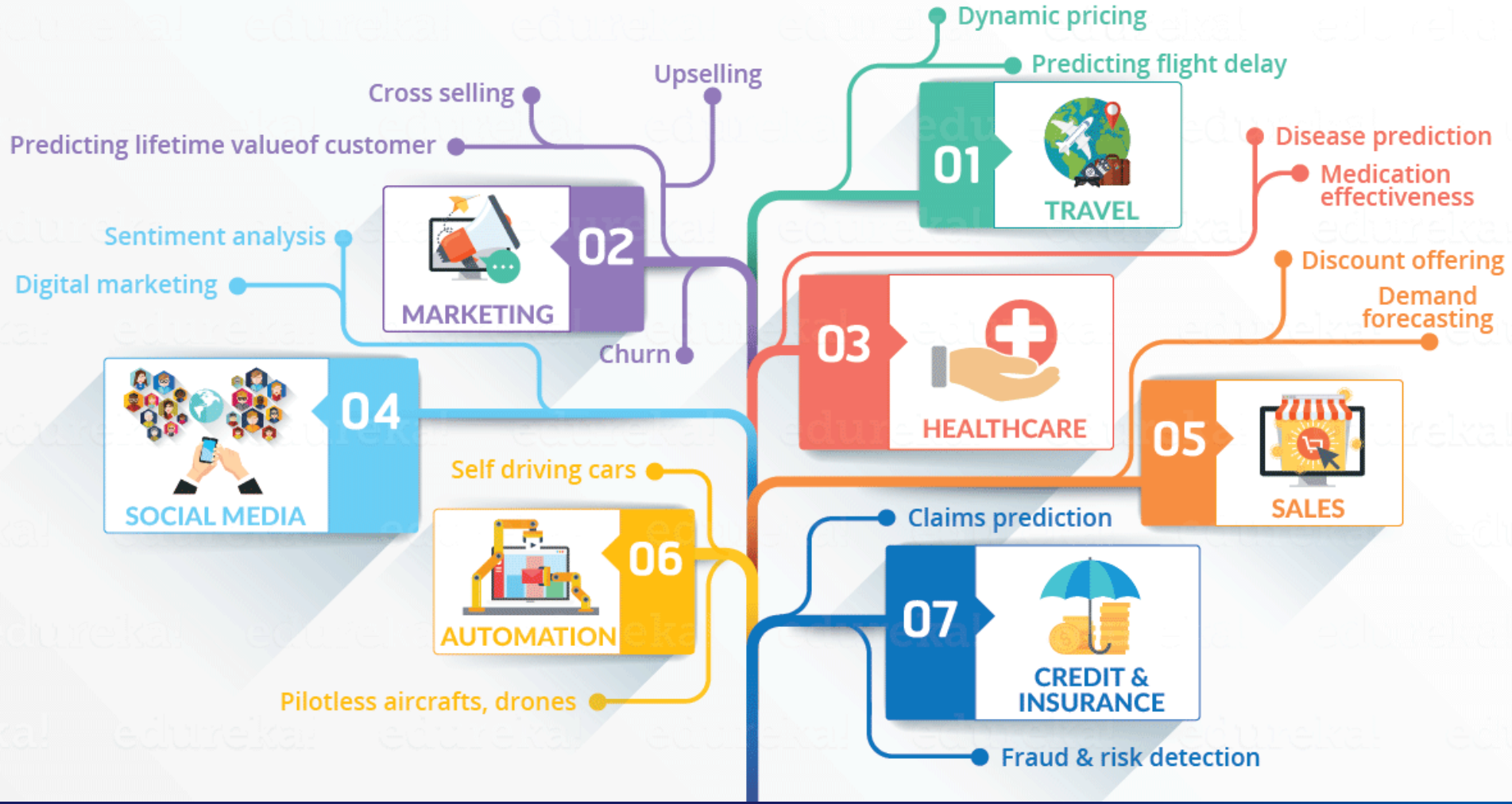
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# What is Data Science?

**Sourav SEN GUPTA**
Lecturer, SCSE, NTU


Dynamic pricing
Predicting flight delay
01
TRAVEL
Disease prediction
Medication effectiveness
Upselling
Cross selling
Predicting lifetime valueof customer
02
MARKETING
Churn
Sentiment analysis
Digital marketing
03
HEALTHCARE
Discount offering
Demand forecasting
05
SALES
04
SOCIAL MEDIA
Self driving cars
06
AUTOMATION
Pilotless aircrafts, drones
Claims prediction
07
CREDIT & INSURANCE
Fraud & risk detection

# Sample COLLECTION

# Practical MOTIVATION

How to Collect the relevant Data?

Does the data Match the Problem?
Does the data Represent Reality?

**How to effectively Sample real Data?**

What is the Real-Life Problem?

Can you relate the Problem to Data?
Would Data help you in Practice?

**How to Identify a Data Science case?**

# Data PREPARATION

# Problem FORMULATION

How to Prepare the relevant Data?

Is the data Clean enough to Analyze?
Is the data Structured for Analysis?

**How to prepare Raw Data for Analysis?**

What is the Data Science Problem?

How do you Formulate it using Data?
How do you Solve it using the Data?

**How to intelligently Construct a Problem?**

# Exploratory ANALYSIS

# Statistical DESCRIPTION

How to Explore the acquired Data?

How to effectively Mine the Data?
How to Compute the vital Statistics?

**How to gain basic Insight from Data?**

How to clearly Describe the Data?

How do you Summarize the Data?
Which vital Statistics are relevant?

**How to succinctly Represent the Data?**

# Analytic VISUALIZATION

# Pattern RECOGNITION

How to clearly Visualize the Data?

How to visually Represent Statistics?
How to highlight “Interesting” Traits?

**How to represent the Data for the Humans?**

How to Identify structure in Data?

Can you See the known Patterns?
Can you Discover unknown Traits?

**How to find Intrinsic insight from the Data?**

# Algorithmic OPTIMIZATION

# Machine LEARNING

How to form “Learning” algorithms?

How to Reduce Errors in Learning?

How to Generalize the Algorithms?

**How to optimally Learn from the Data?**

How to Learn from the Data?

Can you formulate the “Learning”?

Can you automate the “Learning”?

**How to efficiently Learn from the Data?**

Information
# PRESENTATION

Statistical
# INFERENCE

How to present Analysis Outcomes?

How to present Descriptive Analysis?
How to present Inferential Analysis?

**How to Communicate your Data Analysis?**

How to draw Conclusion from Data?

Can you Generalize the “Learning”?
Can you Estimate the Confidence?

**How to confidently Infer from the Data?**

Ethical
# CONSIDERATION

Intelligent
# DECISION

How to conform to Ethical Values?

Does the Analysis violate Legality?
Does the Decision violate Ethics?

## How to Responsibly work in Data Science?

How to take Decisions in Practice?

Can you Decide based on the Data?
Can you Optimize the Outcomes?

## How to Solve a real life problem by Data?


Data Science Pipeline

**Raw Data to Actionable Intelligence**

Real-Life Problems translated in Data

Descriptive and Inferential Analytics

Effective Communication and Decision

**How to optimally solve a problem using data?**

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Data Science Problems

**Sourav SEN GUPTA**
Lecturer, SCSE, NTU


Data Science

# Common Problems

**Five Primary Questions**

How much? How many?

Is it type A or type B?

How is this organized?

Is it a weird behavior?

What should be done next?

https://www.youtube.com/watch?v=0XyV91VYrDs

Data Science

# Common Problems

**Prediction : Numeric**

## How much?
## How many?

What is the expected Sales of the next game of this game franchise?

Is it profitable to make the sequel?

https://www.youtube.com/watch?v=0XyV91VYrDs

Data Science

# Common Problems

**Prediction : Classes**

## Is it type A or type B?

What is the chance that a student will get into NTU in AY2019-2020?

Will an application be successful?

https://www.youtube.com/watch?v=0XyV91VYrDs


Data Science

# Common Problems

**Detection : Structure**

## How is this organized?

Is there any structure apparent within the FairPrice customers?

Which customer group to target?

https://www.youtube.com/watch?v=0XyV91VYrDs

Data Science

# Common Problems

**Detection : Anomaly**

## Is it weird behavior?

Is this Boeing engine behaving in unusual fashion during the flight?

Is the engine still safe to operate?

https://www.youtube.com/watch?v=0XyV91VYrDs


Data Science

# Common Problems

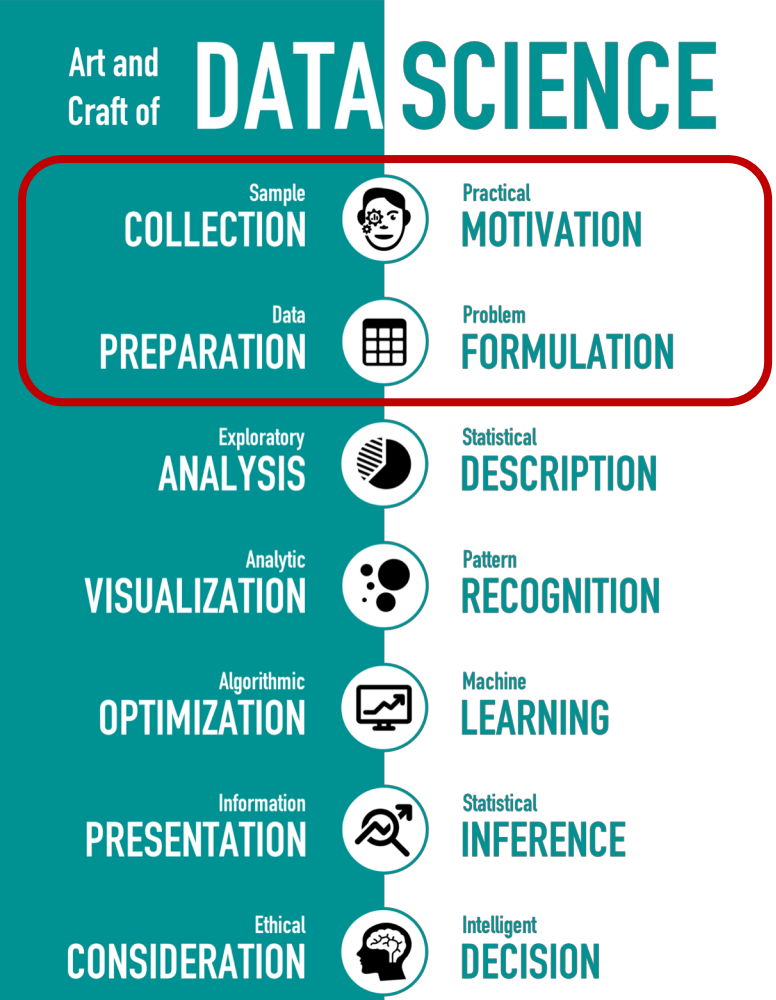
**Decision : Action**

## What should be done next?

Should brake at the Yellow Light or should the car accelerate instead?

Which action will be rewarded?

https://www.youtube.com/watch?v=0XyV91VYrDs




Data Science Pipeline

**Problem Formulation**

What is the practical Motivation?

How to acquire the relevant Data?

How to prepare the acquired Data?

**How to intelligently formulate a problem?**

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Data Science Solutions

**Sourav SEN GUPTA**
Lecturer, SCSE, NTU


Data Science

# Common Problems

**Five Primary Questions**

How much? How many?

Is it type A or type B?

How is this organized?

Is it a weird behavior?

What should be done next?

https://www.youtube.com/watch?v=0XyV91VYrDs

Data Science

# Common Problems

**Prediction : Numeric**

## How much?
## How many?

What is the expected Sales of the next game of this game franchise?

Is it profitable to make the sequel?

Data Science

# Common Solutions

**Prediction : Numeric**

## Regression

Try to find the relationship of Sales of the games with other Variables, like Graphics Quality, Genre, etc.

Model : Sales = $f$ ( Variables )

Data Science

# Common Solutions

**Prediction : Numeric**

## Regression

Model : Sales = $f$ ( Variables )

Linear Regression Models
Tree Models for Regression
Neural Network for Regression

Data Science

# Common Problems

**Prediction : Classes**

## Is it type A or type B?

What is the chance that a student will get into NTU in AY2019-2020?

Will an application be successful?

Data Science

# Common Solutions

**Prediction : Classes**

## Classification

Try to find the Probability of getting admitted to NTU in terms of other Variables, like Scores, Gender, etc.

Model : $\mathcal{P}$ ( Admit ) = $f$ ( Variables )

Data Science

# Common Solutions

**Prediction : Classes**

## Classification

Model : $\mathcal{P}$ ( Admit ) = $f$ ( Variables )

Logistic Regression Model
Tree Models for Classification
Neural Network for Classification


Data Science

# Common Problems

**Detection : Structure**

## How is this organized?

Is there any structure apparent within the FairPrice customers?

Which customer group to target?


Data Science

# Common Solutions

**Detection : Structure**

## Clustering

Try to find Groups of Data Points that are close together but are far from the other Groups of Points.

Close–Far depends on “Distance”


Data Science

# Common Solutions

**Detection : Structure**

## Clustering

Close–Far depends on “Distance”

Distance: Euclidean, Jaccard etc.

k-Means Algorithm for Clustering

Hierarchical Model for Clustering

Data Science

# Common Problems

**Detection : Anomaly**

## Is it weird behavior?

Is this Boeing engine behaving in unusual fashion during the flight?

Is the engine still safe to operate?

Data Science

# Common Solutions

**Detection : Anomaly**

## Anomaly Detection

Try to find Deviations of the Data compared to the Regular Pattern observed through the data model.

Deviations depend on the Model

Data Science

# Common Solutions

**Detection : Anomaly**

# Anomaly Detection

Deviations depend on the Model

Cluster-Analysis based Detection
Nearest Neighbor Detection Model
Support Vector based Detection


Data Science

# Common Problems

**Decision : Action**

## What should be done next?

Should brake at the Yellow Light or should the car accelerate instead?

Which action will be rewarded?


Data Science

# Common Solutions

**Decision : Action**

## Adaptive Learning

Try to model a Profit/Loss Function depending at any given state, and try to Maximize/Minimize the same.

Optimize $f$ ( State, Variables )


Data Science

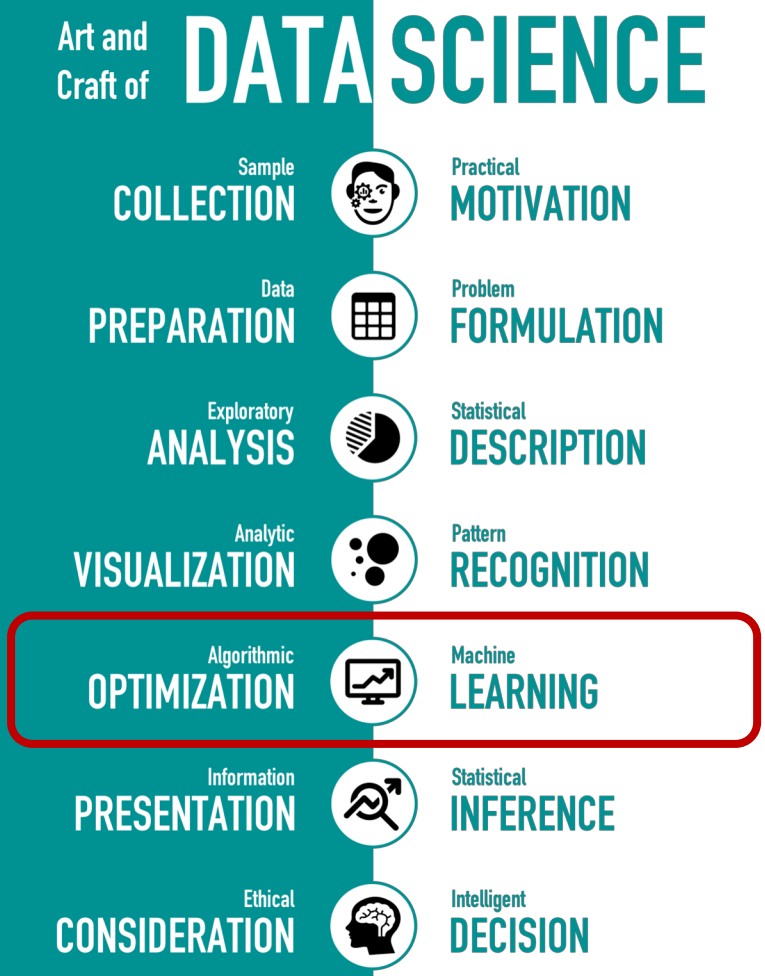
# Common Solutions

**Decision : Action**

## Adaptive Learning

Optimize $f$( State, Variables )

Reinforcement Learning Approach
Monte-Carlo, State-Action-Reward
Q-Learning, Deep Reinforcement




Data Science Pipeline

**Algorithmic Solution**

How to prepare the acquired Data?

How to create a model for the Data?

How to choose the optimal Model?

**How to intelligently infer from the Model?**

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Structured Data in Practice

**Sourav SEN GUPTA**
Lecturer, SCSE, NTU


Data Science

# Common Data Types

## Two Primary Data Types

### Structured Data

Highly Organized, Easy to Analyze

Numeric/Factor, Time Series, Network

### Unstructured Data

Highly Unorganized and Contextual

Text, Image, Voice, Videos

| TV | Radio | Newspaper | Sales |
|---|---|---|---|
| 230.1 | 37.8 | 69.2 | 22.1 |
| 44.5 | 39.3 | 45.1 | 10.4 |
| 17.2 | 45.9 | 69.3 | 9.3 |
| 151.5 | 41.3 | 58.5 | 18.5 |
| 180.8 | 10.8 | 58.4 | 12.9 |
| 8.7 | 48.9 | 75.0 | 7.2 |
| 57.5 | 32.8 | 23.5 | 11.8 |
| 8.6 | 2.1 | 1.0 | 4.8 |
| 199.8 | 2.6 | 21.2 | 10.6 |
| 66.1 | 5.8 | 24.2 | 8.6 |
| 214.7 | 24.0 | 4.0 | 17.4 |
| 23.8 | 35.1 | 65.9 | 9.2 |

Data Science

# Structured Data

## Numeric Data

Highly Organized Data
Clearly Defined Variables
Easy to Mine and Analyze
Numeric Continuous Variables

Example Source

- Spreadsheets (Excel, CSV)
- Standard SQL Databases
- Sensors and Devices

Advertising dataset from ISL by James et al.

| Safety | Doors | Seats | Condition |
|---|---|---|---|
| high | 4 | 2 | unacc |
| med | 5more | more | good |
| high | 5more | more | vgood |
| high | 2 | 2 | unacc |
| high | 2 | 2 | unacc |
| low | 4 | more | acc |
| med | 5more | 2 | unacc |
| high | 4 | 4 | acc |
| med | 2 | more | acc |
| high | 4 | 2 | unacc |
| low | 3 | 4 | unacc |
| high | 3 | 4 | unacc |

Data Science

# Structured Data

## Categorical Data

Highly Organized Data
Clearly Defined Variables
Easy to Mine and Analyze
Factor/Level/Class Variables

Example Source

- Spreadsheets (Excel, CSV)
- Standard SQL Databases
- Sensors and Devices

Car Evaluation dataset from ISL by James et al.

| Price | Doors | Seats | Condition |
|---|---|---|---|
| 230.1 | 4 | 2 | unacc |
| 44.5 | 5more | more | good |
| 17.2 | 5more | more | vgood |
| 151.5 | 2 | 2 | unacc |
| 180.8 | 2 | 2 | unacc |
| 8.7 | 4 | more | acc |
| 57.5 | 5more | 2 | unacc |
| 8.6 | 4 | 4 | acc |
| 199.8 | 2 | more | acc |
| 66.1 | 4 | 2 | unacc |
| 214.7 | 3 | 4 | unacc |
| 23.8 | 3 | 4 | unacc |

Data Science

# Structured Data

## Mixed Data

Highly Organized Data
Clearly Defined Variables
Easy to Mine and Analyze
Numeric and Categorical

Example Source

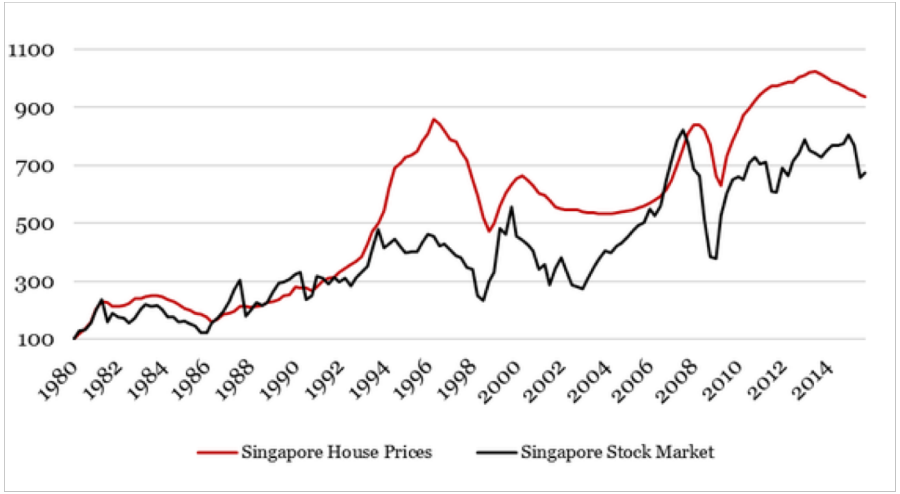
- Spreadsheets (Excel, CSV)
- Standard SQL Databases
- Sensors and Devices

Car Evaluation dataset from ISL by James et al.

Data Science

# Structured Data

## Time Series Data




Highly Organized Data
Clearly Defined Time Axis
Easy to Mine and Analyze
Numeric with Timestamps

Example Source

- Stock and Equity Markets
- Weather Data over Time
- Prices and Promotions

House Prices vs. Stock Data from Bloomberg

Data Science

# Structured Data

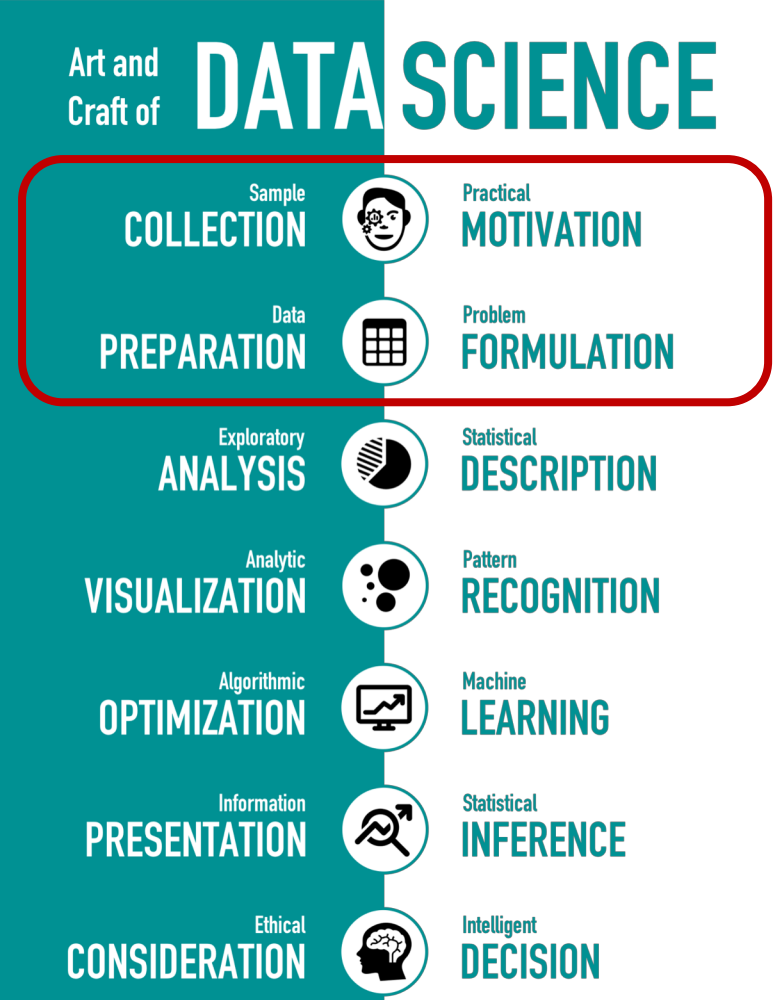
## Network Data


Highly Organized Nodes
Clearly Defined Links/Edges
Easy to Mine and Analyze
Nodes and Connections

Example Source

- Social Networks and Web
- Transport Networks (MRT)
- Financial Transactions

Singapore MRT Network from MRT Website




Data Science Pipeline

**Data Acquisition and Preparation**

What is the type of acquired Data?

How to prepare the acquired Data?

How to analyze the acquired Data?

**How to intelligently handle relevant Data?**

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# Unstructured Data in Practice

**Sourav SEN GUPTA**
Lecturer, SCSE, NTU


Data Science

# Common Data Types

## Two Primary Data Types

### Structured Data

Highly Organized, Easy to Analyze

Numeric/Factor, Time Series, Network

### Unstructured Data

Highly Unorganized and Contextual

Text, Image, Voice, Videos

Data Science

# Unstructured Data

## Text Data

#Cashless payments could soon be a way of life for students in @NTUsg, from the way they #pay to the way they attend classes http://tmsk.sg/aM

Eyeing a Smart Campus: Here's #NTUsg's new leadership team at their first town hall session with the NTU community. They shared how smart technologies will be used at NTU to improve learning and living experiences.

#NTUsg partners Volvo to develop #autonomous #electricbuses in Singapore. NTU is the first university in the world to work with Volvo on self-driving technology for buses. #NTUsgResearch

Highly Unorganized Data
Non-Obvious Variables
Highly Context-Sensitive
Words, Phrases, Emoticons

Example Source

- Social Networks and Web
- Text Messages / WhatsApp
- Books, Wikis, Documents

Twitter Feeds from NTU Singapore

Data Science

# Unstructured Data

## Image Data

Highly Unorganized Data
Non-Obvious Variables
Highly Context-Sensitive
Pixels and Objects

Example Source

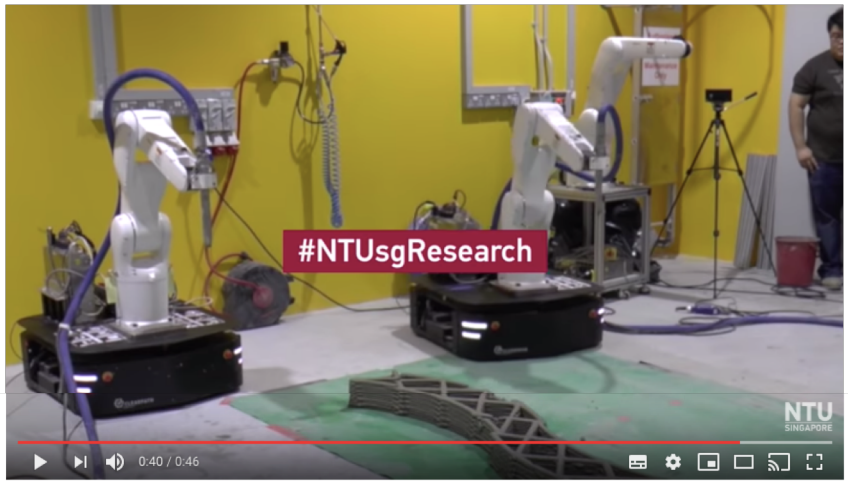
- Social Networks and Web
- Mobile Phone Cameras
- Blogs, Wikis, Documents

Looks like good food! – from the Canteen

Data Science

# Unstructured Data

## Video Data




Highly Unorganized Data
Non-Obvious Variables
Highly Context-Sensitive
Images, Frames, Objects

Example Source

- YouTube and Social Media
- Video Messages and Calls
- Mobile Phone Cameras

Video on 3D Printing from NTU Singapore

Data Science

# Unstructured Data

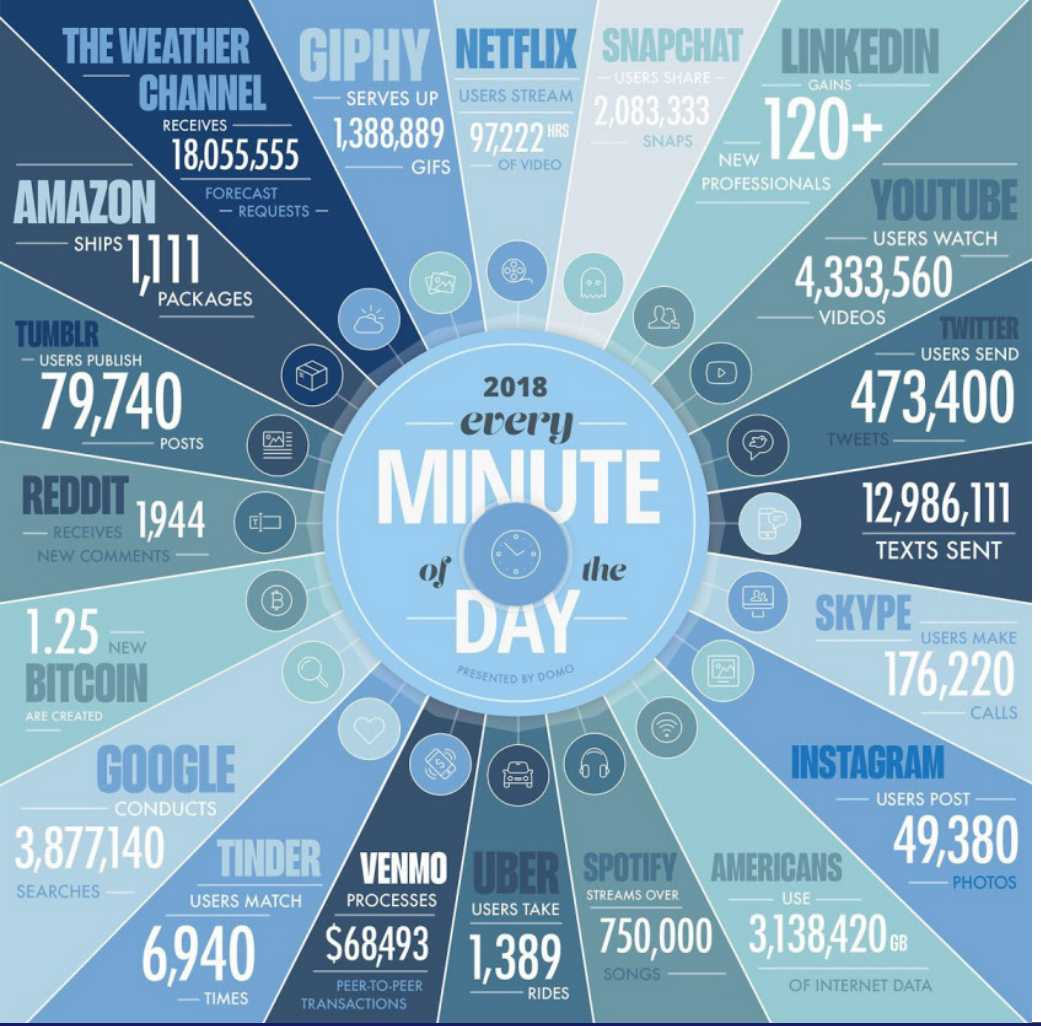
## Voice Data


Highly Unorganized Data
Non-Obvious Variables
Highly Context-Sensitive
Voice Signals and Waves

Example Source

- Songs and Social Media
- Microphones and Cameras
- Recordings, Announcements

Siri on Apple Devices

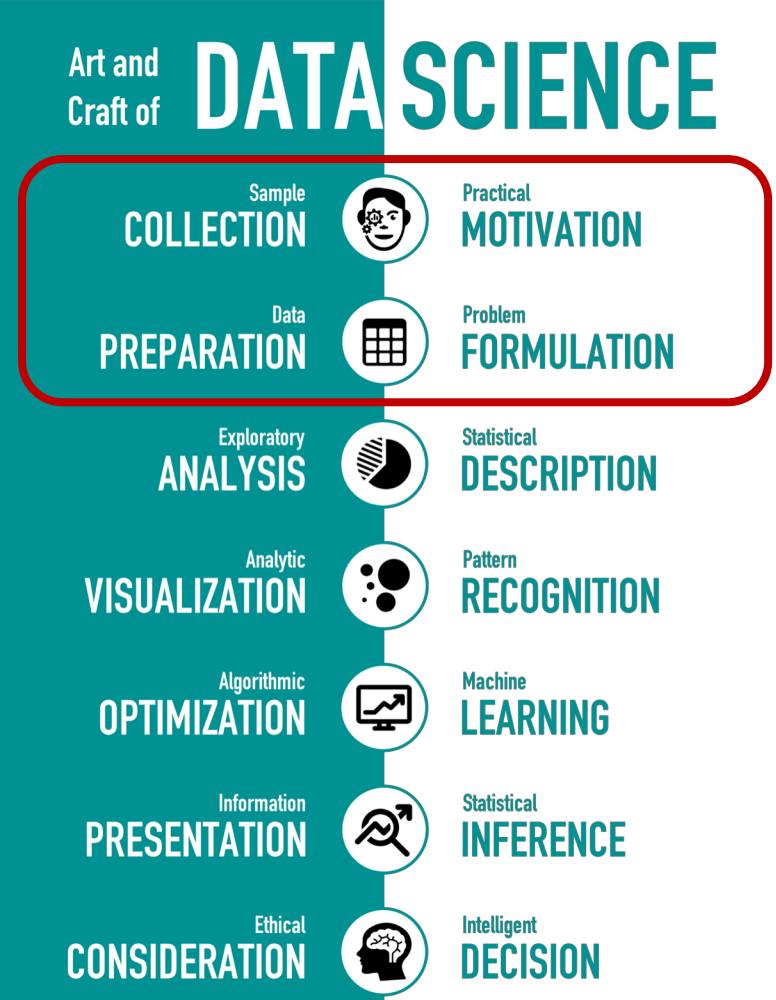



Data Science

# The Rise of Data

| | |
|---|---|
| World | **7.6 Billion** |
| Internet | **4.2 Billion** |
| YouTube | **1.6 Billion** |
| Facebook | **2.2 Billion** |
| Gmail | **1.2 Billion** |
| Instagram | **800 Million** |
| Twitter | **330 Million** |

The figures state active users per month
https://www.domo.com/learn/data-never-sleeps-6




Data Science Pipeline

**Data Acquisition and Preparation**

What is the type of acquired Data?

How to prepare the acquired Data?

How to analyze the acquired Data?

**How to intelligently handle relevant Data?**

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

# The Dataset

**Sourav SEN GUPTA**
Lecturer, SCSE, NTU


Data Science

# The Pokemon Dataset

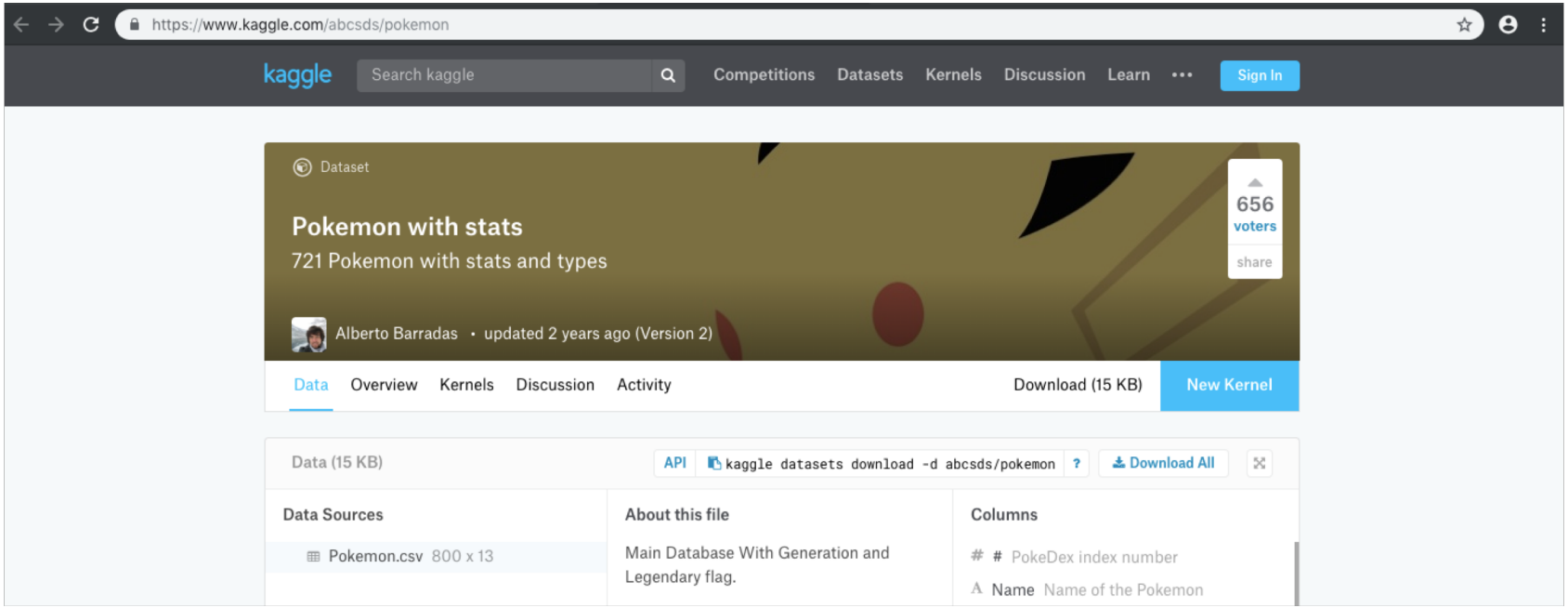
**Structured Data : Mixed**

Highly Organized Data
Clearly Defined Variables
Easy to Mine and Analyze
Numeric and Categorical

Source : Kaggle Datasets
**Pokemon with stats** by Alberto Barradas
https://www.kaggle.com/abcsds/pokemon

Data Science

# The Pokemon Dataset




Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

Data Science

# The Pokemon Dataset

| # | Name | Type 1 | Type 2 | Total | HP | Attack | Defense | Sp. Atk | Sp. Def | Speed | Generation | Legendary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 430 | Honchkrow | Dark | Flying | 505 | 100 | 125 | 52 | 105 | 52 | 71 | 4 | False |
| 338 | Solrock | Rock | Psychic | 440 | 70 | 95 | 85 | 55 | 65 | 70 | 3 | False |
| 32 | Nidoran♂ | Poison | NaN | 273 | 46 | 57 | 40 | 40 | 40 | 50 | 1 | False |
| 442 | Spiritomb | Ghost | Dark | 485 | 50 | 92 | 108 | 92 | 108 | 35 | 4 | False |
| 480 | Uxie | Psychic | NaN | 580 | 75 | 75 | 130 | 75 | 130 | 95 | 4 | True |
| 536 | Palpitoad | Water | Ground | 384 | 75 | 65 | 55 | 65 | 55 | 69 | 5 | False |
| 360 | Wynaut | Psychic | NaN | 260 | 95 | 23 | 48 | 23 | 48 | 23 | 3 | False |
| 478 | Froslass | Ice | Ghost | 480 | 70 | 80 | 70 | 80 | 70 | 110 | 4 | False |
| 76 | Golem | Rock | Ground | 495 | 80 | 120 | 130 | 55 | 65 | 45 | 1 | False |
| 177 | Natu | Psychic | Flying | 320 | 40 | 50 | 45 | 70 | 45 | 70 | 2 | False |

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

Data Science

# The Pokemon Dataset

| | |
|---|---|
| **#** | ID for each Pokemon (runs from 1 to 721) |
| **Name** | Name of each Pokemon |
| **Type 1** | Each Pokemon has a basic Type, this determines weakness/resistance to attacks |
| **Type 2** | Some Pokemons are dual type and have a Type 2 value (set to NaN otherwise) |
| **Total** | Sum of all stats of a Pokemon, a general guide to how strong a Pokemon is |
| **HP** | Hit Points, defines how much damage a Pokemon can withstand before fainting |
| **Attack** | The base modifier for normal attacks by the Pokemon (e.g., scratch, punch etc.) |
| **Defense** | The base damage resistance of the Pokemon against normal attacks |
| **SP Atk** | Special Attack, the base modifier for special attacks (e.g. fire blast, bubble beam) |
| **SP Def** | Special Defense, the base damage resistance against special attacks |
| **Speed** | Determines which Pokemon attacks first each round |
| **Generation** | Each Pokemon belongs to a certain Generation |
| **Legendary** | Legendary Pokemons are powerful, rare, and hard to catch |

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

Data Science

# The Pokemon Dataset

| | |
|---|---|
| **#** | Index (Numeric) |
| **Name** | String (neither Numeric nor Factor) |
| **Type 1** | Categorical (one of 18 Types) |
| **Type 2** | Categorical (one of 18 Types, or NaN) |
| **Total** | Numeric |
| **HP** | Numeric |
| **Attack** | Numeric |
| **Defense** | Numeric |
| **SP Atk** | Numeric |
| **SP Def** | Numeric |
| **Speed** | Numeric |
| **Generation** | Categorical 1 to 6) |
| **Legendary** | Categorical (True or False) |

```
#              int64
Name          object
Type 1        object
Type 2        object
Total          int64
HP             int64
Attack         int64
Defense        int64
Sp. Atk        int64
Sp. Def        int64
Speed          int64
Generation     int64
Legendary       bool
```

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon


Data Science Pipeline

**Exploratory Analysis**

What are the Variables in the Data?

How to characterize the Variables?

How to find relation between them?

**How to intelligently explore acquired Data?**