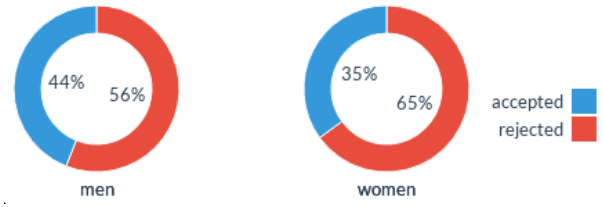
# Fundamental Methods of Data Science

## Class 2

Tomer Libal

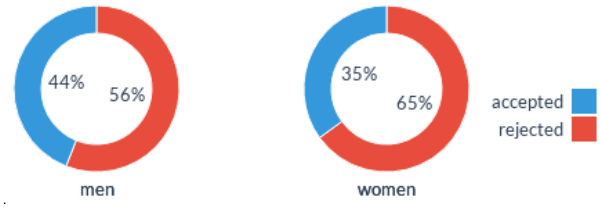
# Simpson's paradox

- What the suers have seen



# Simpson's paradox

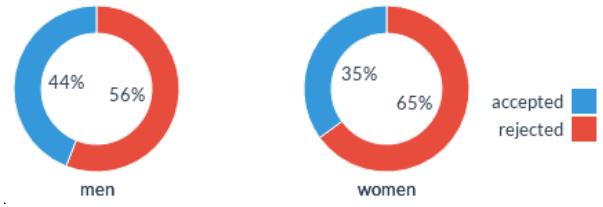
- What the suers have seen



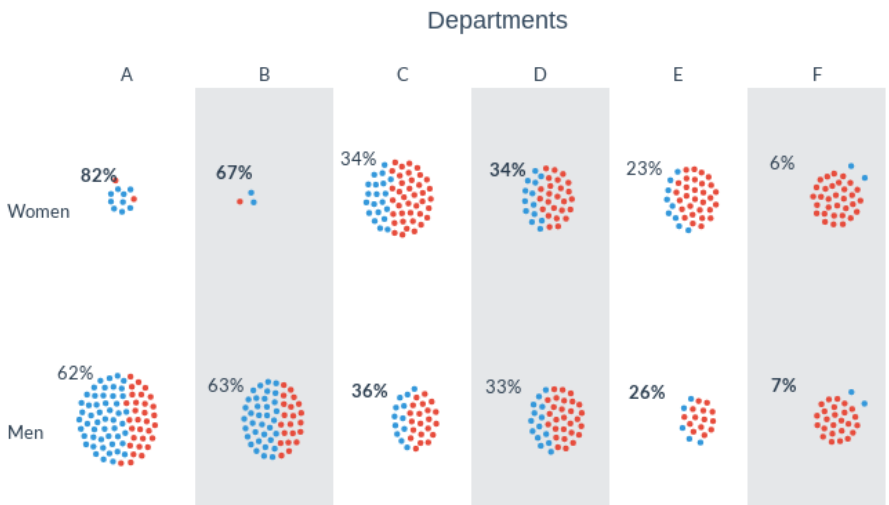

- What they have missed out

# Simpson's paradox

- What the suers have seen




- What they have missed out




- Data is not always as it seems
  - http://vudlab.com/simpsons/

# MegaTelCo: Predicting Customer Churn

- You just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms in the US

# MegaTelCo: Predicting Customer Churn

- You just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms in the US
- They are having a major problem with customer retention in their wireless business
  - In the mid-Atlantic region, 20% of cell phone customers leave when their contracts expire. Communications companies are now engaged in battles to attract each other's customers while retaining their own

# MegaTelCo: Predicting Customer Churn

- You just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms in the US
- They are having a major problem with customer retention in their wireless business
  - In the mid-Atlantic region, 20% of cell phone customers leave when their contracts expire. Communications companies are now engaged in battles to attract each other's customers while retaining their own
- Marketing has already designed a special retention offer
  - Your task is to devise a precise, step-by-step plan for how the data science team should use MegaTelCo's vast data resources to solve the problem

# MegaTelCo: Predicting Customer Churn

- What data you might use?
- How would they be used?
- How should MegaTelCo choose a set of customers to receive their offer in order to best reduce churn for a particular incentive budget?

# MegaTelCo: Predicting Customer Churn

- What data you might use?
- How would they be used?
- How should MegaTelCo choose a set of customers to receive their offer in order to best reduce churn for a particular incentive budget?
- We want to use the data in order to tell which customers are more likely to churn

# MegaTelCo: Predicting Customer Churn

- What data you might use?
- How would they be used?
- How should MegaTelCo choose a set of customers to receive their offer in order to best reduce churn for a particular incentive budget?
- We want to use the data in order to tell which customers are more likely to churn
- But they did not yet churn, what can we do?

# Models

- We can try to predict who might churn, but how?

# Models

- We can try to predict who might churn, but how?
- **Models** are snapshots of the world
- Base on what is known **in the model**, we can try to predict who might churn

# Models

- Assume a model tells us that
  - The three angles of a triangle sum up to 180 degrees
- And we know that
  - 'A' is a triangle
  - 2 angles of 'A' sum up to 100 degrees
  - Can you tell the size of the third angle of 'A'?

# Models

- Assume a model tells us that
  - The three angles of a triangle sum up to 180 degrees
- And we know that
  - 'A' is a triangle
  - 2 angles of 'A' sum up to 100 degrees
  - Can you tell the size of the third angle of 'A'?
- Now assume we have the same model as before except
  - The three angles of a triangle sum up to 200 degrees
  - Can you tell the size of the third angle of 'A'?

# Models

- Assume a model tells us that
  - The three angles of a triangle sum up to 180 degrees
- And we know that
  - 'A' is a triangle
  - 2 angles of 'A' sum up to 100 degrees
  - Can you tell the size of the third angle of 'A'?
- Now assume we have the same model as before except
  - The three angles of a triangle sum up to 200 degrees
  - Can you tell the size of the third angle of 'A'?
- What can you tell about models?

# Model

- **Model** is a simplified version of the world
- **Predictive models** allow us to compute a missing value in the model
- **Instance / example** represents a fact and is described by a set of attributes

# Model

- **Model** is a simplified version of the world
- **Predictive models** allow us to compute a missing value in the model
- **Instance / example** represents a fact and is described by a set of attributes
- What is the model in our triangle example? What is the instance?

# Model induction

- Data science is about creating or **inducting** models from data
- The data used to induct models are called **training data**

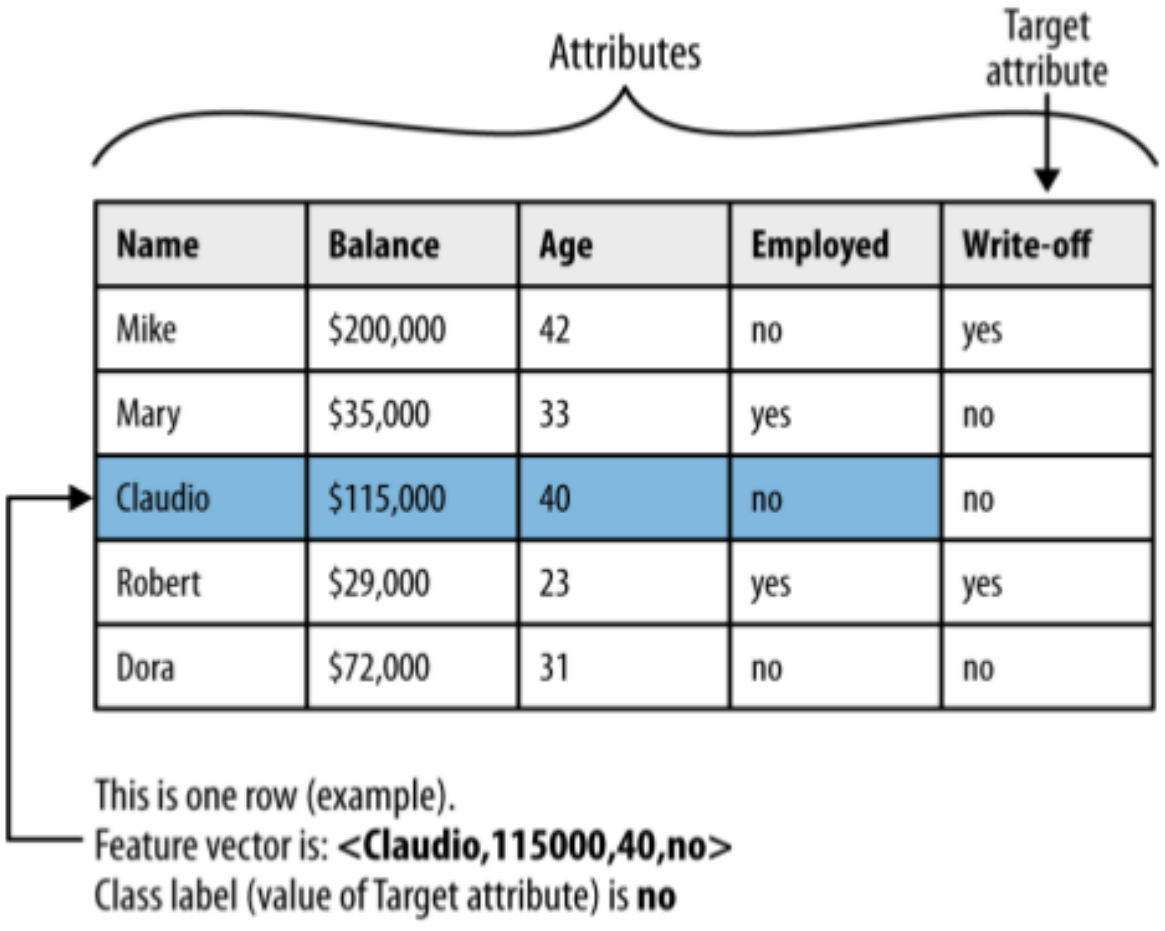
# Model induction

- Data science is about creating or **inducting** models from data
- The data used to induct models are called **training data**
- We induct models for a specific purpose
  - To tell us some attribute of an instance based on other attributes of the instance
  - **Feature vector** contains the attributes we know
  - **Target attribute** is the attribute we want to find

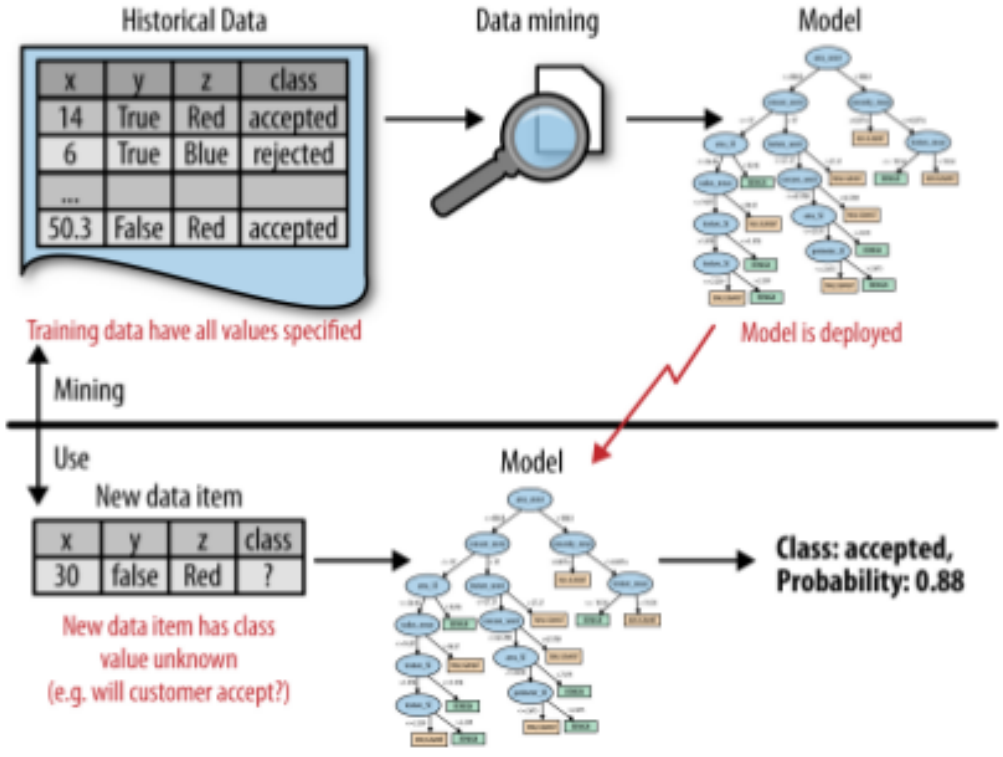
# Model induction

- Data science is about creating or **inducting** models from data
- The data used to induct models are called **training data**
- We induct models for a specific purpose
  - To tell us some attribute of an instance based on other attributes of the instance
  - **Feature vector** contains the attributes we know
  - **Target attribute** is the attribute we want to find
- Training data contain both the feature vector and the target attribute

# Data

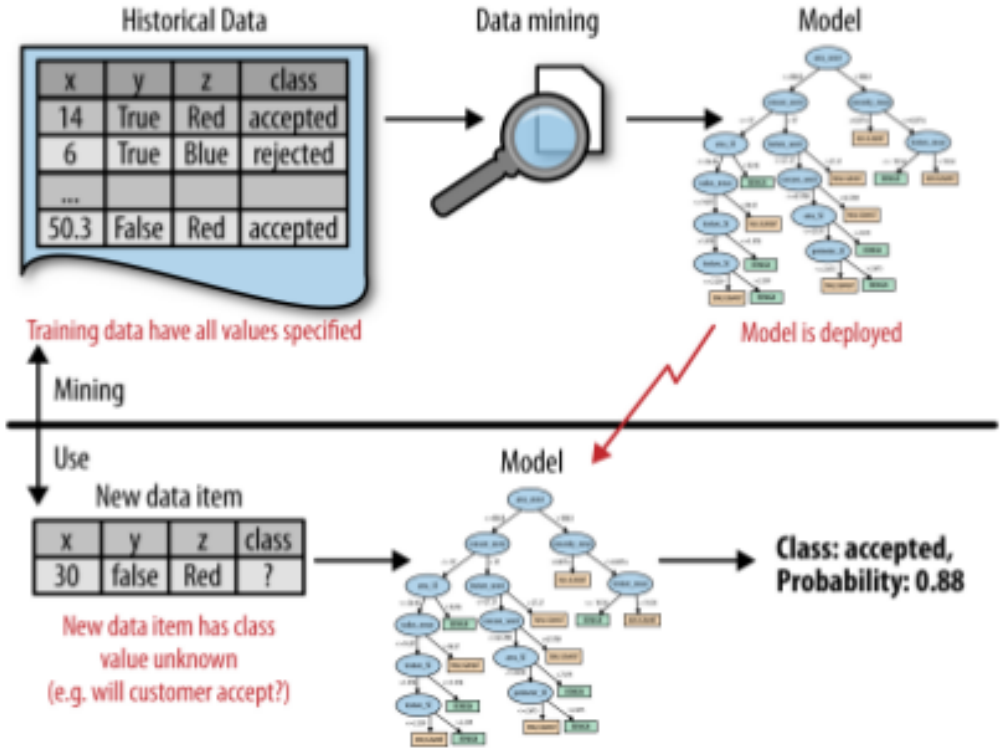

Attributes
Target attribute
Name
Balance
Age
Employed
Write-off
Mike
$200,000
42
no
yes
Mary
$35,000
33
yes
no
Claudio
$115,000
40
no
no
Robert
$29,000
23
yes
yes
Dora
$72,000
31
no
no
This is one row (example).
Feature vector is: <Claudio,115000,40,no>
Class label (value of Target attribute) is no

# Data mining




Provost/Fawcett - Data Science for Business

# Data mining




Provost/Fawcett - Data Science for Business

- We induct a **classification** model base on trees

# Feature Types

- Numeric: anything that has some order
  - Numbers
  - Dates
- Categorical: stuff that does not have an order
  - Binary
  - Text

# Feature Types

- Numeric: anything that has some order
  - Numbers
  - Dates
- Categorical: stuff that does not have an order
  - Binary
  - Text
- Are names numeric or categorical? And ratings?

# Common Data Mining Tasks

- Classification and class probability estimation
  - How likely is this consumer to respond to our campaign?
- Regression
  - How much will she use the service?
- Similarity Matching
  - Can we find consumers similar to my best customers?
- Clustering
  - Do my customers form natural groups?
- Co-occurrence Grouping
  - Also known as frequent itemset mining, association rule discovery, and market-basket analysis
  - What items are commonly purchased together?

# Common Data Mining Tasks

- Profiling (behavior description)
  - What does "normal behavior" look like? (for example, as baseline to detect fraud)
- Data Reduction
  - Which latent dimensions describe the consumer taste preferences?
- Link Prediction
  - Since John and Jane share 2 friends, should John become Jane's friend?
- Causal Modeling
  - Why are my customers leaving?

# Supervised versus Unsupervised Methods

- "Do our customers naturally fall into different groups?"
  - No guarantee that the results are meaningful or will be useful for any particular purpose
- "Can we find groups of customers who have particularly high likelihoods of canceling their service soon after contracts expire?"
  - A specific purpose
  - Much more useful results (usually)
  - Different techniques
  - Requires data on the target
    - The individual's label

# Common Data Mining Tasks

| Task | Supervised methods | Unsupervised methods |
|---|---|---|
| Classification | X | |
| Regression | X | |
| Causal modeling | X | |
| Similarity matching | X | X |
| Link prediction | X | X |
| Data reduction | X | X |
| Clustering | | X |
| Co-occurrence grouping | | X |
| Profiling | | X |

# Common Data Mining Tasks

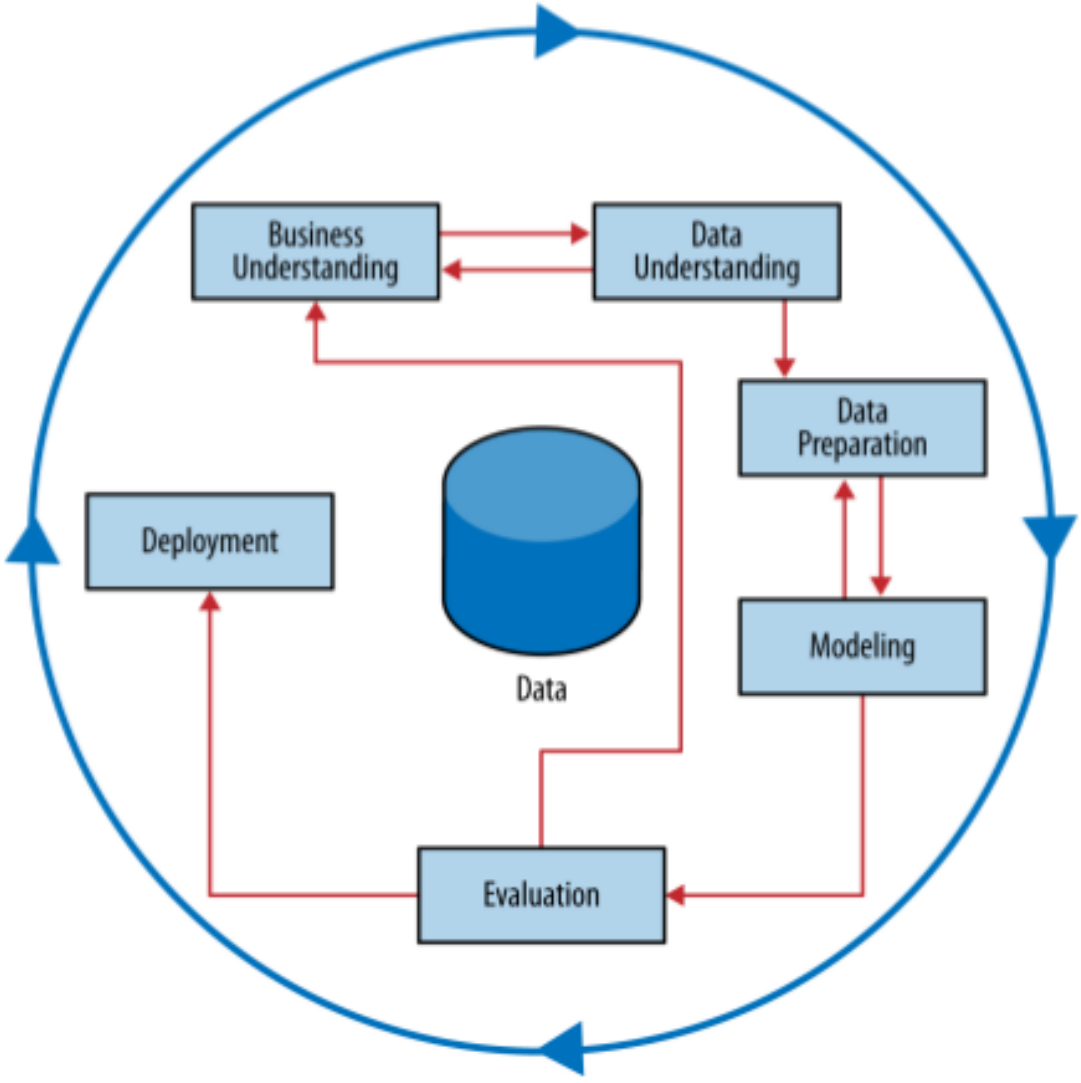
| Task | Supervised methods | Unsupervised methods |
|---|---|---|
| Classification | X | |
| Regression | X | |
| Causal modeling | X | |
| Similarity matching | X | X |
| Link prediction | X | X |
| Data reduction | X | X |
| Clustering | | X |
| Co-occurrence grouping | | X |
| Profiling | | X |

- Which task would you use for the churn example? The Walmart one?

# Supervised Data Mining & Predictive Modeling

- Is there a specific, quantifiable target that we are interested in or trying to predict?
  - Think about the decision
- Do we have data on this target?
  - Do we have **enough** data on this target?
- Do we have relevant data prior to decision?
  - Think timing of decision and action
- The result of supervised data mining is a model that predicts some quantity
- A model can either be used to predict or to understand

# Data mining process




Provost/Fawcett - Data Science for Business

# Jupyter notebook

- **Jupyer Notebook** is a browser-based development and note taking environment
- **Markdown** is a light-weight text formatting language

# Jupyter notebook

- **Jupyer Notebook** is a browser-based development and note taking environment
- **Markdown** is a light-weight text formatting language
- **Exercise**
  - Register to notebooks.azure.com
  - Open class2.ipynb
- **Homework**
  - Finish class2.ipynb