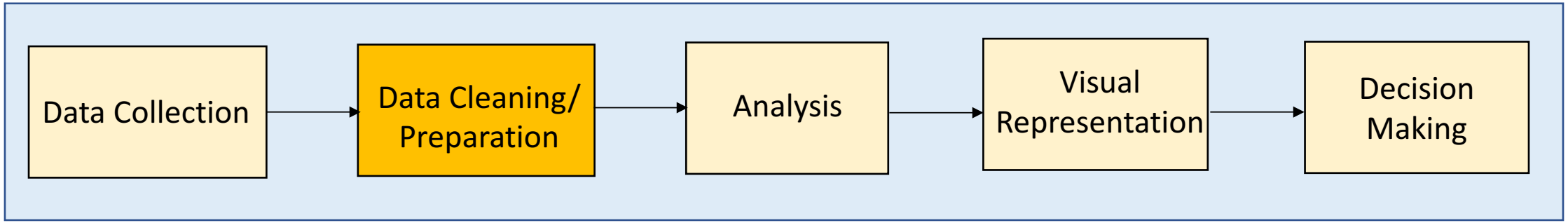
Machine Learning in the Wild

Dealing with Messy Data

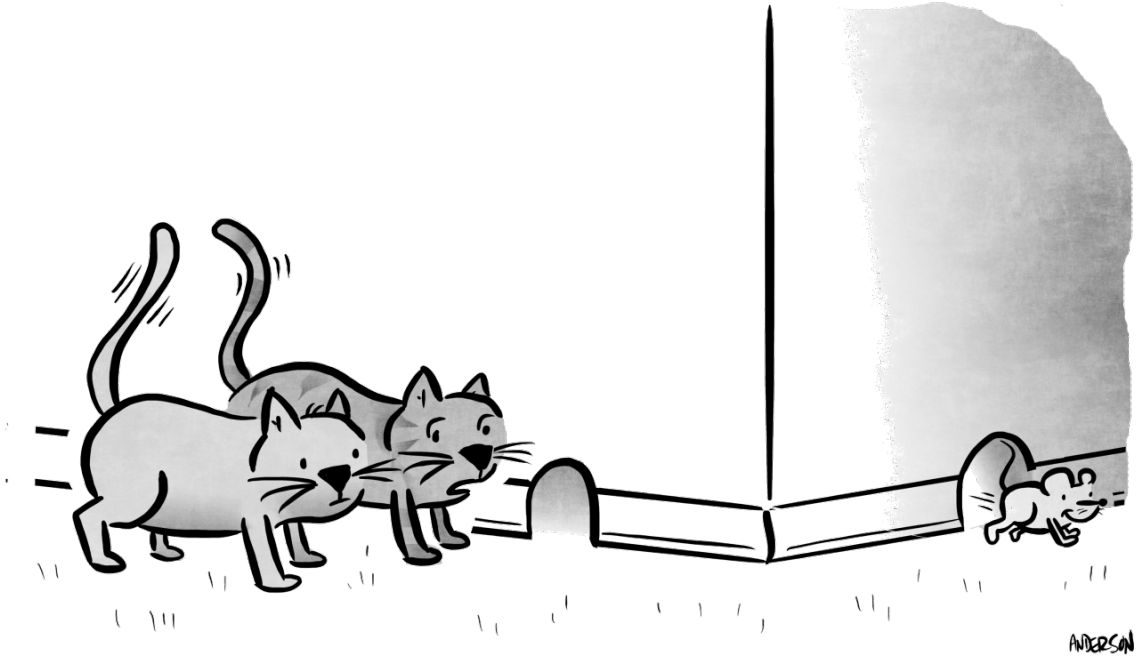
Rajmonda S. Caceres

Analytical Chain: From Data to Actions



What this lecture is about?

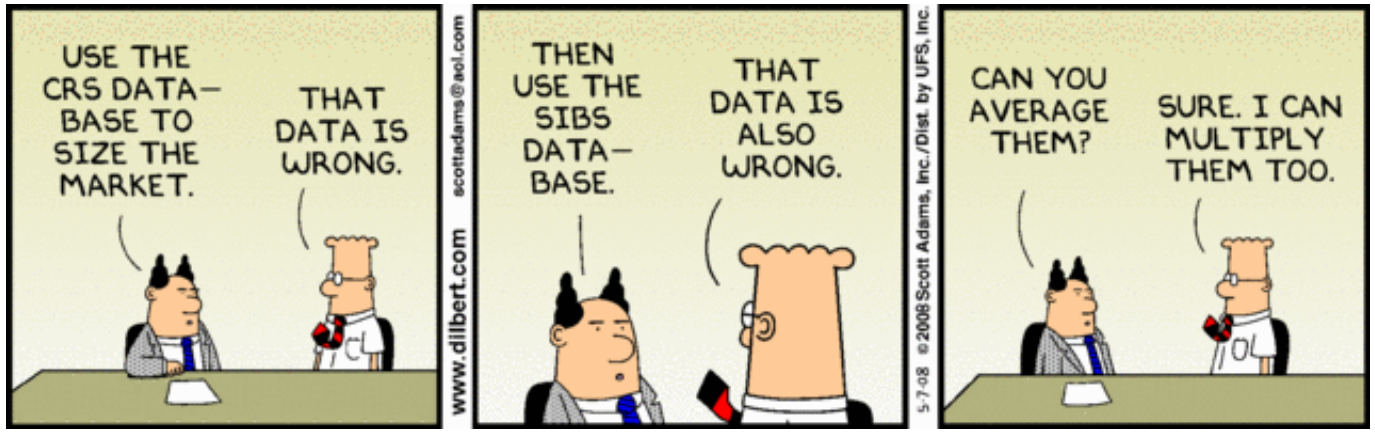
- What are some of data quality issues in real applications?
- Why should we care about data quality?
- How can we leverage statistical techniques to improve the quality of real data?



"According to our current predictive analytics solution, the mouse should be exiting from this hole in 3… 2… 1…" #betterdata

Data Quality Issues In Real Applications

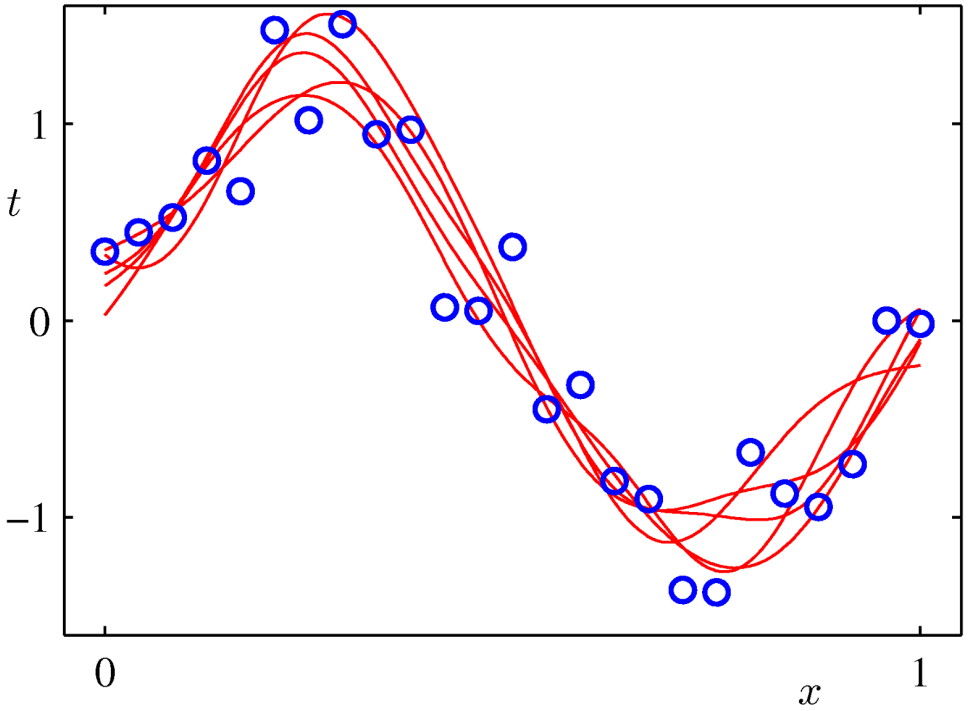
- **Missing data**
- **Outliers**
- Noisy data
- Data mismatch
- Curse of dimensionality



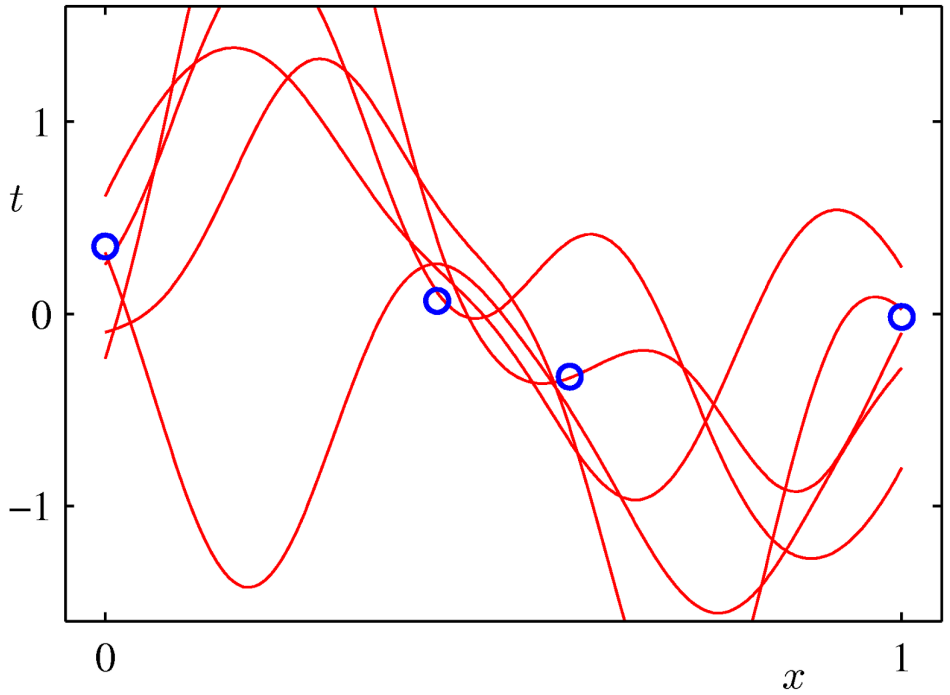
- Data quality greatly effects analysis and insights we draw from data
  - Introduces bias
  - Causes information loss

Effects of Missing Data On Regression

- Which regression model should we pick?



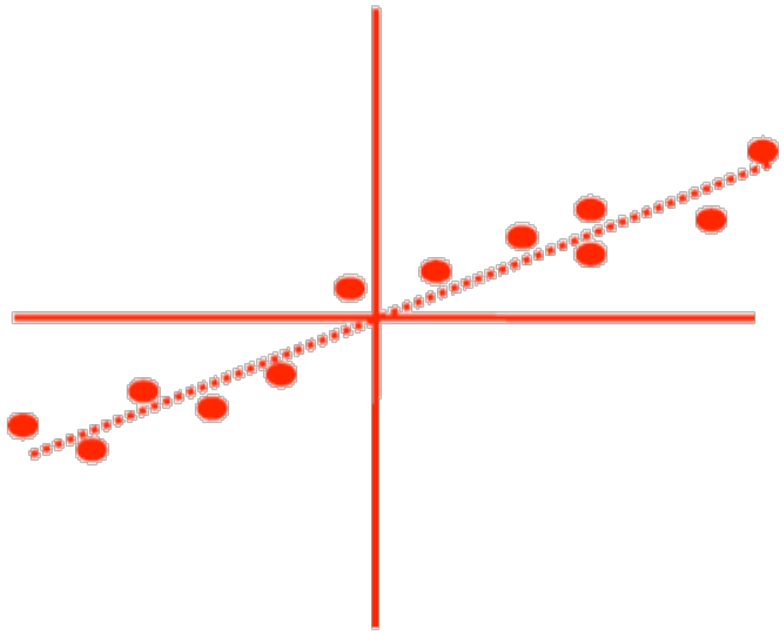
Sufficient data points to perform regression



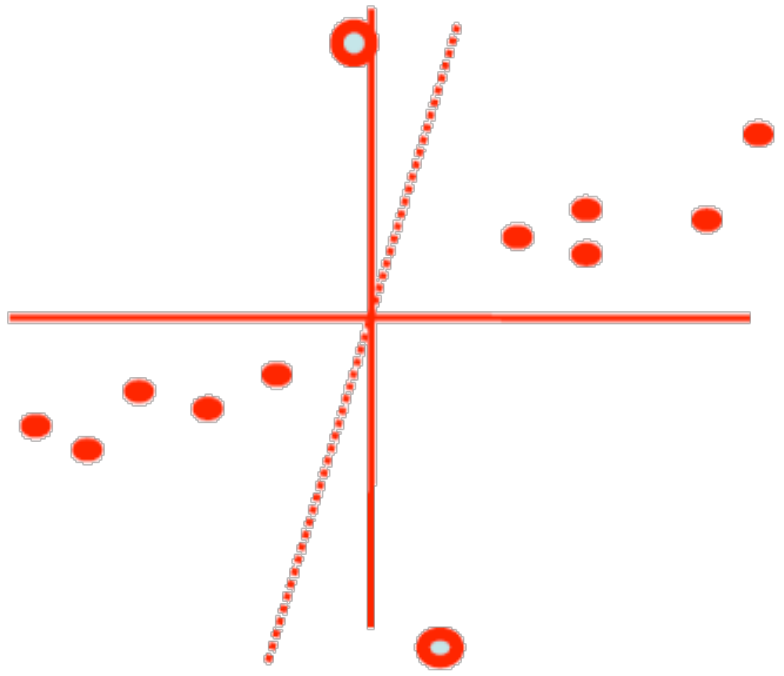
Insufficient data points to perform regression

Figures are from Chris Bishop's Book "Pattern Recognition and Machine Learning"

Effects of Outliers & Missing Data On PCA



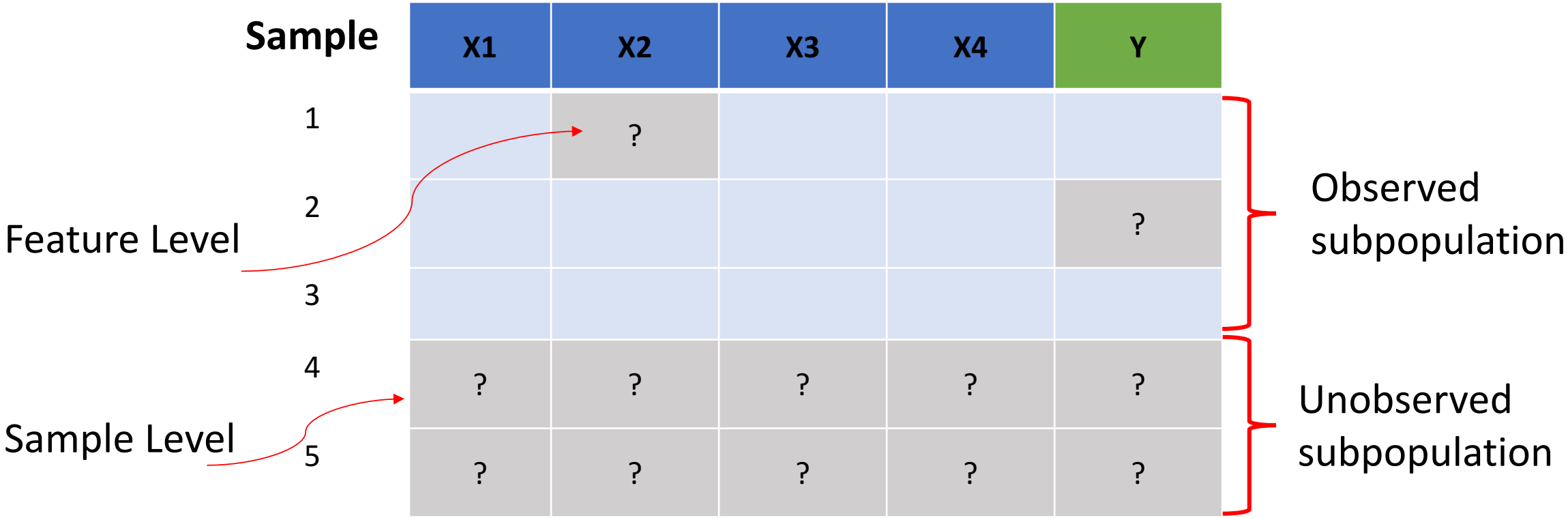
Case 1



Case 2

- Poor quality data greatly affects dimensionality reduction methods

Missing Data



- How much missing data is too much?
  - **Rough guideline:** If less < 5-10%, likely inconsequential, if greater need to impute

Mechanisms of Missing Data (Rubin 1976)

- Missing Completely At Random(MCAR): missing data does not depend on observed and unobserved data
  - P(missing|complete data) = P(missing)
- Missing At Random (MAR): missing data depends only on observed data
  - P(missing|complete data) = P(missing|observed data)
- Missing Not At Random (MNAR):
  - P(missing|complete data) $\not=$ P (missing|observed data)

Mean Imputation

1. Assume the distribution of missing data is the same as observed data
   - Replace with mean, median or other point estimates
   - Advantages:
     - Works well if MCAR
     - Convenient, easy to implement
   - Disadvantages:
     - Introduces bias by smoothing out the variance
     - Changes the magnitude of correlations between variables

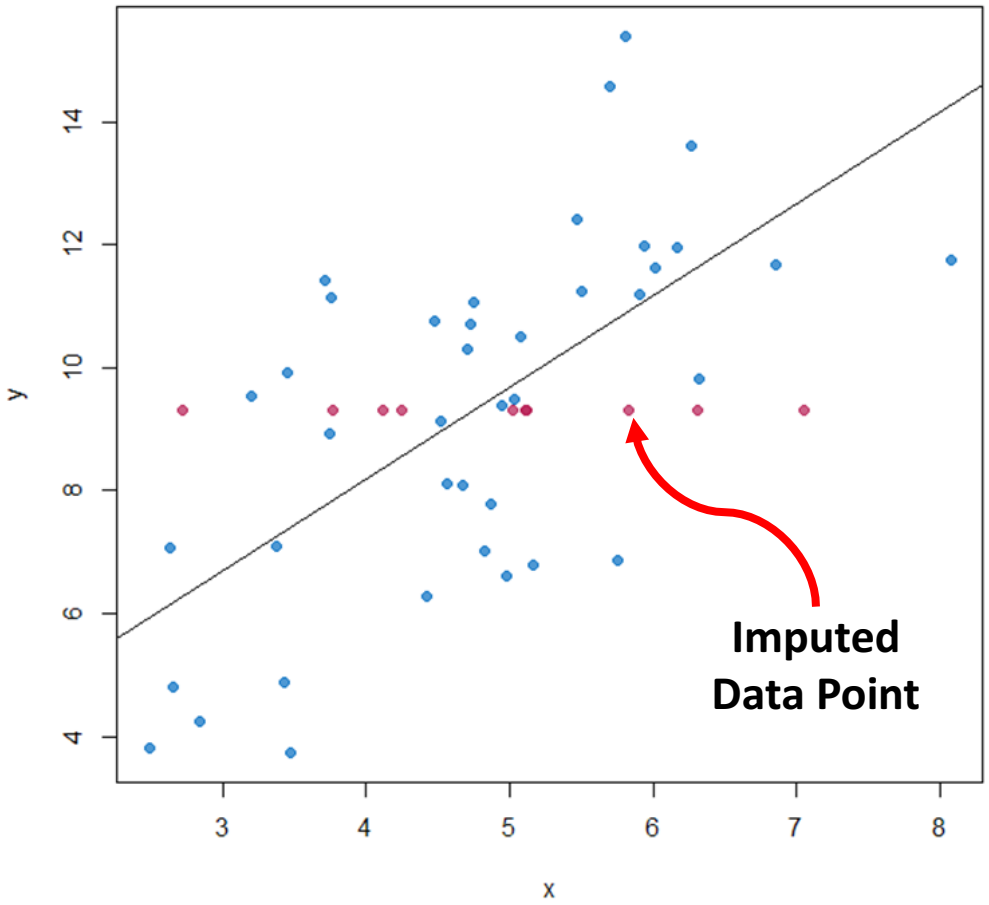


Figure taken from:
https://www.iriseekhout.com/missing-data/missing-data-methods/imputation-methods/

Regression Imputation

- Better imputation - leverage attribute relationships
- Monotone missing patterns
- Replace missing values with predicted scores from regression equation

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$

- Stochastic regression: add an error term
- Advantage:
  - Use information from observed data
- Disadvantage:
  - Overestimate model fit, weakens variance

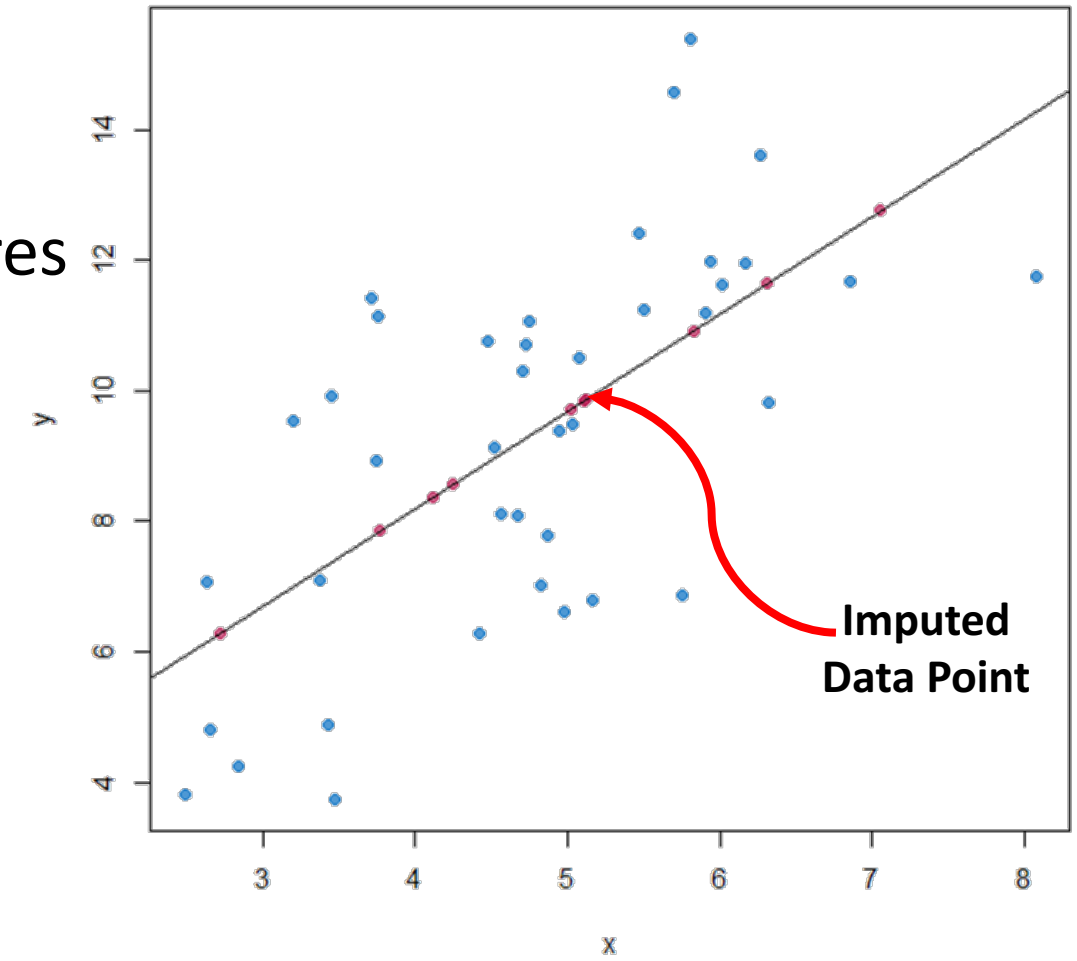
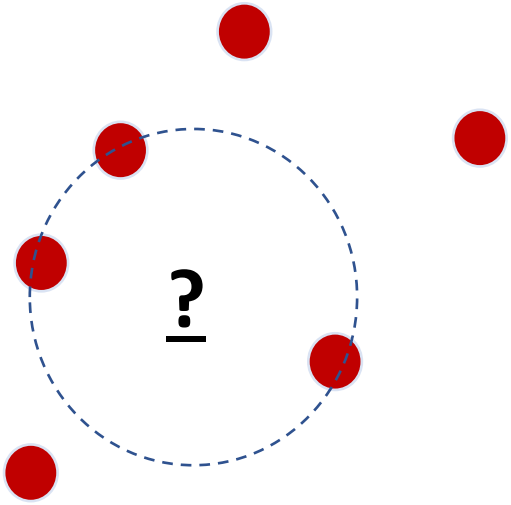


Figure taken from:
https://www.iriseekhout.com/missing-data/missing-data-methods/imputation-methods/

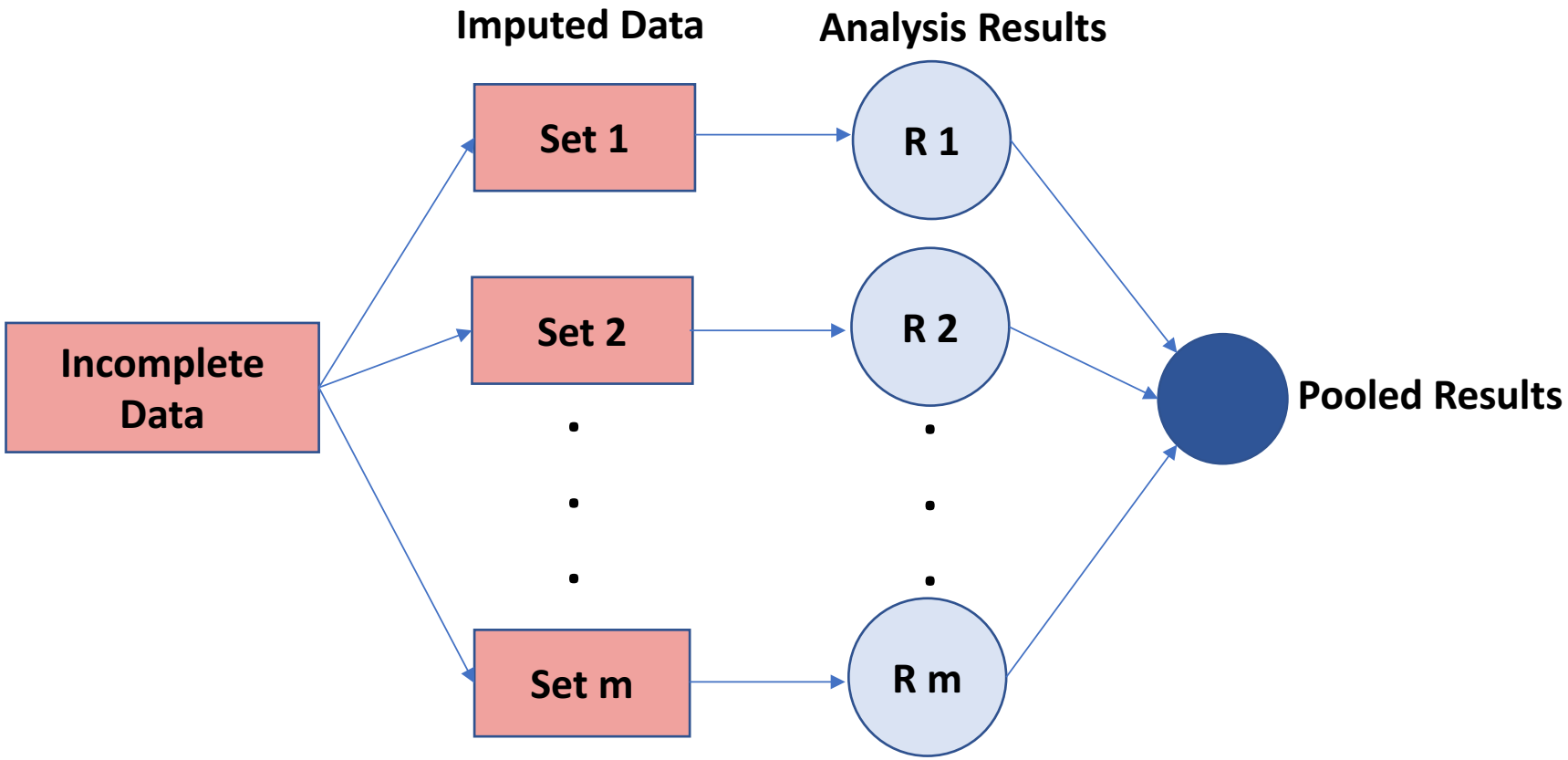
k-Nearest Neighbor (kNN) Imputation

- Leverage similarities among different samples in the data
- Advantages:
  - Doesn’t require a model to predict the missing values
  - Simple to implement
  - Can capture the variation in data due to its locality
- Disadvantages:
  - Sensitive to how we define what similar means



kNN Imputation, k=3

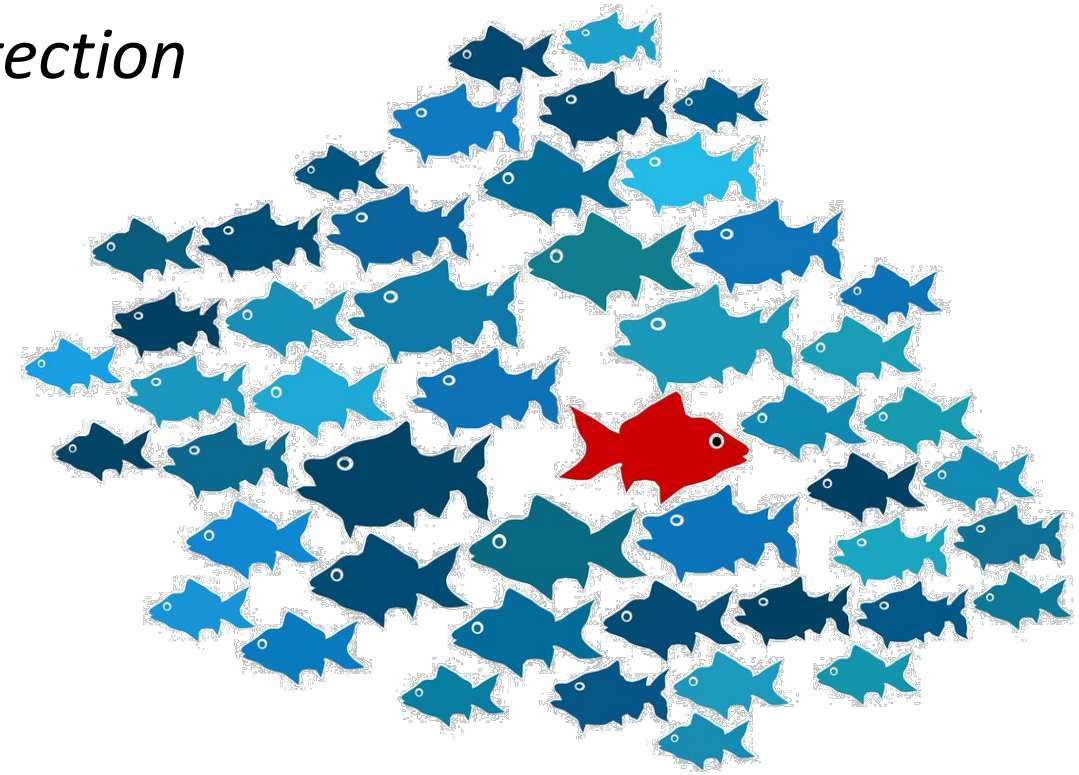
Multiple Imputation



- Treat the missing value as a random variable and impute it **m** times
- Minimize the bias introduced by data imputation through averaging

Outliers

- **Outlier**: Observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism", Hawkins(1980)
- Outliers vs noise
- Outlier detection vs. *novelty detection*



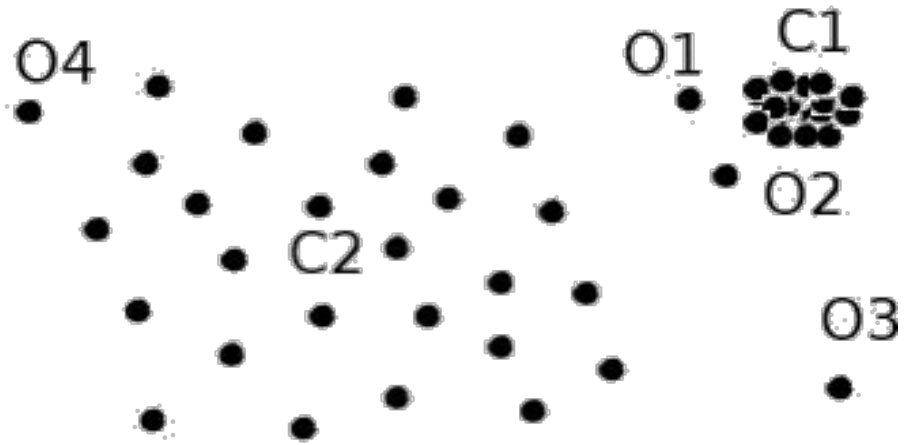
Outlier Detection Techniques

- Make assumptions about normal data and outliers
  - **Statistical**
  - **Proximity-based**
  - Clustering-based methods

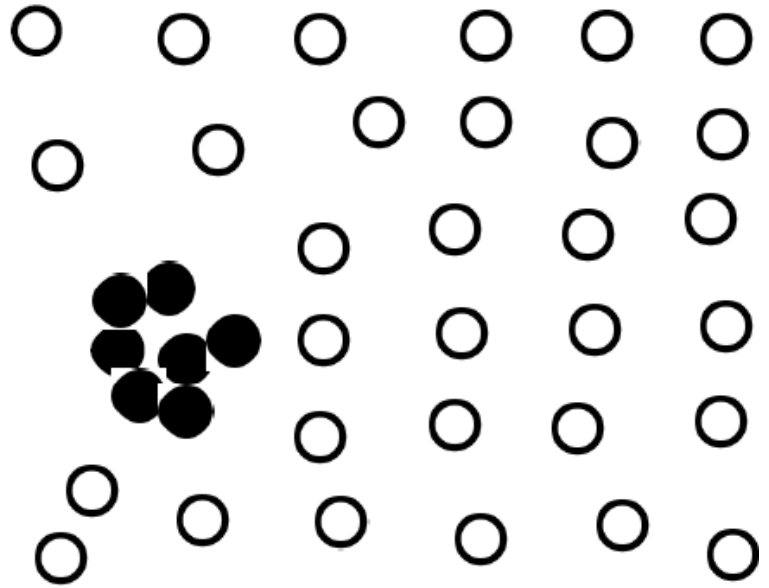
- Have access to labels of outlier examples
  - **Supervised**
  - Semi-supervised
  - Unsupervised

Types of Outliers

- **Universal vs contextual**
  - O1, O2 local outliers (relative to cluster C1)
  - 03 global outlier

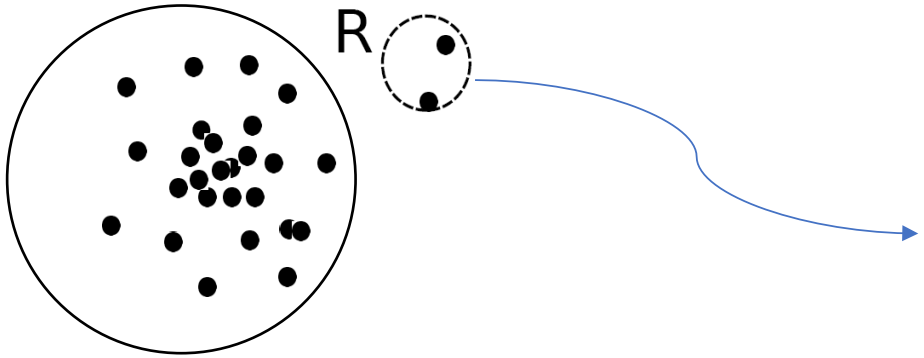


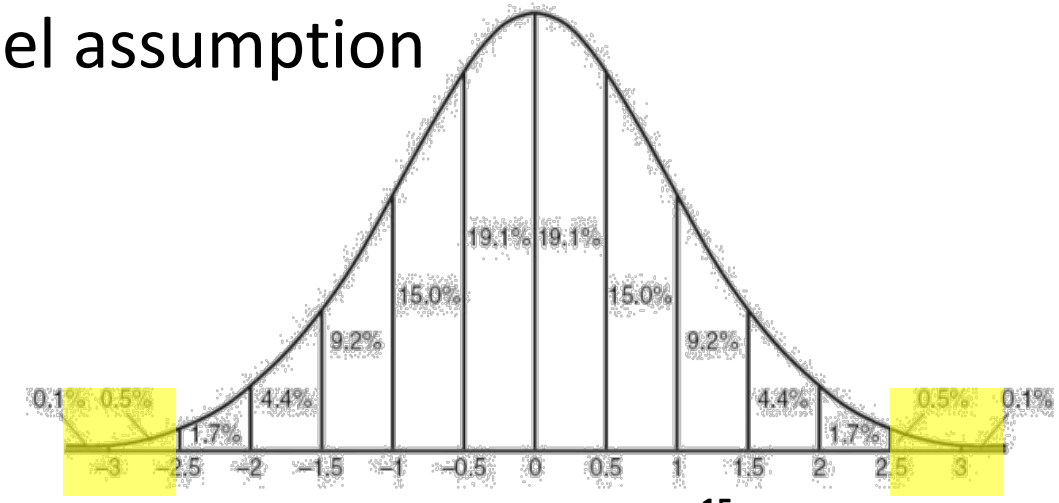
- **Singular vs collective**



Statistical Methods

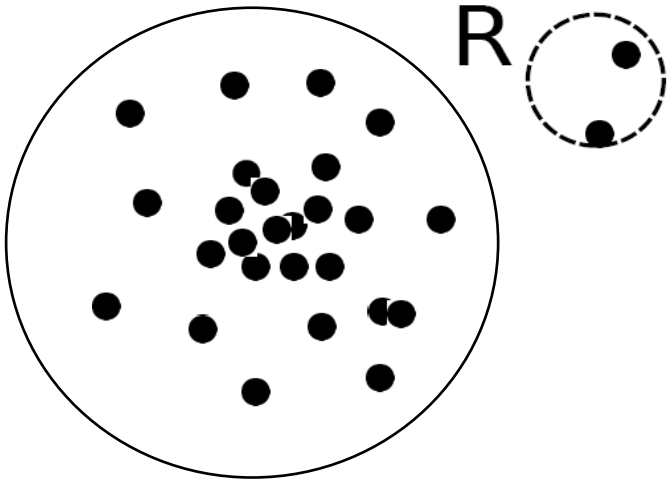
- Assume ``normal'' observations follow some statistical model
  - Example: Assume the normal points come from a Gaussian distribution.
- Learn the parameters of the model from the data
- Data points not following the model are considered outliers
  - Example 1: Throw away points that fall at the tails
  - Example 2: Throw away low probability points
- Disadvantage: Method is depended on model assumption





Classification-Based Method

- If labels of outlier data points exist we could treat the problem as a classification problem
- Train a classifier to separate the two classes, “normal’ and outliers class
  - Usually heavily biased toward the normal class: “Unbalanced classification problem”
  - Cannot detect “unseen” outliers
  - Often unrealistic to assume we have labels for outlier points
- One class classification: learn the boundary for the normal class. Points outside the boundary are considered outliers
  - Can detect new outliers



Proximity-Based Methods

- If the near-by points are far away, consider the data point an outlier
- No assumption on labels or models of “normal” distribution
- No free lunch though: we rely on the robustness of the proximity measure
- **Distance-based methods:**
  - An observation is an outlier if its neighborhood does not have enough other observations
- **Density-based methods:**
  - An observation is an outlier if its density is relatively much lower than that of its neighbors

Density-based Outlier Detection

- Local Outlier Factor (LOF) Algorithm (Breuning 2000)
- For each point, compute the k nearest neighbors N(j)
- Compute the point density

$$f(i) = \frac{k}{\sum_{j \in N(j)} d(i,j)}$$

- Compute the local outlier score:

$$LOF(i) = \frac{1/k \sum_{j \in N(j)} f(j)}{f(i)}$$

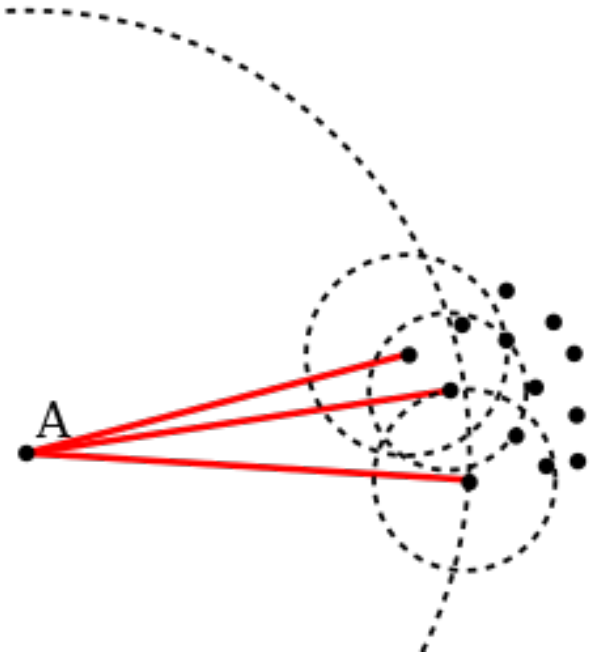


Figure taken from:
https://commons.wikimedia.org/wiki/File:LOF-idea.svg

- As the name suggests, LOF is robust in detecting local outliers

Main Takeaways

- Explore and understand as much as possible the quality of your data
- Identify appropriate techniques that can mitigate some of the issues of data quality
  - **Some of the same techniques you are learning in this course can be leveraged to improve the quality of the training data**
- Know when you need more data
- Understand biases introduced by data imputation techniques
- Approach data science as an iterative process - all the components are connected

Your analysis is only as good as your data

References

- Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3): 581-592.
- Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, J. Wiley & Sons.
- Little, R. J. and D. B. Rubin (2002). Statistical Analysis with Missing Data. Hoboken, NJ, John Wiley & Sons.
- Breunig, M. M. , Kriegel, H. , Ng, R. T. , Sander, J. (2000). LOF: Identifying Density-Based Local Outliers