Data Mining Methods

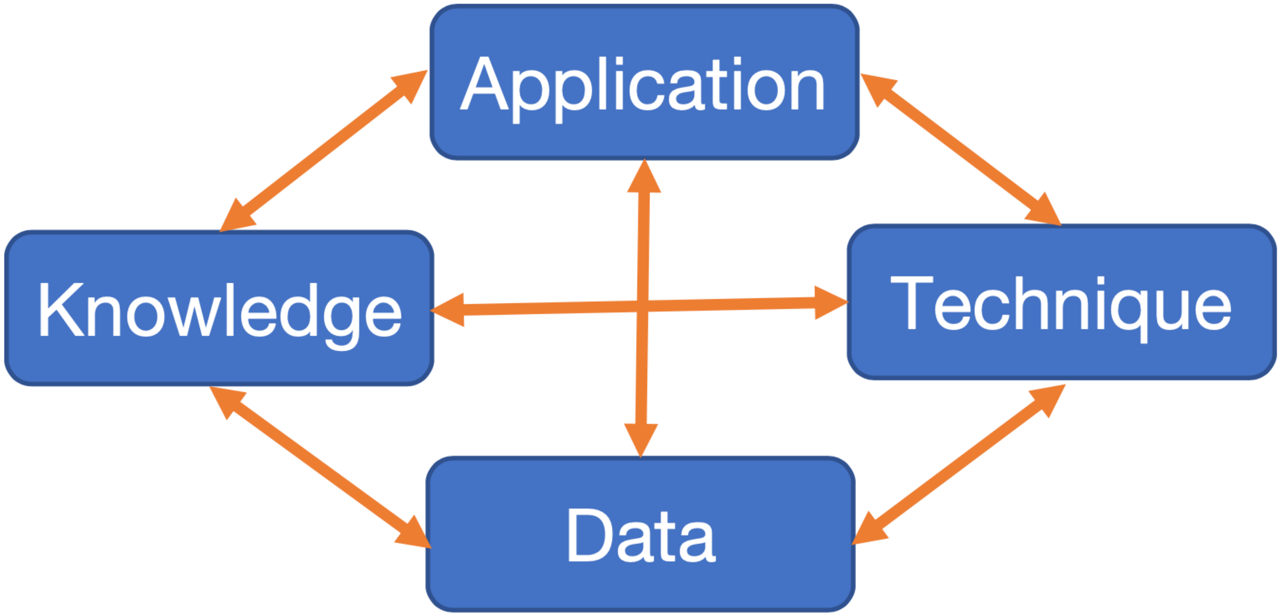
from data to insight

There is no such a thing as insight
without a clear and concise question,
as well as having a way to measure the success or failure
of the answer obtained

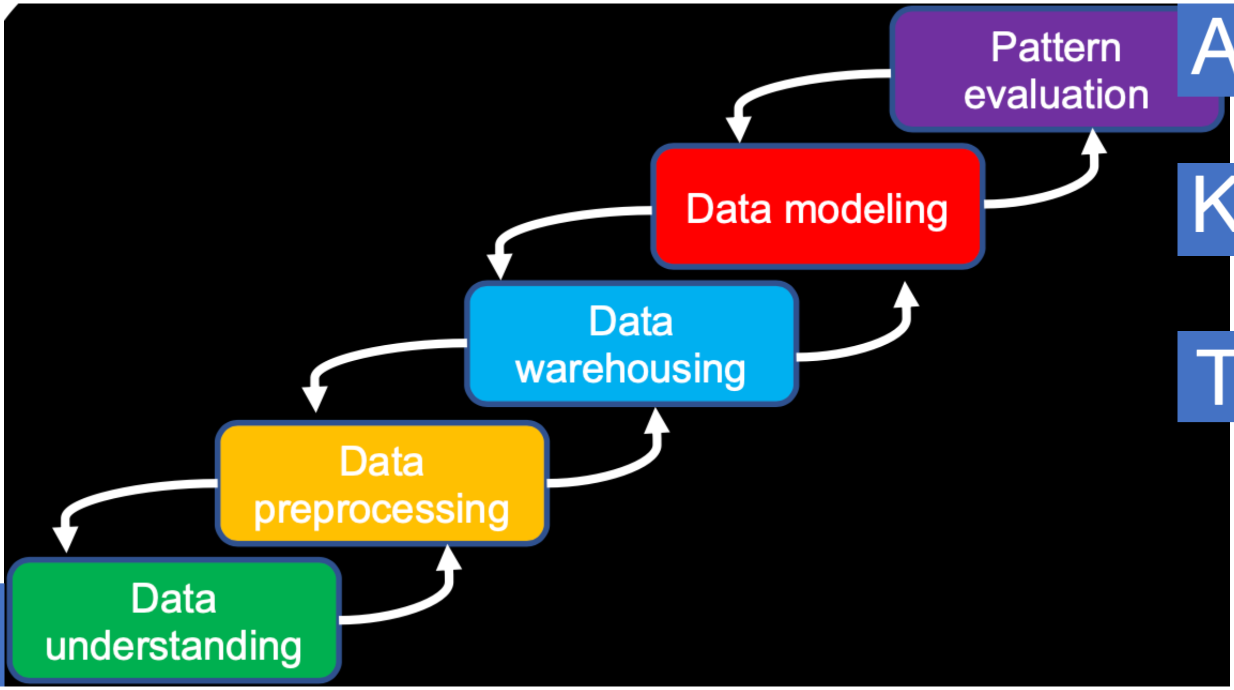
Learning objective:

- Identify the core functionalities of data modelling in the data mining pipeline.
- Apply the Apriori algorithm for frequent itemset mining, among others

Data Mining: Four Views



Data Mining Pipeline



Application

Knowledge

Technique

Data

Technique View

- Frequent pattern analysis
- Classification & prediction
- Clustering
- Anomaly detection
- Trend and evolution analysis

Frequent Pattern Analysis

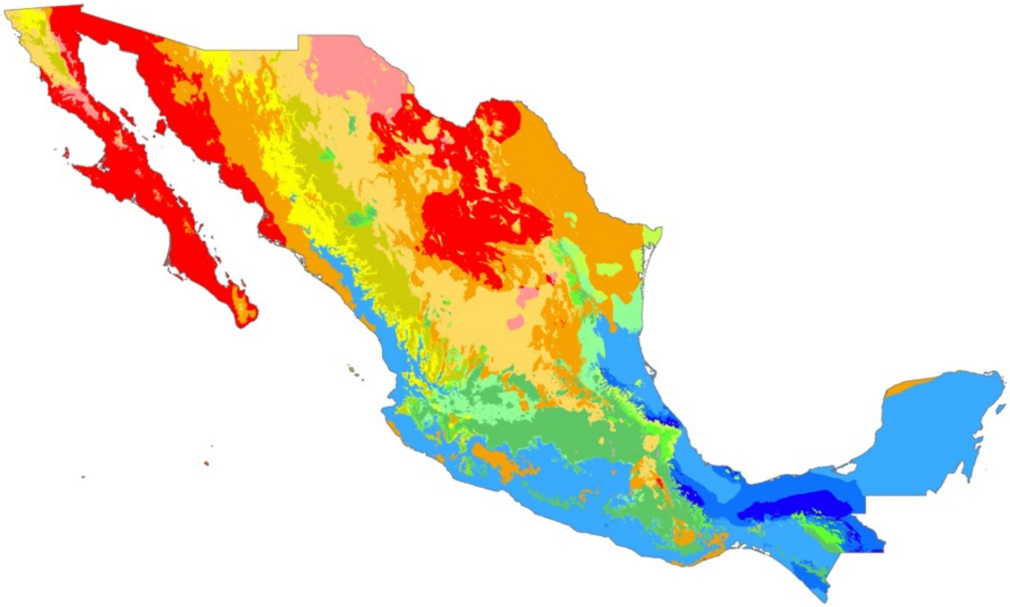
- Frequent itemset
- Frequent sequence
- Frequent structure
- Association rules
- Correlation analysis

Classification

- Pre-defined classes
- Need training data
- Build model to distinguish classes

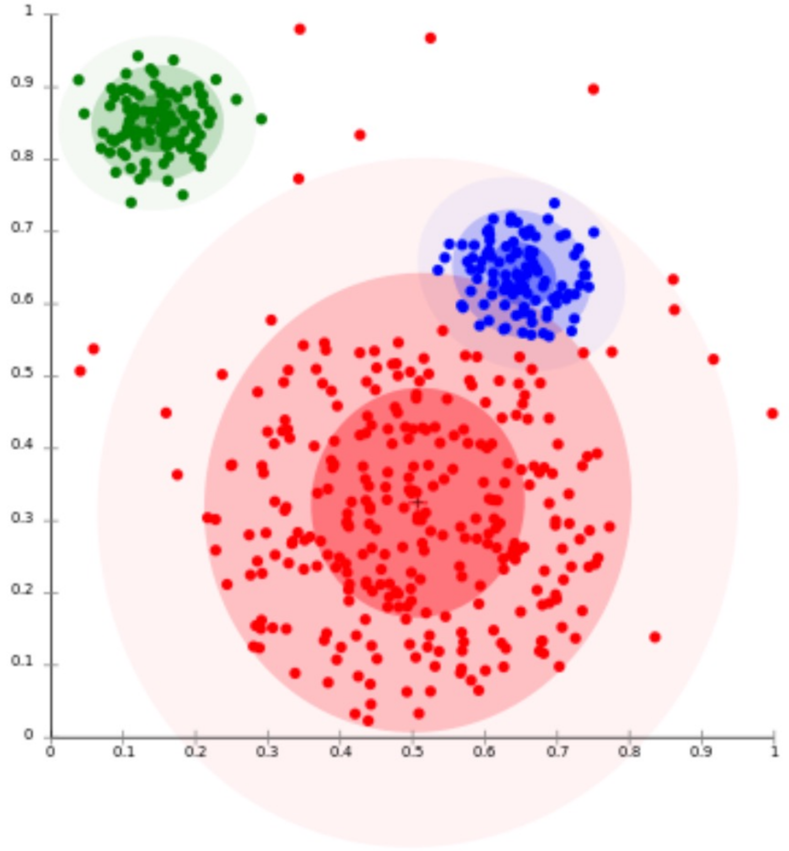
Prediction

- Numerical prediction (continuous value)
  - E.g., weather, stock price, traffic



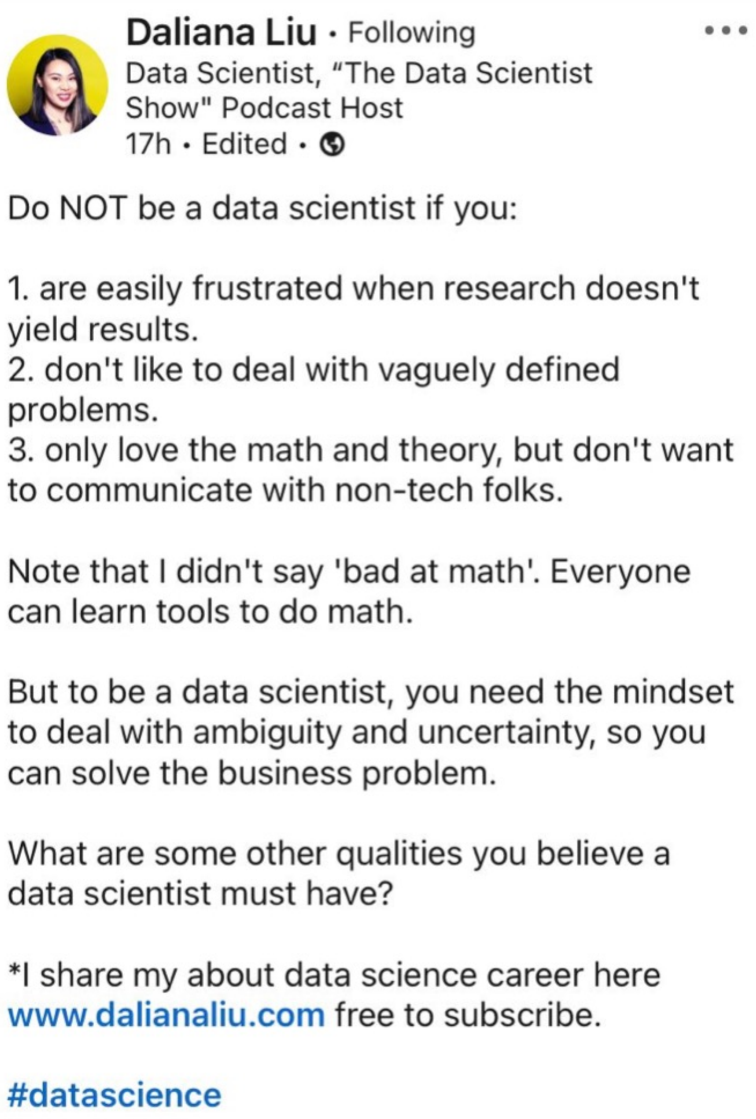
Clustering

- No predefined classes
- Intra-cluster similarity
- Inter-cluster dissimilarity



Anomaly Detection

- Anomaly/outlier - Differ from the “norm”
  - E.g., error, noise, fraud, extreme events



Do NOT be a data scientist if you:

1. are easily frustrated when research doesn't yield results.
2. don't like to deal with vaguely defined problems.
3. only love the math and theory, but don't want to communicate with non-tech folks.

Note that I didn't say 'bad at math'. Everyone can learn tools to do math.

But to be a data scientist, you need the mindset to deal with ambiguity and uncertainty, so you can solve the business problem.

What are some other qualities you believe a data scientist must have?

*I share my about data science career here www.dalianaliu.com free to subscribe.

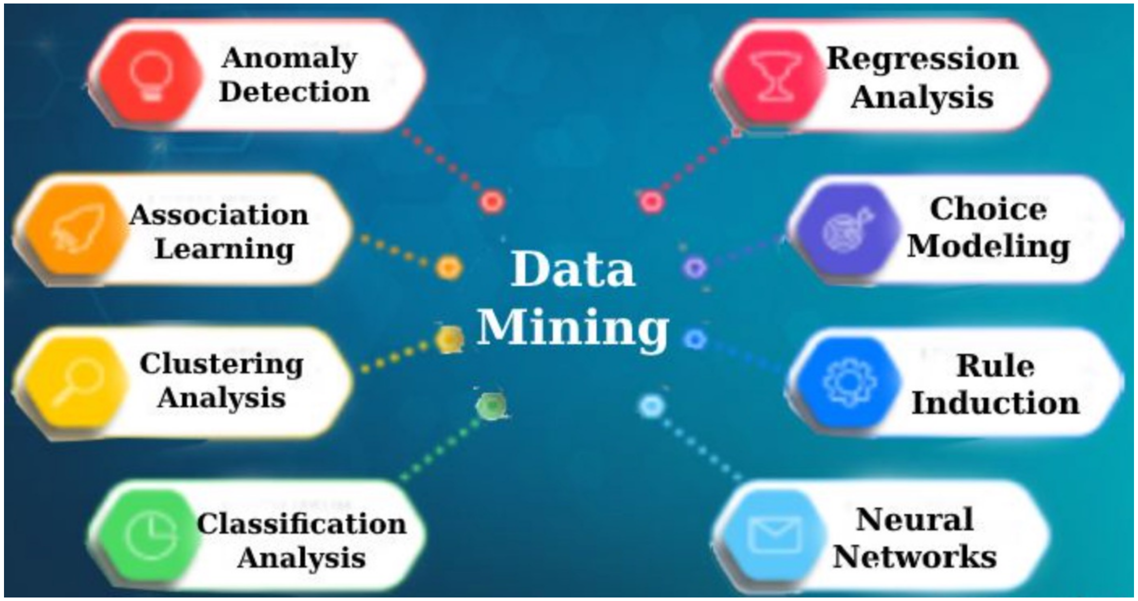
#datascience

Trend and Evolution Analysis

- Changes over time
  - ✓ Overall trend
  - ✓ Periodical patterns
  - ✓ Anomalies

Data Mining Methods

- Frequent pattern analysis
- Classification
- Clustering
- Outlier analysis



Data Mining Methods

Data science skills

Range from

- programming to design
- mathematics to storytelling

The motivation – Data mining

- Deriving valuable insights from data
  - ✓ widely welcomed by businesses

Typical questions :

- What product will sell better in conjunction with another popular product?
- How can customers be encouraged to spend a longer time in an online portal?
- What advertisement should be placed on what site?
- How to determine if a retail transaction is valid ?

The skills

- Both statistics and a strong business acumen
- Foundations in computer science, mathematics, modelling and programming
- Good communication skills & inquisitive mind

Inter alia

Sexiest job of the 21st century

- data scientist
- it is hard to find people with the right skills to fill in these roles
- this has lead to branding

  data scientists as Unicorns.

Good data scientist

a linear combination of some of the following traits:

- Curiosity
- Grasp of machine learning
- Data product building and management
- Effective communication of data insights
- Programming and data visualisation abilities
- Knowledge of statistics and probability
- Healthy skepticism, in the scientific tradition

Four pillars

1. Identify who the main stakeholders and clarify the lines of reporting.
2. To be able to work independently and productively
3. Identify the data to tackle a problem
    - proper interpretation is not necessarily easy, and misrepresentation of the results can be very damaging.
4. Have the outcome always in mind

Technologies

- Data Framework - MapReduce, BigQuery, Hadoop, Spark
- Streaming data collection - Kafka, Flume, Scribe
- Job scheduling - Azkaban, Oozie
- Big Data Query languages - Pig, Hive
- Data stores - Voldemort, Cassandra, Neo4j, Hbase

Open Source Tools

- Python: Data manipulation, prototyping, scripting
- **Apache Hadoop:** Framework for processing big data
- Apache Mahout: Scalable machine-learning algorithms for Hadoop
- Spark: Cluster-computing framework for data analytics
- R Project for Statistical Computing: Data manipulation and graphing
- Julia: High-performance technical computing
- GitHub, Subversion: Software and model management tools
- Ruby, Perl, OpenRefine: Prototyping and production scripting languages

The steps

- Question identification
- Data acquisition
- Data munging – wrangling - data janitor
- Model construction
- Representation
- Interaction