Data Types and Formats

DR. AMI GATES

Understanding Data Types and Levels of Measurement

Before conducting any analysis, we need to:

- Understand what **type of data** we have – **qualitative versus quantitative**.
- Understand the **different representations** of the data. (M, male, MALE, 0)
- Determine whether the different values are **discrete or continuous**
- Determine if the attribute values are **nominal, ordinal, interval, ratio**, etc.
- Preprocess the data for data analytics

Levels of Measurement: Examples

- Nominal
  - Examples: ID numbers, eye color, zip codes
- Ordinal
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
- Interval
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- Ratio
  - Examples: temperature in Kelvin, length, time, counts

- The types of transformations allowed will differ. For example, transformations on ordinal data must preserve the actual order.

Nature of Data - Record Data

Data that consists of a collection of records, each of which consists of a fixed set of attributes - **what types and levels are here**?

ID	Birth Date	Marital Status	Income	Height
1	3/12/1966	Single	125K	tall
2	2/17/1945	Divorced	100K	short
3	1/12/1990	Married	120K	short
4	7/13/1985	Single	90K	medium
5	8/30/1994	Single	150K	tall

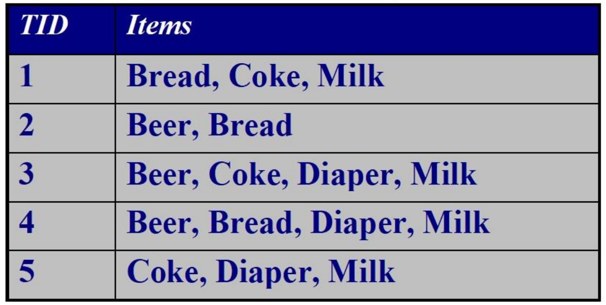
Nature of Data – Word Count in Documents

Each document can be viewed as a **term (word)** vector. In a term vector:

- Each term is a component (attribute/variable) of the vector,
- The value of each component is the number of times the corresponding word occurs in the document.
- The first vector here is <3, 0, 5, 0, 1, 4, 0, 1, 1, 3>

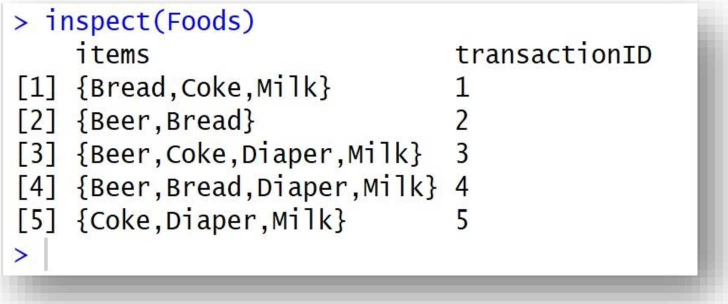
	dog	cat	bark	sleep	eat	play	ball	tree	fall	bird
DOC 1	3	0	5	0	1	4	0	1	1	3
DOC 2	0	10	0	2	2	1	0	0	0	3
DOC 3	0	1	0	0	3	4	1	5	0	0

Transaction Data



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

Introduction to Data Mining, 2nd Edition



```
> inspect(Foods)
    items                     transactionID
[1] {Bread,Coke,Milk}         1
[2] {Beer,Bread}              2
[3] {Beer,Coke,Diaper,Milk}   3
[4] {Beer,Bread,Diaper,Milk}  4
[5] {Coke,Diaper,Milk}        5
>
```

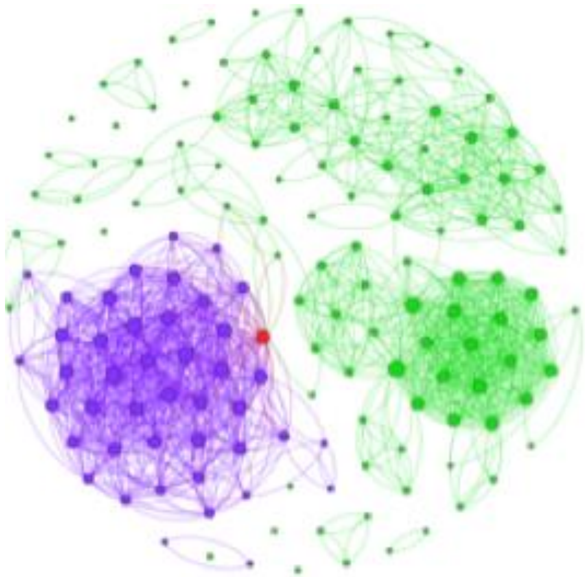
1	Bread
1	Coke
1	Milk
2	Beer
2	Bread
3	Beer
3	Coke
3	Diaper
3	Milk
4	Beer
4	Bread
4	Diaper
4	Milk
5	Coke
5	Diaper
5	Milk

TID	Bread	Coke	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

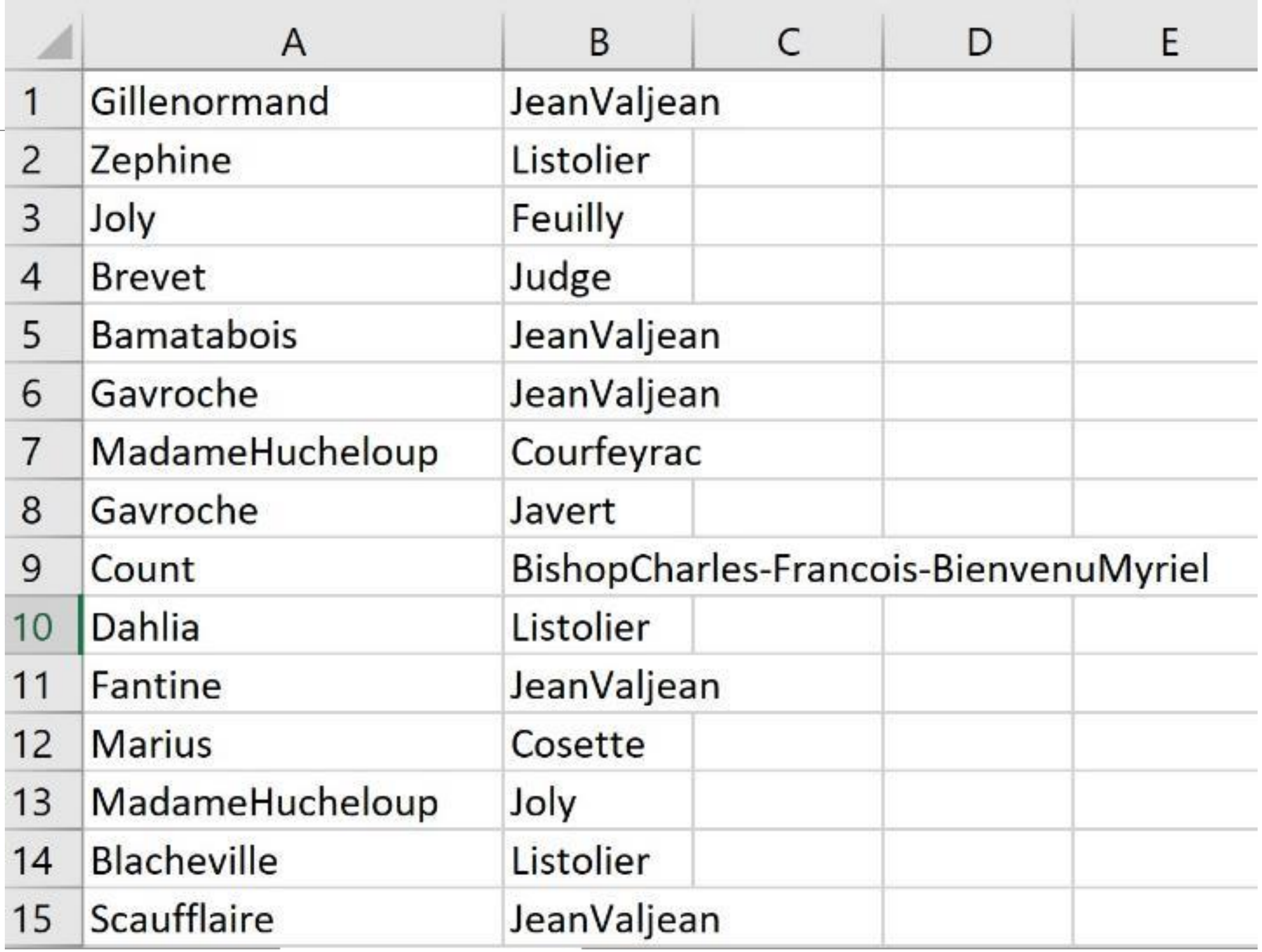
Nature of Data – Graph/Network Data

A graph can be generated from any data that has objects and connections between those objects.

Network data must be clearly define vertices (nodes) and relationships (edges).

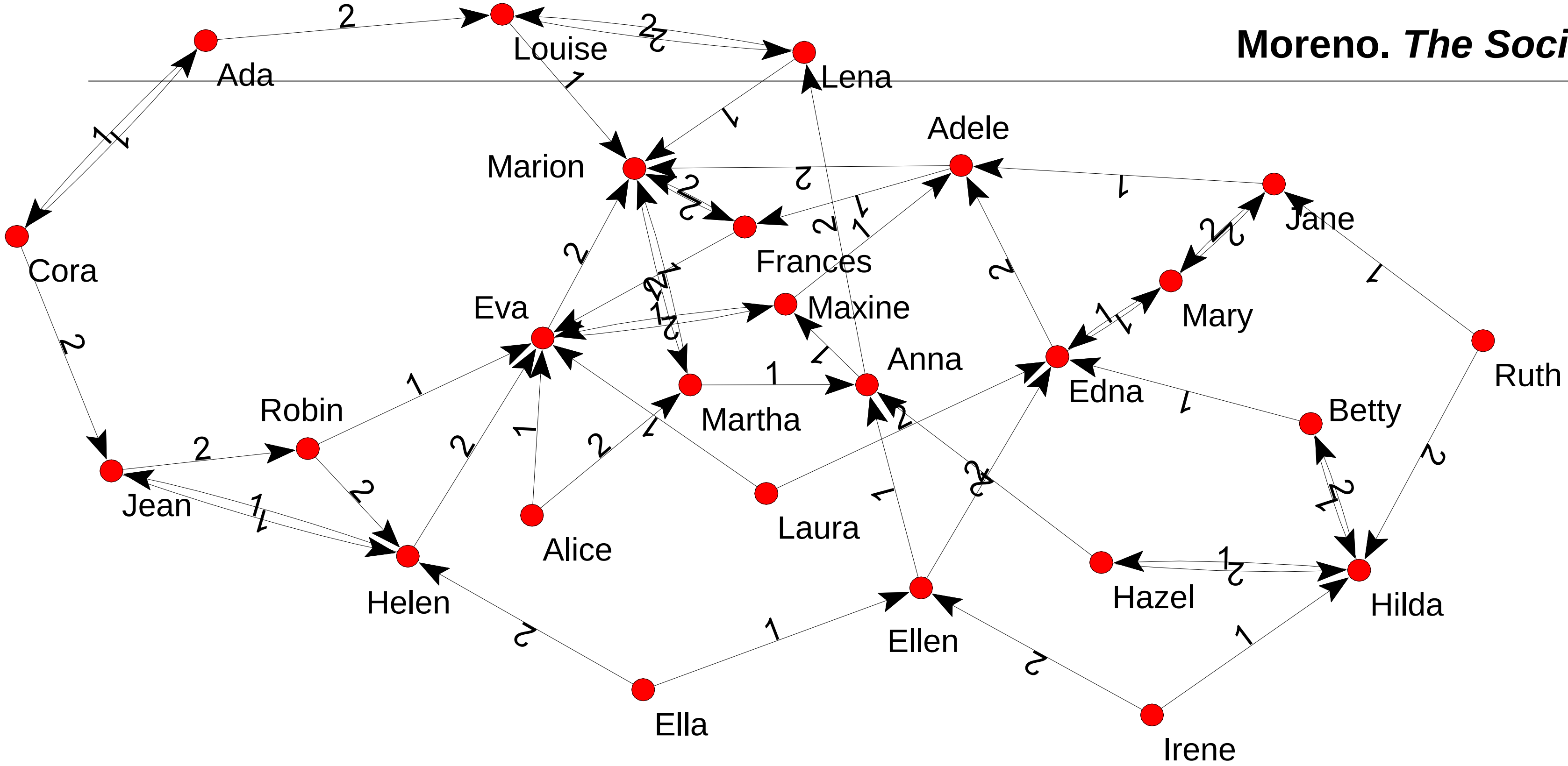


Network Data Example



	A	B	C	D	E
1	Gillenormand	JeanValjean			
2	Zephine	Listolier			
3	Joly	Feuilly			
4	Brevet	Judge			
5	Bamatabois	JeanValjean			
6	Gavroche	JeanValjean			
7	MadameHucheloup	Courfeyrac			
8	Gavroche	Javert			
9	Count	BishopCharles-Francois-BienvenuMyriel			
10	Dahlia	Listolier			
11	Fantine	JeanValjean			
12	Marius	Cosette			
13	MadameHucheloup	Joly			
14	Blacheville	Listolier			
15	Scaufflaire	JeanValjean			

Network Example



Moreno. *The Sociometry Reader*. 1960

- Girls' school dormitory dining-table **partner choices** (this is the relationship)
- First and second choices shown as weighted edges.
- The Girls are the nodes (vertices)

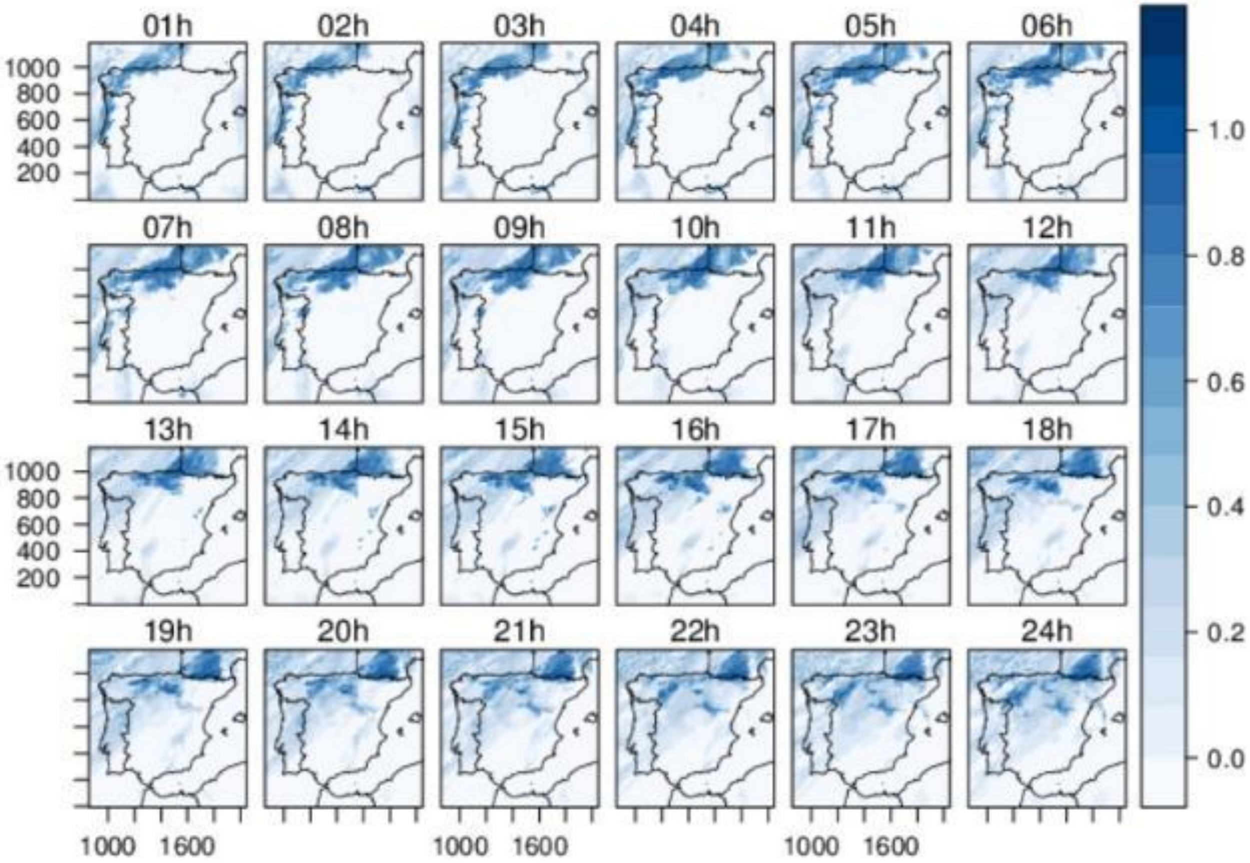
Nature of Data –
Ordered Data

Data that contains an **ordering** that is important to preserve.

One example is **genomic data**.
The order is critical.

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

Nature of Data – Spatio-Temporal Data



http://oscarperpinan.github.io/spacetime-vis/spacetime.html

Nature of Data – Numeric Data Matrix Concept

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.

longitude	latitude	temperature	humidity	pressure
139	135	89.5	78	1013
200	33	12.5	32	244

as a 2 by 5 matrix

139	**135**	**89.5**	**78**	**1013**
200	**33**	**12.5**	**32**	**244**

Where we get data?

1) **Experiments**

2) **Observational studies**

3**) API Data Gathering**

4) Public data examples:

- Government agencies, e.g. Census Bureau
    - CDC, Bureau of Labor Statistics, …
- Online statistics
- Online markets, e.g. stock market
- Companies built on open data, e.g. Twitter, White Pages, Wikis
- Public profile information, e.g. LinkedIn, Facebook
- Online newspapers/blogs
- Public image galleries
- Kaggle