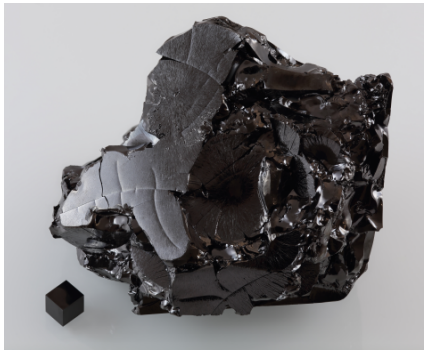
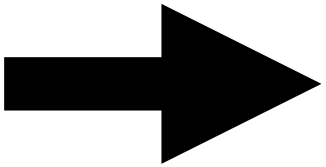
The Data Science Process

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."*





BIG DATA UNIVERSITY

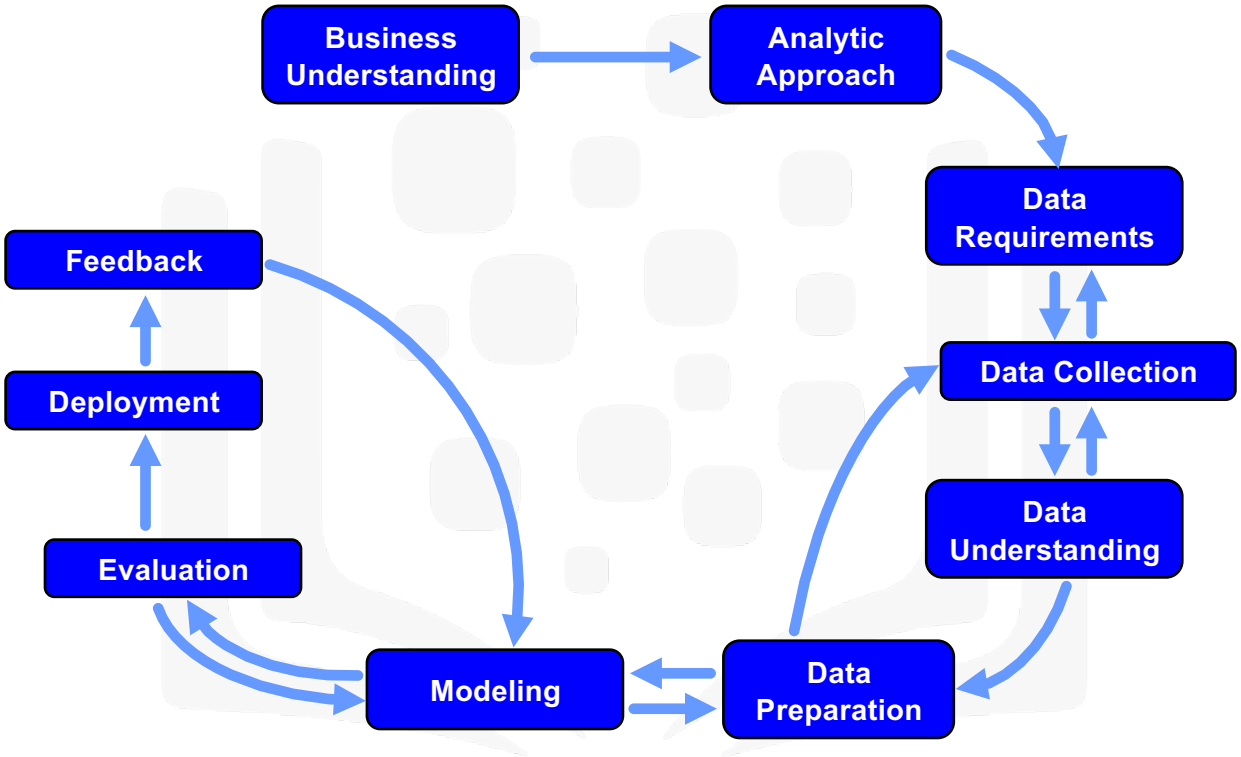
Data science

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



1. 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

Traffic:

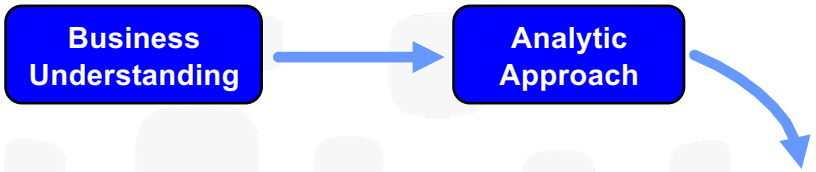
Problem: Traffic congestion wastes time and money

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



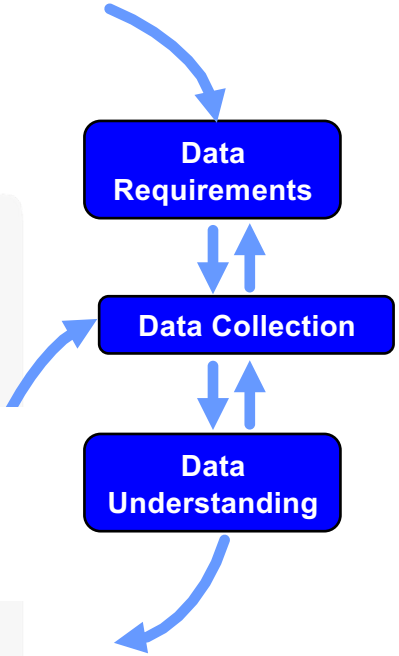
- Express problem in context of statistical and machine learning techniques
  - **Regression**:
    - "Predicting revenue in the next quarter?"
  - **Classification**:
    - "Does this patient have cancer A, cancer B, or are they healthy?"
  - **Clustering**:
    - "Are there groups of users that seem 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
- …

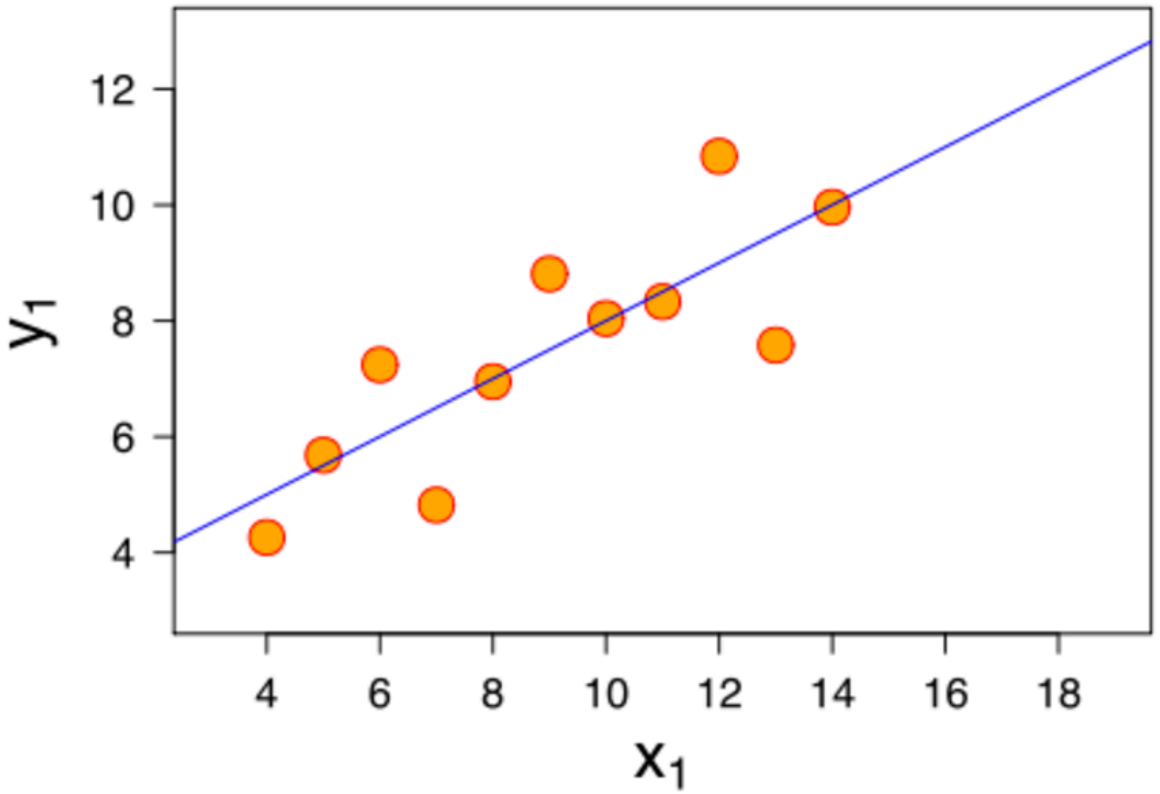
Data compilation

- 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



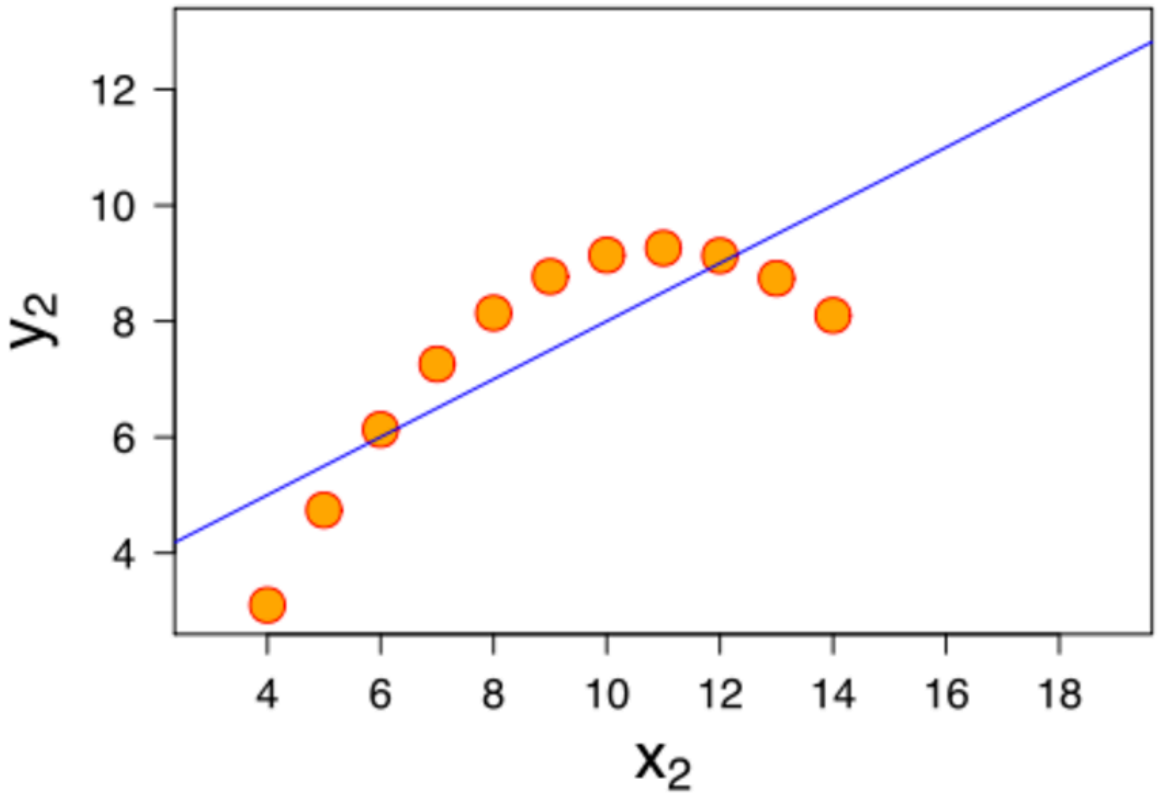
BIG DATA UNIVERSITY

#1 What can you tell me about this data?



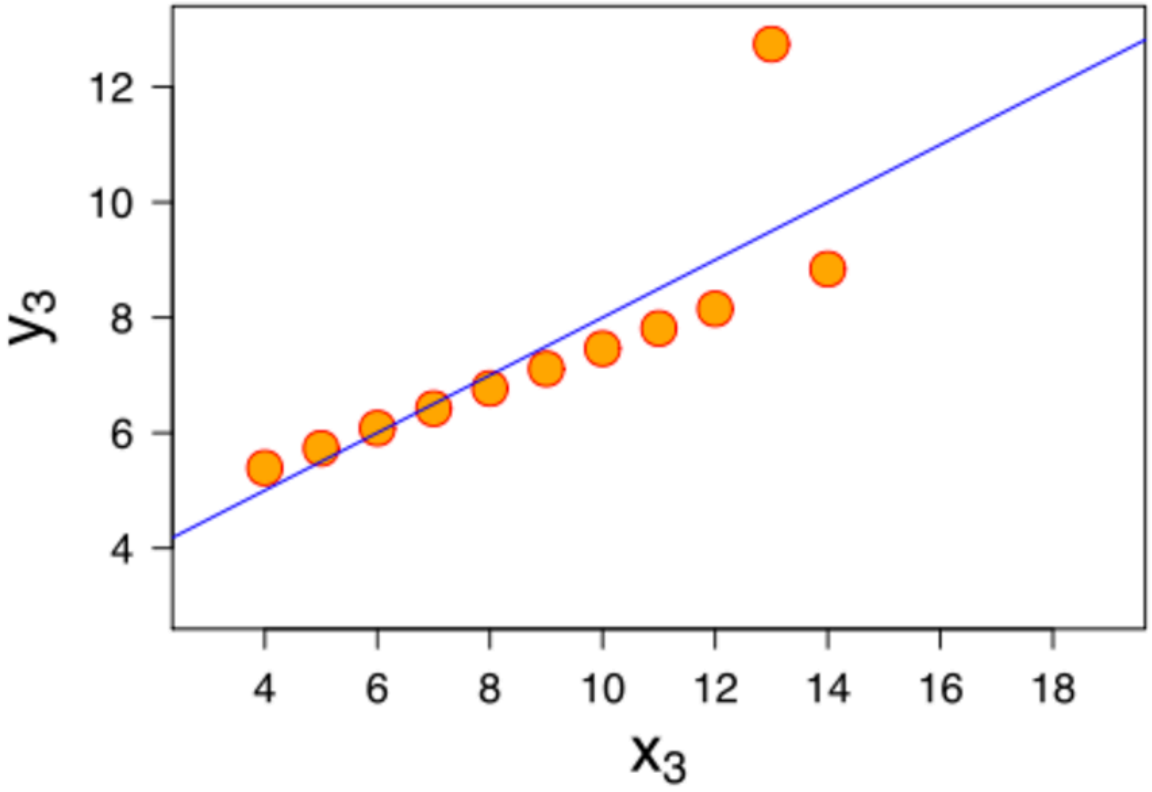
BIG DATA UNIVERSITY

#2 What can you tell me about this data?



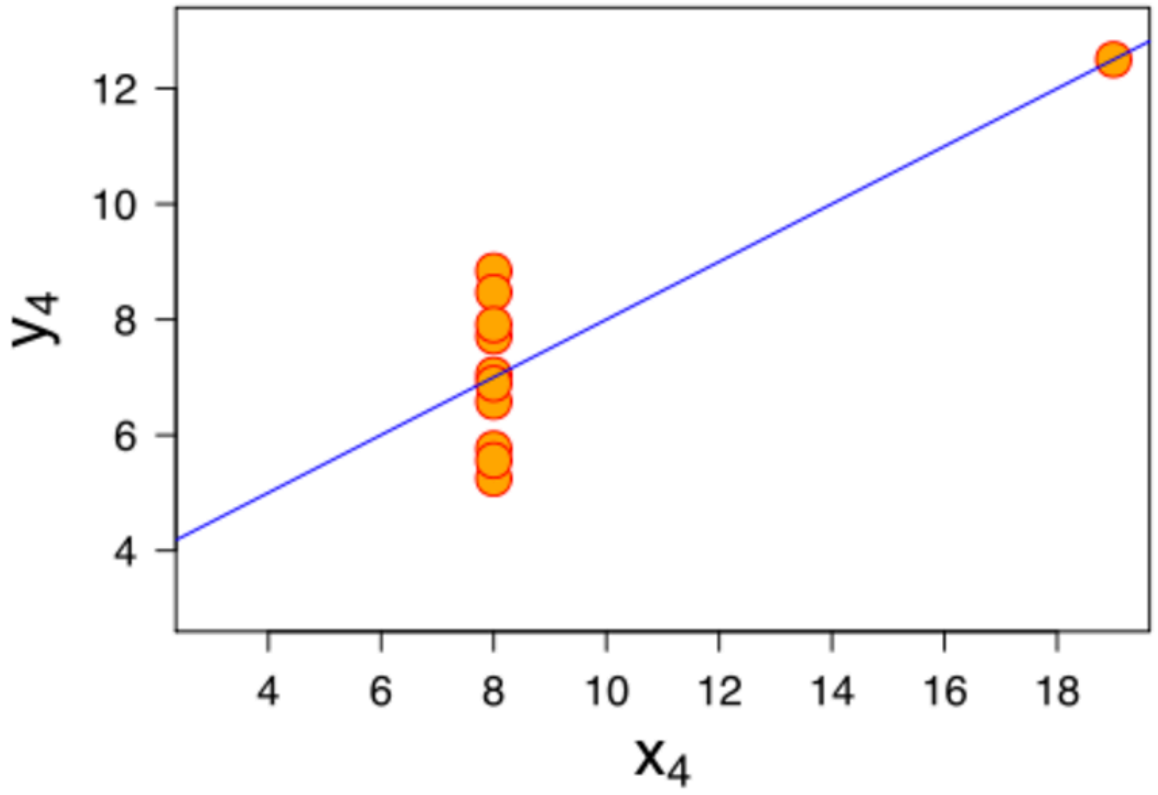
BIG DATA UNIVERSITY

#3 What can you tell me about this data?



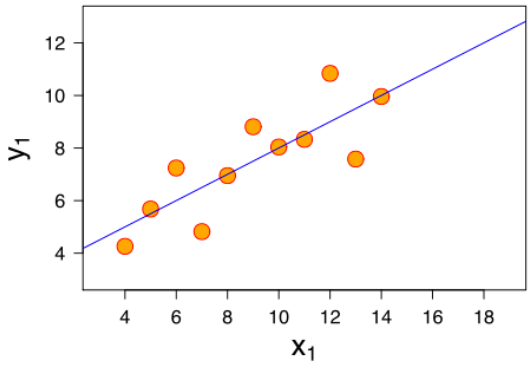
BIG DATA UNIVERSITY

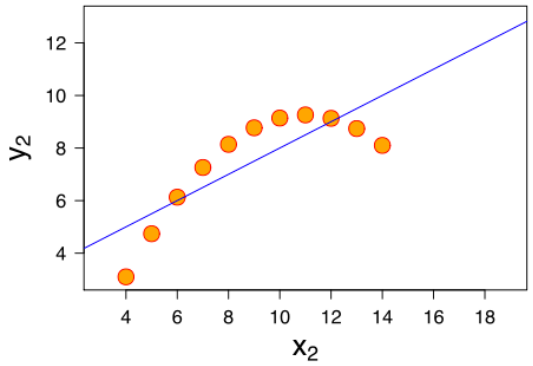
#4 What can you tell me about this data?

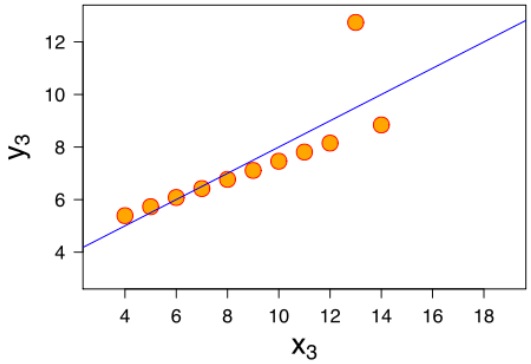


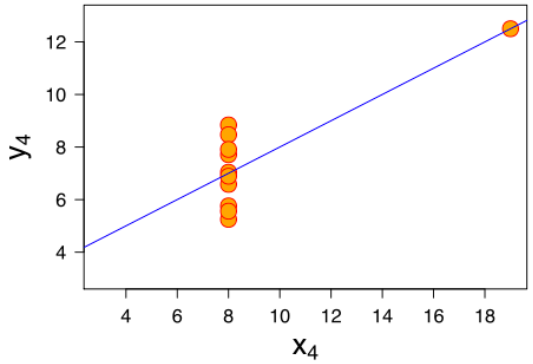
BIG DATA UNIVERSITY

Importance of Visualization









Same properties:

mean(x) = 9

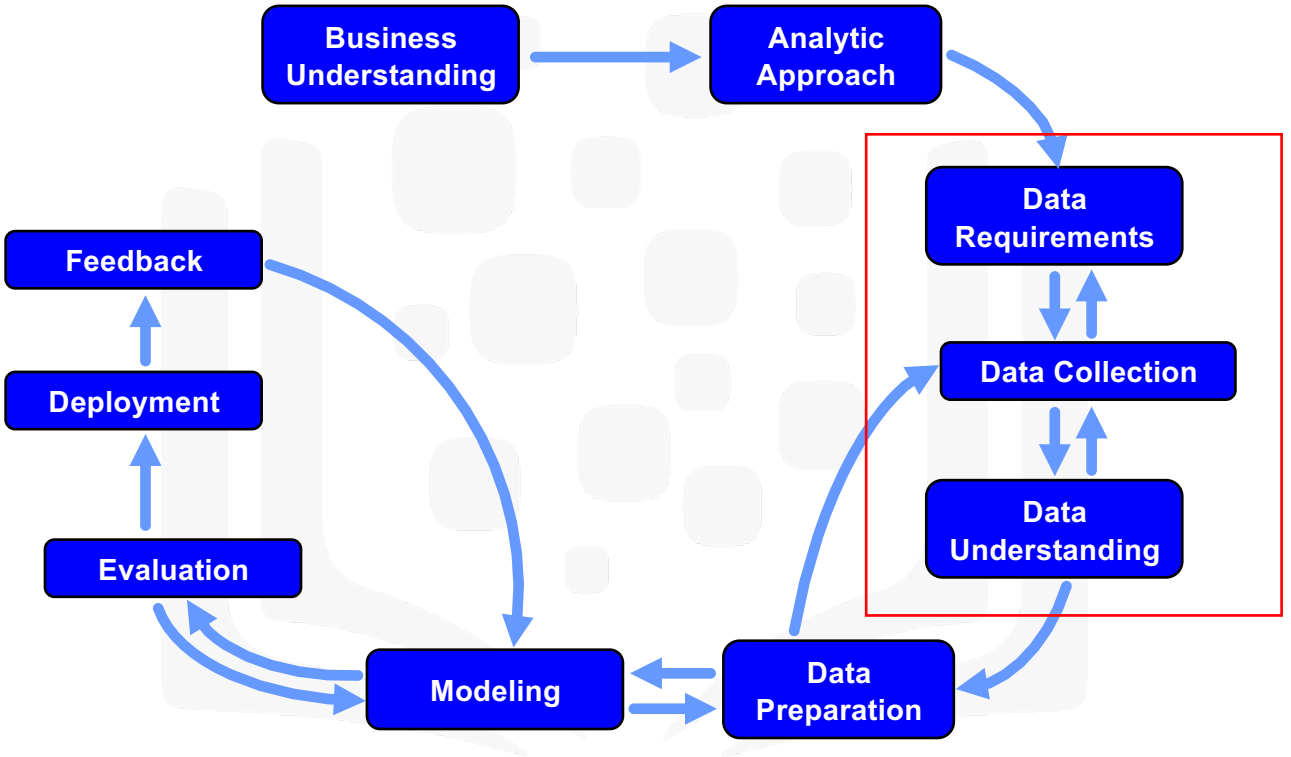
mean(y) = 7.5

y = 3.00 + 0.500x

corr(x,y) = 0.816

Anscombe's Quartet

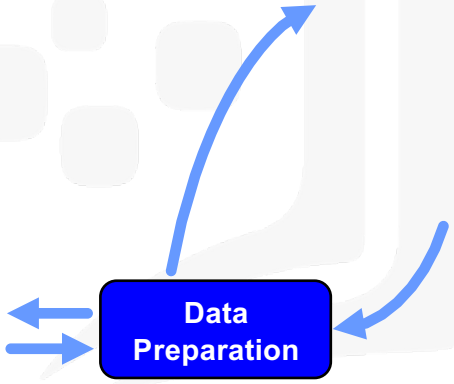
CRISP-DM Methodology diagram



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

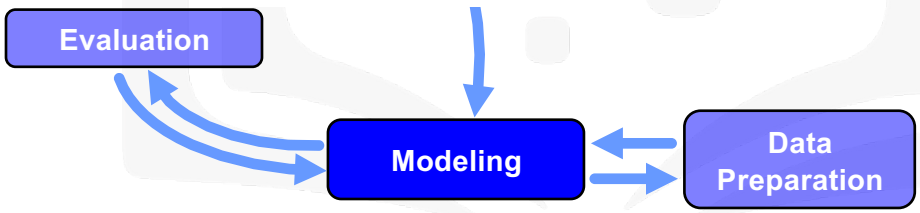
- Arguably the most time-consuming step
- "80% of the entire DS process is in data cleaning and preparation"



Modeling

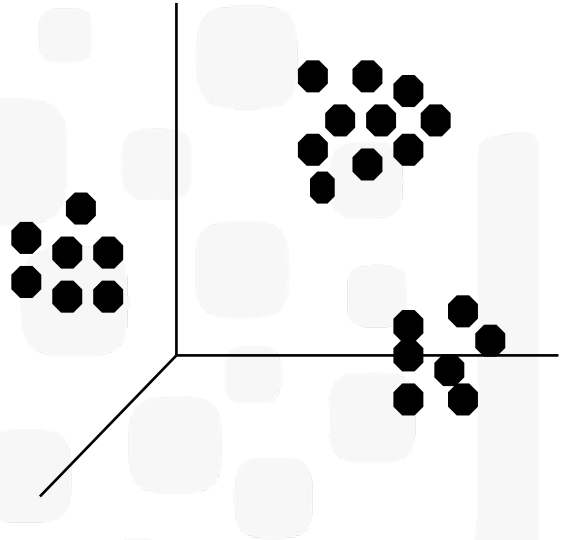
- **Modeling**:
  - Developing predictive or descriptive models
  - May try using multiple algorithms

- Highly iterative process



Example: Clustering

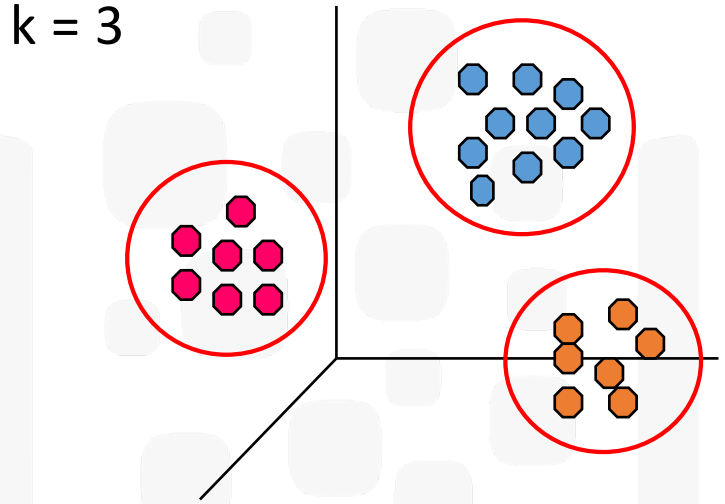
K-means Clustering



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

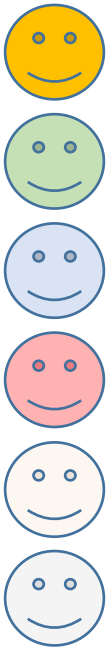
Example: Clustering

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**

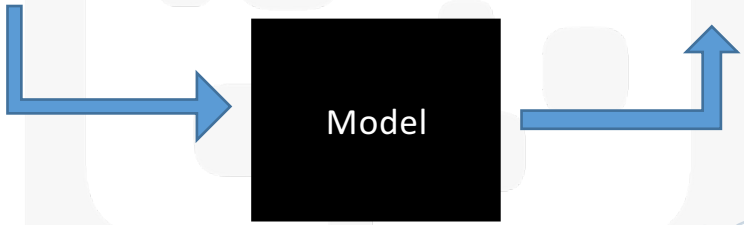
BIG DATA UNIVERSITY

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**: _____]

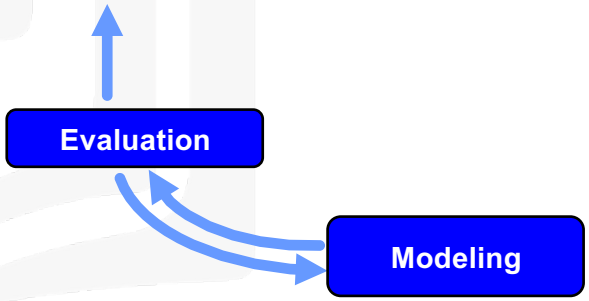
Disorder 1

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

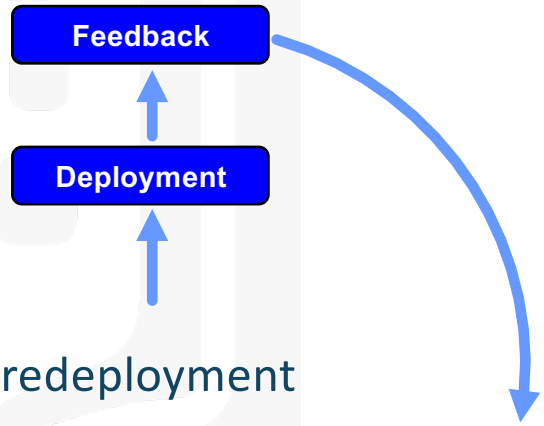


BIG DATA UNIVERSITY

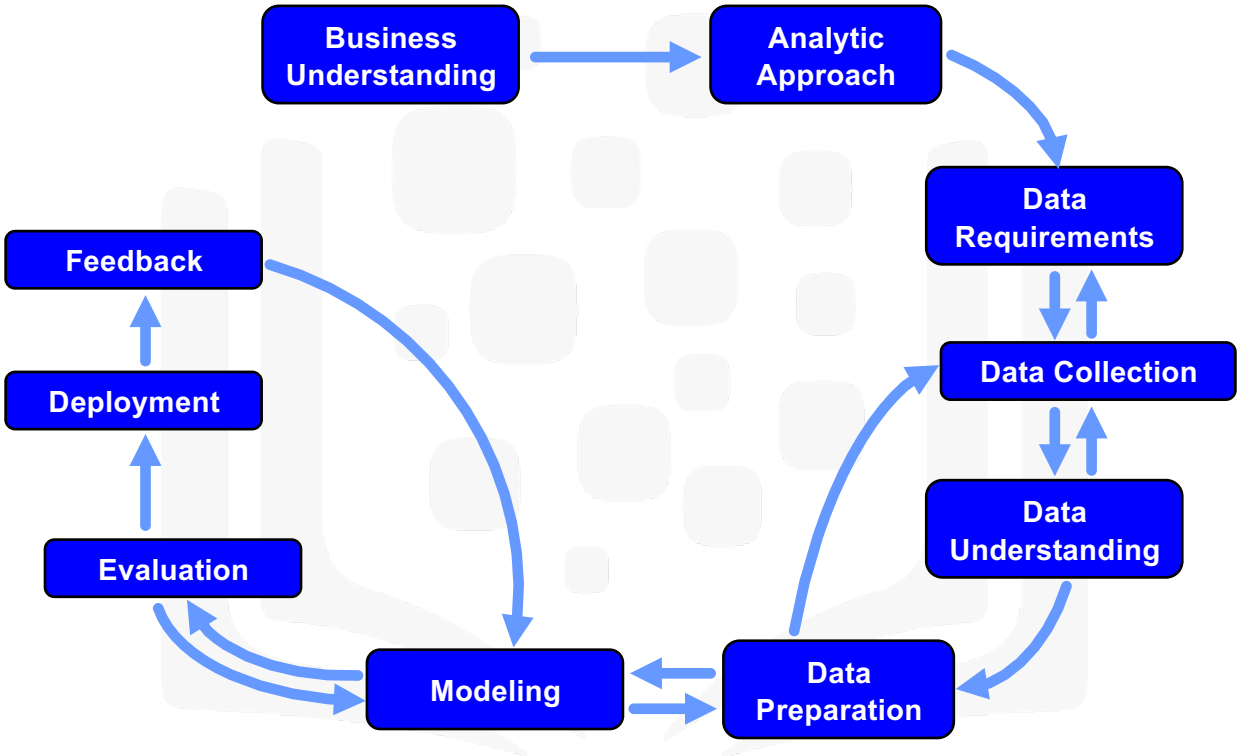
Deployment and feedback

- 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*

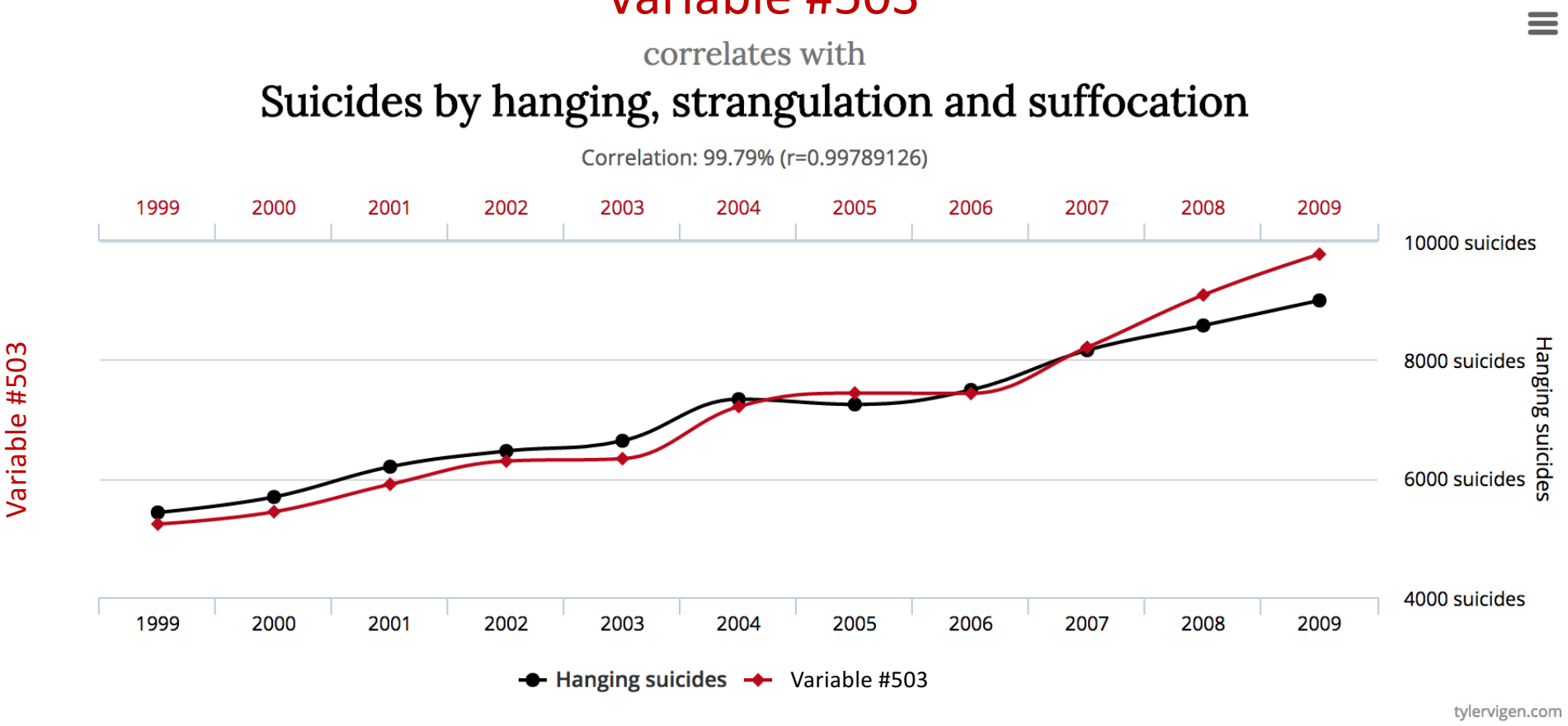
Suicides by hanging, strangulation and suffocation

Variable #503

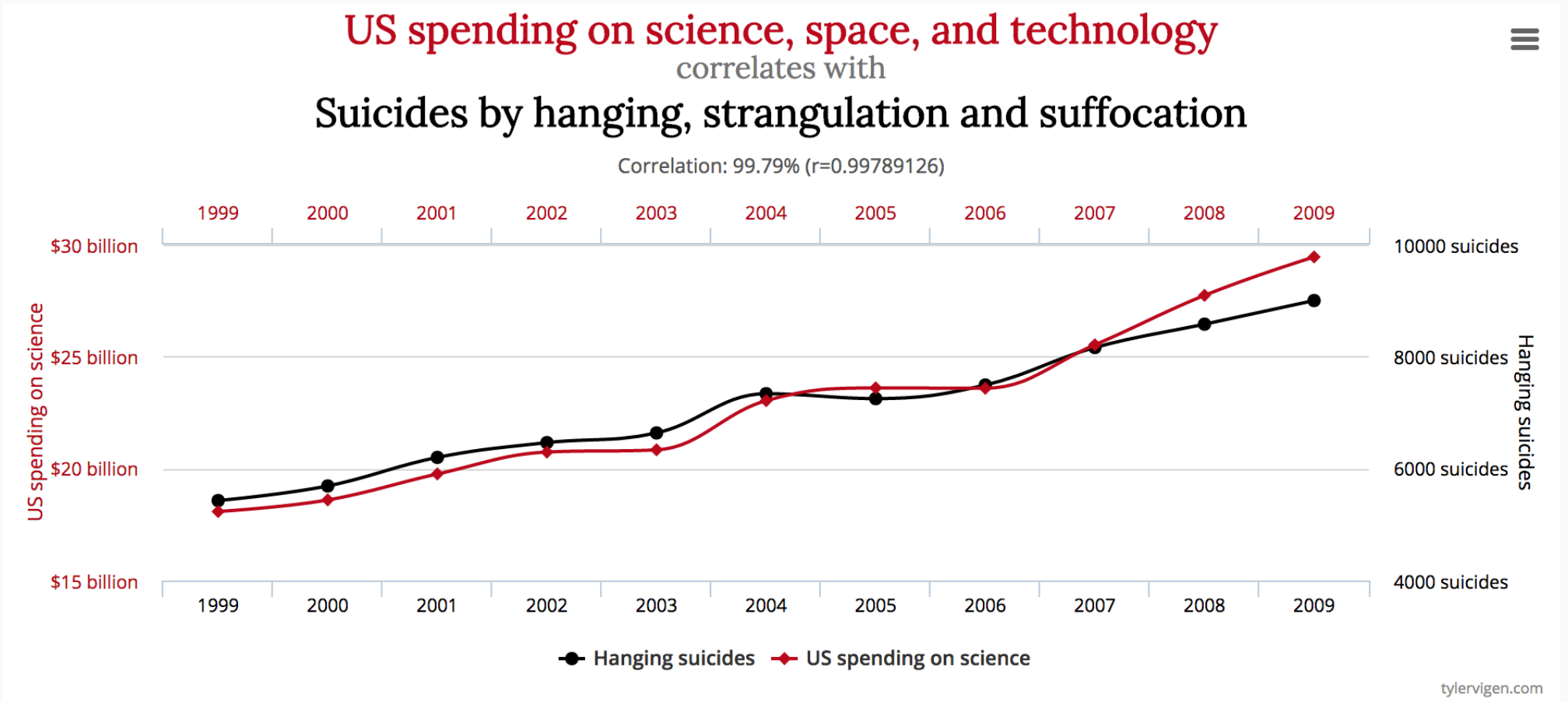
correlates with

Suicides by hanging, strangulation and suffocation

Variable #503



Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention



Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

Learning More About Data Science

Where can you learn more about data science?

BIG DATA UNIVERSITY

COURSES

Big Data 101
Fireside Analytics Inc. **BD0101EN**
Beginner

Data Science 101
Fireside Analytics Inc. **DS0101EN**
Beginner

Hadoop 101
BDU **BD0111EN**
Beginner

Python 101
BDU **PY0101EN**
Beginner

R 101
BDU **RP0101EN**
Beginner

Scala 101
LightBend **SC0101EN**
Beginner

Deep Learning 101
DeepLearning.TV **ML0115EN**
Intermediate

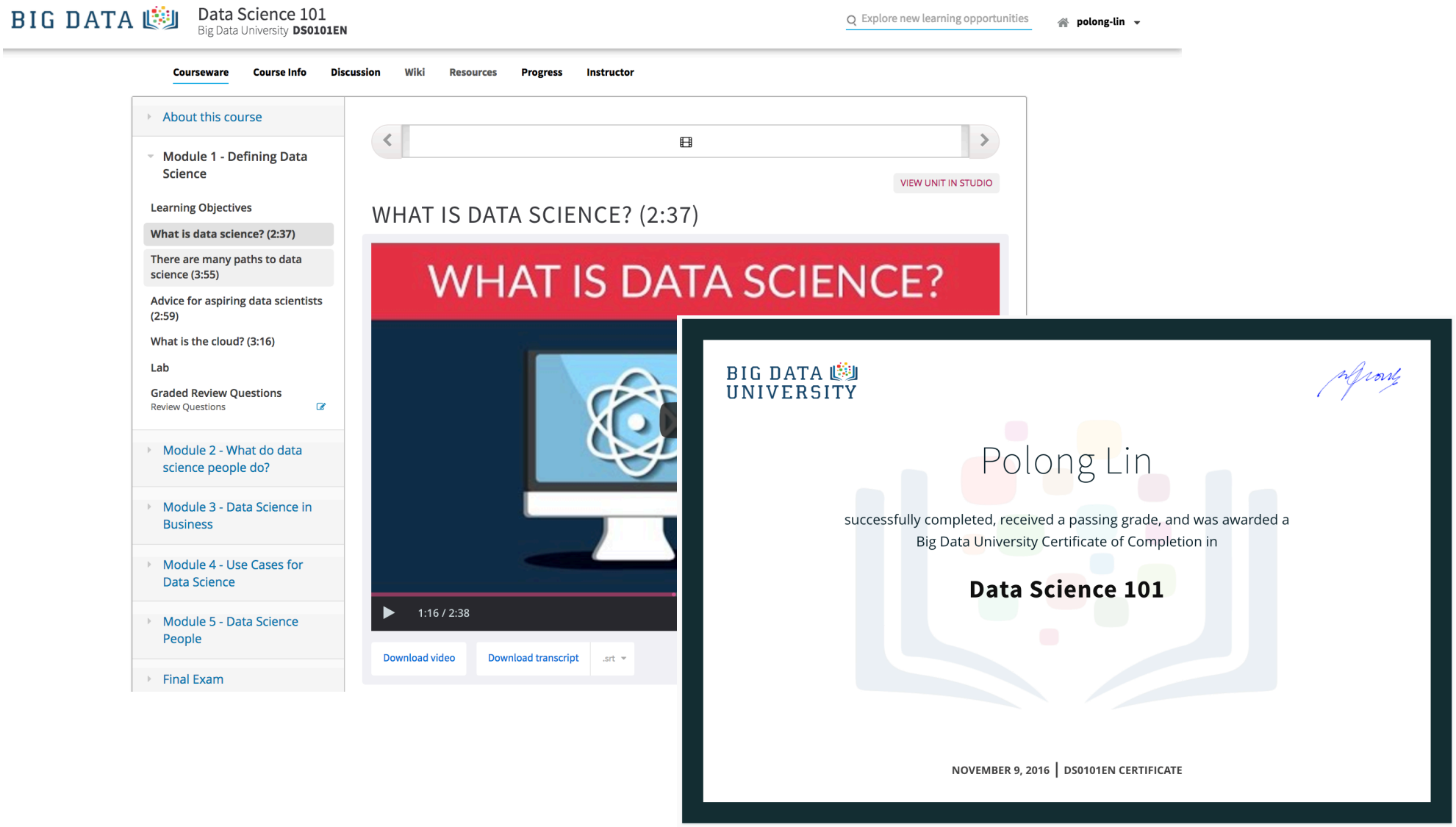
Deep Learning with TensorFlow
ML0120EN
Advanced

Text Analytics 101
BDU **TA0105**
Beginner

Watson Analytics 101
BDU **WA0101EN**
Beginner

Spark Fundamentals I
BDU **BD0211EN**
Beginner

Machine Learning with Python
ML0101EN
Beginner


BIG DATA
Data Science 101
Big Data University DS0101EN
Explore new learning opportunities
polong-lin
Courseware
Course Info
Discussion
Wiki
Resources
Progress
Instructor
About this course
Module 1 - Defining Data Science
Learning Objectives
What is data science? (2:37)
There are many paths to data science (3:55)
Advice for aspiring data scientists (2:59)
What is the cloud? (3:16)
Lab
Graded Review Questions
Review Questions
Module 2 - What do data science people do?
Module 3 - Data Science in Business
Module 4 - Use Cases for Data Science
Module 5 - Data Science People
Final Exam
VIEW UNIT IN STUDIO
WHAT IS DATA SCIENCE? (2:37)
WHAT IS DATA SCIENCE?
1:16 / 2:38
Download video
Download transcript
.srt
BIG DATA UNIVERSITY
Polong Lin
successfully completed, received a passing grade, and was awarded a Big Data University Certificate of Completion in
Data Science 101
NOVEMBER 9, 2016 | DS0101EN CERTIFICATE

BIG DATA UNIVERSITY

Data Science Methodology

In this course you'll learn how data science is practiced from start to end. This course is a must-know for aspiring data scientists as it provides the strategic framework to all data science problems.

Enroll

ABOUT THIS COURSE

Learn how data scientists think!

- Learn the major steps involved in tackling a data science problem.
- Learn the major steps involved in practicing data science, with interesting real-world examples at each step: from forming a concrete business or research problem, to collecting and analyzing data, to building a model, and understanding the feedback after model deployment.

COURSE SYLLABUS

- **Lesson 1: Introduction to Data Science Methodology**
- **Lesson 2: Business Understanding**
- **Lesson 3: Analytic Approach**
- **Lesson 4: Data Compilation**
- **Lesson 5: Data Preparation**
- **Lesson 6: Data Modeling**
- **Lesson 7: Model Evaluation**
- **Lesson 8: Model Deployment and Feedback**

GENERAL INFORMATION

- This course is free.

TELL YOUR FRIENDS

COURSE CODE:
DS0103EN

AUDIENCE:
Data Scientists, Data Engineers, Anyone with interest in Data Science

COURSE LEVEL:
Beginner

TIME TO COMPLETE:
3 Hours

BigDataUniversity.com

Free courses!

- Data Science
- Big Data
- Data Engineering

Earn badges!

Learn anytime!

For your organizations

- We can create **dedicated portals** for your employees to gain skills in data science