Introduction

Paolo Boldi
DSI
LAW (Laboratory for Web Algorithmics)
Università degli Studi di Milan

What is this course about

- A miscellanea of topics
- Most of them still at an embrionic stage of development.
- In some cases, on the border between different research areas.
- On top of this, it is a *highly personal* anthology: no pretense to make a complete overview.
- It is best characterized by some *common* features.

Size matters

- Data size (large, huge) is one of these features
- Web data is the playground of asymptotics. . .
- . . . and often also constants matter!
- Most problems would be easy (or trivial) if size was *not* as large
- Keyword: ALGORITHMS

Digital wilderness

- Data are generated in an uncontrolled, totally decentralized, heterogeneous way
- This is rather *new* to computer science
- Presence of noise, adversarial behaviors, different levels of quality, highly unstructured data...
- ...prompt for techniques and methods more similar to those adopted by physicists
- Keyword: EXPERIMENTAL

User-generated content

- Data are generated by *users*
- Typically: thousands to millions of users
- Their collective behavior (*wisdom of the crowds*) is a treasure trove of knowledge
- This raises privacy / anonymization issues
- Moreover: user-generated data are very precious and rarely made available to public scrutiny
- Keyword: DATA (UN)AVAILABILITY