Information Visualization

Data sense making (2)

Lesson 6

Marilena Daquino
Assistant Professor

Department of
Classical Philology
and Italian Studies

marilena.daquino2@unibo.it

Table of contents

01 Other selected tasks

Geo-spatial, topical, network analysis

02 Hands-on

Data analysis with python

01

Selected tasks

Maps, flows, and graphs

Statistics

Profiling (micro-meso-macro level)

Temporal

WHEN: evolution of variables over time

Geo-spatial

WHERE: trajectories and space dimension of variables

Topical

WHAT: analysis of categorical variables

Network

WITH WHOM: relations and distance between data points

01

Geospatial analysis

Maps maps maps!

Geographic maps

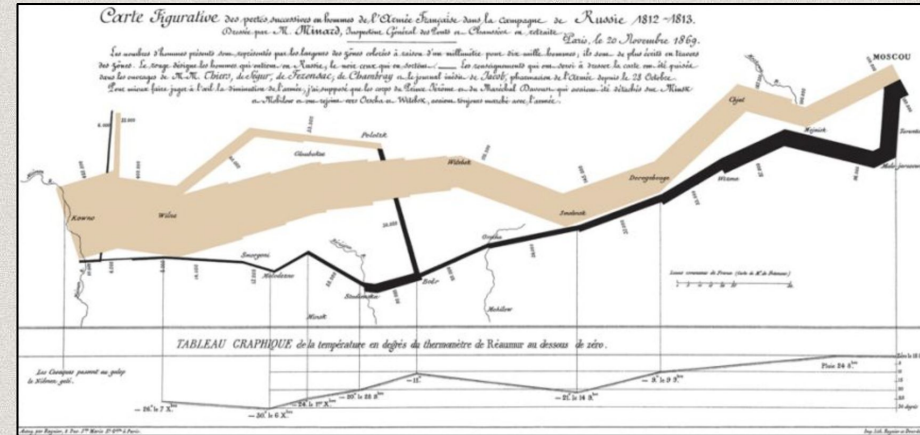## Purpose

_____

From the simple representation of entities on a map, e.g. **pin** locations on a map, to more complex tasks, e.g. trace the path of historical events.

The interpretation of the map highlights patterns (e.g. distribution, dispersion, clusters, trajectories)

We see this today!

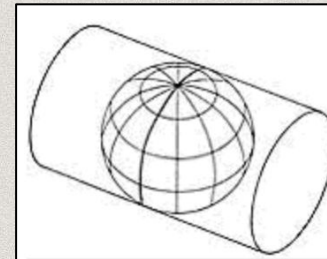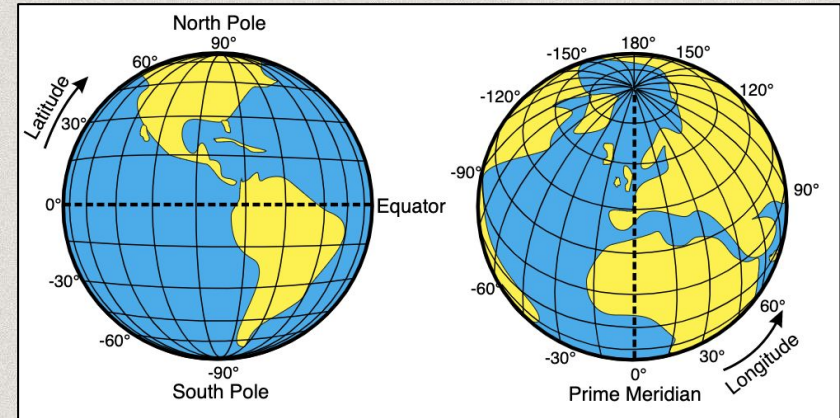*The advance (tan) and disastrous retreat (black) of Napoleon's Grande Armee through Russia.*

Geographic maps

Fundamentals

Data must be accompanied by geospatial information, such as **latitude and longitude** used to position a **shape** on an abstraction of the earth, based on a chosen **projection** system.

Universal Transverse Mercator (UTM) Grid. The equator is used as reference point for defining North and South.

Geographic maps

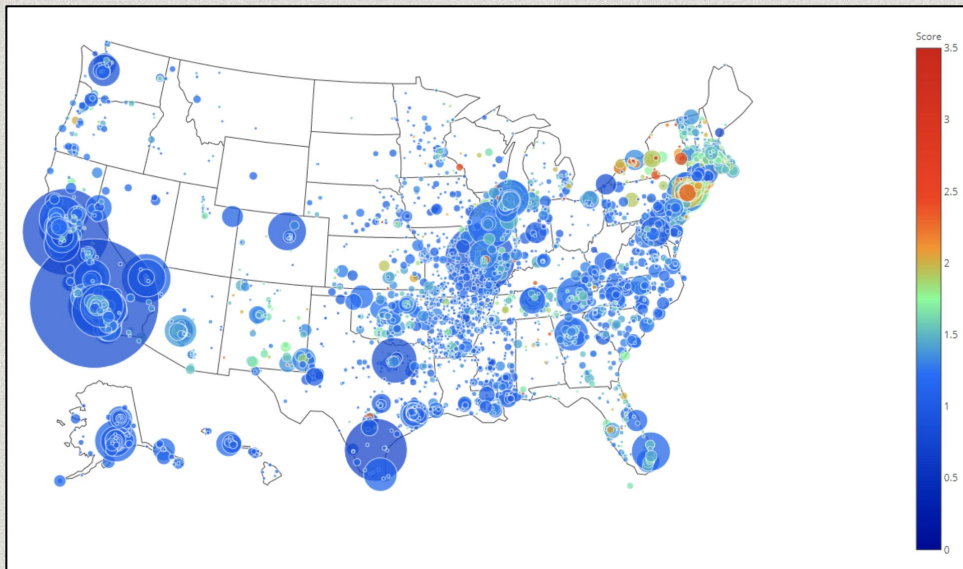## Fundamentals
_____

A shape can be a **point** (or a glyph), identified by two coordinates placed on cartesian axes

x=long; y=lat

Other variables can be represented by the size (numeric), the color (categorical), or the gradation (ordinal).
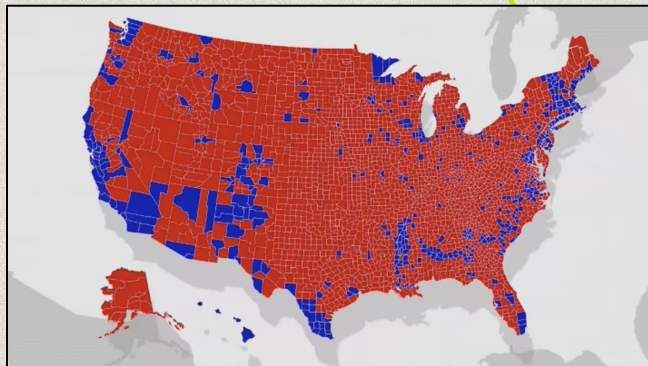
Geographic maps

Geospatial analysis
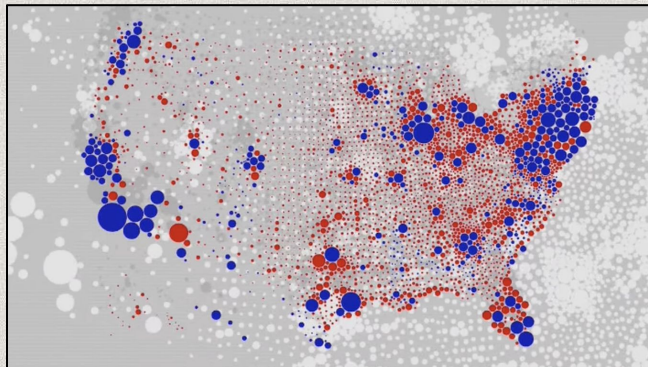
Fundamentals

A shape can be an **area**, colored according to categorical and numeric values (e.g. US election votes).

However, areas can be misleading, since larger areas are associated to high numbers (but it might not be the case).

2020 US elections, countries by winning party.

2020 US elections, countries by number of voters.

O1

Topical analysis

Machine learning tasks

Topical analysis

Topical analysis

Purpose

Analyse the **occurrences** of a topic in texts, images, or structured data to understand distribution, dependencies with other topics, correlation, etc.

Typical tasks include **classification** (e.g. of documents, images), **recommendation** (e.g. similar docs), and **prediction** (e.g. new docs).

Topical analysis is often used to refer to news and social media analysis, where occurrences of a topic in media is analysed to understand a phenomenon (e.g. propaganda, fake news, terrorism).

Topical analysis

Methods

Mainly based on Machine Learning algorithms

- **Texts**. Natural Language Processing
- **Images**. Computer vision
- **Structured data**. Other probabilistic and rule-based models
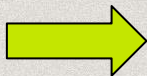
Natural Language Processing

Basic tasks

- **Tokenization.** Extraction of single words from sentences.
- **Lemmatization.** Extraction of lemmas from words
- **Part Of Speech (POS) tagging.** Annotation of linguistic and morphological elements in sentences.
- **Named Entity Recognition (NER).** Identify strings that represent real-world entities (people, organisations, places, dates, etc.)
- **Co-referencing and relation extraction.** Identify named entities not explicitly mentioned, and their relation.

We see this in the next class!

Natural Language Processing

Topical analysis

More sophisticated tasks

- **Text classification.** Annotate documents with categories. Methods include topic modelling.
- **Sentiment analysis.** Categorize texts based on the valence of sentiment (positive, negative, neutral)
- **Text summarisation.** Reduces the length of one or more texts to smaller, meaningful, summaries.
- **Question answering.** Retrieve information based on a natural language question.
- **Text generation.** Generate texts from image/text inputs.
- And more…
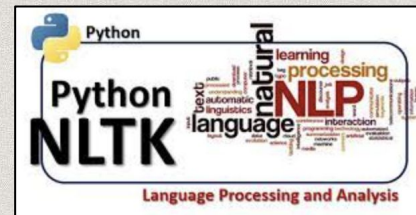
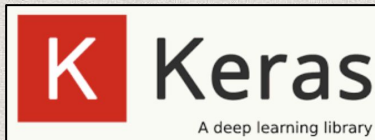# Computer vision

General tasks

- **Object detection.** Identify the presence and the contour of objects in an image.
- **Image classification.** Associate categories to images, based on their subjects.
- **Object recognition.** Identify a specific instance of a class (e.g. a person, a building)
- **Image restoration.** Reconstruction of original color palette.

Machine Learning

General tasks

- **Classification**. Associate a class or category to texts, images, data.
- **Recommendation**. Based on classification, provides a ranking of data based on categories.
- **Prediction**. Estimate next values in a series of continuous values (e.g. prices of artworks next month)
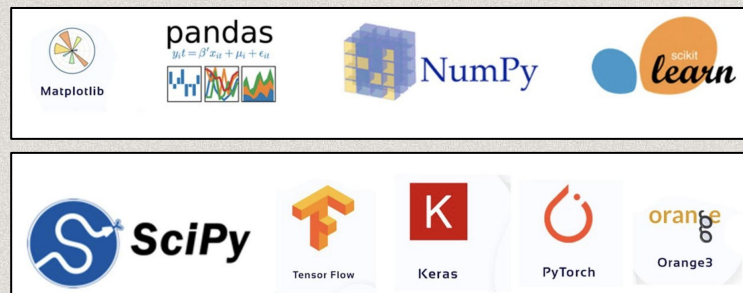
Machine Learning

Some tasks
_____

- **Clustering.** Highlight groupings, e.g. identify groups of people w/ a common behaviour.
- **Outlier detection.** Extract anomalies, e.g. online frauds. Looks for differences rather than similarities.
- **Association rule mining.** Extract patterns that identify groups, e.g. products frequently bought together.

We see this today!

If you are interested, have a look at this tutorial

Apriori algorithm

What

Look for groups of entities that are **likely** to occur together

What input

Groups of co-occurring items in the form of lists (of different lengths)

What is not

It does not look for similarities or differences, but for **correlation**

What output

Groups of co-occurring items in the form of lists, called **rules**, ranked by **likelihood**

Apriori algorithm

An algorithm to find association rules.

It tries all the combinations of items (w/ a minimum frequency).

It generates a rule representing the **probability** of co-occurrence.

Apriori algorithm

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Apriori algorithm

We have 5 customers in a supermarket, and we know what is in their basket.

What are the items that are mostly bought together?

Apriori algorithm

Rules
_____

Rules are in the form

IF {Antecedent} THEN {Consequent}

    A -> C

Caveat
_____

Even a small dataset creates a huge number of association rules, which have to be interpreted and pruned.

Pruned rules include trivial rules or/and inexplicable rules.

You must see <u>this</u>!

Measures
_____

SUPPORT. Measures how *frequently* items of the rule occur together in the dataset.

CONFIDENCE. Measures the conditional **probability** that the consequent will occur when there is the antecedent.

Rules with high support and confidence are interesting.

02

Hands-on

Maps and tables

What you'll do

Hands-on

Geospatial

Pin points on a
map

Topical

Association rule
mining

Exercise

Assignment

Review

Review the tutorial

Exercise

Solve the problems
(time to code!)

Fill in the <u>form</u>
with your answers

TODO

Nothing, next time we
will work on web
technologies only

Thanks!

Do you have any questions?

marilena.daquino2@unibo.it

https://github.com/marilenadaquino/information_visualization