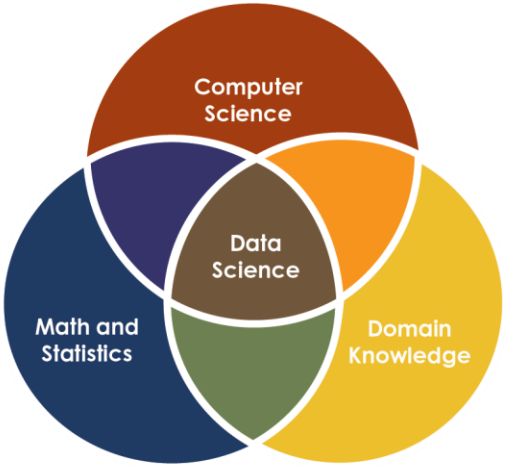


Lecture 1: Introduction to Data Science

NATIONAL INSTITUTE OF ELECTRONICS AND INFORMATION TECHNOLOGY, CHANDIGARH

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data

Lecture:1 Outline

- Data, Big Data and Challenges
- Data Science
  - Introduction
  - Why Data Science
- Data Scientists
  - What do they do?
- Major/Concentration in Data Science
  - What courses to take.

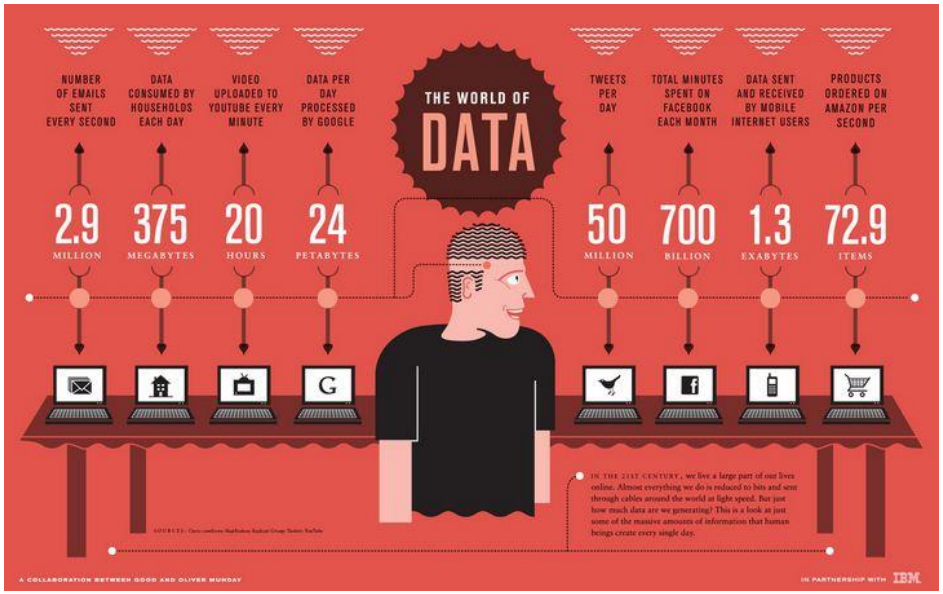
Data All Around

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network

How Much Data Do We have?

- Google processes 20 PB a day (2008)
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 1000 genomes project: 200 TB

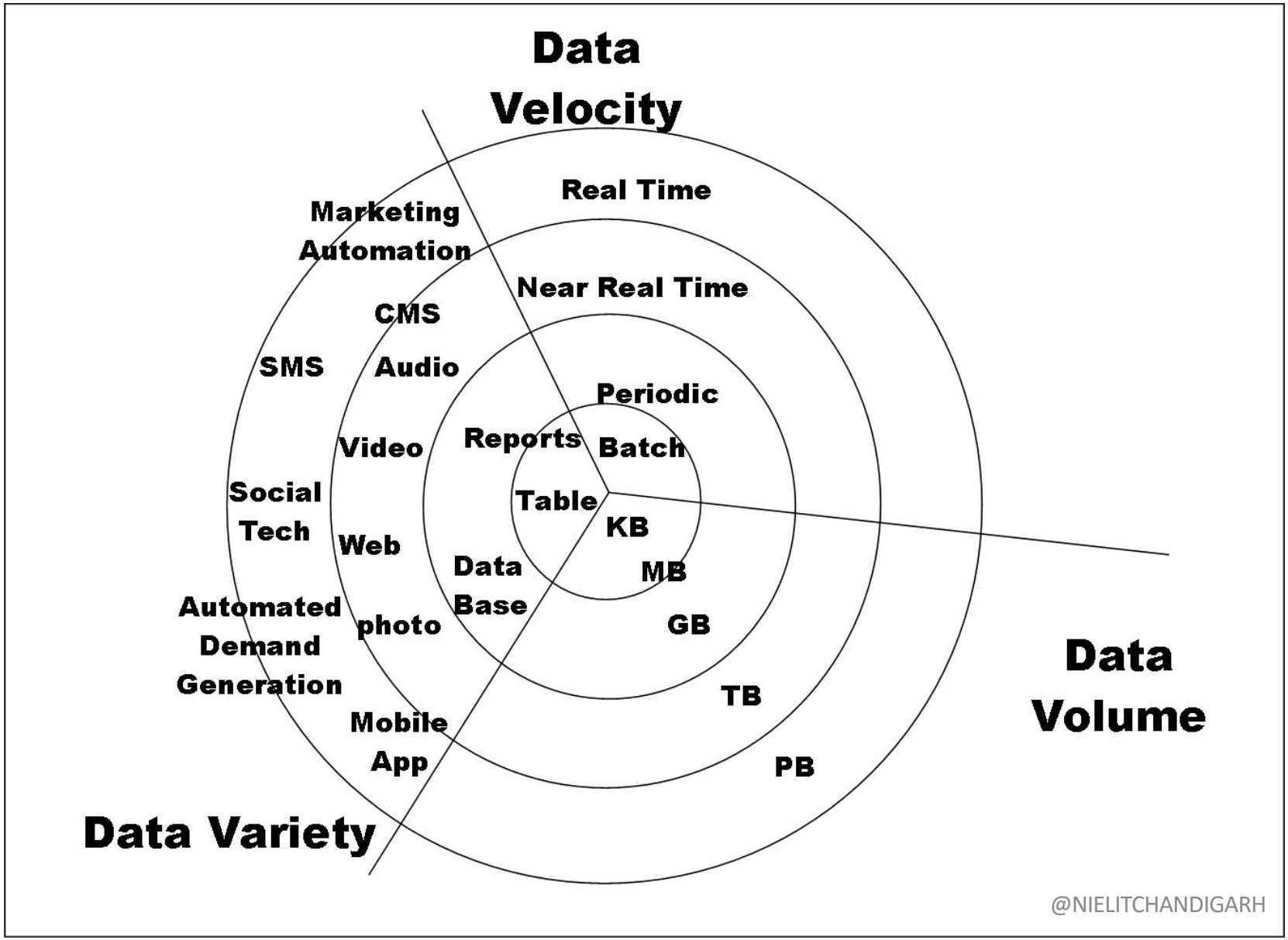
- Cost of 1 TB of disk: $35
- Time to read 1 TB disk: 3 hrs
  (100 MB/s)



Big Data

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume
    - The size of the data
  - Velocity
    - The latency of data processing relative to the growing demand for interactivity
  - Variety and Complexity
    - the diversity of sources, formats, quality, structures.

Big Data


Data
Velocity
Real Time
Near Real Time
Periodic
Batch
Marketing
Automation
CMS
SMS
Audio
Video
Reports
Table
Social
Tech
Web
KB
MB
Data
Base
GB
Automated
Demand
Generation
photo
TB
Mobile
App
PB
Data
Volume
Data Variety
@NIELITCHANDIGARH

Types of Data We Have

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- You can afford to scan the data once

What To Do With These Data?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

Big Data and Data Science

- "… the Hottest job in the next 10 years will be statisticians," Hal Varian, Google Chief Economist
- The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute's June 2011
- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
  - One proposal (elsewhere) for an MS in "Big Data Science"

What is Data Science?

- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

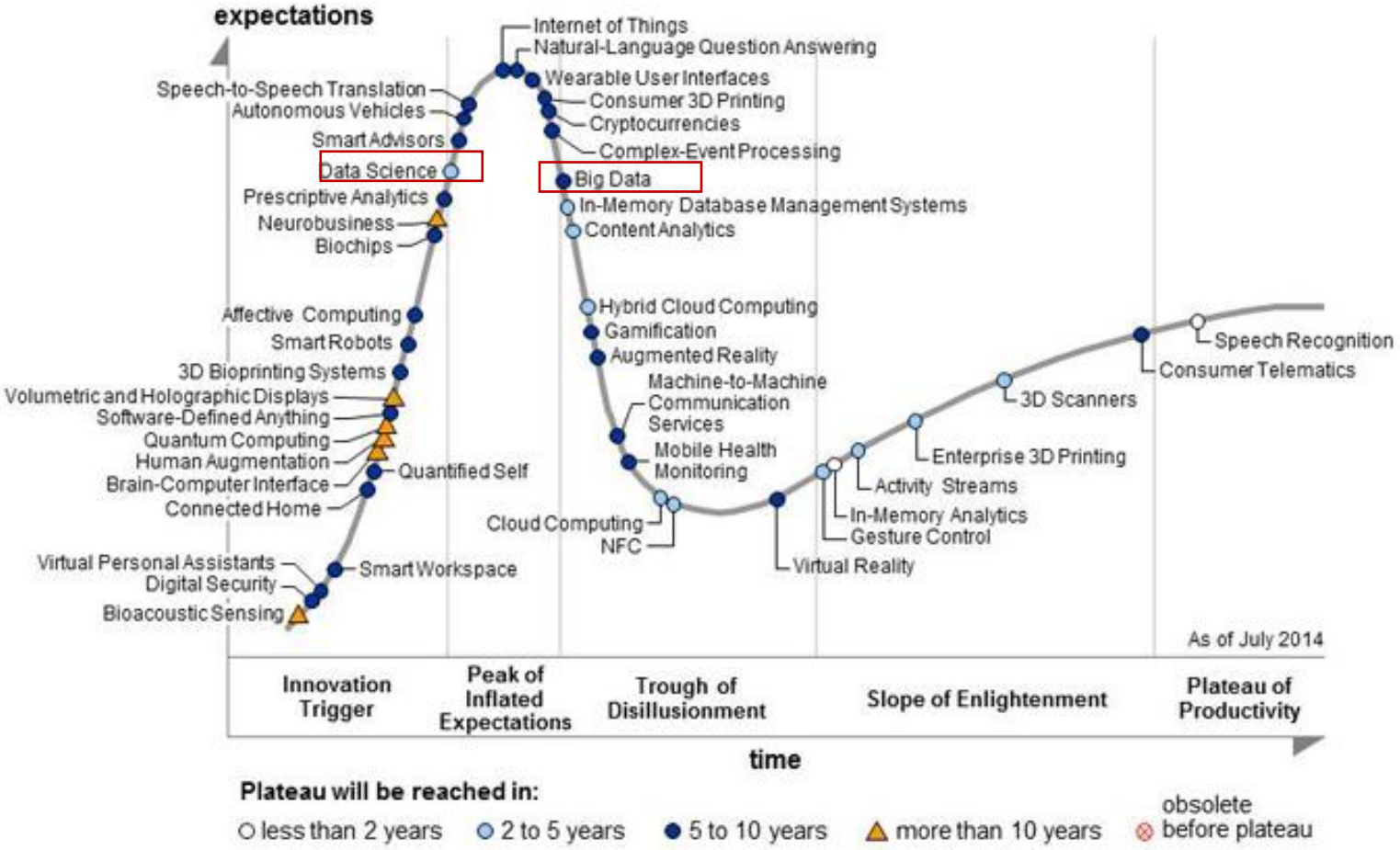
https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

What is Data Science?

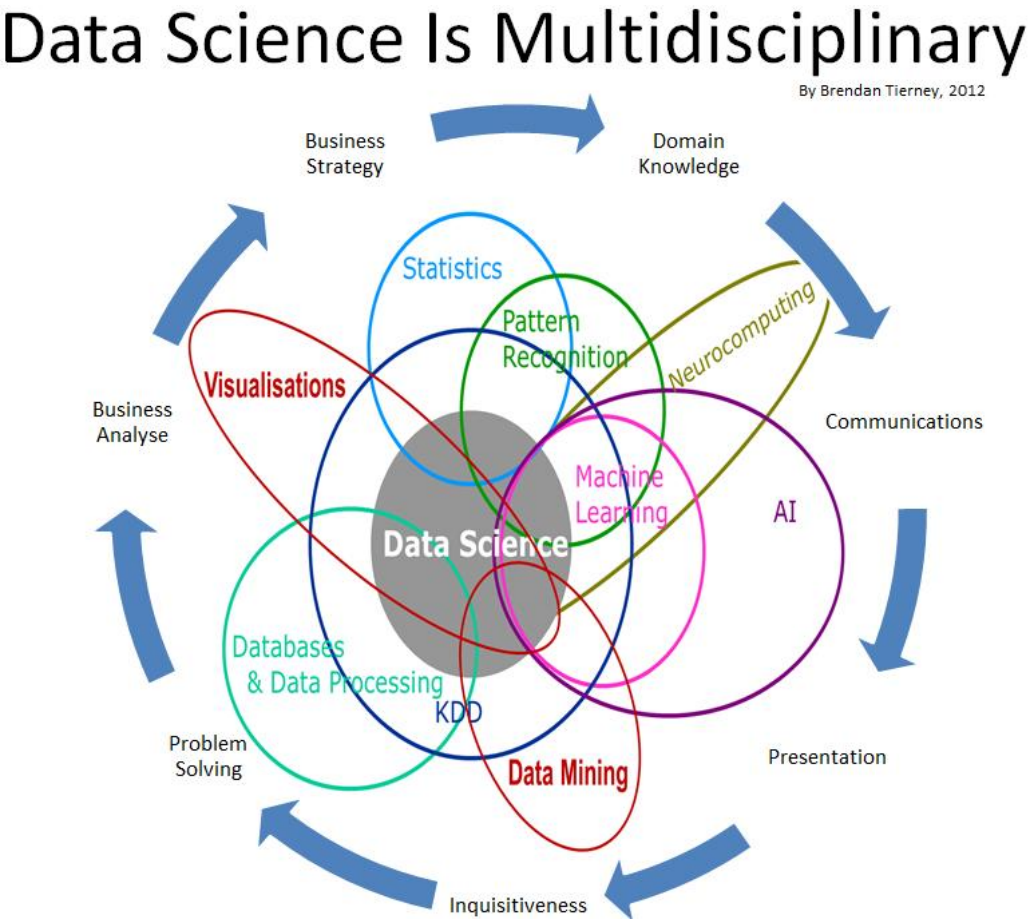
- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.

Why is it HOT?

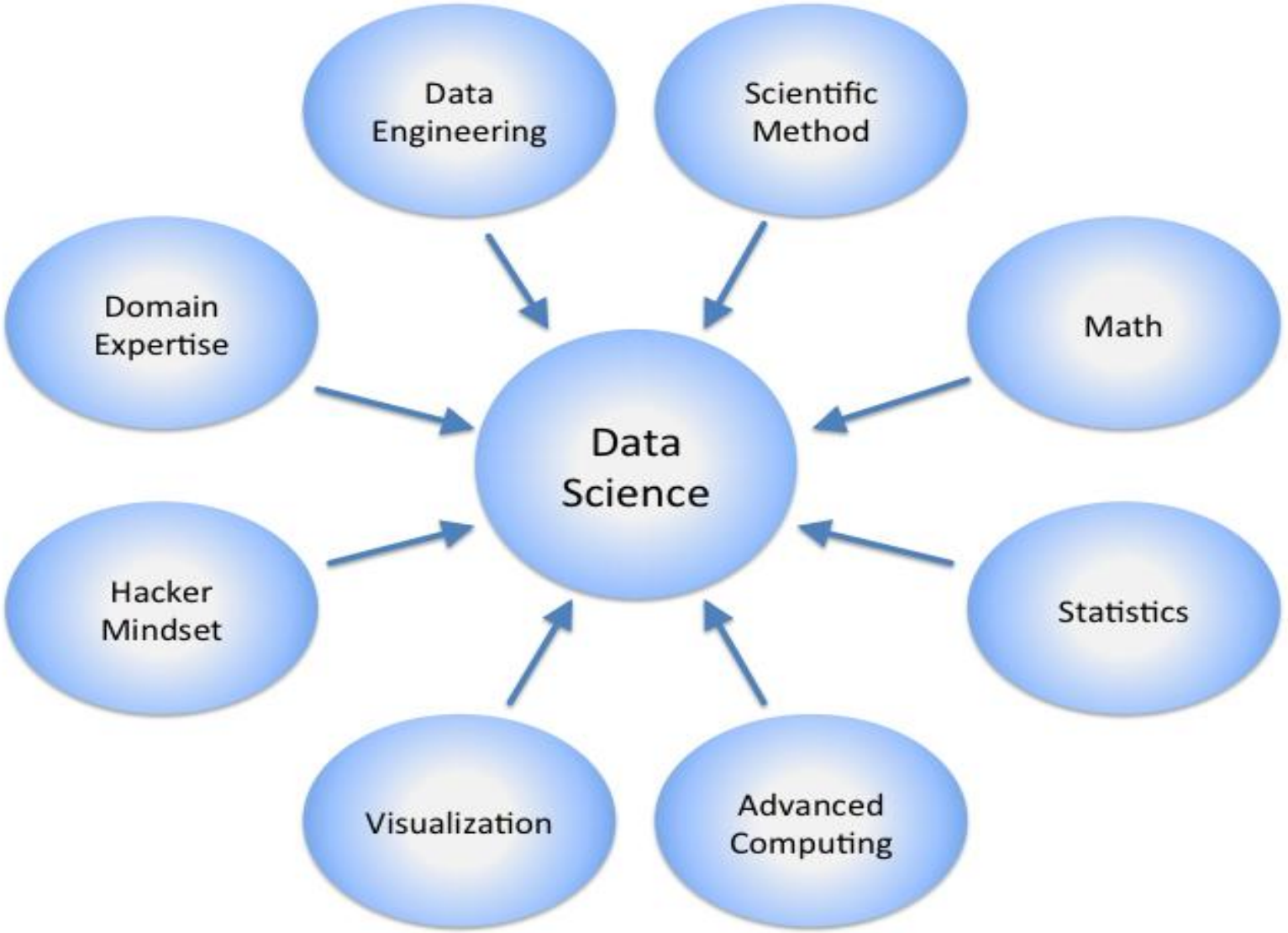
- Gartner's 2014 Hype Cycle



Data Science



Data Science

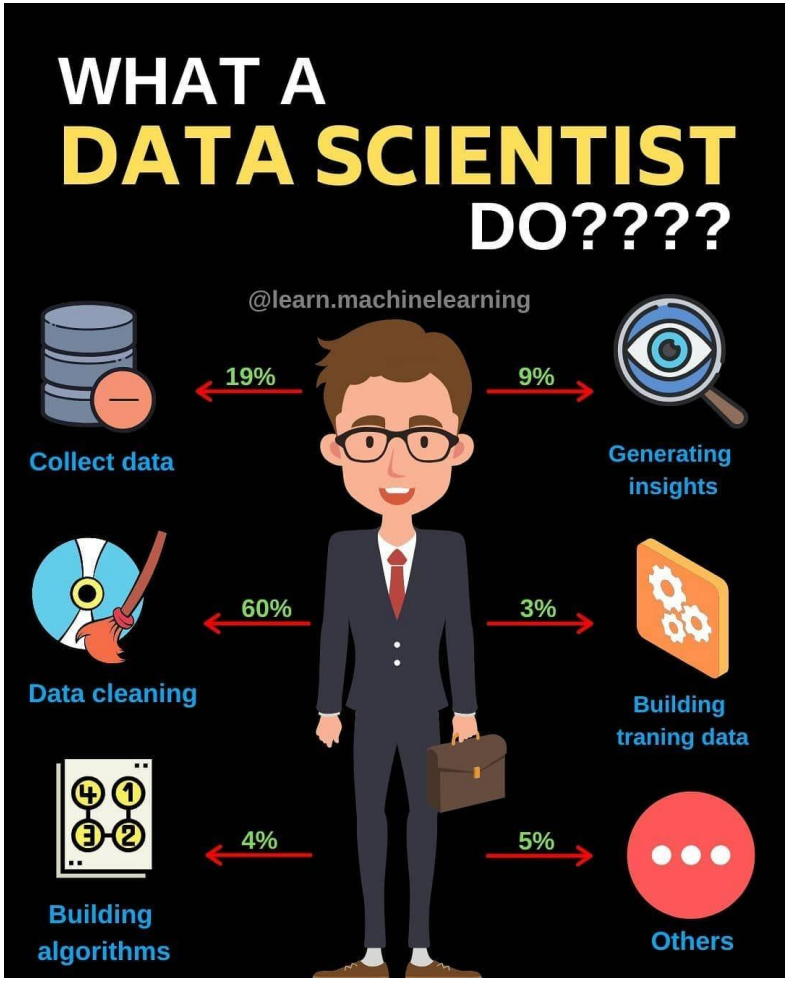


Real Life Examples

- Companies learn your secrets, shopping patterns, and preferences
  - For example, can we know if a woman is pregnant, even if she doesn't want us to know? Target case study
- Data Science and election (2008, 2012)
  - 1 million people installed the Obama Facebook app that gave access to info on "friends"

Data Scientists

- Data Scientist
  - The Prestigious Job of the 21st Century
- They find stories, extract knowledge. They are not reporters



Data Scientists

- Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions

What do Data Scientists do?

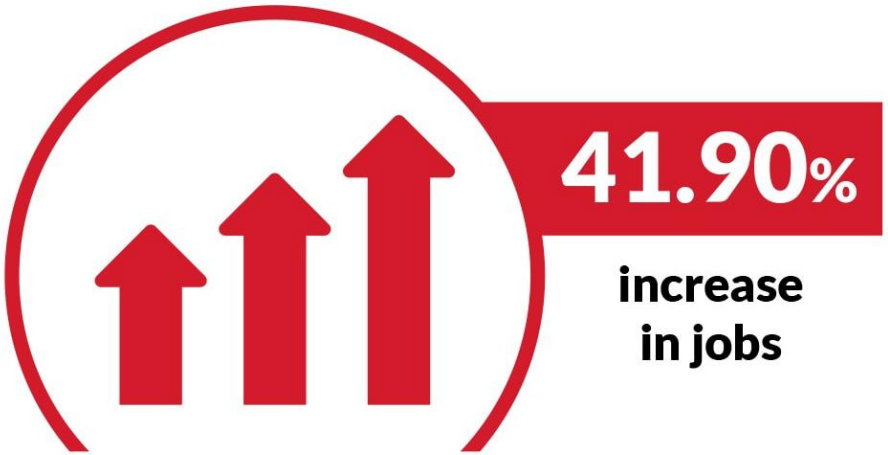
- National Security
- Cyber Security
- Business Analytics
- Engineering
- Healthcare
- And more ….

Concentration in Data Science

- Mathematics and Applied Mathematics
- Applied Statistics/Data Analysis
- Solid Programming Skills (R, Python, Julia, SQL)
- Data Mining
- Data Base Storage and Management
- Machine Learning and discovery

DATA SCIENCE & MACHINE LEARNING

LECTURE 2: DS & ML



NATIONAL INSTITUTE OF ELECTRONICS AND INFORMATION TECHNOLOGY, CHANDIGARH

रा.इ.सू.प्रौ.सं
NIELIT

@NIELITCHANDIGARH

WHAT IS DATA SCIENCE?

- “Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.”

- “Data science intends to analyze and understand actual phenomena with ‘data’. In other words, the aim of data science is to reveal the features or the hidden structure of complicated natural, human, and social phenomena with data from a different point of view from the established or traditional theory and method.”

Source: Wikipedia

WHAT IS DATA SCIENCE?

...solving problems with data...

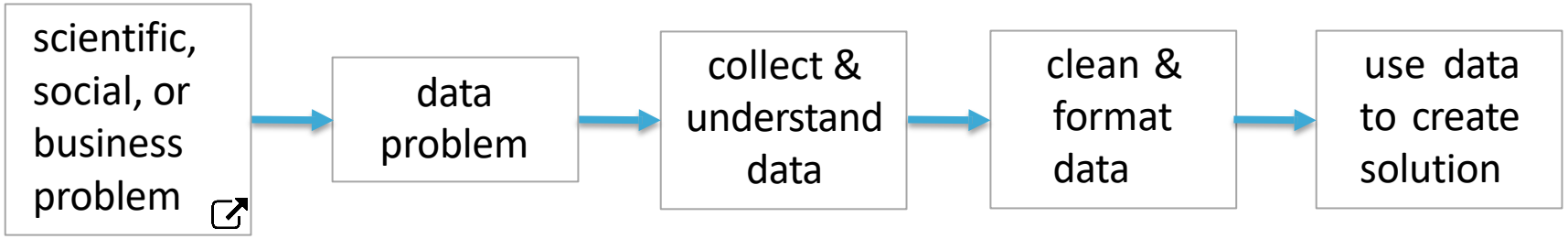
...sounds cool!

What makes a good data scientist?

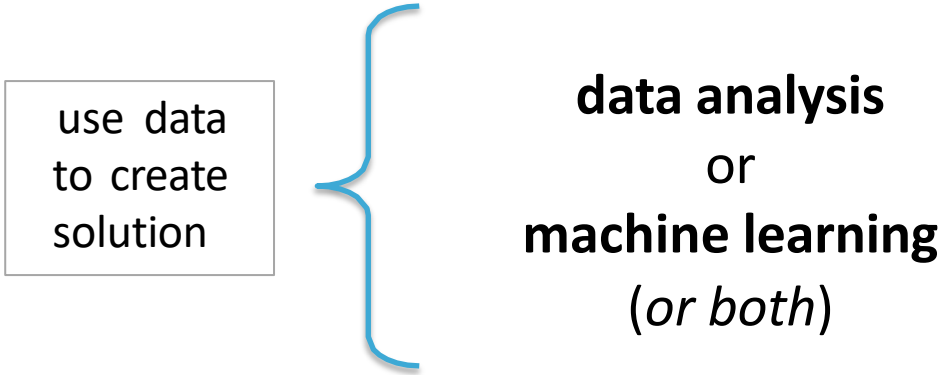
[PDSH] p xi
[DSFS] p 1-2

WHAT IS DATA SCIENCE?

…solving problems with data…



…which step is most challenging?

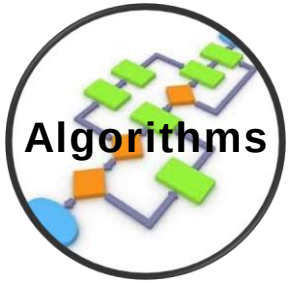


WHAT IS DATA ANALYSIS?

…using data to discover useful information…

- **data**: anything you can *measure* or *record*

- **statistics**: summarize (and visualize) *main characteristics* of the data

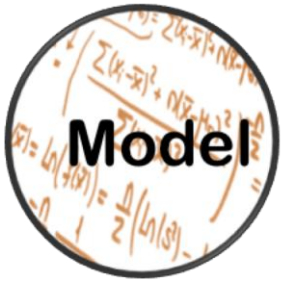


- **algorithms**: apply algorithms to find *patterns* in the data

WHAT IS MACHINE LEARNING?

…creating and using models that learn from data…

- **data**: anything you can *measure* or *record*

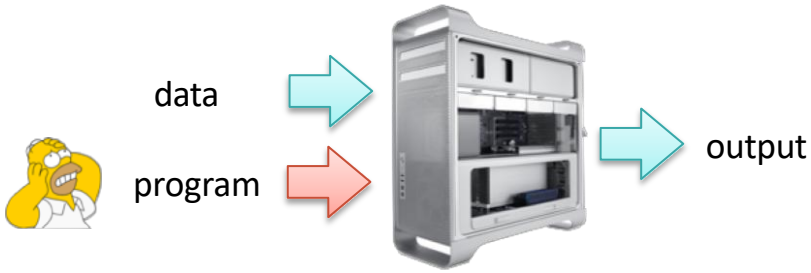


- **model**: specification of a (mathematical) *relationship* between different variables

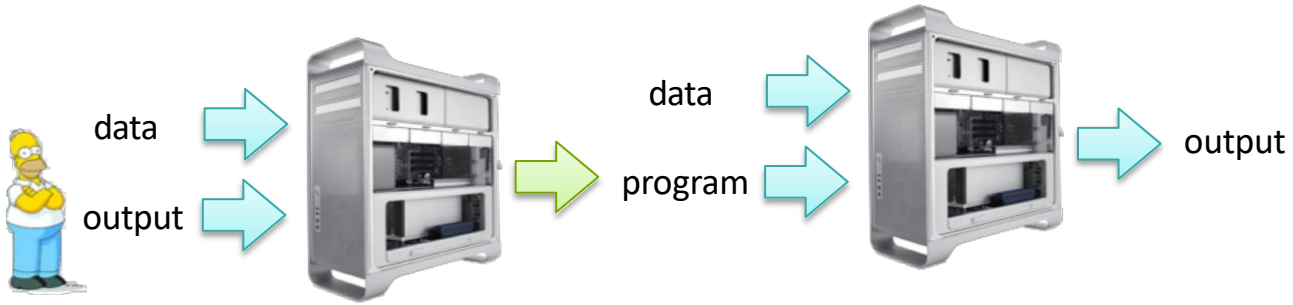
- **evaluation**: how well does the model *work*?

WHAT IS MACHINE LEARNING?

- Traditional CS



- Machine Learning

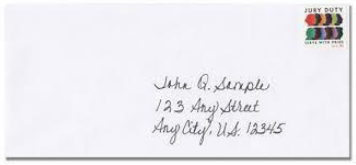


WHAT IS MACHINE LEARNING?

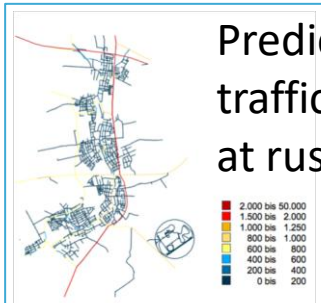
- …creating and using models that learn from data…

- Examples

Identifying zip code from handwritten digits



Detecting communities in social networks



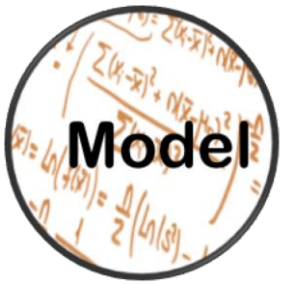
Predicting the traffic volume at rush hour

Detecting fraudulent credit card transactions

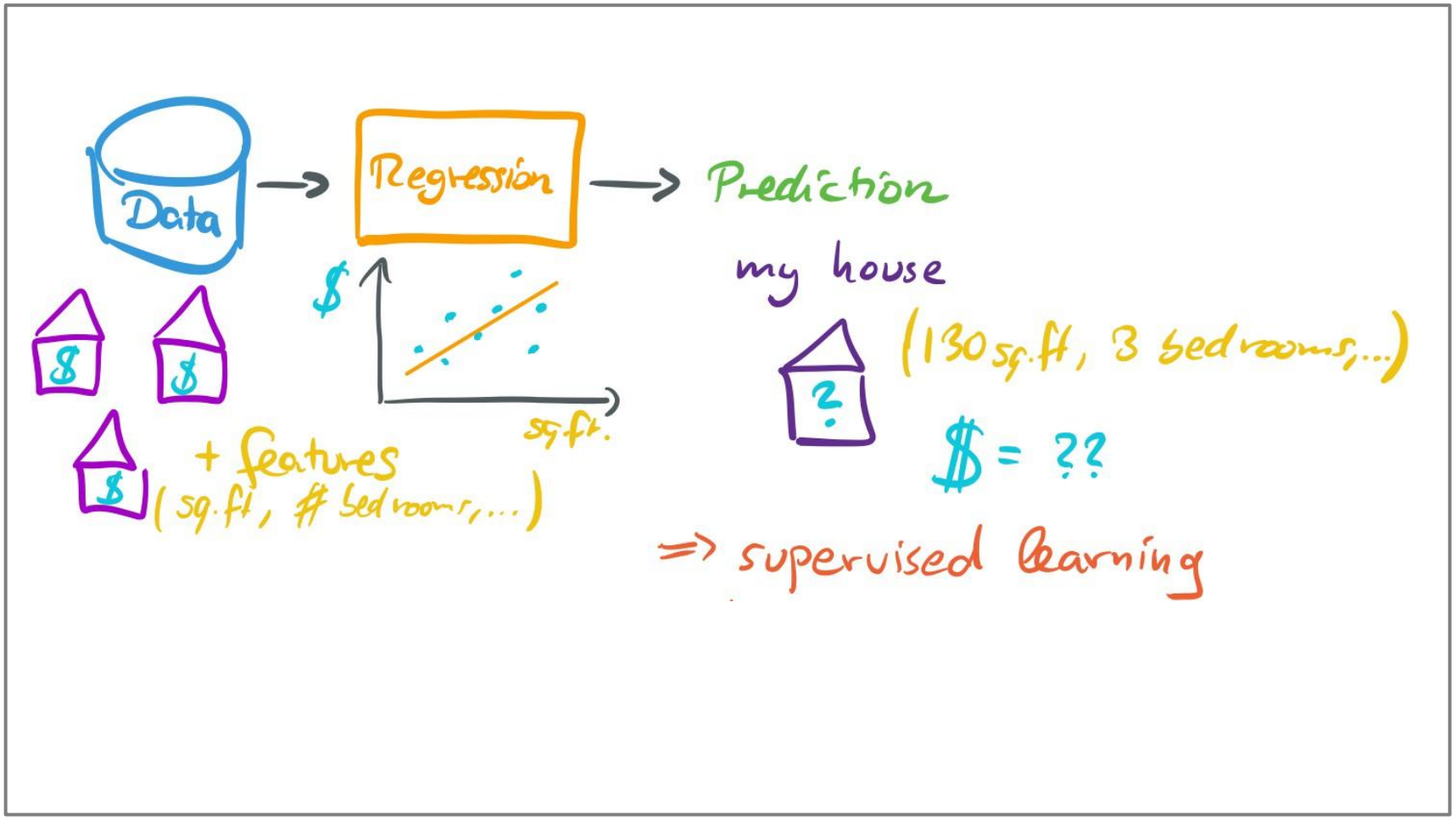
Determining the location of distribution centers based on customers' residence

[DSFS] p 3-13

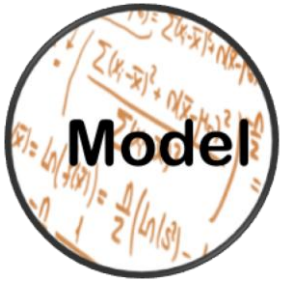
LEARNING FROM DATA



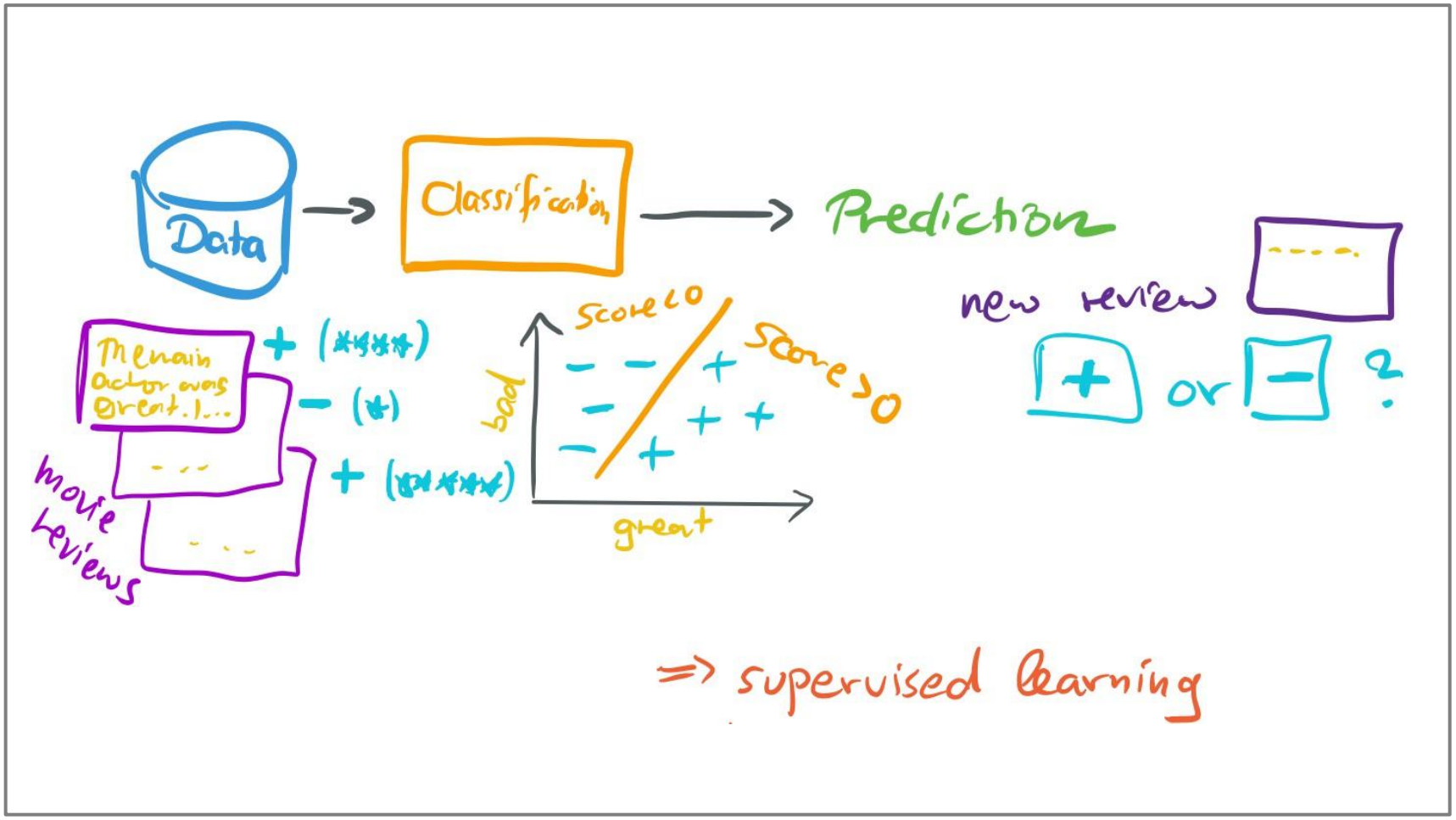
- Regression



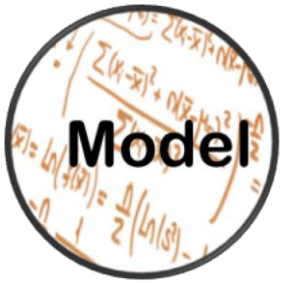
LEARNING FROM DATA



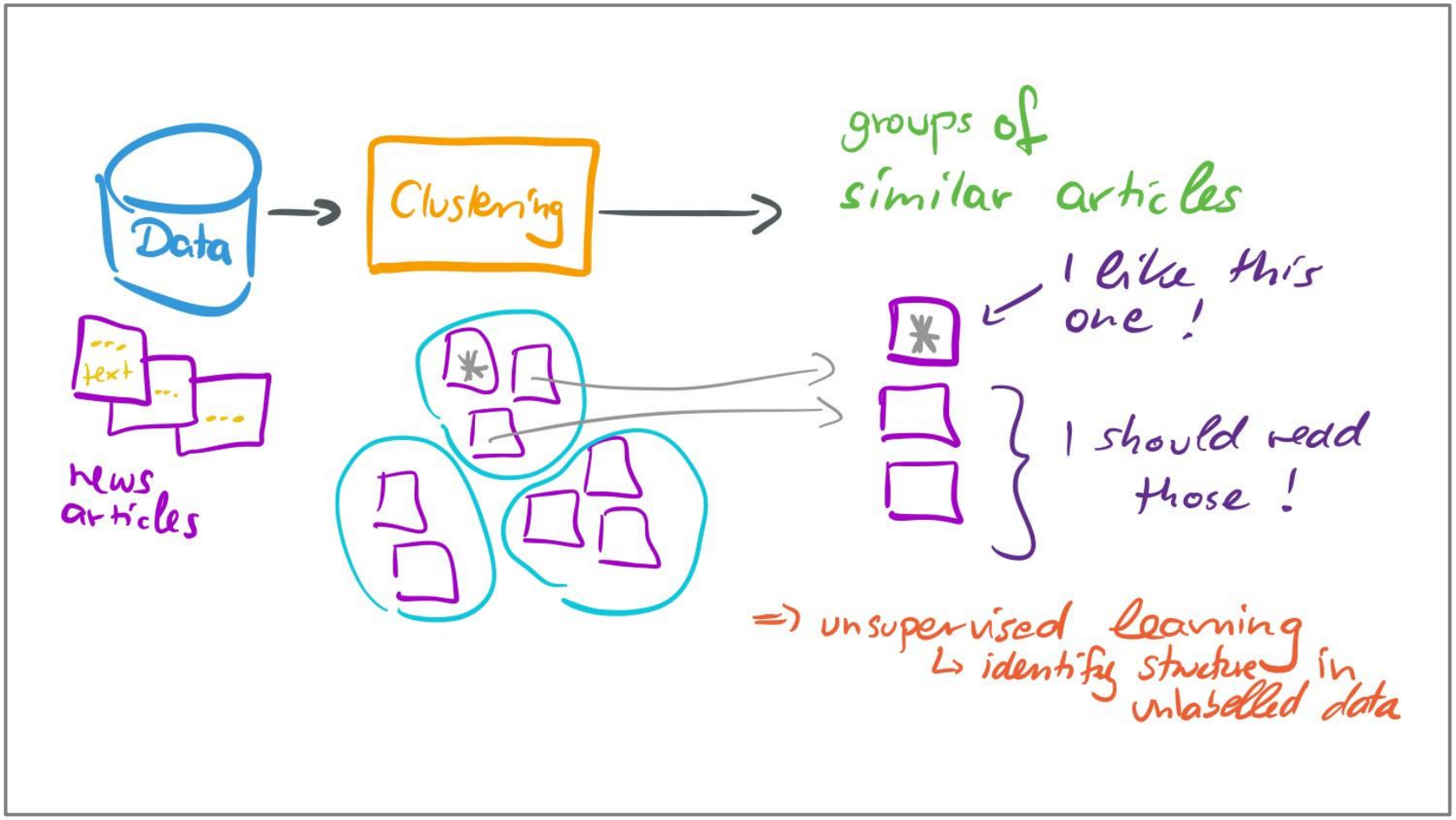
- Classification



LEARNING FROM DATA



- Clustering



WHAT IS **MACHINE LEARNING**?

regression, classification, clustering

…creating and using models that learn from data…

→ **supervised learning**/predictive modelling

- come up with predictions
- extract knowledge/insights

→ **unsupervised learning**/data mining

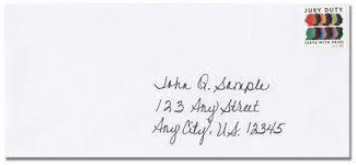
[PDSH] p 332-342
[DSFS] p 141-142

ACTIVITY 1

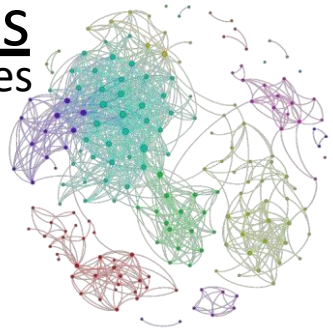
regression, classification, clustering

- …creating and using models that learn from data…
- Categorize these Examples

Identifying zip code from handwritten digits



Detecting communities in social networks



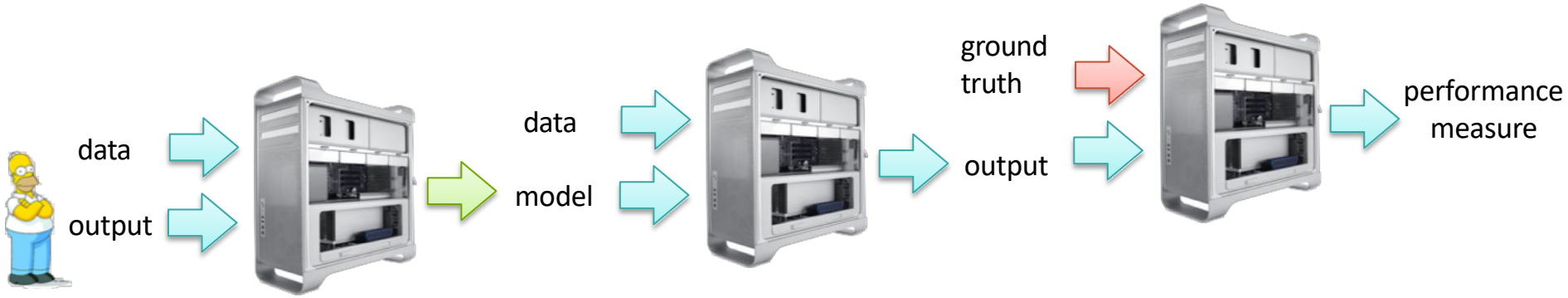
Predicting the traffic volume at rush hour

Detecting fraudulent credit card transactions

Determining the location of distribution centers based on customers' residence

MACHINE LEARNING WORKFLOW

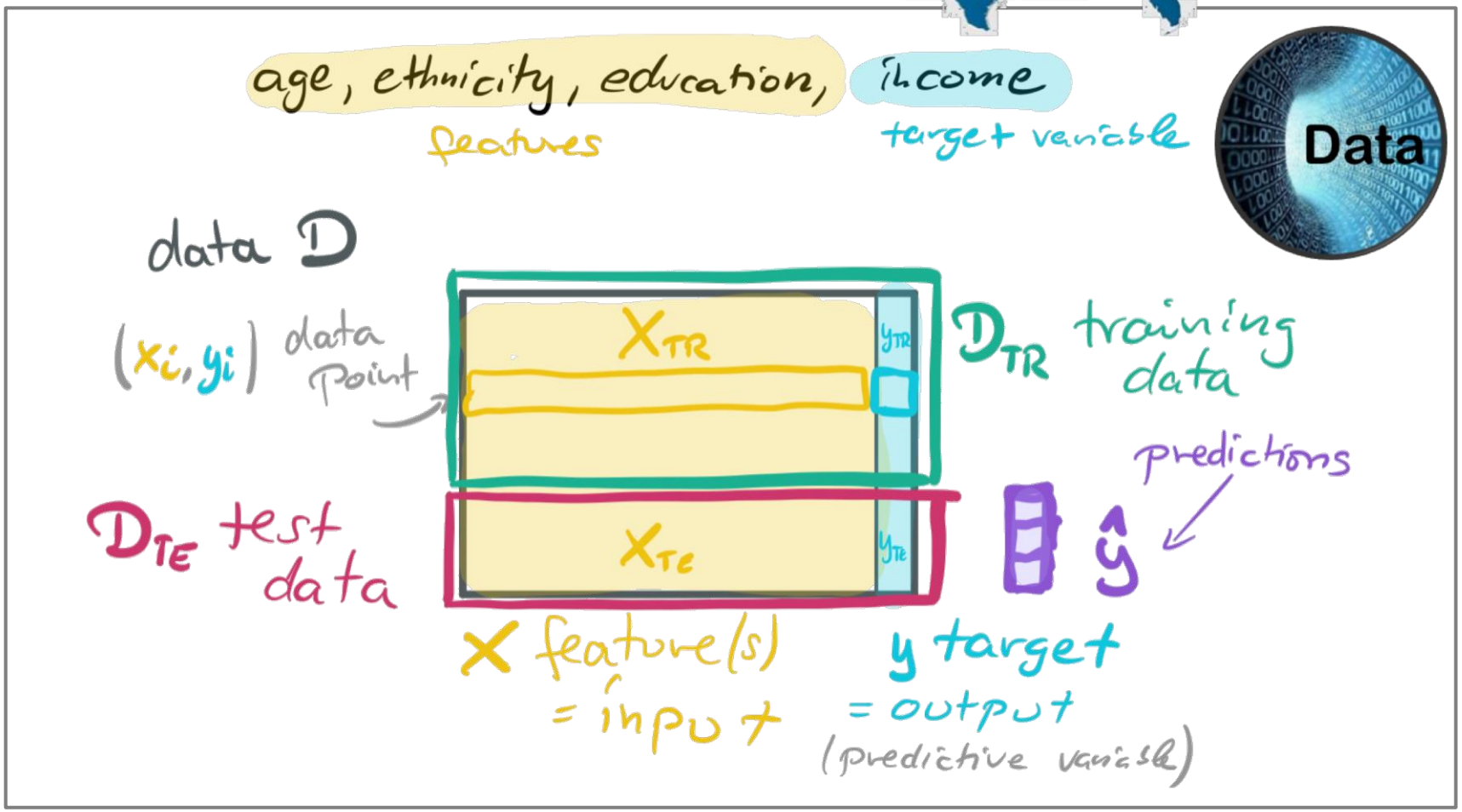
- *training phase*, *test phase*, *evaluation phase*



→ let's have a closer look at the *data* we are using

ACTIVITY 2

- Example: Census Data



- *training data* and *test data*

DATA

- Notation:
  - D all observed data
  - X all features
  - y observations
  - $\square_{TE}$ test
  - $\square_{TR}$ training
  - $\hat{y}$ predictions

Helper Notation:

n number of data points

d number of features

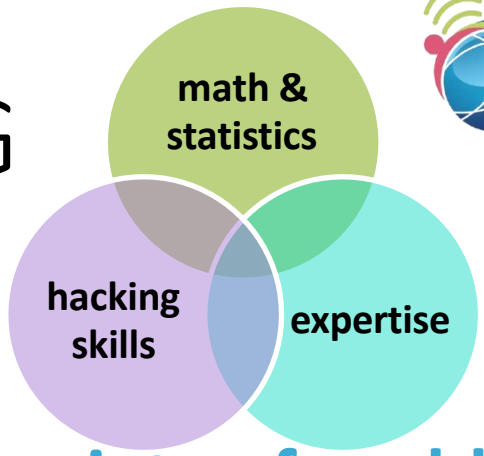
m number of training points

$\square_{1,\ldots,i,\ldots,n}$: indices for data points

$\square_{1,\ldots,j,\ldots,d}$: indices for features

- What data structure to use?
  - *set*, *list*, or *array*?

SUMMARY & READING



- *Data Science* is about

data, **models**, and **evaluation**

- *Data Science* can solve a wide **variety of problems** – once we have the *right* data *and* model!