# Lecture 1

## Implications of Big Data for ML

# Lecture I – Big Data in Machine Learning

Marina Meilă

Department of Statistics
University of Washington

STAT 548/CSE 547
Spring, 2025

# Big Data

How Big?

- Lobster 100's TB 7000$ ~10 GB (free)
- cell painting ~ 100's TB
- real time social media 50 TB
- brain plugs data + connectome 14 TB
- 10000 in financial
- youtube ~ 100's hours
- health care claims $\sim 10^9$
- $10^5$ particle trajectories - in britain
- $10^5$ London marathon data

$n \Longrightarrow$

$n \times D$

High dimensional

- images 4K x 4K ~ D
  - 1 pixel = 3 channels x 256 levels
  - medical, bio, astro
- econ
- financial
- biological sequences $10 - 10^6$
- music, audio
- video
- atmospheric
- sports data

How used?

- pattern recognition → find pattern
  - classification
  - detection
  - ~~clustering~~ completion
  - testing
  - regression
- discover dependencies - relationship
- find outliers
- find neighbors similar points to some X (retrieval)

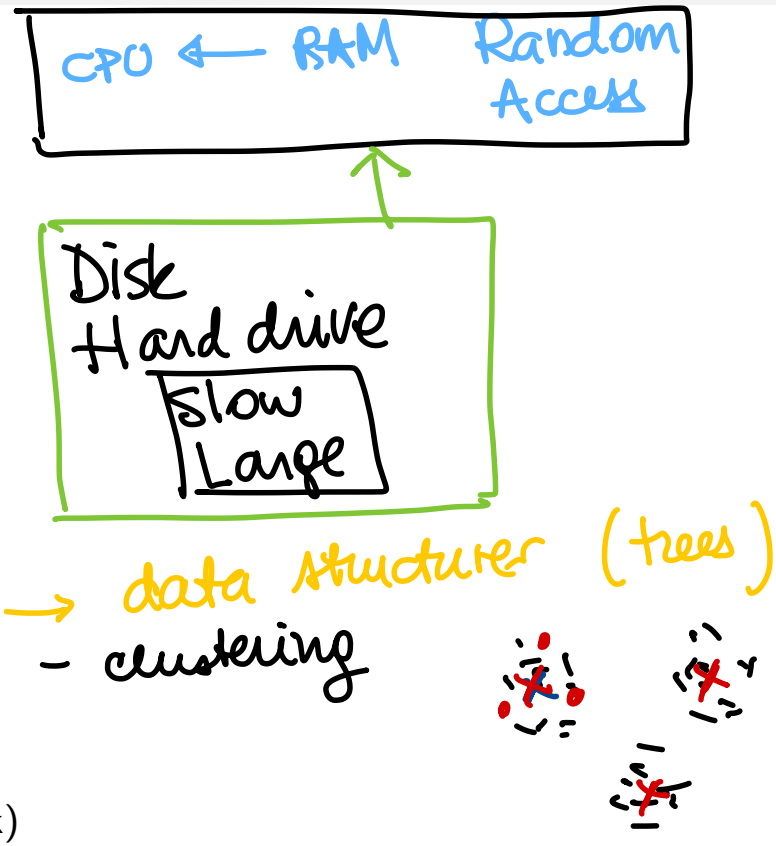
Time scales?

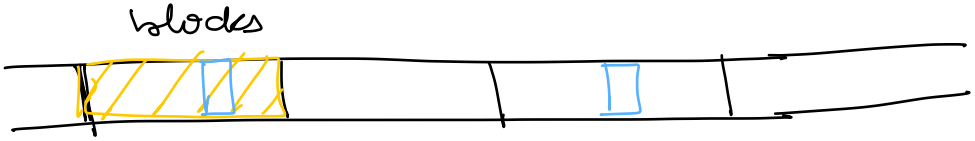
Repeated?

# Big data and Machine Learning I

Big Data has implications for ML at many levels

- Storage
  - may not fit in local memory
  - expensive/slow to move around
  - I/O expensive/slow
- Access
  - serial/by block, not random
- Indexing
  - Preprocessing steps that allow faster access during
- Computing
  - Parallelization when possible
  - Automation of resource management (Hadoop, Spark)
- Algorithms
  - predominantly sub-quadratic, i.e. $\mathcal{O}(n), \tilde{\mathcal{O}}(n)$
  - sub-linear, i.e. $o(n)$ when possible – sampling, Stochastic Gradient Descent (SGD)






slow → read/write at many random

- sorting
- random permutation
- Stoch Grad Descent