ORIE 4741: Learning with Big Messy Data

Introduction

Professor Udell

Operations Research and Information Engineering
Cornell

October 16, 2021

Outline

ORIE 4741/5741: Learning with Big Messy Data

want to take this class?

- **ASAP:**
  - enroll (or drop) (or get on wait list)
  - fill out course survey
  - sign up for discussion forum
  - sign up for iClicker
- **Thursday 9/2/2021:** homework 0

links on course website:
`https://people.orie.cornell.edu/mru8/orie4741/`

Course staff

- Prof. Madeleine Udell
- TA: Richard Phillips (CS PhD)
- TA: Tao Jiang (ORIE PhD)
- TA: Connor Lawless (ORIE PhD)
- TA: Yuanping Du (ORIE MEng)
- TA: Max de Ledebur (ORIE MEng)
- TA: Jody Zhu (ORIE+CS Undergraduate)
- TA: Tara Khanna (ORIE Undergraduate)
- TA: Kevin Jiang (CS Undergraduate)

Tech stack

- In person or Zoom for lectures, section, and office hours
- Course website for course materials
  (syllabus, schedule, homework, project, etc)
- iClicker for polls
- Zulip for Q&A
- Gradescope for quizzes, submitting homework, grades, solutions
- Github for code (demos, projects, and hw starter code)

Course requirements and grading

course website:

(grading, course requirements, lectures, homework, etc)

`https://people.orie.cornell.edu/mru8/orie4741/`

- (15%) Participation: for every lecture (after this one), use
  - iClicker for sync lectures
  - participation form for async lectures
- (30%) Homework
  - due every two weeks or so
  - first one due next Thursday
- (15%) Quizzes
  - 30 min quiz every week or so
- (40%) Project

Course requirements and grading

course website:
(grading, course requirements, lectures, homework, etc)
`https://people.orie.cornell.edu/mru8/orie4741/`

- (15%) Participation: for every lecture (after this one), use
  - iClicker for sync lectures
  - participation form for async lectures
- (30%) Homework
  - due every two weeks or so
  - first one due next Thursday
- (15%) Quizzes
  - 30 min quiz every week or so
- (40%) Project

FAQ:

- yes, you can take the class online (even async)
- yes, you can take section online (even async), or not take the section

ORIE 5741 vs 4741

- 5741 has same course material as 4741

ORIE 5741 vs 4741

- 5741 has same course material as 4741
- expectations and rubric for course project differ
  - more business-oriented project
  - more detailed problem formulation
  - project presentation required (in addition to report)

ORIE 5741 vs 4741

- 5741 has same course material as 4741
- expectations and rubric for course project differ
  - more business-oriented project
  - more detailed problem formulation
  - project presentation required (in addition to report)
- rubric will reflect MEng learning outcomes:
  1. Mastery and Application of Core Disciplinary Knowledge
  2. Problem Formulation and Organization and Planning of the Solution Process
  3. Collaborative Problem Solving and Issue Resolution
  4. Communication of Knowledge, Ideas, and Decision Justification
  5. Self-Directed Learning and Professional Development

ORIE 5741 vs 4741

- 5741 has same course material as 4741
- expectations and rubric for course project differ
  - more business-oriented project
  - more detailed problem formulation
  - project presentation required (in addition to report)
- rubric will reflect MEng learning outcomes:
  1. Mastery and Application of Core Disciplinary Knowledge
  2. Problem Formulation and Organization and Planning of the Solution Process
  3. Collaborative Problem Solving and Issue Resolution
  4. Communication of Knowledge, Ideas, and Decision Justification
  5. Self-Directed Learning and Professional Development
- Everyone in a project group with 5741 students will be graded according to 5741 rubric.

Questions

during lecture:

- ask out loud
- zoom chat (to everyone, or to a TA)

outside of lecture:

- ask at office hours
- ask on discussion forum
- don't send email

Outline

Logistics

Stories

Definitions

Kinds of learning

Syllabus

Oh, you work with big messy data? Maybe you could help us out...?

My career in big data

academic

- B.S. in Mathematics and Physics at Yale
- Ph.D. in Computational and Mathematical Engineering at Stanford
- postdoctoral fellow at the Center for the Mathematics of Information at Caltech
- professor in ORIE at Cornell

My career in big data

academic

- B.S. in Mathematics and Physics at Yale
- Ph.D. in Computational and Mathematical Engineering at Stanford
- postdoctoral fellow at the Center for the Mathematics of Information at Caltech
- professor in ORIE at Cornell

applied work

- finance: Goldman Sachs, BlackRock, Capital One, Schonfeld, Two Sigma, ...
- tech: Google, Retina.ai, Marketing Attribution
- cybersecurity: DARPA, Expanse (formerly Qadium)
- healthcare: Apixio, Ontario
- clean energy: Aurora
- politics: Obama 2012

Data table: politics

age	gender	state	income	education	voted?	support	$\cdots$
29	F	CT	\$53,000	college	yes	Biden	$\cdots$
57	?	NY	\$19,000	high school	yes	?	$\cdots$
?	M	CA	\$102,000	masters	no	Trump	$\cdots$
41	F	NV	\$23,000	?	yes	Trump	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Data table: politics

age	gender	state	income	education	voted?	support	⋯
29	F	CT	\$53,000	college	yes	Biden	⋯
57	?	NY	\$19,000	high school	yes	?	⋯
?	M	CA	\$102,000	masters	no	Trump	⋯
41	F	NV	\$23,000	?	yes	Trump	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

goals:

- detect demographic groups?
- find typical responses?
- identify related features?
- impute missing entries?

Medicine

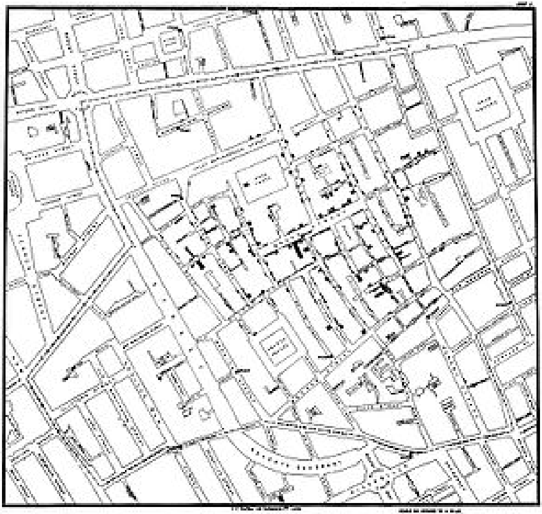
Noble,Jam - Armstrong - SRS EHR v8.3.313.0 - Workstation AMANDA
File Drawers View Mail Reports Tools Help
Appointments
6/20/2012
Armstrong, Max; Multiple Locatio
Time Name Appointment Type
09:00 AM Noble, James Consult
09:30 AM Casper, Carolyn Exam- New Patient
10:00 AM Harris, Richard Follow Up
10:30 AM Meril, Leslie Exam- Est. Patient
10:30 AM Moss, Pete Consult
10:45 AM Copeland, Elizabeth Consult
11:00 AM Derek, Martin Exam- New Patient
01:00 PM Paltrow, Mary Consult
01:15 PM Gold, Alan Exam - Est. Patient
01:30 PM Dipierro, Drew Consult
01:30 PM Glass, Tyler Exam - Est. Patient
02:15 PM Thompson, Brian Exam - Est. Patient
02:30 PM Carter, James Consult
03:30 PM Abner, Darlene Follow Up
04:00 PM Brown, Kevin Exam - Est. Patient
04:15 PM Newsome, Gina Exam - Est. Patient
04:30 PM Newsome, Jenna Exam - Est. Patient
Mail Status
Normal 11
Rx 2
Transcription 1
Clinical Summary
Noble, James
Demographics
Address: 4516 West Huron Street
Chicago, IL 60607
Email: jnoble@gmail.com
DOB: 04.05.1962
Age: 50y
Patient ID: 88501
Primary Ins: Aetna U.S. Healthcare - Master
Pharmacy Directions
CVS/pharmacy #8671
510 College Mall Rd
Bloomington, IN 47401 P (812) 336-7306
F (812) 335-9347
User Defined Fields
Name Value
Injury Shoulder pain
Employer Name WBI
Employer Fax 201-867-5309
Insurance Carrier GEICO
Adjuster Fax Num... 201-555-8735
Diagnosis 727.3
Position Truck Driver
CareTracker ID # 36817221
Dictations
Appt Time Completed
06.20.2012 09:00 AM
Alert
Tests
Description Date
MRI On Order
MRI 11.17.2011
Diagnosis
Diagnosis Status Date ICD-9 Notes
ACL Tear 06.20.2012 844.2
Lumbago 11.17.2011 724.2
Vitals
Smoking Status
Status ID Date
Never smoker 4 06.18.2012
Procedures
73000 73090 99211 99212 99213 99214 99215
Procedure Date CPT Type Status
Knee Arthroscopy/Surg 06.20.2012 29881 Internal Performed
Rx History Active Discontinued Intolerant Del/Cancel
Status Date Drug Strength Instructions
06.18.2012 Keflex 250 MG 1 tab po BID
11.04.2010 CeleBREX 100 MG 1 tablet by mouth in the AM
04.13.2006 OxyCODONE HCl 20 MG 1 tab po q4h prn pain
Non-Drug Allergies
Description Reaction Notes
Latex Severe rash
Shellfish Hives
Family History
Relationship Deceased Notes
Sister No JRA
Maternal Grandmother Yes Osteoarthritis
Surgeries
Description Date Surgeon Notes
Arthroscopy 12. 18. 2011 Armstrong N/A
Appointments
Date Time Doctor Reason Type Location Notes
Current encounter: 6/20/2012 9:00:00 AM
Transfer Encounter
Noble, James
All Messages Flow Sheet Rx
Desktop Message Ctr Scan Place Forms
Chart Notes
Radiology Reports
History
Digital X-Rays
Correspondence Received
Consult Letters Sent
Orders
Injections
Op-Reports
Discharge Summaries
PT Assessments
Charge Capture
SRS PACS
Surgery Documents
Surgery
DASH

Data table: medicine

age	gender	heart disease	statins?	$\cdots$
29	F	yes	no	$\cdots$
57	?	no	no	$\cdots$
?	M	no	no	$\cdots$
41	F	yes	yes	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

- find similar patients?
- understand systemic healthcare needs?
- use symptoms to detect which patients have COVID-19?
- detect patients who had series of mini-strokes?

Pollution

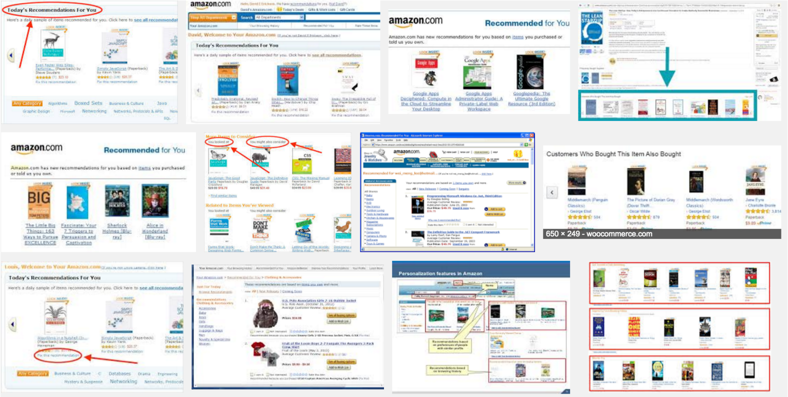


[Snow, 1854]

Pollution

location	time	CO2	O2	O3	$\cdots$
1	1	.7	.9	?	$\cdots$
1	2	.5	.7	?	$\cdots$
1	3	.4	.5	1.4	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$		

Marketing


Today's Recommendations For You
amazon.com
Recommended for You
Amazon.com has new recommendations for you based on items you purchased or told us you own.
Customers Who Bought This Item Also Bought
650 × 249 - woocommerce.com
Personalization features in Amazon

Marketing

customer	product 1	product 2	product 3	$\cdots$
1	yes	?	yes	$\cdots$
2	yes	yes	?	$\cdots$
3	?	?	yes	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Finance


1D
5D
1M
6M
YTD
1Y
5Y
MAX
8,000%
6,000%
4,000%
2,000%
0%
-2,000%
Sep 2020
Nov 2020
Jan 2021
Mar 2021
May 2021
Jul 2021
GameStop Corp.
$206.59
+$201.61
↑4,048.39%
AMC Entertainmen...
$45.71
+$40.17
↑725.09%

Finance

ticker	t_1	t_2	$\cdots$
AAPL	.05	-.21	$\cdots$
GOOG	-.11	.24	$\cdots$
FB	.07	-.18	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$

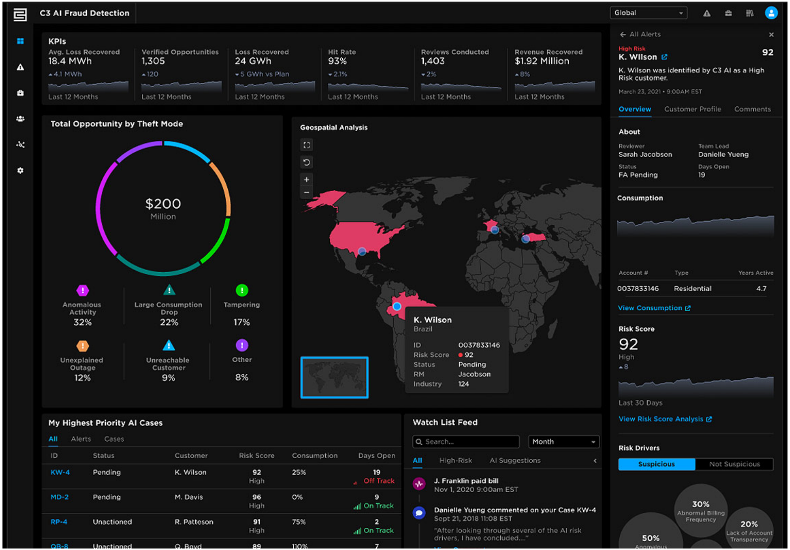
Environmental, social and governance (ESG) data

- one row for each asset at each time
- one column per key performance indicator (KPI)
  - carbon emissions
  - e-waste management
  - climate change risk
  - worker safety
  - ...
- values: numerical ratings $1, \ldots, 10$ or boolean $\{0, 1\}$
- triangular missing pattern: KPI/asset coverage increases with time

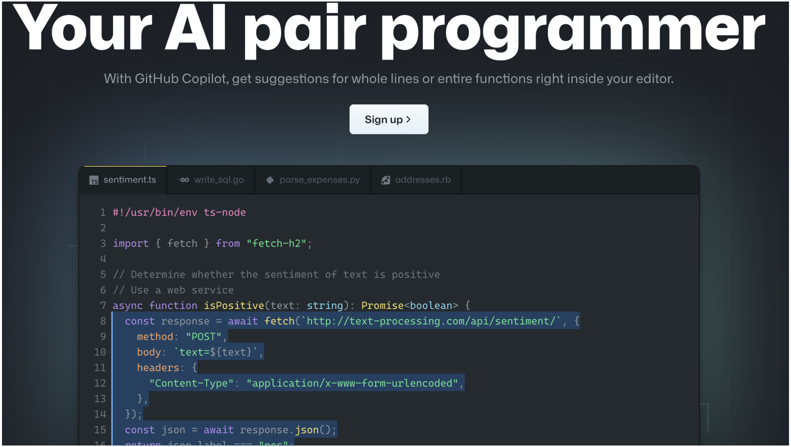
goals for ESG analysis:

- impute missing items?
- audit/improve on vendor data quality?
- predict long term returns?

Fraud detection



Autocomplete



Application areas

- health
- politics
- governance
- advertising
- retail
- ecommerce
- finance
- ...

Outline

Big

- NASA, 1997: "taxing the capacities of main memory, local disk, and even remote disk"

[1] image courtesy of Kim Minor @ IBM

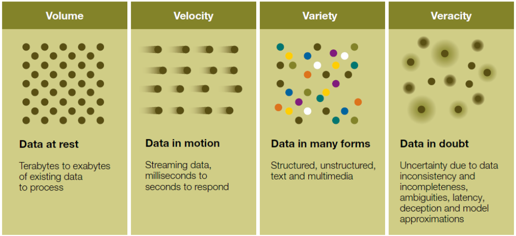
Big

- NASA, 1997: "taxing the capacities of main memory, local disk, and even remote disk"
- OED, 2015: "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges"

[1] image courtesy of Kim Minor @ IBM

Big

- NASA, 1997: "taxing the capacities of main memory, local disk, and even remote disk"
- OED, 2015: "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges"
- 4 Vs:

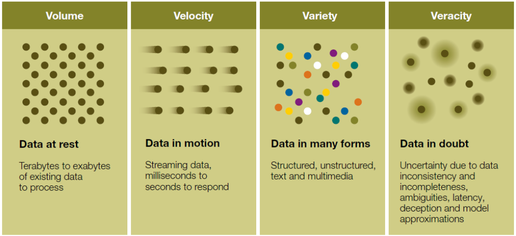


[1]

[1] image courtesy of Kim Minor @ IBM

Big

- NASA, 1997: "taxing the capacities of main memory, local disk, and even remote disk"
- OED, 2015: "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges"
- 4 Vs:



[1]

- 5th V: value

[1] image courtesy of Kim Minor @ IBM

Big: our definition

Definition

An algorithm for **big data** is one with computational and memory requirements that scale **linearly** (or nearly linearly) in the size of the data.

Big: our definition

Definition

An algorithm for **big data** is one with computational and memory requirements that scale **linearly** (or nearly linearly) in the size of the data.

why this definition? independent of

- hardware
- business

Big: our definition

Definition

An algorithm for **big data** is one with computational and memory requirements that scale **linearly** (or nearly linearly) in the size of the data.

why this definition? independent of

- hardware
- business

if you use only algorithms for **big data**, then you're working with **big data**

Messy

- noisy: some (or all) values suffer errors, inaccuracies, or malicious corruption

Messy

- noisy: some (or all) values suffer errors, inaccuracies, or malicious corruption
- missing: some values are missing, inconsistent, not recorded, or lost

Messy

- noisy: some (or all) values suffer errors, inaccuracies, or malicious corruption
- missing: some values are missing, inconsistent, not recorded, or lost
- heterogeneous: values of many different types
  - continuous values (*e.g.*, 4.2, π)
  - discrete values (*e.g.*, 0, 4, 994)
  - nominal values (*e.g.*, apple, banana, pear)
  - ordinal values (*e.g.*, rarely, sometimes, often)
  - graphs or networks (*e.g.*, person 1 is friends with person 2)
  - text (*e.g.*, doctor's note describing symptoms)
  - sets (*e.g.*, items purchased)

Learning

Learning

- machine learning?

Learning

- machine learning?
- human learning?

Learning

- machine learning?
- human learning?
- when data is **big** and **messy**,
  machine help is essential for human learning!

Outline

Data table

n examples (patients, respondents, households, assets)
d features (tests, questions, sensors, times)

$$\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nd} \end{bmatrix}$$

- a_i is ith row of A: feature vector for ith example
- $a_{:j}$ is jth column of A: values for jth feature across all examples
- a_{ij} is jth feature of ith example

Supervised learning

- identify one column of data that we want to predict

$$\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1\,d-1} & y_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \cdots & x_{n\,d-1} & y_n \end{bmatrix} = \begin{bmatrix} & & \\ X & & y \\ & & \end{bmatrix}$$

- $x_i \in \mathcal{X}$ for $i = 1, \ldots, n$ are rows of X
- $y_i \in \mathcal{Y}$ for $i = 1, \ldots, n$ are entries of y

Supervised learning

- identify one column of data that we want to predict

$$\begin{bmatrix} \quad & A & \quad \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1\,d-1} & y_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \cdots & x_{n\,d-1} & y_n \end{bmatrix} = \begin{bmatrix} X & & y \end{bmatrix}$$

- $x_i \in \mathcal{X}$ for $i = 1, \ldots, n$ are rows of X
- $y_i \in \mathcal{Y}$ for $i = 1, \ldots, n$ are entries of y
- we believe there is a mapping $f : \mathcal{X} \to \mathcal{Y}$

$$y_i \approx f(x_i)$$

- our goal is to learn f

Example: supervised learning for credit decisioning

- goal: decide which credit card applicants should be approved
- input space: entries of $\mathcal{X} \in \mathbf{R}^d$ correspond to fields in credit application
  - *e.g.*, salary, years in residence, outstanding debt, number of credit lines, ...
- output space: $\mathcal{Y} = \{+1, -1\}$
  - $+1$ means approve
  - -1 means reject
- data: $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$
  applications of previous customers, and credit approval decisions made by humans

Example: supervised learning for credit decisioning

- goal: decide which credit card applicants should be approved
- input space: entries of $\mathcal{X} \in \mathbf{R}^d$ correspond to fields in credit application
  - *e.g.*, salary, years in residence, outstanding debt, number of credit lines, ...
- output space: $\mathcal{Y} = \{+1, -1\}$
  - $+1$ means approve
  - -1 means reject
- data: $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$
  applications of previous customers, and credit approval decisions made by humans

Q: what are potential problems with using a model built with this data?

Example: supervised learning for credit decisioning

- goal: decide which credit card applicants should be approved
- input space: entries of $\mathcal{X} \in \mathbf{R}^d$ correspond to fields in credit application
  - *e.g.*, salary, years in residence, outstanding debt, number of credit lines, ...
- output space: $\mathcal{Y} = \{+1, -1\}$
  - $+1$ means approve
  - -1 means reject
- data: $\mathcal{D} = (x_1, y_1), \dots, (x_n, y_n)$
  applications of previous customers, and credit approval decisions made by humans

Q: what are potential problems with using a model built with this data?

A: wrong objective: human decision may not be correct decision; covariate shift: future data may look unlike past data; ...

Exercise: formalizing real problems

- identify a prediction goal
- identify the input space $\mathcal{X}$
- identify the output space $\mathcal{Y}$
- identify the data $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$ you'd like to use
- what kinds of noise do you expect in the data?

Outline

Course objectives (I)

- plot
- predict
- cluster
- impute
- denoise
- recommend
- understand

Course objectives (II)

this course is about

- algorithms for big messy data
- learning to ask the right questions

at the end of the course, you should have learned

- at least one method to solve any problem
- machine learning is not magic; it's math
- when **not** to trust your solution

Course objectives (II)

this course is about

- algorithms for big messy data
- learning to ask the right questions

at the end of the course, you should have learned

- at least one method to solve any problem
- machine learning is not magic; it's math
- when **not** to trust your solution

the rest you can learn online. . .

ORIE 4741: Learning with Big Messy Data

want to take this class?

- **ASAP:**
  - enroll (or drop) (or get on wait list)
  - fill out course survey
  - sign up for discussion forum
  - sign up for iClicker
- **Thursday 9/2/2021:** homework 0

links on course website:
`https://people.orie.cornell.edu/mru8/orie4741/`