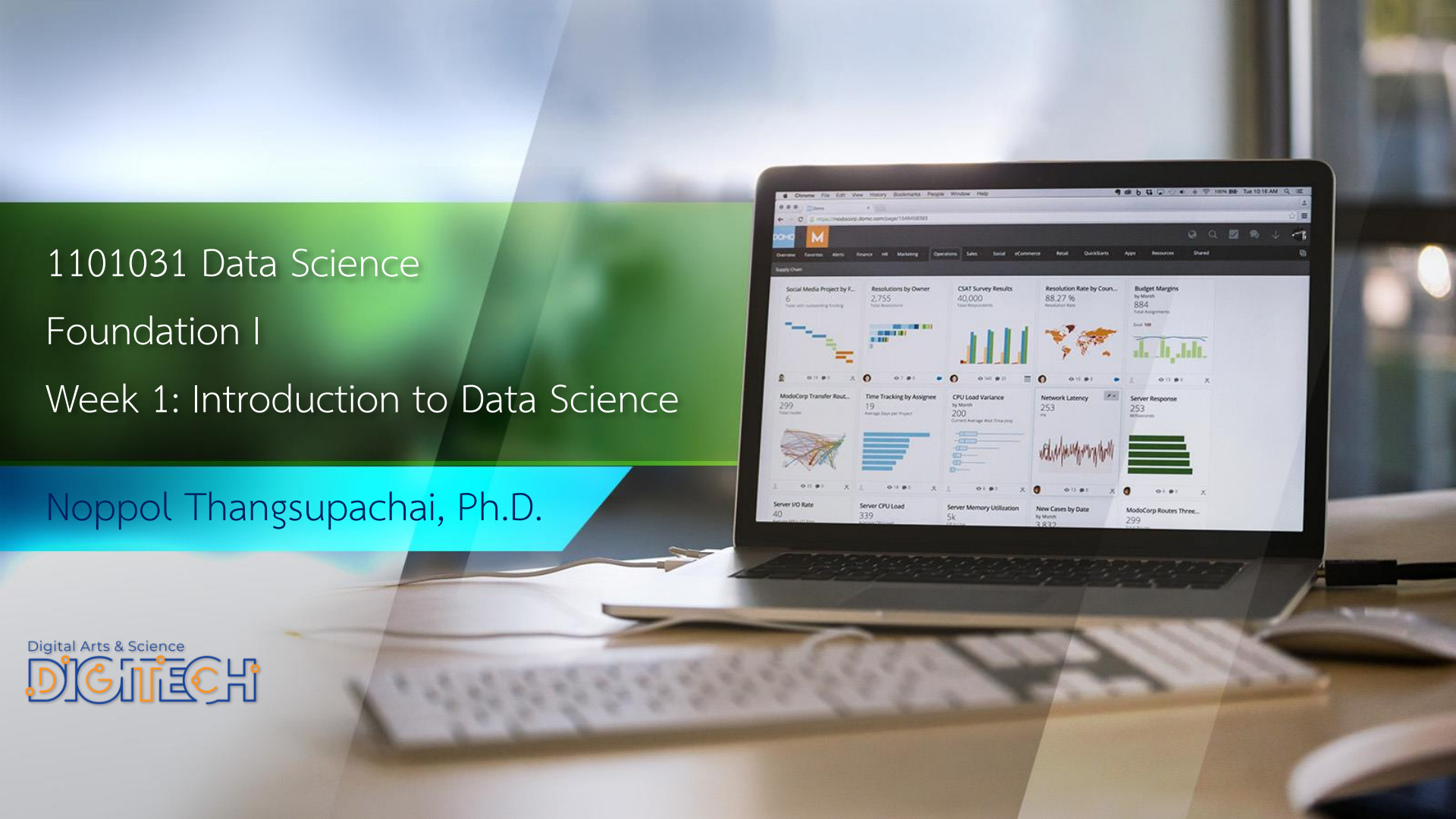
1101031 Data Science

Foundation I

Week 1: Introduction to Data Science

Noppol Thangsupachai, Ph.D.



Digital Arts & Science
DIGITECH

Topics

- Course syllabus roadmap
- Introduction
- What is data science
- Who is data scientist
- Data science process
- Data Sourcing

Digital Arts & Science
DIGITECH

Course syllabus

- 1101031 Data Science Foundation I
- 3 Credit (2-2-5)
- Modular with 1101032 Data Science Foundation II
- Lab and Assignment 60%
- Final Examination 40%

Digital Arts & Science DIGITECH

Introduction

- Motivation case:

Introduction

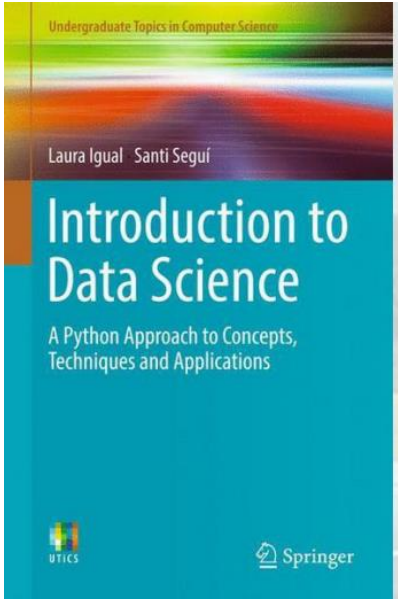
- What is Data science:

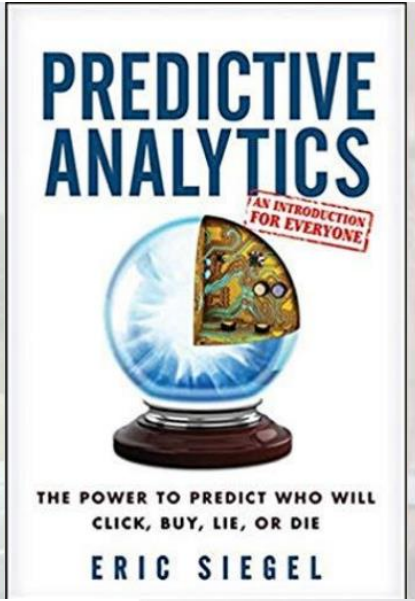
Introduction

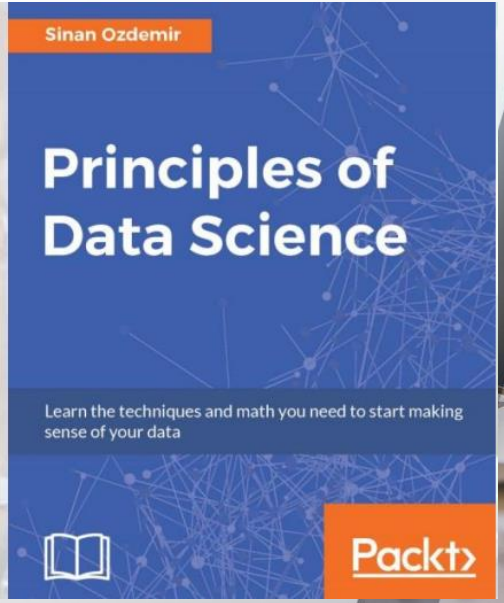
- What is Data science:

Introduction

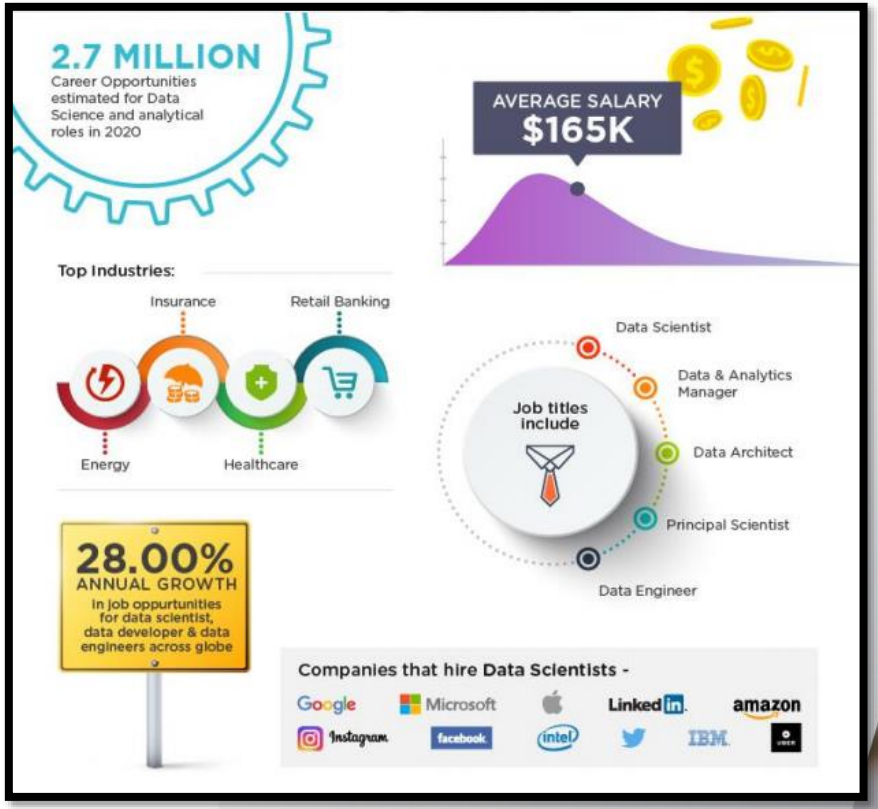
- Textbook and References:



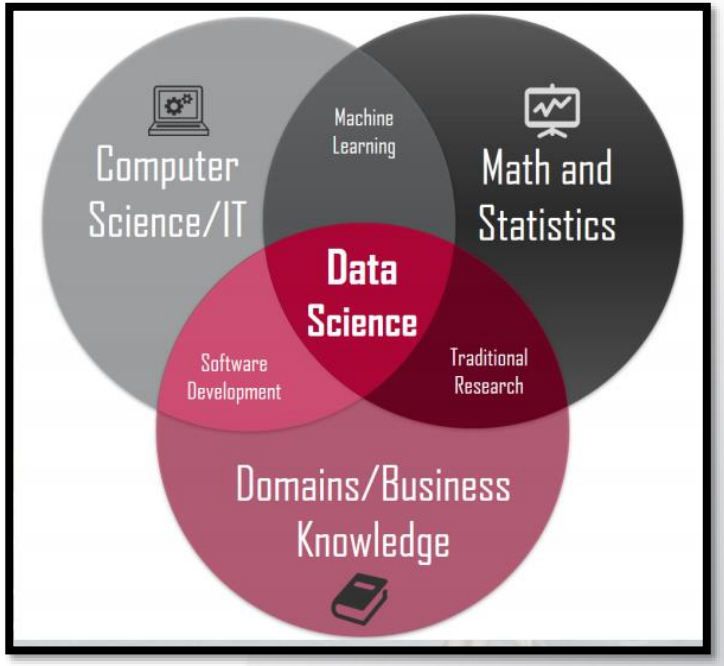




Why Data Science



Why Data Science



Digital Arts & Science DIGITECH

Why Data Science

- is the study of the generalizable extraction of knowledge from data (Wikipedia)
- is getting predictive and/or actionable insight from data (Neil Raden)
- Involves extracting, creating, and processing data to run it into business value (Vincent Granville)

Why Data Science

- Data science is not new, Data science is just modernizing existing reporting solution, analytics solution, data warehousing solution, business intelligence solution, and even data management solution. (Jothi Periasamy)

So, Data science is…

- New thinking
- New ideas
- New data source
- New data structure
- New data architecture
- New data processing mechanism
- Innovation on data
- New way of solving problems

Why Data Scientist

Why Data Science

Why Data Scientist

Data Scientist
KASIKORNBANK
Bangkok Metropolitan Area, Thailand
Good programming skills (Python, R, or Scala preferred) solve the specified problems using various ...

Data Scientist (Transformation)
Siam Commercial Bank
Bangkok Metropolitan Area, Thailand
Scripting experience in (Python, R, JavaScript, forms (data warehouses/SQL, unstructured da

Associate Data Scientist, SAS Global Pre-Sales Academy
SAS
Bangkok, TH
Experience programming with languages such as R, Python, SAS, Lua or simil determined based on the applicant's education, ... global-sas.icims.com

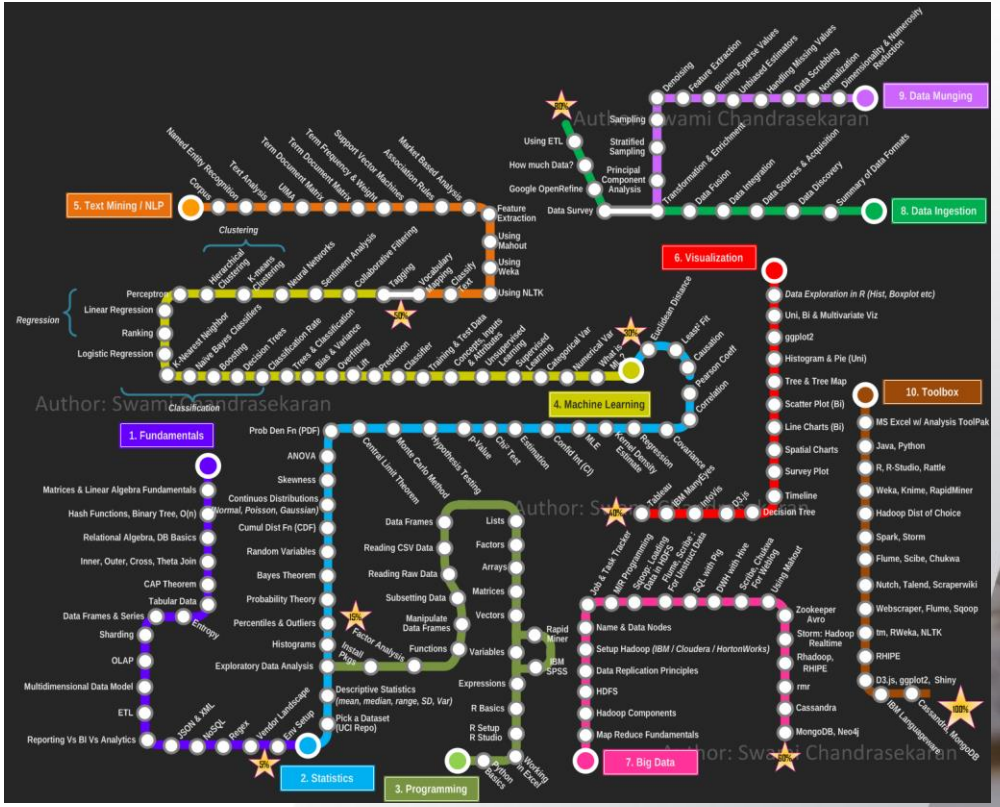
Source: Linkedin Jobs

Data Scientist
DTAC
Bangkok, Bangkok City
Working knowledge of SQL and relational databases and analysis tool data analysis and develop effective machine ...

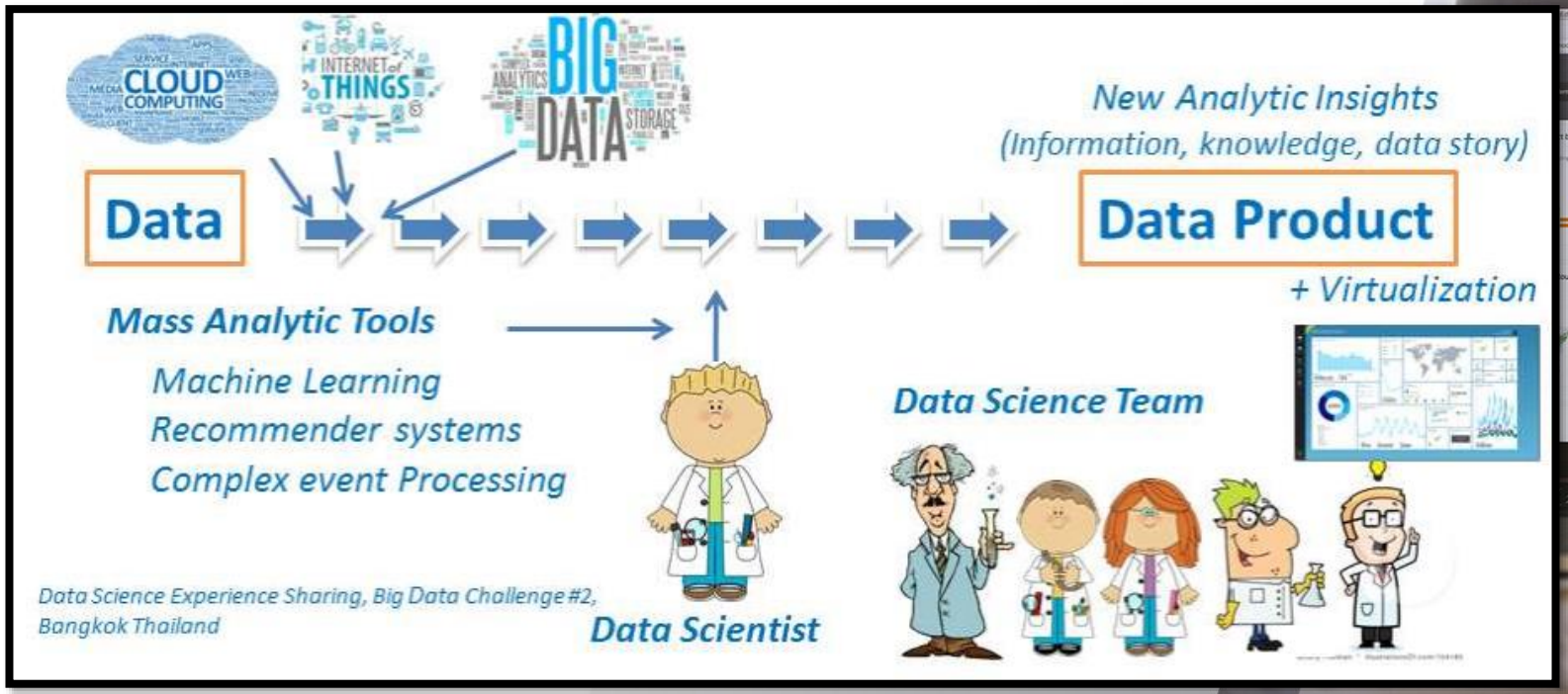
Data Engineer - MIS
TMB Bank PCL
Bangkok Metropolitan Area, Thailand
Technical Skills : Python, R, Tableau, SQL. I data from Data Warehouse , a wide variet

Do you want to grab any of these hot jobs?

Become a Data Scientist



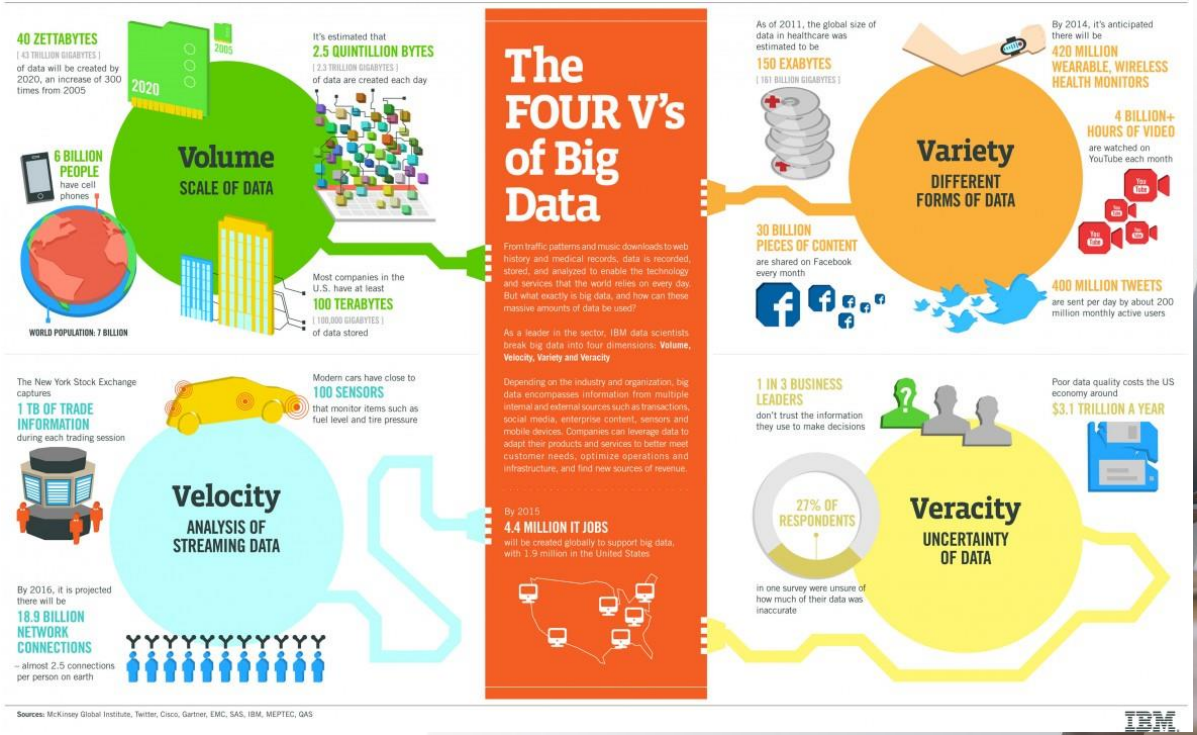
Become a Data Scientist



Why cloud computing?

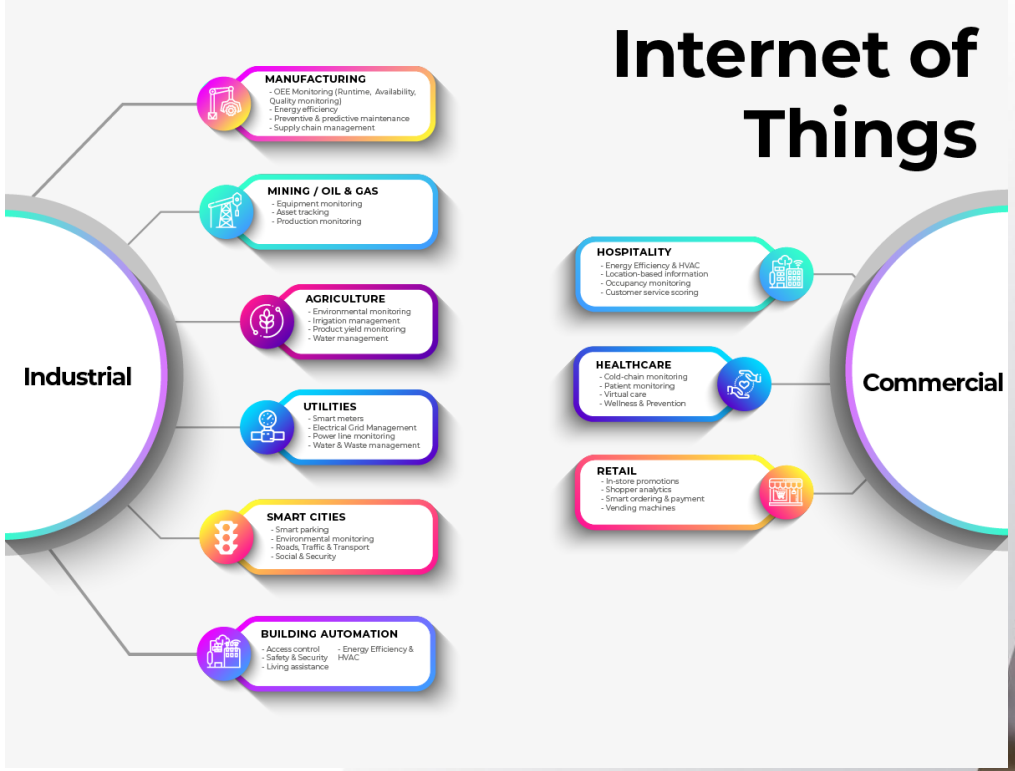
https://www.thumbsup.in.th/cloud-computing-thailand-survey-by-imc

Why Big Data?



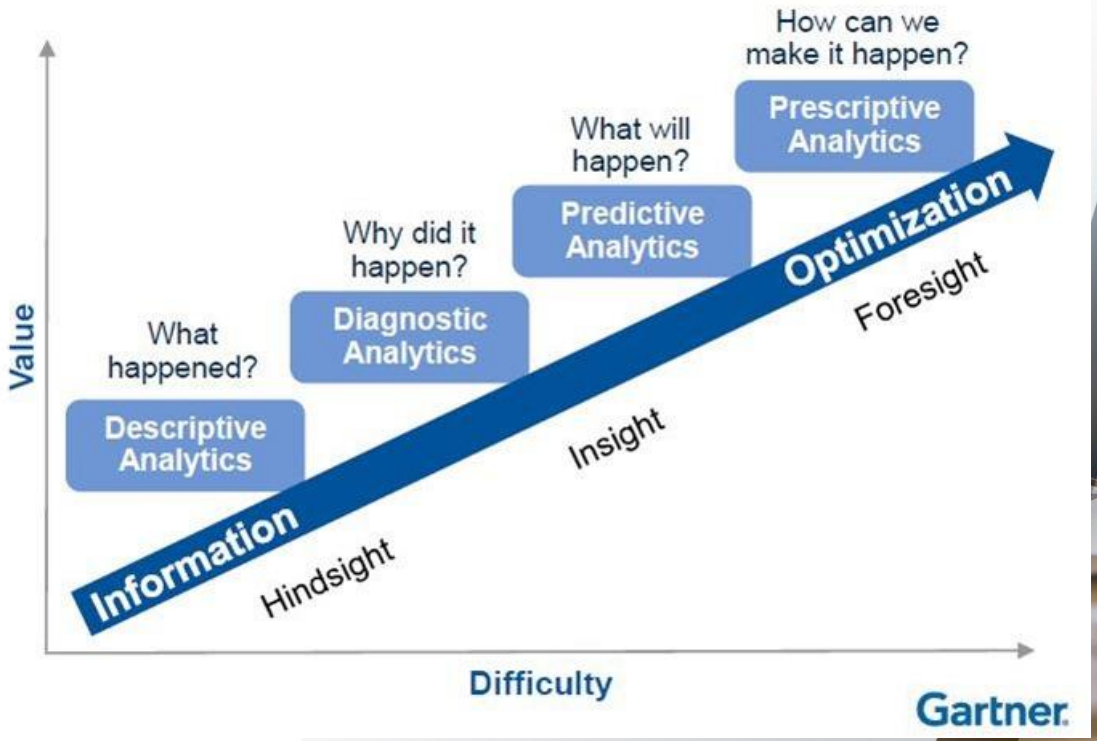
https://www.ibmbigdatahub.com/infographic/four-vs-big-data

Why Internet of Thing?



https://ubidots.com/blog/iot-consumer-vs-commercial-vs-industrial-main-overview/

Data Analytics Levels



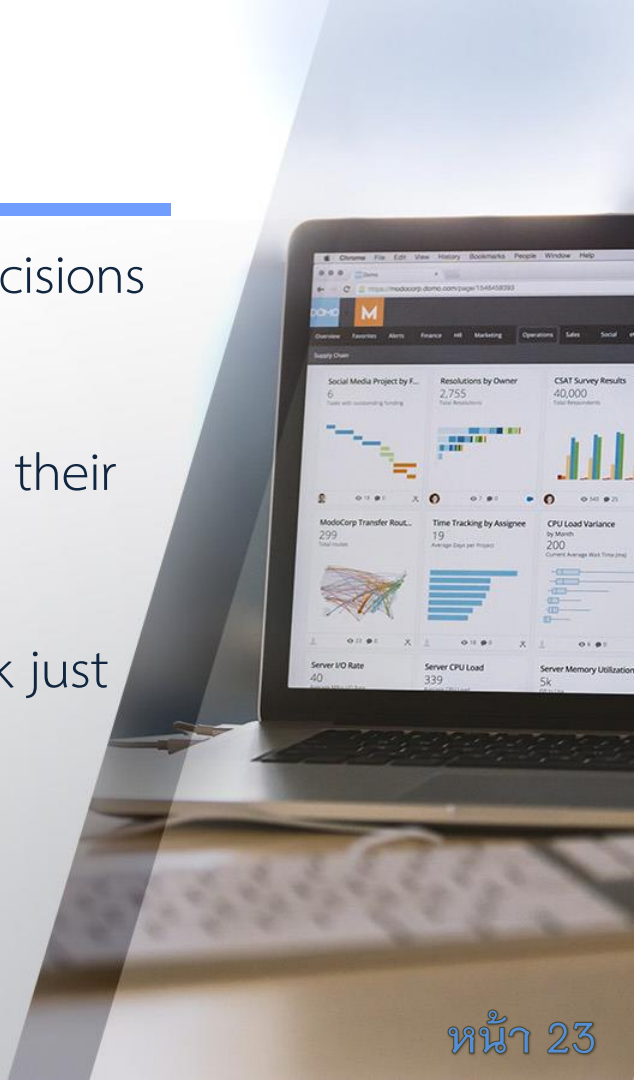
https://datafloq.com/read/prescriptive-analytics-getting-ahead-of-the-curve/2202

Trend of Predictive Analytics

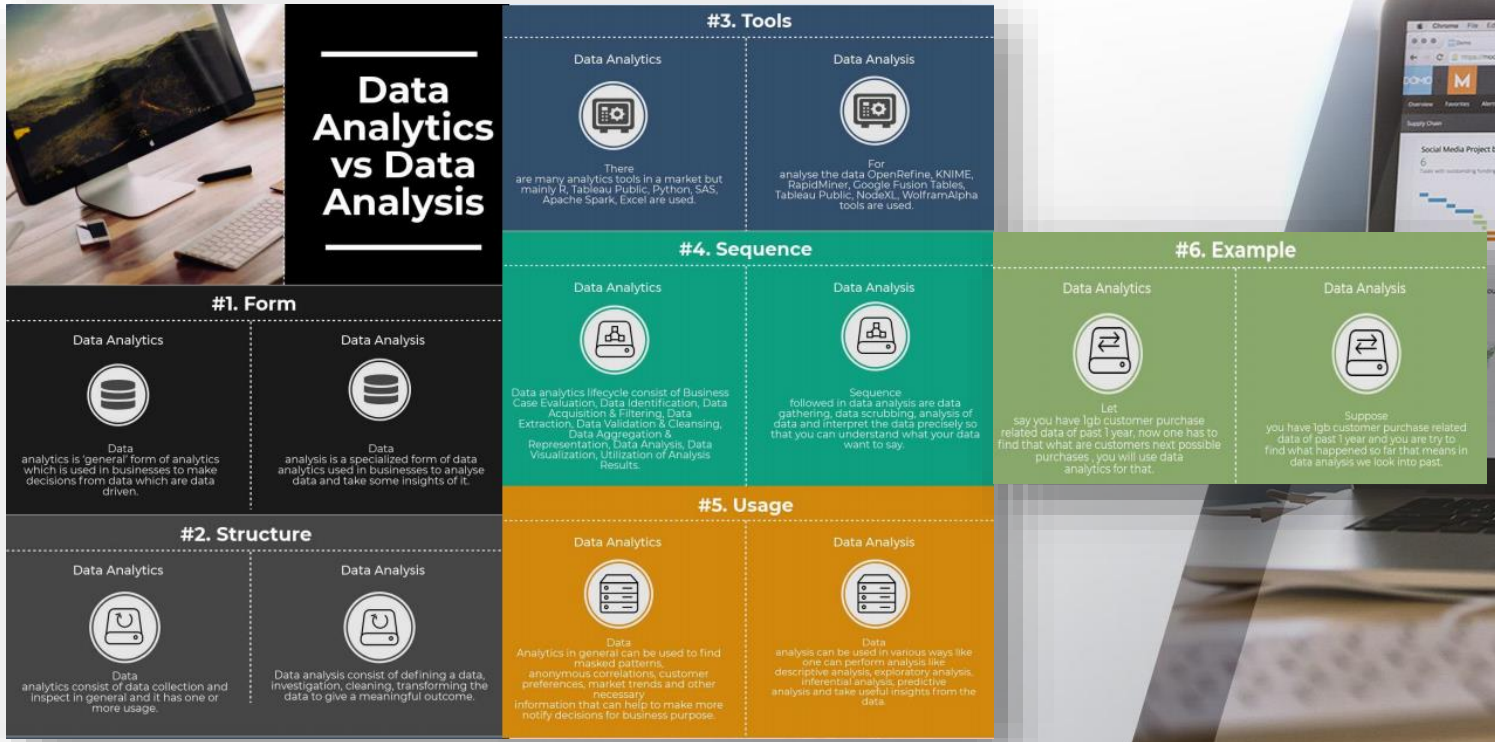
- A lot of companies want to do predictive analytics, but have yet to master basic reporting – *Deloitte Consulting's Miller*
- Predictive analytics is forward looking using past events to anticipate the future
- A set of Business Intelligence technologies that uncovers relationships and patterns within large volumes of data that can be used to predict behavior and events.

Challenges to Decision Maker

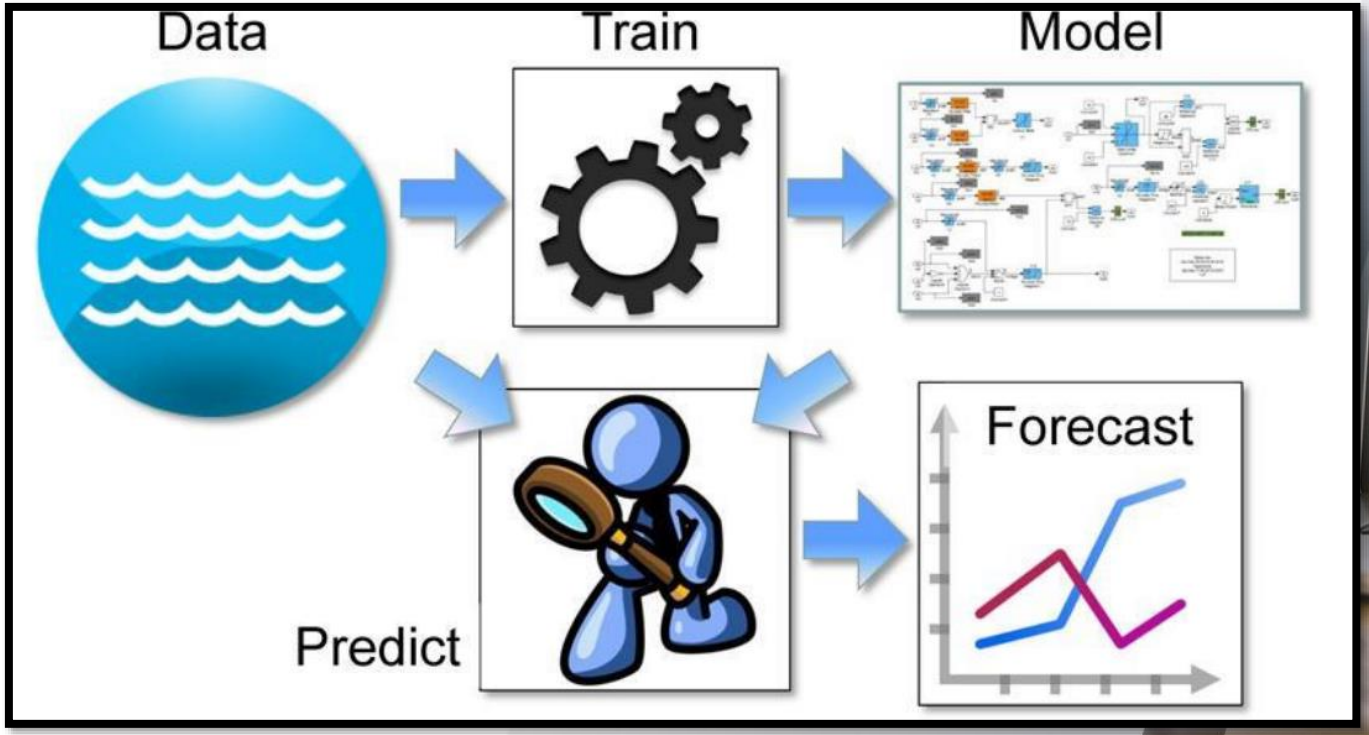
- 1 in 3 business leaders frequently make critical decisions without the information they need
- 1 in 2 don’t have access to the information across their organization needed to do their jobs
- 19+ Hours spent by knowledge workers each week just searching for and understanding information

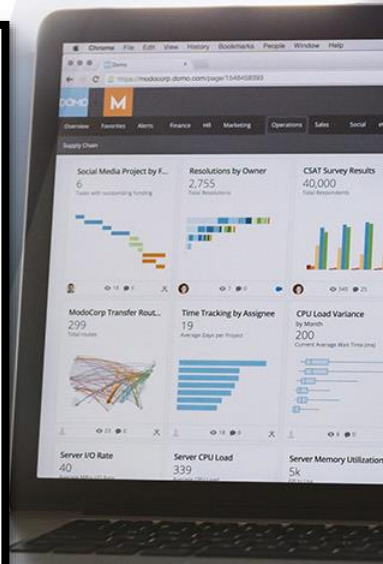


Why Data Science



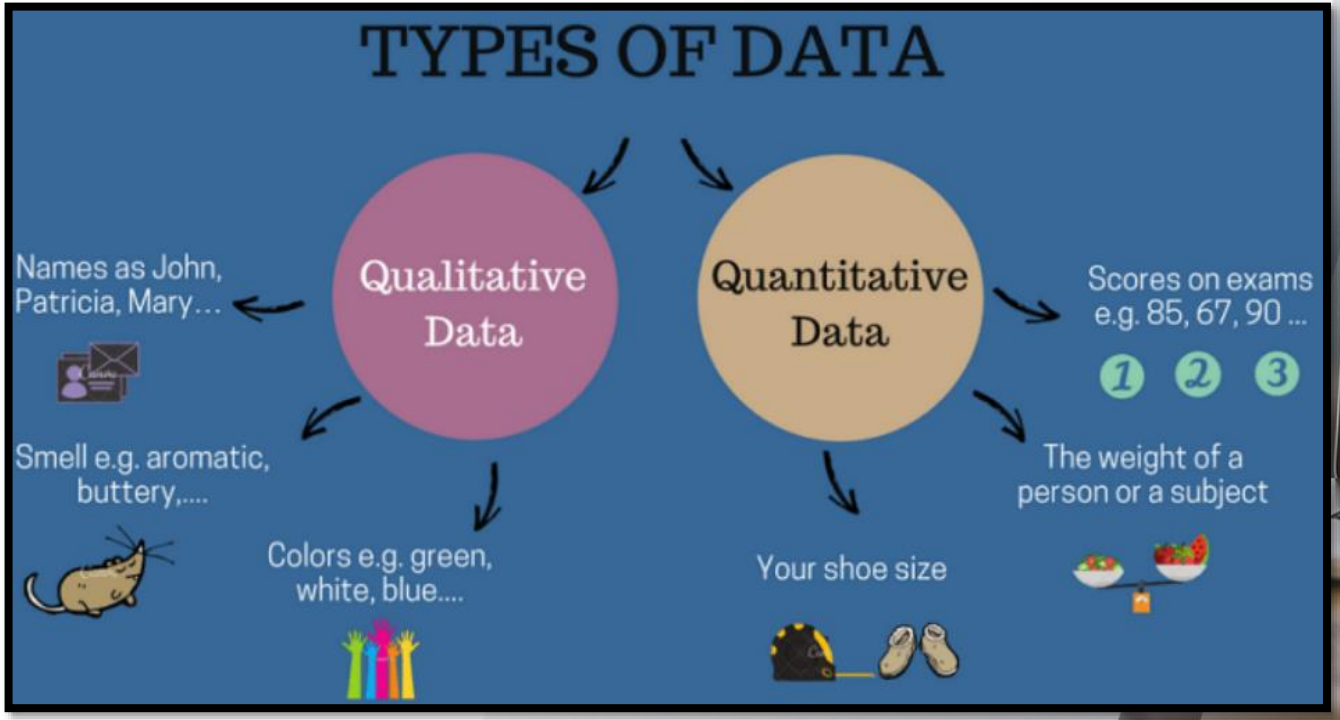
Trend of Predictive Analytics





Digital Arts & Science
DIGITECH

Type of Data


TYPES OF DATA
Qualitative Data
Quantitative Data
Names as John, Patricia, Mary…
Smell e.g. aromatic, buttery,....
Colors e.g. green, white, blue....
Scores on exams e.g. 85, 67, 90 ...
1 2 3
The weight of a person or a subject
Your shoe size

Quantitative Data

Key characteristics of quantitative data:

- It can be quantified and verified.
- Data can be counted.
- Data type: number and statistics.
- It answers questions such as “how many, “how much” and “how often”.

Examples of quantitative data:

- Scores on tests and exams e.g. 85, 67, 90 and etc.
- The weight of a person or a subject.
- The number of hours of study.
- Your shoe size.
- The square feet of an apartment.
- The temperature in a room.
- The volume of a gas and etc.

Type of Quantitative Data

- **Discrete data** – a count that involves integers. Only a limited number of values is possible. The discrete values cannot be subdivided into parts. For example, the number of children in a school is discrete data. You can count whole individuals. You can’t count 1.5 kids.
- **Continuous data** – information that could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have almost any numeric value. For example, you can measure your height at very precise scales — meters, centimeters, millimeters and etc.

Type of Qualitative Data

Key characteristics of qualitative data:

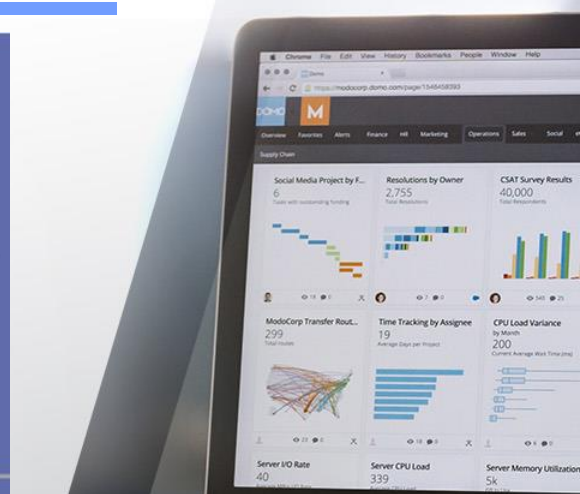
- It cannot be quantified and verified.
- Data cannot be counted.
- Data type: words, objects, pictures, observations, and symbols.
- It answers questions such as “how this has happened” or and “why this has happened”.

Examples of qualitative data:

- Your socioeconomic status
- Colors e.g. the color of the sea
- The Smell e.g. aromatic, buttery, camphoric and etc.
- Your favorite holiday destination such as Hawaii, New Zealand and etc.

Qualitative VS Quantitative

Basis for Comparison	Qualitative Data	Quantitative Data
Definition	Qualitative data is information that can't be expressed as a number	Quantitative data is data that can be expressed as a number or can be quantified
Can data be counted?	NO	YES
Data type	Words, objects, pictures, observations, and symbols	Number and statistics
Questions that data answer	How and why this has happened?	"how many, "how much" and "how often"

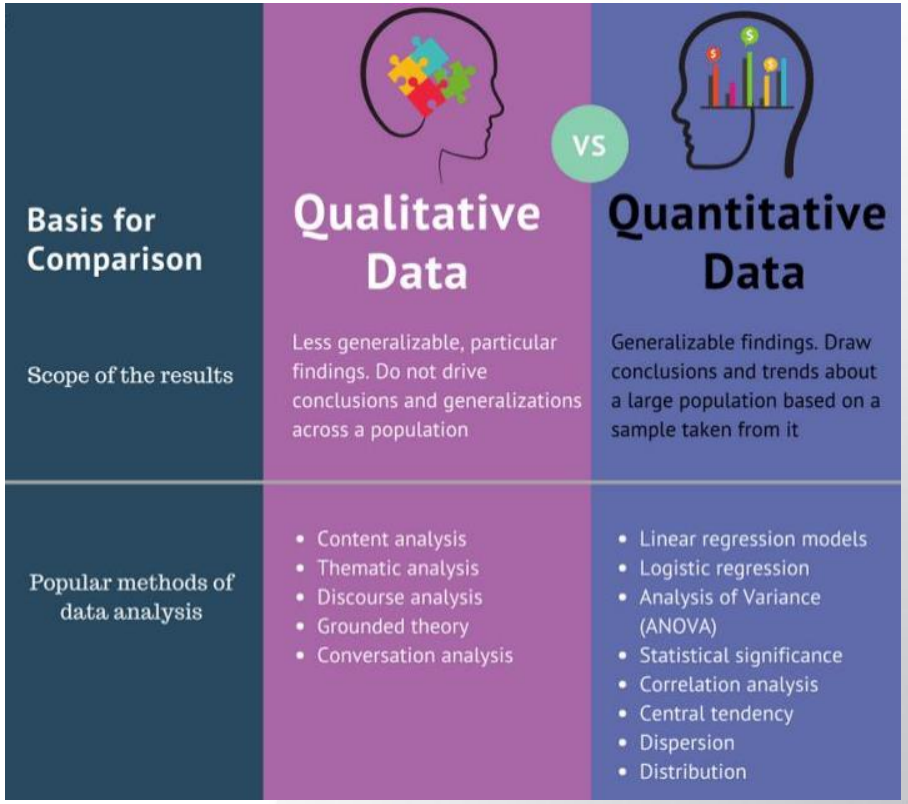


Qualitative VS Quantitative

Basis for Comparison	Qualitative Data	Quantitative Data
Examples	• Names as John, Maria,… • Ethnicity such as American Indian, Asian, etc. • Colors e.g. green, white, blue	• Scores on tests and exams e.g. 85, 67, 90 and etc. • The weight of a person or a subject • Your shoe size
Purposes of data analysis	Understand, explain, and interpret social interactions and patterns	Test hypothesis, develop predictions for the future, check cause and effect
Types of data analysis	Patterns, characteristics, theme identification	Statistical relationship identification

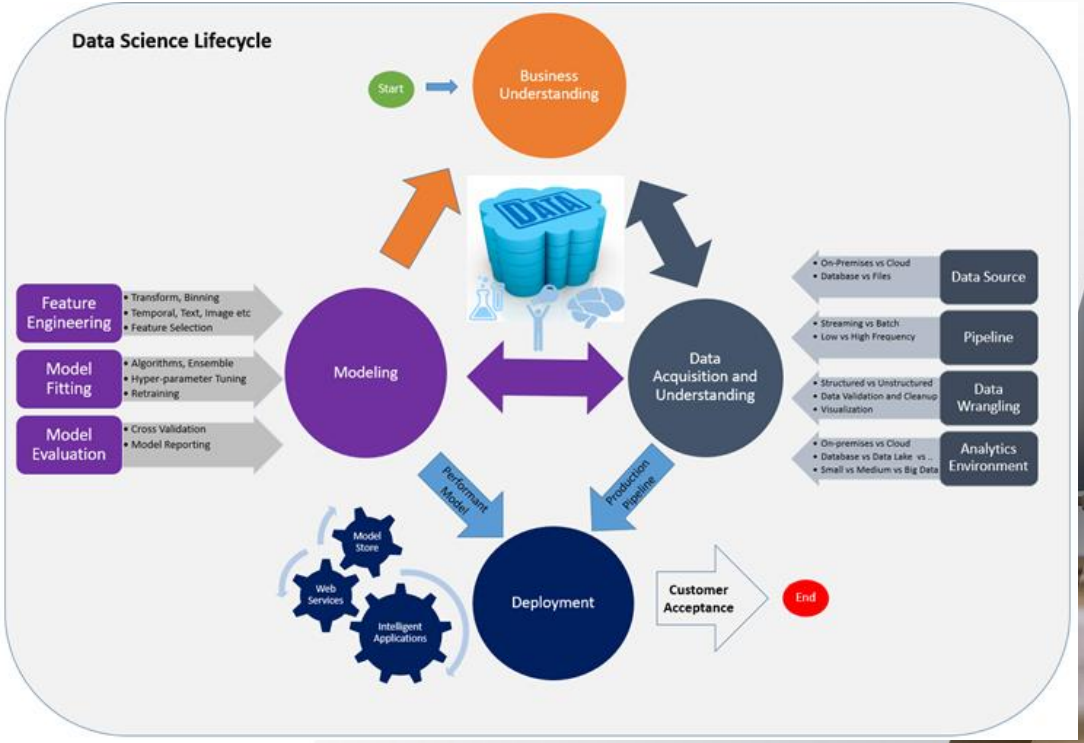
Digital Arts & Science DIGITECH

Qualitative VS Quantitative



Basis for Comparison	Qualitative Data	Quantitative Data
Scope of the results	Less generalizable, particular findings. Do not drive conclusions and generalizations across a population	Generalizable findings. Draw conclusions and trends about a large population based on a sample taken from it
Popular methods of data analysis	• Content analysis • Thematic analysis • Discourse analysis • Grounded theory • Conversation analysis	• Linear regression models • Logistic regression • Analysis of Variance (ANOVA) • Statistical significance • Correlation analysis • Central tendency • Dispersion • Distribution

Data Science Life Cycle



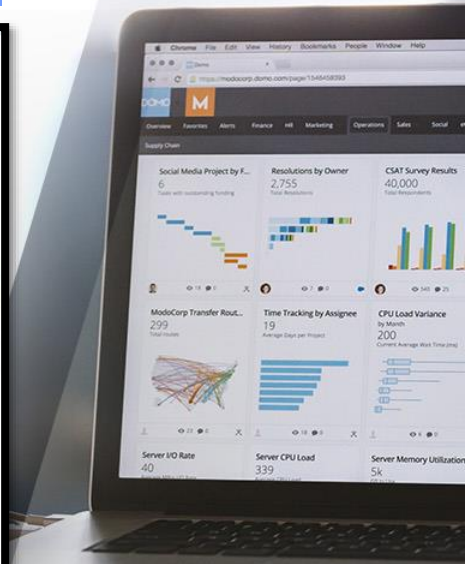
https://blog.revolutionanalytics.com/2016/10/the-team-data-science-process.html

Business Understanding

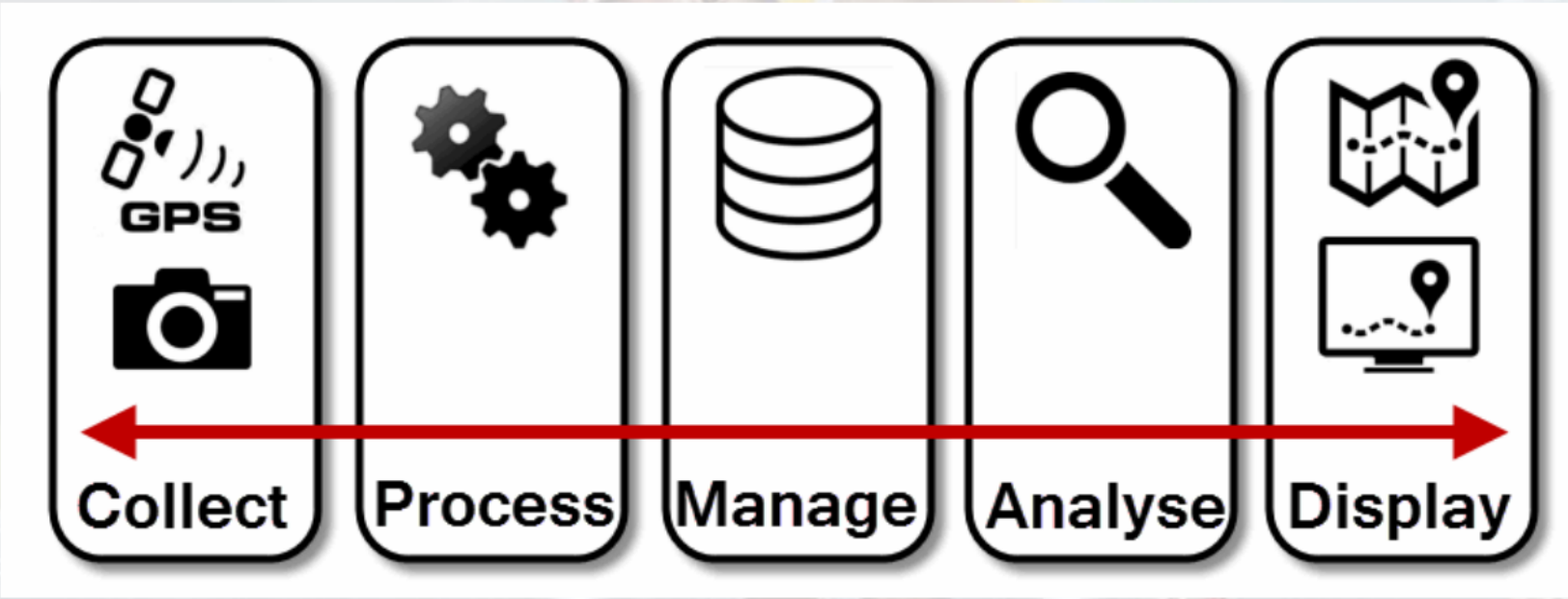
BUSINESS UNDERSTANDING

Determining Business Objectives

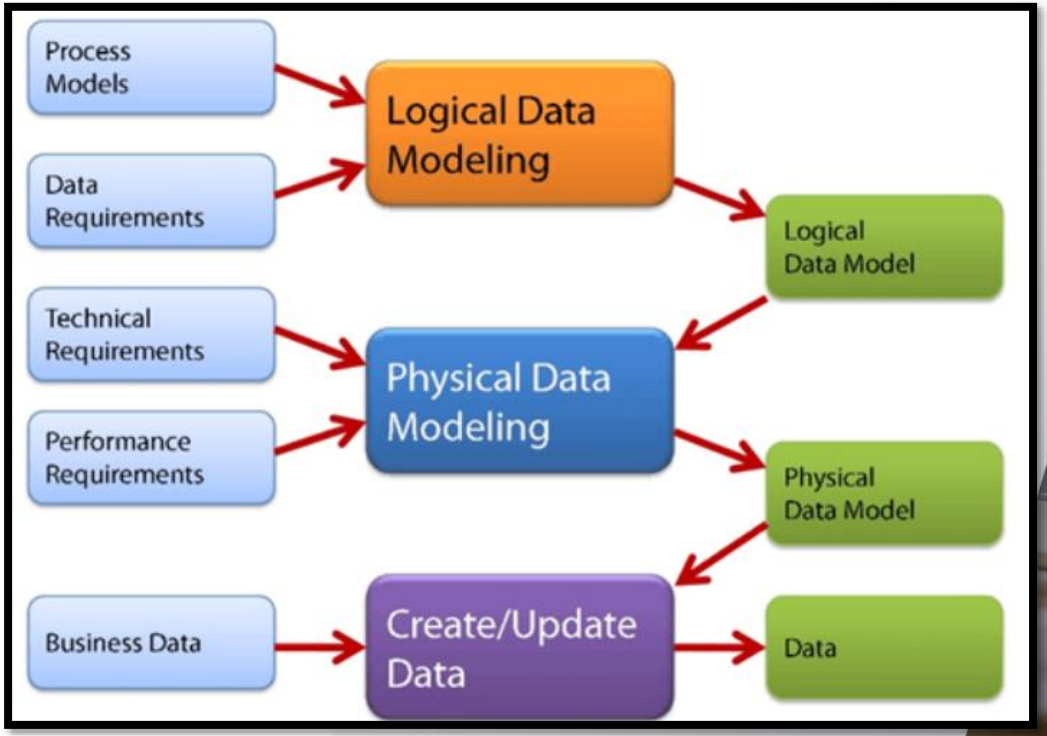
1. Gather background information
   - Compiling the business background
   - Defining business objectives
   - Business success criteria
2. Assessing the situation
   - Resource Inventory
   - Requirements, Assumptions, and Constraints
   - Risks and Contingencies
   - Cost/Benefit Analysis
4. Determining data science goals
   - Data science goals
   - Data science success criteria
4. Producing a Project Plan



Data Acquisition and Understanding



Data Modeling



Data Modeling

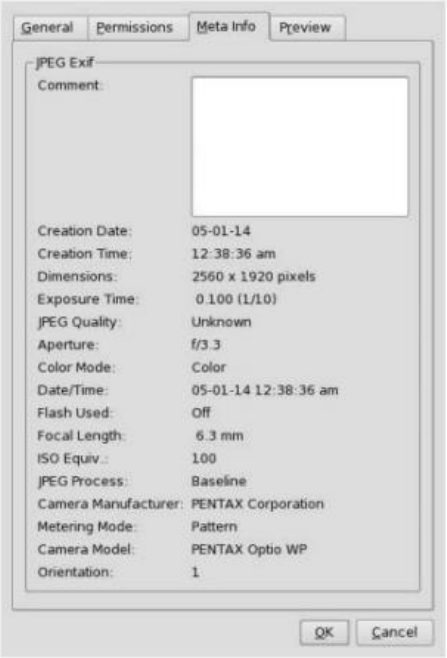
Type of Data Models

- **Conceptual:** This Data Model defines WHAT the system contains. This model is typically created by Business stakeholders and Data Architects. The purpose is to organize, scope and define business concepts and rules.
- **Logical:** Defines HOW the system should be implemented regardless of the DBMS. This model is typically created by Data Architects and Business Analysts. The purpose is to developed technical map of rules and data structures.
- **Physical:** This Data Model describes HOW the system will be implemented using a specific DBMS system. This model is typically created by DBA and developers. The purpose is actual implementation of the database.

Data VS Metadata

DATA

METADATA



General | Permissions | Meta Info | Preview

JPEG Exif

Comment:	
Creation Date:	05-01-14
Creation Time:	12:38:36 am
Dimensions:	2560 x 1920 pixels
Exposure Time:	0.100 (1/10)
JPEG Quality:	Unknown
Aperture:	f/3.3
Color Mode:	Color
Date/Time:	05-01-14 12:38:36 am
Flash Used:	Off
Focal Length:	6.3 mm
ISO Equiv.:	100
JPEG Process:	Baseline
Camera Manufacturer:	PENTAX Corporation
Metering Mode:	Pattern
Camera Model:	PENTAX Optio WP
Orientation:	1

OK | Cancel

Data Model Representation

Classes of **entities** **(kinds of things)** about which a company wishes to know or hold information

WHO	*Person, Employee, Vendor, Customer, Department, Organisation, …*
WHAT	*Product, Service, Raw Material, Training Course, Flight, Room, …*
WHEN	*Time, Day, Date, Calendar, Reporting Period, Fiscal Period, …*
WHERE	*Geographic location, Delivery address, Storage Depot, Airport, …*
WHY	*Order, Complaint, Inquiry, Transaction, …*
HOW	*Invoice, Policy, Contract, Agreement, Document, Account, …*

Data Model Representation

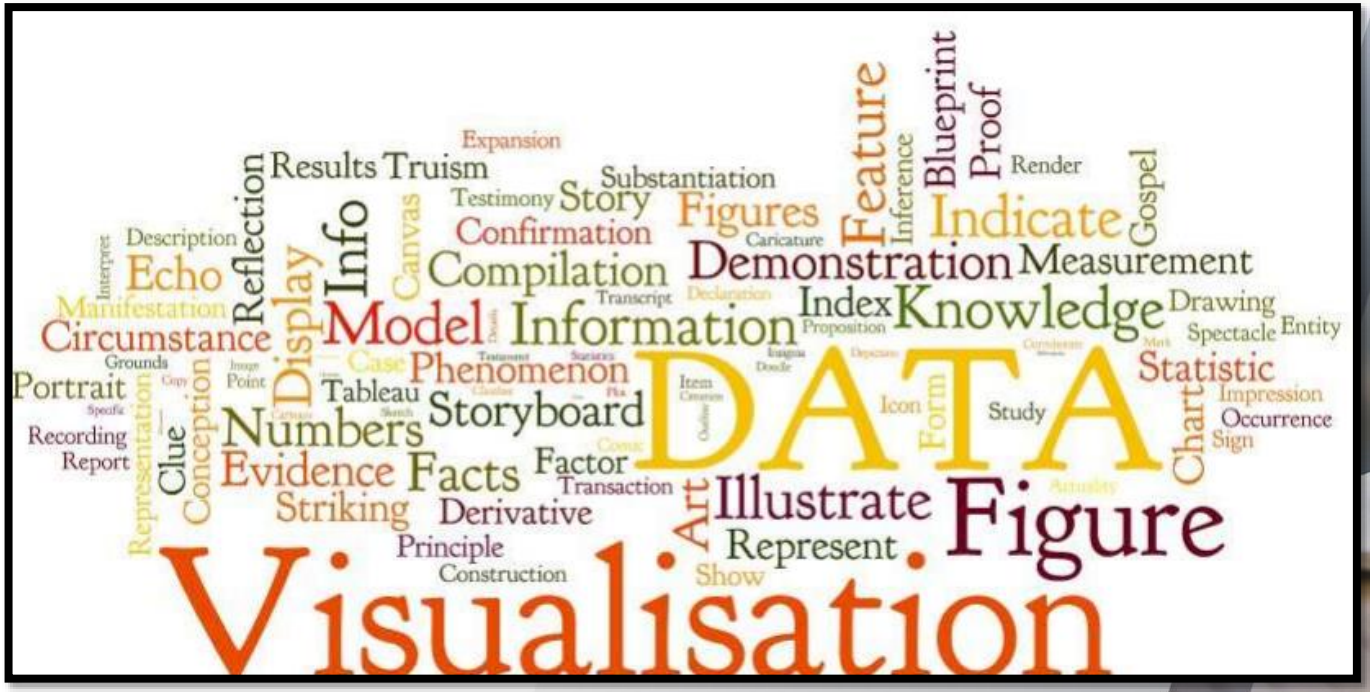
Classes of **entities** **(kinds of things)** about which a company wishes to know or hold information

the **attributes** of that information **(facts about things)**

relationships among those entities and **(often implicit)** relationships among those attributes

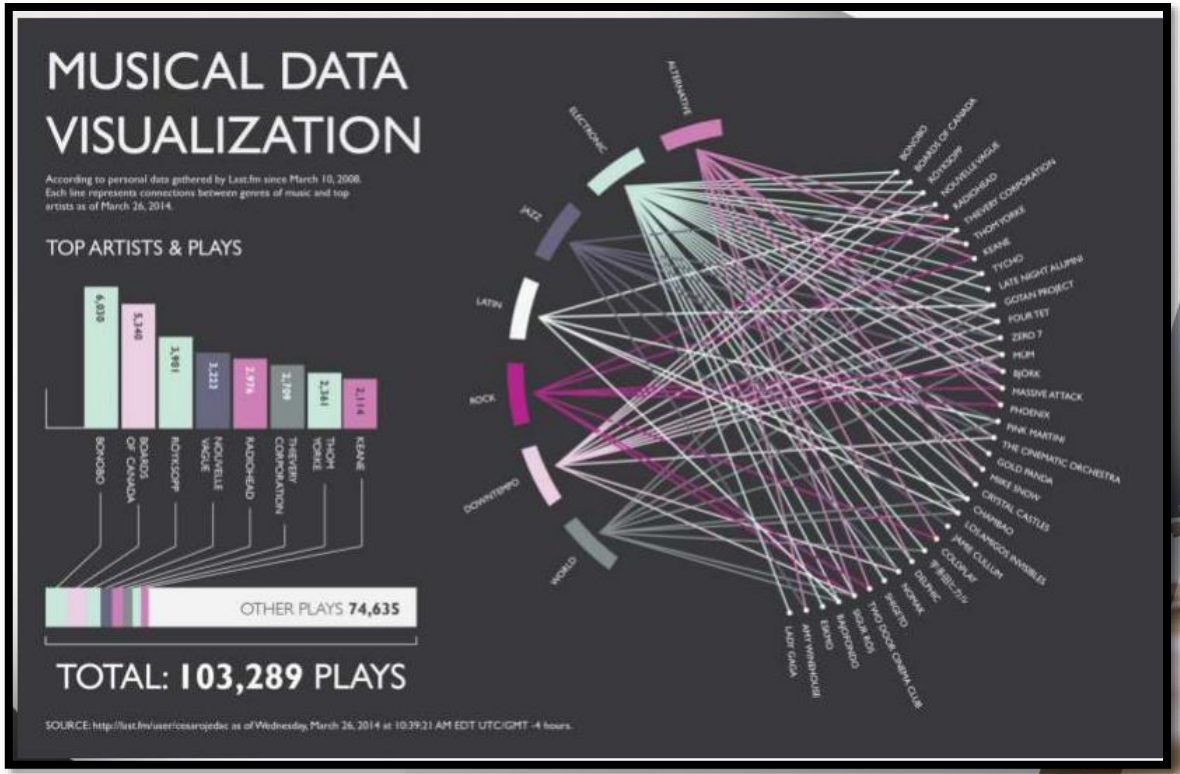
The model describes the organization of the data irrespective of how data might be represented in a computer system

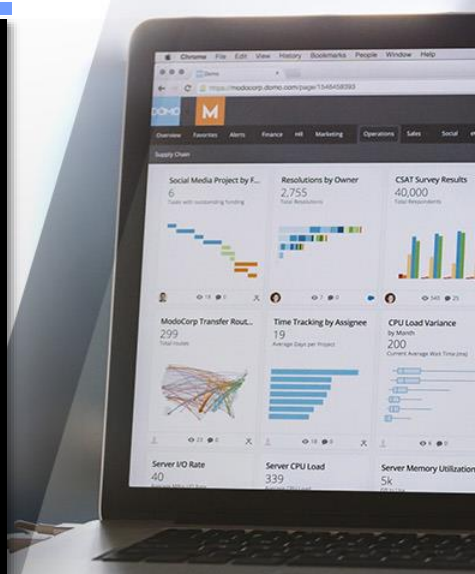
Data Visualisation



Data Visualisation

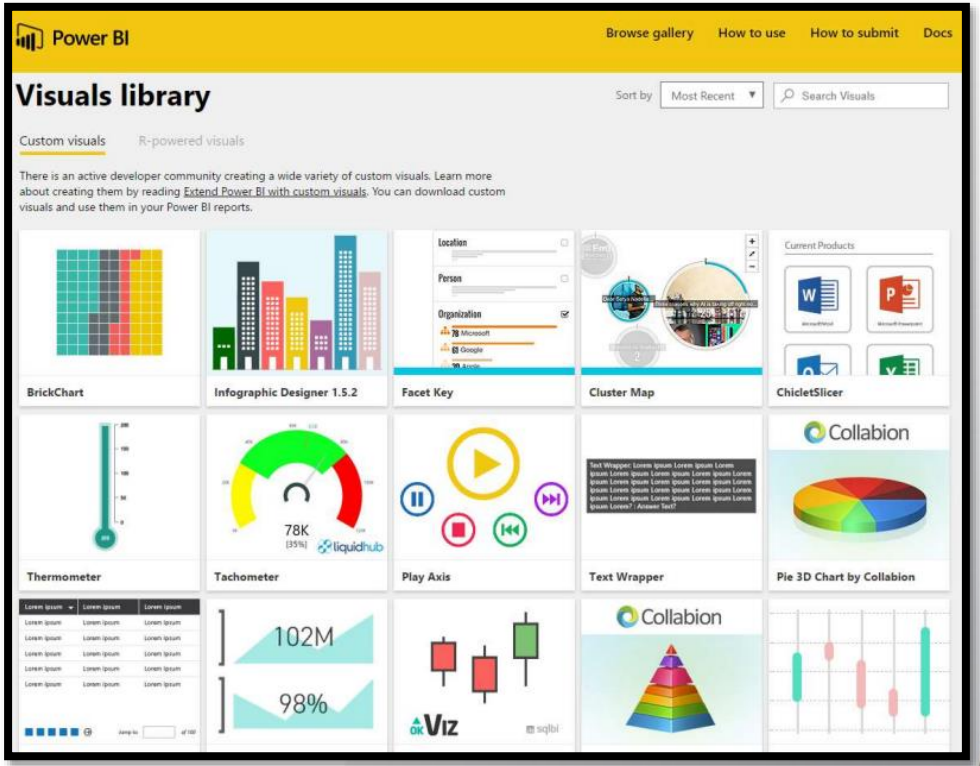
Data Visualisation





Digital Arts & Science
DIGITECH

Data Visualisation



Data Sourcing

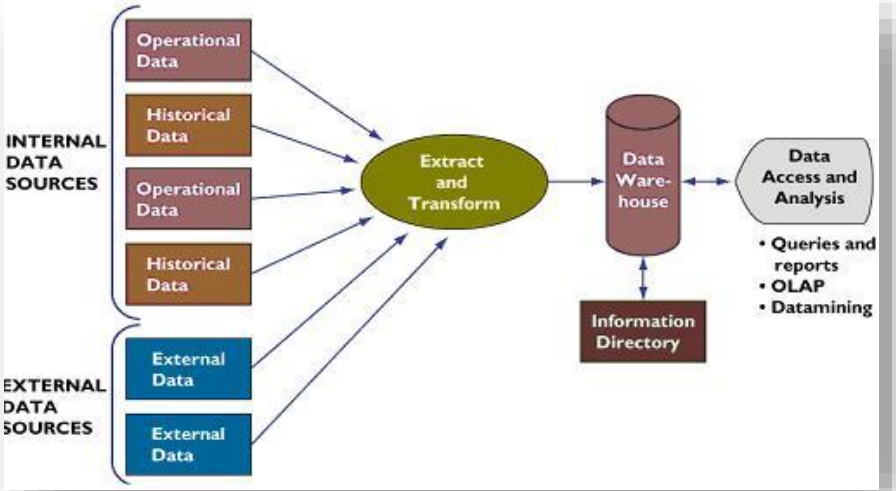
- กำหนดแหล่งข้อมูลที่มีความสำคัญต่อองค์กร
- รูปแบบ ลักษณะ ที่มาของข้อมูล
- กำหนดวิธีการสกัด (Extract) ข้อมูลเพื่อองค์กรสามารถใช้งานเมื่อต้องการ
- กำหนดนโยบายในการนำเข้า แปลง (Transform) และจัดเก็บข้อมูลในคลังข้อมูล (Load)

Building A Data Landscape, 2013
https://online-behavior.com/analytics/data-landscape

Data Sourcing (cont.)

แหล่งข้อมูลที่องค์กรสามารถนำมาใช้ในการดำเนินงานประกอบด้วย

- แหล่งข้อมูลภายในองค์กร
  - ระบบสารสนเทศในองค์กร
  - ข้อมูลการปฏิบัติงาน
  - ระบบบันทึกข้อมูลการทำงานอัตโนมัติ
- แหล่งข้อมูลภายนอกองค์กร
  - ข้อมูลผู้ที่เกี่ยวข้องกับองค์กร
  - ข้อมูลจากสื่อสังคมออนไลน์



DATA WAREHOUSE + DATA MINING, 2013
https://9chooknow.blogspot.com/2013/03/data-warehouse-data-mining.html

Data Sourcing (cont.)

Business requirement mapping to source systems

- KPI dimension matrix
  - Profile Source Systems for Relevant Datasets
  - Define Source Extract Mechanisms
  - Provide Source Extract Files for Information Integration:
    - structure data file
    - naming convention data heading
    - frequency of generate data
    - mode of delivery

	COMMON DIMENSIONS						
BUSINESS PROCESSES	Date	Product	Warehouse	Store	Promotion	Customer	Employee
Issue Purchase Orders	X	X	X				
Receive Warehouse Deliveries	X	X	X				X
Warehouse Inventory	X	X	X				
Receive Store Deliveries	X	X	X	X			X
Store Inventory	X	X		X			
Retail Sales	X	X		X	X	X	X
Retail Sales Forecast	X	X		X			
Retail Promotion Tracking	X	X		X	X		
Customer Returns	X	X		X	X	X	X
Returns to Vendor	X	X		X			X
Frequent Shopper Sign-Ups	X			X		X	X

The Matrix revisited, 2005
https://www.kimballgroup.com/2005/12/the-matrix-revisited/

Data Sourcing (cont.)

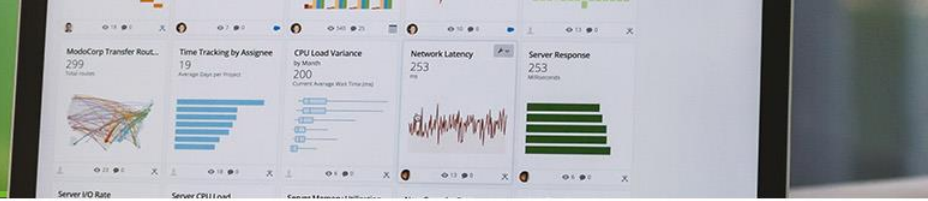


Table 3-2. *Key Differences Between Push and Pull Mechanisms*

Parameters	Push Mechanism	Pull Mechanism
Nature of extraction	Source system team provides the source data extracts in the interface formats provided by the information integration team.	The information integration team is provided read access to source tables to query and pick up the relevant data sets for further processing.
Source system knowledge	The source system team has extensive knowledge of the source systems and provides the source data extracts as per the interface formats agreed with the information integration team.	The information integration team has to build knowledge of the source system and **extract** the relevant data from the source tables based on the access provided by the source systems team.
Source system changes	In case of push mechanism, there is no impact of source system structure changes as the source system team generates the extract files. The information integration process is insulated from the source system changes.	In case of pull mechanism, the source system structure changes have to be understood by the information integration team and there will be changes to the information integration jobs that access the source systems to pull relevant data.

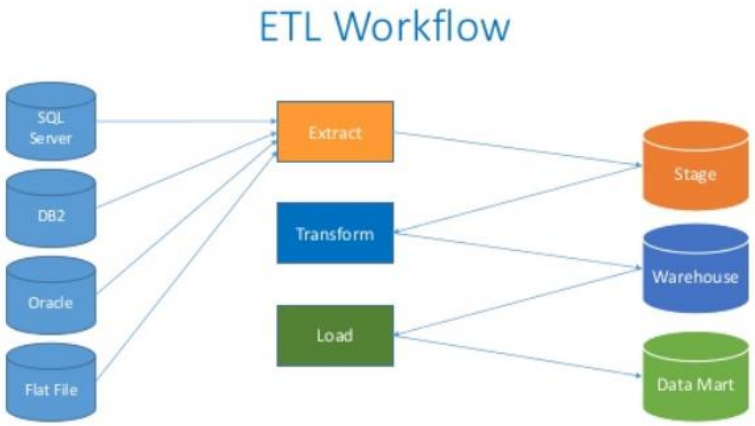
ที่มา:

Enterprise Information Management in Practice: Managing Data and Leveraging Profits in Today's Complex Business Environment, Saumya Chaki, 2015)

Data Sourcing (cont.)

Information Sourcing Patterns and Challenges

- Logical Data Extraction
  - Full extraction
  - Incremental extraction
  - Change data capture
- Physical Data Extraction
- Automated Data Extraction

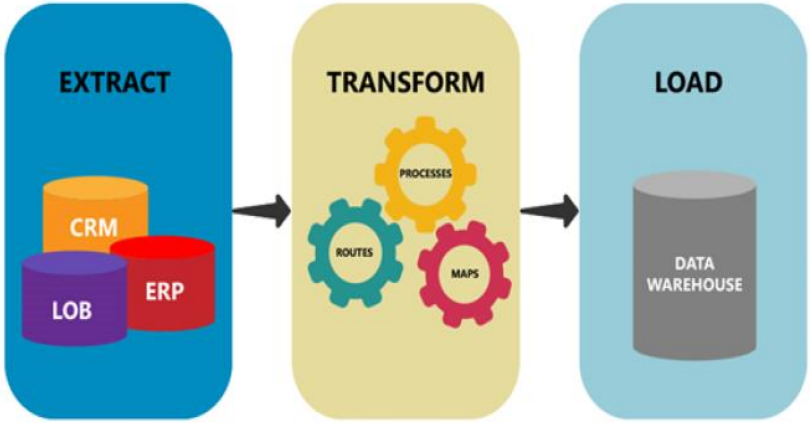


Which Data Extraction Approach is Best for Your Data Warehouse?, 2018
https://datawarehouseinfo.com/data-warehouse-data-extraction-models/

Data Sourcing (cont.)

Information Sourcing Patterns and Challenges

- Data conversion challenges
- Metadata gaps
- Mergers and acquisitions
- Manual data
- Real-time source data extrac

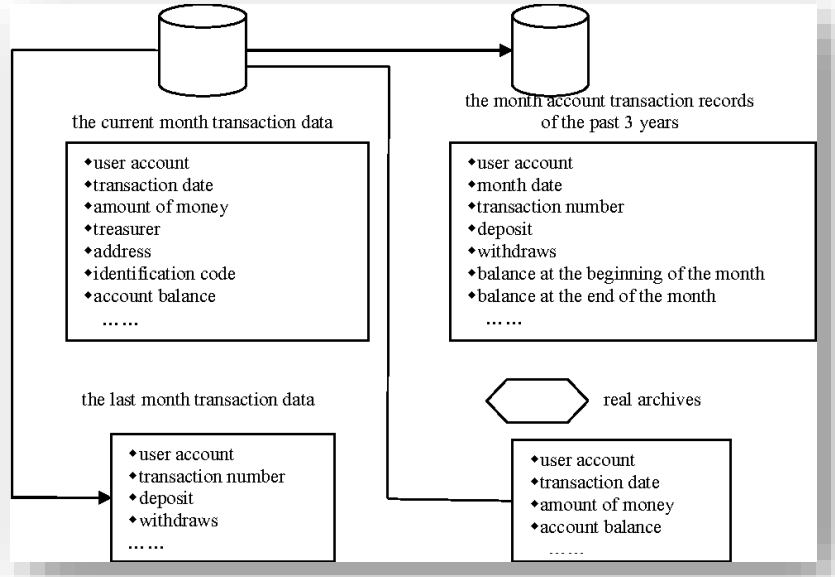


ETL vs. ELT: Transform First or Transform Later?, 2018
https://datawarehouseinfo.com/etl-vs-elt-transform-first-or-transform-later/

Data Sourcing (cont.)

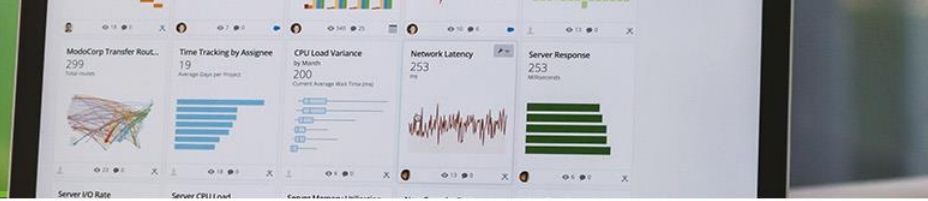
Data Granularity

- Data volumes and storage costs
- Query performance
- Source data availability
- Batch performance impact

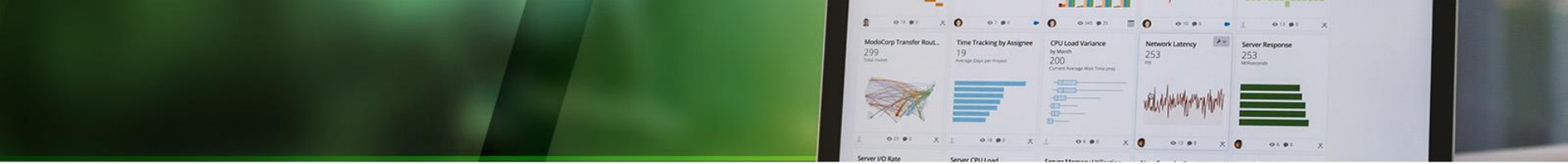


Classification of Data Granularity in Data Warehouse, 2017
https://www.semanticscholar.org/paper/Classification-of-Data-Granularity-in-Data-Lv-Zhou/aea746ba5fcdd504c51ace554dc343d55d2c024b

Activities 10 points



1. Categorize Qualitative and quantitative data. Work in groups. Take data online and show 5 examples each for both types of data.
2. Download Excel file from E-learning and using excel graphs, provide infographics on the data


ModoCorp Transfer Rout...
299
Time Tracking by Assignee
19
CPU Load Variance
200
Network Latency
253
Server Response
253
Server I/O Rate


Q
A