Data Mining: Introduction

Lecture Notes for Chapter 1

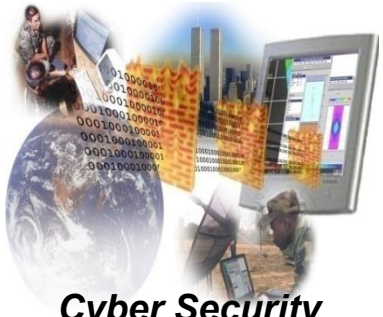
Introduction to Data Mining, 2nd Edition

by

Tan, Steinbach, Karpatne, Kumar

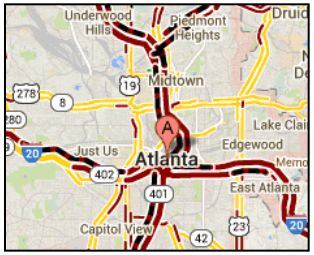
Large-scale Data is Everywhere!

- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Reúna todos os dados que puder quando e onde possível.
- Expectations
  - Os dados coletados terão valor para a finalidade coletada ou para uma finalidade não prevista.

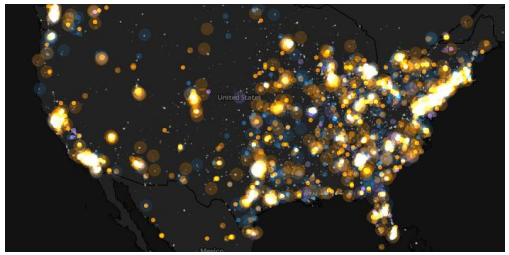


Cyber Security

E-Commerce

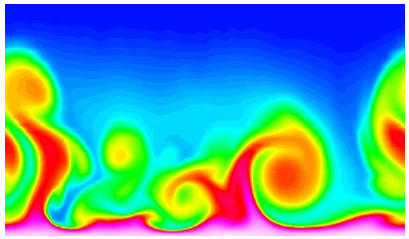


Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

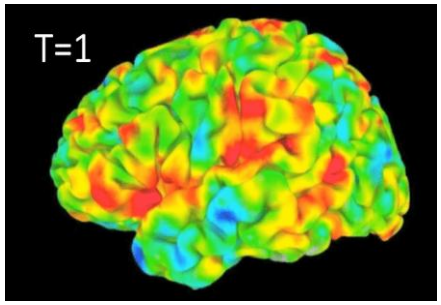
Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - Yahoo has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/ grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

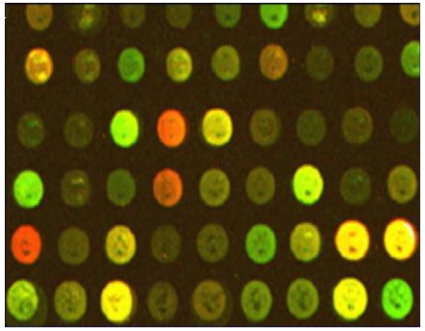
Why Data Mining? Scientific Viewpoint

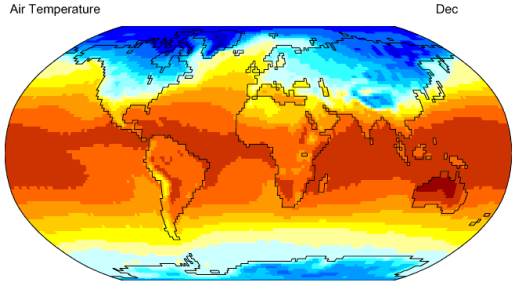
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours



fMRI Data from Brain


Sky Survey Data


Gene Expression Data

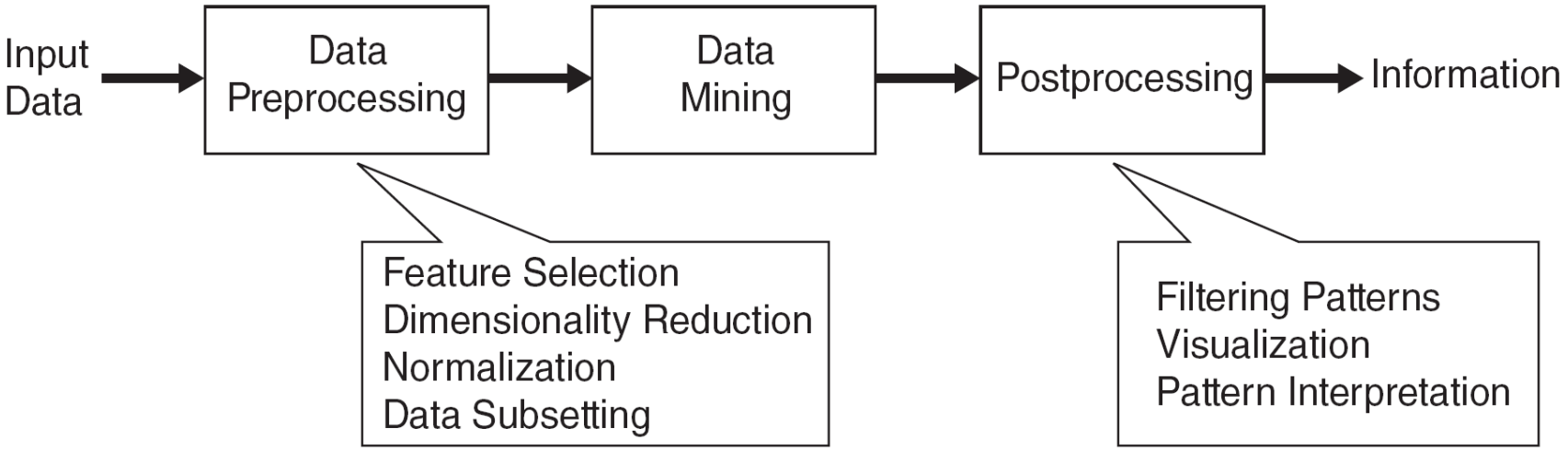


Surface Temperature of Earth

- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

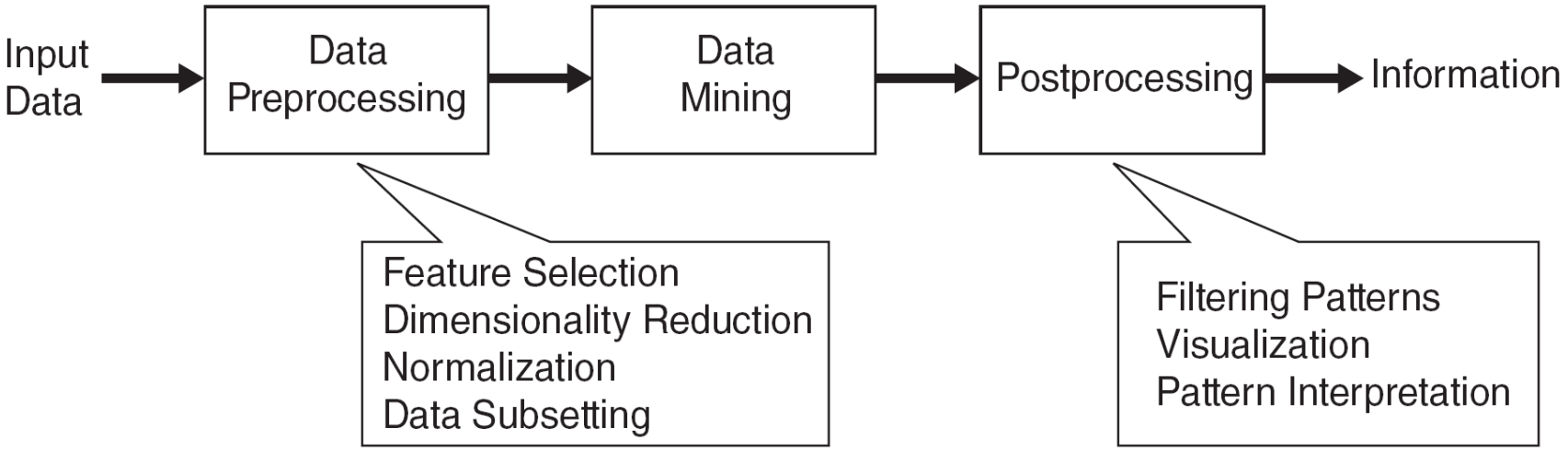
What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



What is Data Mining?

- A key component of the emerging field of data science and data-driven discovery



Data Mining Tasks

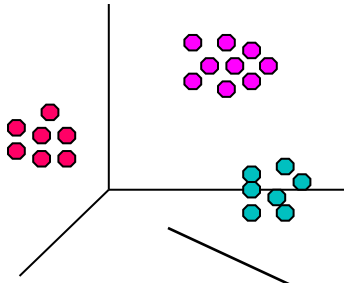
- Prediction Methods [Supervised]
  - Use some variables to predict unknown or future values of other variables.

- Description Methods [Unsupervised]
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks ...

Supervised / Unsupervised



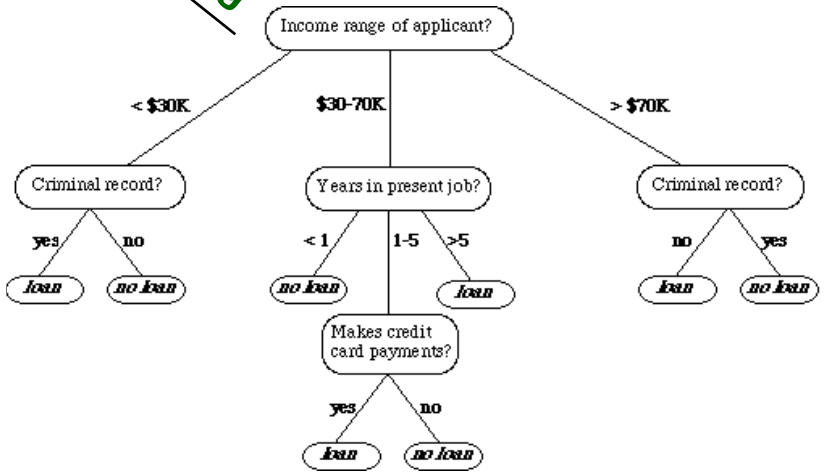
Clustering

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Predictive Modeling

Association



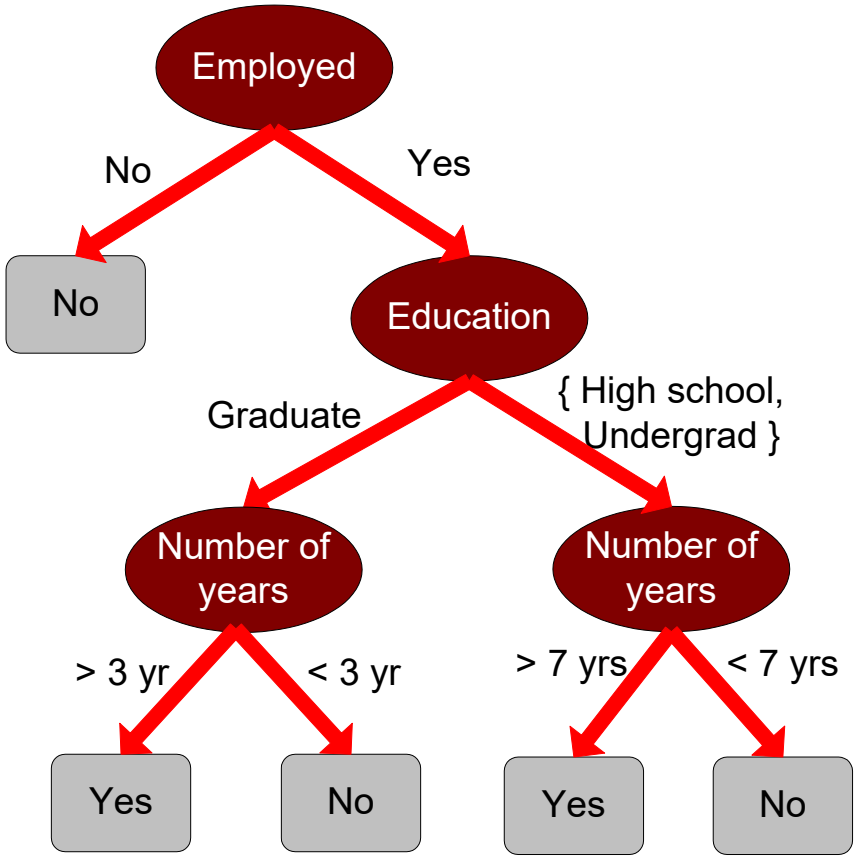
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness

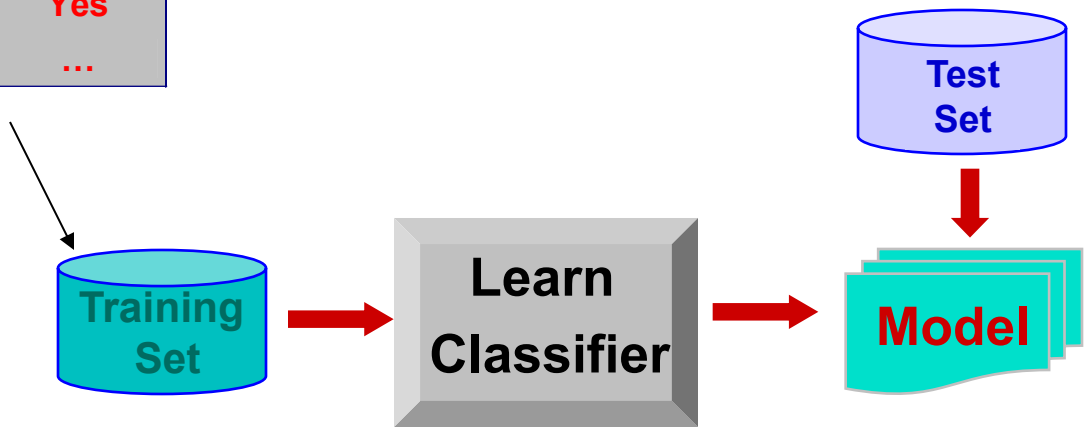


Classification Example

categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

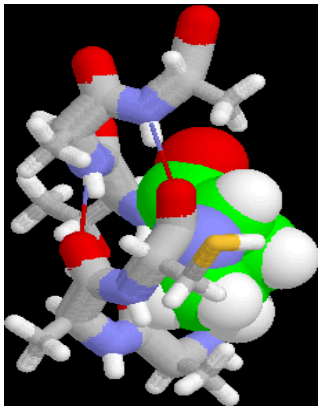


Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent

- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data

- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant



- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

Classification: Application

- Fraud Detection
  - **Goal:** Predict fraudulent cases in credit card transactions.
  - **Approach:**
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Article

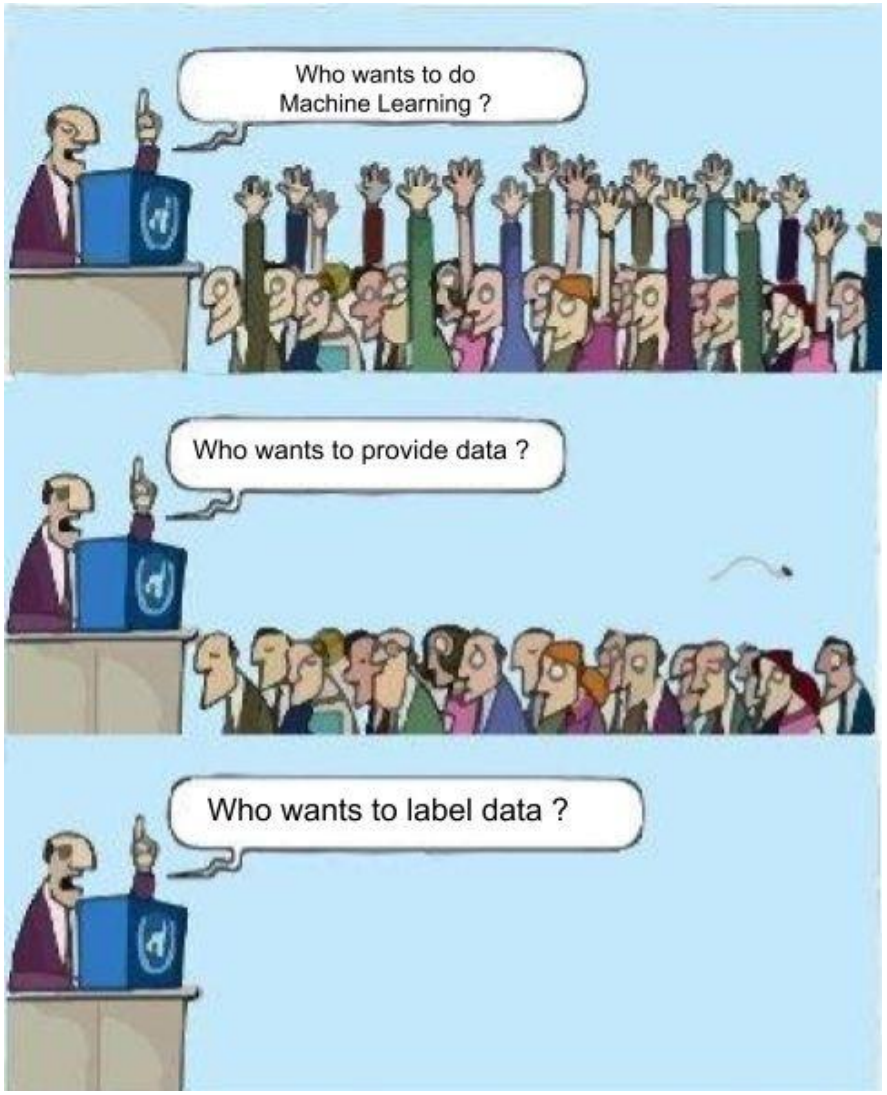
Zargari Khuzani, A., Heidari, M. & Shariati, S.A. *Sci Rep* **11,** 9887 (2021). https://doi.org/10.1038/s41598-021-88807-2

- COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images []
  - Usou classificação para distinguir, de maneira confiável, imagens CXR (radiografia de tórax) de pacientes com COVID-19 de outras formas de pneumonia.
  - Objetivo é usar o classificador COVID, em conjunto com outros testes, para alocação de recursos hospitalares por meio de uma triagem rápida de casos não COVID-19.

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

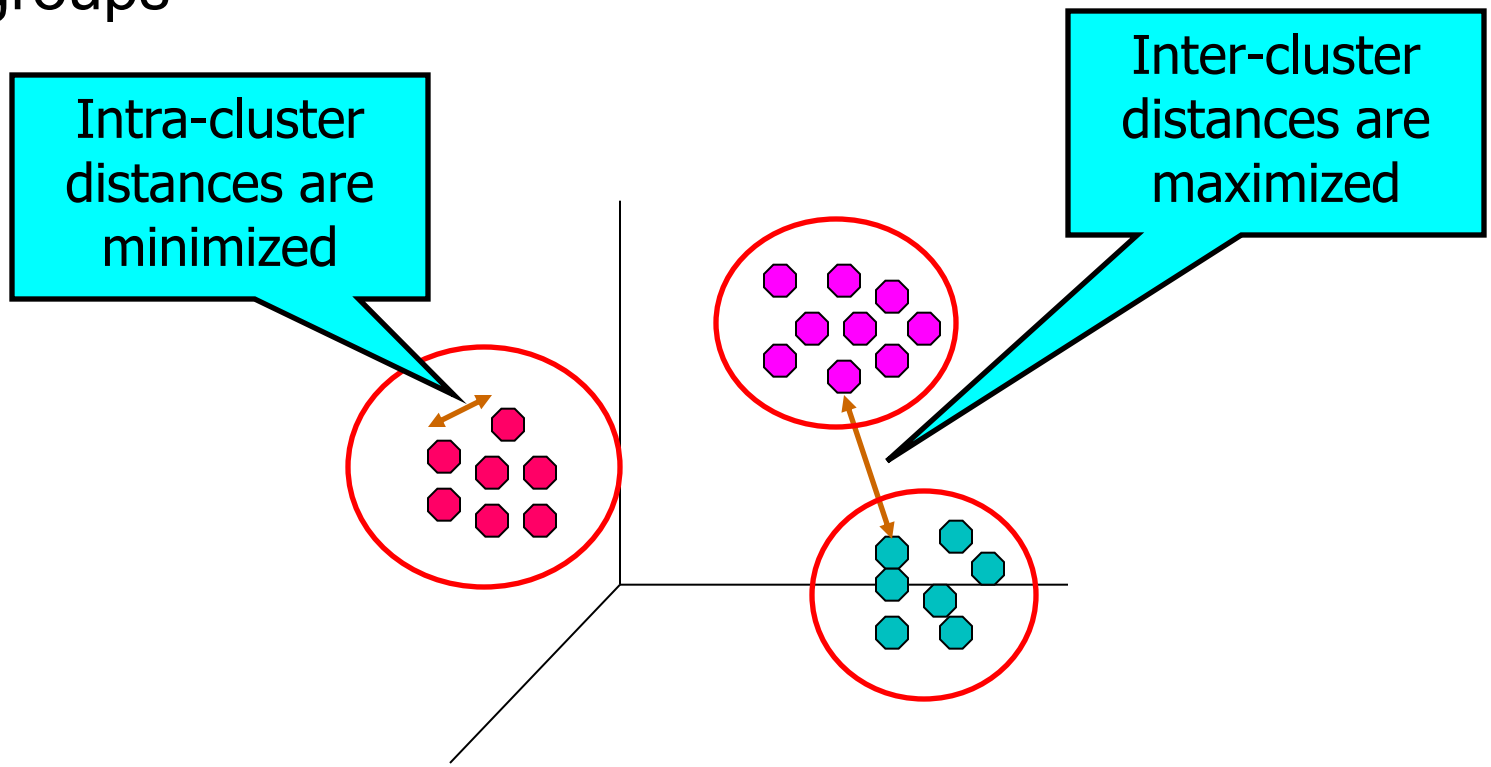
Predictive Modeling / Supervised



Ref. ?

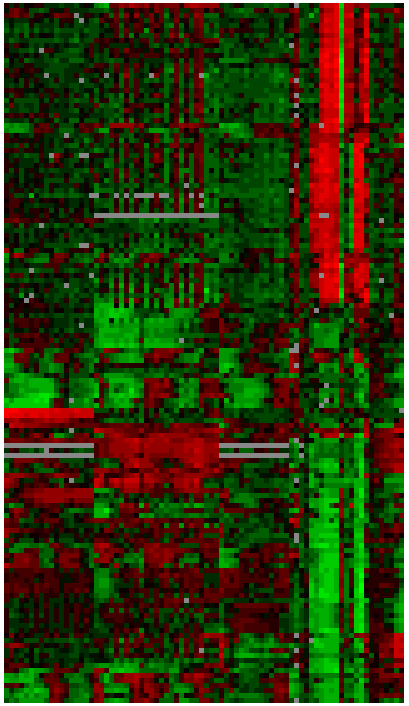
Clustering / Unsupervised

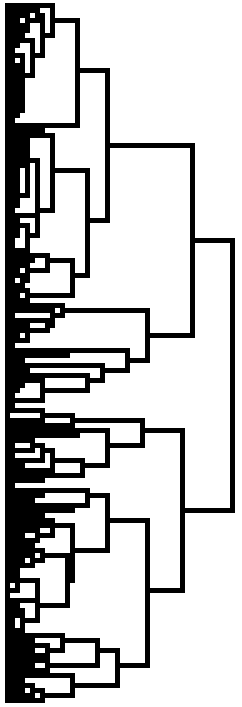
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



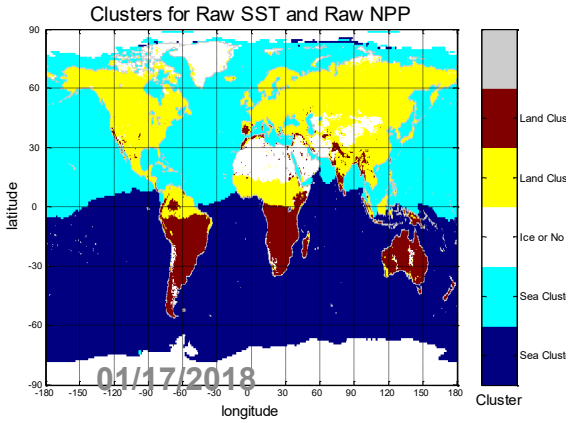
Applications of Cluster Analysis

- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets

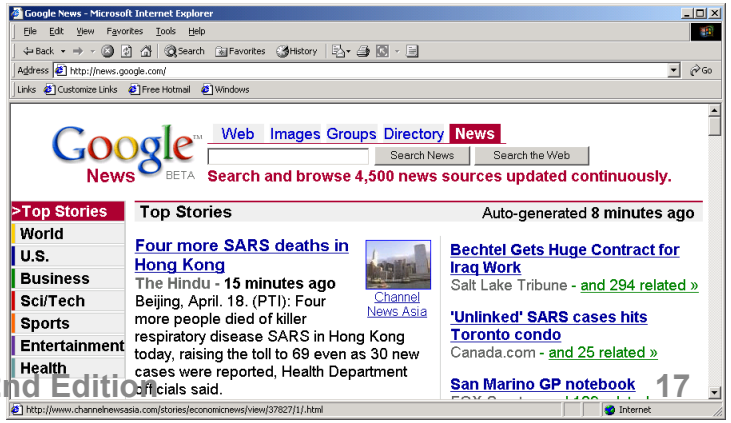




Courtesy: Michael Eisen



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



Clustering: Application

- Market Segmentation:
  - **Goal:** subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - **Approach:**
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Article

ALVES, H. J. de P. .; FERNANDES, F. A. .; LIMA, K. P. de; BATISTA , B. D. de O. .; FERNANDES , T. J. **Research, Society and Development**, *[S. l.]*, v. 9, n. 10, p. e5829109059, 2020. DOI: 10.33448/rsd-v9i10.9059.

- The COVID-19 pandemic in Brazil: an application of the k-means clustering method
  - Usou agrupamento para agrupar as UFs por similaridade devido a algumas características (coeficientes epidemiológicos), a fim de observar as medidas de combate ao COVID-19 realizadas em cada um desses grupos
  - O agrupamento pode se apresentar como um recurso adicional para sinalizar quais locais e quais medidas devem ser adotadas em cada um dos grupos e/ou onde determinadas medidas foram efetivas.

Association Rule Discovery / Unsupervised

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk, Bread} --> {Coke}

{Diaper} --> {Beer}

Association Analysis: Applications

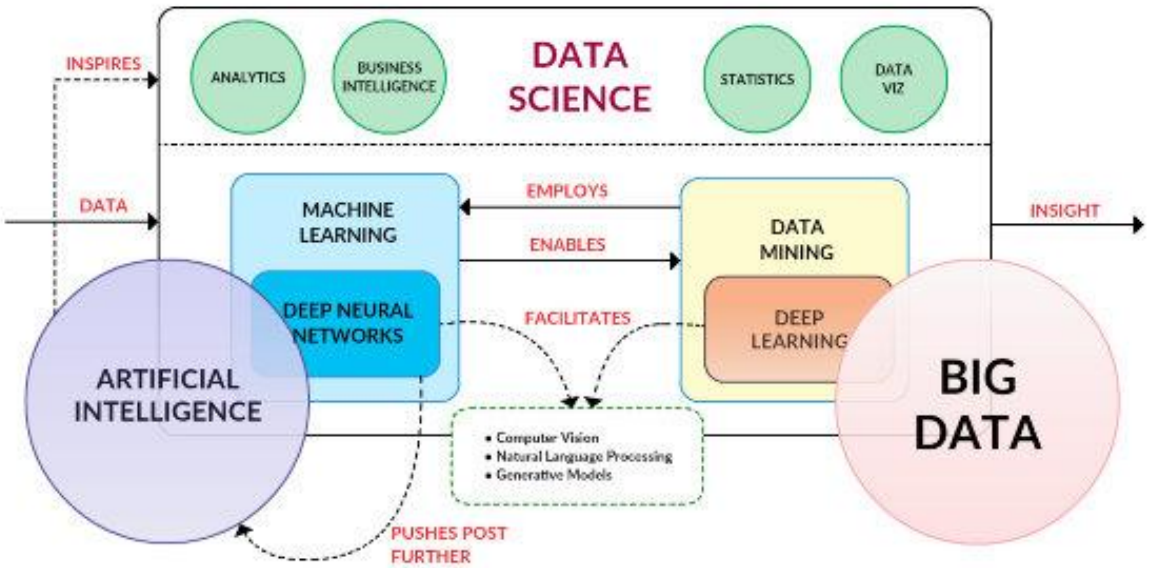
- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Association Analysis: Article

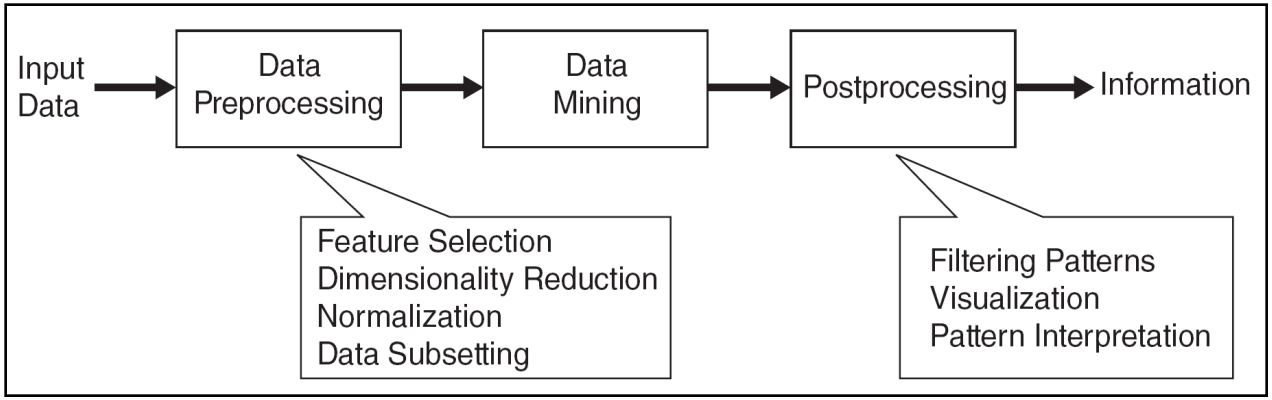
- Association mining based approach to analyze COVID-19 response and case growth in the United States
  - Usou associação para entender os efeitos de diferentes intervenções não farmacêuticas na contenção da taxa de infecção por COVID-19.

Katragadda, S., Gottumukkala, R., Bhupatiraju, R.T. *et al*. *Sci Rep* **11,** 18635 (2021). https://doi.org/10.1038/s41598-021-96912-5

Finalizando...



https://www.kdnuggets.com/2016/03/data-science-puzzle-explained.html



Finalizando...

- https://machinelearningmastery.com/types-of-learning-in-machine-learning/