Data Science Education: 2020 Vision and Versions

Leading educators from multiple universities and disciplines will discuss their visions and initiatives in providing general data science education on their campuses, followed by Q&A.

Chair: Dustin Tingley
Harvard University

Michael Jordan and Ani Adhikari
Computer Science & Statistics, UC Berkeley

Michael Franklin
Computer Science, University of Chicago

Matthew Jones
History, Columbia University

Joseph Blitzstein
Statistics, Harvard University

Alison Gibbs
Statistics, University of Toronto
Data Science Education: 2020 Vision and Versions

Michael Jordan and Ani Adhikari
University of California, Berkeley
Our Vision

● All students should be able to make decisions based on data
● Data science education should be a conceptual blend of traditions in statistics and computer science, with a strong outward-looking component
● All students should be able to tackle significant problems, the complexity increasing with the students’ background knowledge and interests
The Vision, Realized

- *Foundations of Data Science* (data8.org)
  - Connector courses
- Human context and ethics
- Some math and computing
- *Principles and Techniques of Data Science* (data100.org)
- *Probability for Data Science* (prob140.org)
- *Data, Inference, and Decisions* (data102.org)
Data, Inference, and Decisions
The Syllabus

- Review: Frequentist and Bayesian Decision Making
- False Discovery Rate Control
- Online FDR
- Probability Interpretation of Linear and Logistic Models
- Bayesian Hierarchical Models
- Gibbs Sampling and Importance Sampling
- Confidence and Credible Intervals
- Chernoff and Hoeffding Bounds
- Causal Inference
- Rudiments of Experimental Design
- Bandits: Greedy and UCB Algorithms
- Thompson Sampling
- Dynamic Programming and Q-Learning
- Introduction to LQR and Control Theory
- Matching Markets
- Nonparametric Methods
- Privacy
- Real-world Consequences of Decisions
- Communication: Presenting Results and Conclusions
The Students

2018-2019
● 2800 in Data 8
● 1600 in Data 100
● 500 in Prob 140

The first Data Science majors graduated in Spring 2019.

Over 300 DS majors will graduate in Spring 2020, along with several hundred who will get the DS minor.
Inaugural Symposium

hdsr.mitpress.mit.edu

Data Science Education: 2020 Vision and Versions

Michael Franklin
University of Chicago
Data Science at U Chicago

Michael Franklin
Liew Family Chairman of Computer Science
Sr. Advisor to the Provost for Computing and Data Science

October 25, 2019
HDSR Education Panel
UChicago Data Science Context

Recruitment of key senior faculty:
- Luis Bettencourt, Pritzker Director of the Mansueto Institute for Urban Innovation and Professor of Ecology and Evolution
- Nick Feamster, Neubauer Professor of CS and Faculty Director of CDAC
- Sendhil Mullainathan, University Professor at Chicago Booth
- Rebecca Willett, Professor of Statistics and Computer Science
- Ben Zhao, Neubauer Professor of CS
- Heather Zheng, Neubauer Professor of CS

Many excellent junior hires in CS, Stats, Biological Sciences, Business, Econ, Molecular Engineering, Soc. Sci, …

Currently: Open search for Data Science Faculty (joint CS/Stats and others)
# Center for Data & Computing - 2019 Grants

<table>
<thead>
<tr>
<th>Category</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>When Technology Transforms Society:</td>
<td>The <strong>Societal and Ethical Impacts</strong> of Quantum Computing and AI</td>
</tr>
<tr>
<td>A Data Processing Pipeline to Transcribe</td>
<td><strong>Broadcast Police Communications</strong></td>
</tr>
<tr>
<td>N-body Networks for</td>
<td><strong>Jet Physics at the Energy Frontier</strong></td>
</tr>
<tr>
<td>Seminar Series: Rising Stars in Data Science</td>
<td>(<strong>diversity initiative</strong>)</td>
</tr>
<tr>
<td>Racial Inequality in Financial Resilience</td>
<td></td>
</tr>
<tr>
<td>Rational <strong>Protein Engineering</strong> Using Data-Driven Generative models</td>
<td></td>
</tr>
<tr>
<td>Computational Modeling to Quantify</td>
<td><strong>Social Determinants of Cardiovascular Disease</strong></td>
</tr>
<tr>
<td>The role of rapid <strong>bacterial evolution</strong> in human health and disease</td>
<td></td>
</tr>
<tr>
<td><strong>Empowering Doctors</strong> in Areas with No-Internet Coverage with a Mobile Decision-making Interface</td>
<td></td>
</tr>
<tr>
<td>Engineering Opportunity: Optimal Skill and Relationship Mixtures for Individual, Enterprise &amp; Regional Prosperity</td>
<td></td>
</tr>
<tr>
<td>Predicting Shifts in Biological Growth Driven by <strong>Climate Change</strong>: A Geometric Deep Learning Approach</td>
<td></td>
</tr>
<tr>
<td>Disentangling <strong>Visual Style</strong> and Content</td>
<td></td>
</tr>
<tr>
<td><strong>Clinical Behavior Analysis</strong> via Unsupervised Learning</td>
<td></td>
</tr>
</tbody>
</table>
Data Science: An Emerging New Discipline
(results from a campus-wide committee on Data Science)
Data Science: An Emerging New Discipline
(results from a campus-wide committee on Data Science)
Data Science: An Emerging New Discipline (results from a campus-wide committee on Data Science)
The Data Science Lifecycle

Acquire
- Create, capture, gather from:
  - Lab
  - Fieldwork
  - Surveys
  - Devices
  - Simulations
  - etc

Clean
- Organize
- Filter
- Annotate
- Clean

Use / Reuse
- Analyze
- Mine
- Model
- Derive data
- Visualize
- Decide
- Act
- Drive:
  - Devices
  - Instruments
  - Computers

Publish
- Share
- Data
- Code
- Workflows
- Disseminate
- Aggregate
- Collect
- Create portals, databases, etc
- Couple with literature

Preserve / Destroy
- Store to:
  - Preserve
  - Replicate
  - Ignore
- Subset, compress
- Index
- Curate
- Destroy

{Ethics, Policy, Regulatory, Stewardship, Platform, Domain} Environment

Realizing the Potential of Data Science

Communications of the ACM, April 2018

By Francine Berman, Rob Rutenbar, Brent Hailpern, Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, Margaret Martonosi, Padma Raghavan, Victoria Stodden, Alexander S. Szalay
Data Science Undergraduate Minor

- Jointly administered between CS and Statistics
- Six course program suitable for all majors*

1. Introductory Sequence (four courses required):

<table>
<thead>
<tr>
<th>Course Code</th>
<th>Course Title</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMSC/STAT 11800</td>
<td>Introduction to Data Science, I</td>
<td>100</td>
</tr>
<tr>
<td>CMSC/STAT 11900</td>
<td>Introduction to Data Science, II</td>
<td>100</td>
</tr>
<tr>
<td>CMSC 25300</td>
<td>Mathematical Foundations of Machine Learning</td>
<td>100</td>
</tr>
<tr>
<td>CMSC 25900</td>
<td>Ethics, Data, and Privacy</td>
<td>100</td>
</tr>
</tbody>
</table>

2. Elective Sequence (two courses required):

Two of the following

<table>
<thead>
<tr>
<th>Course Code</th>
<th>Course Title</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMSC 13600</td>
<td>Data Engineering</td>
<td>100</td>
</tr>
<tr>
<td>CMSC 23900</td>
<td>Data Visualization</td>
<td>100</td>
</tr>
<tr>
<td>STAT 22200</td>
<td>Design of Experiments</td>
<td>100</td>
</tr>
</tbody>
</table>

Total units 200

*Special “concentrations” for Computer Science, Economics and Statistics Students
Data Science Undergraduate Minor

- Jointly administered between CS and Statistics
- Six course program suitable for all majors*

1. Introductory Sequence (four courses required):

<table>
<thead>
<tr>
<th>Course Code</th>
<th>Course Title</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMSC/STAT 11800</td>
<td>Introduction to Data Science, I</td>
<td>100</td>
</tr>
<tr>
<td>CMSC/STAT 11900</td>
<td>Introduction to Data Science, II</td>
<td>100</td>
</tr>
<tr>
<td>CMSC 25300</td>
<td>Mathematical Foundations of Machine Learning</td>
<td>100</td>
</tr>
<tr>
<td>CMSC 25900</td>
<td>Ethics, Data, and Privacy</td>
<td>100</td>
</tr>
</tbody>
</table>

2. Elective Sequence (two courses required):

<table>
<thead>
<tr>
<th>Course Code</th>
<th>Course Title</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMSC 13600</td>
<td>Data Engineering</td>
<td>100</td>
</tr>
<tr>
<td>CMSC 23900</td>
<td>Data Visualization</td>
<td>100</td>
</tr>
<tr>
<td>STAT 22200</td>
<td>Design of Experiments</td>
<td>100</td>
</tr>
</tbody>
</table>

Total units 200

*Special “concentrations” for Computer Science, Economics and Statistics Students
Data Science Undergraduate Courses

- CS/Stats 118/119 Introduction to Data Science I and II
  - 118 based on Berkeley Data 8 with similar prerequisite assumptions
    - Uses the Berkeley datascience library
  - 119 covers additional analytics/ML, wrangling, scalability
    - Uses real tools: Pandas, Scikit Learn, Apache Spark, SQL, …
    - Increased focus on the Data Science Lifecycle
  - Currently jointly taught by the Chairs of the CS and Stats departments

- New Courses on Data Engineering, Responsible Data Science, Math for ML
- Minor serves as a foundation for development of a Data Science Major
- CS, Stats and Econ majors have Data Science “concentrations”
Foundational Research Questions in Data Science

- A Theory of Data Integration and Fusion
- Weaving the Statistical and Logic-based viewpoints
- Data Curation and Reproducible Research
- Data Ethics, Privacy and Security
- Value and Economic Laws of the Data Economy
- Decision Making With and Quantification of Uncertainty
- Data Visualization and Cognitive Biases of Interpretation
- …
Next Steps

• Data Science Major under development
• **Hiring initial cadre of faculty** to form core team for Data Science
  • Currently bootstrapping with CS and Statistics faculty
  • Data Science search is underway
• Establishing a Data Science research agenda and collaborations
• Establishing a Ph.D. program
• “Training” for Ph.D. students in other areas
Data Science Education: 2020 Vision and Versions

Alison Gibbs
University of Toronto
NAVIGATING WHITEWATER:
UNDERGRADUATE STATISTICS IN THE ERA OF DATA SCIENCE

Alison L. Gibbs
October 25, 2019
UNDERGRADUATES IN STATISTICS PROGRAMS OF STUDY AT U OF T
UNDERGRADUATES IN STATISTICS PROGRAMS OF STUDY AT U OF T

4500


0 500 1000 1500 2000 2500 3000 3500 4000 4500 5000

all
major
minor
specialist
applied specialist
"I keep saying the sexy job in the next ten years will be statisticians."

Hal Varian, Google chief economist
McKinsey Quarterly
"Data Scientist: The Sexiest Job of the 21st Century"

Thomas H. Davenport and D.J. Patil
Harvard Business Review
UNDERGRADUATES IN STATISTICS PROGRAMS OF STUDY AT U OF T

Interesting
Useful
Complements other study
Job opportunities
Enjoyed first course
Combining statistics with a major in another discipline

Combining statistics with a focus in another discipline
AN EXPERT

- Has extensive knowledge
- Has skills to apply that knowledge quickly and efficiently
- In settings seen before

AN ADAPTIVE EXPERT

- Can transfer knowledge in novel or complex situations
- Can create new procedures in novel or complex situations
- Is flexible
- Able to innovate
- Continuous learner
- Seeks challenges
- Creative
- Has better developed meta-cognitive skills
DEVELOPING ADAPTIVE EXPERTISE

- Need to spend more time on:
  
  - Learning that emphasizes understanding
  - Opportunities to explore, discover, and struggle
  - Experience with lots of variations on problems
A FIRST COURSE  (STA130.UTSTAT.UTORONTO.CA)

**Mondays**
- Large lecture sections
- Data stories
- A new idea or method each week
- R by example to answer data questions

**Fridays**
- Practice problems
- R Markdown template to start solutions
- One solution brought to tutorial for grading and discussion
- Oral or written communication exercise each week

**Project**

**Fridays**
- Small group tutorials
- Random group assignments each week
STA130 Project Poster Fair
A NEW VIEW OF STATISTICS PROGRAMS OF STUDY

Statistics
Domain expertise
Computation
Statistical Science encompasses methods and tools for obtaining knowledge from data and for understanding the uncertainty associated with this knowledge.

The purpose of our undergraduate programs are to:

1. equip students with a general framework for obtaining knowledge from data;
2. give students skills that they are able to flexibly apply to a variety of problems; and
3. provide students with the ability to learn new methods as needs, data sources, and technology change
END GOALS FOR STUDENTS IN OUR STATISTICS PROGRAMS OF STUDY

20 Program Learning Outcomes organized into 5 themes:

- Statistical Theory, including Probability
- Methods and Applications
- Computational Thinking
- Professional Practice
- Problem Solving
Data Science Education: 2020 Vision and Versions

Joseph Blitzstein
Harvard University
Inaugural Symposium

Data Science Education: 2020 Vision and Versions

Matthew Jones
Columbia University
Data: Past, Present, Future

Matthew Jones
Department of History, Columbia
@nescioquid
joint work with:
Chris Wiggins

data-ppf.github.io
aka
what should future statisticians CEOs, and senators know about the history and ethics of data?

Matthew Jones
Department of History, Columbia
@nescioquid
joint work with:
chris.wiggins@columbia.edu
data-ppf.github.io
what should future statisticians CEOs, and senators know about the history and ethics of data?

1. hypotheses
2. why ethics?
3. what we taught
4. what we learned
Hypotheses

there is important material being taught neither to future statisticians nor to future senators
Hypotheses

capabilities are teachable without prerequisite
  Functional
  Rhetorical
  Critical
Hypotheses

pair intellectual changes with political and ethical context
what powers motivated this advance?
how did this advance rearrange power?
Historians Politely Remind Nation To Check What's Happened In Past Before Making Any Big Decisions

9/28/11 9:00am • SEE MORE: SCIENCE & TECHNOLOGY

Trying to avoid repeating bad things we did in the past is a good idea, historians say.
2. why ethics?

2017-09-05: cathy o’neil

2018-01-08: safiya noble

2018-01-23: virginia eubanks

2019-01-15: shoshana zuboff

(among increasingly many others)

something is wrong on the internet
2. why ethics?

1. fuzzies
2. techies
what should future statisticians CEOs, and senators know about the history and ethics of data?

0. preamble: class origin story
1. why history?
2. why ethics?
3. what we taught
4. what we learned
data 1770s-present: capabilities & intents

- 1770s
  - Qualitative statistics
  - "Vulgar" statistics

- 1830s
  - Mathematical statistics

- 1900
  - Regression + Hypothesis testing

- WWII
  - Computation + Crypto

- 1960s
  - "intelligence"

- 1980s
  - AI winter

- 1990s
  - Machine learning

- You are here
  - AI renaissance
  - (surveillance economy)
  - (defense + advertising)
  - (data wars (for funding))
  - (data @ war)
  - (math vs. science)
  - (policy + eugenics)
  - (statecraft)
data 1770s-present: capabilities & intents

(really this is just weeks 3-11)
data 1770s-present: capabilities & intents

(really this is just weeks 3-11)
close each week with:
- how did new capabilities rearrange power? (who can now do what, from what, to whom?)
- role of rights
- harms
- justice

(week 1 & 2 had plenty of harms+injustice)
3. what we taught: 14 weeks: Tuesday discussion

1 intro
2 setting the stakes
3 risk and social physics
4 statecraft and quantitative racism
5 intelligence, causality, and policy
6 data gets real: mathematical baptism
7 WWII, dawn of digital computation
8 birth and death of AI
10 data science, 1962-2017
11 AI2.0
12 ethics
13 present problems & VC-backed attention economy
14 future solutions

3. what we taught: 14 weeks: Thursday Labs

1. first steps in Python interrogating the UCI dataset
2. EDA with the UCI dataset
3. Quetelet and GPAs
4. Galton
5. statistics and society; Yule, Spearman, Simpson
6. p-hacking; Fisher
7. the first data science
8. AI 1.0; Expert systems; Perceptron
9. databases and recsys; the Netflix Prize story
10. trees along with in-lab lecture on trees
11. interactive: 3 ML’s; FAT 1.0 disparate impact, disparate treatment, and
12. normative+technical approaches to defining and defending privacy; our own
13. problems along with in-lab lecture on NSA history
14. solutions

e.g., week 3 “risk and social physics”
e.g., week 4 regression & quantitative racism
e.g., week 4 regression & quantitative racism

Data: Past, Present, Future | Lab 4 | 2/14/2019
describing and predicting: Galton, regression, inventing error, sur

Galton and

Now it's your turn!

can you the regression for

1. everybody and his/her mother?
2. males and fathers
e.g., week 5 IQ, policy and causality

"GENERAL INTELLIGENCE."

the general diminution caused by the impurity: Classics o.60, French o.56, English o.45, and Mathematics o.39.

DAVID FREEDMAN

From association to causation: some remarks on the history of statistics

"Journal de la société française de statistique", tome 140, no 3 (1999), p. 5-32

SOME REMARKS ON THE HISTORY OF STATISTICS

that welfare outside the poor-house creates paupers – the estimated coefficient on the out-relief ratio is positive.
e.g., week 5 IQ, policy and causality

for more, see "Paradoxes" chapter in Pearl 2018
Breaking Codes and Finding Trajectories: Women at the Dawn of the Digital Age

I was a Colossus operator, which we considered to be the crème de la crème. We felt we were “at the sharp end,” where there was a great tension and flow of adrenaline ... operating those incredible machines.

—Jean Beech, Colossus operator¹

I don’t know if you can picture how exciting the ENIAC was to all of us. And we didn’t talk socially or any other time about anything else. It was—we discussed it almost all the time.

—Jean Jennings, ENIAC programmer²

¹ Jean Beech
² Jean Jennings
e.g., week 7 “sigint”

Lab 7: Let's be Bayesian Cryptologists

Before we get started, figure out what a Vigenère cypher is. Use the google or the bing or the duckduckgo.

Now, encipher, with a Vigenère cypher, your lastname (e.g. "wiggins") using the key "lego." You can use a Vigenère square to help you.

In [1]:
   from itertools import starmap, cycle

   def encrypt(message, key):
      # convert to uppercase.
      # strip out non-alpha characters.
      message = filter(str.isalpha, message.upper())

      # single letter encryption.
      def enc(c, k):
         return chr(((ord(k) + ord(c) - 2*ord('A')) % 26) + ord('A'))

      return ''.join(starmap(enc, zip(message, cycle(key))))

In [2]: encrypt("wiggins", "optimusprime")
Out[2]: 'QDFUANQ'

In [3]: encrypt("wiggins", "optimusprime")
Out[3]: 'QDFUANQ'

Try your own!
e.g., week 10 “data science”, 1962-present
“three kinds of ML” (interactive)

...forests are A+ predictors. But their mechanism for producing a prediction is difficult to understand. Trying to delve into the tangled web that generated a plurality vote from 100 trees is a Herculean task. So on interpretability they rate an F.
e.g., week 12 the ethics of data

How will AI change your life? AI Now Institute founders Kate Crawford and Meredith Whittaker explain.
e.g., week 12 the ethics of data

history: Tuskegee -> Belmont
e.g., week 12 the ethics of data

1. articulate principles
2. articulate tensions among them
3. design to support them - interaction design - process design (in this case, the IRB process)
e.g., week 12 the ethics of data

history: Tuskegee -> Belmont

1. articulate ethics as principles
2. articulate tensions among them
3. articulate design to support them
e.g., week 12 ethics lab:
- database of ruin
- k-anonymity
- terms of service

_kdnuggets.com_ (2014):
“Big Data Comic Explains the Current State of Privacy”
4. what we learned

1. history + ethics: how to integrate throughout a “tech” education
2. draw parallels to today
3. capabilities rearrange power
4. story of “data” is story of truth+power
   - contested
Data: Past, Present, Future

Matthew Jones
Department of History, Columbia
@nescioquid
joint work with:
Chris Wiggins

data-ppf.github.io