INTRODUCTION TO DATA MINING

BEIYU LIN

COURSE INFORMATION

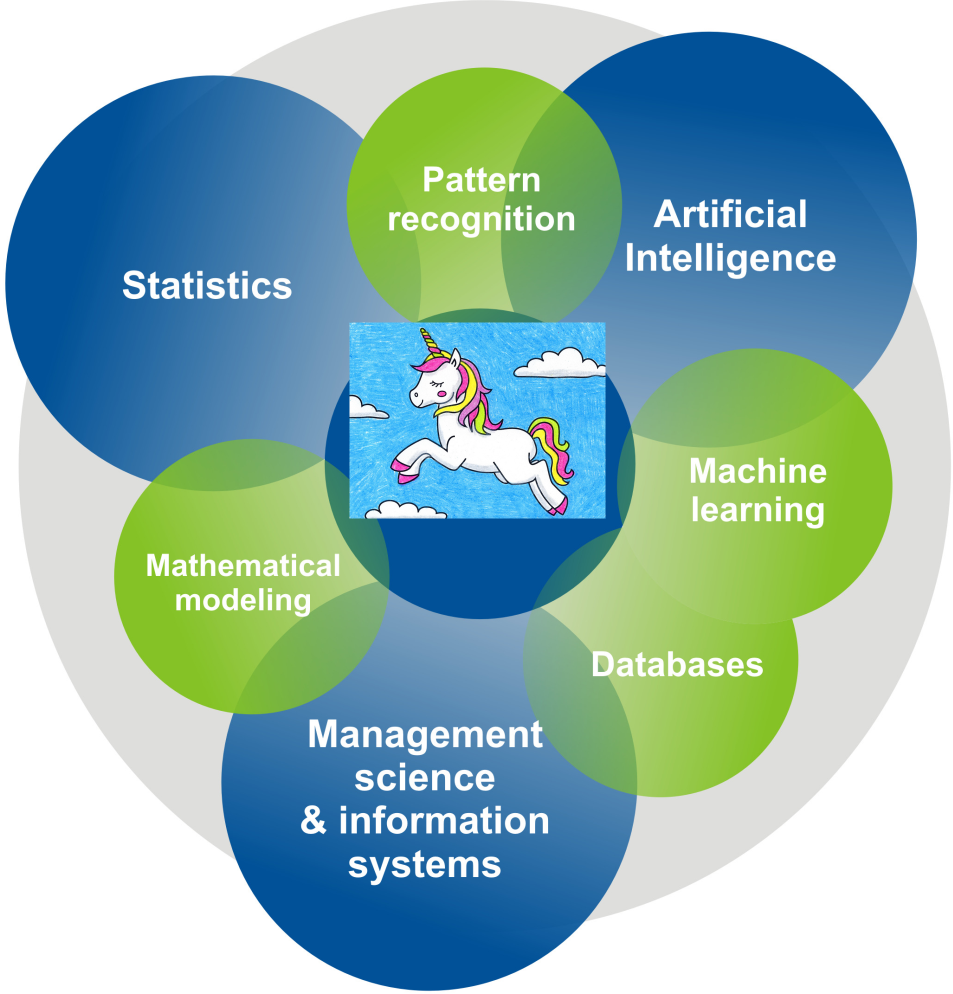
Outline: https://beiyulincs.github.io/teach/fall_21/dm.html

Syllabus: https://beiyulincs.github.io/teach/fall_21/syllabus_cs_458.pdf

OUTLINE

- **What is Data Mining?**
- **Why Data Learning is important?**
- **Data Learning and its Applications**
- **Real Life Examples**

DATA MINING


Pattern recognition
Artificial Intelligence
Statistics
Machine learning
Mathematical modeling
Databases
Management science & information systems

WHAT IS DATA MINING

The process of discovering meaningful **new**

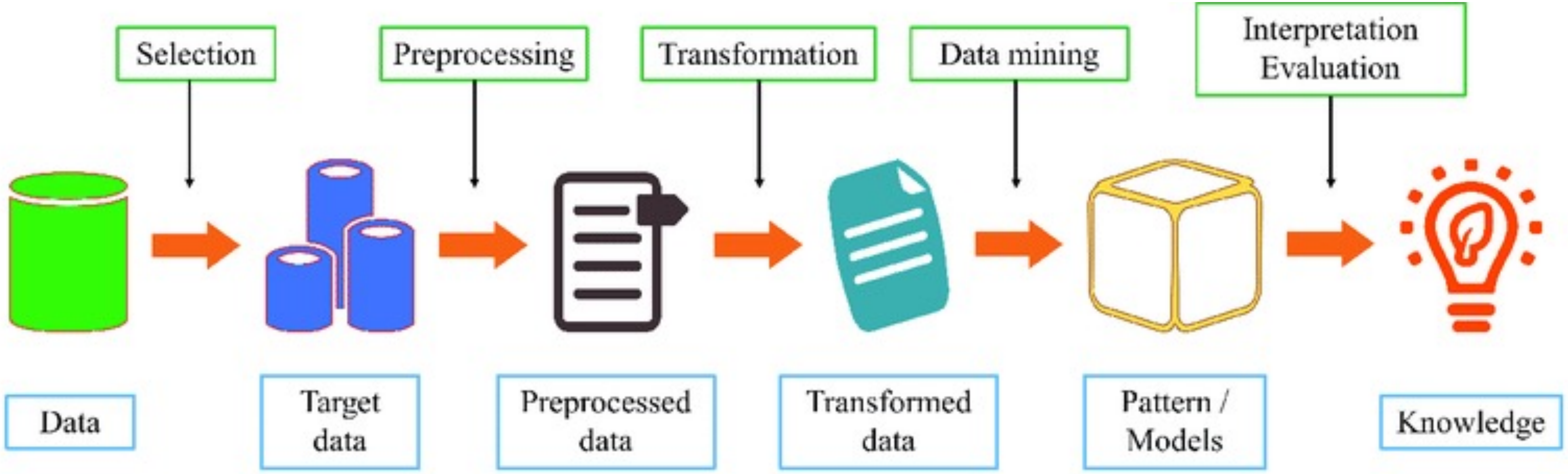
- correlations
- Patterns
- Trends

By learning from large amounts of stored data

Via

- Pattern recognition
- Statistical
- Mathematical
- Machine learning methods

DATA MINING PROCESS



WHAT IS DATA MINING

- Data mining (knowledge discovery from data (KDD))
  - – Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - – Data mining: a misnomer (outliner)?
- • Alternative names
  - – Knowledge discovery in data,
  - – knowledge extraction,
  - – data/pattern analysis,
  - – data archeology,
  - – data dredging, information harvesting, business intelligence, etc.

DATA MINING IN REAL LIFE

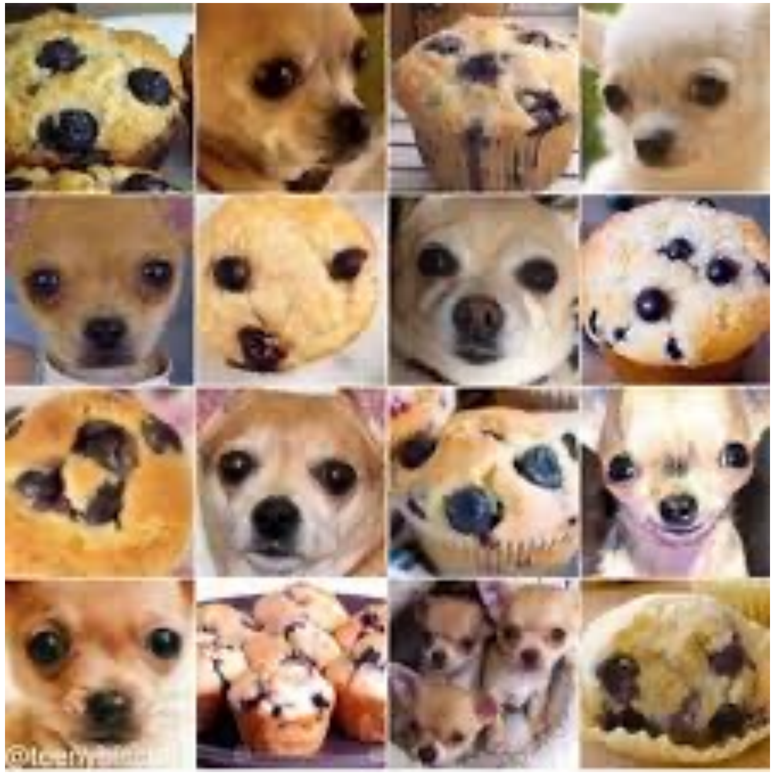


Image Classification

DATA MINING IN REAL LIFE

Document Categorization

Sign In Help FAQs
SportsNews
THE LATEST NEWS IN THE WORLD OF SPORTS
News | Weather | Travel | TV | Radio
Search
NFL MLB NBA NHL NCAA FB NCAA BB NASCAR SOCCER
Scores | Schedule | Standings | Stats | Power | Rankings | Teams | Players | Rumors | Transactions
READ MORE

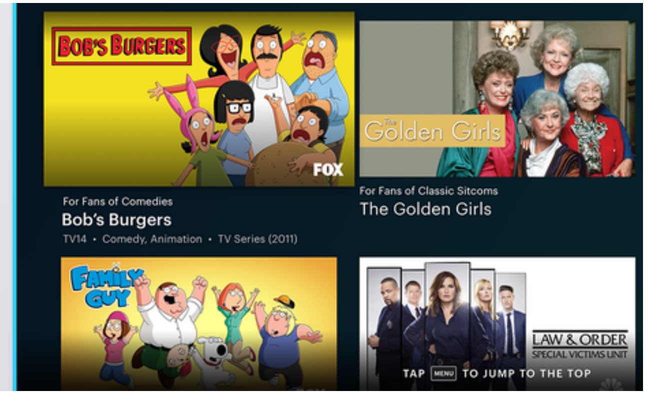
2011 Senior Bowl: January 27, 2011

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

2011 Senior Bowl: January 27, 2011

Who Can Dethrone The Yankees?

Faltering Flyers slip on fallen Leafs


BOB'S BURGERS
FOX
The Golden Girls
For Fans of Comedies
Bob's Burgers
TV14 • Comedy, Animation • TV Series (2011)
For Fans of Classic Sitcoms
The Golden Girls
FAMILY GUY
LAW & ORDER SPECIAL VICTIMS UNIT
TAP MENU TO JUMP TO THE TOP

Speech Recognition

Protein Classification

Fraud Detection

Playing Games

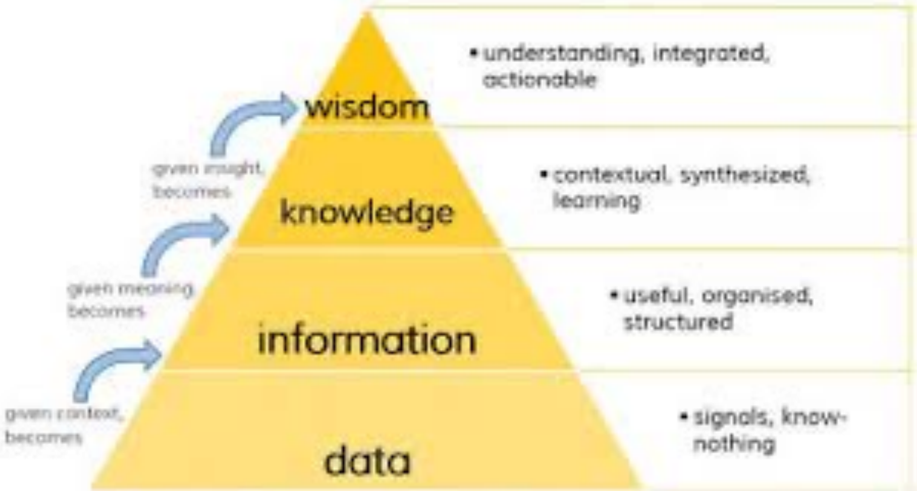
Spam Detection

WHY DATA MINING

"The world is one big data problem."
(by Andrew McAfee, co-director of the MIT Initiative)

"Data is the new science. Big Data holds the answers." (Pat Gelsinger, CEO, VMWare)

transforming raw data into useful knowledge



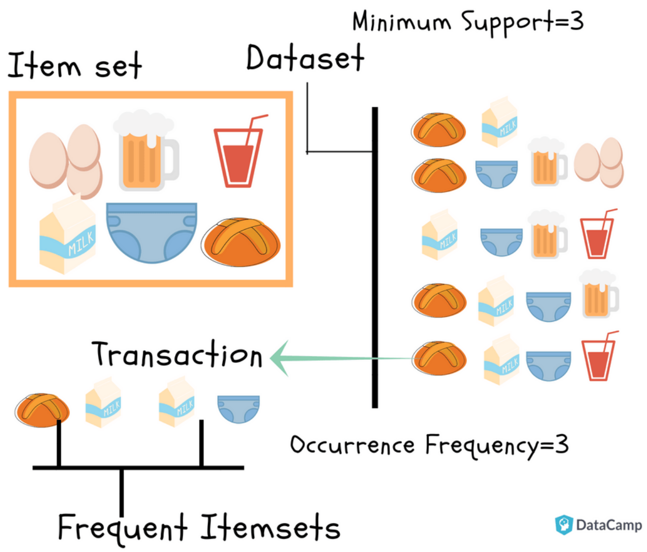
WHY DATA MINING

- Data analysis and decision support
  - – Market analysis and management
    - • Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - – Risk analysis and management
    - • Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - – Fraud detection and detection of unusual patterns (outliers)

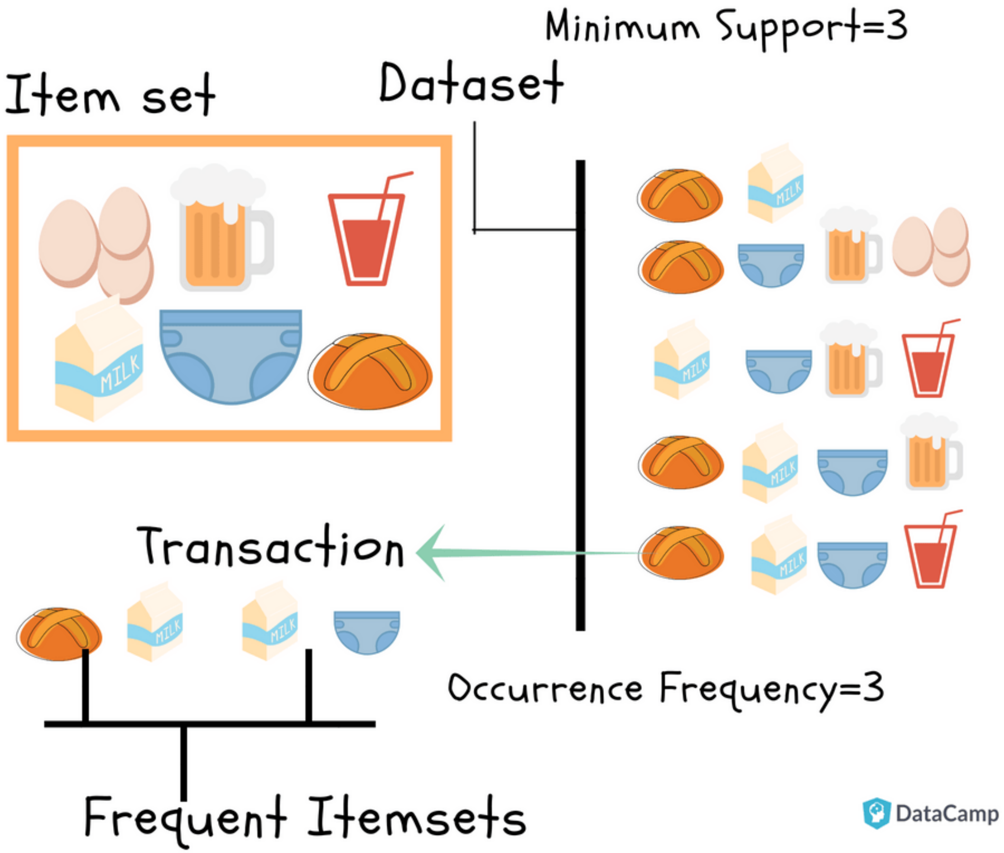
- Other Applications
  - – Text mining (news group, email, documents) and Web mining
  - – Stream data mining
  - – DNA and bio-data analysis

DATA MINING ALGORITHMS

- Association rule
- Categorization (supervised learning)
- Clustering (unsupervised learning)
- Mining Internet of Things (IoT) data



ASSOCIATION RULE



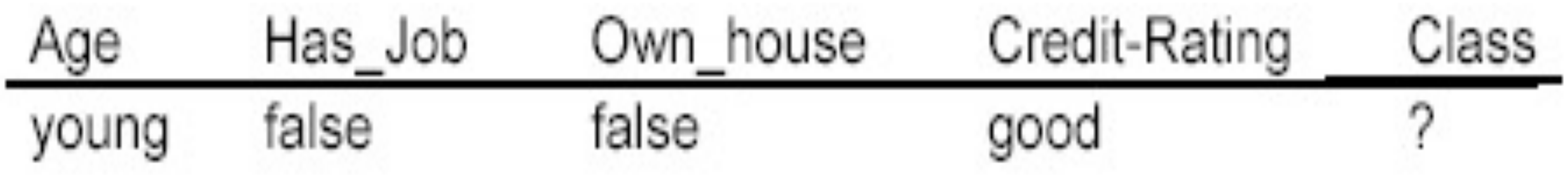
SUPERVISED CLASSIFICATION

Decide whom credit card application should be approved.

Goal: use a person’s information seen so far to produce good prediction rule for future applications.

MODELS – SUPERVISED LEARNING

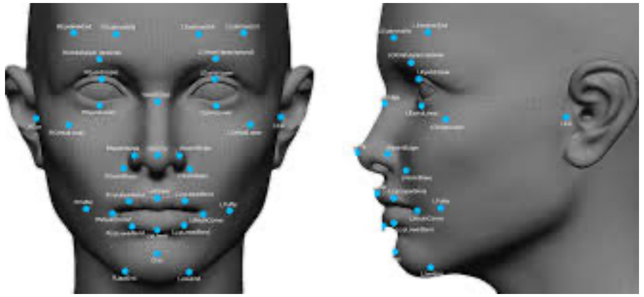
- Learn a classification model from the data
- Use the model to classify future loan applications into
  - Yes (approved) and
  - No (not approved)
- What is the class for following case/instance?

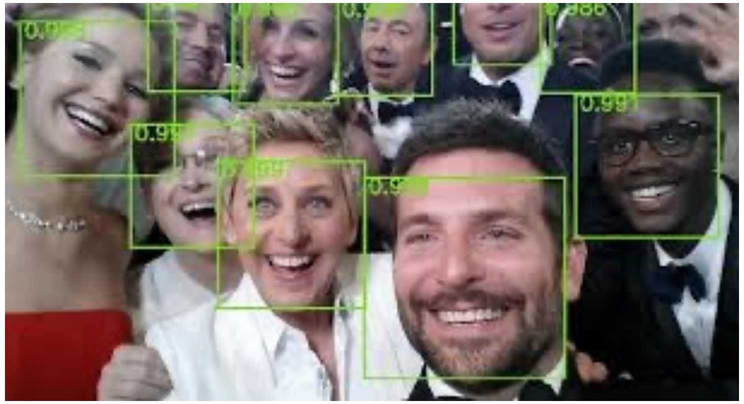


Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?

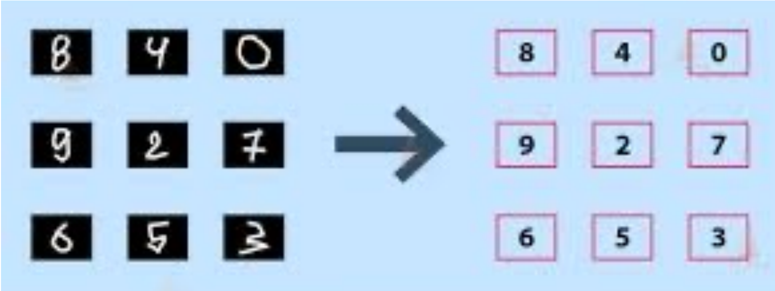
MODELS – SUPERVISED LEARNING (IMAGE CLASSIFICATION)

Face Detection and Recognition



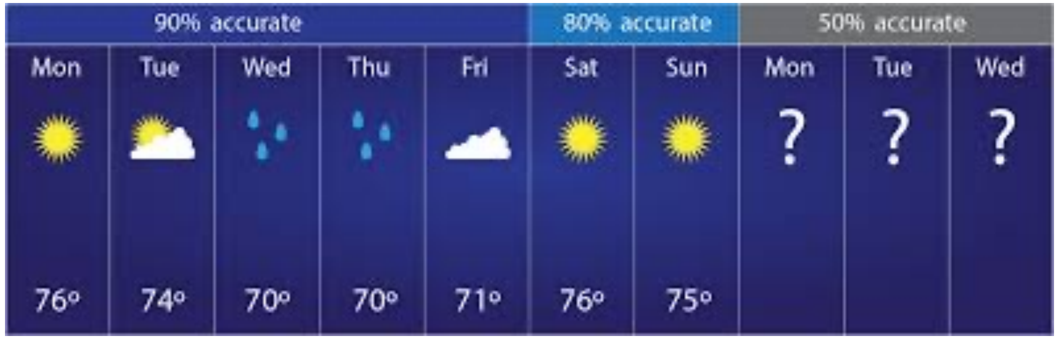


Handwritten digit recognition (convert hand-written digits to characters 0..9)



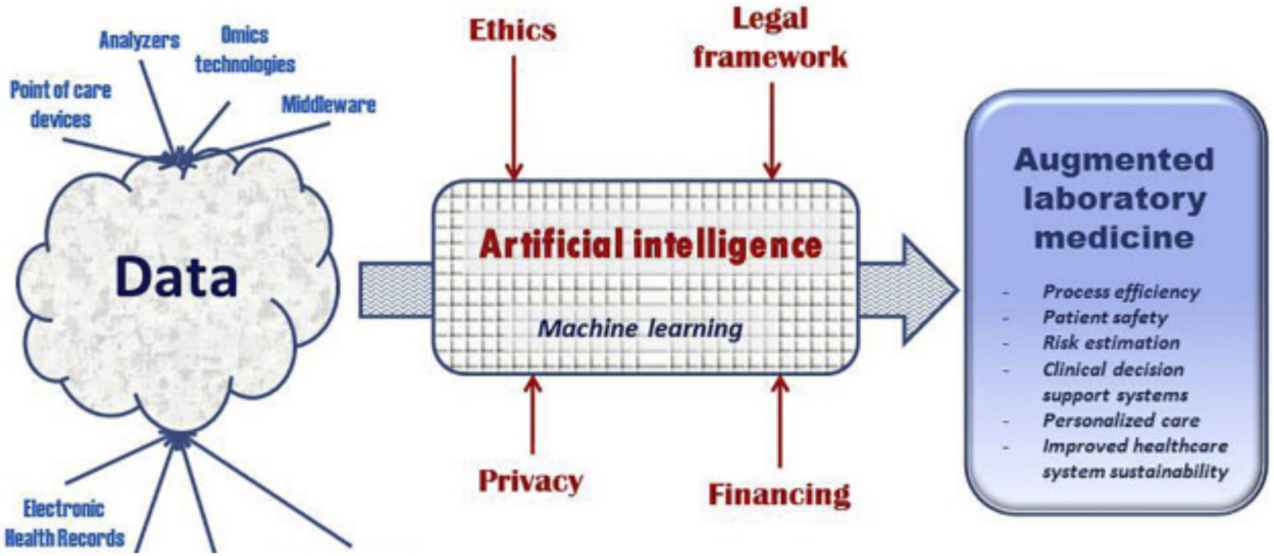
MODELS – SUPERVISED LEARNING (OTHER EXAMPLES)

Weather Prediction



90% accurate					80% accurate		50% accurate		
Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed
☀	⛅	💧	💧	☁	☀	☀	?	?	?
76°	74°	70°	70°	71°	76°	75°			

Medicine



Analyzers

Omics technologies

Point of care devices

Middleware

Data

Electronic Health Records

Ethics

Legal framework

Artificial intelligence

Machine learning

Privacy

Financing

Augmented laboratory medicine

- Process efficiency
- Patient safety
- Risk estimation
- Clinical decision support systems
- Personalized care
- Improved healthcare system sustainability

Computational Economics:

– predict if a stock will rise or fall

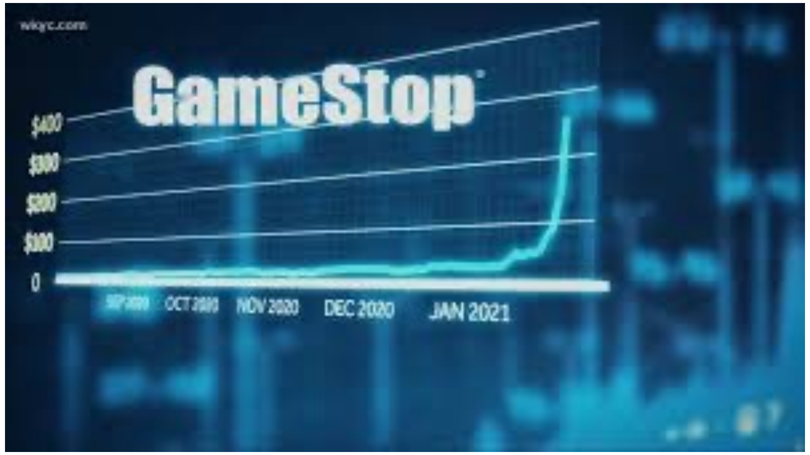
– predict if a user will click on an ad or not

- in order to decide which ad to show

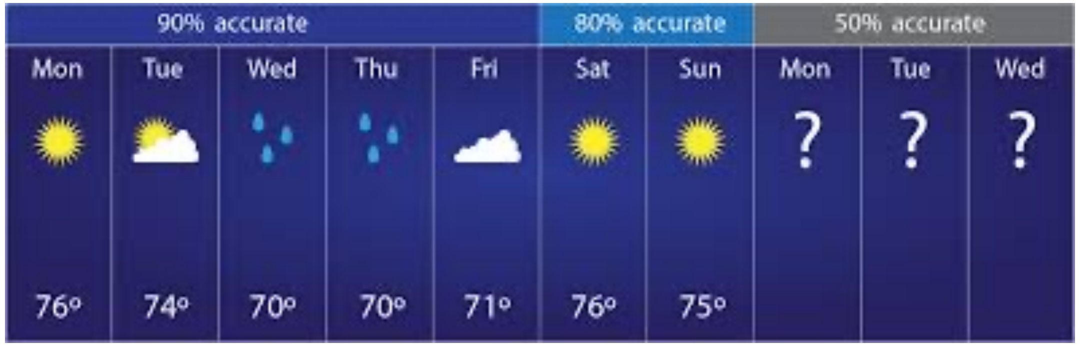
MODELS – SUPERVISED LEARNING (REGRESSION)

Regression: Predicting a numeric value

Stock market

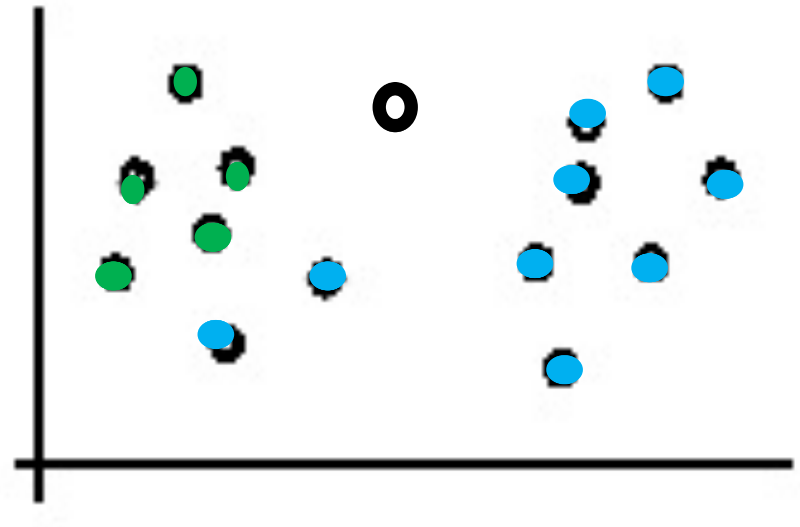


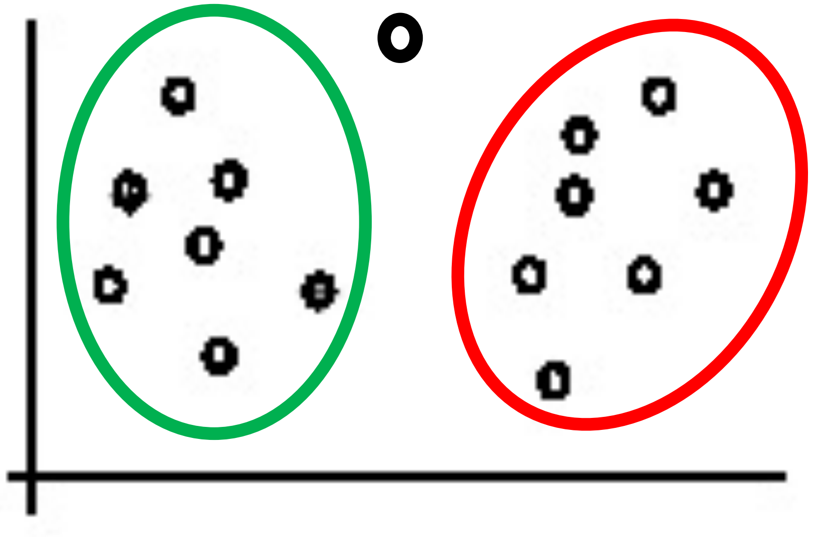
Weather prediction



90% accurate					80% accurate		50% accurate		
Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed
☀	⛅	💧	💧	☁	☀	☀	?	?	?
76°	74°	70°	70°	71°	76°	75°			

MODELS – UNSUPERVISED LEARNING (CLUSTERING)





Thank you!