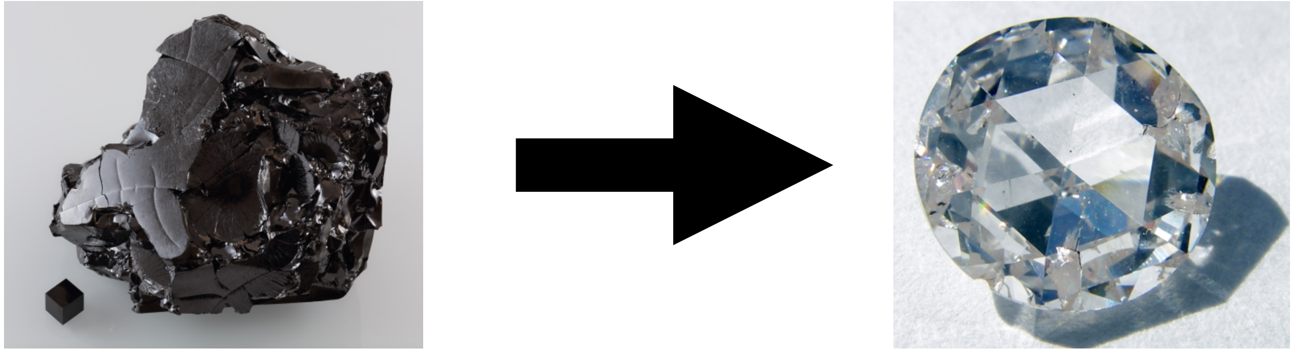
The Data Science Process

Polong Lin
Big Data University Leader & Data Scientist
IBM
polong@ca.ibm.com

BIG DATA
UNIVERSITY

*“**Every day**,*
*we create **2.5 quintillion bytes** of data —*
*so much that **90% of the data** in the world today*
*has been created in the **last two years** alone.”*



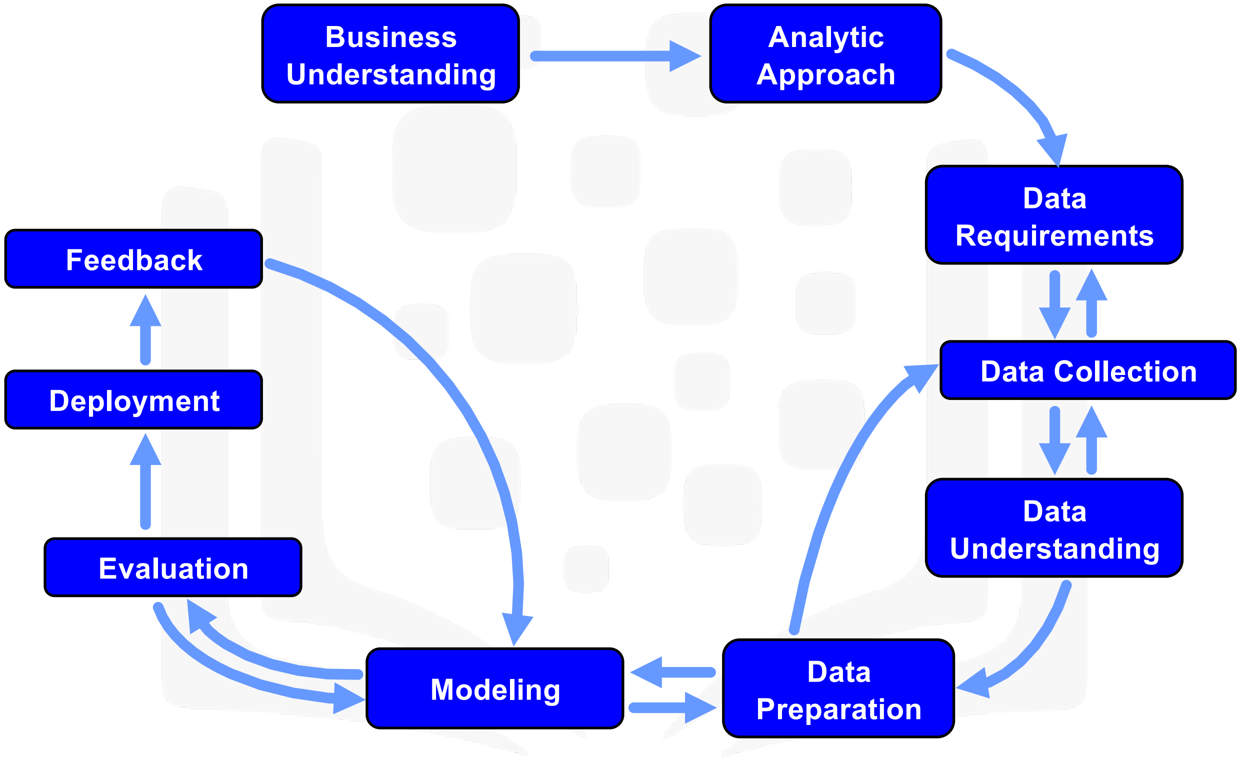
Data science

The interest in data science

- Solve problems and answer questions using data
- Goal to improve future outcomes

What is the data science process?

CRISP-DM Methodology diagram



Cross Industry Standard Process for Data Mining

1. Business understanding

Business Understanding

Every project begins with **business understanding**.

- Project objective?
- Business sponsors play the most critical role
- What are we trying to do – what is the goal?
- How do you define “success” and how can you measure it?

1. Business understanding

Business Understanding

Traffic:

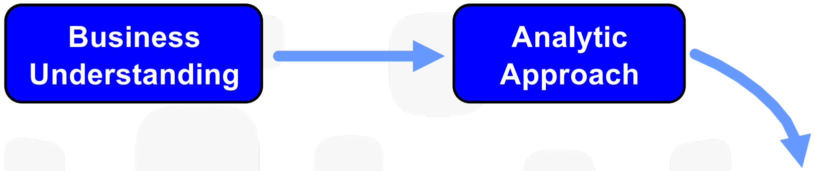
Problem: Traffic congestion wastes time and money

Clear question: How can we optimize traffic light duration using data on traffic patterns, weather, and pedestrian traffic?

Measurable outcomes:

- % decrease in commute time
- % decrease in length/duration of traffic jams

2. Analytic Approach



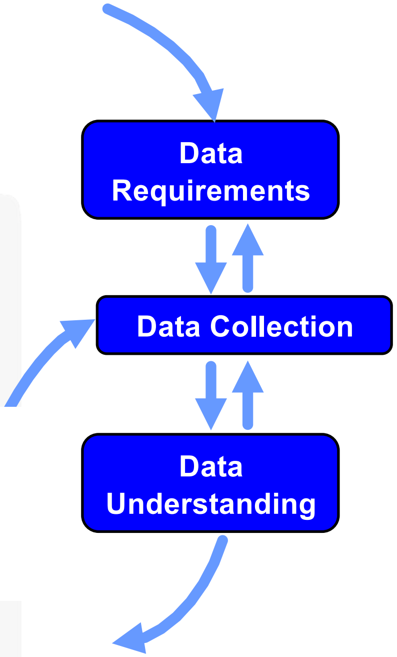
- Express problem in context of statistical and machine learning techniques
  - **Regression**:
    - “Predicting revenue in the next quarter?”
  - **Classification**:
    - “Does this patient have cancer A, cancer B, or are they healthy?”
  - **Clustering**:
    - “Are there groups of users that seem to behave similarly to each other?”
  - **Recommendation/Personalization**:
    - “How can I target discounts to specific customers?”
  - **Outlier Detection**

Statistical / machine learning technique(s)

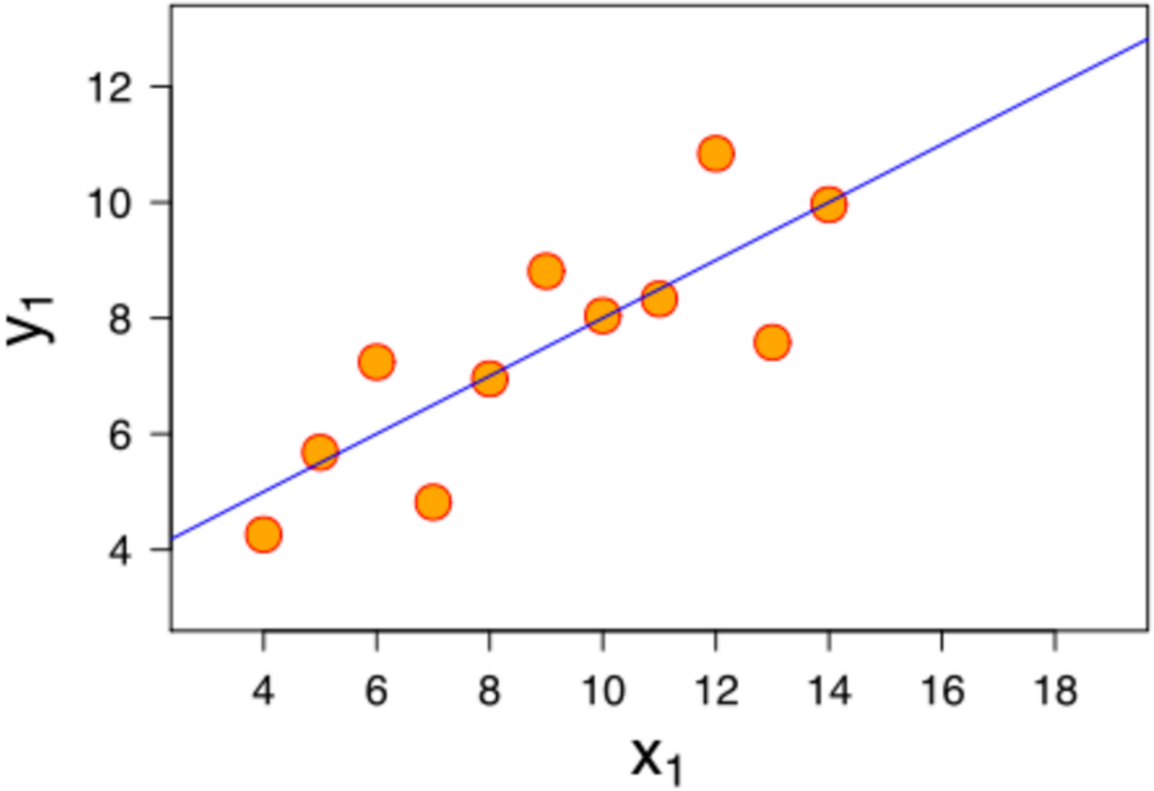
- Linear regression
- Logistic regression
- Clustering
  - K-means
  - Hierarchical
  - Density-based
- Classification Trees
- Random Forests
- Neural networks
- Text mining (natural language processing)
- Principal component analysis
- Support Vector Machines
- Hidden Markov Models
- …

Data compilation

- The chosen analytic approach determines the **data requirements**.
    - Content, formats, representations
- Initial **data collection** is performed.
    - Available Data?
    - Obtain data?
    - Revise data requirements or collect more data?
- Then **data understanding** is gained.
    - Initial insights about data
    - Descriptive statistics and visualization
    - Additional data collection to fill gaps, if needed

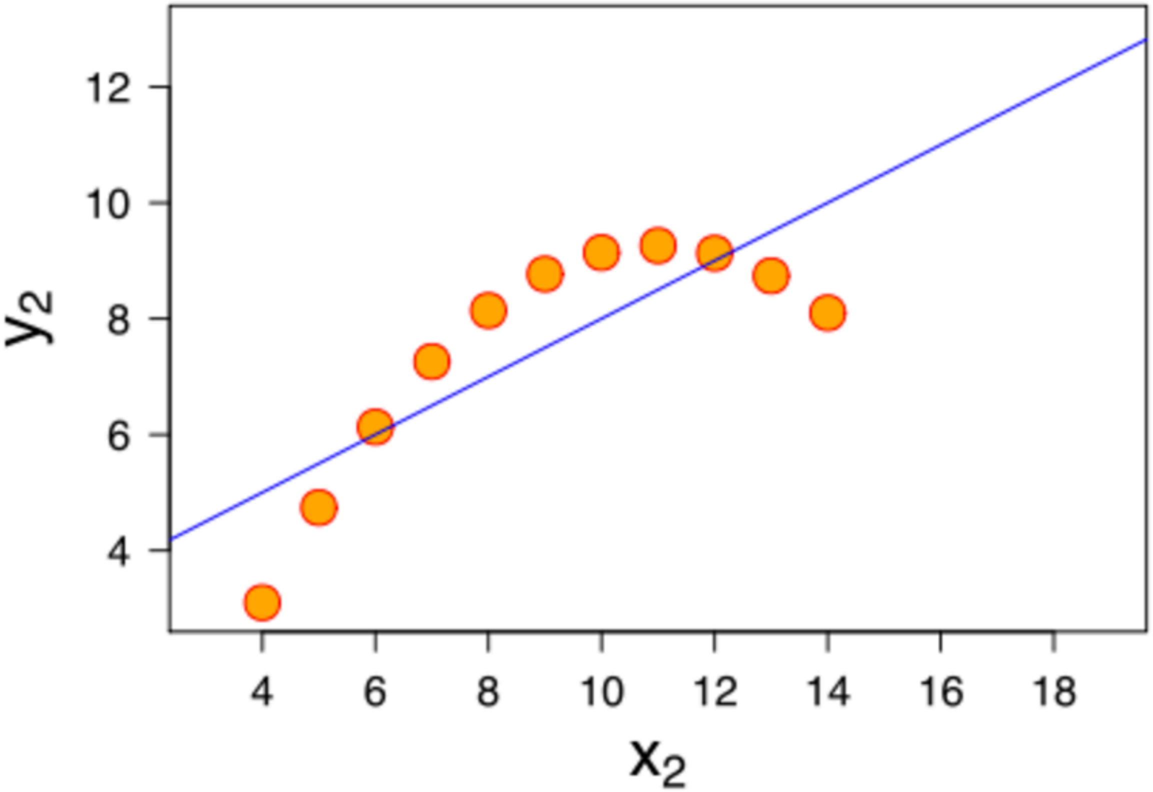


#1 What can you tell me about this data?

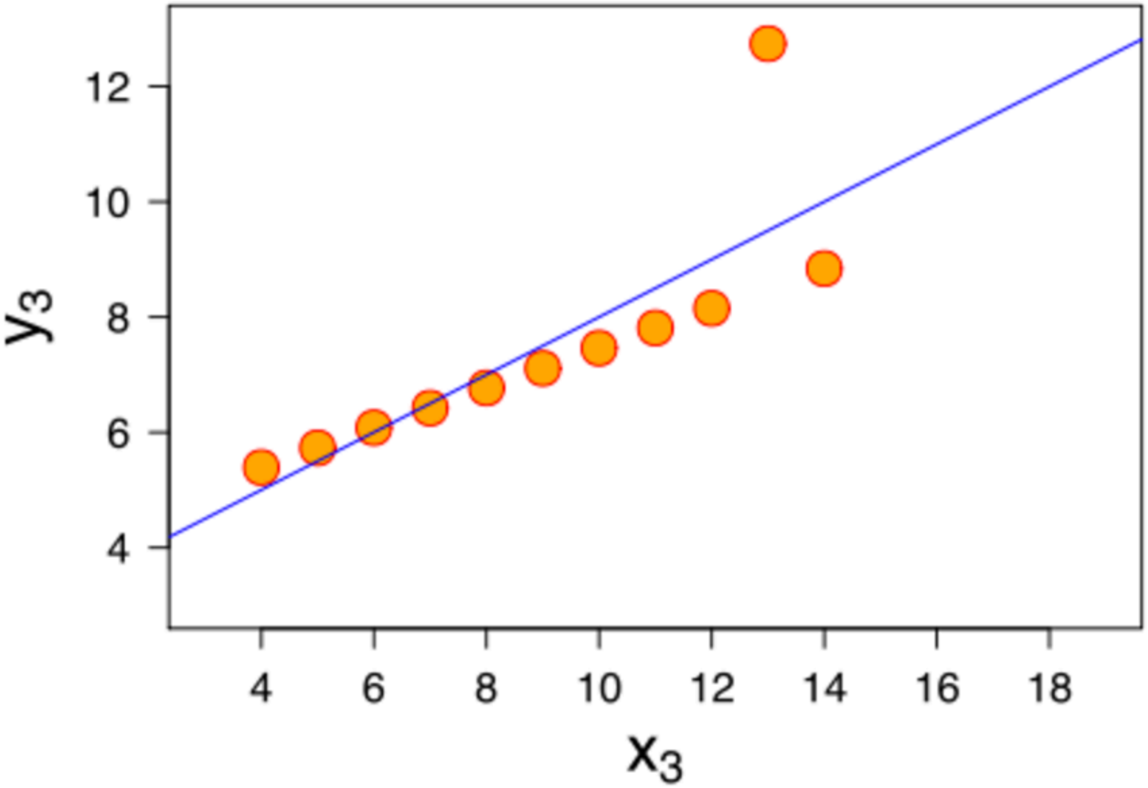


BIG DATA UNIVERSITY

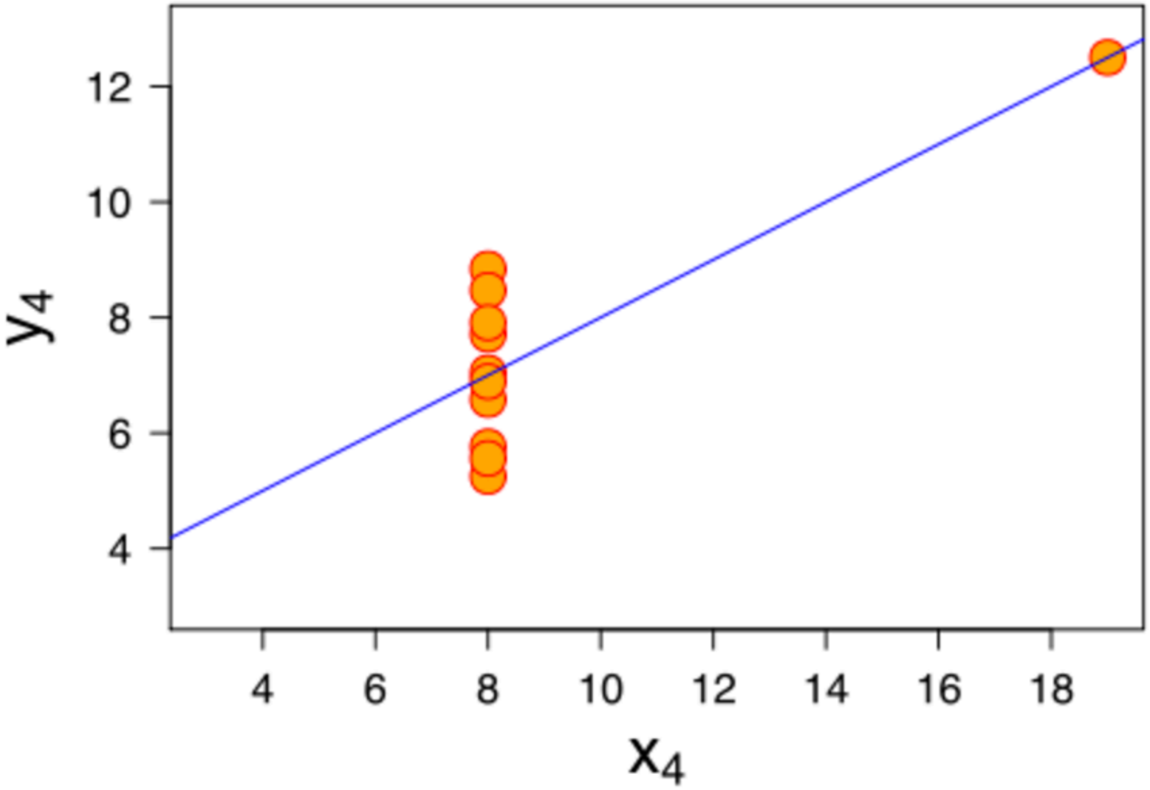
#2 What can you tell me about this data?



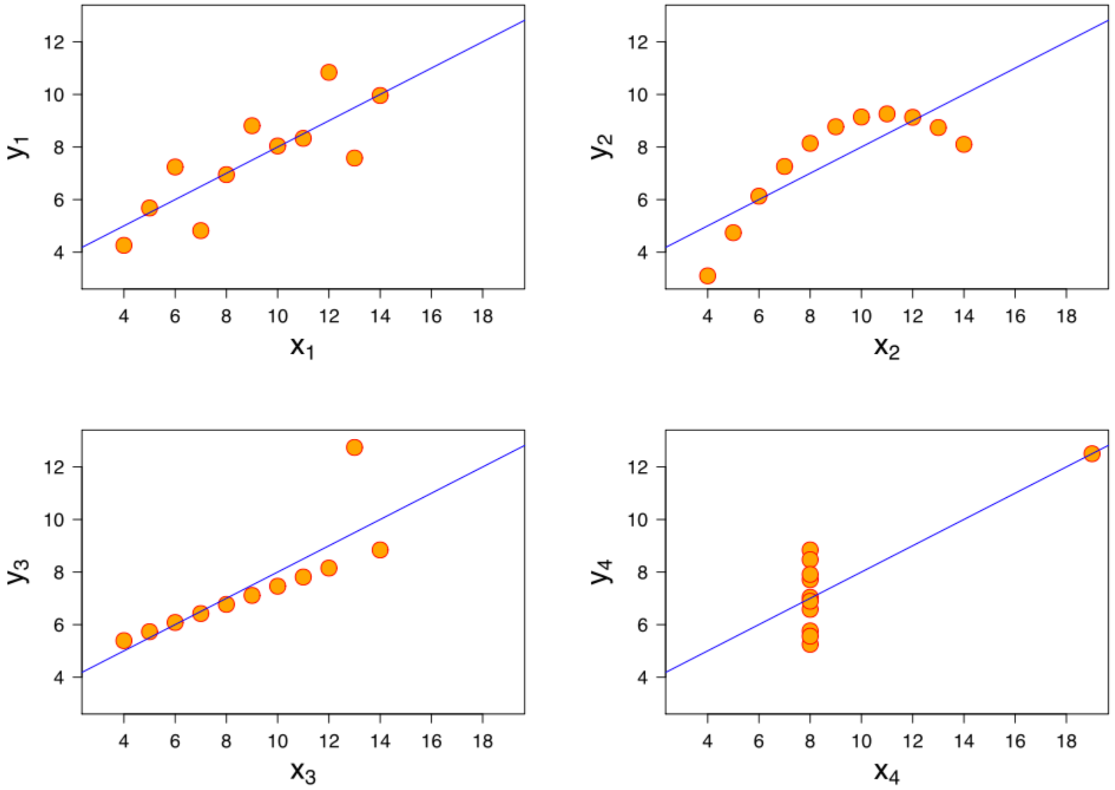
#3 What can you tell me about this data?



#4 What can you tell me about this data?



Importance of Visualization



Same properties:

mean(x) = 9

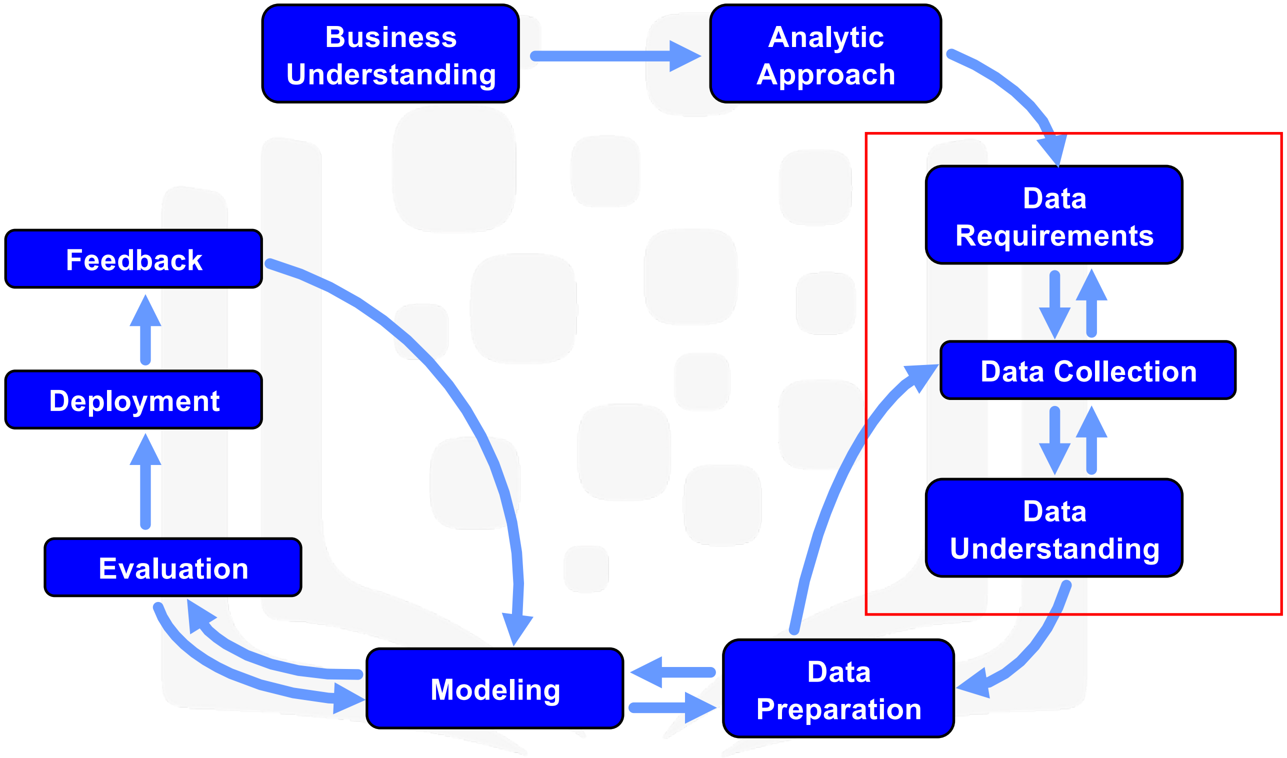
mean(y) = 7.5

y = 3.00 + 0.500x

corr(x,y) = 0.816

Anscombe's Quartet

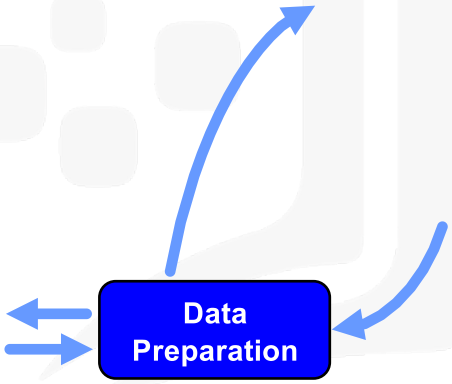
CRISP-DM Methodology diagram



Data preparation

- **Data preparation** encompasses all activities to construct and clean the data set.
  - Data cleaning
    - Missing or invalid values
    - Eliminating duplicate rows
    - Formatting properly
  - Combining multiple data sources
  - Transforming data
  - Feature engineering
  - Text analysis
- Accelerate data preparation by automating common steps

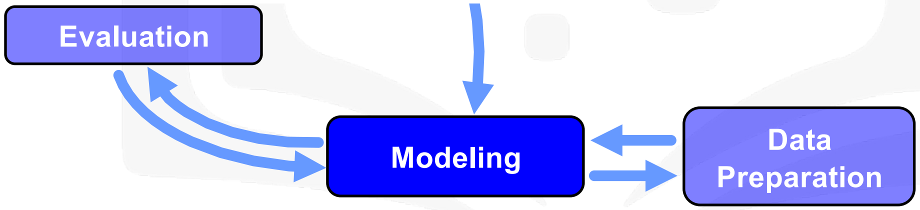
- Arguably the most time-consuming step
- “80% of the entire DS process is in data cleaning and preparation”



Modeling

- **Modeling**:
  - Developing predictive or descriptive models
  - May try using multiple algorithms

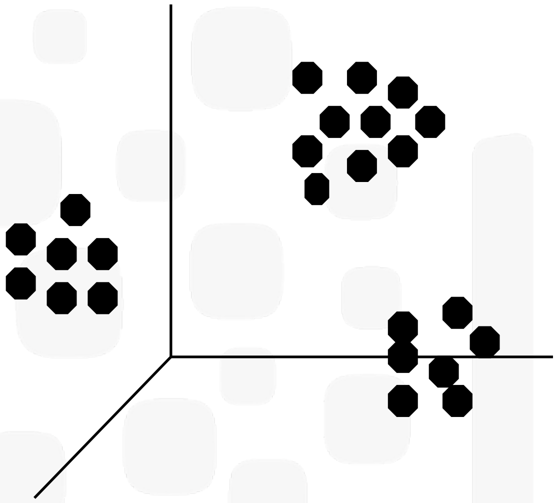
- Highly iterative process



BIG DATA UNIVERSITY

Example: Clustering

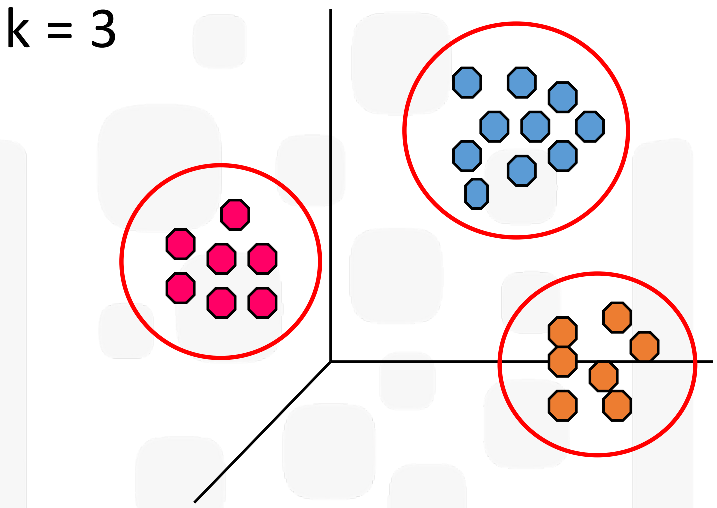
K-means Clustering



Group similar cuisines together into ***k*** number of clusters.

Example: Clustering

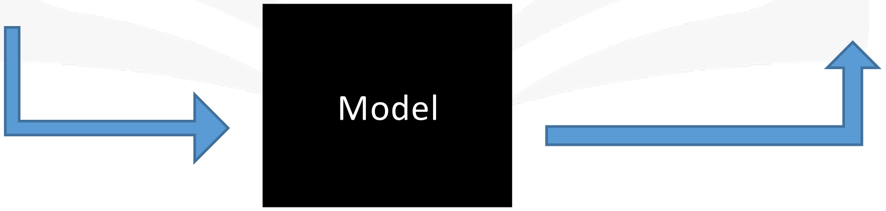
K-means Clustering



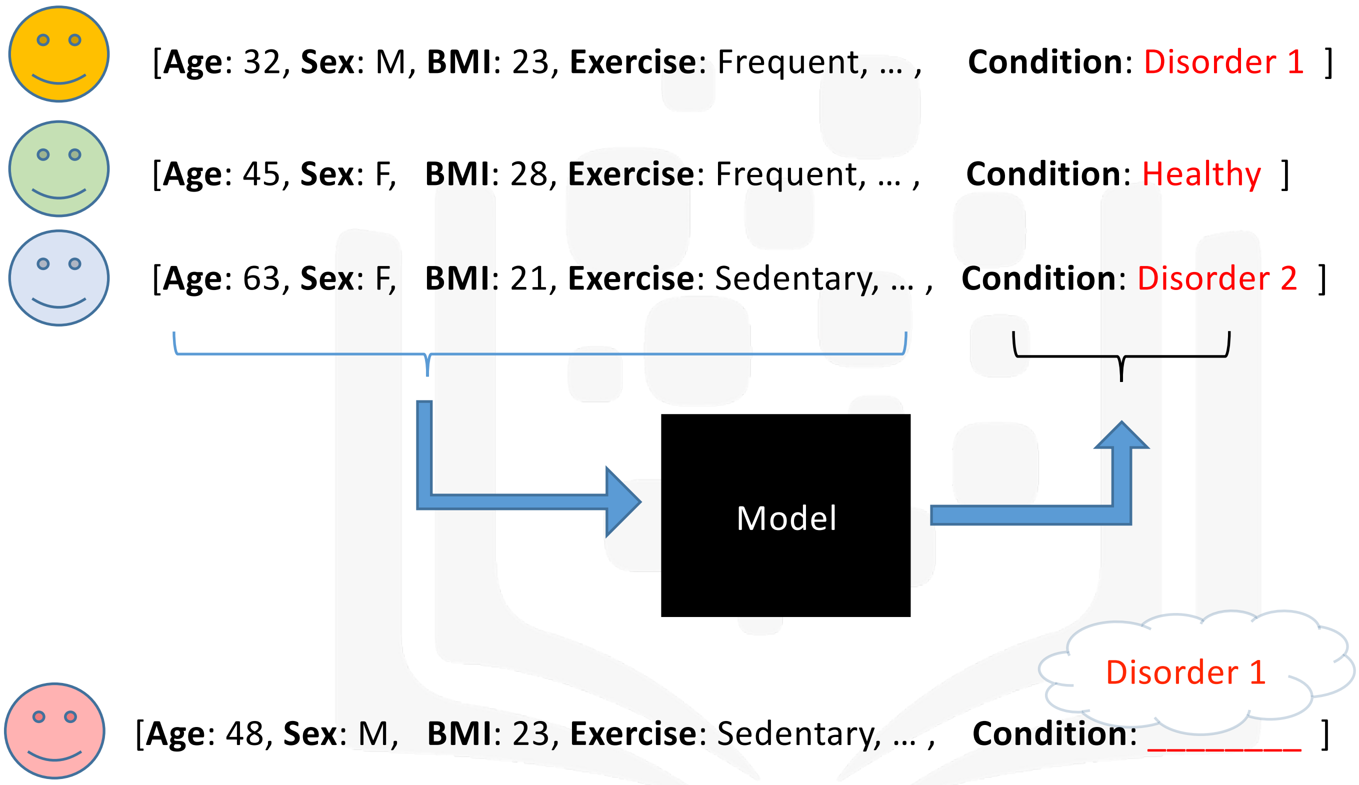
Group similar cuisines together into ***k*** number of clusters.

Example: Clustering

[**Age**: 18, **Sex**: M, **BMI**: 23, **Exercise**: Frequent, **Hobbies**: Golf, …]	CLUSTER A
[**Age**: 45, **Sex**: F, **BMI**: 28, **Exercise**: Frequent, **Hobbies**: Baseball, …]	CLUSTER B
[**Age**: 83, **Sex**: F, **BMI**: 25, **Exercise**: Sedentary, **Hobbies**: Gymnastics, …]	CLUSTER C
[**Age**: 28, **Sex**: M, **BMI**: 23, **Exercise**: Normal, **Hobbies**: Softball, …]	CLUSTER B
[**Age**: 30, **Sex**: F, **BMI**: 25, **Exercise**: Normal, **Hobbies**: Golf, …]	CLUSTER A
[**Age**: 15, **Sex**: M, **BMI**: 22, **Exercise**: Frequent, **Hobbies**: Golf, …]	CLUSTER A



Example: Classification



[**Age**: 32, **Sex**: M, **BMI**: 23, **Exercise**: Frequent, … , **Condition**: Disorder 1]

[**Age**: 45, **Sex**: F, **BMI**: 28, **Exercise**: Frequent, … , **Condition**: Healthy]

[**Age**: 63, **Sex**: F, **BMI**: 21, **Exercise**: Sedentary, … , **Condition**: Disorder 2]

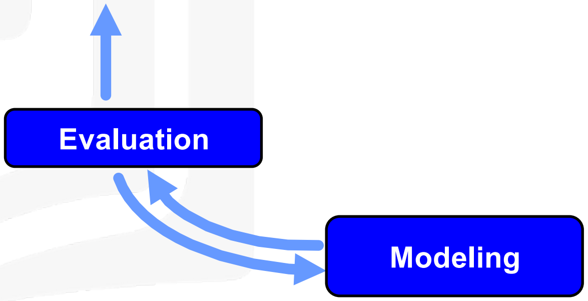
[**Age**: 48, **Sex**: M, **BMI**: 23, **Exercise**: Sedentary, … , **Condition**: ________]

Model evaluation

- Model **evaluation** is performed during model development and before model deployment.
  - Understand the model's quality
  - Ensure that it properly addresses the business problem

- Diagnostic measures
  - Suitable to the modeling technique used
  - Training/Testing set
  - Refine model as needed

- Statistical significance tests

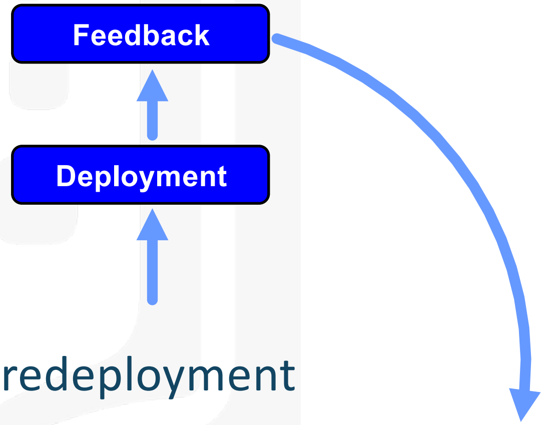


Deployment and feedback

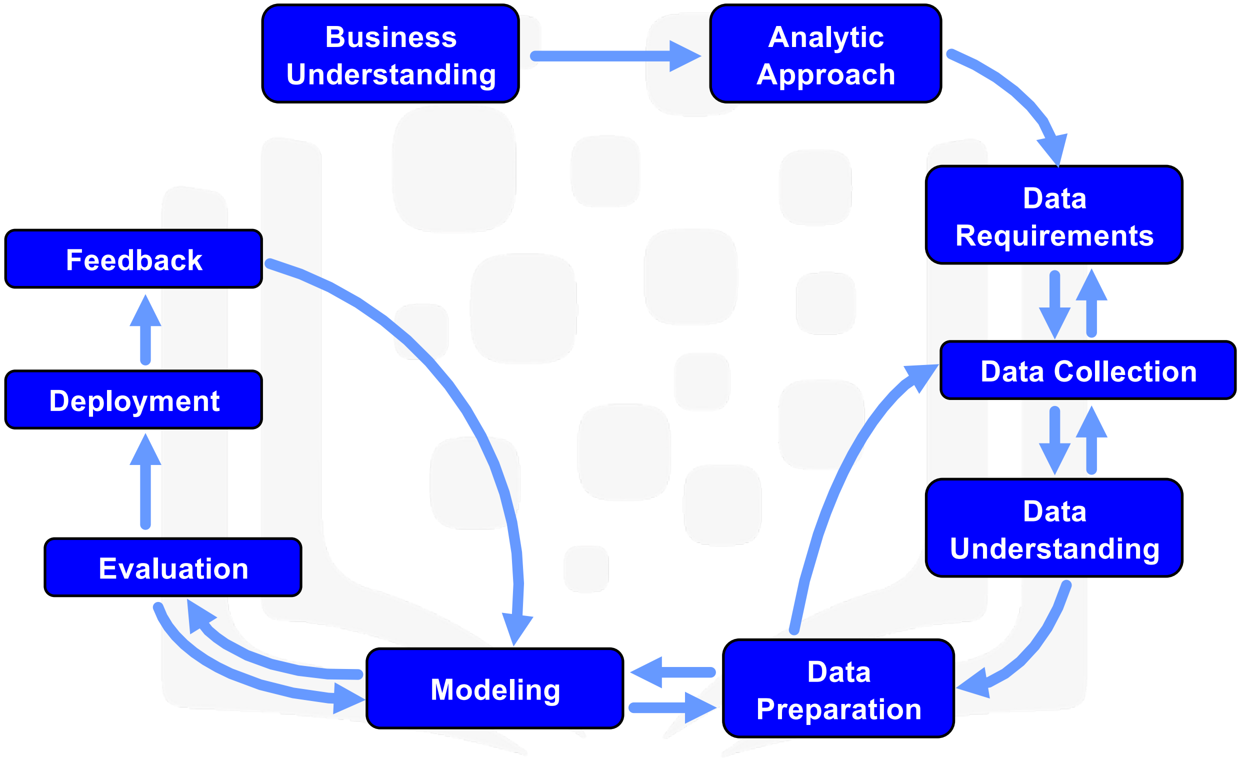
- Once finalized, the model is **deployed** into a production environment.
  - May start in a limited / test environment
  - Involves other roles:
    - Solution owner
    - Marketing
    - Application developers
    - IT administration
- Getting **Feedback :**
  - How well did the model perform?
  - Iterative process for model refinement and redeployment
  - A/B testing

Big Data University:

- Inactive -> Active



CRISP-DM Methodology diagram



“All models are wrong but some are useful” *– George Box, Statistician*