Describing and visualising data

Silvio Peroni
silvio.peroni@unibo.it – https://orcid.org/0000-0003-0530-4305 – @essepuntato

Data Science (A.Y. 2021/2022)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA

Summary of the previous lectures (1/2)

A **datum** is a declarative statement subject-predicate-object that, through the predicate, either **attributes** a **literal** (i.e. a value such as a string, a number, etc.) to a subject **entity** *or* it **relates** such a subject **entity** with another **entity**

Each entity, being used either as subject or object of a statement, is characterised by **a unique identifier**

The **same entity** can be used as subject or object in one or more data, while a literal **cannot be used** as subject in any datum

An attribute is intrinsically **part of** the entity to which it is associated – modifying the value of an attribute affect **only** the entity to which it refers to

A **data model** is an abstract, simplified and formal representation of some data related to a system or a real domain, and enables us to describe what a data collection is about and to check data correctness

A data model permit one to specify **classes** of entities, their **attributes** and **relations**

Summary of the previous lectures (2/2)

Depending on the structure in which data are stored (or exposed), you need to approach the queries to datasets from a different angle

- With **tabular data**, often you have to combine tables between them to obtain bigger tables which contain the query requirements and the related answer
- With **graph data**, you explore the graph starting from fixed points (i.e. known entities, values, predicates) to find a pattern that is compliant with the query

A **database** as a **collection of data** which organised, stored and accessed electronically, which can be created through a database management system (DBMS)

A **transaction** is a unit of work performed (compliant with **ACID properties**) within a DBMS against a database and usually represents any change in a database

SQL and SPARQL are a **query languages** used and designed for managing data in **relational** and **graph-based** database management systems respectively, and allows one to **create** data and to **query** them

Any question about the previous lecture?

Descriptive statistics

Descriptive statistics are a series of statistics which aim at describing quantitatively a collection of data

Such statistics do not infer new information from a given population, since it does not use probability at all, but it provides measure to summarise data as they are

Often, such statistics are accompanied by visual graphs that enable a reader to understand simply some of the aspects of a collection of data

Different kinds of measures:

- measures of central tendency: mean, median, and mode
- measures of variability: minimum, maximum, and standard deviation

Mean

In mathematics and statistics, the arithmetic mean or, simply, the mean is the sum of a collection of numbers divided by the count of numbers in the collection

For instance, consider the following years of publication of 10 articles

1962, 2005, 2007, 2011, 2011, 2013, 2014, 2016, 2019, 2022

the mean is

(1962 + 2005 + 2007 + 2011 + 2011 + 2013 + 2014 + 2016 + 2019 + 2022) / 10 =

2008

Median

The median is the value separating the higher half from the lower half of a data sample – it may be thought of as "the middle" value

Basic feature: it is not skewed by a small proportion of extremely large or small values, and therefore provides a better representation of a typical value

How to calculate it:

- If the count of numbers *n* in a collection is odd, the median value is at index *(n-1)/2* (starting indexing items from 0, as in Python list)
- If the count of numbers *n* in a collection is even, the median value is the mean of the value at index *(n/2)-1* and *n/2* (starting indexing items from 0, as in Python list)

0 1 2 3 4 5 6 7 8 9
1962, 2005, 2007, 2011, 2011, 2013, 2014, 2016, 2019, 2022
(n/2)-1 n/2 → (2011 + 2013) / 2 = 2012

Mode

The mode is the value that appears most often in a set of data values

For instance, consider the following years of publication of 10 articles

1962, 2005, 2007, 2011, 2011, 2013, 2014, 2016, 2019, 2022

the mode is 2011

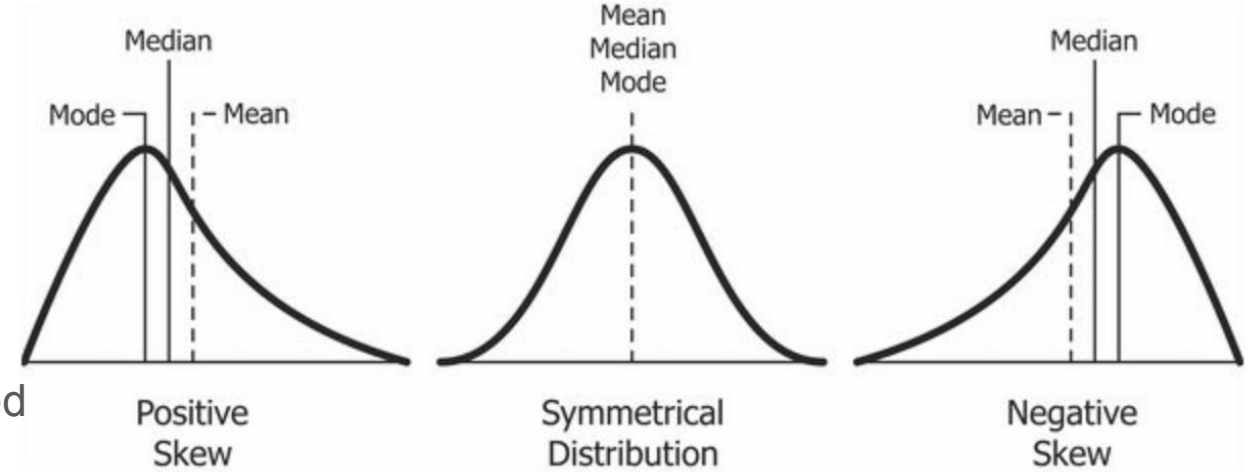
For a sample where each value occur precisely once, the usual practice is to discretize the data by assigning frequency values to intervals of equal distance, as for making a histogram, effectively replacing the values by the midpoints of the intervals they are assigned to

Why mean, median and mode can be different

The mean is largely affected by outliers, i.e. either small or large values that differ significantly from other observations

The median can be used as a measure of location when one thinks extreme values are of minimal or no importance, e.g. because a distribution is skewed

The mode is the same as that of the mean and median in a normal distribution, but it may be very different in highly skewed



Skewness. Wikipedia EN. https://en.wikipedia.org/wiki/Skewness

Minimum and maximum

The maximum and minimum are the values of the greatest and least elements of a collection

For instance, consider the following years of publication of 10 articles

1962, 2005, 2007, 2011, 2011, 2013, 2014, 2016, 2019, 2022

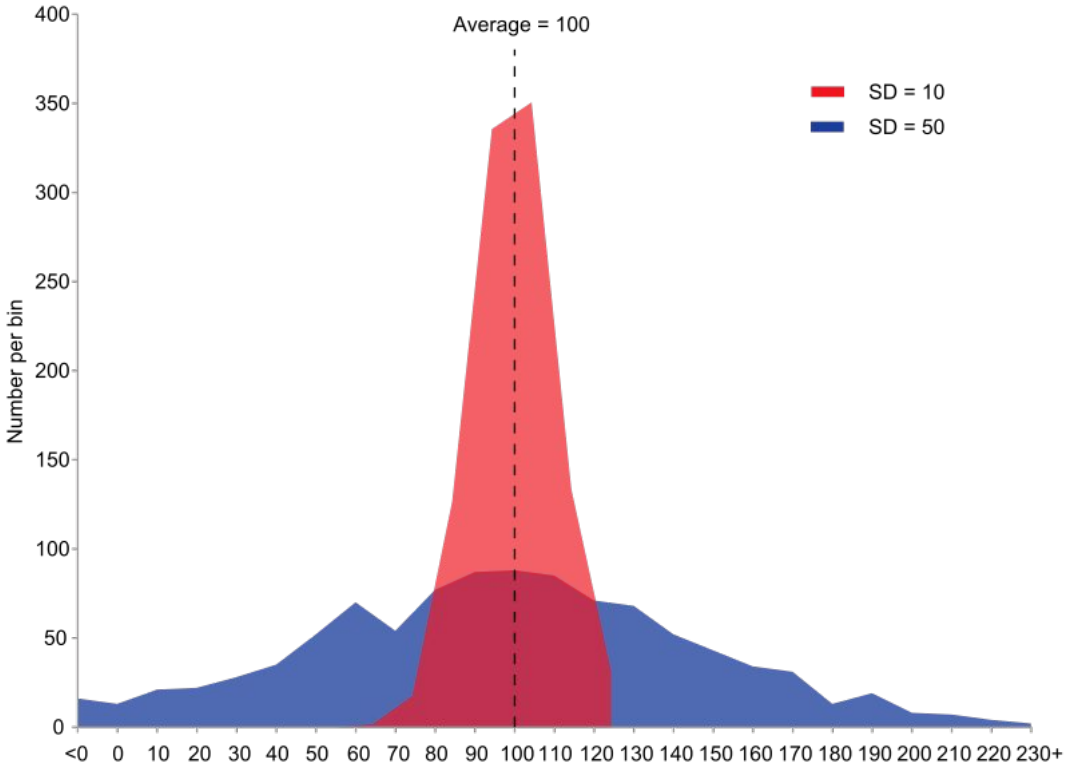
the maximum is 2022 and the minimum is 1962

If the sample has outliers, they necessarily include the sample maximum or sample minimum, or both

Standard deviation

The standard deviation measures the amount of dispersion of a set of values

- Low standard deviation: the values tend to be close to the mean
- High standard deviation: the values are spread out over a wider range



Standard deviation. Wikipedia EN. https://en.wikipedia.org/wiki/Standard_deviation

Visualisation

Visualisation techniques are of crucial important to effectively communicate a message to humans

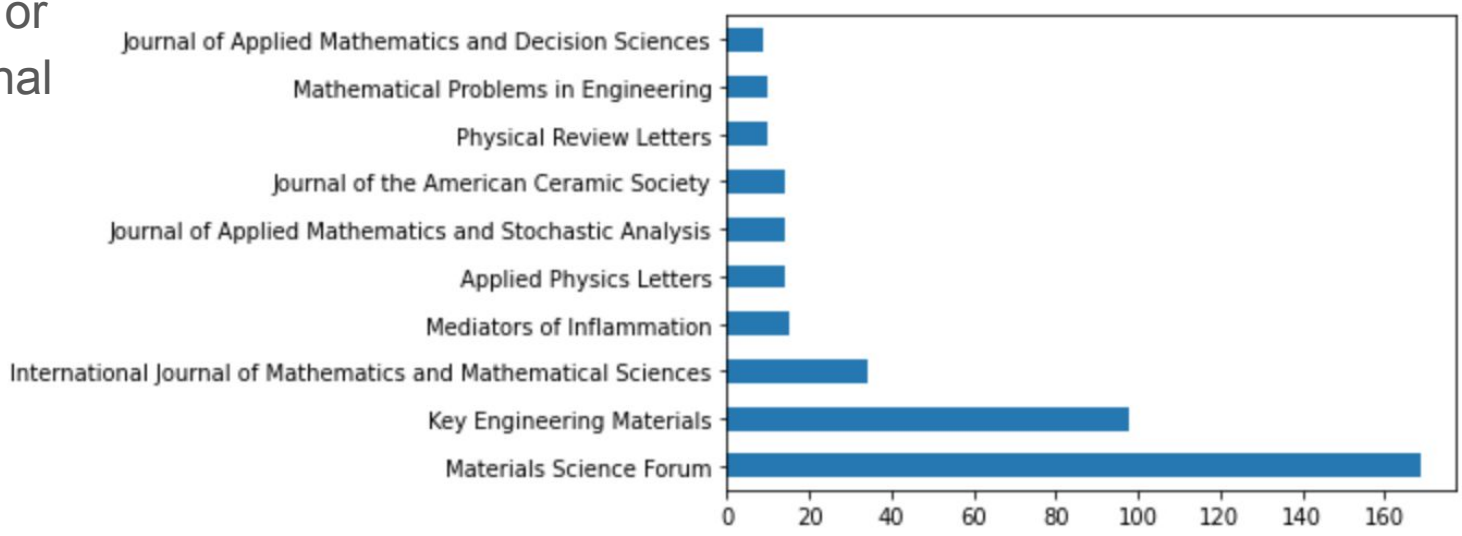
Data visualisation concerns the techniques used to communicate (often statistical) information about data, that can be categorised according to specific labels or shown as an time-oriented evolution of observations

Data visualisation techniques may be combined, when needed, with information visualisation techniques, that are used to represent visually numerical and non-numerical data (e.g. text or geolocated information) to support human comprehension of a phenomenon

Bar charts

Bar charts are used to present categorical data – i.e. a variable that can take on one of a limited, and usually fixed, number of possible values – with rectangular bars with heights or lengths proportional to the values that they represent

The bars can be plotted vertically or horizontally

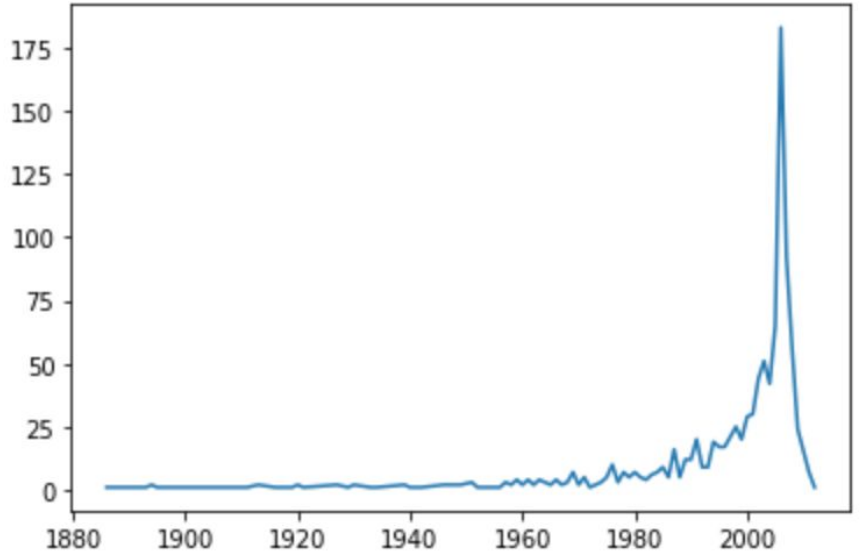


Time series

A time series is a series of data points indexed and ordered according to time, usually depicted as x-axis of a two dimensional graph

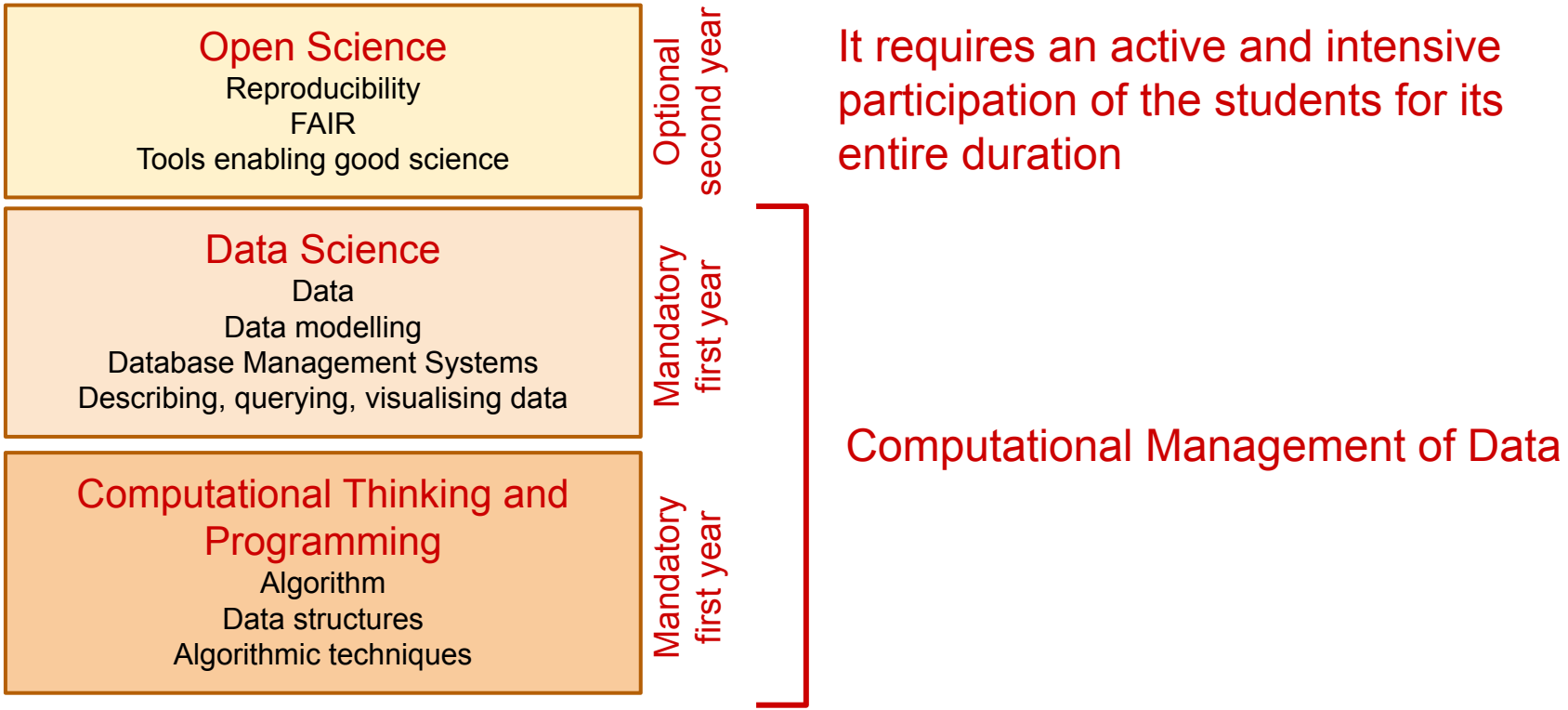
Commonly, a time series is a sequence taken at successive equally spaced points in time (e.g. years of publication)

A time series is very frequently plotted via a run chart (which is a temporal line chart).



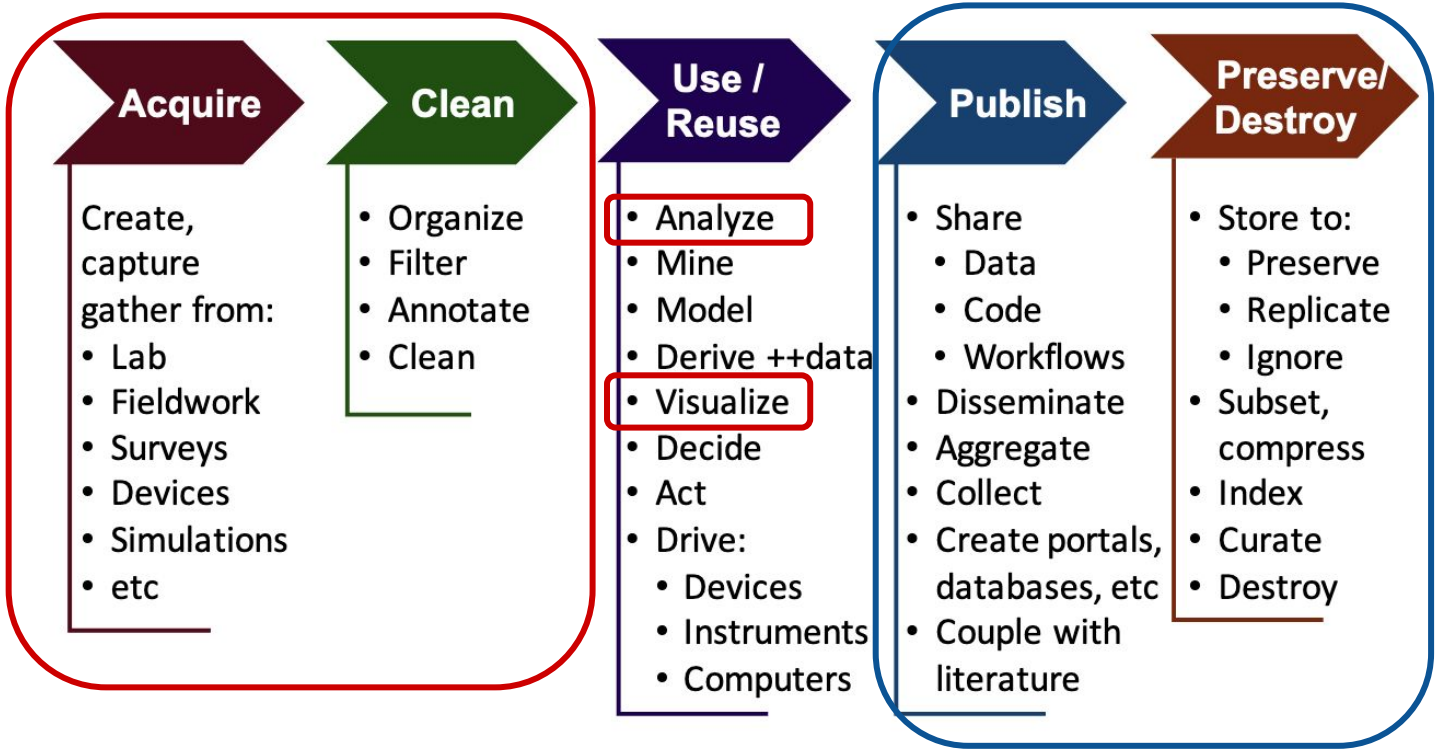
Continuing working on data

Possible paths after Data Science



Data Science life cycle (a reprise)

DHDK Data Science course (this course)



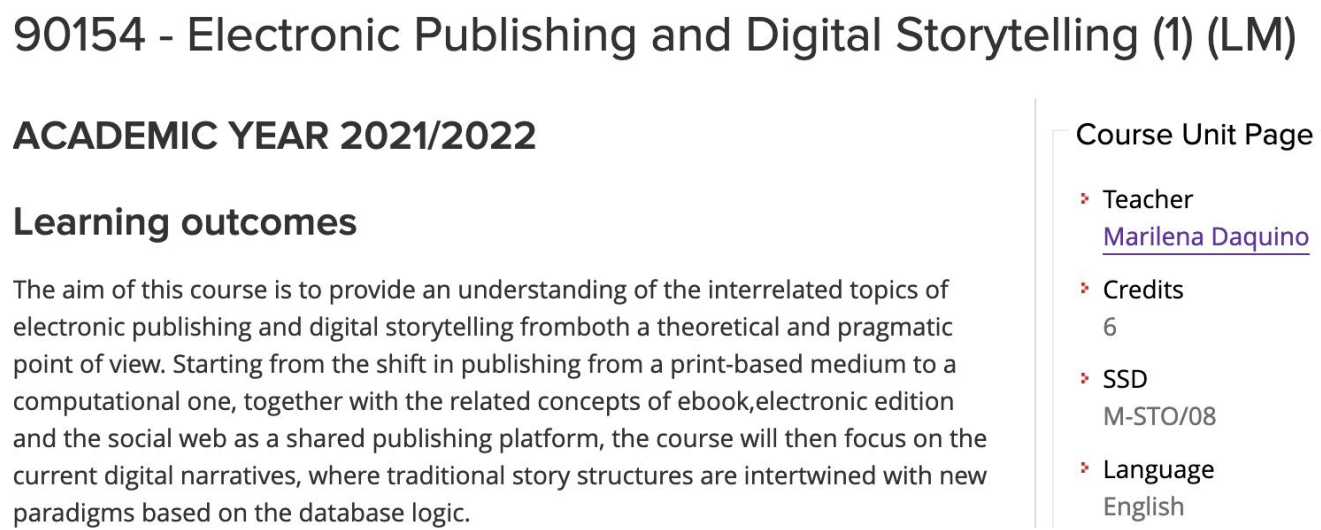
DHDK Open Science course (optional 2nd year course)

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., Stodden, V., & Szalay, A. S. (2018). Realizing the potential of data science. Communications of the ACM, 61(4), 67–72. https://doi.org/10.1145/3188721

About narratives and visualisations

DHDK course of the second year

It will change name in Information Visualisation



90154 - Electronic Publishing and Digital Storytelling (1) (LM)

ACADEMIC YEAR 2021/2022

Learning outcomes

The aim of this course is to provide an understanding of the interrelated topics of electronic publishing and digital storytelling fromboth a theoretical and pragmatic point of view. Starting from the shift in publishing from a print-based medium to a computational one, together with the related concepts of ebook,electronic edition and the social web as a shared publishing platform, the course will then focus on the current digital narratives, where traditional story structures are intertwined with new paradigms based on the database logic.

Course Unit Page

- Teacher
  Marilena Daquino
- Credits
  6
- SSD
  M-STO/08
- Language
  English

Descriptive and inference statistics

DHDK course of the second year

Crucial also for having an appropriate introduction to descriptive statistics…

… but, in particular, for having tools about probability and inference from an existing sample

92987 - Basic Analytics (1) (Lm)

ACADEMIC YEAR 2021/2022

Learning outcomes

techniques concerning the analysis of data bases. In particular the student is expected to learn: - probability s tools - measures of variance - index numbers

Course contents

1. Data collection, management and visualization
2. Descriptive statistics
3. Foundations of probability
4. Statistical inference
    - 4.1 Sampling
    - 4.2 Estimation
    - 4.3 Hypothesis testing
5. Simple regression

Course Unit Page

- Teacher
  Luca Trapin
- Credits
  6
- SSD
  SECS-S/02
- Language
  English
- Campus of Bologna
- Degree Programme
  Second cycle degree programme (LM) in Digital Humanities and Digital Knowledge (cod. 9224)

Studying networks (i.e. data)

DHDK course of the second year

It will change name in Network Analysis

93288 - Web Science (1) (Lm)

ACADEMIC YEAR 2021/2022

Learning outcomes

At the end of the course, students gain knowledge on the Web as a socio-technical system involving specific processes, entities, and behaviours, using interdisciplinary methods that blend computer science, sociology, ethnography, economics, linguistics, etc. The students are able to analyse the Web phenomena similarly to typical objects from natural sciences, distinguishing between data and applications, agents from computationally generated profiles, and addressing the characteristics of networks of entities emerging from the informationl, physical, social, and conceptual spaces constituting the Web.

Course Unit Page

- Teacher
  Saverio Giallorenzo
- Credits
  6
- SSD
  INF/01
- Language
  English
- Campus of Bologna

Machine learning

There are several course in UNIBO (external to DHDK) dedicated to machine learning, that you may include in your CV

93279 - APPLIED MACHINE LEARNING
Daniele Bonacorsi
Credits: 10
Area: Pharmacy and Biotechnology
Campus of Bologna
SSD: FIS/01
Second cycle degree programme (LM) in Bioinformatics (cod. 8020)

93279 - APPLIED MACHINE LEARNING (Modulo 1)
Daniele Bonacorsi
Course Timetable from **Mar 10, 2022** to **Apr 08, 2022**

93279 - APPLIED MACHINE LEARNING (Modulo 2)
Daniele Bonacorsi
Course Timetable from **Apr 28, 2022** to **Jun 13, 2022**

93280 - APPLIED MACHINE LEARNING - BASIC
Daniele Bonacorsi
Credits: 4
Area: Pharmacy and Biotechnology
Campus of Bologna
SSD: FIS/01
Second cycle degree programme (LM) in Bioinformatics (cod. 8020)
Course Timetable from **Mar 10, 2022** to **Apr 08, 2022**

95662 - INTRODUCTION TO MACHINE LEARNING
Giovanni Della Lunga
Credits: 3
Area: Statistics
Campus of Bologna
SSD: SECS-S/06
Second cycle degree programme (LM) in Quantitative Finance (cod. 8854)
Course Timetable from **Feb 26, 2022** to **Mar 12, 2022**

Ask a thesis

Main topic

Semantic publishing: the latest revolution in scholarly publishing

The term has been introduced for the very first time by David Shotton, and concerns the enhancement of scholarly publications by the use of modern web standards to improve interactivity, openness and usability, including the use of ontologies to encode rich semantics in the form of machine-readable RDF metadata

What is an open citation

Citation: conceptual directional link from a citing entity to a cited entity

The citation data related to a particular citation must include:

- the *representation* of such a conceptual directional link
- the *basic metadata* of the citing entity and the cited entity, i.e. sufficient information to create or retrieve textual bibliographic references

A bibliographic citation is an open citation when the data needed to define the citation are: structured, separate, open, identifiable, available

Peroni, S., & Shotton, D. (2018). Open Citation: Definition. Figshare. https://doi.org/10.6084/m9.figshare.6683855

OpenCitations

OpenCitations is an infrastructure organization for open scholarship dedicated to the publication of open citation data as Linked Open Data using Semantic Web technologies, thereby providing a disruptive alternative to traditional proprietary citation indexes

Currently, OpenCitations' Index is COCI, that contains

- 1,271,360,867 citations
- 71,337,645 bibliographic resources

Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. Quantitative Science Studies, 1(1), 428–444. https://doi.org/10.1162/qss_a_00023

Opportunities

Plenty of – and open to new – ideas and new developments in terms of infrastructure (i.e. programming), and also in terms of analytical studies (i.e. bibliometrics and scientometrics), theoretical analysis (i.e. understanding the functions of citations), and modelling (i.e. SPAR Ontologies)

The Research Centre for Open Scholarly Metadata and the Digital Humanities Advanced Research Centre allow students to do internships in this fascinating and challenging domain, working on one of the most cited and used Open Science projects of the past two years

End

Describing and visualising data

Silvio Peroni
silvio.peroni@unibo.it – https://orcid.org/0000-0003-0530-4305 – @essepuntato

Data Science (A.Y. 2021/2022)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA