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

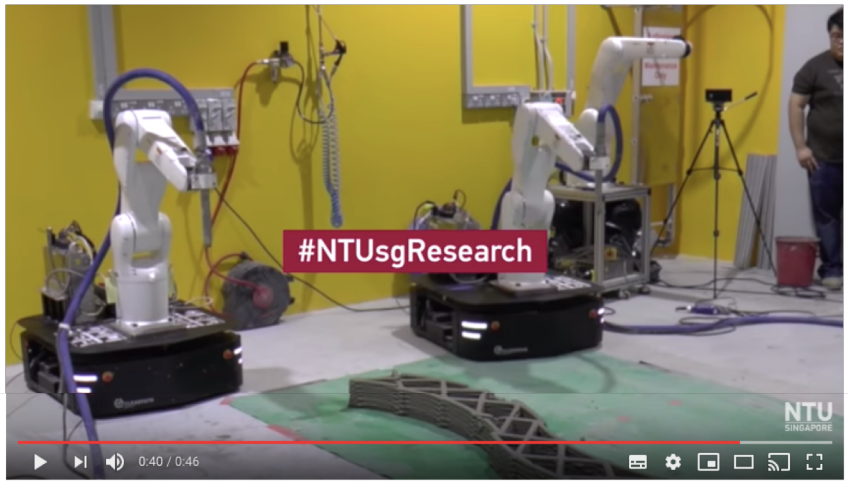
- Social Networks and Web
- Mobile Phone Cameras
- Blogs, Wikis, Documents

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

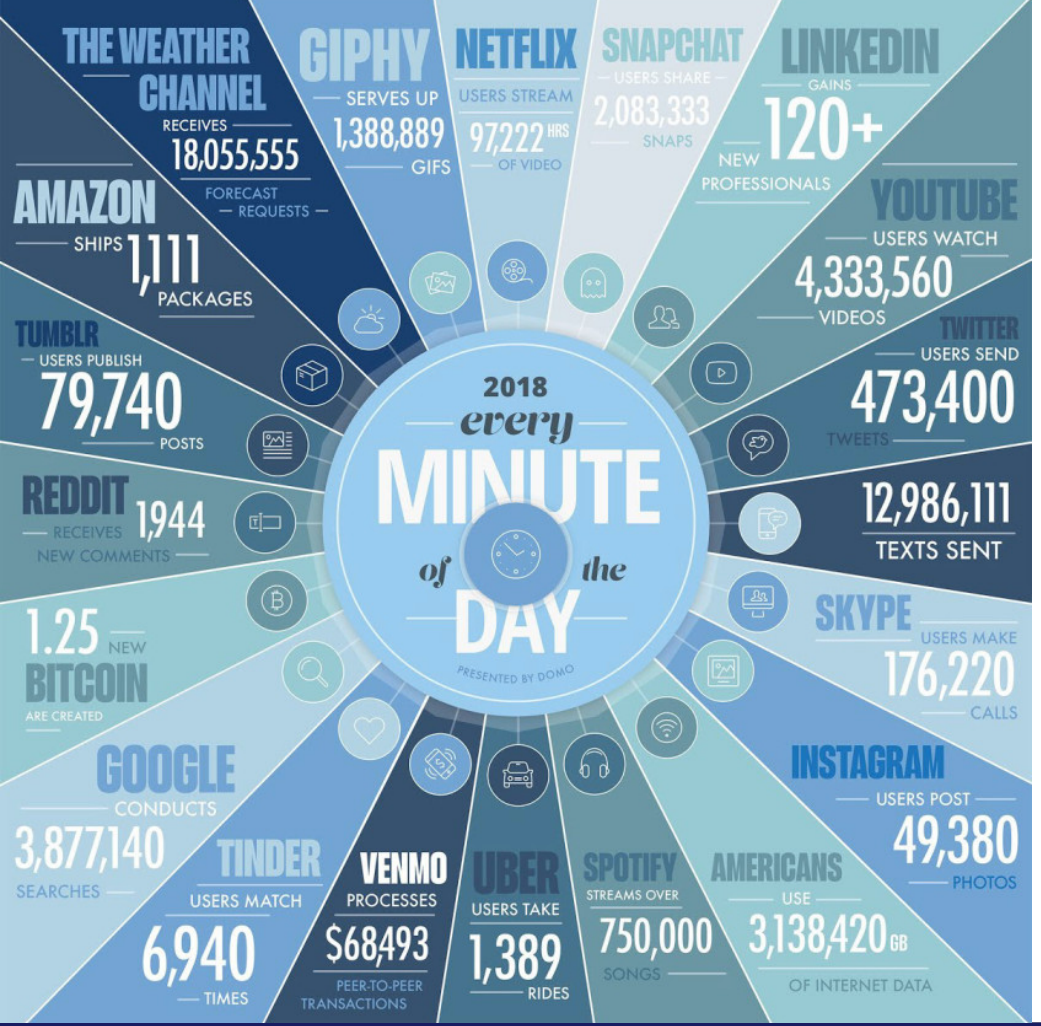
Voice Data

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

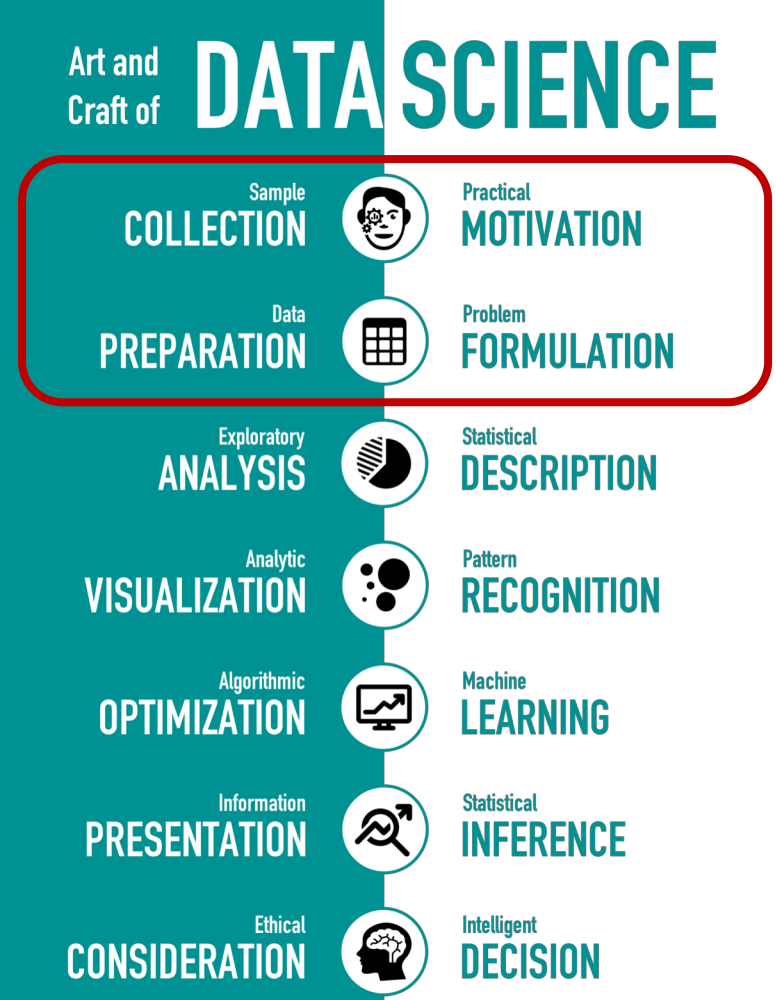


Data Science

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?