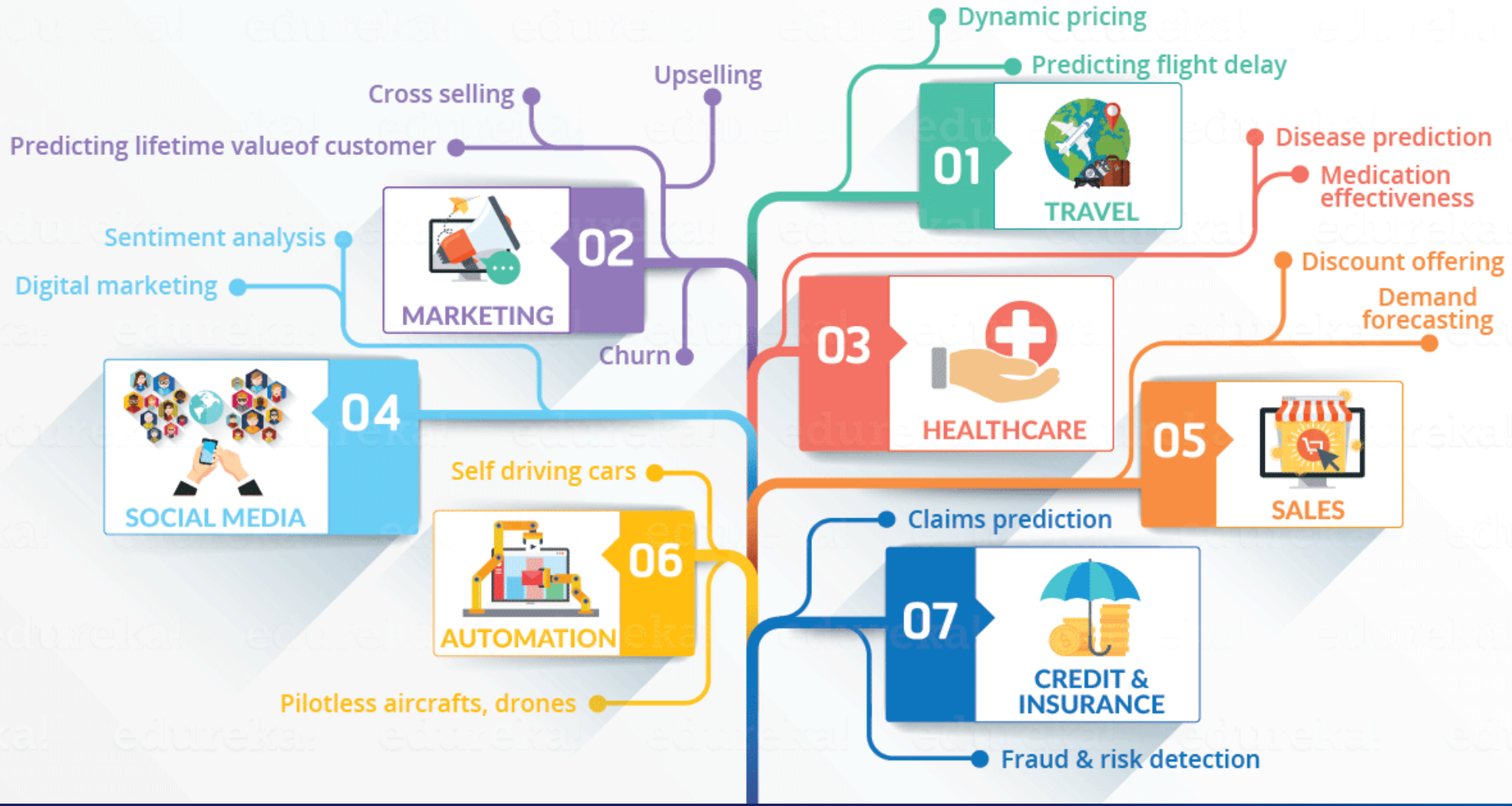
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Data Science

What is everyone talking about?

Sourav SEN GUPTA, Lecturer

School of Computer Science and Engg.
Nanyang Technological University


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

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?

Data Science

Structured Data

Numeric Data

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

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

Example Source

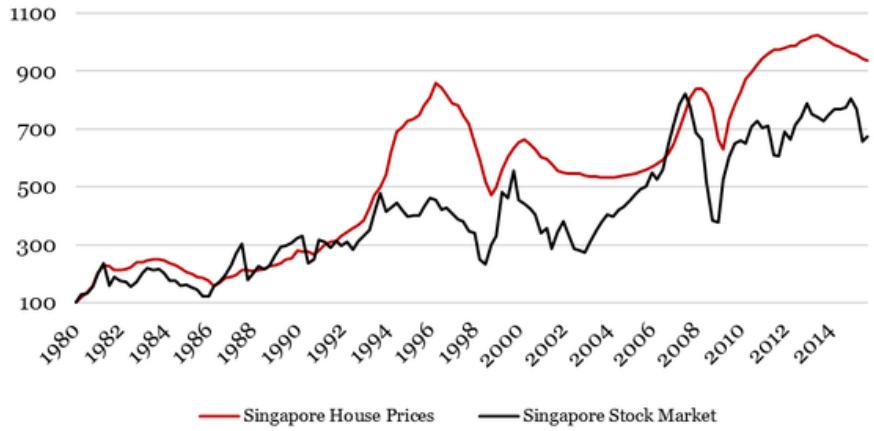
- Spreadsheets (Excel, CSV)
- Standard SQL Databases
- Sensors and Devices

Advertising dataset from ISL by James et al.

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

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

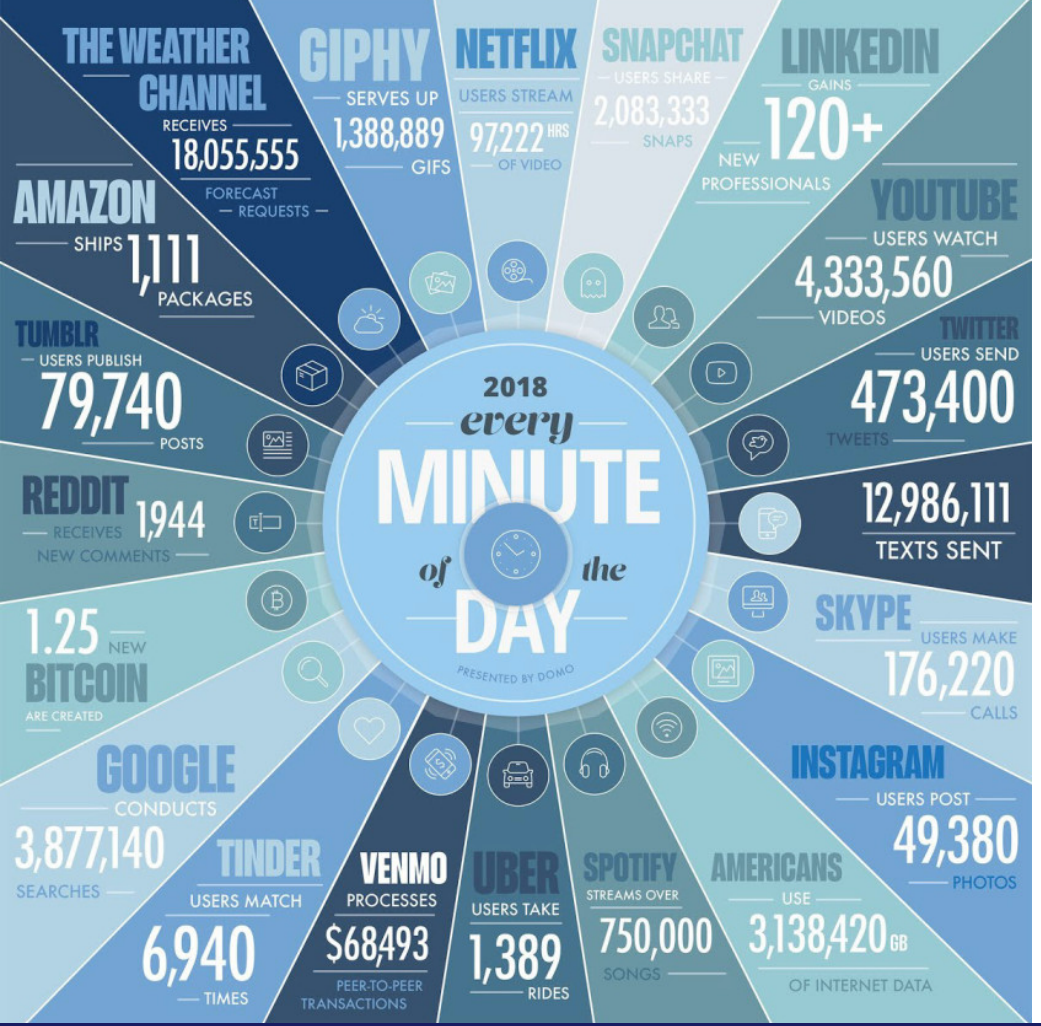
Voice Data

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

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 customer will default a loan from the bank?

Will loan application be approved?

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 Netflix viewer profile?

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 aircraft 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 is **Interdisciplinary**

1. Computational Thinking
2. Fundamentals of Statistics
3. Machine Learning Models
4. Multivariate Linear Algebra
5. Data Mining / Wrangling
6. Visualization and Design

- **Data-Analytic Thinking and Computations**
- **Domain Knowledge and Application Areas**