Introduction to Big Data Analytics

Anna Monreale

Computer Science Department

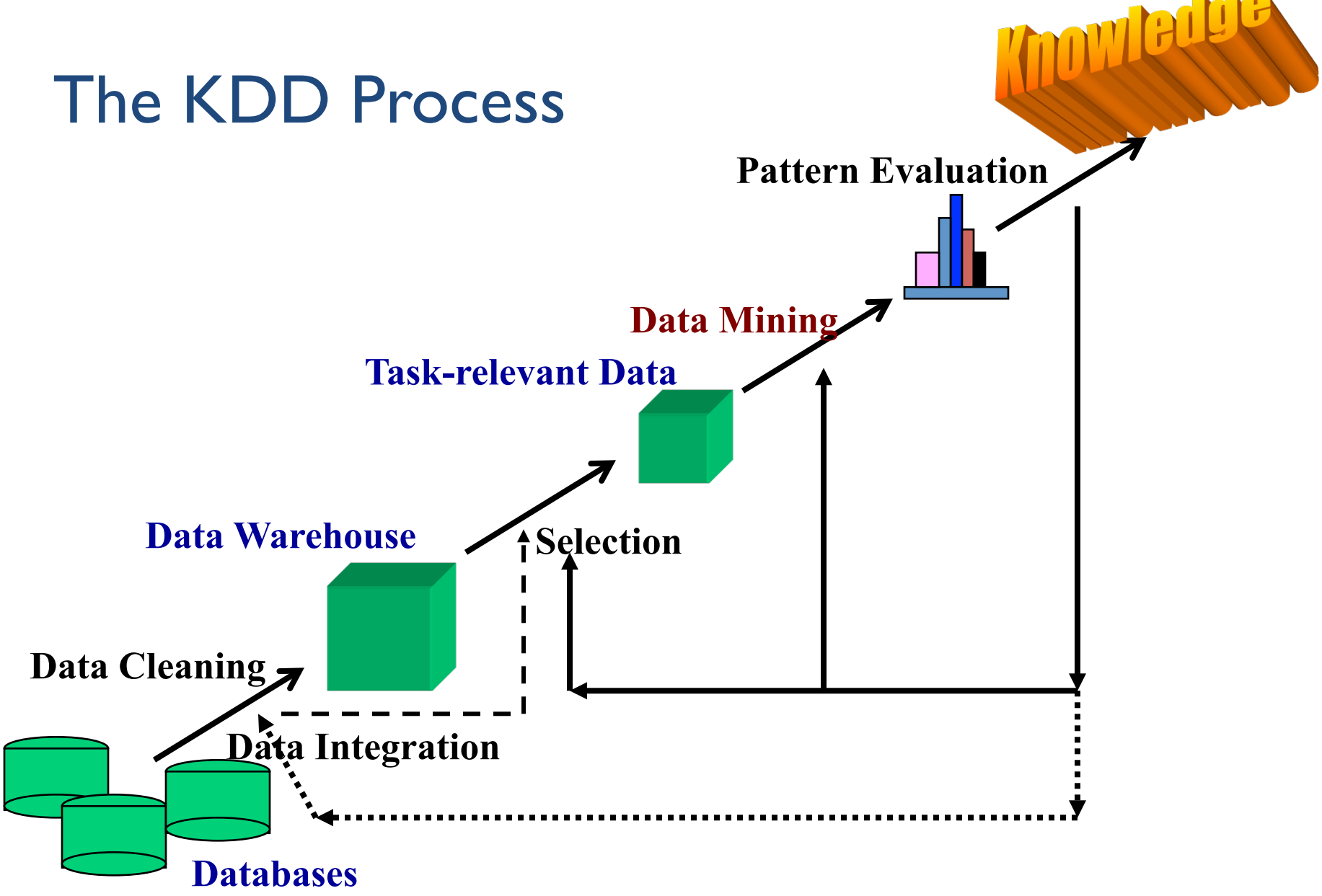
What is Big Data Analytics?

Big data analytics is the process of **collecting** and **analysing** large data sets from traditional and digital sources to **extract** trends and patterns that can be used to **support decision-making.**

What is Data Mining?

It is the use of **efficient** techniques for the analysis of **very large collections of data** and the **extraction** of useful and possibly unexpected patterns in data (hidden knowledge).

The KDD Process

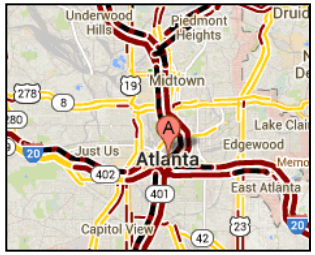


Large-scale Data is Everywhere!

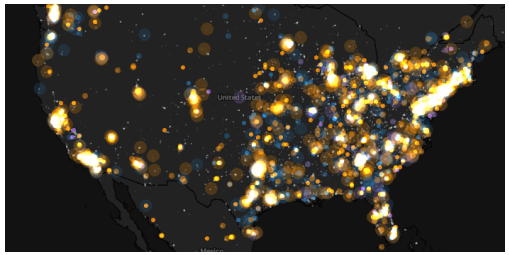
- **Enormous data growth in both commercial and scientific databases**
  - due to advances in data generation and collection technologies
- **New mantra**
  - Gather whatever data you can whenever and wherever possible
- **Expectations**
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

Cyber Security

E-Commerce

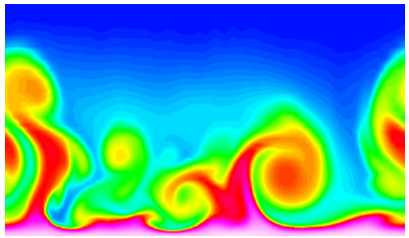


Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

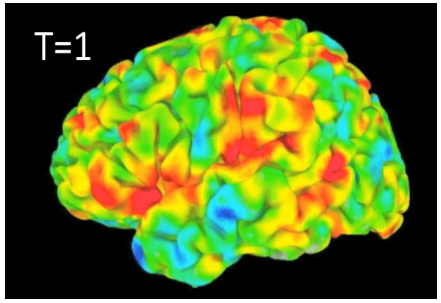
Why Data Mining? Commercial Viewpoint

- **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)

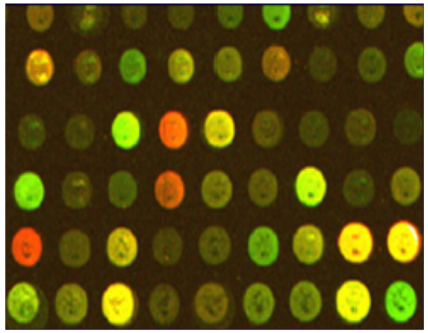
Why Data Mining? Scientific Viewpoint

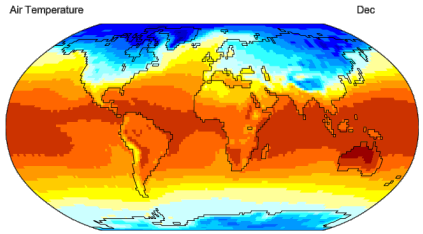
- **Data collected and stored at enormous speeds**
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- **Data mining helps scientists**
  - in automated analysis of massive datasets
  - In hypothesis formation



fMRI Data from Brain


Sky Survey Data


Gene Expression Data

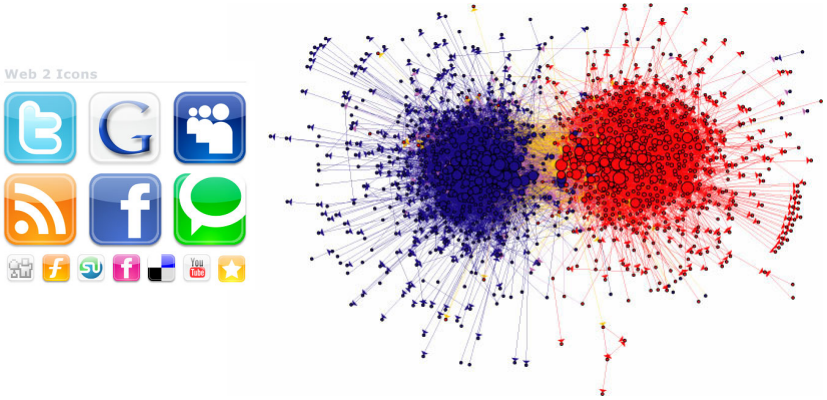


Surface Temperature of Earth

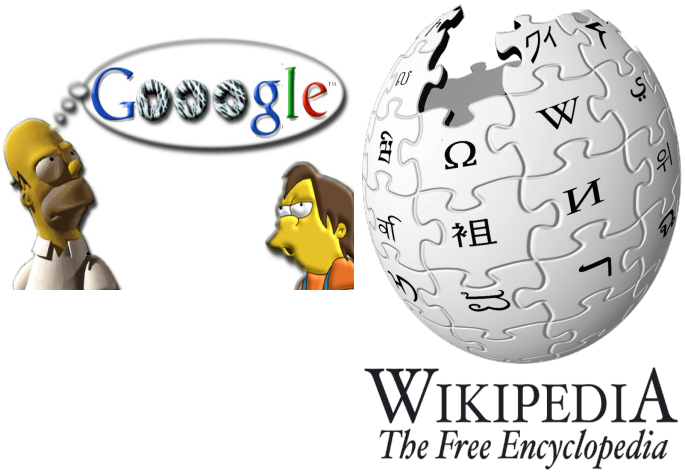
Big data proxies of social life

Shopping patterns & lyfestyle

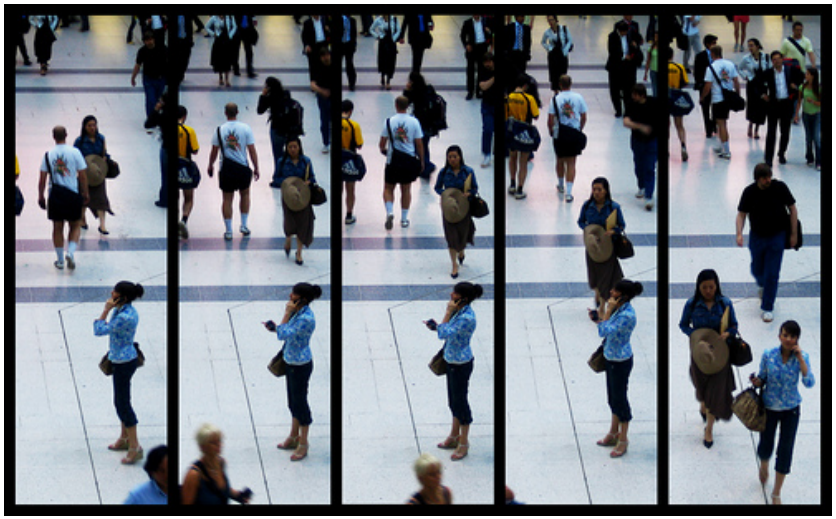
RELATIONSHIPS & SOCIAL TIES

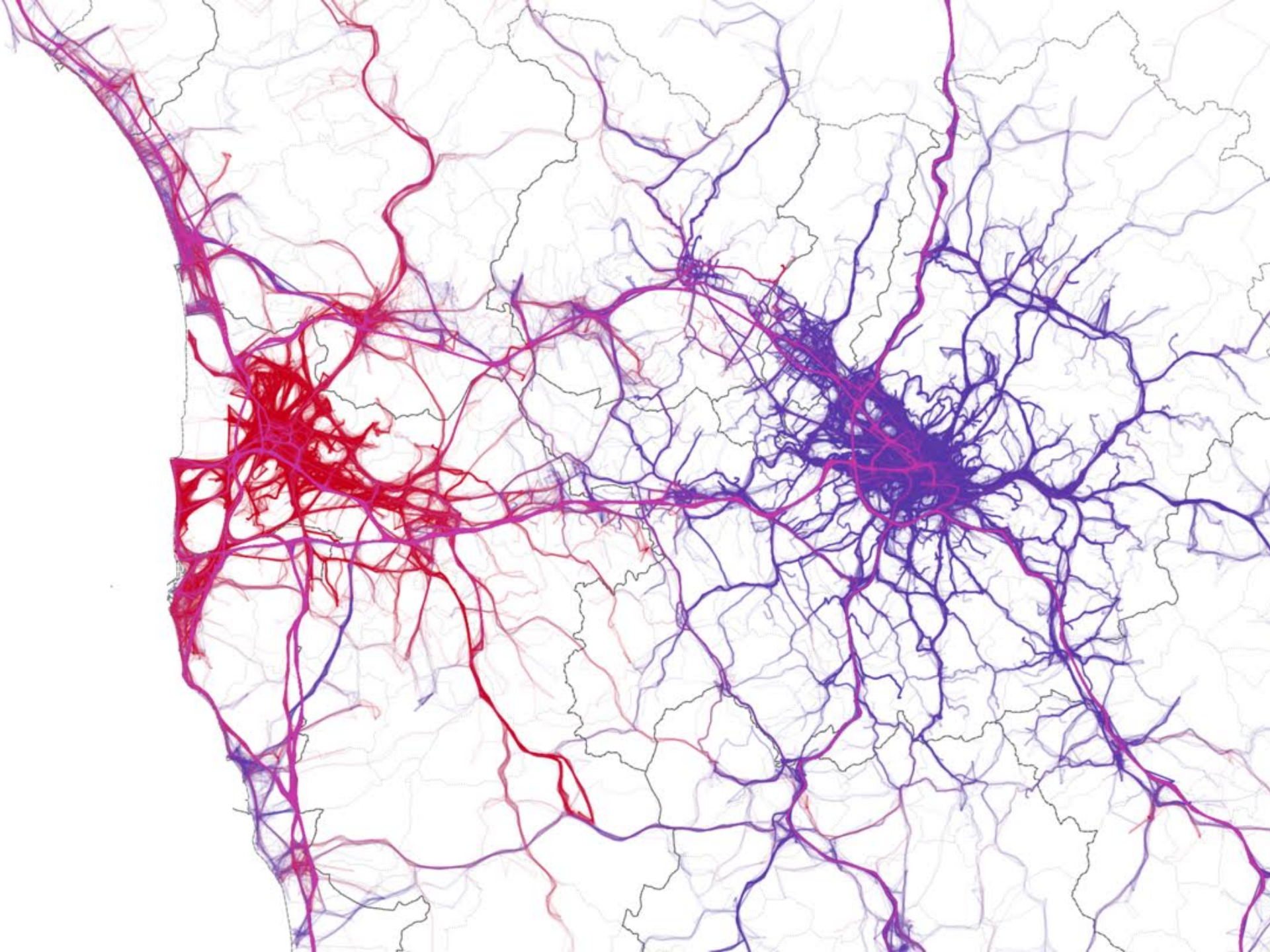


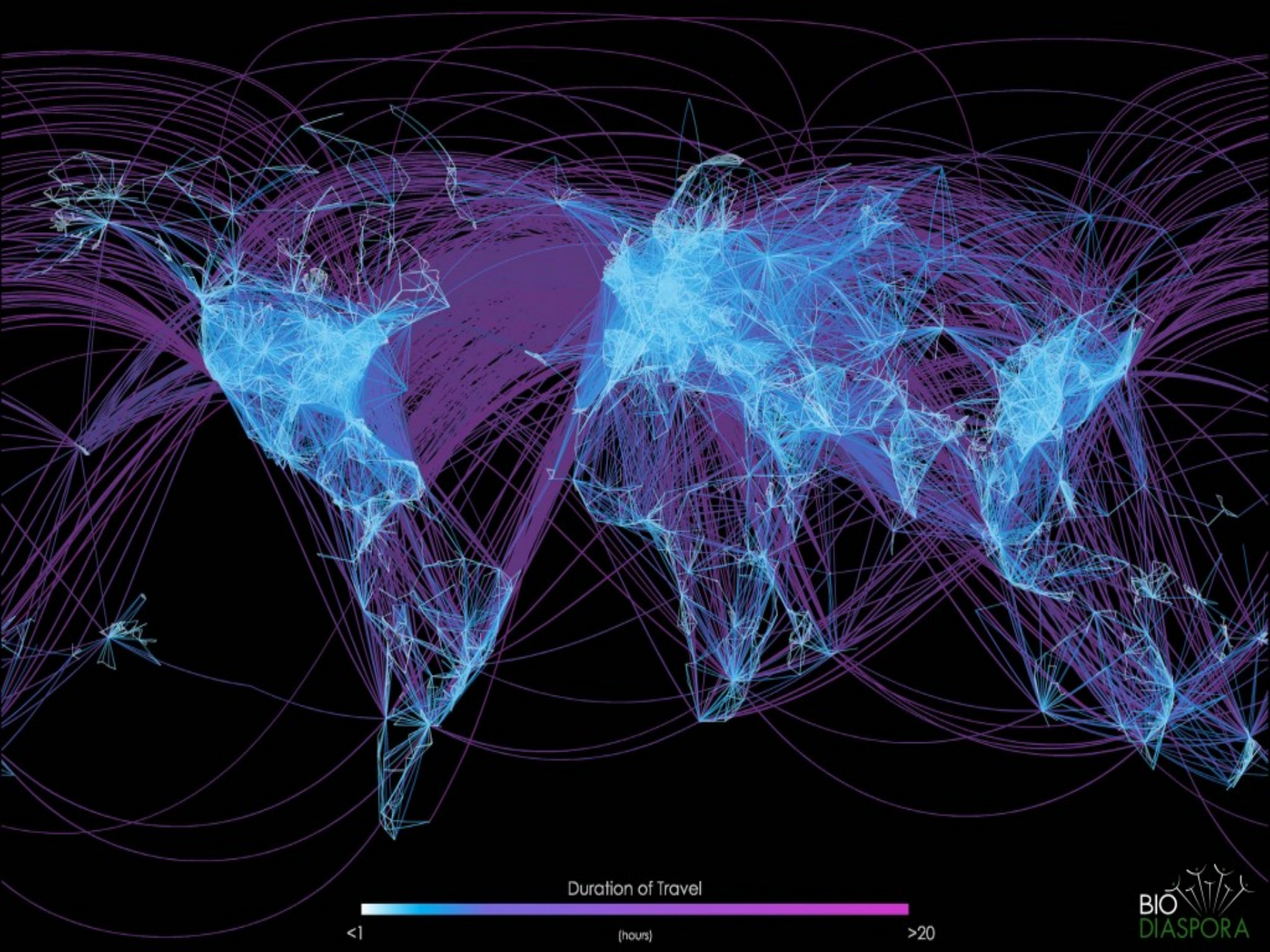
DESIRES, OPINIONS, SENTIMENTS



MOVEMENTS



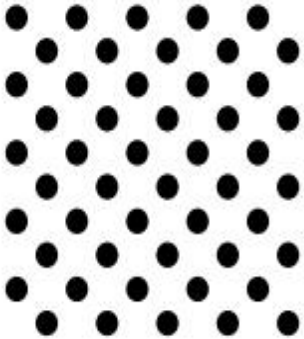
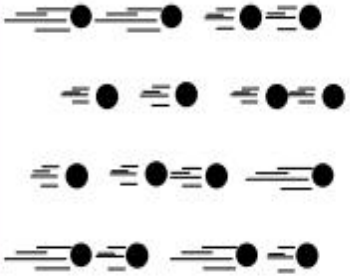
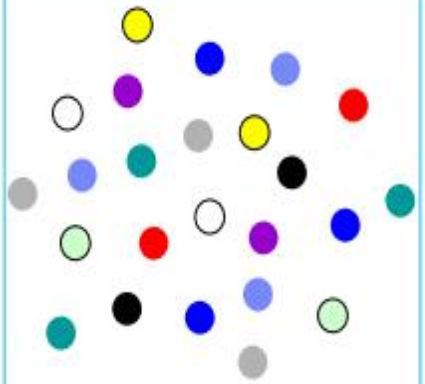
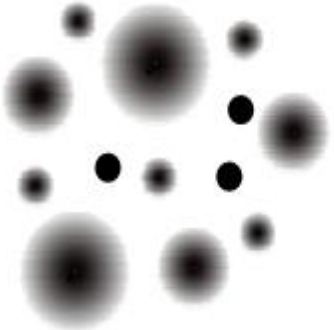



Duration of Travel
<1
(hours)
>20
BIO
DIASPORA

Primary Data

- **Original data** that has been collected for a specific purpose
- Primary data **is not altered** by humans

The Four V's of Big Data

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Varierty of Data Sources

Varierty of Data Sources

ARCHIVES

Archives of scanned documents, statements, insurance forms, medical record and customer correspondence, paper archives, and print stream files that contain original systems of record between organizations and their customers

DOCS

XLS, PDF, CSV, email, Word, PPT, HTML, HTML 5, plain text, XML, JSON, etc.

MEDIA

Images, videos, audio, Flash, live streams, podcasts, etc.

DATA STORAGE

SQL, NoSQL, Hadoop, doc repository, file systems, etc.

BUSINESS APPS

Project management, marketing automation, productivity, CRM, ERP content management systems, HR, storage, talent management, procurement, expense management, Google Docs, intranets, portals, etc.

PUBLIC WEB

Government, weather, competitive, traffic, regulatory, compliance, health care services, economic, census, public finance, stock, OSINT, the World Bank, SEC/Edgar, Wikipedia, IMDb, and other Web services

SOCIAL MEDIA

Twitter, LinkedIn, Facebook, Tumblr, Blog, SlideShare, YouTube, Google+, Instagram, Flickr, Pinterest, Vimeo, Wordpress, IM, RSS, Review, Chatter, Jive, Yammer, etc.

MACHINE LOG DATA

Event logs, server data, application logs, business process logs, audit logs, call detail records (CDRs), mobile location, mobile app usage, clickstream data, etc.

SENSOR DATA

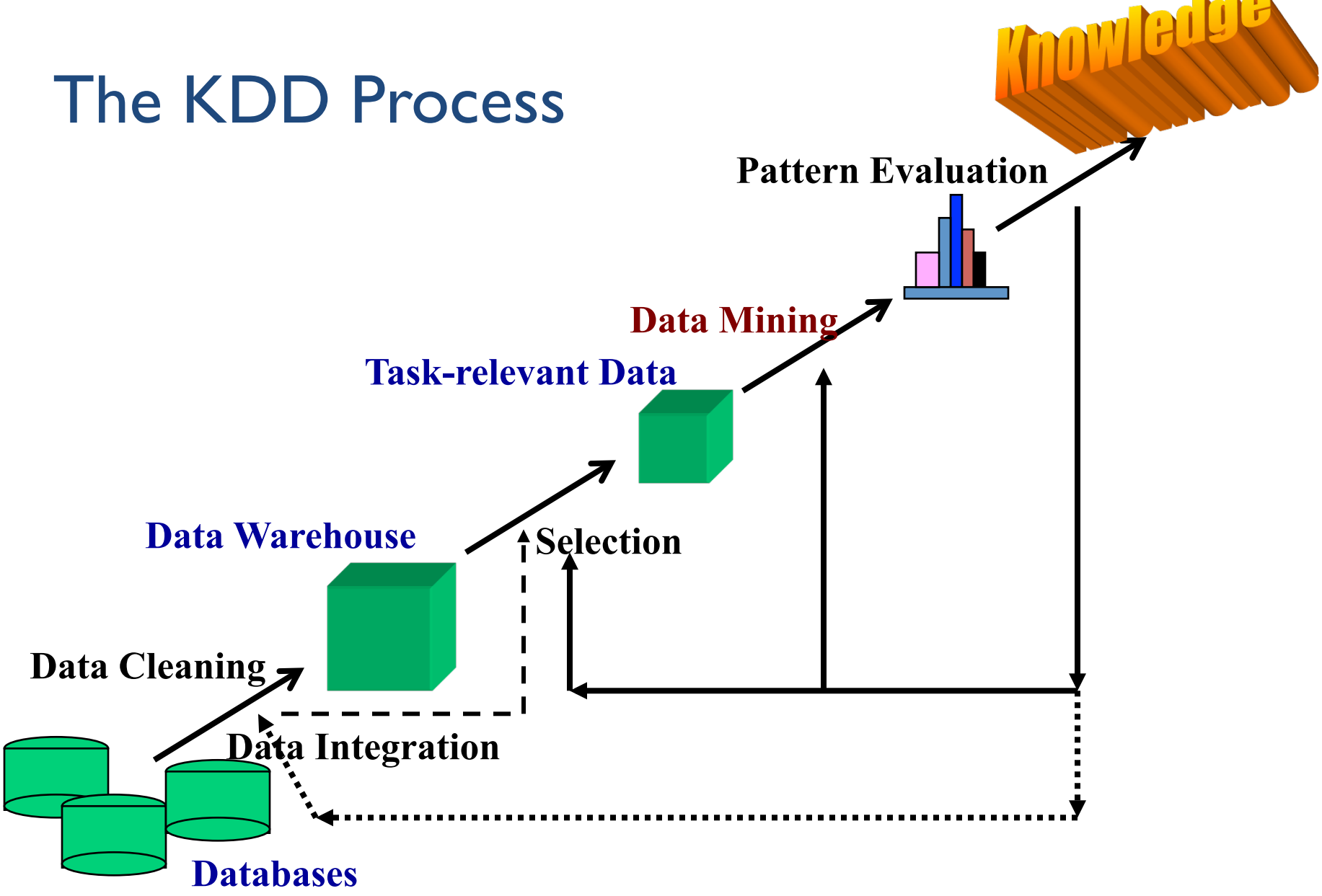
Medical devices, smart electric meters, car sensors, road cameras, satellites, traffic recording devices, processors found within vehicles, video games, cable boxes or household appliances, assembly lines, office buildings, cell towers and jet engines, air conditioning units, refrigerators, trucks, farm machinery, etc.

Secondary Data

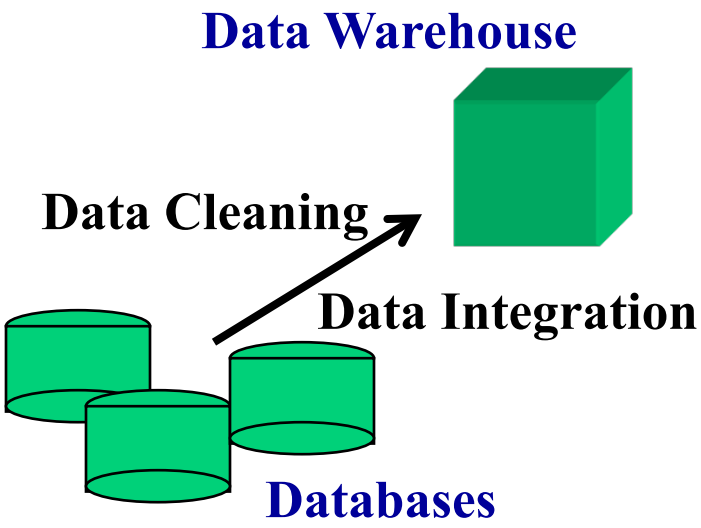
- Secondary data is the data that has been already collected and made available for other purposes

- Secondary data may be obtained from many sources, including literature, industry surveys, compilations from computerized databases and information systems, and computerized or mathematical models of environmental processes

The KDD Process



Data Integration and Preparetion

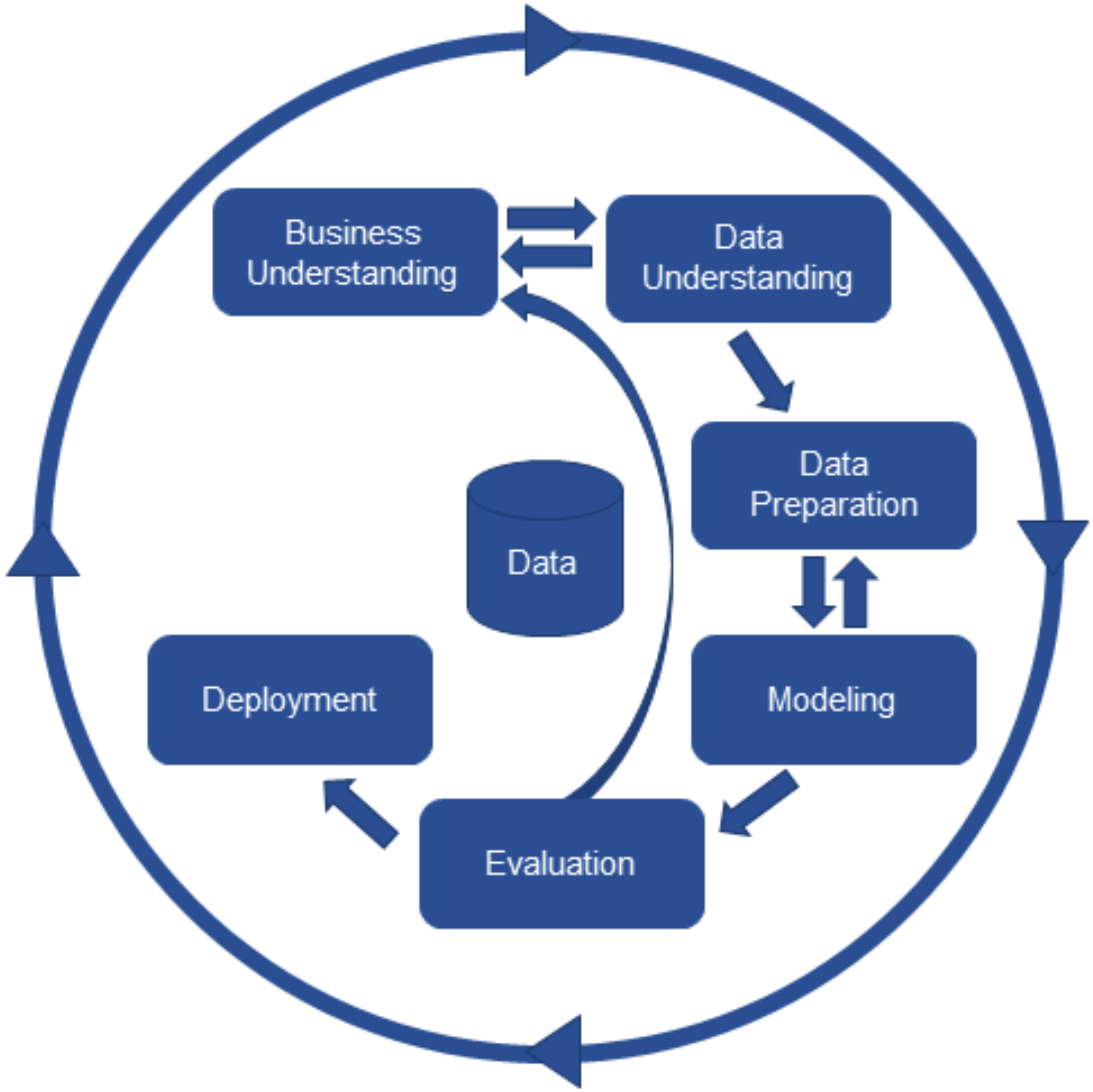


Data Integration involves the process of data understanding, data cleaning, merging data coming from multiple sources and transfoming them to load them into a **Data Warehouse**

Data Warehouse is a database targeted to answer **specific business questions**

Developing a big data analytics project requires the **BUSINESS UNDERSTANDING**

CRISP Model


Business Understanding
Data Understanding
Data Preparation
Data
Modeling
Deployment
Evaluation

DATA UNDERSTANDING

Which is the type of data?

Types of data sets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

What is Data?

- Collection of **data objects** and their **attributes**
- An **attribute** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an **object**
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

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

Data Matrix

- If data objects have **the same fixed set of numeric attributes**, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an ***m* by *n* matrix**, where there are *m* rows, one for each object, and *n* columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Each document becomes a ‘term’ vector
  - Each term is a component (attribute) of the vector
  - The value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

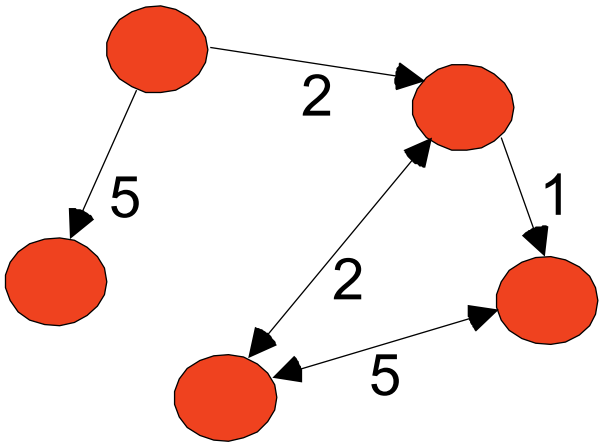
Transaction Data

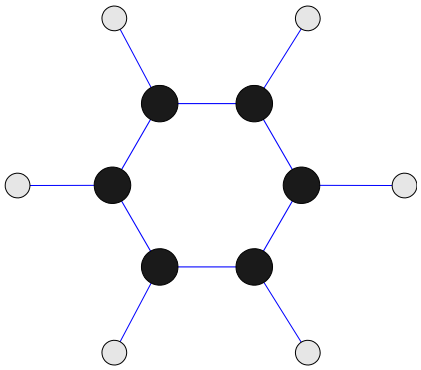
- A special type of record data, where
  - Each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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

Graph Data

- Examples: Generic graph, a molecule, and webpages





Benzene Molecule: C6H6

Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

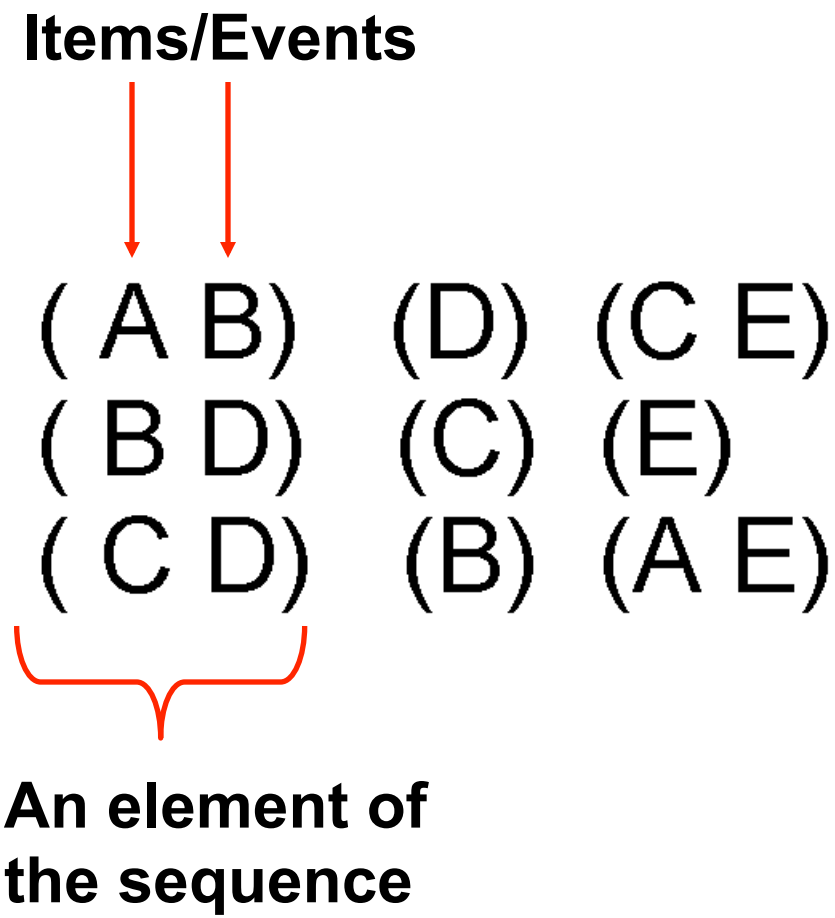
General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Ordered Data

- Sequences of transactions



Ordered Data

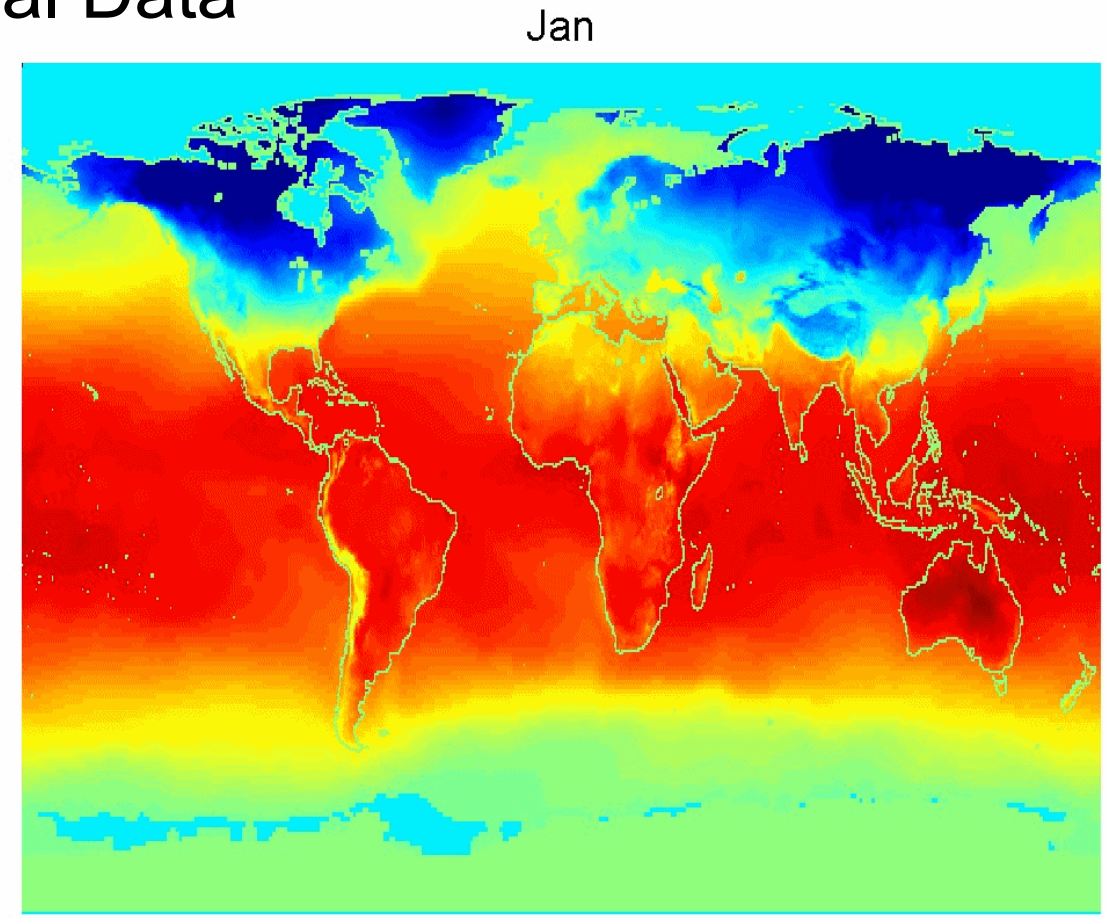
- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Data Understanding Goals

Gain insight in your data

- with respect to your project goals
- and in general

Find answers to the questions

- What kind of attributes do we have?
- How is the data quality?
- Does a visualization helps?
- Are attributes correlated?
- What about outliers?
- How are missing values handled?

A checklist for data understanding

- Determine the **quality of the data**
- Compare **statistics** and **distributions** with the expected behavior
- **Find outliers**
- Discover new or confirm expected **dependencies** or **correlations** between attributes.
- Detect and examine **missing values**
- **Check** specific application dependent **assumptions** (e.g. the attribute follows a normal distribution)

Attribute Type

- **categorical (nominal):** finite domain
  - The values of a categorical attribute are often called classes or categories.
  - Examples: {female,male}, {ordered,sent,received}

- **ordinal:** finite domain with a linear ordering on the domain
  - Examples: {B.Sc., M.Sc., Ph.D.}

- **numerical:** values are numbers

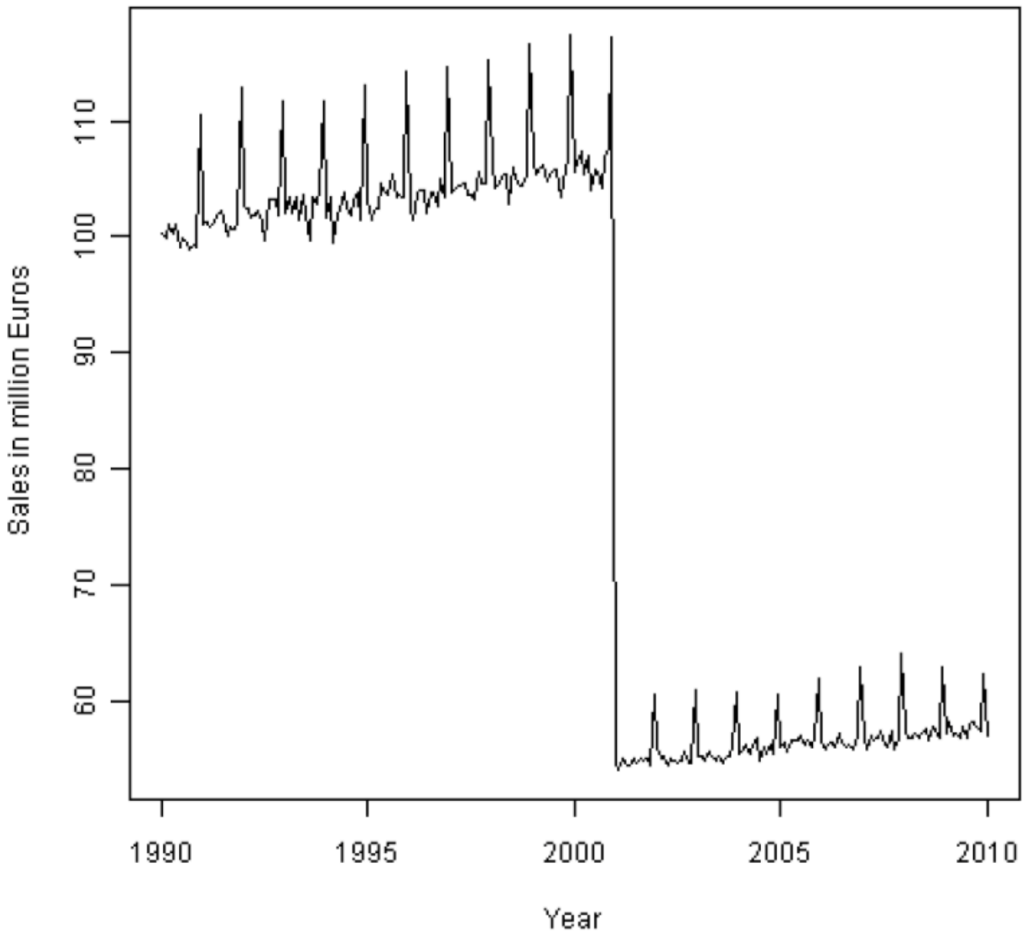
- **discrete**: categorical attribute or numerical attribute whose domain is a subset of the integer number.

- **continuous:** numerical attribute with values in the real numbers or in an interval

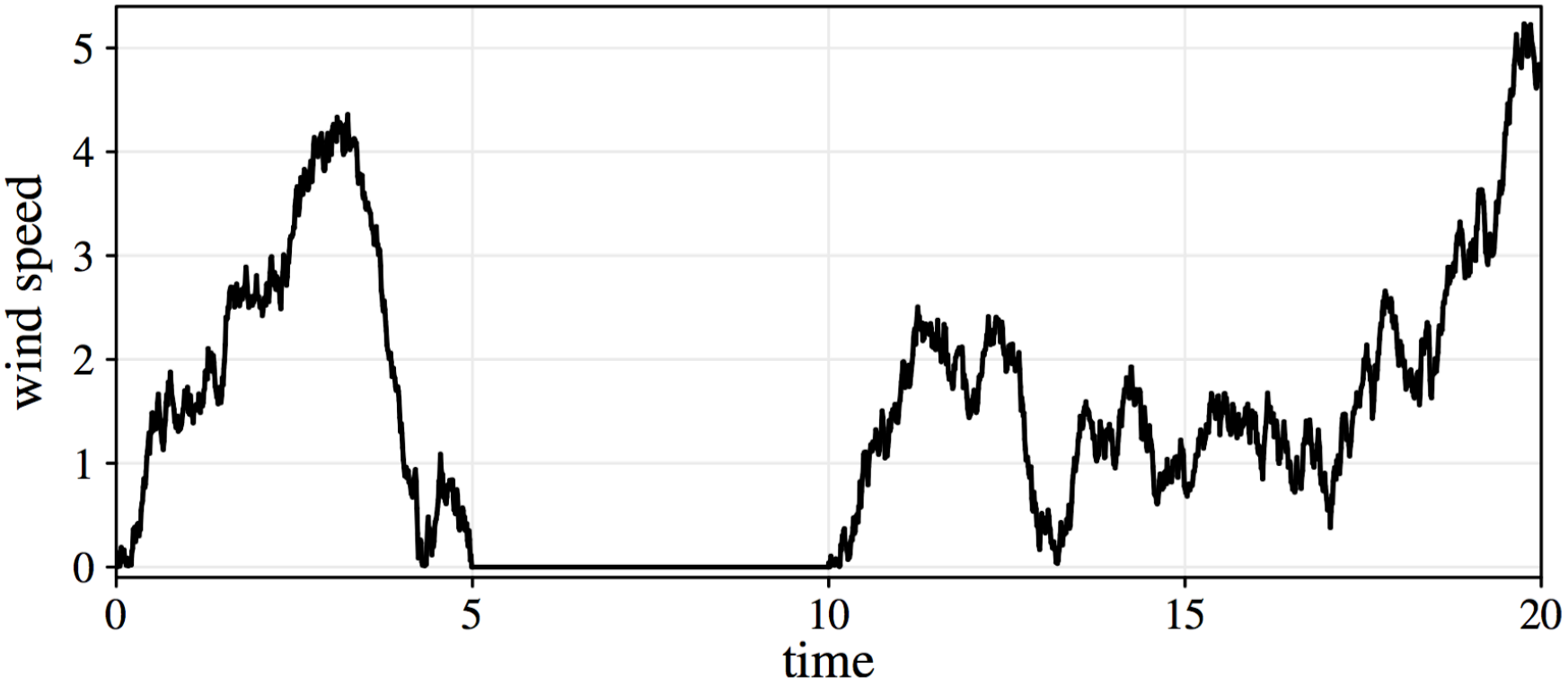
Data Quality

- **Syntactic accuracy:** Entry is not in the domain.
  - **Examples**: fmale in gender, text in numerical attributes, ... Can be checked quite easy.
- **Semantic accuracy:** Entry is in the domain but not correct
  - **Example**: John Smith is female
  - Needs more information to be checked (e.g. "business rules").
- **Completeness**: is violated if an entry is not correct although it belongs to the domain of the attribute.
  - **Example**: Complete records are missing, the data is biased (A bank has rejected customers with low income.)
- **Unbalanced data:** The data set might be biased extremely to one type of records.
  - **Example**: Defective goods are a very small fraction of all.
- **Timeliness:** Is the available data up to date?

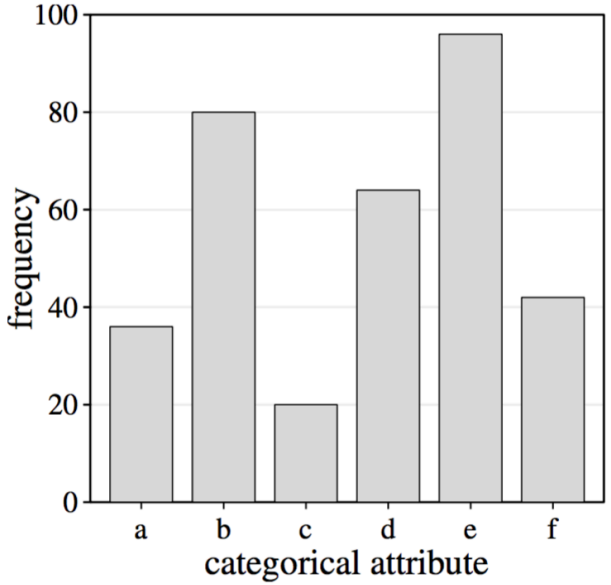
Data Visualization



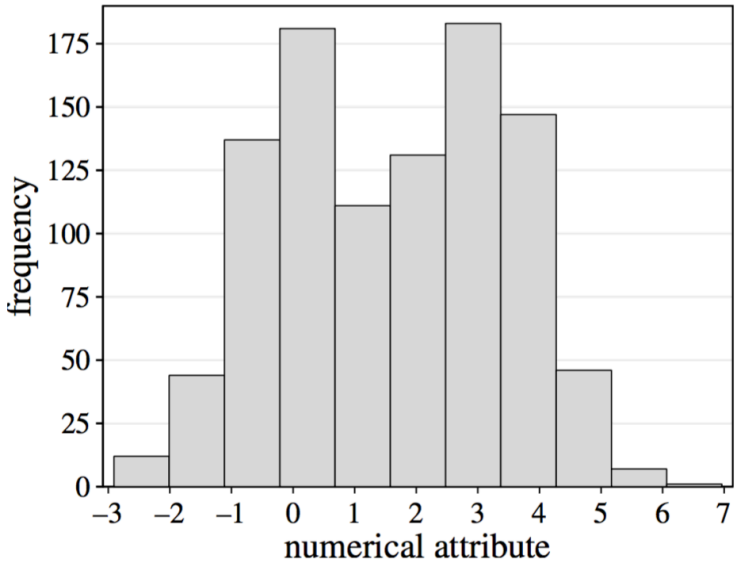
Data Visualization



Distributions



Bar Chart

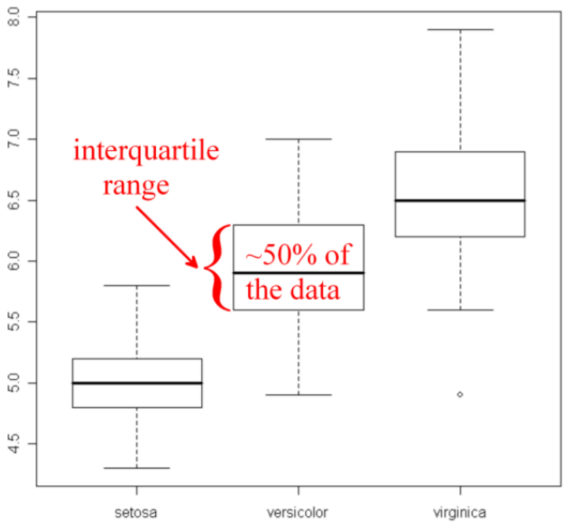


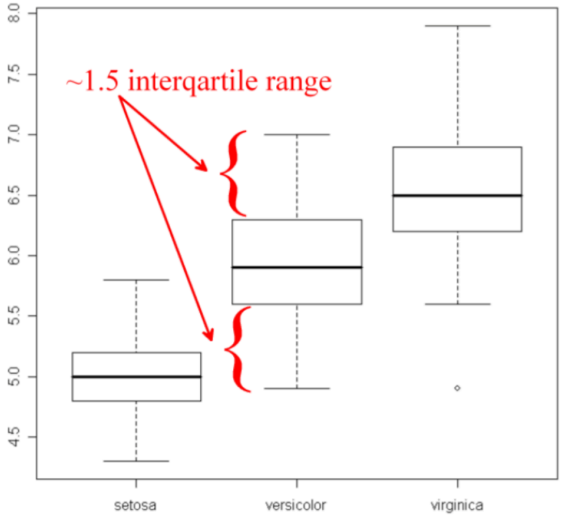
Histogram

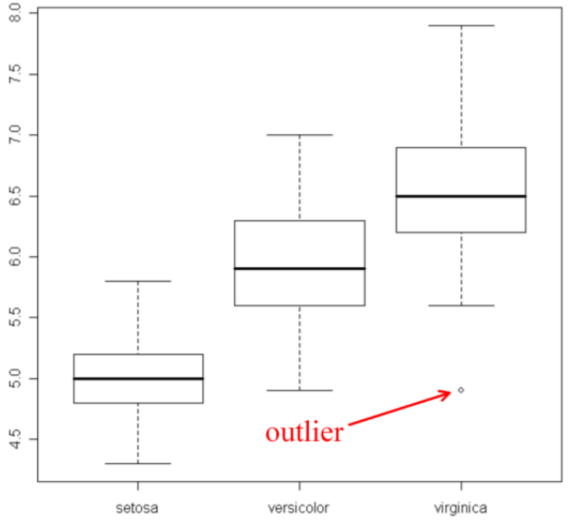
Statistics

- **Mean & Standard deviation**
  - Useful in determining the overall trend of a data set and disperision of the data (std dev)
- **q%-quantile** (0 < q < 100): The value for which q% of the values are smaller and 100-q% are larger.
- Median is the 50%-quantile.
- 25%-quantile (1st quartile), median (2nd quantile), 75%-quantile (3rd quartile).
- Interquartile range (IQR): 3rd quantile - 1st quantile.

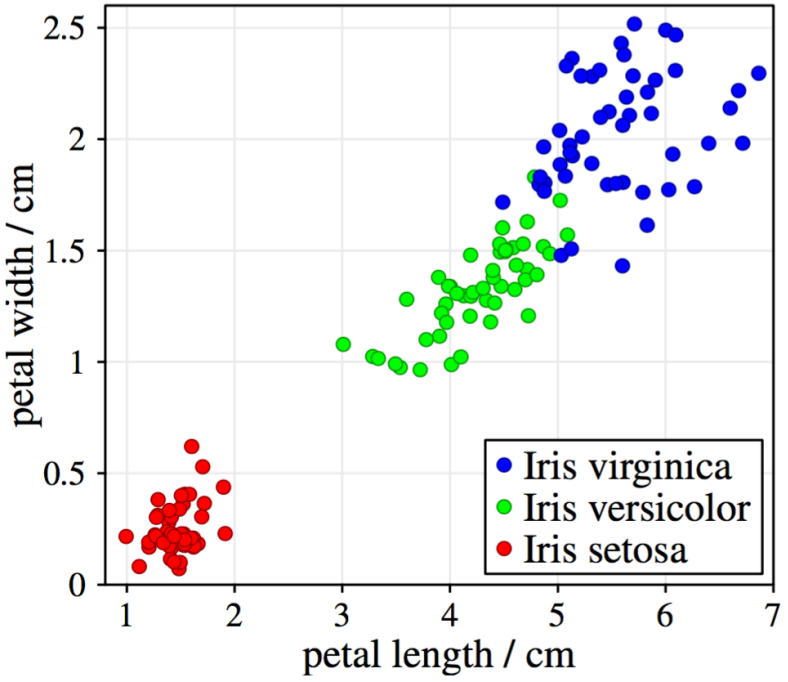
Box Plot

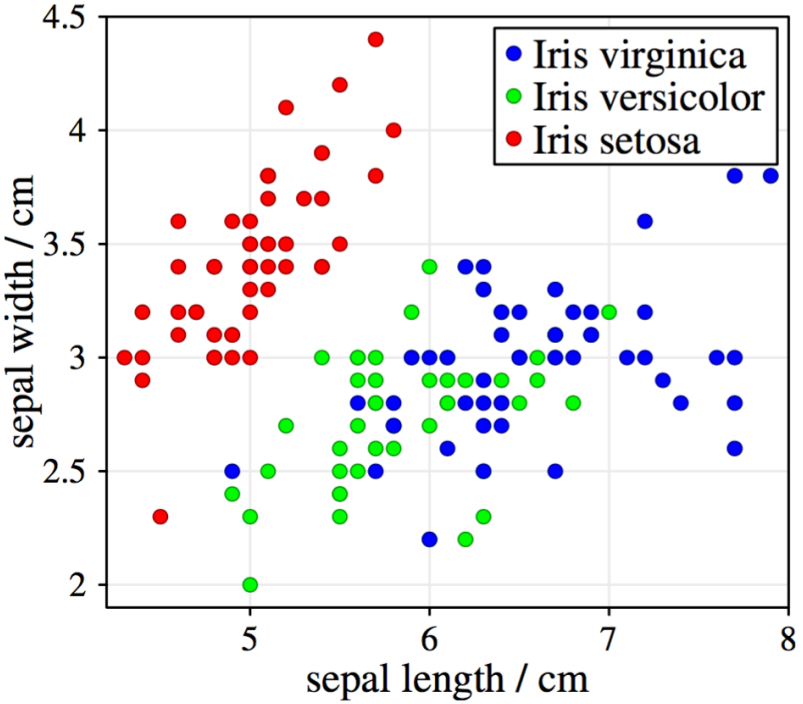

interquartile
range
~50% of
the data
setosa
versicolor
virginica


~1.5 interqartile range
setosa
versicolor
virginica

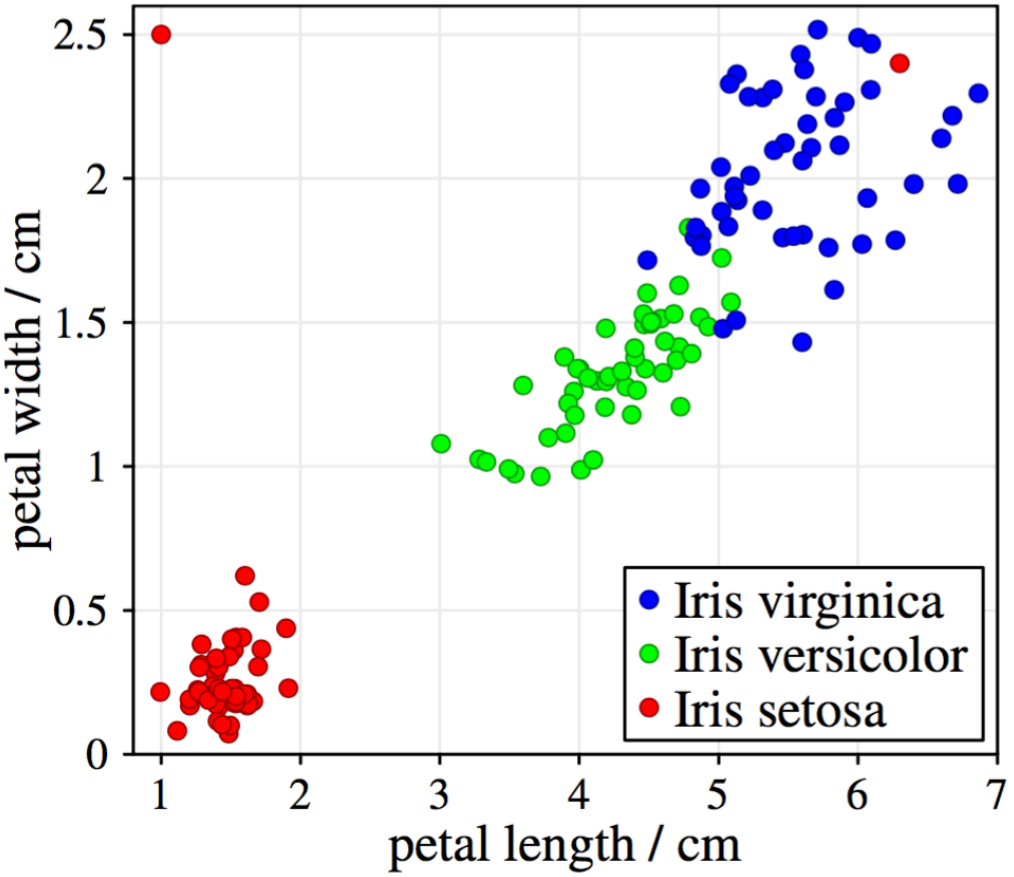

outlier
setosa
versicolor
virginica

Scatter Plot





Scatter Plot & Outliers

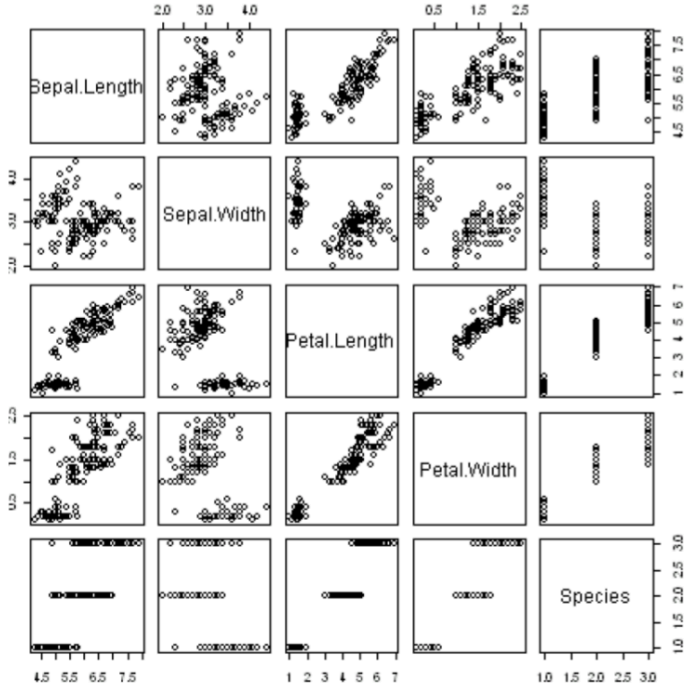


Pearson Correlation

The (sample) **Pearson's correlation** coefficient is a measure for a linear relationship between two numerical attributes X and Y and is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Pearson Correlation and Scatter Matrix



	sepal length	sepal width	petal length	petal width
sepal length	1.000	−0.118	0.872	0.818
sepal width	−0.118	1.000	−0.428	−0.366
petal length	0.872	−0.428	1.000	0.963
petal width	0.818	−0.366	0.963	1.000

Missing Values

- For some instances values of single attributes might be missing.
- Causes for missing values:
  - broken sensors
  - refusal to answer a question
  - irrelevant attribute for the corresponding object (pregnant (yes/no) for men)
- Missing value might not necessarily be indicated as missing (instead: zero or default values).

DATA CLEANING & PREPARATION

Manage Missing Values

- Elimination of record
- Substitution of values

  Note: it can influence the original distribution of values

  - Use mean/median/mode
  - Estimate missing values **using the probability distribution** of existing values
  - Data Segmentation and using mean/mode/median of each **segment**
  - Data Segmentation and using **the probability distribution within the segment**

Data Reduction

Reducing the amount of data

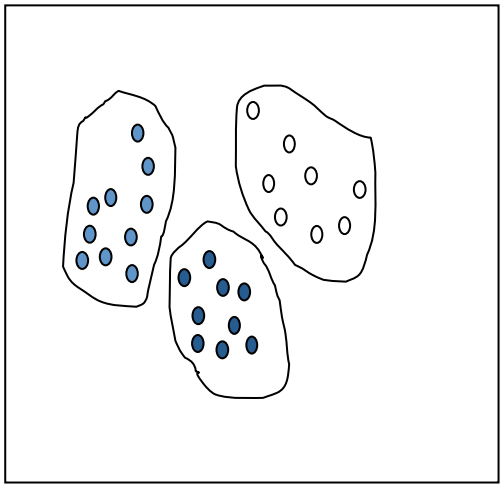
- Reduce the number of **records**
  - Data Sampling: select a subset of **representative** data
- Reduce the number of **columns** (attributes)
  - Select a subset of attributes
  - Generate a new (a smaller) set of attributes

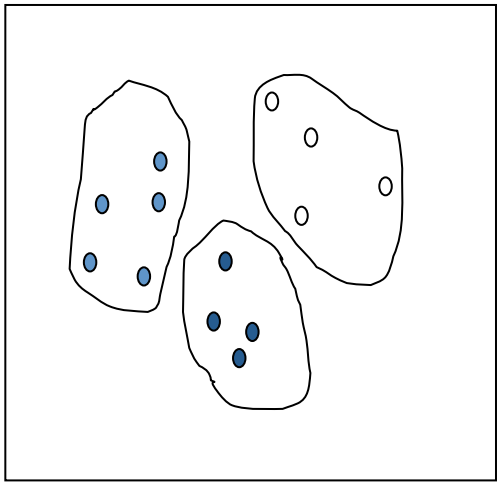
Sampling

- **Random sampling:** it can generate problems due to the possible peaks in the data

- **Stratified sampling**:
  - Approximation of the percentage of each class
  - Suitable for distribution with peaks: each peak is a **layer**

Raw Data





Stratified Sample

Removing irrelevant/redundant features

- For **removing irrelevant features**, a performance measure is needed that indicates how well a feature or subset of features performs w.r.t. the considered data analysis task

- For removing **redundant features**, either a performance measure for subsets of features or a correlation measure is needed.

Modeling Attributes

- Sometimes the attributes selected are **raw attributes**
  - They have a meaning in the original domain
- They could not be good enough to obtain accurate predictive models
- Need of a series of manipulation to generate new attributes showing better properties that will help the predictive power of the model

The new attributes are usually named

modeling variables or analytic variables.

Data Normalization

Min-Max Normalization

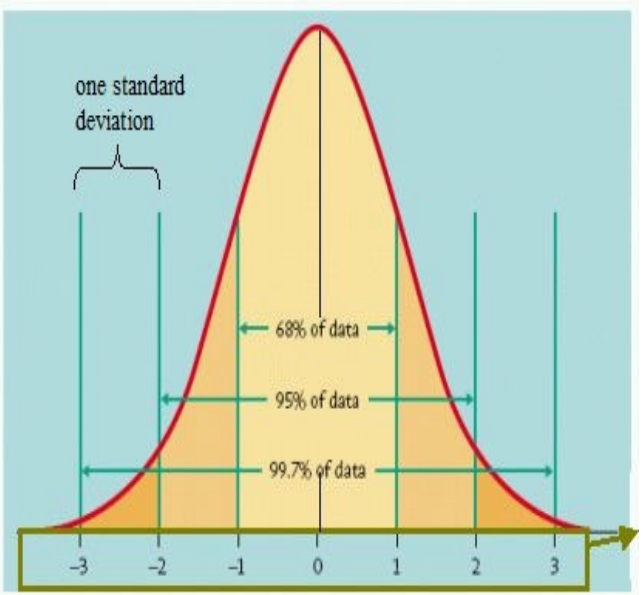
- The min-max normalization aims to scale all the numerical values *v of a numerical* attribute *A to a specified range denoted* by [new − minA, new − maxA].
- The following expression transforms *v to* the new value *v'*:

$$v' = \frac{v - min_A}{max_A - min_A}(new - max_A - new - min_A) + new - min_A$$

Data Normalization

Z-score Normalization

- If minimum or maximum values of attribute *A* are not known, or the data is noisy, or is skewed, the *min-max* normalization is good
- Alternative: normalize the data of attribute A to obtain a new distribution with mean 0 and std. deviation equal to 1



$$v' = \frac{v - \overline{A}}{\sigma_A}.$$