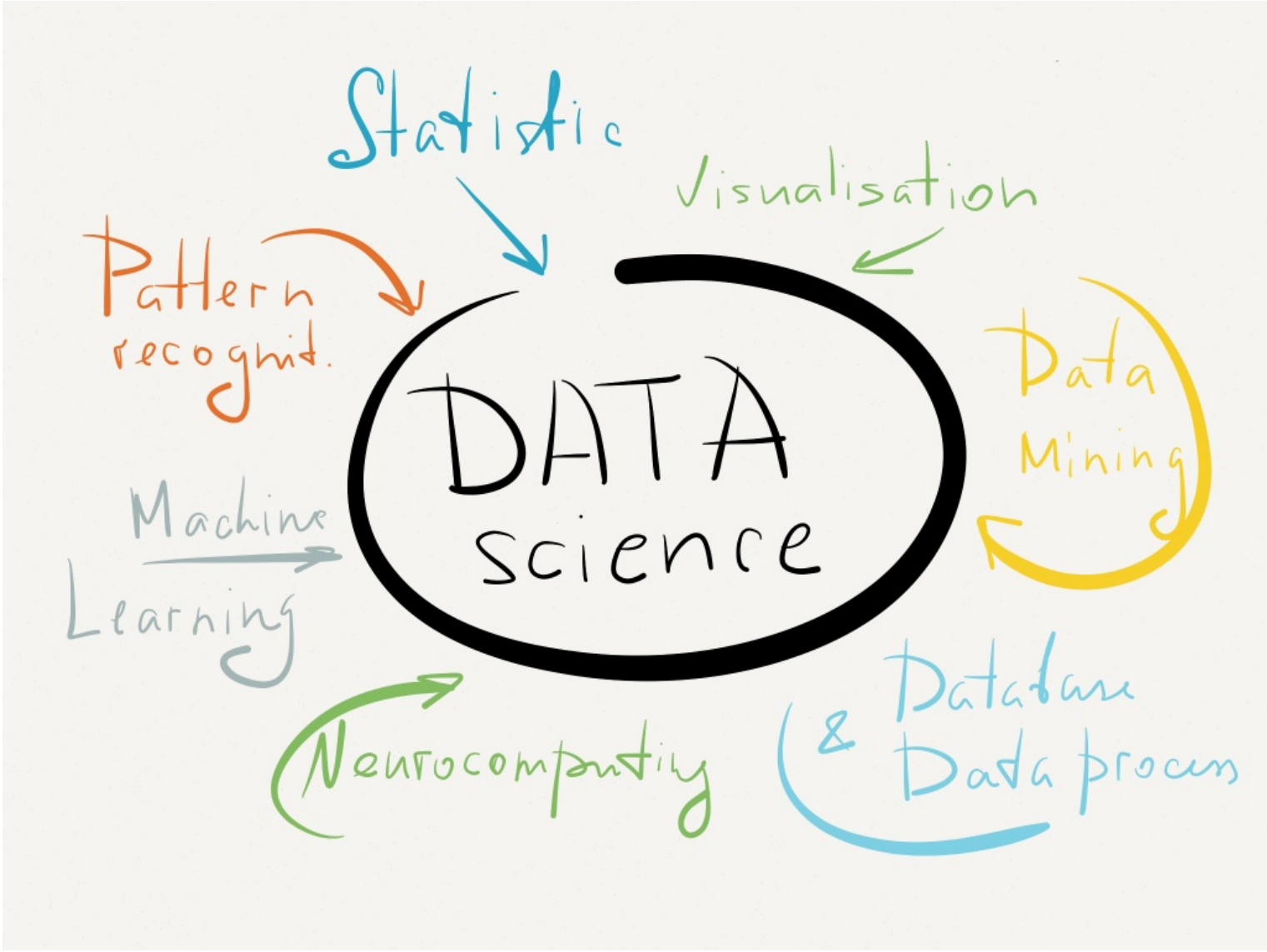
Data Science

- The term "data science" has exploded in business environments



Data Science

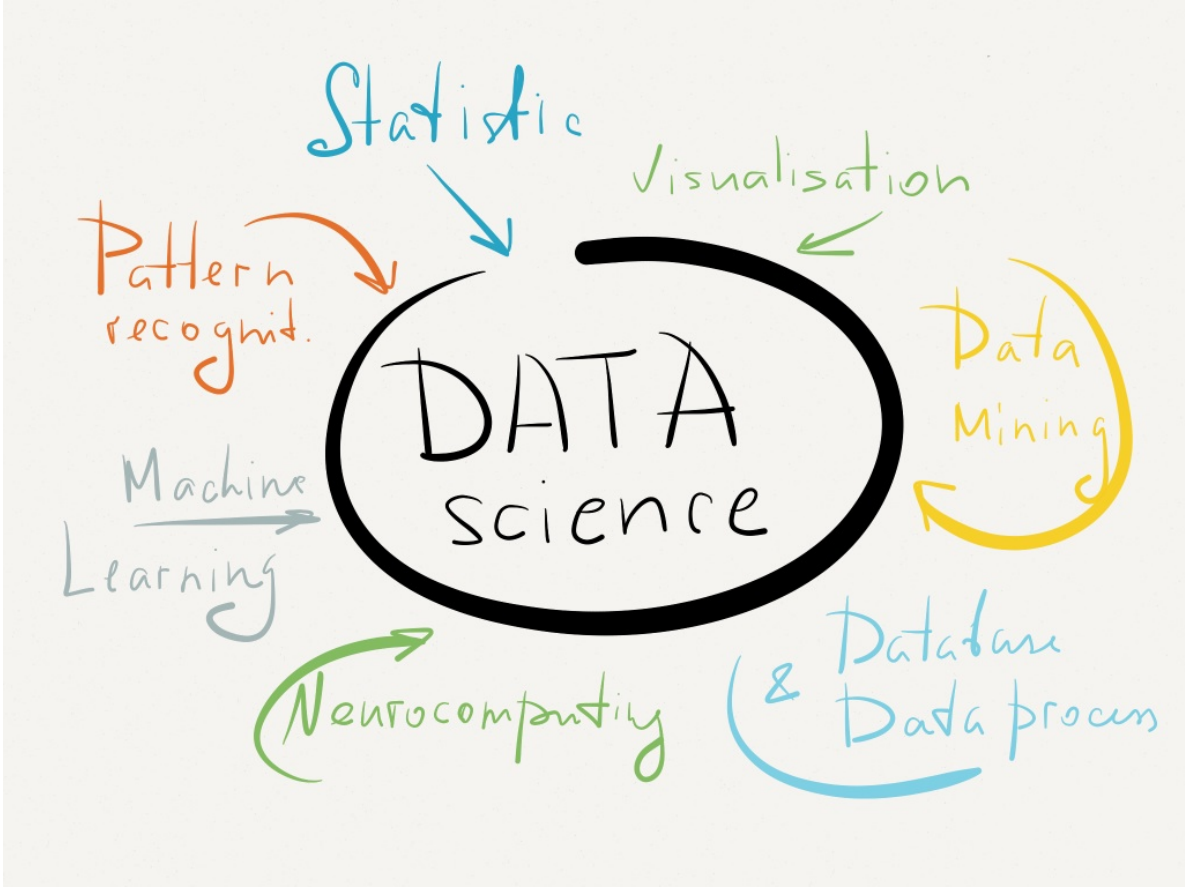
- Many academics and journalists see no distinction between data science and statistics



Data Science

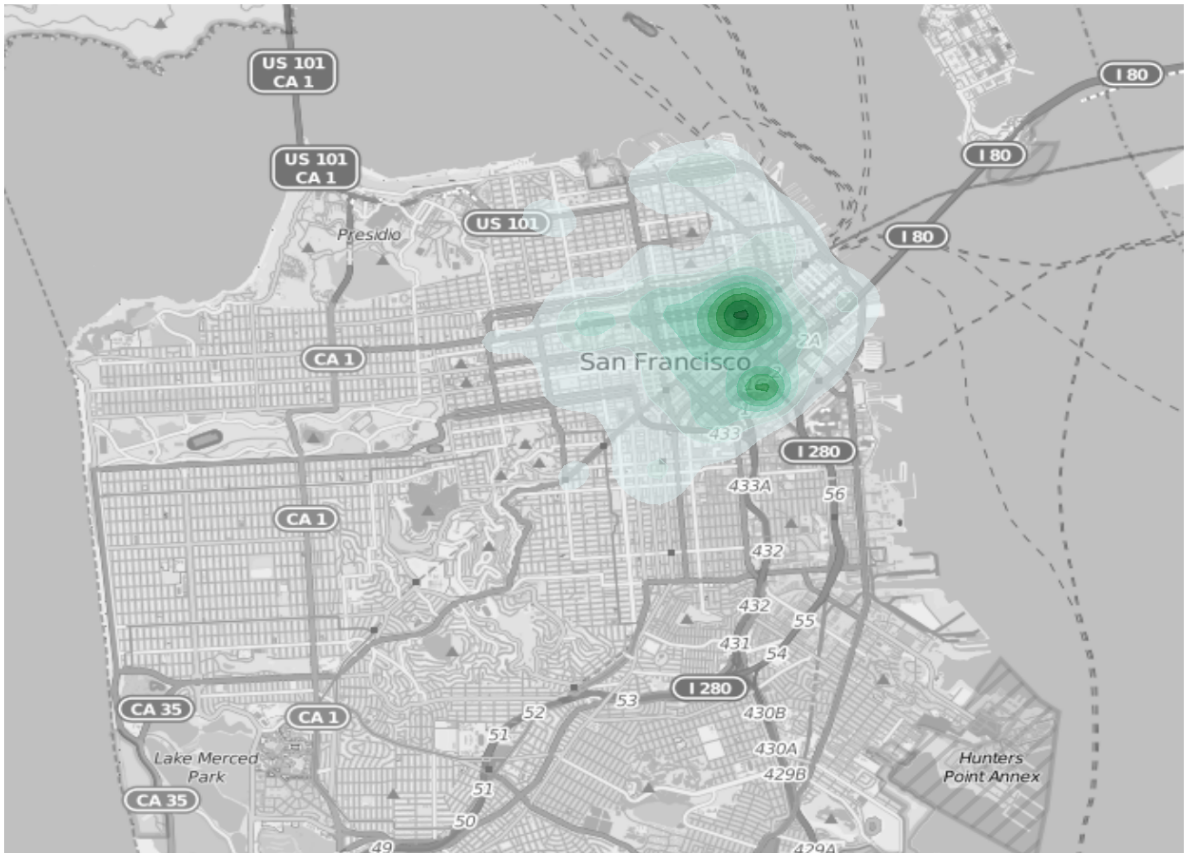
Nate Silver

- Sexed-up term for statistics. Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn't berate the term statistician



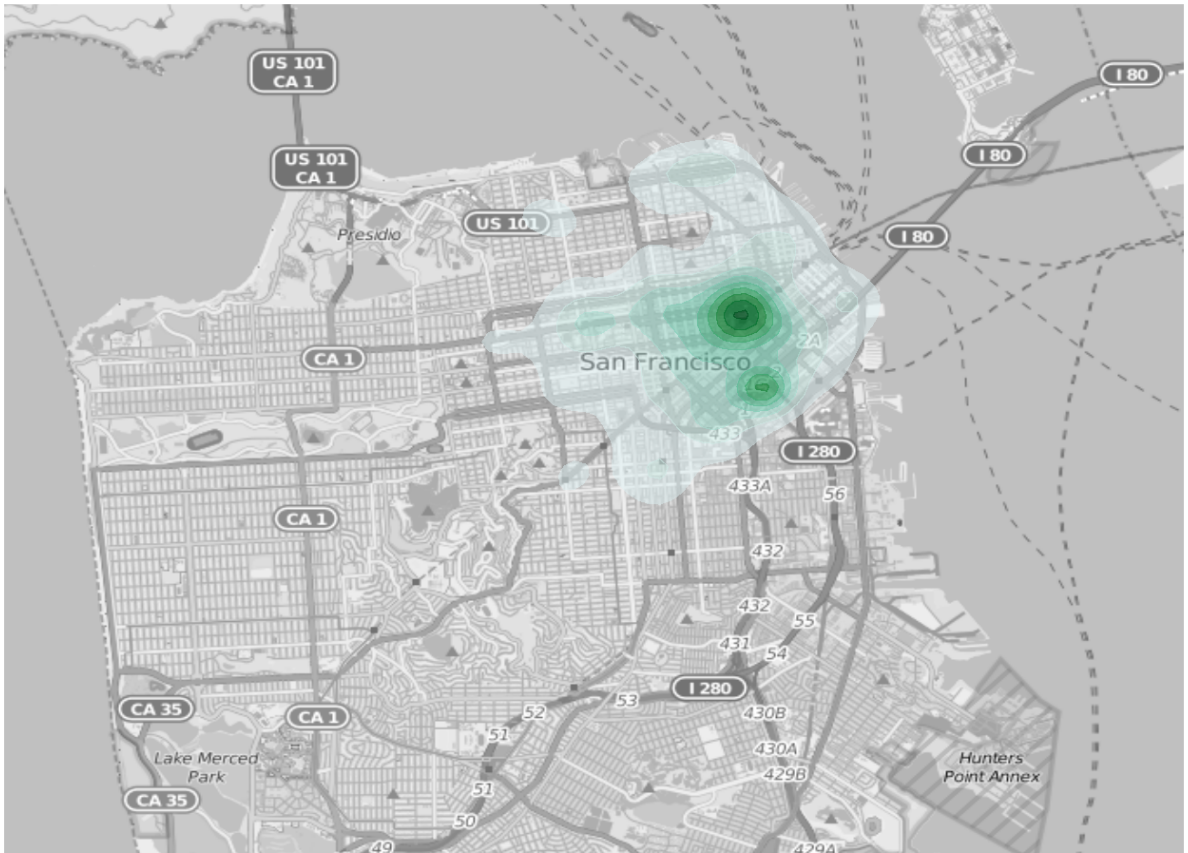
Data Science

- Find and interpret rich data sources
- Create visualizations to aid in understanding data
- Data Scientists are people who turn data into applications



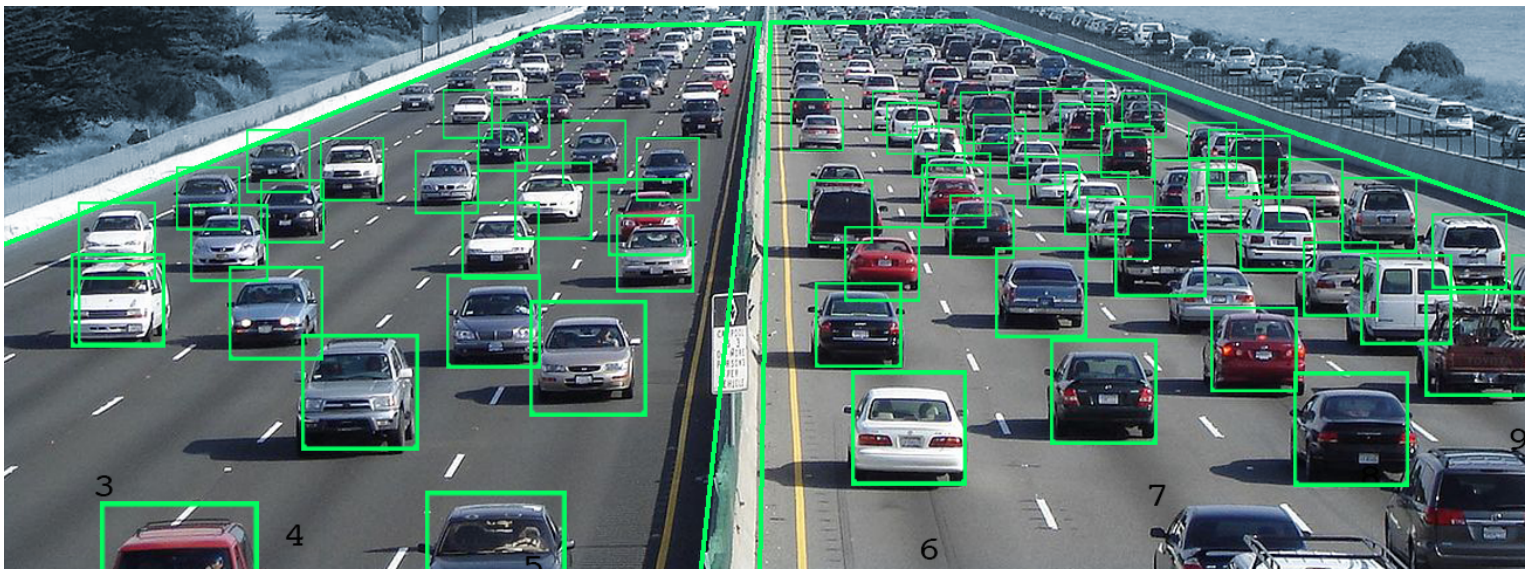
Data Science

- Find and interpret rich data sources
- Create visualizations to aid in understanding data
- Data Scientists are people who turn data into applications



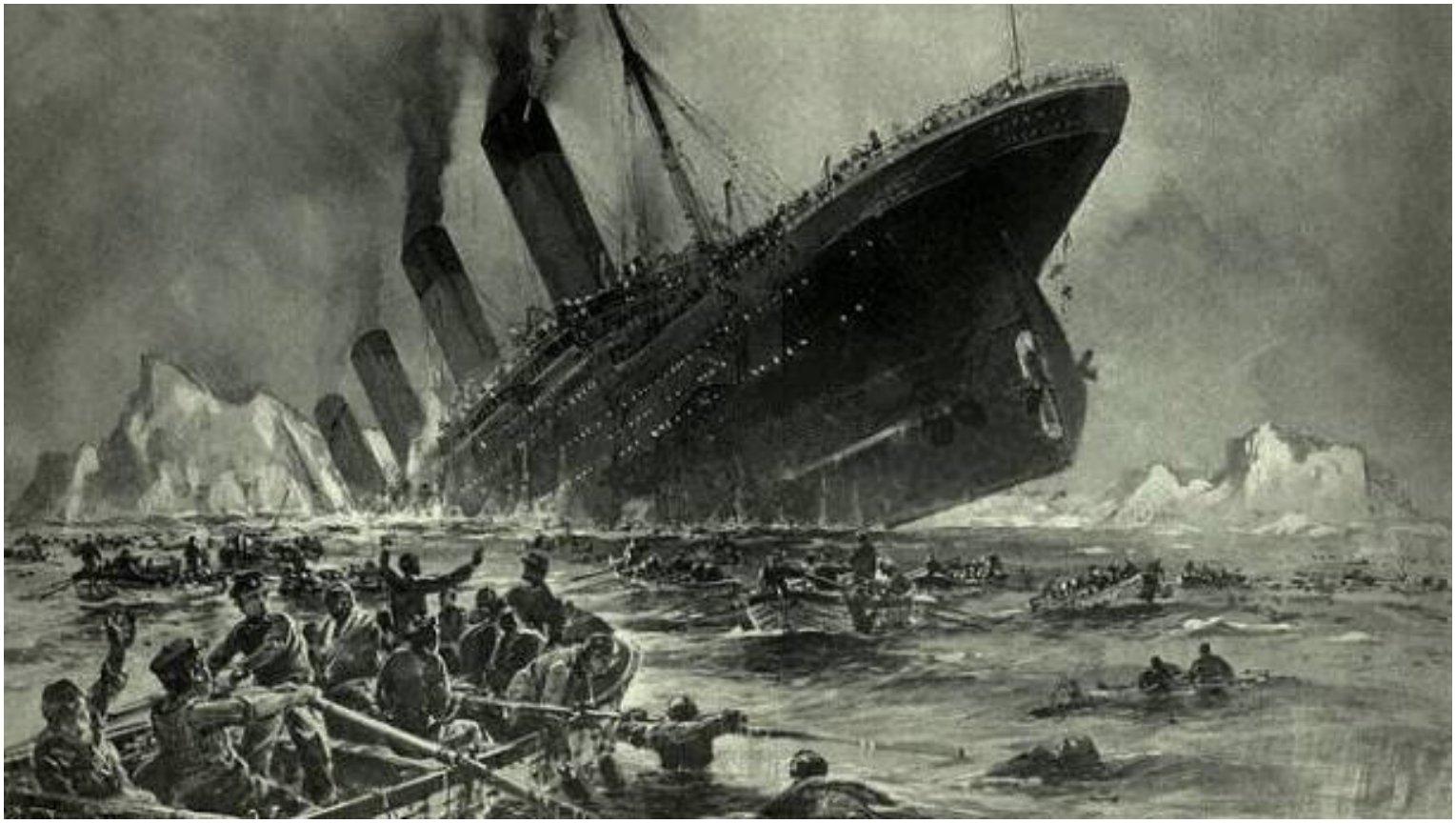
Data Science

- Find and interpret rich data sources
- Create visualizations to aid in understanding data
- Data Scientists are people who turn data into applications

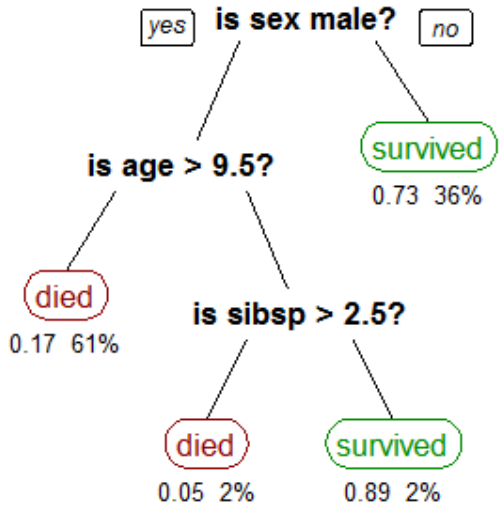


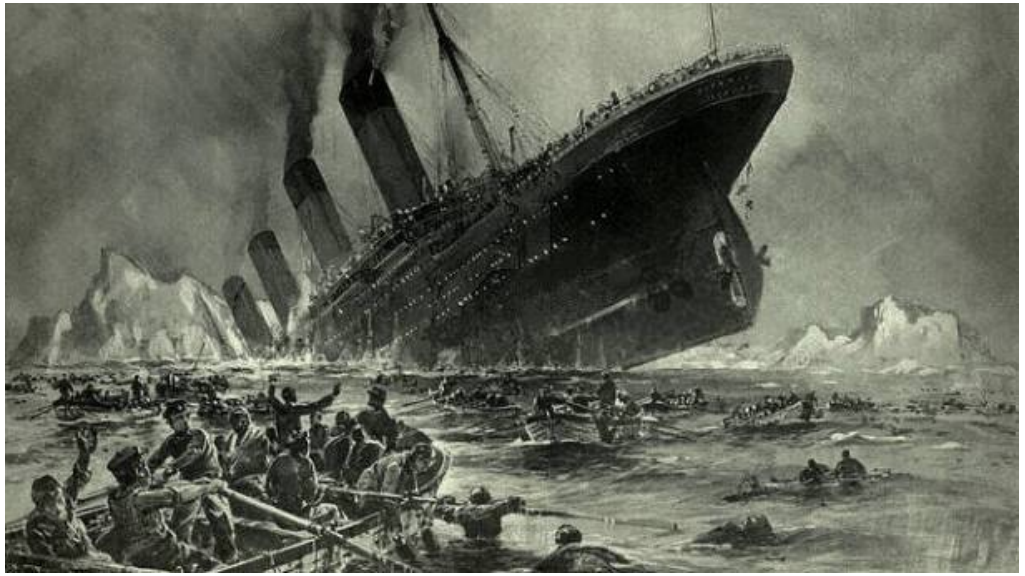
Data Science

- Find and interpret rich data sources
- Create visualizations to aid in understanding data
- Data Scientists are people who turn data into applications



Data Science



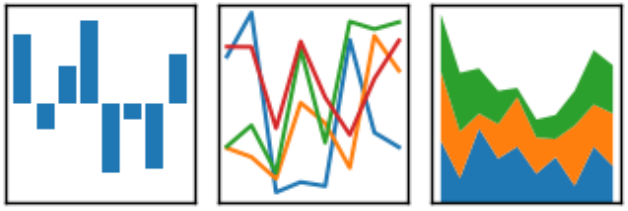


Python libraries for data science

- Pandas (for data manipulation and visualization)

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



- Scikit-learn (for machine learning)

scikits learn

machine learning in Python

Titanic Dataset

On April 15, 1912

- the Titanic sank after colliding with an iceberg
  - killing 1502 out of 2224 passengers and crew.
  - There were not enough lifeboats for the passengers and crew.
- Some groups of people were more likely to survive than others, such as women, children, and the upper-class.

Task

- What sorts of people were likely to survive ?
- Use Data Science or Machine Learning to predict which passengers survived the tragedy.

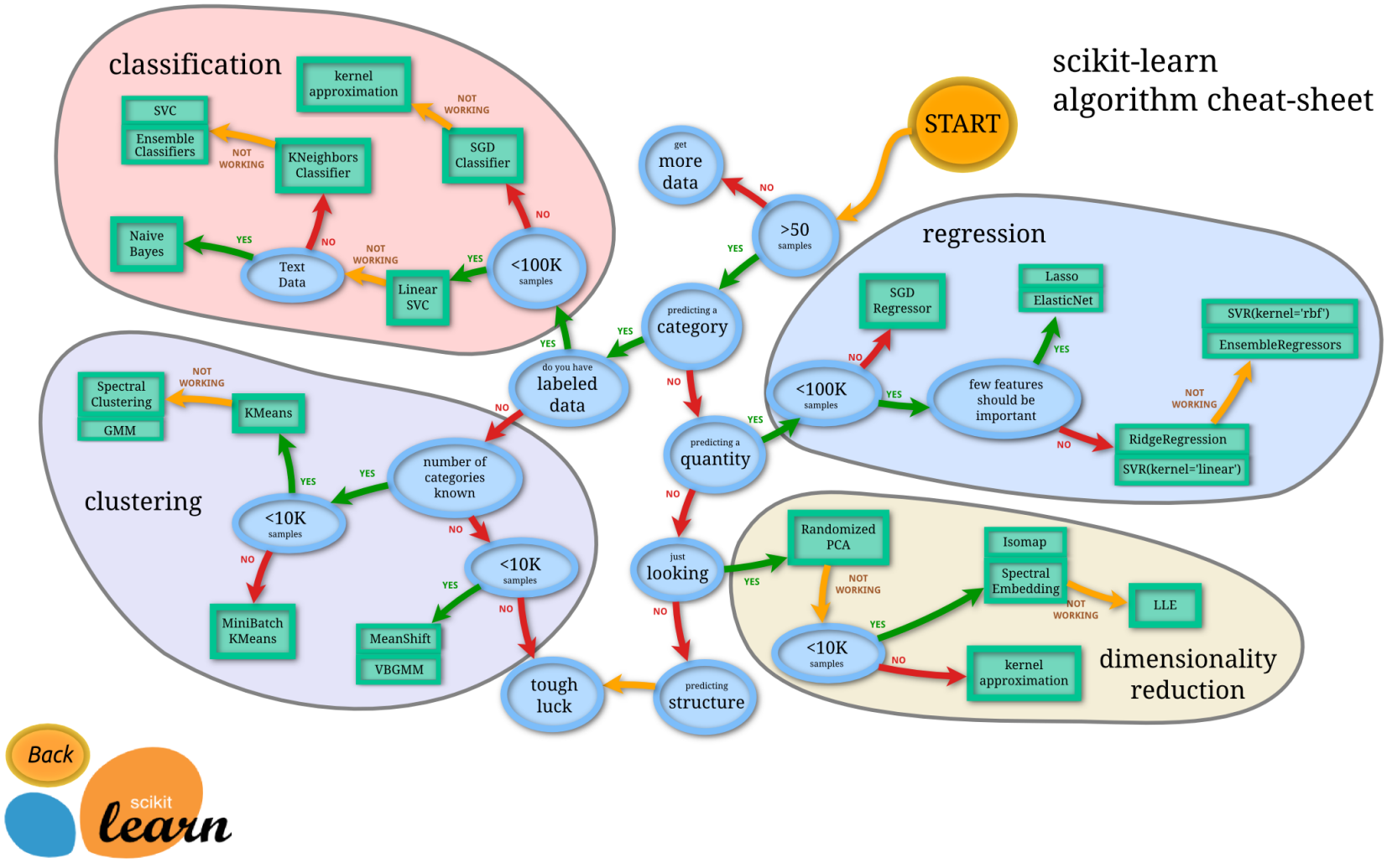
Titanic Dataset

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southamp

- Download link: goo.gl/oF5GBc

Sklearn


scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
kernel approximation
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K samples
few features should be important
RidgeRegression
SVR(kernel='linear')
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
<10K samples
MeanShift
VBGMM
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
<10K samples
kernel approximation
YES
NO
NOT WORKING
Back
scikit learn

Check out the exercises at the bottom of the jupyter notebook

Exercise 1:
Try different depths of the decision tree (see code in the notebook)

Exercise 2:
Use three sklearn methods and see which gives the highest score (see code in the notebook)