Real World Applications of Data Science

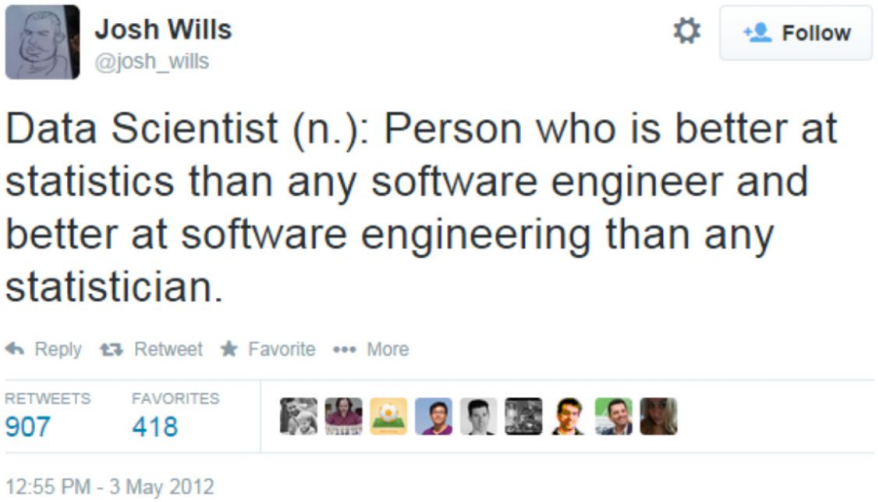
In partnership with:
Proscia Inc, Betamore, Spark B-more

Lecture 1 Notes: Intro DS + Python Basics + Intro ML

What is a data scientist?

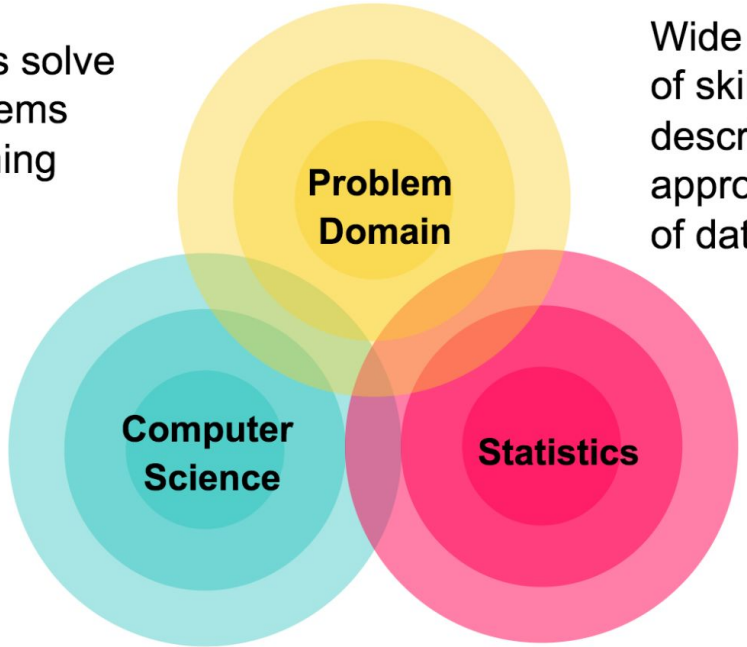
Data scientist definitions

Data scientist definitions



Data scientist roles

Data Scientists solve complex problems using data mining techniques



Wide variance in terms of skillsets: many job descriptions are more appropriate for a team of data scientists

The Value of Data Scientists

The Value of Data Scientists

Data scientists add values to companies by doing one of 3 things:

1) Predicting the good
2) Identifying the bad
3) Automating existing processes

Let’s take a look at some real world applications of data science..

Predicting Neonatal Infection

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 predicts infection before it occurs

Data: 16 vital signs such as heart rate, respiration rate, blood pressure, etc…

Impact: Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

Image: http://www.babycaretips4u.com/wp-content/uploads/2014/03/premature-baby.jpg
Case Study: http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695

Automating government tasks

Problem: Processing disability claims at the Social Security Administration is a time-intensive process, with many claims taking over 2 years to adjudicate

Goal: Automate the approval of a subset of the "simplest" disability claims

Data: Free text in the claims form

Impact: Able to fully automate 20% of the simplest claims. Rating accuracy of the algorithm is higher than the average claims examiner.

Predicting grade of cancer

Problem: Interpathologist conccurrence rates as low as 55% for prevalent diseases like Prostate Cancer Adenocarcinoma

Goal: Increase diagnostic accuracy

Data: Labeled, digitized biopsy images

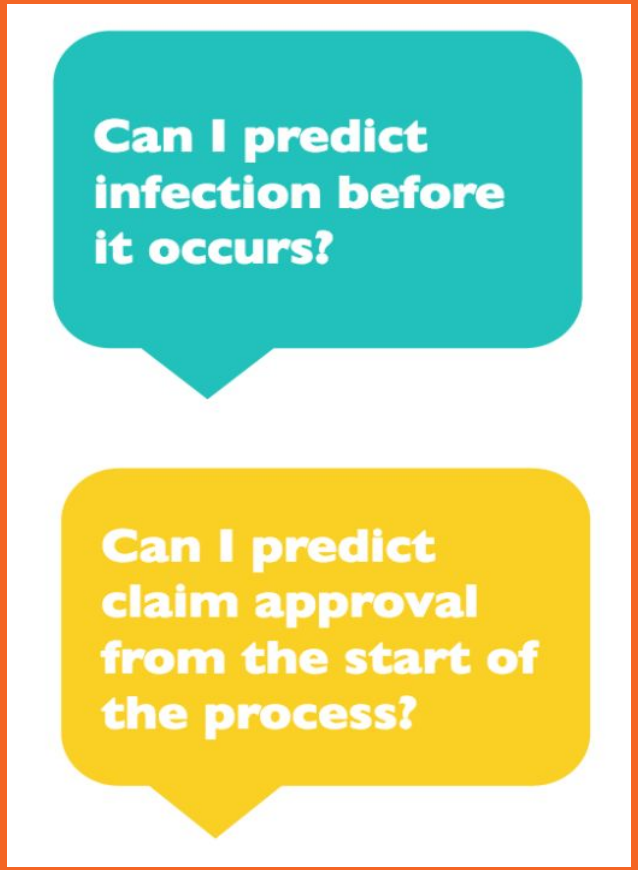
Impact: Much higher concurrence with prognostic consensus

Data Science Workflow

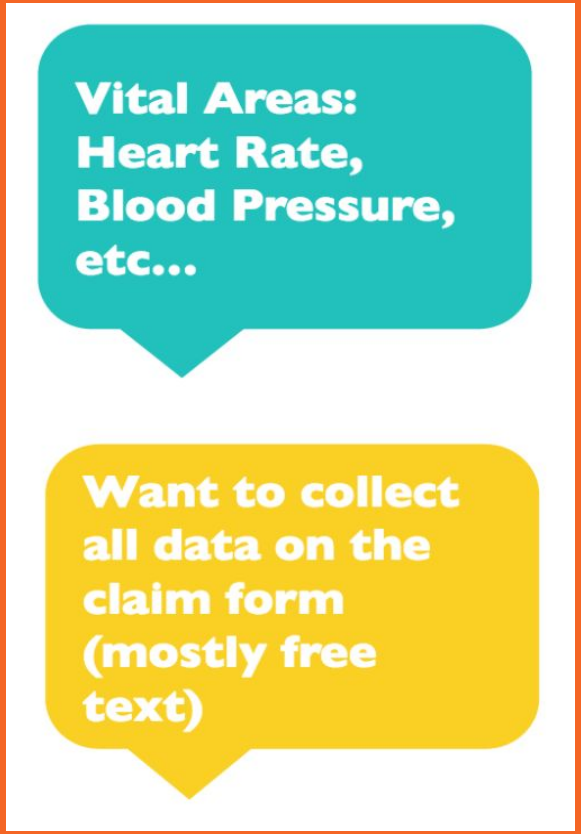
How to think like a data scientist

1) Define the problem/question
2) Identify and collect data
3) Explore and prepare data
4) Build and evaluate model
5) Communicate results

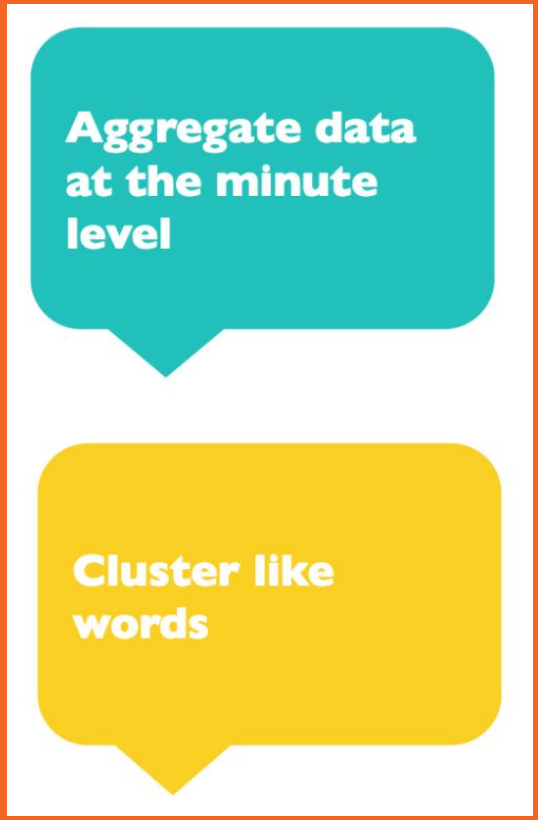
1) Define the Problem/Question



2) Identify and Collect Data

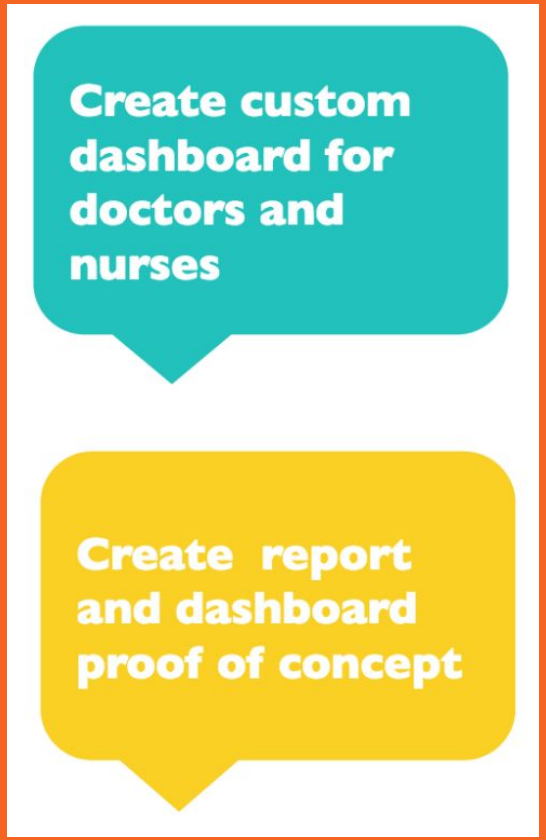


3) Explore and Prepare Data



4) Build and Evaluate Models

5) Communicate Results



Qualities of a Good Data Scientist

- Asks Rational Questions
- Understands Pros/Cons of different techniques
- Communicates Clearly
- Retains Intellectual Humility

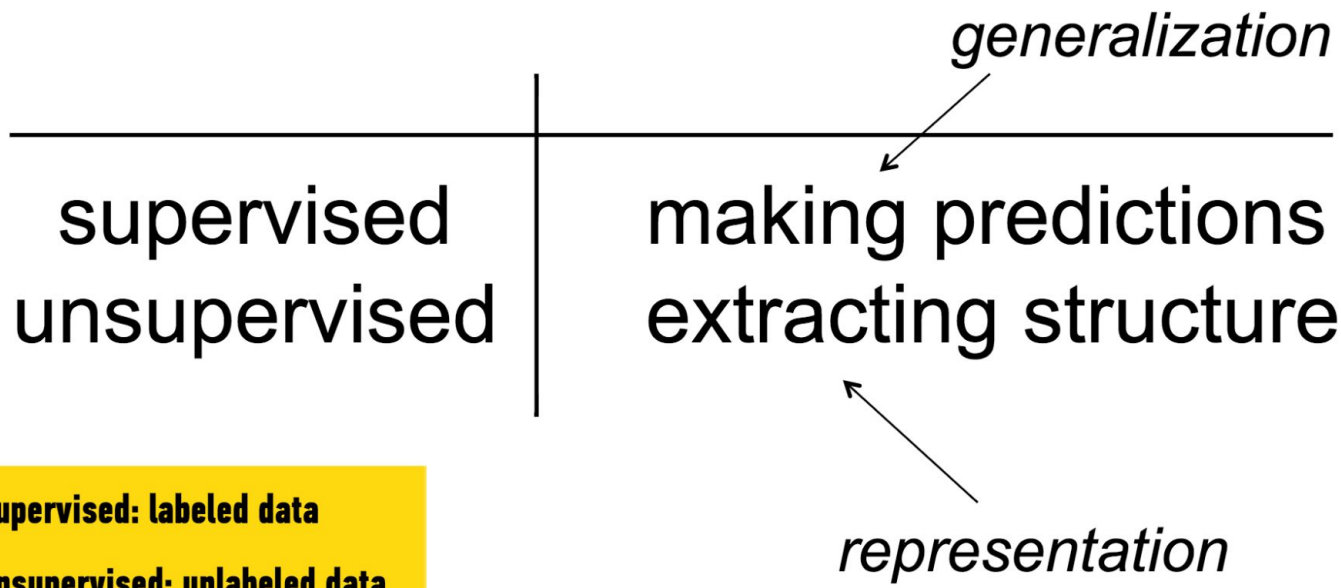
What the hell is machine learning?

What is Machine Learning?

“ A field of study that gives computers the ability to learn without being explicitly programmed” (1959)

- Machine learning is a class of algorithms that are data-driven. Unlike classical algorithms, it’s the data that defines a “good” answer.
- The core of machine learning deals with: Representation, and generalization

Types of ML problems



Supervised: labeled data

Unsupervised: unlabeled data

Supervised Learning

- “Vector” list of Predictors X
  - Features, independent variables, inputs, regressors, covariates, attributes
- Response Y
  - Outcome, dependent variable, label, target
- If Y is continuous: **Regression**
  - Price, blood pressure..
- If Y is categorical (values in finite, unordered set): **Classification**
  - Digits 0-9, cancer grades of tissue
- Data is composed of observations (predictors and associated response)
  - Samples, examples

Predicting Neonatal Infection

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 predicts infection before it occurs

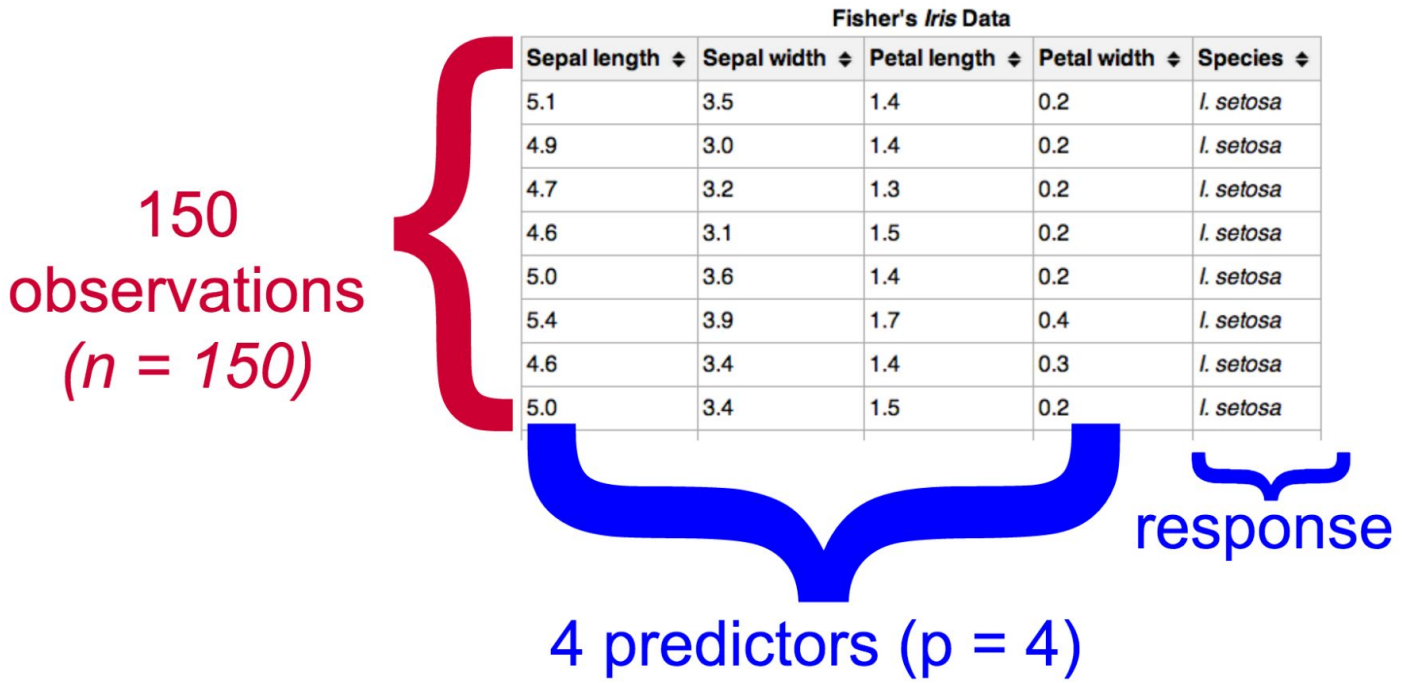
Data: 16 vital signs such as heart rate, respiration rate, blood pressure, etc…

Impact: Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

predictors

Sample response: Did the child develop an infection? True/False

Iris Data Set Intro



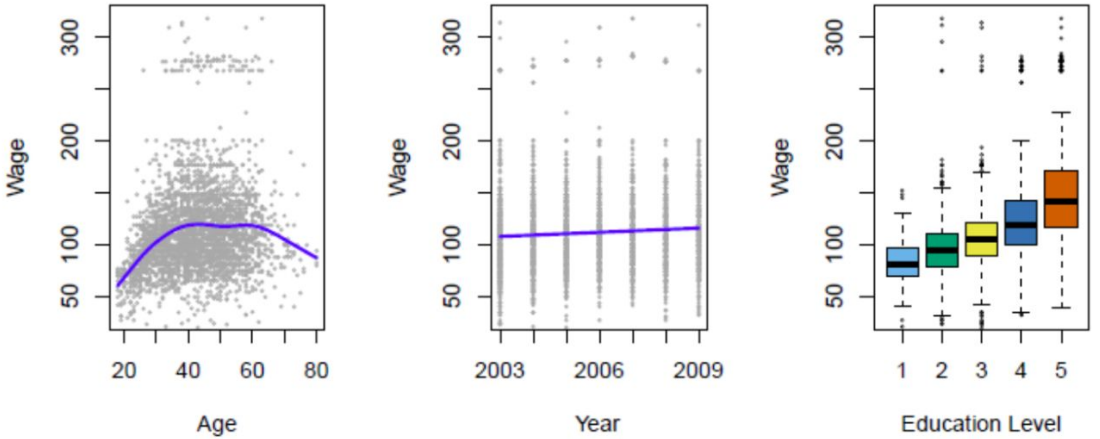
Supervised Learning

Supervised learning uses known **labeled/training** cases to:

- Accurately predict unseen test cases
- Understand which predictors affect response, and how
- Assess the quality of our predictions

Regression Example

Establish the relationship between salary and demographic variables in population survey data

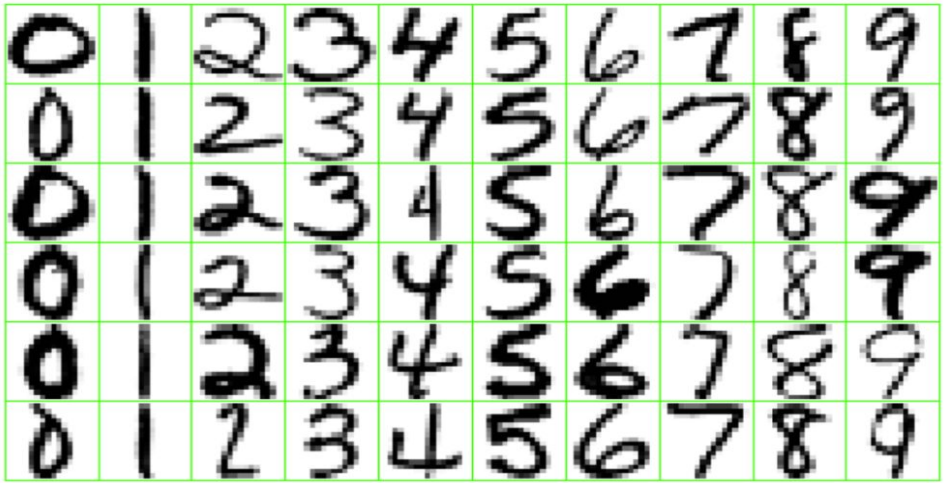


Income survey data for males from the central Atlantic region of the USA in 2009

Source: https://class.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf

Classification Example

Identify the numbers in a handwritten zip code



Source: https://class.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf

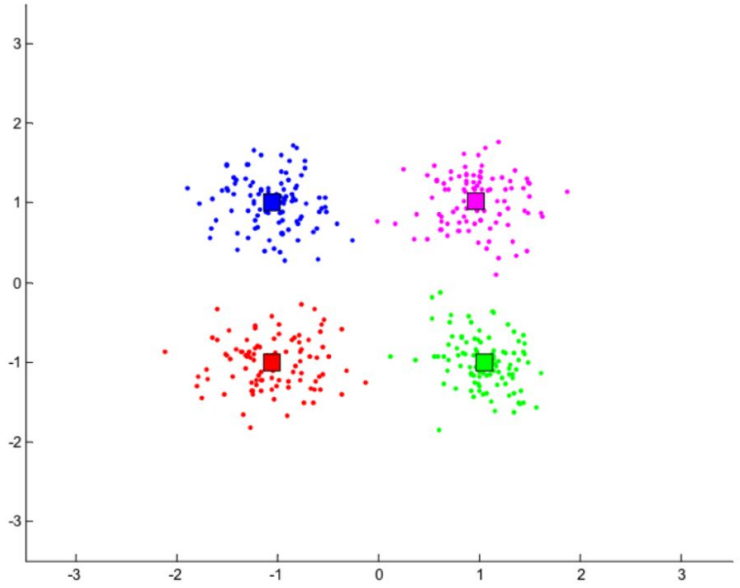
Unsupervised Learning

- No response variable Y; Just predictor X
- Objective is more open:
  - Find groups of observations that behave similarly
  - Find predictors that behave similarly
  - Find combinations of features that explain behavior of data
- Sometimes useful as preprocessing step for supervised
  - Clustering, Principal Component Analysis

Supervised v. Unsupervised

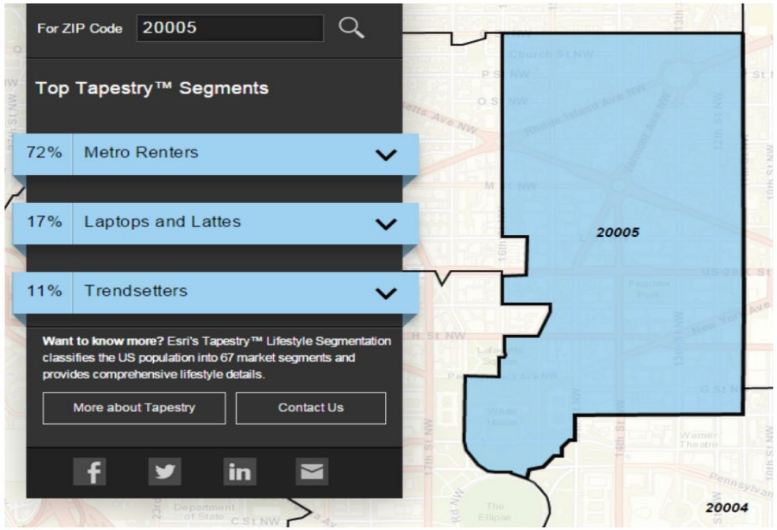
	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

Clustering example



Clustering example

Classify US residential neighborhoods into 67 unique segments based off of demographic and socioeconomic information



Example of cluster: **Metro Renters:**

- Young, mobile, educated, or still in school
- Live alone or with a roommate
- Works long hours
- Buys groceries at Whole Foods and Trader Joe's
- Shops at Banana Republic, Nordstrom, and Gap
- Loves yoga, go skiing, and attend Pilates sessions.

Source: http://www.esri.com/landing-pages/tapestry/

Supervised v. Unsupervised

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

Classification Example

Q: How does a classification problem work?
A: Data in, predicted labels out.

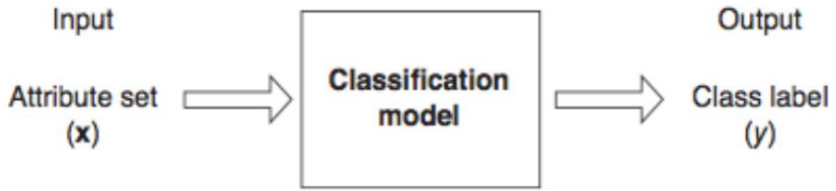


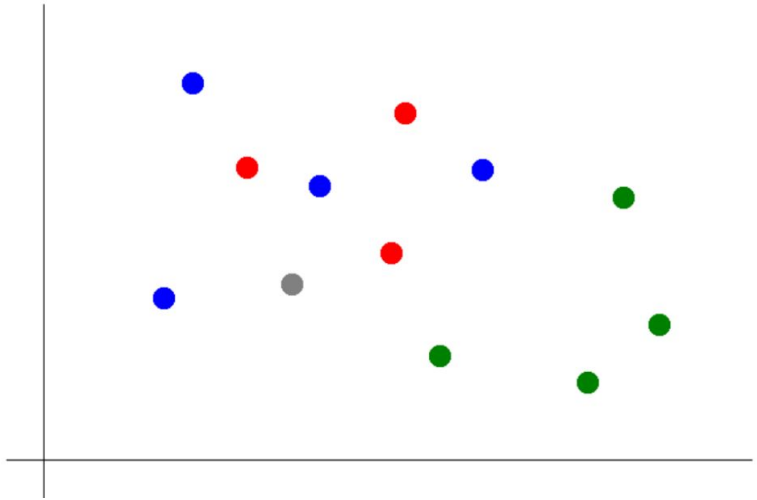
Figure 4.2. Classification as the task of mapping an input attribute set $\mathbf{x}$ into its class label y.

Classification via KNN

Suppose we want to predict the color of the gray dot.

QUESTION:

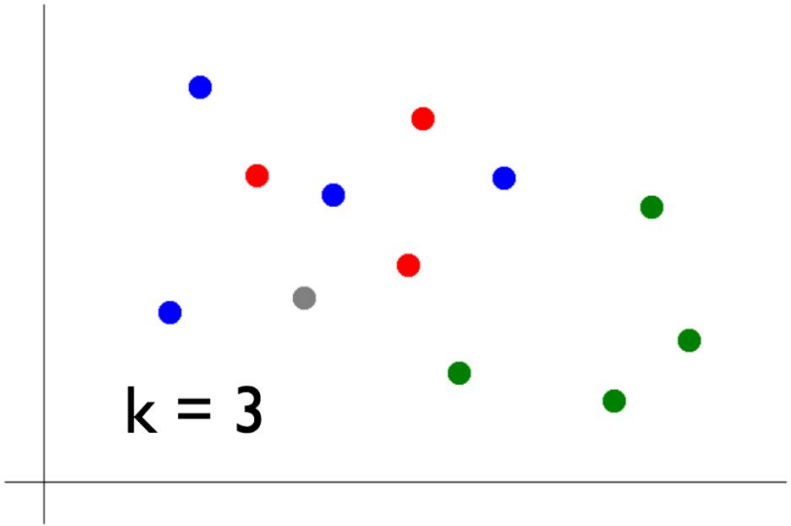
What are the predictors?
What is the response?



Classification via KNN

Suppose we want to predict the color of the gray dot.

1) Pick a value for k.

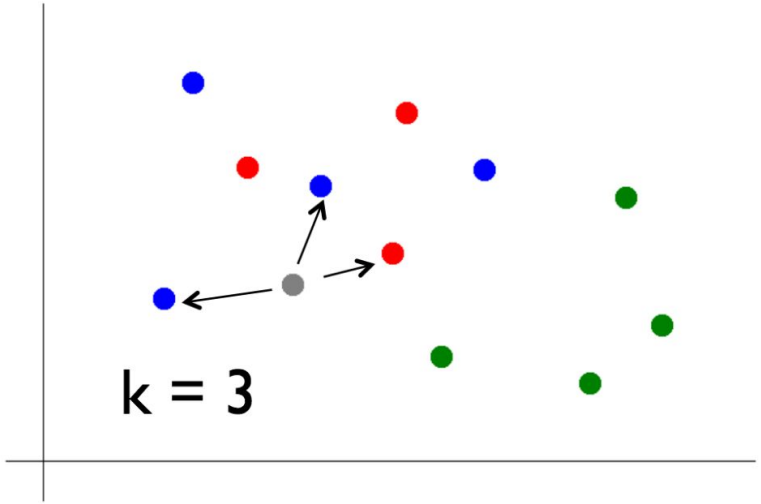


Classification via KNN

Suppose we want to predict the color of the gray dot.

1) Pick a value for k.

2) Find colors of k nearest neighbors.



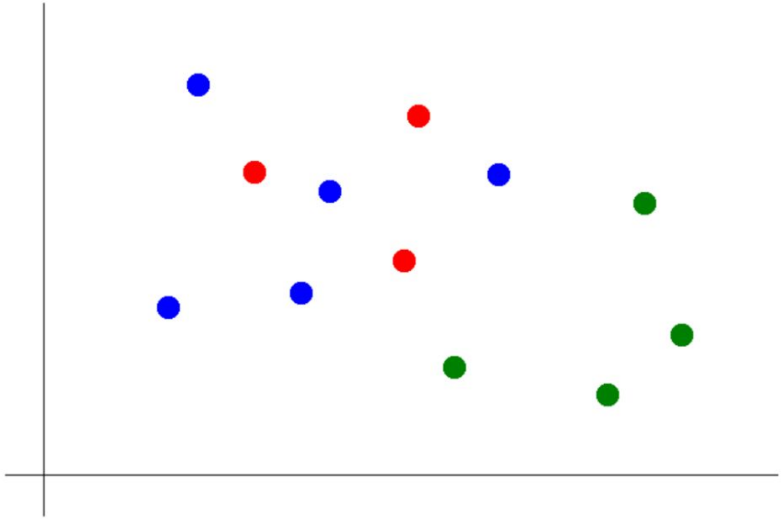
Classification via KNN

Suppose we want to predict the color of the gray dot.

1) Pick a value for k.

2) Find colors of k nearest neighbors.

3) Assign the most common color to the gray dot.



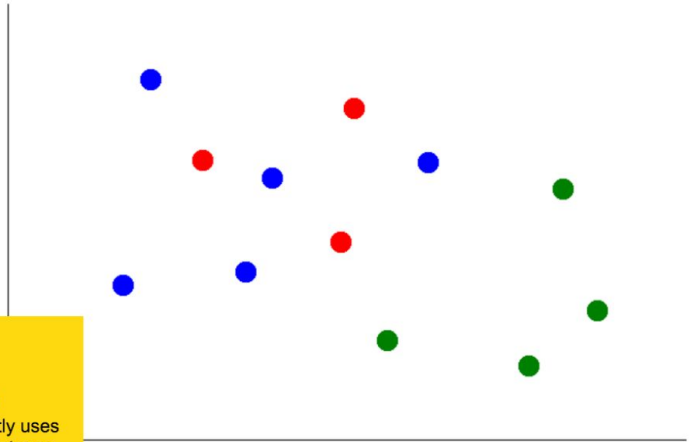
Classification via KNN

Suppose we want to predict the color of the gray dot.

1) Pick a value for k.

2) Find colors of k nearest neighbors.

3) Assign the most common color to the gray dot.



NOTE:

Our definition of "nearest" implicitly uses the *Euclidean distance function*.

K-Nearest Neighbors specs

Advantages

- Simple to understand and explain
- Model training phase is fast (low complexity)
- Non-parametric (no assumed decision boundary)

Disadvantages

- Prediction phase slow when n is very large
- Sensitive to irrelevant features

Advanced algorithms available for study

Any questions?