III Data types

Many ways to characterize data types

- structured or unstructured
- dependency-oriented or nondependency-oriented
- numerical, categorical or mixed
- static ↔ temporal; spatial; spatio-temporal

Structured vs. Unstructured

- **Structured**
  - has a predefined structure (e.g., rows and features)
  - e.g., multidimensional, graph-formed, time series
- **Unstructured**
  - no pre-defined format, just a string
  - e.g., text, audio, video, signal data
- **Semistructured**
  - contains internal tags that identify separate data elements
  - e.g., XML documents, emails

Dependency-orientation

- Nondependency-oriented: no specified dependencies between objects or attributes
- Dependency-oriented: data objects or values related temporally, spatially or through network links
  1. **explicit dependencies**
     - relationships in graph or network data
  2. **implicit dependencies**
     - known to typically occur
     - e.g., consecutive temperature readings likely similar

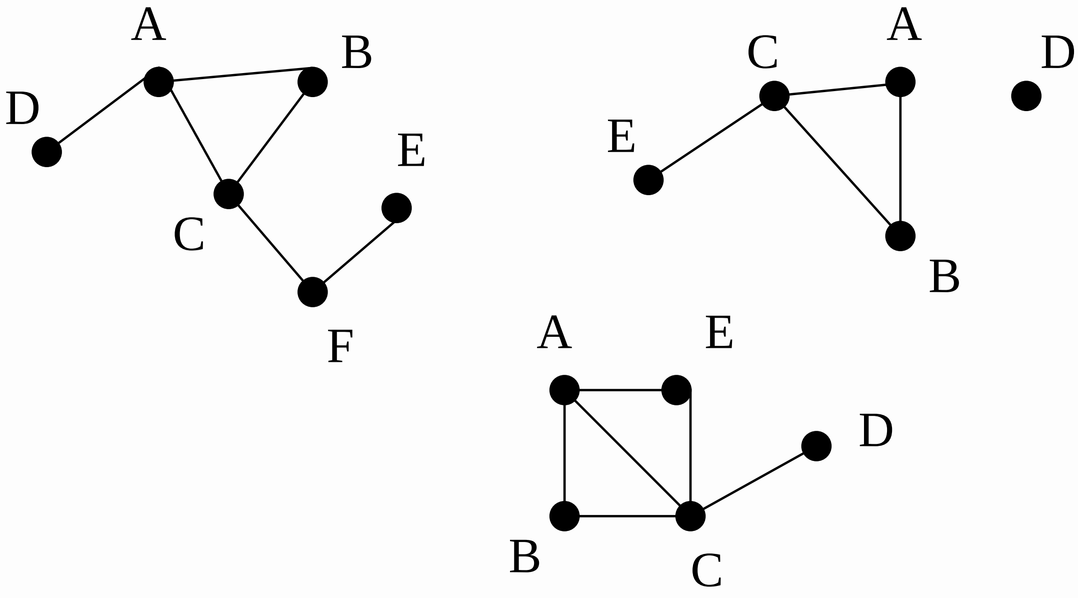
Difference: dependencies in data type vs. patterns in data instances

DATA TYPE ↔ DATA INSTANCES

graph

Dependencies in data structure:

edges present relationships



Discovered dependency: clique of A, B and C

occurs frequently

Implicit dependencies harder to separate from patterns!

Basic data type: Multidimensional data

- a set of records, whose fields are features
- notate $\mathcal{D} = \{\overline{X_1}, \ldots, \overline{X_n}\}$, where $\overline{X_i} = (x_i^1, \ldots, x_i^d)$
  - n rows (records, data points, instances, objects) and d features (fields, attributes, dimensions)
- suitable for a relational database, e.g., cow data:

name	race	weight	parity	milk/d	activity
Rose	Holstein	640	2	35	4800
Daisy	Ayrshire	675	3	37	5100
Strawberry	Finncattle	615	4	28	7200
Molly	Ayrshire	650	1	32	6300

Numerical, categorical or mixed?

Depending on the type of variables, data may be called numerical (quantitative), categorical or mixed (both).

Variables can be classified by measurement scales:

1 Categorical

1.1 **Nominal**: values are only labels, **no order**

- e.g., gender (binary), colour, home city, occupation
- mode (most common value) is defined

1.2 **Ordinal**: values have an **order**

- e.g., satisfaction with services: very unsatisfied, unsatisfied, neutral, satisfied, very satisfied
- mode and median (the middle value) defined

Measurement scales (cont'd)

2 Numerical

2.1 **Interval scale**: difference between values is defined, but **not ratio**

- no true zero point
- temperature 20°C is not twice as warm as 10°C!
- mean and standard deviation defined

2.2 **Ratio scale**: also **ratio** is defined

- absolute zero = absence of the measured property
- temperature in Kelvins, length, weight, duration
- mean, standard deviation, geometric mean ($(\prod x_i)^{1/n}$), coefficient of variation (σ/μ) defined

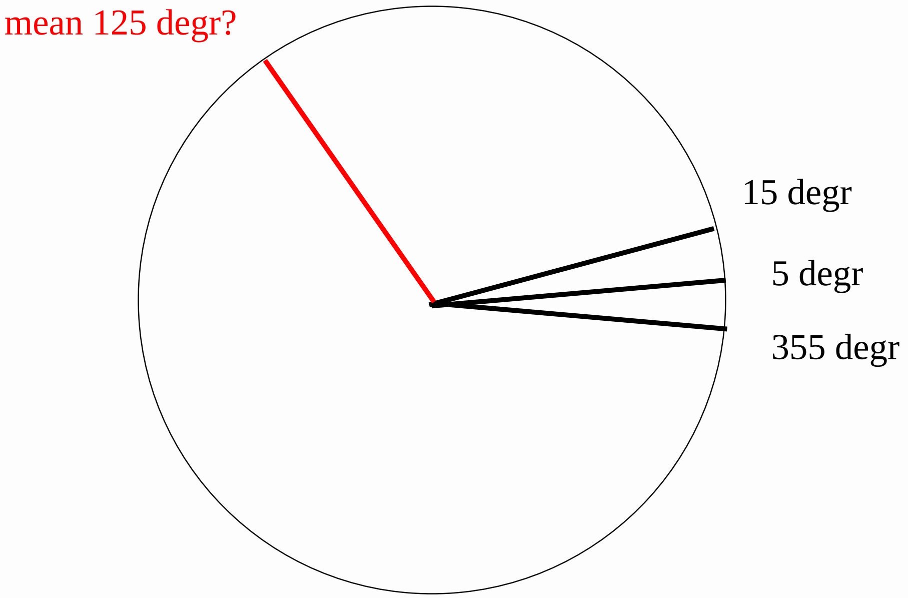
Circular variables

Idea: Values are ordered categories, where the last category precedes the first

1. Interval circular
   - e.g., compass direction (angles), time of day, day of year
   - zero on the measurement scale not meaningful!
2. Ordinal circular
   - e.g., days of the week (Mon, Tue,...), compass aspect (N, NE, E,...)

Be careful! E.g, cannot calculate arithmetic mean or normal correlation.

Example: What is the mean angle??



NO! Real mean angle 5 degr!

- present angles α_i by $(cos(\alpha_i), sin(\alpha_i))$
- $S = \sum_i sin(\alpha_i)$, $C = \sum_i cos(\alpha_i)$
- $\theta = arctan\left(\frac{S}{C}\right)$, if $S \geq 0$, $C > 0$
- $\theta = arctan\left(\frac{S}{C}\right) + \pi$, if $C < 0$
- $\theta = arctan\left(\frac{S}{C}\right) + 2\pi$, if $S < 0$, $C \leq 0$
- $\theta = \pi/2$, if $S > 0$, $C = 0$
- undefined, if $S = 0$, $C = 0$

Present other circular variables first as angles (e.g., $\alpha = \frac{h*2\pi}{24}$)

Warning: Number codes $\neq$ numerical variables

Categorical values have often arbitrary numerical codes that can't be interpreted as numbers!

Gender: 1 = Female, 2 = Male

Cow's race: 0 = Holstein, 1 = Ayrshire, 2 = Finncattle

- cannot measure distance or ratio or calculate mean or Pearson correlation
- you can get numerical presentation by creating dummy (binary indicator) variables for each value
  - e.g., $I_{Holstein}$=1, if race=Holstein, and 0 otherwise

Warning (cont’d)

The same holds for ordinal variables:

Opinion: 1 = fully disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = fully agree

- if fully ordinal and distances between categories equal, variable may be treated as numerical (but not always optimal)
- more typical when many categories (≥ 7)
- Be careful!

Opinion: 0 = Don’t know, 1 = fully disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = fully agree

Other data types

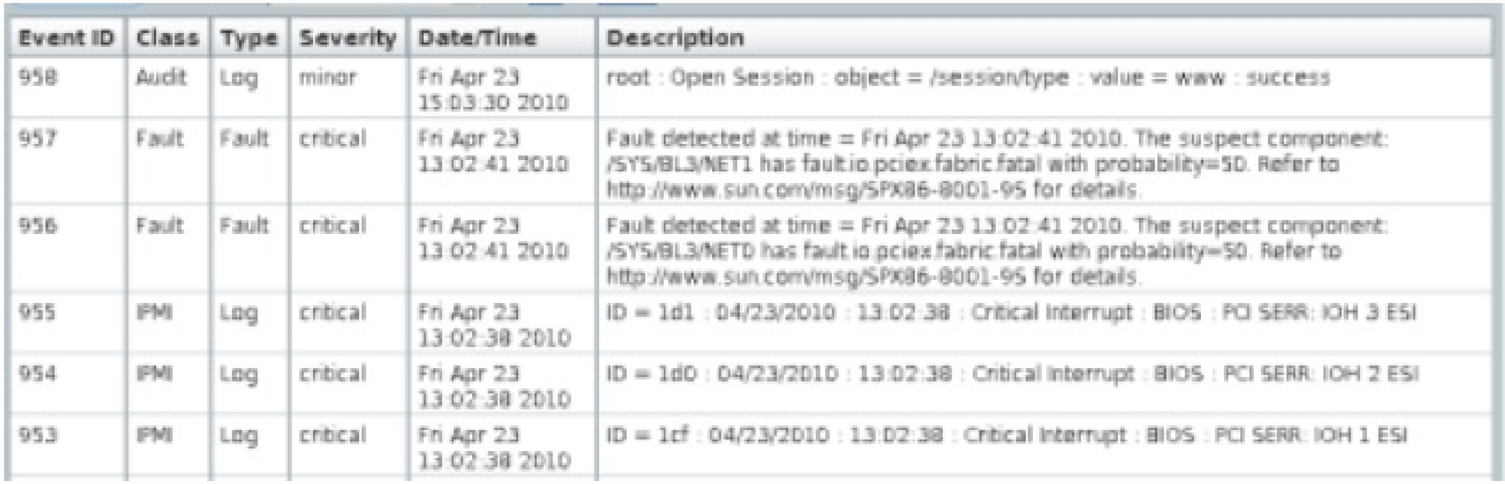
- time series
- discrete sequences
- spatial data
- network and graph data
- text

Time series

- continuous measurements over time
- e.g., from environmental sensors, health monitoring devices, ECG
- at time stamps $t_1, \ldots, t_n$ measurements $(Y_1, \ldots, Y_n)$
- may also be multivariate time series $(\overline{Y_1}, \ldots, \overline{Y_n})$, where $\overline{Y_i} = (y_i^1, \ldots, y_i^d)$
- e.g., heart rate, oxygen saturation, diastolic and systolic blood pressure at every minute
- often temporal correlations (like dependencies between consecutive values or periodic patterns)

Discrete sequences

- like time series, but sequences of categorical variables
- special case: strings (no time stamps, but positions)
- e.g., event logs, strings of nucleotides (DNA, genes)



Event ID	Class	Type	Severity	Date/Time	Description
958	Audit	Log	minor	Fri Apr 23 15:03:30 2010	root : Open Session : object = /session/type : value = www : success
957	Fault	Fault	critical	Fri Apr 23 13:02:41 2010	Fault detected at time = Fri Apr 23 13:02:41 2010. The suspect component: /SYS/BL3/NET1 has fault.io.pciex.fabric.fatal with probability=50. Refer to http://www.sun.com/msg/SPX86-8001-9S for details.
956	Fault	Fault	critical	Fri Apr 23 13:02:41 2010	Fault detected at time = Fri Apr 23 13:02:41 2010. The suspect component: /SYS/BL3/NET0 has fault.io.pciex.fabric.fatal with probability=50. Refer to http://www.sun.com/msg/SPX86-8001-9S for details.
955	IPMI	Log	critical	Fri Apr 23 13:02:38 2010	ID = 1d1 : 04/23/2010 : 13:02:38 : Critical Interrupt : BIOS : PCI SERR: IOH 3 ESI
954	IPMI	Log	critical	Fri Apr 23 13:02:38 2010	ID = 1d0 : 04/23/2010 : 13:02:38 : Critical Interrupt : BIOS : PCI SERR: IOH 2 ESI
953	IPMI	Log	critical	Fri Apr 23 13:02:38 2010	ID = 1cf : 04/23/2010 : 13:02:38 : Critical Interrupt : BIOS : PCI SERR: IOH 1 ESI

Figure from `https://docs.oracle.com/cd/E19140-01/html/821-0796/gjfwa.html`

Difficulty: how to combine temporal data when the measuring frequency varies?

Example from a cow-house:

- body temperature and rumen acidity are measured every minute
- activity device records average activity every 15 min
- milk production (amount, protein and fat contents etc.) is measured daily
- feeding automaton event log contains time stamp, automaton id, cow id, feed type, amount and duration for every visit
- drinking automaton event log contains time stamp, cow id, amount of water and duration

Spatial and spatiotemporal data

- spatial: measurements of non-spatial attributes in spatial locations (typically 2D)
  - e.g. sea surface temperature
- spatiotemporal data
  - e.g., temperature over time or ship trajectories

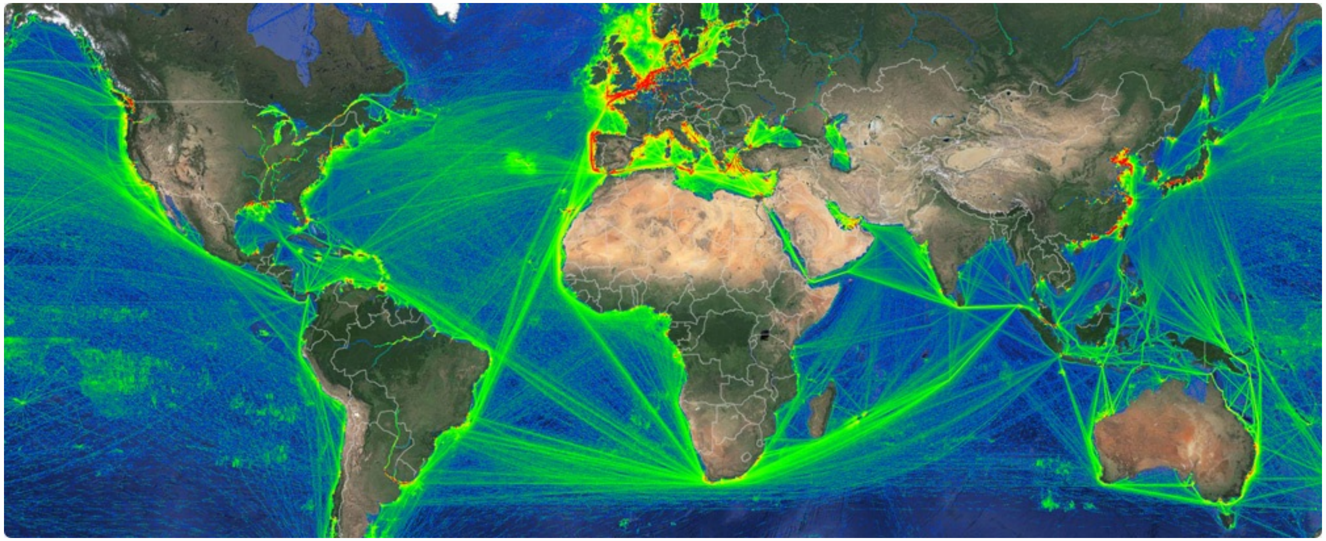


Figure from `http://www.elane.com/EN/Detail106.html`

Spatiotemporal data: contextual and behavioural attributes

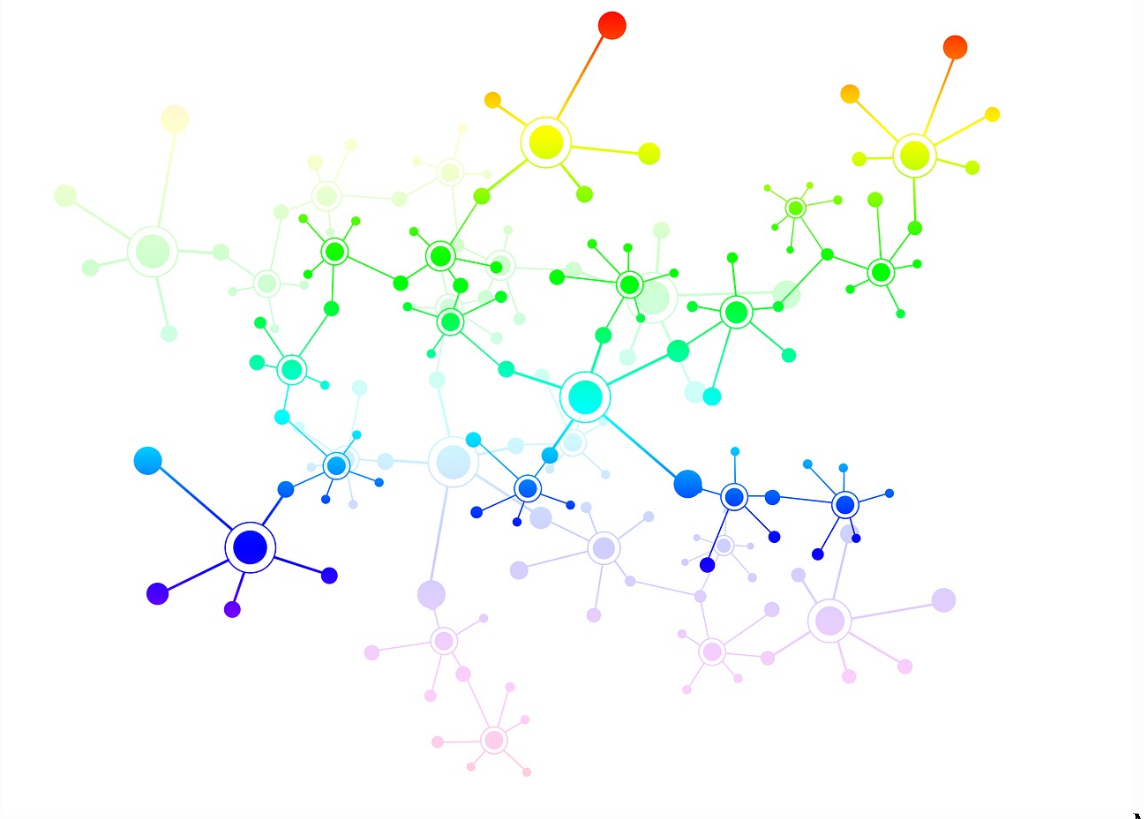
Contextual attributes define the context
Behavioural attributes are measured in this context

Two main types of spatiotemporal data:

1. Both spatial and temporal attributes define the context where some behavioural attribute (like temperature) is measured
2. Temporal attribute is contextual and spatial attributes are behavioural (e.g., trajectory analysis)

Network and graph data

- nodes correspond objects and edges relationships
  + attributes may be associated with nodes or edges
- directed (web structure) or undirected (social network)



Example: wikipedia hyperlink structure

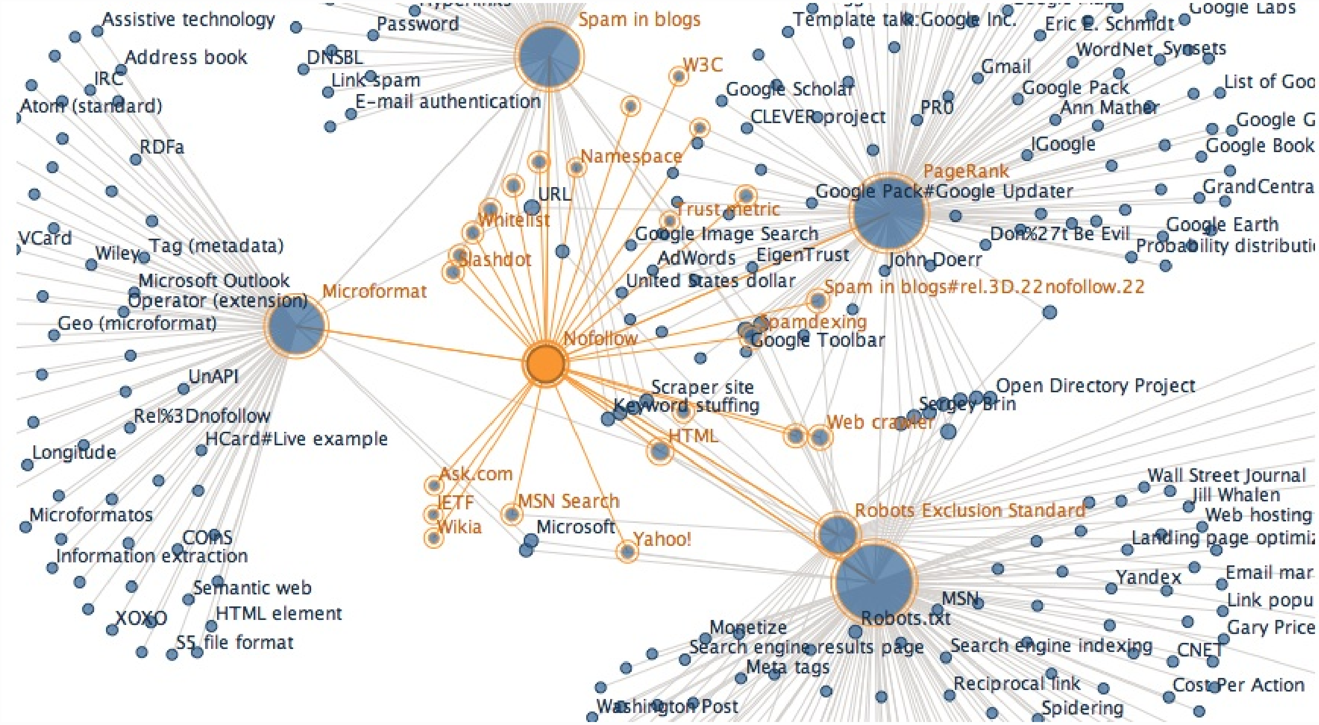
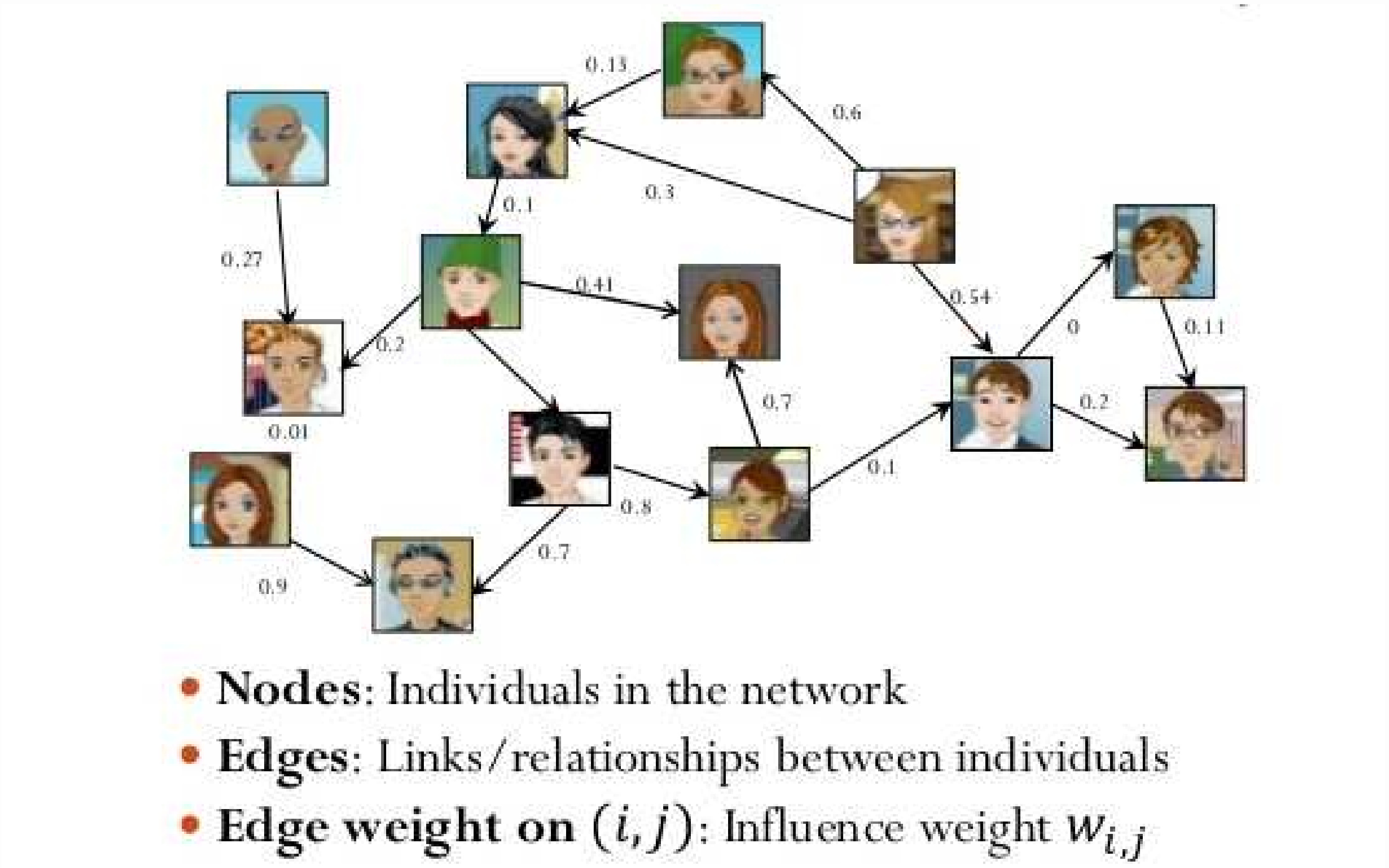


Figure from https://wiki.digitalmethods.net/Dmi/WikipediaAnalysis

Example: social network structure



- **Nodes**: Individuals in the network
- **Edges**: Links/relationships between individuals
- **Edge weight on** (i, j): Influence weight $w_{i,j}$

Source: Lu and Lakshmanan ICDM 2012 `https://www.slideshare.net/WeiLu12/profit-maximization-over-social-networks`

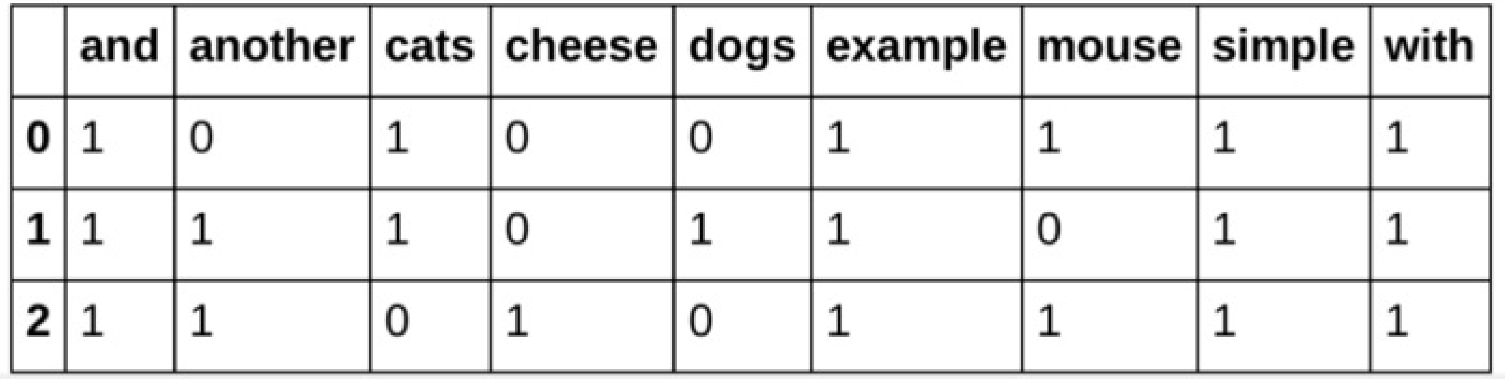
Text data

- raw text is a string, i.e., dependency-oriented
- often represented as a **bag-of-words** or **document-term matrix** (nondependency-oriented)
- which can be presented in vector space (as multidimensional data)
  - how often terms occur in document? ⇒ numerical features for term frequencies
  - ⇒ often transformed to tf-idf values (contains weighting + log scaling)

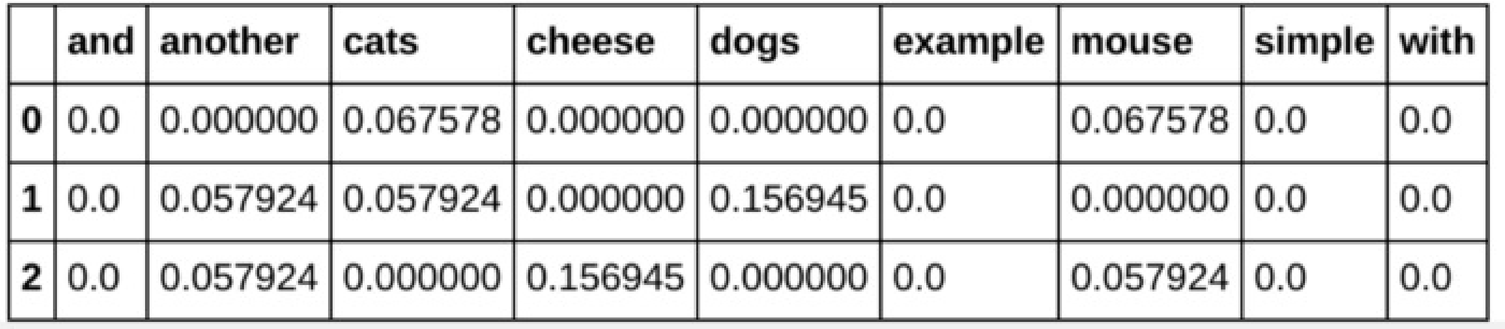
More on the text mining lecture!

Example: tf-idf presentation of sentences

d0: Simple example with cats and mouse
d1: Another simple example with dogs and cats
d2: Another simple example with mouse and cheese



	and	another	cats	cheese	dogs	example	mouse	simple	with
0	1	0	1	0	0	1	1	1	1
1	1	1	1	0	1	1	0	1	1
2	1	1	0	1	0	1	1	1	1



	and	another	cats	cheese	dogs	example	mouse	simple	with
0	0.0	0.000000	0.067578	0.000000	0.000000	0.0	0.067578	0.0	0.0
1	0.0	0.057924	0.057924	0.000000	0.156945	0.0	0.000000	0.0	0.0
2	0.0	0.057924	0.000000	0.156945	0.000000	0.0	0.057924	0.0	0.0

Example from `https://medium.com/@MSalnikov/text-clustering-with-k-means-and-tf-idf-f099bcf95183`