CS 591, Lecture 10
Data Analytics: Theory and Applications
Boston University

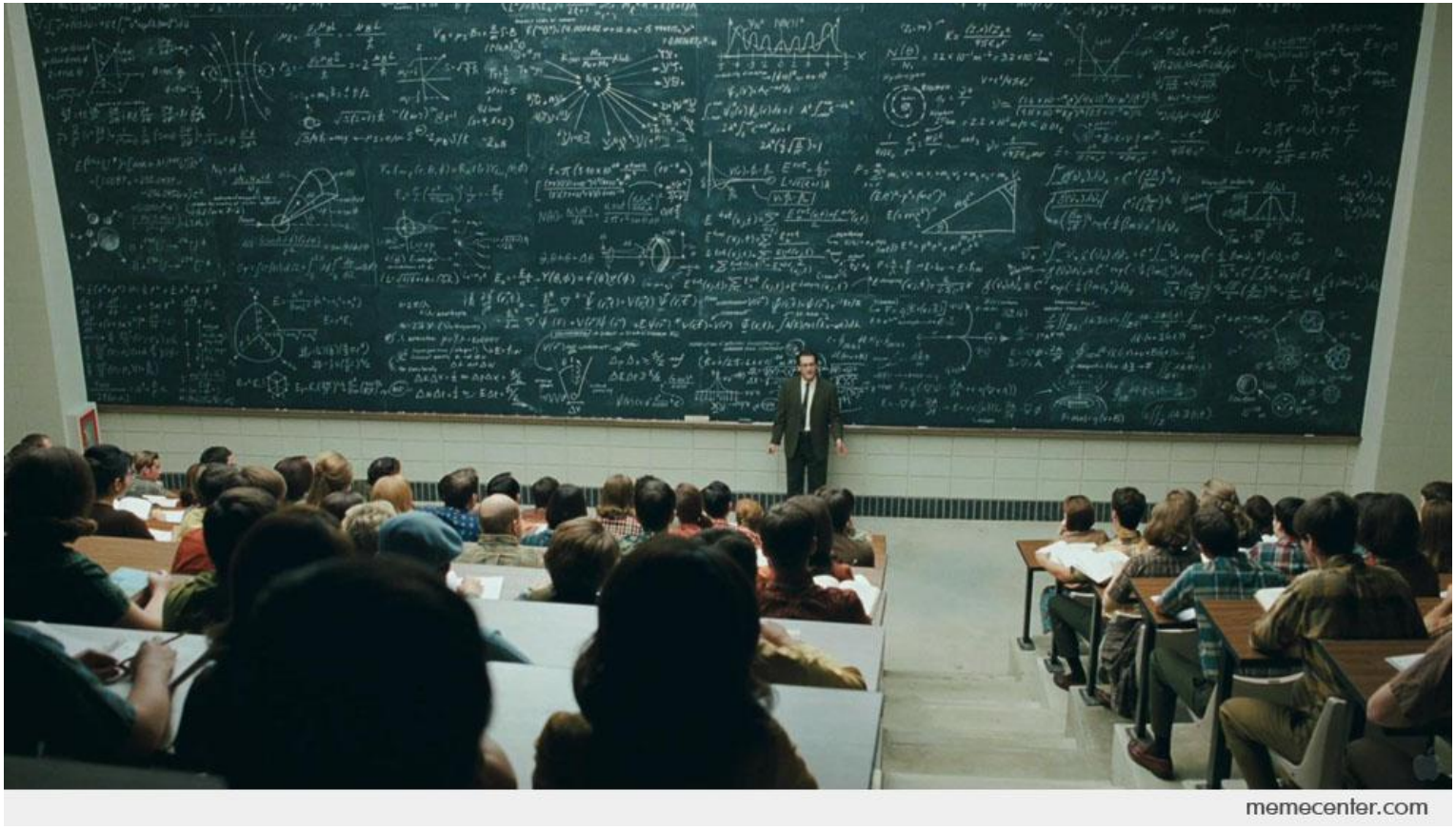
Babis Tsourakakis

February 27th, 2017

Today's lecture

- F_2 Moment Estimation (AMS sketch) (PART I)
  - Dimensionality reduction
- Some of you asked for project ideas (PART II)

AMS Sketch, Blackboard



Project ideas

Class Topics

1. **Search** (Hashing, kd-trees, LSH)
2. **Data streams**
3. **Graph Mining**
4. **Machine Learning**

Related Top-Tier Conferences

1. **NIPS, ICML, KDD**
2. **SIGIR, WWW**
3. **VLDB, SIGMOD**
4. **CVPR, ECCV, ICCV**
5. **AAAI, IJCAI**
6. **STOC, FOCS, SODA, ICALP, ESA**

Where to find datasets?

Tons of interesting datasets are available.

- Kaggle
- ICON
- Zurich Building Image Database
- Yale face database
- UIUC Image Database for Car Detection

Where to find datasets?

- Flower Datasets
- Tiny Images Dataset
- Synthia
- Action DB
- Sensor data
- Facebook graphs

Where to find datasets?

- Datahub
- UCI ML datasets
- Im2Latex
- NYC police data
- Konect networks
- Web Graphs

Where to find datasets?

- FB wall posts
- Human centric sensing
- Stackexchange dump
- MovieLens dataset
- LastFM
- **Many more out there...**

Project Ideas: Search

- Explore Cuckoo, and Tabulation hashing
  - Project Example
  - Some slides from Stanford
- SHA1 broken announcement, SHA1 attack Web site
- Hashing for Machine Learning
  - Feature Hashing for Large Scale Multitask Learning
  - Accelerated Large Scale Optimization by Concomitant Hashing

Project Ideas: Search

method	input sim.	hash function	dist. measure	optimization criteria
spectral hashing [135]	E	LE	HD	CC + BB + BU
kernelized spectral hashing [37]	S, E	KE	HD	CC + BB + BU
Hypergraph spectral hashing [153], [89]	S	CL	HD	CC + BB + BU
Topology preserving hashing [145]	E	LI	HD	CC + CCD + BB + BU
hashing with graphs [83]	S	KE	HD	CC + BB
ICA Hashing [35]	E	LI, KE	HD	CC + BB + BU + MIM
Semi-supervised hashing [125], [126], [127]	S, E	LI	HD	CC + BB + PU
LDA hash [122]	S	LI	HD	CC + PU
binary reconstructive embedding [63]	E	LI, KE	HD	MDD
supervised hashing with kernels [82]	E, S	LI, KE	HD	MDS
spec hashing [78]	S	CL	HD	MDSD
bilinear hyperplane hashing [84]	ACS	BILI	HD	MDS
minimal loss hashing [101]	E, S	LI	HD	HL
order preserving hashing [130]	E	LI	HD	ROL
Triplet loss hashing [103]	E, S	Any	HD, AHD	TL
listwise supervision hashing [128]	E, S	LI	HD	TL
Similarity sensitive coding (SSC) [114]	S	CL	WHD	CE
parameter sensitive hashing [115]	S	CL	WHD	CE
column generation hashing [75]	S	CL	WHD	CE
complementary projection hashing [55]*	E	LI, KE	HD	SP + CP + PBB
label-regularized maximum margin hashing [96]*	E, S	KE	HD	SP + MM + BB
Random maximum margin hashing [57]*	E	LI, KE	HD	SP + MM + BB
spherical hashing [38]*	E	SF	NHD	SP + PBB
density sensitive hashing [79]*	E	LI	HD	SP + BB
multi-dimensional spectral hashing [134]	E	LE	WHD	CC + BB + BU
Weighted hashing [131]	E	LI	WHD	CC + BB + BU
Query-adaptive bit weights [53], [54]	S	LI (all)	QWHD	CE
Query adaptive hashing [81]	S	LI	QWHD	CE

Source: Hashing for Similarity Search: A Survey

Project Ideas: Search

- Hashing for Machine Learning (cont.)
  - Hash-SVM: Scalable Kernel Machines for Large-Scale Visual Classification
  - Learning to hash with binary reconstructive embeddings
- Fast k-NN graph construction using LSH, benchmarks on SSL
- Detecting approximate duplicate documents using MinHash
- Predictive Indexing

Project Ideas: Data Streams

- Implement streaming graph algorithms
- Summarizing Data using Bottom k-Sketches
- Implement a spell-checker using space-efficient data structures
- Implement an autocomplete algorithm
- Implement a space-efficient hangman game

Project Ideas: Graph Mining

- Predicting signed edges
- Semi-supervised learning with graph embeddings
- Influence Maximization
- Co-evolutionary dynamics in social networks: a case study of Twitter
- Sketch-based Influence Maximization and Computation: Scaling up with Guarantees
- Hypergraph/tensor mining on fMRI data

Project Ideas: Machine Learning

- Matrix factorization and collaborative filtering
- Robust PCA
- Transfer learning (related to self driving cars)
- GANs
- Structure from Motion
- Location recovery

Project Ideas: Machine Learning

- Sketching and Neural Networks
- RL for combinatorial games
- description2code
- Handwritting recognition
- Tensor Kart
- Generate images
- Awesome deep learning papers