data science @ The New York Times

chris.wiggins@columbia.edu
chris.wiggins@nytimes.com
@chrishwiggins

biology: 1892 vs. 1995
biology: 1892 vs. 1995

new toolset, new mindset
“These are indeed exciting times, not unlike the early days of recombinant DNA in the 1970s, in which a revolutionary new technology permitted entirely new questions about the nature of genes to be raised. This challenge is new to biology, and its resolution will require, in addition to existing paradigms of molecular biology, new sets of analytical tools... disciplines outside of biology will be required to collaborate on this problem.”
genetics: 1837 vs. 2012

ML toolset; data science mindset
genetics: 1837 vs. 2012
ML toolset; data science mindset

arxiv.org/abs/1105.5821 ; github.com/rajanil/mkboost
data science: mindset & toolset
news: 20th century

church

state
2015: The Year in Visual Stories and Graphics
THE 2016 RACE

‘Serious Voter Fraud’? Um, No

There are many reasons Donald J. Trump’s claim doesn’t withstand scrutiny.

8h ago · By NATE COHN
The Upshot
A New York Times website with analysis and data visualizations about politics, policy and everyday life.

- New York, NY

Repositories:
- **4thdownbot-model**
  - The model behind @NYT4thDownBot
  - Python
  - 93 Stars
  - 14 Forks
  - Updated on Oct 16

- **2016-upshot-siena-polls**
  - Data from the 2016 NYT Upshot/Siena voter file polls
  - R
  - 3 Stars
  - 3 Forks
  - Updated on Sep 23

Top languages:
- R
- Python
- Ruby
- HTML

People:
- Jeremyjbowers
  - Jeremy Bowers
news: 20th century

church  state
news: 21st century

church

state

data
newspapering: 1851 vs. 1996

1851

1996
1,615,934 site-wide views over the last hour
1,257,958 average Sunday New York Times print circulation
554 stories written over the last 24 hours
206 countries with visitors in the past 25 minutes
243,192 words written in the last 24 hours
65 New York Times newspaper print sites globally
733 page views from India in the last 10 minutes
BASEBALL

World Series Drop

OCT. 30, 2014

The New York Times
Trendy cafe chain offering upscale coffee drinks & pastries, plus beans & brewing equipment.

160 Berry St, Brooklyn, NY 11249
Open today 7:00 am – 7:00 pm

(510) 653-3394

Popular times: Tuesdays
"...social activities generate large quantities of potentially valuable data...The data were not generated for the purpose of learning; however, the potential for learning is great"
"...social activities generate large quantities of potentially valuable data...The data were not generated for the purpose of learning; however, the potential for learning is great” - J Chambers, Bell Labs, 1993, “GLS”
data science: the web
data science: the web

is your “online presence”
data science: the web

is a microscope
data science: the web

is an experimental tool
newspapering: 1851 vs. 1996 vs. 2008

The New York Times Introduces a Web Site

The New York Times begins publishing daily on the World Wide Web today, offering readers around the world immediate access to most of the daily newspaper's contents.

The New York Times on the Web, as the electronic publication is known, contains most of the news and feature articles from the current day's printed newspaper, classified advertising, reporting that does not appear in the newspaper, and interactive features including the newspaper's crossword puzzle.
“a startup is a temporary organization in search of a repeatable and scalable business model” –Steve Blank
every publisher is now a startup
every publisher is now a startup
Advertisers adjusted spending accordingly. In the first quarter of 2016, 85 cents of every new dollar spent in online advertising will go to Google or Facebook, said Brian Nowak, a Morgan Stanley analyst.

every publisher is now a startup
news: 21st century

church

state

data
news: 21st century

church

state

data
learnings
learnings

- descriptive modeling
- predictive modeling
- prescriptive modeling
(actually ML, shhhhh...)
Supervised Learning

Reinforcement Learning

Unsupervised Learning

2012; h/t michael littman
learnings

- descriptive modeling
- predictive modeling
- prescriptive modeling
<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive</td>
<td>Specify $x$; learn $z(x)$ or $p(z</td>
</tr>
<tr>
<td>Predictive</td>
<td>Specify $x$ and $y$; learn to predict $y$ from $x$</td>
</tr>
<tr>
<td>Prescriptive</td>
<td>Specify $x$, $y$, and $a$; learn to prescribe $a$ given $x$ to maximize $y$</td>
</tr>
</tbody>
</table>
descriptive modeling, e.g.,

cf. daeilkim.com ; import bnpy
driving question: what content should we translate in which countries?

cf. daeilkim.com ; import bnpy
recommendation as inference

<table>
<thead>
<tr>
<th>Most Emailed</th>
<th>Most Viewed</th>
<th>Recommended for You</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. THE OUTLAW OCEAN A Renegade Trawler, Hunted for 10,000 Miles by Vigilantes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Campus Suicide and the Pressure of Perfection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. As Tech Booms, Workers Turn to Coding for Career Change</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Prison Worker Who Aided Escape Tells of Sex, Saw Blades and Deception</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Under Oath, Donald Trump Shows His Raw Side</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. American Hunter Killed Cecil, Beloved Lion That Was Lured Out of Its Sanctuary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. A Creature on the Loose Puts Milwaukee Residents on Edge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. NFL Upholds Tom Brady’s Ban; Cellphone’s Fate Helped Make the Call</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Escalator Death in China Sets Off Furor Online</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. DAVID BROOKS The Structure of Gratitude</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
recommendation as inference

bit.ly/AlexCTM
CTM generative model:

r: clicks
w: words
u: user-topic association
v: article-topic association

Chong & Blei, SIGKDD 2011
related: learning phenotypes from EHR

Learning probabilistic phenotypes from heterogeneous EHR data.

Pivovarov R¹, Perotte AJ², Grave E³, Angiolillo J⁴, Wiggins CH⁵, Elhadad N⁶.
Learning probabilistic phenotypes from heterogeneous EHR data.

Pivovarov R¹, Perotte AJ², Grave E³, Angiolillo J⁴, Wiggins CH⁵, Elhadad N⁶.
driving question:
what should we show users?

1. THE OUTLAW OCEAN
   A Renegade Trawler, Hunted for 10,000 Miles by Vigilantes

2. Campus Suicide and the Pressure of Perfection

3. As Tech Booms, Workers Turn to Coding for Career Change

4. Prison Worker Who Aided Escape Tells of Sex, Saw Blades and Deception

5. Under Oath, Donald Trump Shows His Raw Side

6. American Hunter Killed Cecil, Beloved Lion That Was Lured Out of Its Sanctuary

7. A Creature on the Loose Puts Milwaukee Residents on Edge

8. N.F.L. Upholds Tom Brady’s Ban; Cellphone’s Fate Helped Make the Call

9. Escalator Death in China Sets Off Furor Online

10. DAVID BROOKS
    The Structure of Gratitude
learnings

- descriptive modeling
- predictive modeling
- prescriptive modeling
Air Bag Flaw, Long Known to Honda and Takata, Led to Recalls

By HIROKO TABUCHI  SEPT. 11, 2014

The air bag in Jennifer Griffin’s Honda Civic was not among the recalled vehicles in 2008. Jim Keely

work w/Daeil Kim & Hiroko Tabuchi
The most predictive words / features

Learned Weights - Regularized @ 0.3

After training the model, we then applied this on the full dataset.

We looked for comments that Hiroko didn't label as being suspicious, but the algorithm did to follow up on (374 / 33K total).

**Result:** 7 new cases where a passenger was injured were discovered from those comments she missed.

work w/Daeil Kim & Hiroko Tabuchi
work w/Daeil Kim & Hiroko Tabuchi
driving question: which records should she investigate?

The most predictive words / features

Learned Weights - Regularized @ 0.3

- Predictive of a normal comment.
- Predictive of a suspicious comment

Result: 7 new cases where a passenger was injured were discovered from those comments she missed.

work w/Daeil Kim & Hiroko Tabuchi
predictive modeling, e.g.,

cf. modelingsocialdata.org
predictive modeling, e.g.,

“the funnel”

cf. modelingsocialdata.org
interpretable predictive modeling

cf. modeling.socialdatalab.org
<table>
<thead>
<tr>
<th>TFNAME</th>
<th>DB-MOTIF</th>
<th>MOTIF</th>
<th>DBNAME</th>
<th>d(p,q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBF1</td>
<td>CACGTG</td>
<td></td>
<td>YPD</td>
<td>0.092635</td>
</tr>
<tr>
<td>CGG everted repeat</td>
<td>CGGN'CCG</td>
<td></td>
<td>YPD</td>
<td>0.092821</td>
</tr>
<tr>
<td>MSN2</td>
<td>AGGG</td>
<td></td>
<td>TRANSFAC</td>
<td>0.085626</td>
</tr>
<tr>
<td>HSF1</td>
<td>TTCNNNGAA</td>
<td></td>
<td>SCPD</td>
<td>0.102410</td>
</tr>
<tr>
<td>YRP1</td>
<td>AGGAC</td>
<td></td>
<td>TRANSFAC</td>
<td>0.145611</td>
</tr>
<tr>
<td>STE12</td>
<td>AGAAC</td>
<td></td>
<td>TRANSFAC</td>
<td>0.296750</td>
</tr>
<tr>
<td>GCN4</td>
<td>TGACT</td>
<td></td>
<td>SCPD</td>
<td>0.282221</td>
</tr>
<tr>
<td>TBP</td>
<td>TTATA</td>
<td></td>
<td>TRANSFAC</td>
<td>0.376801</td>
</tr>
<tr>
<td>HAP1</td>
<td>CGGNNNTWNCGG</td>
<td></td>
<td>YPD</td>
<td>0.423004</td>
</tr>
<tr>
<td>RAP1</td>
<td>RNNACCCA</td>
<td></td>
<td>SCPD</td>
<td>0.526059</td>
</tr>
<tr>
<td>mPAC</td>
<td>AAAATT</td>
<td></td>
<td>AlignACE</td>
<td>0.552493</td>
</tr>
<tr>
<td>mRRRE</td>
<td>AAAAAA</td>
<td></td>
<td>AlignACE</td>
<td>0.630740</td>
</tr>
<tr>
<td>PHO4</td>
<td>CGGTCA</td>
<td></td>
<td>TRANSFAC</td>
<td>0.672261</td>
</tr>
<tr>
<td>YAP1</td>
<td>GGATCA</td>
<td></td>
<td>TRANSFAC</td>
<td>0.777816</td>
</tr>
<tr>
<td>MIG1</td>
<td>CCCCCAAA</td>
<td></td>
<td>YPD</td>
<td>0.799412</td>
</tr>
<tr>
<td>MET31,32</td>
<td>AAACCTGG</td>
<td></td>
<td>YPD</td>
<td>0.84893</td>
</tr>
<tr>
<td>HAP2,3,4</td>
<td>CTGTGG</td>
<td></td>
<td>TRANSFAC</td>
<td>1.070837</td>
</tr>
</tbody>
</table>

cf. modelingsocialdata.org

arxiv.org/abs/q-bio/0701021
driving question: can we influence these?

cf. modelingssocialdata.org
optimization & learning, e.g.,

optimization & prediction, e.g.,

“newsvendor problem,” literally (+prediction+experiment)
driving question: how many copies to send?

"newsvendor problem," literally (+prediction+experiment)
Introduction

The New York Times is winning at journalism. Of all the challenges facing a media company in the digital age, producing great journalism is the hardest. Our daily report is deep, broad, smart and engaging — and we’ve got a huge lead over the competition.

At the same time, we are falling behind in a second critical area: the art and science of getting our journalism to readers. We have always cared about the reach and impact of our work, but we haven’t done enough to crack that code in the digital era.
Colin Russel 10:21 AM
!blossom facebook? all

blossombot BOT 10:21 AM ⭐
Blossom has the following suggestions for your next Facebook posts:

nytopinion: http://www.nytimes.com/2015/08/12/opinion/when-innocence-is-no-defense.html
tmagazine: Blossom currently has no suggestions
nytfood: Blossom currently has no suggestions
w/“audience development” team
leverage methods which are predictive yet performant
A decision-theoretic generalization of on-line learning and an application to boosting

Yoav Freund Robert E. Schapire

AT&T Labs
180 Park Avenue
Florham Park, NJ 07932
{yoav, schapire}@research.att.com

December 19, 1996
driving question: which content should we promote, where and when?

Colin Russel 10:21 AM
!blossom facebook? all

blossombot BOT 10:21 AM ✨
Blossom has the following suggestions for your next Facebook posts:

nytopinion: http://www.nytimes.com/2015/08/12/opinion/when-innocence-is-no-defense.html

tmagazine: Blossom currently has no suggestions


nytfood: Blossom currently has no suggestions

driving questions:

- which stories to translate?
- which stories to recommend to which people?
- which records to investigate?
- which behaviors should we try to influence?
- how many copies to send to which stores?
- which stories to promote, when + where?
driving questions:

- which stories to translate?
- which stories to recommend to which people?
- which records to investigate?
- which behaviors should we try to influence?
- how many copies to send to which stores?
- which stories to promote, when + where?

- what drug/treatment should this patient get?
- what gene should I knock out?
prescriptive modeling

descriptive: specify $x$; learn $z(x)$ or $p(z|x)$ where $z$ is “simpler” than $x$
predictive: specify $x$ and $y$; learn to predict $y$ from $x$
prescriptive: specify $x$, $y$, and $a$; learn to prescribe $a$ given $x$ to maximize $y$
prescriptive modeling

1. slow:
   - via generative modeling & importance sampling

2. fast:
   - via generative modeling & online optimization
prescriptive modeling

- “a causes y” $\iff$ $\exists$ family $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$
- define off-policy/exploration distribution
  $p_-(y, a, x) = p(y|a, x)p_-(a|x)p(x)$
- define exploitation distribution
  $p_+(y, a, x) = p(y|a, x)p_+(a|x)p(x)$
- Goal: Maximize $E_+(Y)$ over $p_+(a|x)$ using data drawn from $p_-(y, a, x)$.

E.g., set *your* hospital’s good (personalized) policy using health records from *their* hospital (w/different policy & outcomes, but same patient distribution)
“off policy value estimation”
(cf. “causal effect estimation”)
(cf. “inverse propensity weight”)

\[ \hat{\mu} = E_+(Y) = \frac{1}{N} \sum_{i=1}^{N} y_i \frac{1[a_i = h(x_i, \theta)]}{p_-(a_i|x_i)} \]

cf. Langford `08-`16;
Horvitz & Thompson `52;
Holland `86
“off policy value estimation”
(cf. “causal effect estimation”)
(cf. “inverse propensity weight”)

\[
\hat{\mu} = E_+(Y) = \frac{1}{N} \sum_{i=1}^{N} y_i \frac{1[a_i = h(x_i, \theta)]}{p_-(a_i|x_i)}
\]

Vapnik’s razor
“When solving a (learning) problem of interest, do not solve a more complex problem as an intermediate step.”
prescriptive modeling

1. slow:
   - via generative modeling & importance sampling

2. fast:
   - via generative modeling & online optimization
real-time A/B -> “bandits”

cf. modelingsocialdata.org
ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES.

By WILLIAM R. THOMPSON. From the Department of Pathology, Yale University.

616 citations, 1933
summary:
pay attention to:

1. people
2. ideas
3. things

cf. USAF
descriptive:

predictive:

prescriptive:

Explore

Learning

Test

Optimizing

Reporting
data skills

data science and...

- data engineering
- data analytics
- data embeds
- data product
- data governance
- data multiliteracies

cf. “data scientists at work”, ch 1
data skills

data science and...

- data analytics
- data engineering
- data embeds
- data product
- data multiliteracies

nota bene!

these are 3 *separate* skill sets; academia often conflates the 3. In industry these are*
- 3 related functions
- 3 collaborating teams

cf. “data scientists at work”, ch 1

* or at least @ NYT, 2016
data science: mindset & toolset

drew conway, 2010
Information Platforms and the Rise of the Data Scientist

Jeff Hammerbacher

modern history: 2009
At Facebook, we felt that traditional titles such as Business Analyst, Statistician, Engineer, and Research Scientist didn’t quite capture what we were after for our team. The workload for the role was diverse: on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization in a clear and concise fashion. To capture the skill set required to perform this multitude of tasks, we created the role of “Data Scientist.”
Mission

The Data Science Institute at Columbia University is training the next generation of data scientists and developing innovative technology to serve society. With more than 150 faculty working in a wide range of disciplines, the Institute seeks to foster collaboration in advancing techniques to gather and interpret data, and to address the urgent problems facing society. The Institute works closely with industry to bring promising ideas to market.
The current Data Science Institute course schedule may be viewed [here](#).

The following is a list of data science-related courses. Please refer to the Directory of Courses for the most current course offerings and information.

**STATISTICS & COMPUTER SCIENCE**

**STCS W4242 (formerly STAT W4242)**

**Introduction to Data Science**

Professor Anasf Salleb-Aouissi ([Syllabus](#))

Data Science is a dynamic and fast-growing field at the interface of Statistics and Computer Science. The emergence of massive datasets containing millions or even billions of observations provides the primary impetus for the field. Such datasets arise, for instance, in large-scale retailing, telecommunications, astronomy, and internet social media. This course will emphasize practical techniques for working with large-scale data. Specific topics covered will include statistical modeling and machine learning, data pipelines, programming languages, "big data" tools, and real world topics and case studies. The use of statistical and data manipulation software will be required. Course intended for non-quantitative graduate-level disciplines. **This course will not count towards degree requirements for graduate programs such as Statistics, Computer Science, or Data Science.** Students should
Data Science Institute

Certification of Professional Achievement in Data Sciences

The Certification of Professional Achievement in Data Sciences prepares students to expand their career prospects or change career paths by developing foundational data science skills.

ELIGIBILITY REQUIREMENTS

- Undergraduate degree
- Prior quantitative coursework (calculus, linear algebra, etc...)
- Prior introductory to computer programming coursework

APPLICATION REQUIREMENTS

- Online application
- Uploaded transcripts from every post-secondary institution attended
- Three recommendation letters
- Personal statement
- Curriculum vitae / résumé
Data Science Free Online Courses (edX)

Data science is making us smarter and more innovative in so many ways. How does it all work?

In this Data Science and Analytics XSeries you will gain insight into the latest data science tools and their application in finance, health care, product development, sales and more. With real world examples, we will demonstrate how data science can improve corporate decision-making and performance, personalized medicine and advance your career goals.

Taught by a distinguished team of professors at Columbia University’s Data Science Institute, this XSeries is perfect for anyone who wants to understand basic concepts in data science without getting into the weeds of programming. Aimed at organization leaders, business managers, health care professionals and anyone considering a career in data science, this series will steep learners in the fundamentals of statistics, machine learning and algorithms. It will also introduce emerging technologies such as the Internet of Things, or wirelessly connected products, and techniques that allow computers to summarize mountains of text, audio and video. Concrete examples provided throughout the series will ensure that
Data Science Institute

Master of Science in Data Science

The Master of Science in Data Science allows students to apply data science techniques to their field of interest, building on four foundational courses offered in our Certification of Professional Achievement in Data Sciences program. Our students have the opportunity to conduct original research, included in a capstone project, and interact with our industry partners and faculty. Students may also choose an elective track focused on entrepreneurship or a subject area covered by one of our six centers.

ELIGIBILITY REQUIREMENTS

- Undergraduate degree
- Prior quantitative coursework (calculus, linear algebra, etc...)
- Prior introductory to computer programming coursework

WHO SHOULD APPLY?
Data Science Institute

Data Sciences & Entrepreneurship

Encouraging entrepreneurship and developing an entrepreneurial ecosystem for the Institute's faculty and staff who are interested in starting companies is an important component of the Institute's mission and has emerged as a central educational theme within Columbia Engineering. Columbia Engineering promotes engineering innovation and engaged entrepreneurship. Its entrepreneurship programs provide education and support for Columbia Engineering students and faculty, socially engaged
Explorations in Data Science

Columbia University Science Honors Program

Spring 2016

About this course

In this course, our students (aka, the explorers) will carry out a series of explorations in data science to learn about statistical thinking principles and data analysis skills used in data science.

These explorations will cover topics including but not limited to

- descriptive statistics,
- sampling and estimation,
- association,
- regression analysis, etc.

high school!
complementary/experiential education?
Hackathons are proof that it's all going to be OK:

see also: http://bit.ly/hackNY15vid