36-462/36-662, Data Mining

Cosma Shalizi

Lecture 1, 18 January 2022 — Welcome to the course

Agenda for today

- Course mechanics
  - All of the details are in the syllabus
- General orientation to the course

What is statistical learning?

- Statistical learning:
  - how to fit predictive models
  - to training data
  - usually by solving an optimization problem
  - so the model will probably predict well
  - on average
  - on new data

Course mechanics

- Class meetings
- Readings
- Homework
- Class homepage: [http://www.stat.cmu.edu/~cshalizi/dm/22]
  - Full syllabus with all the details
  - Links to course assignments, due dates, etc.
  - What to read when
- Gradescope: submitting almost all your work
- Canvas: submitting reading questions, gradebook, solutions
- Piazza: question-answering

Class meetings

- Lecture: me explaining and demonstrating stuff, you asking questions
- In-class exercises: you checking your understanding
- No electronics when we're in person
- No recordings

In-class exercises

- Short (< 20) minute problem-solving exercises related to lecture *and* homework
- Pencil-and-paper, not electronics
- Groups of up to 4 when we're in person
- Most if not all class meetings, due via Gradescope the next day

Reading

- Most class meetings will have **key** reading: do it!
- Many will have *suggested* reading: try to do some of it
- Most will have background reading: if you get interested

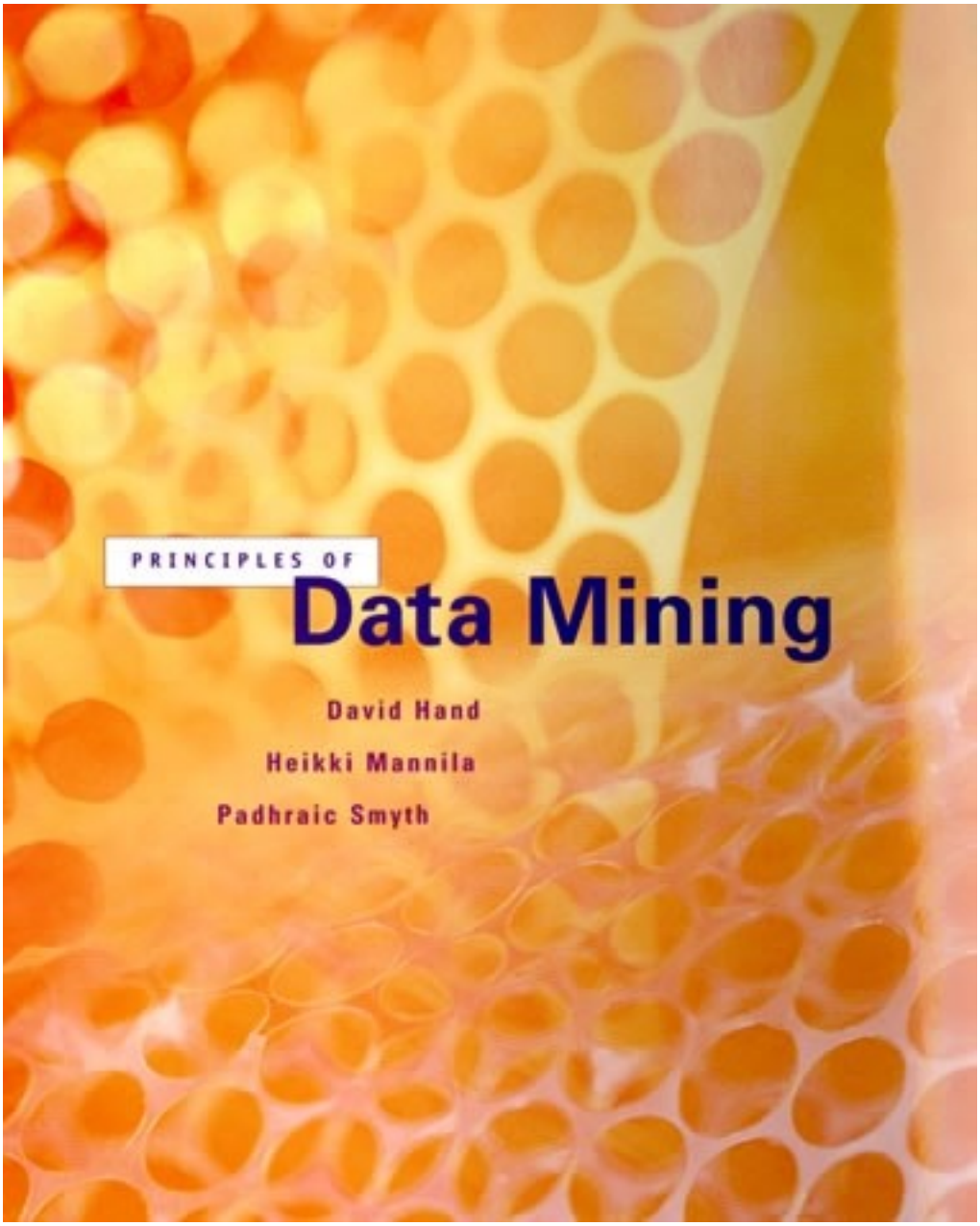
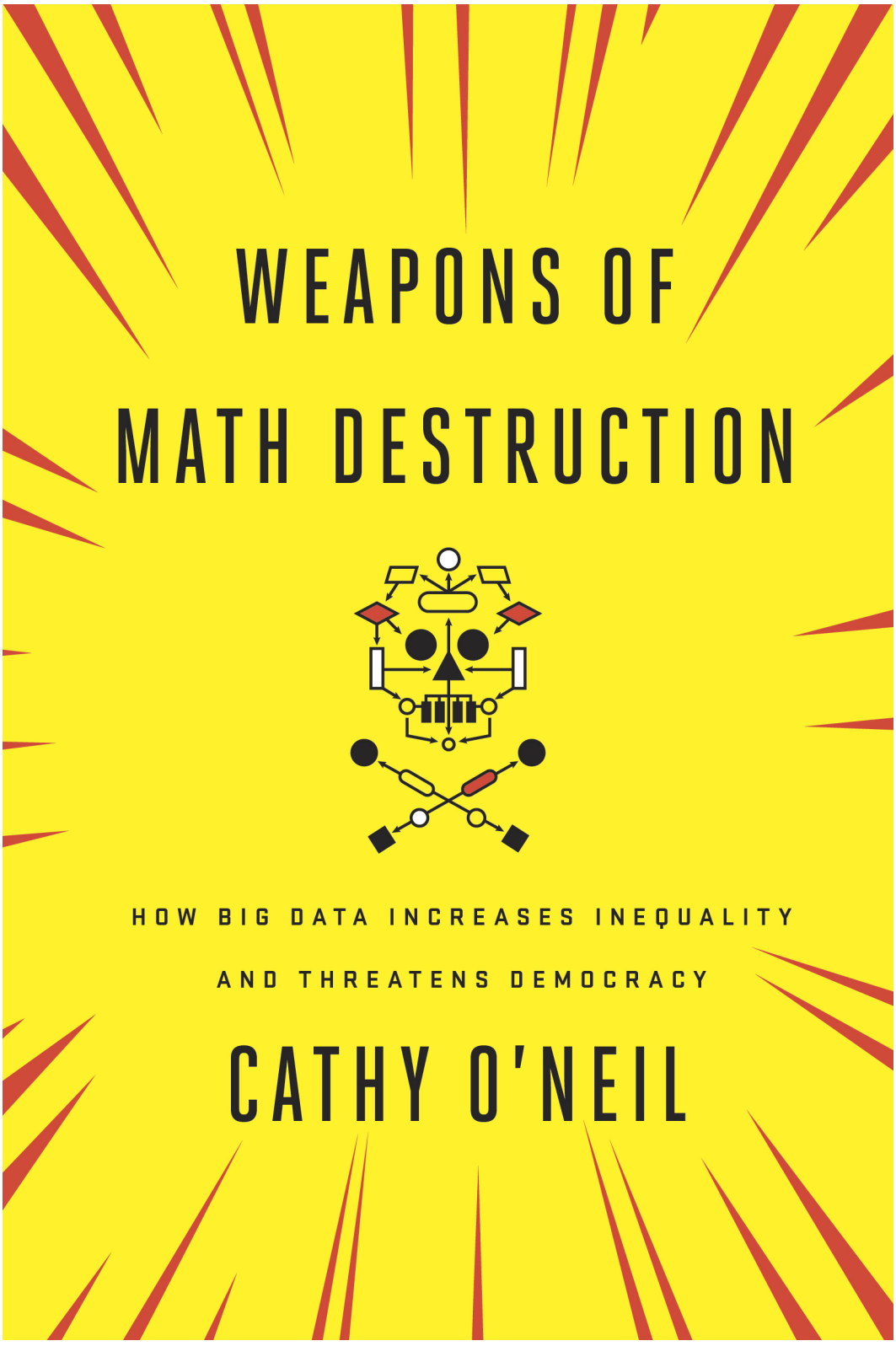


Figure 1: *Principles of Data Mining*

Reading: Textbook

Reading: Textbook



Reading: Textbook

Reading: Textbook

Reading: Recommended

Homework

- Implementing methods on actual data
- Working out some of the mathematical details
- Practicing interpreting and communicating the results
- One assignment per week, 12 in all
  - Released by Friday each week (sometimes earlier)
  - Usually due Thursdays at 6 pm via Gradescope

Homework

- 10% of each homework will be graded on the quality & clarity of your communication
  - There will be a rubric for this on each assignment
- Most (if not all) homeworks will also have an online assignment
  - Easy if you've done the reading
  - Usually due Mondays at 6 pm via Canvas
  - Online assignments will be about 10% of each homework

Grading

- Homework: 90% of your grade
  - **Lowest 3** homework grades dropped automatically
  - **No late homework**
  - If you do all homeworks with a minimum grade of $\geq 60\%$, **lowest 4** grades get dropped
- In-class exercises: 10% of your grade
  - **Lowest 5** dropped automatically
  - If you do all exercises with minimum of $\geq 60\%$, lowest **6** dropped
- No exams
- Grade boundaries: 90 for an A, 80 for a B, etc.
  - 662: 97 for an A+, etc.

Time expectations

- This is a 9 **credit-hour** class
- You spend 3 hours in lecture each week
- $\Rightarrow$ **6 hours** working on the class outside of lecture each week
  - averaged over the semester
  - Talk to me if it's taking much longer than that

Cheating, collaboration & plagiarism

- Don't
- You can talk to each other, you can read whatever you like, but everything you turn in **must** be your own work

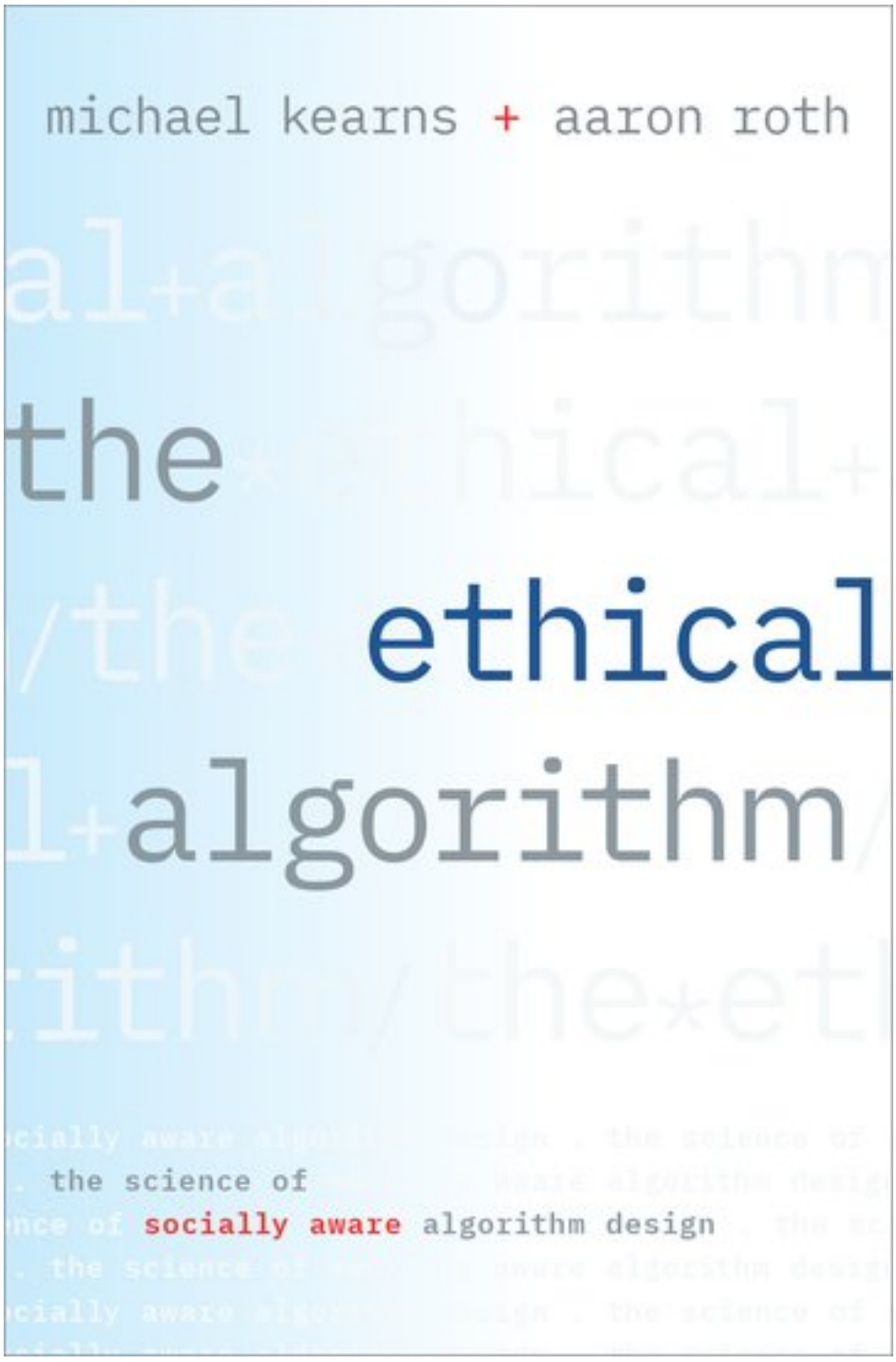


Figure 2: *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*

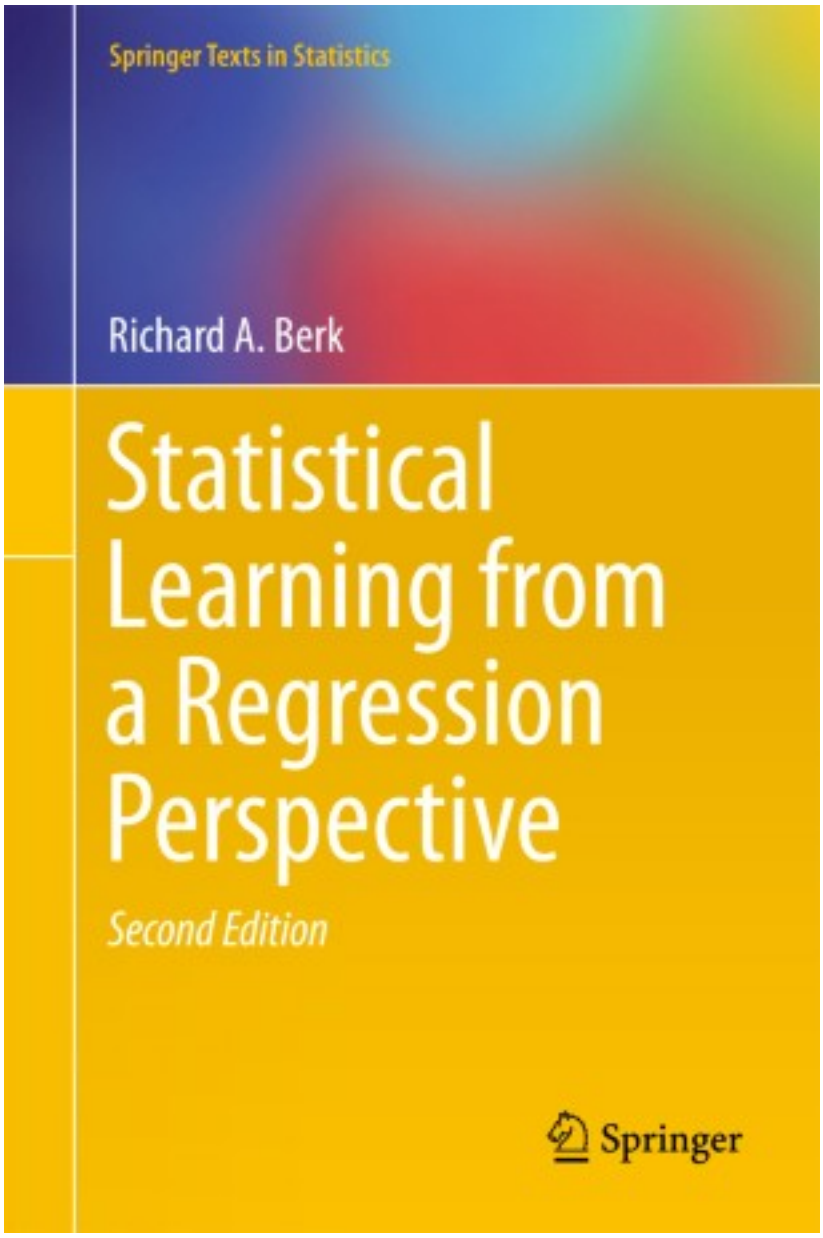


Figure 3: *Statistical Learning from a Regression Perspective*

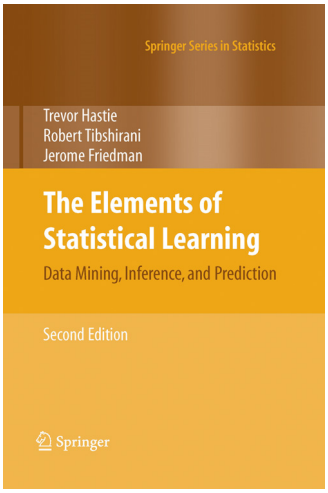


Figure 4: *Elements of Statistical Learning*

  - Exception to "read whatever you like": Don't read old solutions, or share this year's
  - Exception to "all work must be your own: Working together is OK for in-class exercises
- Full policy in the syllabus
- You will need to do a HW 0 about the class cheating policy before anything else will be graded

Homework format

- We will use R Markdown to integrate your code directly in to your writing
  - Write a source file that's mostly ordinary text, plus the R code you want
  - "Knit" to an HTML or PDF with the text plus the output of the code (figures, tables, numbers)
- Ensures **computational reproducibility**: your results really came from the code that you say/think they came from
- Keep your raw R Markdown; expect to be randomly picked to turn it in about once this semester

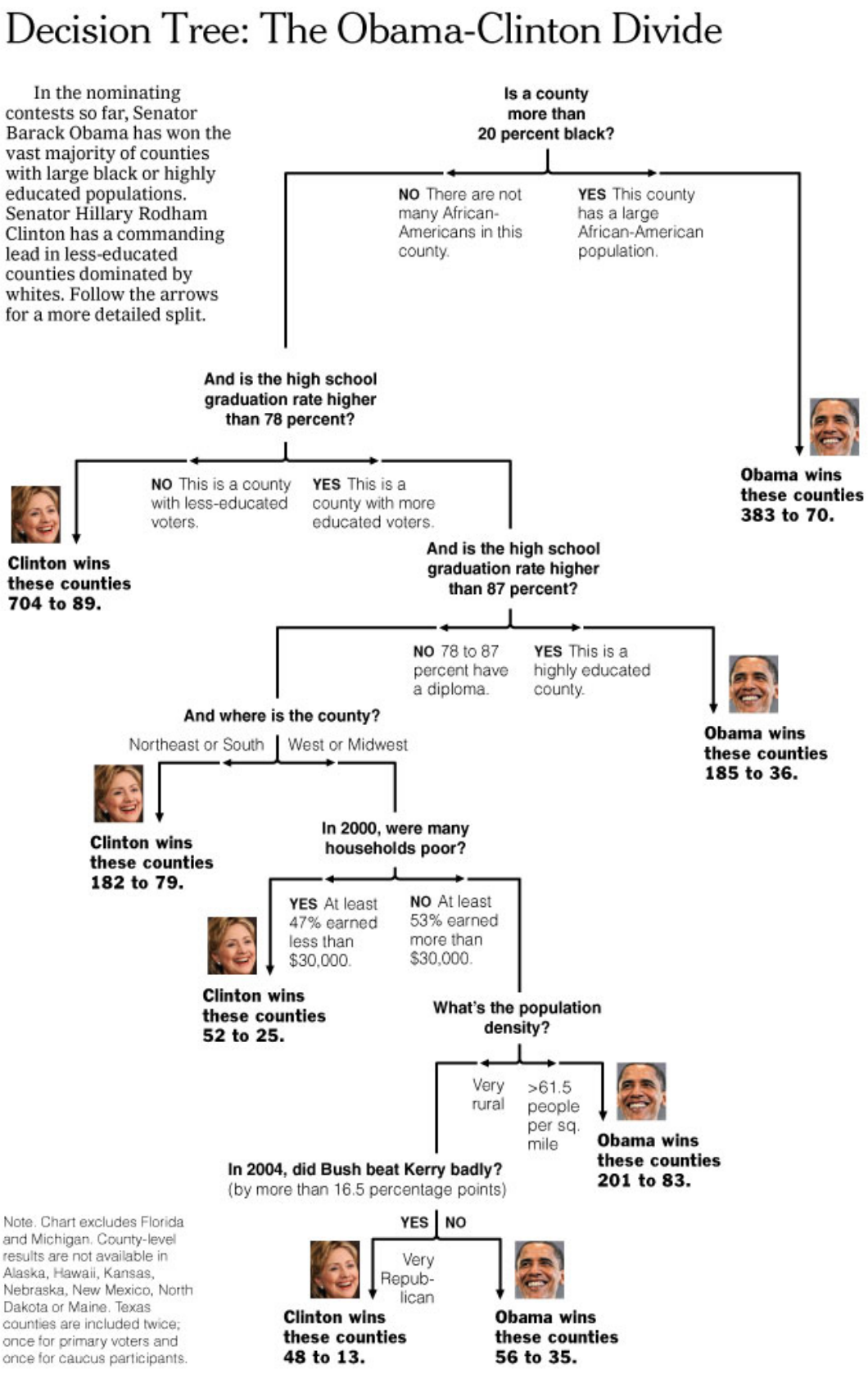
What are we going to learn about

So many things!

Nearest neighbors

- "This new case will do what similar cases did" is surprisingly powerful
- Need good ways to define "similar"
- Need good ways to find similar cases in big data sets

Prediction and decision trees



- Simple, binary-choice models for prediction
- Plus ways of combining many trees to get "forests"

Nonlinear features and kernels

- Recycle everything we know about linear models by using new, nonlinear features of the data
- Avoid having to actually calculate the transformations by using tailored similarity measures ("kernels")
- *Random* nonlinear functions are surprisingly powerful

Dimension reduction

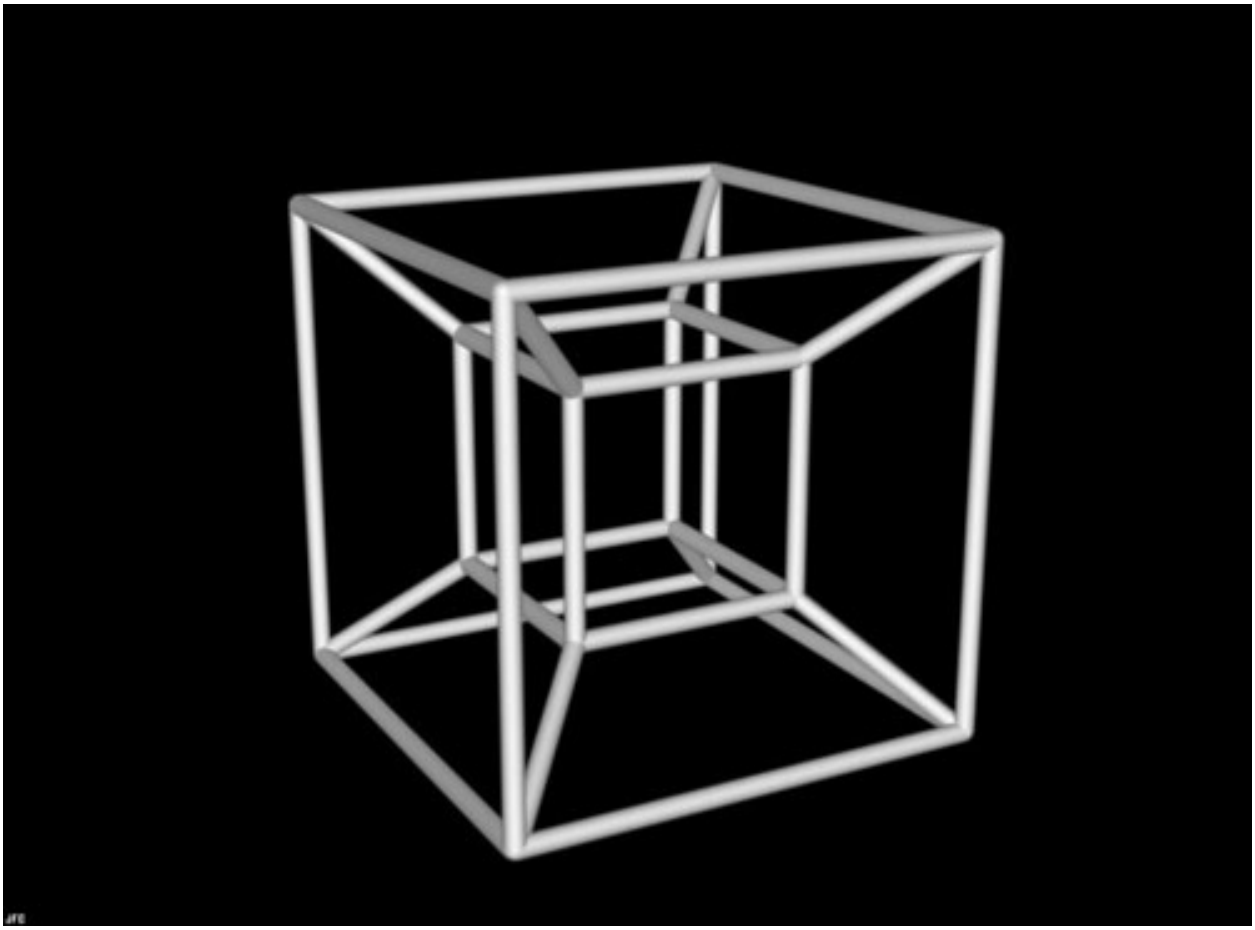


Figure 5: https://live.staticflickr.com/3560/3487720211_1df38f25e8.jpg

- More than 3D is very hard for us to grasp
  - *Maybe* 5 if you use color and animation well
- Somehow reduce the huge number of features to something more manageable *but* still intelligible

Clustering

- Divide the data into groups, without knowing what the right groups are

- Using probability models so this isn't *totally* arbitrary

Mathematical tools to make all this work

- **Decision theory** to think carefully about what's a good prediction
- **Optimization** to actually fit good models
- **Regularization** to keep the models from getting too weird

Checking our guesses

- Estimates of prediction error on *new* data
  - Models are *always* optimistic
- Approximate formulas based on optimization theory
- Cross-validation for seeing how well our models *actually* predict
  - Divide the data, fit the model to one part, evaluate predictions on the other

Applications

- Could have looked at astronomy, biology, marketing, medicine, supply-chain management, social welfare services...
- Will instead focus on four:
  - Recommendation engines
  - Credit scoring and loan decisions
  - Criminal risk/recidivism prediction
  - Predictive policing
  - You live with the first two and hopefully will never be on the receiving end of the others

Recommendation engines

- "You may also like"
- Uses: nearest neighbors, clustering, dimension reduction, classification, ...
- Abuses/unintended consequences...

Fairness in prediction

- How do we keep our models from just reproducing the injustices of the society around us?
  - *Can* we keep our models from doing this? *Should* we?
  - What counts as an injustice?
- Case study: credit scores, predicting who will repay loans, and loan-making
- Case study: how risky is it to release this just-arrested person, before their trial?
- Case study: what happens when the police try to predict *where* there will be crime? when they try to predict *who* will do the crimes?

Waste, fraud and abuse

- Sometimes statistical learning just won't work
  - Bad data
  - No useful predictions to be made
  - Overwhelming data and the curse of dimensionality

Figure 6: https://web.archive.org/web/20080901072600if_/http://failblog.files.wordpress.com/2008/01/camerafail.jpg

Waste, fraud and abuse

Figure 7: https://web.archive.org/web/20080901072600if_/http://failblog.files.wordpress.com/2008/01/camerafail.jpg

- Sometimes statistical learning is just the wrong thing to do

Where did this come from?

- Statistics worked out lots of ideas about how to *make* predictions and how to *evaluate* predictions
  - Regression, especially regression by matching and by nonlinear functions
  - Principal components and factor analysis
  - Classifiers and discriminant analysis
  - Clustering and mixture models
  - Cross-validation
  - Many of these ideas weren't very practical in the 1960s, or 1920s, or even 1800s...
- Computer scientists had been interested in getting machines to learn almost from the first computers in the 1940s
- Between 1980 and 1995, some computer scientists started using those statistical tools, and statisticians started using models and algorithms from CS
  - Or from theoretical biology ("neural networks"), physics, etc.

- "Data mining", "Knowledge discovery in data bases", "Statistical learning", "Machine learning"...

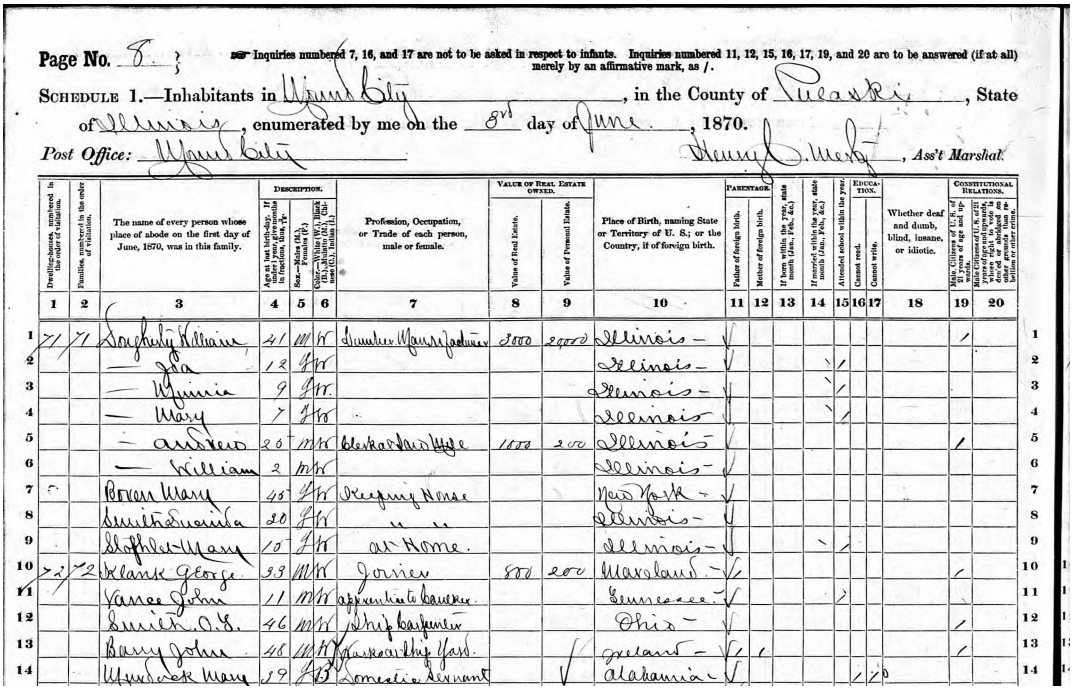
What will you need to know?

- 36-401, modern regression
- = Linear statistical models **in R**
- = Actual experience with predictive modeling of data
  - + Mathematical statistics (for notions of inference and error)
  - + Probability (for notions of distributions and risk)
  - + Linear algebra through eigenvalues and eigenvectors (**essential** for multivariate data)
  - + Calculus (**essential** for optimization)

Next time: The truth about linear regression

- A review, without the mythology, to set us up for more powerful prediction methods
- Do the reading!

Backup: Where did this *really* come from?



Page No. 8

Inquiries numbered 7, 16, and 17 are not to be asked in respect to infants. Inquiries numbered 11, 12, 15, 16, 17, 19, and 20 are to be answered (if at all) merely by an affirmative mark, as /.

SCHEDULE 1.—Inhabitants in Ward's Grove, in the County of Pulaski, State of Illinois, enumerated by me on the 8th day of June, 1870.

Post Office: Ward's Grove — Henry S. Merry, Ass't Marshal.

	Dwelling-houses numbered in the order of visitation.	Families numbered in the order of visitation.	The name of every person whose place of abode on the first day of June, 1870, was in this family.	Age at last birth-day. If under 1 year, state whether in months or days.	Sex—Males (M.), Females (F).	Color—White (W.), Black (B.), Mulatto (M.), Chinese (C.), Indian (I.).	Profession, Occupation, or Trade of each person, male or female.	Value of Real Estate.	Value of Personal Estate.	Place of Birth, naming State or Territory of U. S.; or the Country, if of foreign birth.	Parentage: Father of foreign birth.	Parentage: Mother of foreign birth.	If born within the year, state months (Jan., Feb., &c.)	If married within the year, state month (Jan., Feb., &c.)	Education: Attended school within the year.	Education: Cannot read.	Education: Cannot write.	Whether deaf and dumb, blind, insane, or idiotic.	Male Citizens of U. S., of 21 years of age and upwards.	Male Citizens of U. S., of 21 years of age and upwards, whose right to vote is denied or abridged on other grounds than participation in rebellion or crime.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	71	71	Loughridge William	41	M	W	Farmer	3000	2000	Illinois	/								/	
2			— Ida	12	F	W				Illinois	/					/				
3			— Minnie	9	F	W				Illinois	/					/				
4			— Mary	7	F	W				Illinois	/					/				
5			— Andrew	20	M	W	Clerk at Merchandise	1000	200	Illinois	/								/	
6			— William	2	M	W				Illinois	/									
7			Bower Mary	48	F	W	Keeping House			New York	/									
8			Smith Lucinda	20	F	W	" "			Illinois	/									
9			Stoffle Mary	10	F	W	at Home			Illinois	/					/				
10	72	72	Blank George	33	M	W	Farmer	800	200	Maryland	/								/	
11			Vance John	11	M	W	Apprentice to Carpenter			Tennessee	/					/				
12			Smith O. B.	46	M	W	Ship Carpenter			Ohio	/								/	
13			Barry John	48	M	W	Blacksmith Ship Yard		/	Ireland	/	/							/	
14			Woodcock Mary	19	F	B	Domestic Servant		/	Alabama	/					/	/	D		

- "The state, the coldest of all cold monsters"

Backup: Where did this *really* come from?

- We invented data-processing machines, *before* computers, because we were already keep tabs on so much about so many people
- We have traditions of data-driven prediction and decision-making going back hundreds of years

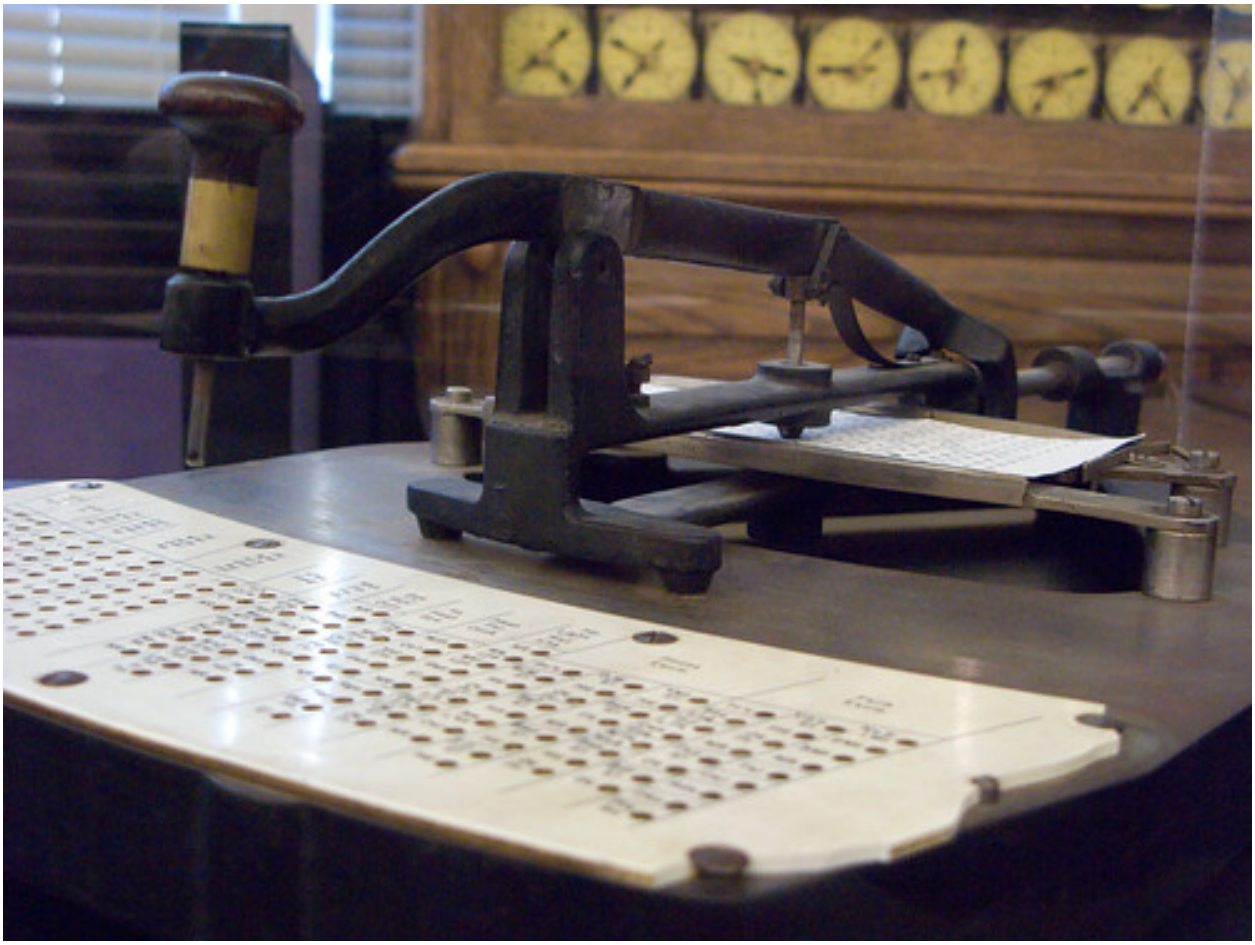


Figure 8: http://farm3.static.flickr.com/2411/2404562785_5b887699de.jpg

Backup: Where did this *really* come from?

Figure 9: https://live.staticflickr.com/369/19194898203_6f9b3bba5f_c.jpg

- Computers made it *really easy* to **create**, **store** and **analyze** data
  - Create: sensors (phones, cameras, cash registers...), digitization (doing everything via computer)
  - Store: hard disks, disk arrays, data-bases...
  - Analyze: that's where we come in
  - Organizations create and store the data even when they don't have particular analyses in mind

Backup: Where did this *really* come from?

- Tension: flexibility to find many different patterns vs. vulnerability to noise and coincidence
  - All the data is great: brilliant ideas from 50–100 years ago become practical!
  - Why we look a lot of different ways of finding patterns
  - All the data is a problem: If you run a gazillion analyses, (gazillion/20) will be significant at the 5% level
  - Why we look at how to avoid fooling ourselves

Backup: Where did this *really* come from?

- If we hadn't spent centuries re-shaping our societies to collect and act on data, dropping in people skilled in statistical learning wouldn't accomplish very much
- Even dropping in statistical learners *and computers* wouldn't accomplish very much