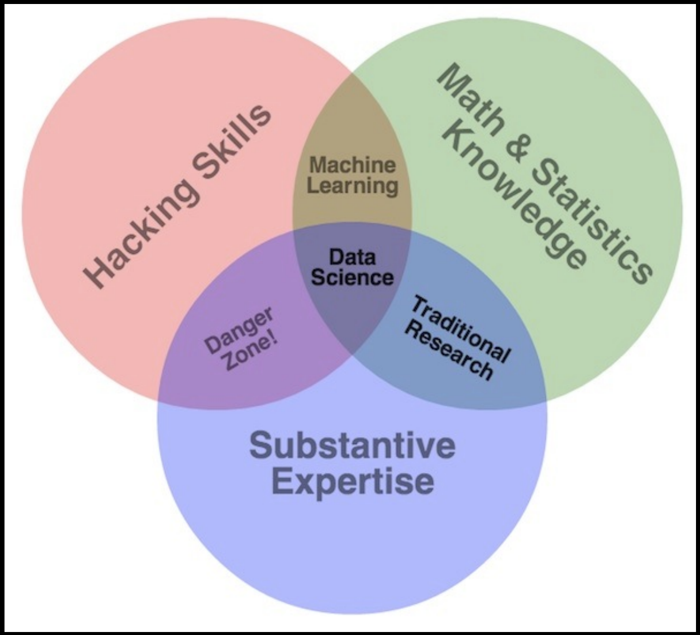
data science

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

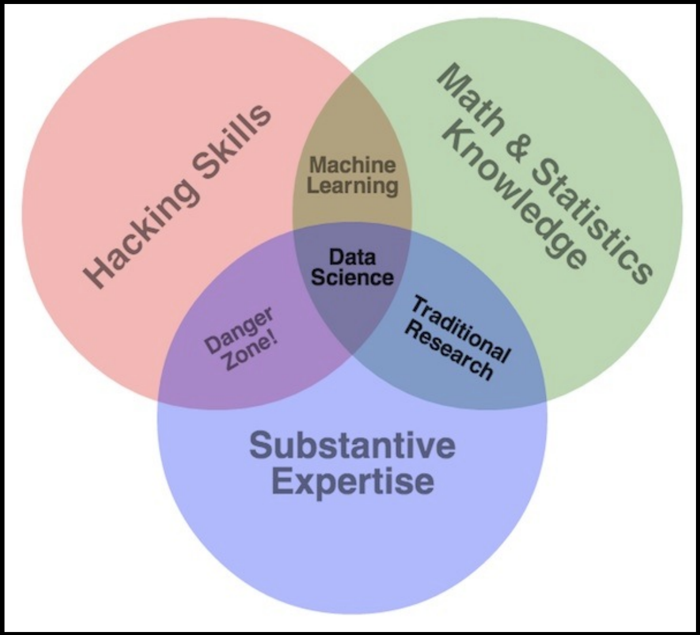
data science: 2 definitions

applied machine learning



drew conway, 2010

applied machine learning



drew conway, 2010

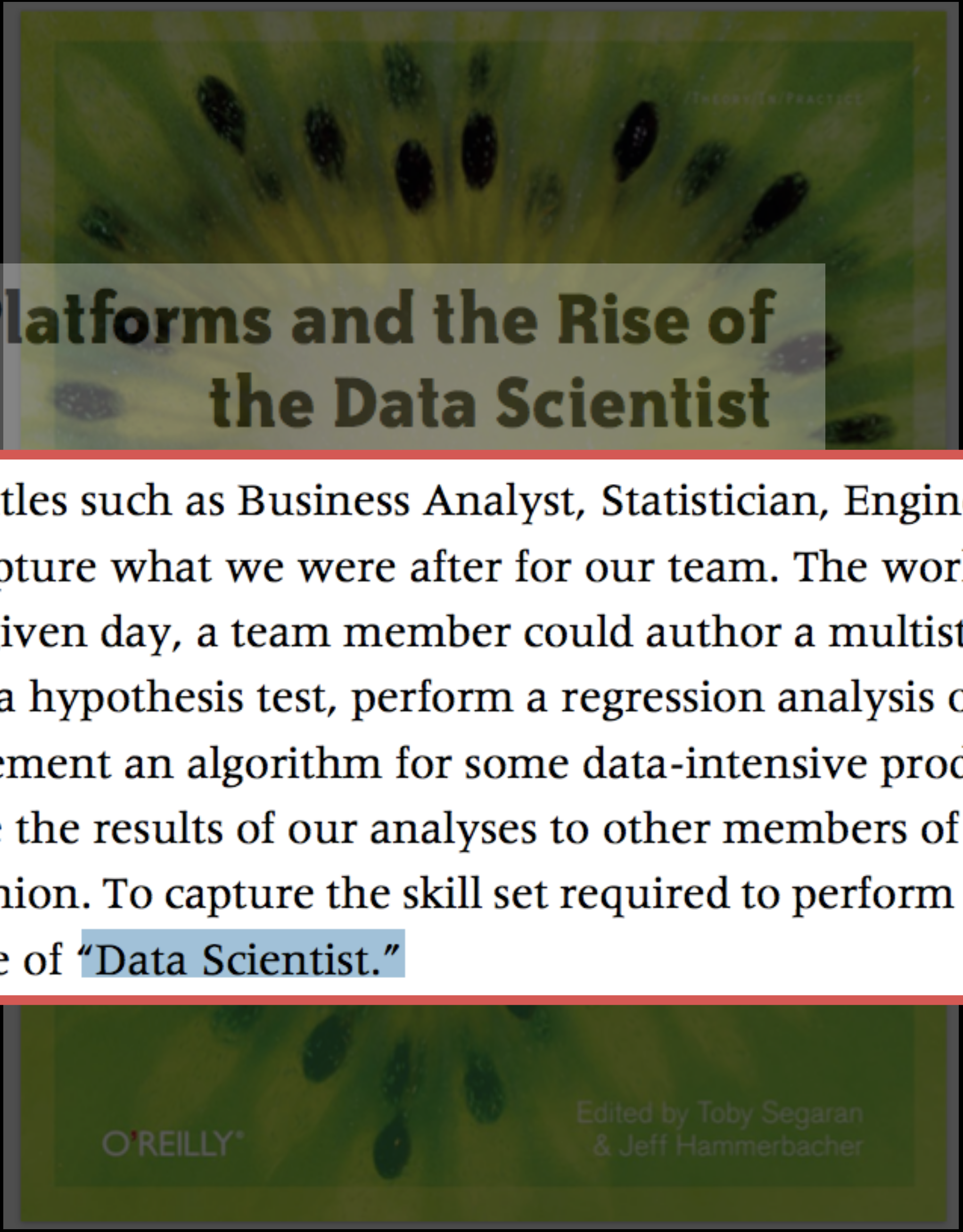
(closer to DS in academia)

Information Platforms and the Rise of the Data Scientist

Jeff Hammerbacher

modern history:
2009

Information Platforms and the Rise of the Data Scientist



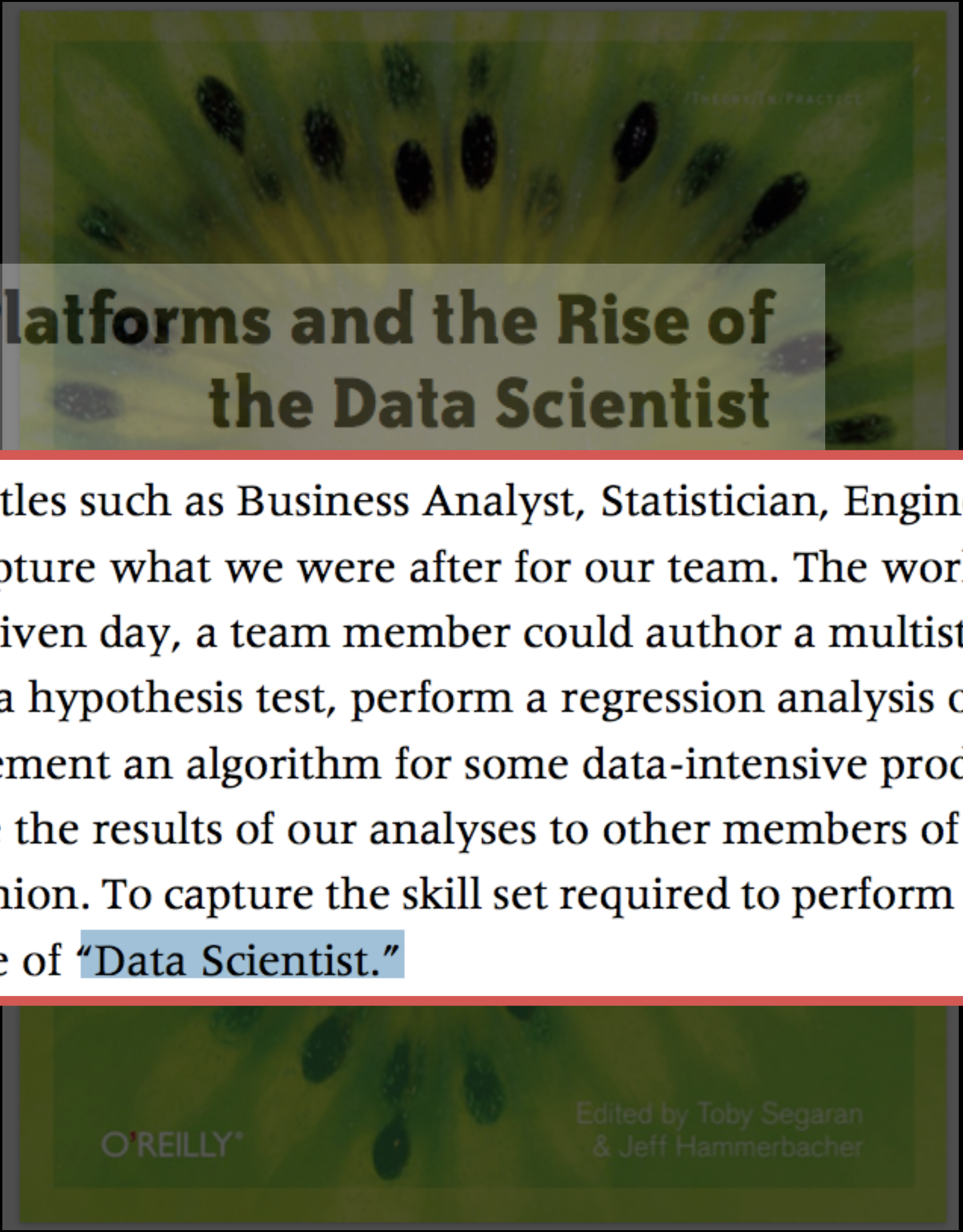
At Facebook, we felt that traditional titles such as Business Analyst, Statistician, Engineer, and Research Scientist didn't quite capture what we were after for our team. The workload for the role was diverse: on any given day, a team member could author a multistage processing pipeline in Python, design 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."

2009

Information Platforms and the Rise of the Data Scientist

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."

2009



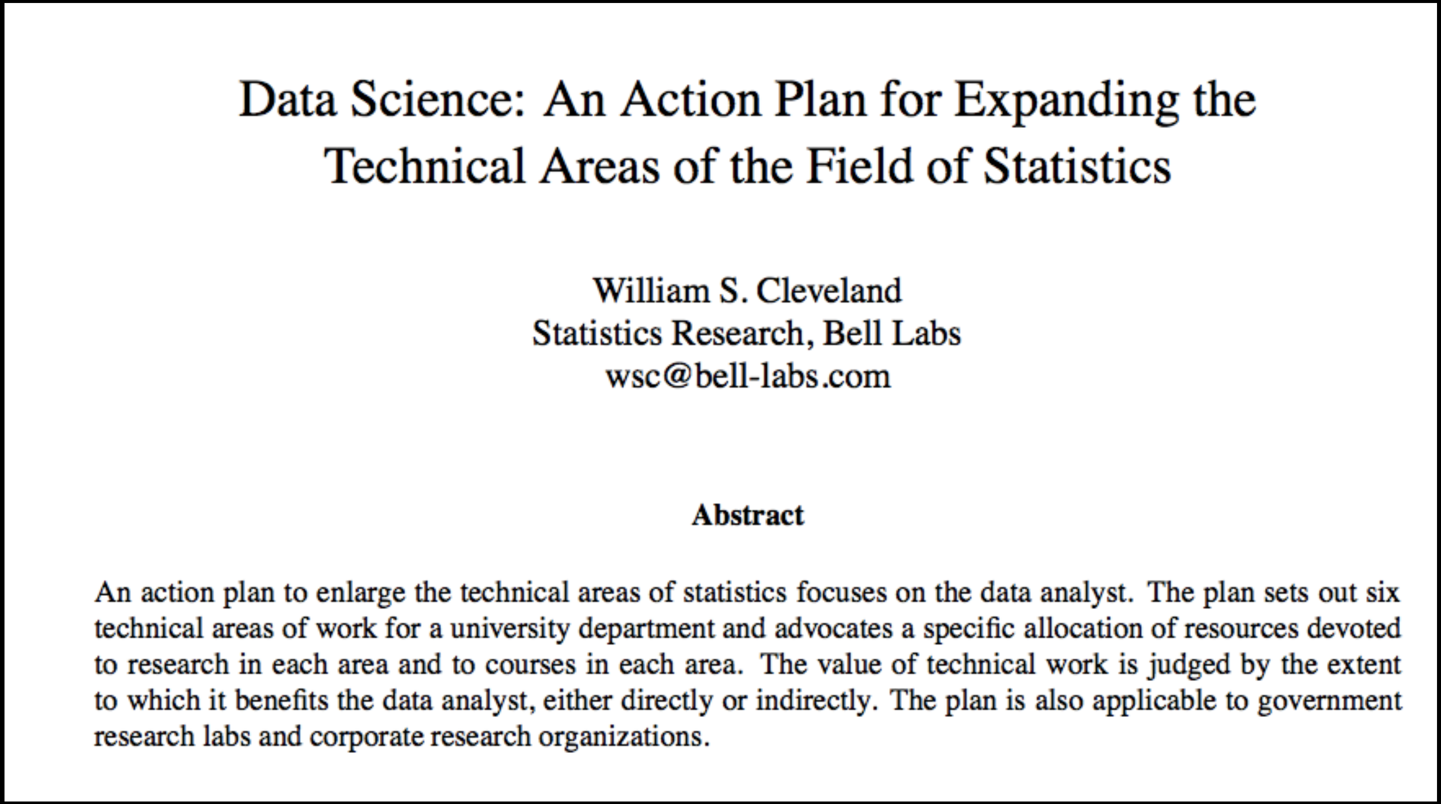
(closer to DS in industry)

data science: pre-2009 history

GOOG: “igert in data science”
http://www.stat.ucla.edu/~cocteau/ds3.pdf (2007)

“data science”

ancient history:



Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland
Statistics Research, Bell Labs
wsc@bell-labs.com

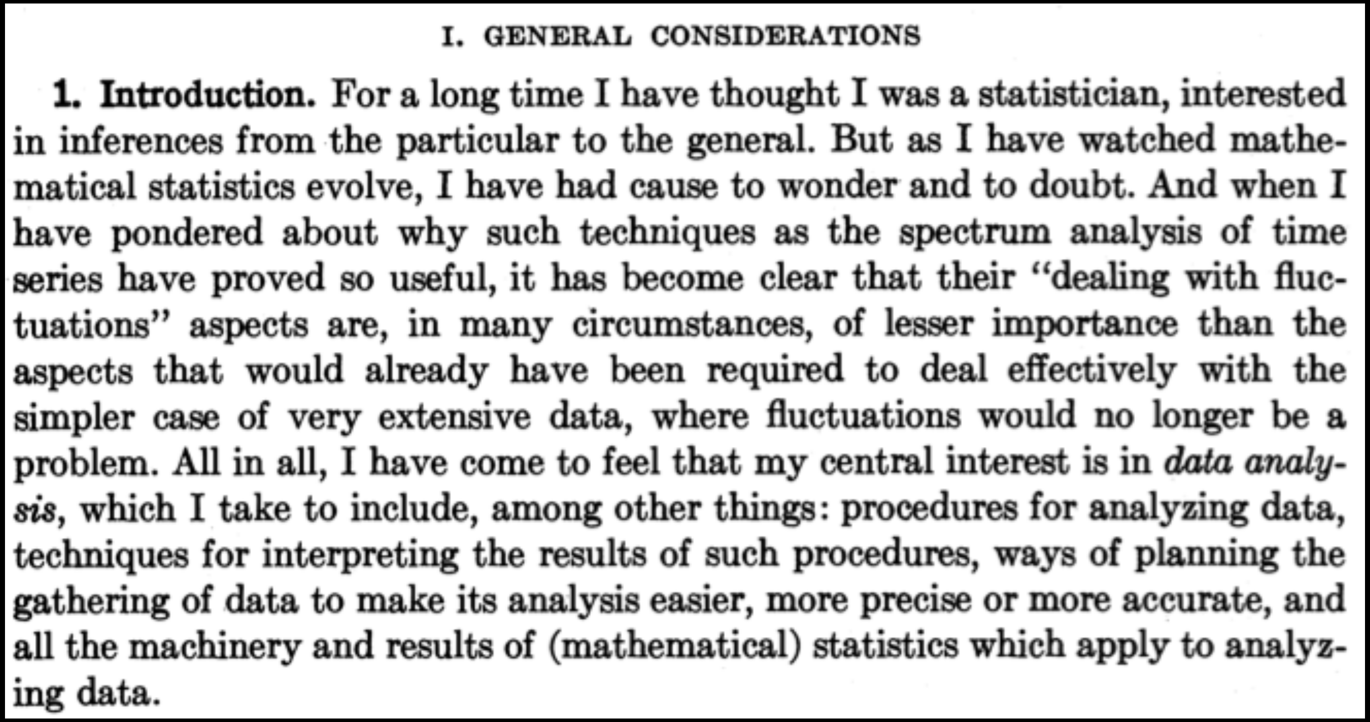
Abstract

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

“data science”
ancient history: 2001

data science
context

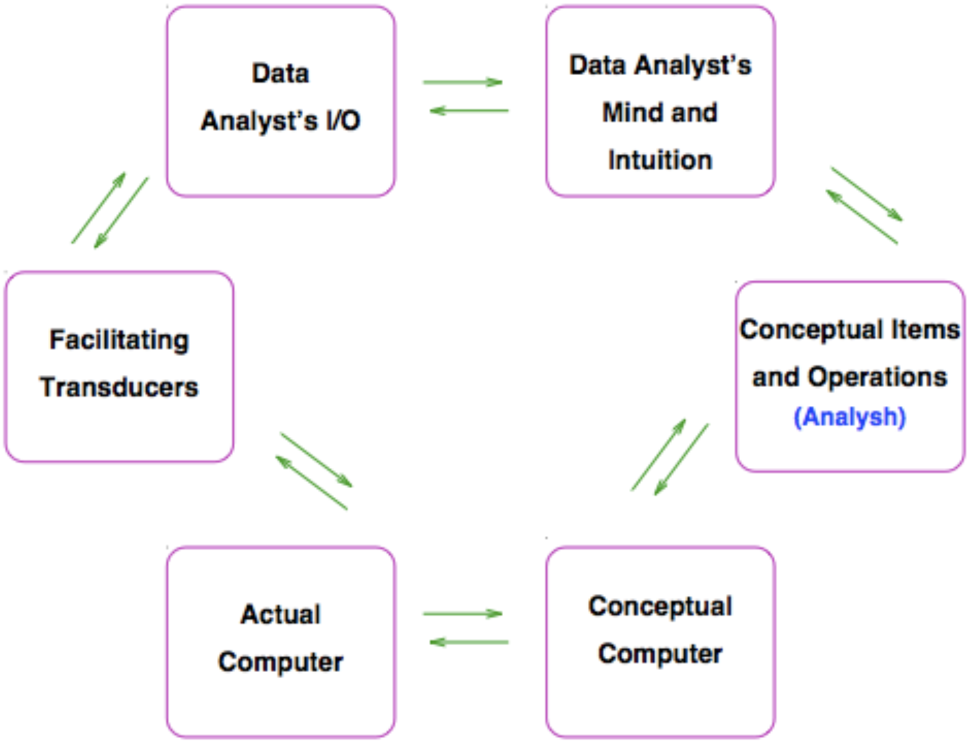
“the progenitor of data science.” - @mshron



I. GENERAL CONSIDERATIONS

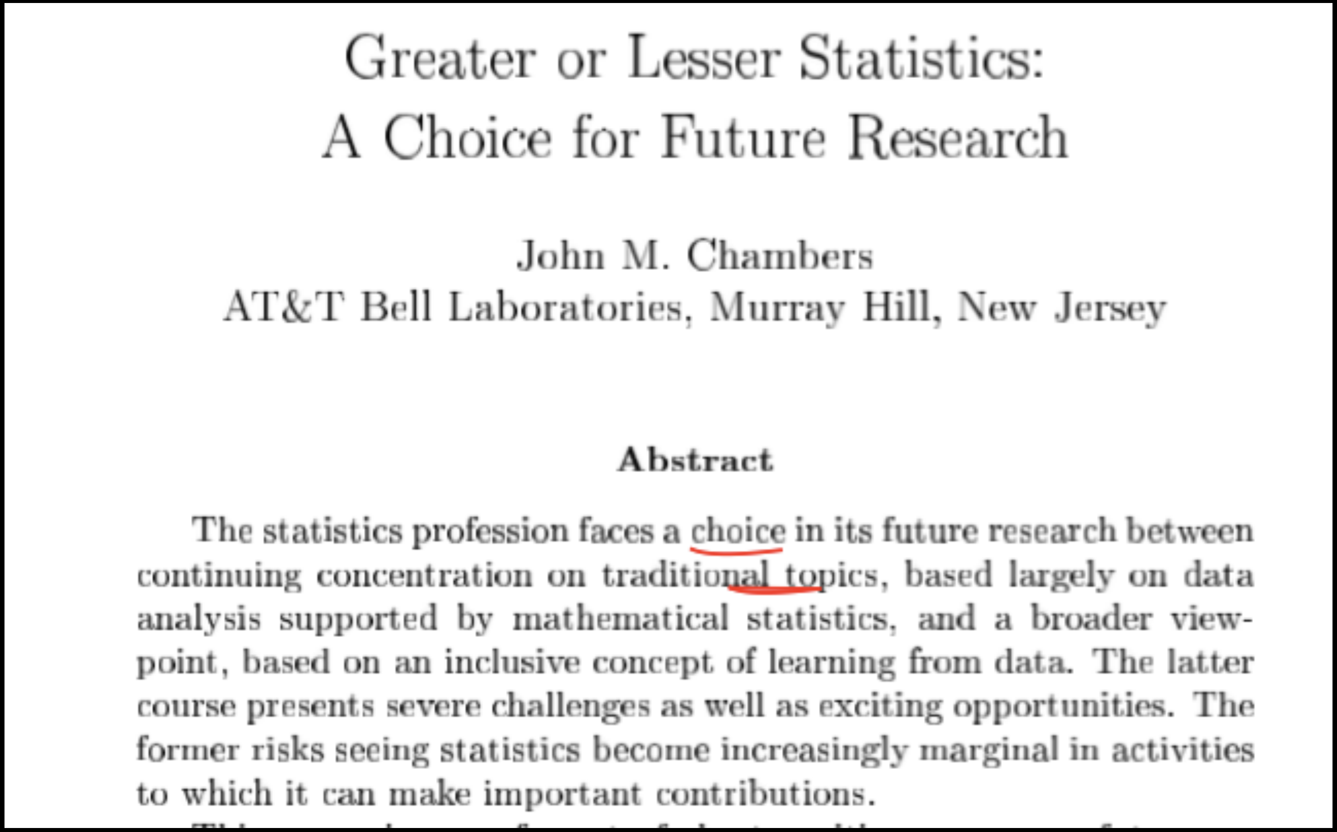
1. Introduction. For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. And when I have pondered about why such techniques as the spectrum analysis of time series have proved so useful, it has become clear that their "dealing with fluctuations" aspects are, in many circumstances, of lesser importance than the aspects that would already have been required to deal effectively with the simpler case of very extensive data, where fluctuations would no longer be a problem. All in all, I have come to feel that my central interest is in *data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

"The Future of Data Analysis," 1962
John W. Tukey



John W. Tukey (Feb. 3, 1965)

Tukey 1965, via John Chambers



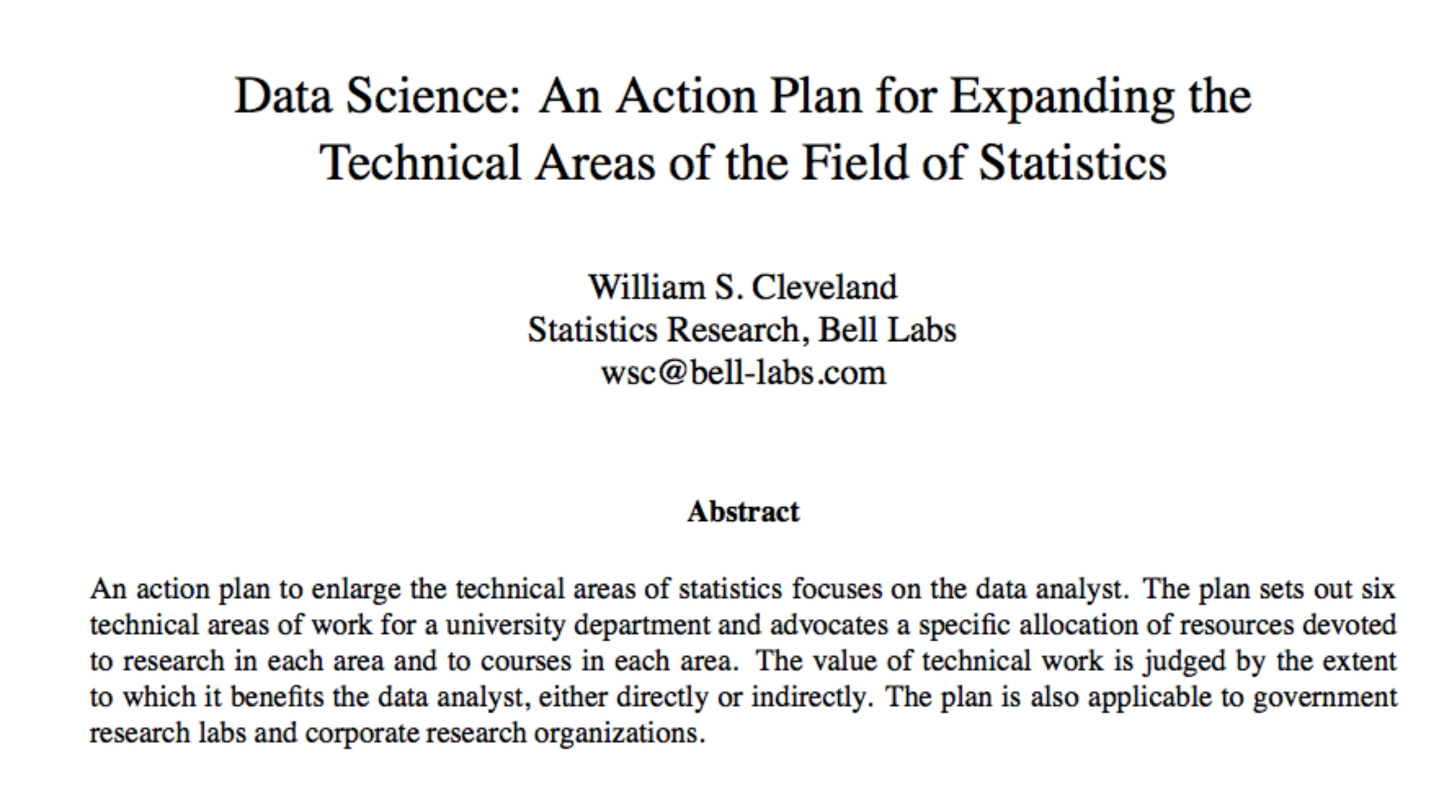
Greater or Lesser Statistics: A Choice for Future Research

John M. Chambers
AT&T Bell Laboratories, Murray Hill, New Jersey

Abstract

The statistics profession faces a choice in its future research between continuing concentration on traditional topics, based largely on data analysis supported by mathematical statistics, and a broader viewpoint, based on an inclusive concept of learning from data. The latter course presents severe challenges as well as exciting opportunities. The former risks seeing statistics become increasingly marginal in activities to which it can make important contributions.

fast forward -> 1993



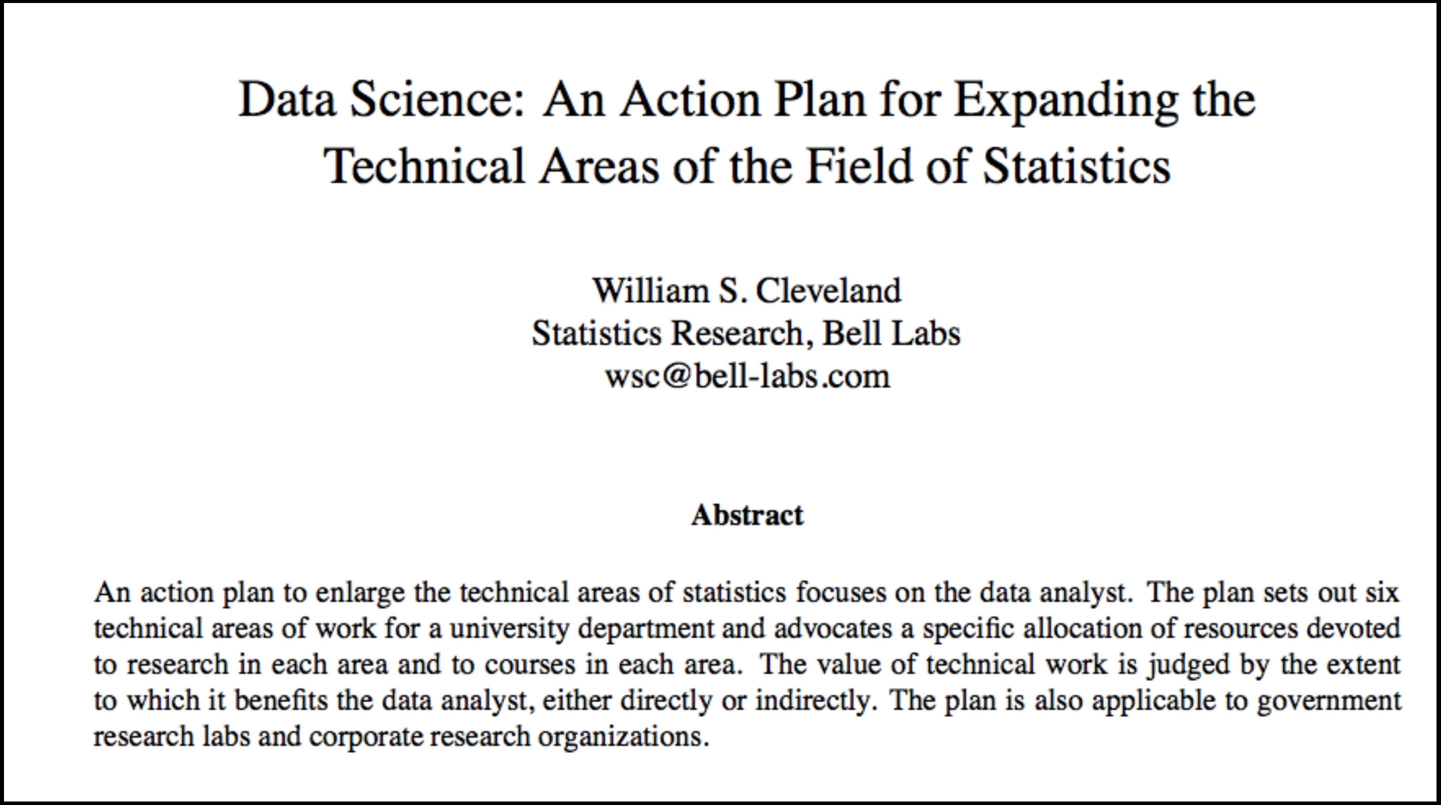
Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland
Statistics Research, Bell Labs
wsc@bell-labs.com

Abstract

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

fast forward -> 2001



Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

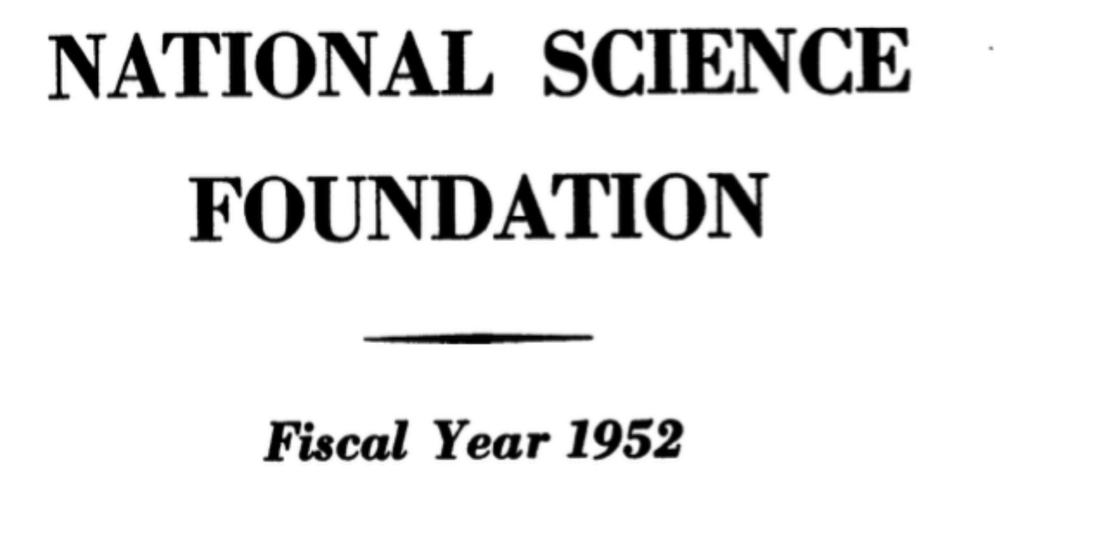
William S. Cleveland
Statistics Research, Bell Labs
wsc@bell-labs.com

Abstract

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

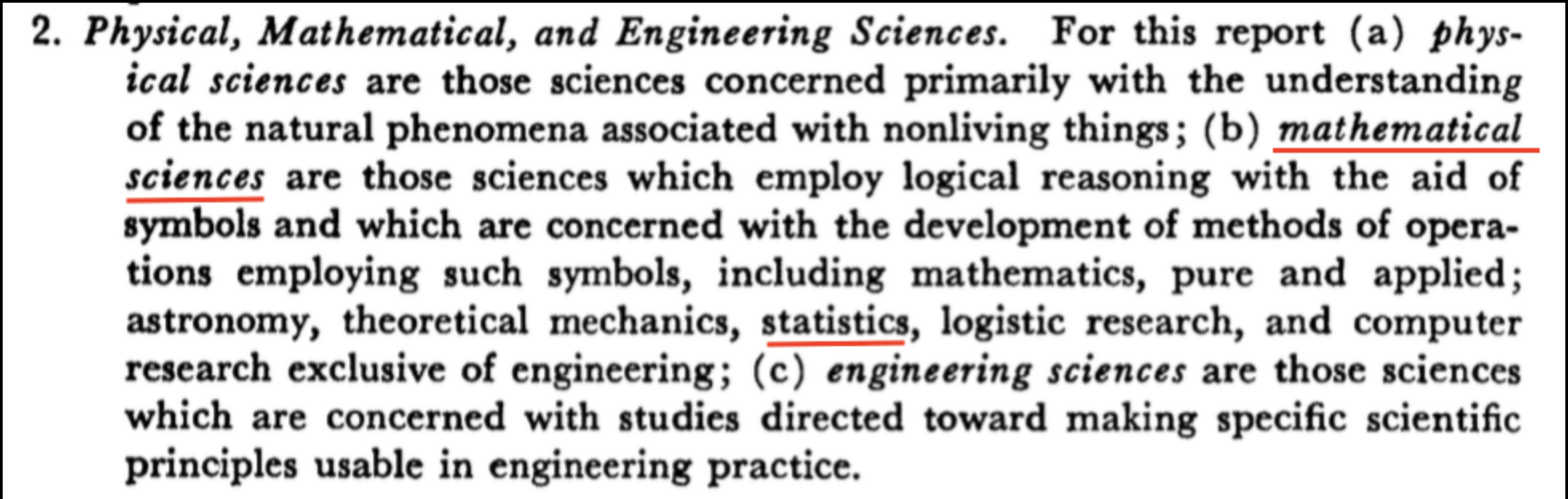
“The primary agents for change should be university departments themselves.”

NB: placement of stats \in math



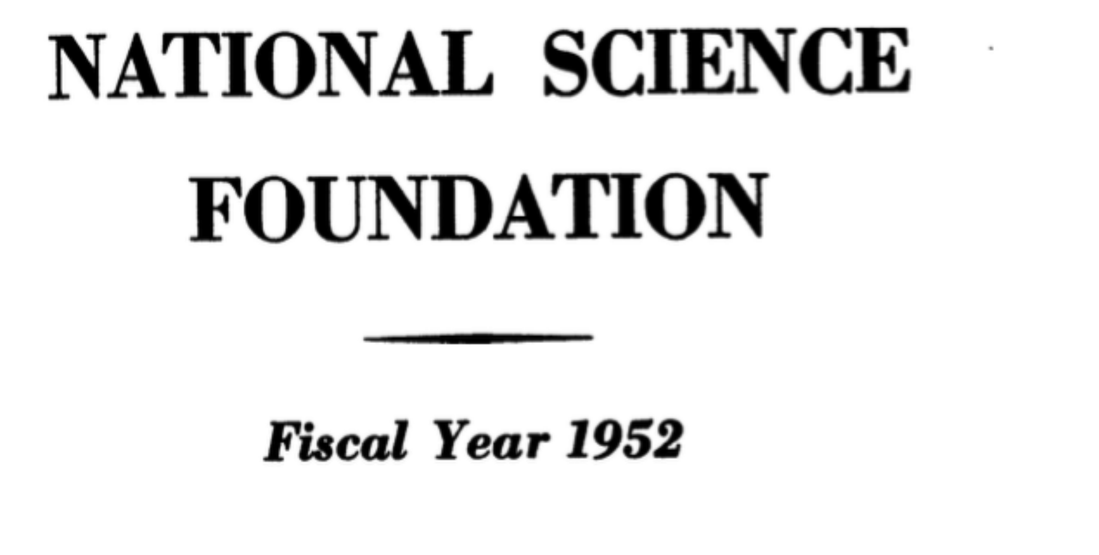
NATIONAL SCIENCE FOUNDATION

Fiscal Year 1952



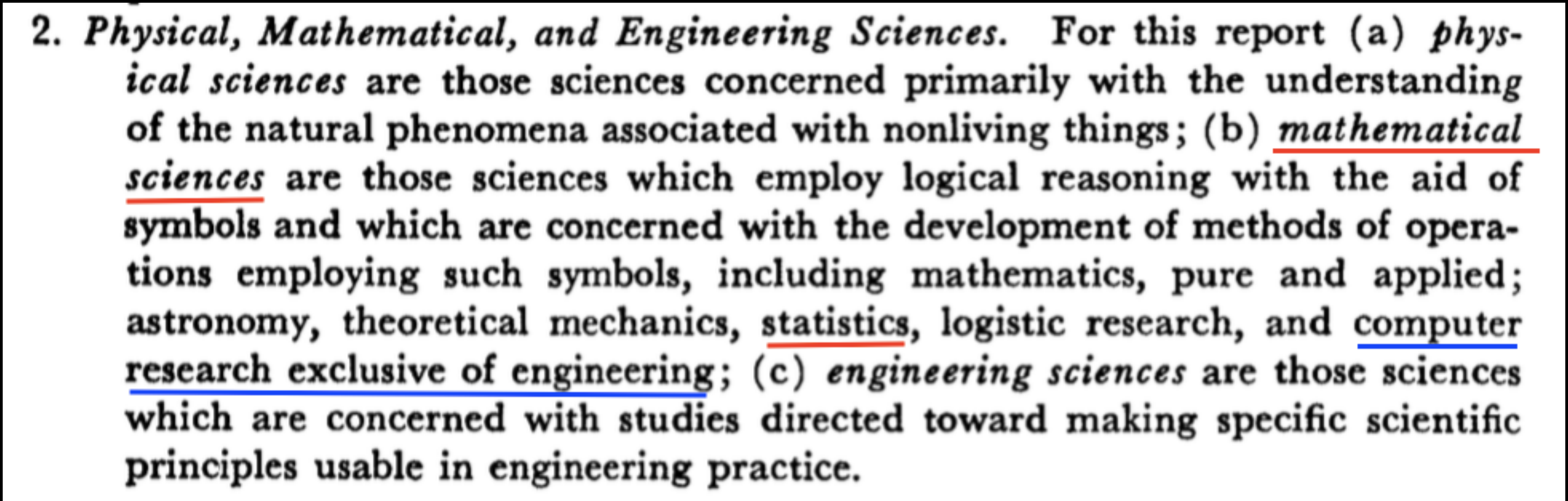
2. *Physical, Mathematical, and Engineering Sciences.* For this report (a) *physical sciences* are those sciences concerned primarily with the understanding of the natural phenomena associated with nonliving things; (b) *mathematical sciences* are those sciences which employ logical reasoning with the aid of symbols and which are concerned with the development of methods of operations employing such symbols, including mathematics, pure and applied; astronomy, theoretical mechanics, statistics, logistic research, and computer research exclusive of engineering; (c) *engineering sciences* are those sciences which are concerned with studies directed toward making specific scientific principles usable in engineering practice.

NB: placement of stats \in math

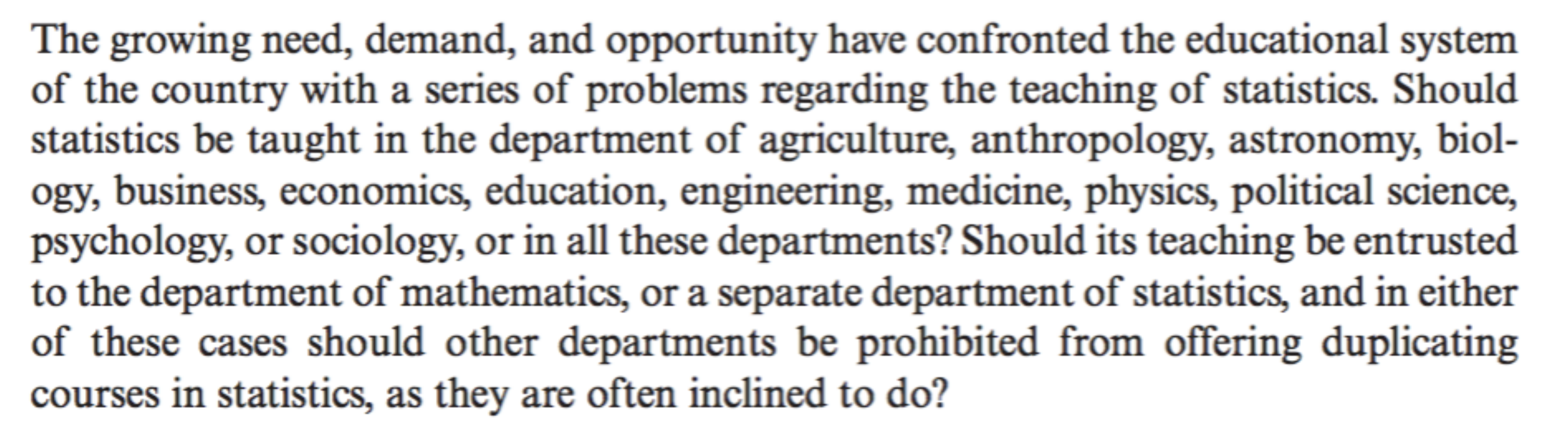


NATIONAL SCIENCE FOUNDATION

Fiscal Year 1952

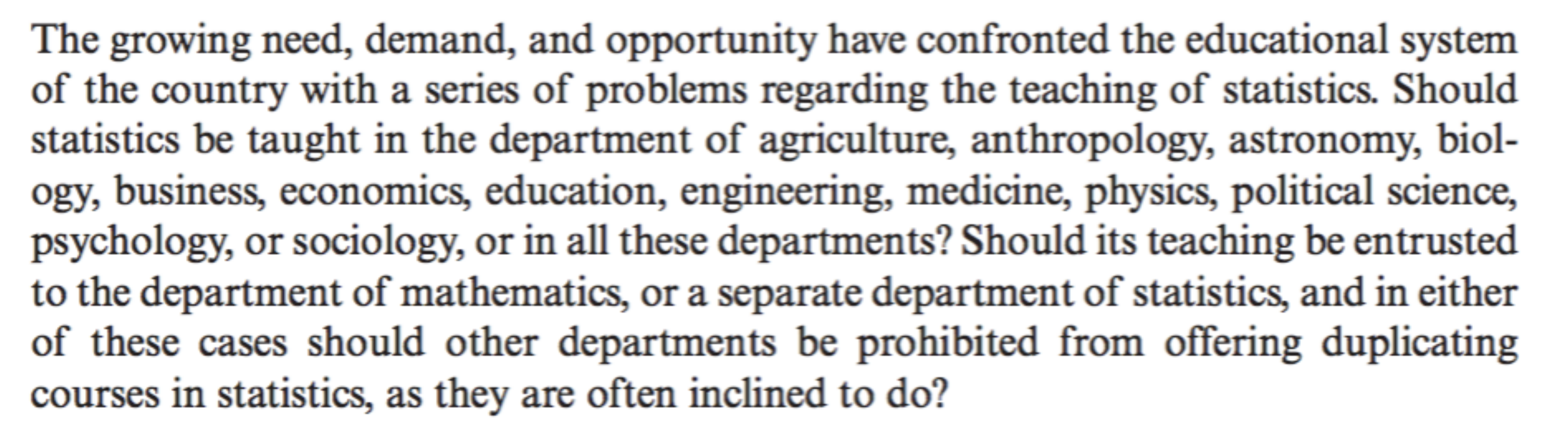


2. *Physical, Mathematical, and Engineering Sciences.* For this report (a) *physical sciences* are those sciences concerned primarily with the understanding of the natural phenomena associated with nonliving things; (b) *mathematical sciences* are those sciences which employ logical reasoning with the aid of symbols and which are concerned with the development of methods of operations employing such symbols, including mathematics, pure and applied; astronomy, theoretical mechanics, statistics, logistic research, and computer research exclusive of engineering; (c) *engineering sciences* are those sciences which are concerned with studies directed toward making specific scientific principles usable in engineering practice.



The growing need, demand, and opportunity have confronted the educational system of the country with a series of problems regarding the teaching of statistics. Should statistics be taught in the department of agriculture, anthropology, astronomy, biology, business, economics, education, engineering, medicine, physics, political science, psychology, or sociology, or in all these departments? Should its teaching be entrusted to the department of mathematics, or a separate department of statistics, and in either of these cases should other departments be prohibited from offering duplicating courses in statistics, as they are often inclined to do?

not always obvious, cf. Hoteling 1945



The growing need, demand, and opportunity have confronted the educational system of the country with a series of problems regarding the teaching of statistics. Should statistics be taught in the department of agriculture, anthropology, astronomy, biology, business, economics, education, engineering, medicine, physics, political science, psychology, or sociology, or in all these departments? Should its teaching be entrusted to the department of mathematics, or a separate department of statistics, and in either of these cases should other departments be prohibited from offering duplicating courses in statistics, as they are often inclined to do?

data science: 2 definitions

1. in academia: machine learning applied, science/research focus

2. in industry: machine learning applied, product/software focus

data science history: 2 epochs

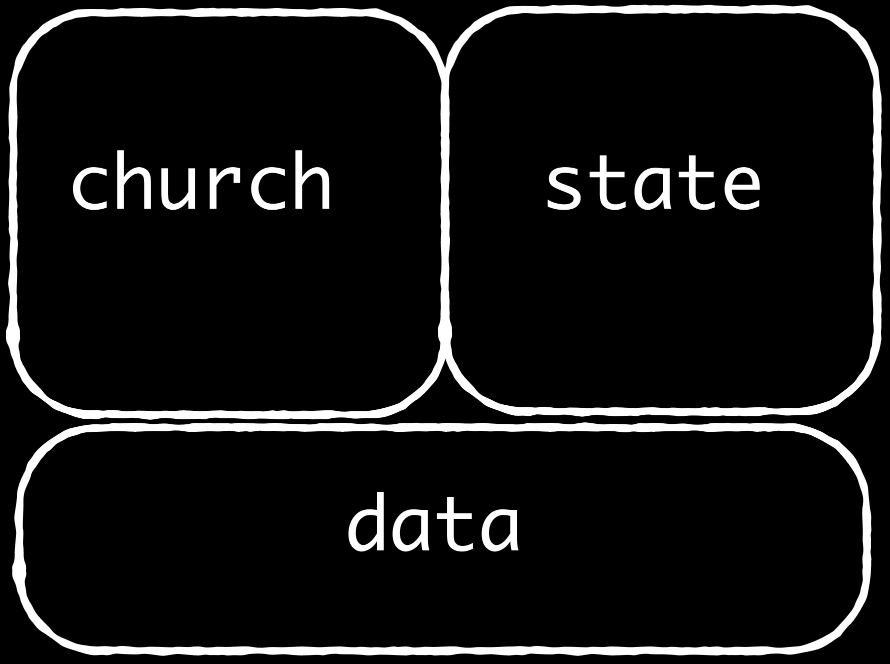
1. slow burn @Bell: as heretical statistics (see also Breiman)

2. caught fire 2009-now: as job description

data science @ The New York Times

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

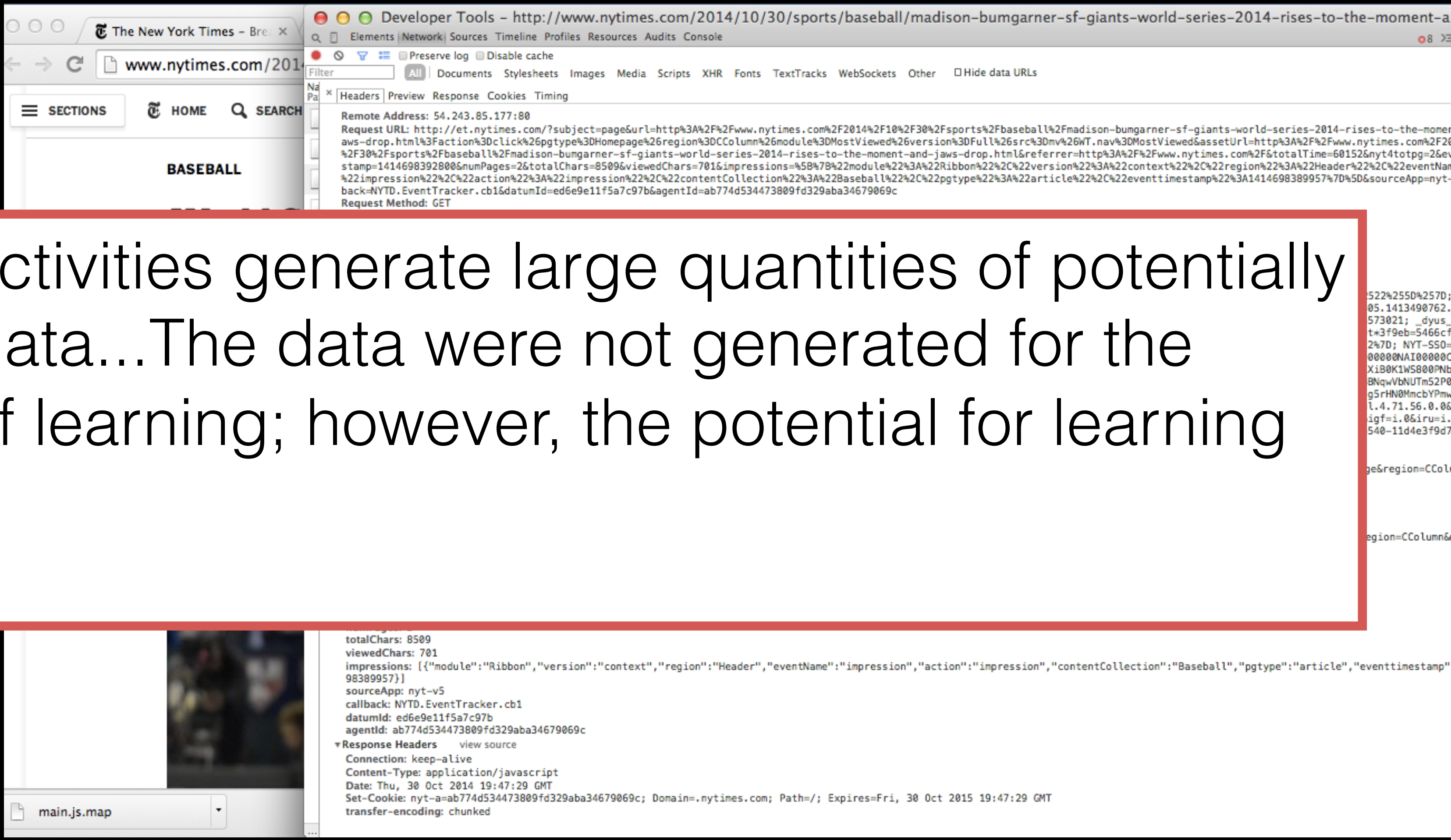
news: 21st century



church

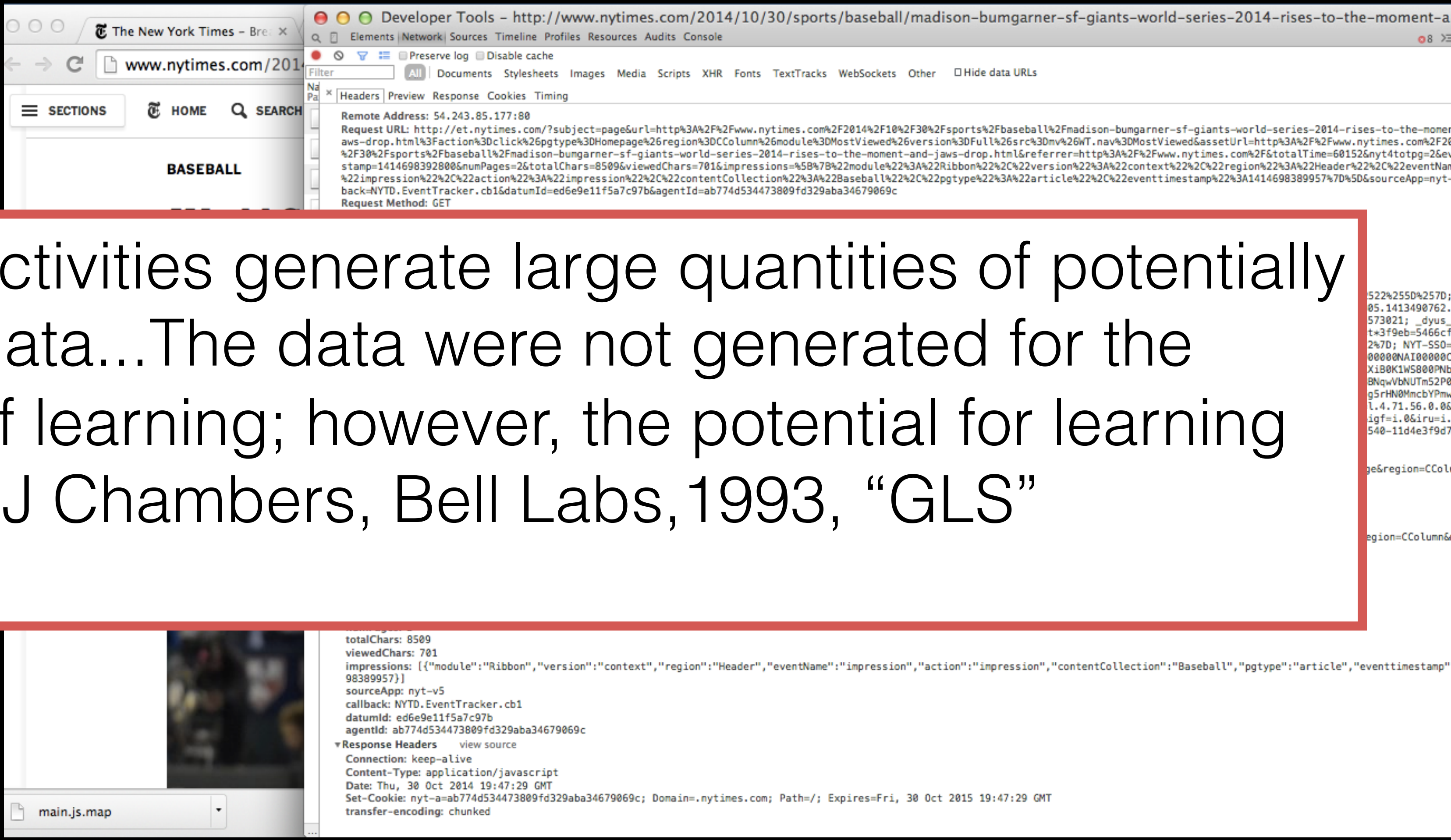
state

data



```
Remote Address: 54.243.85.177:80
Request URL: http://et.nytimes.com/?subject=page&url=http%3A%2F%2Fwww.nytimes.com%2F2014%2F10%2F30%2Fsports%2Fbaseball%2Fmadison-bumgarner-sf-giants-world-series-2014-rises-to-the-momen
aws-drop.html%3Faction%3Dclick%26pgtype%3DHomepage%26region%3DCColumn%26module%3DMostViewed%26version%3DFull%26src%3Dmv%26WT.nav%3DMostViewed&assetUrl=http%3A%2F%2Fwww.nytimes.com%2F20
%2F30%2Fsports%2Fbaseball%2Fmadison-bumgarner-sf-giants-world-series-2014-rises-to-the-moment-and-jaws-drop.html&referrer=http%3A%2F%2Fwww.nytimes.com%2F&totalTime=60152&nyt4totpg=2&ev
stamp=1414698392800&numPages=2&totalChars=8509&viewedChars=701&impressions=%5B%7B%22module%22%3A%22Ribbon%22%2C%22version%22%3A%22context%22%2C%22region%22%3A%22Header%22%2C%22eventNam
%22impression%22%2C%22action%22%3A%22impression%22%2C%22contentCollection%22%3A%22Baseball%22%2C%22pgtype%22%3A%22article%22%2C%22eventtimestamp%22%3A1414698389957%7D%5D&sourceApp=nyt-
back=NYTD.EventTracker.cb1&datumId=ed6e9e11f5a7c97b&agentId=ab774d534473809fd329aba34679069c
Request Method: GET

522%255D%257D;
05.1413490762.
573021; _dyus_
t*3f9eb=5466cf
2%7D; NYT-SSO=
00000NAI00000C
XiB0K1WS800PNb
BNqwVbNUTm52P0
g5rHN0MmcbYPmw
l.4.71.56.0.0&
igf=i.0&iru=i.
540-11d4e3f9d7

ge&region=CCol

egion=CColumn&

totalChars: 8509
viewedChars: 701
impressions: [{"module":"Ribbon","version":"context","region":"Header","eventName":"impression","action":"impression","contentCollection":"Baseball","pgtype":"article","eventtimestamp"
98389957}]
sourceApp: nyt-v5
callback: NYTD.EventTracker.cb1
datumId: ed6e9e11f5a7c97b
agentId: ab774d534473809fd329aba34679069c
▾Response Headers     view source
Connection: keep-alive
Content-Type: application/javascript
Date: Thu, 30 Oct 2014 19:47:29 GMT
Set-Cookie: nyt-a=ab774d534473809fd329aba34679069c; Domain=.nytimes.com; Path=/; Expires=Fri, 30 Oct 2015 19:47:29 GMT
transfer-encoding: chunked
```

"...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"

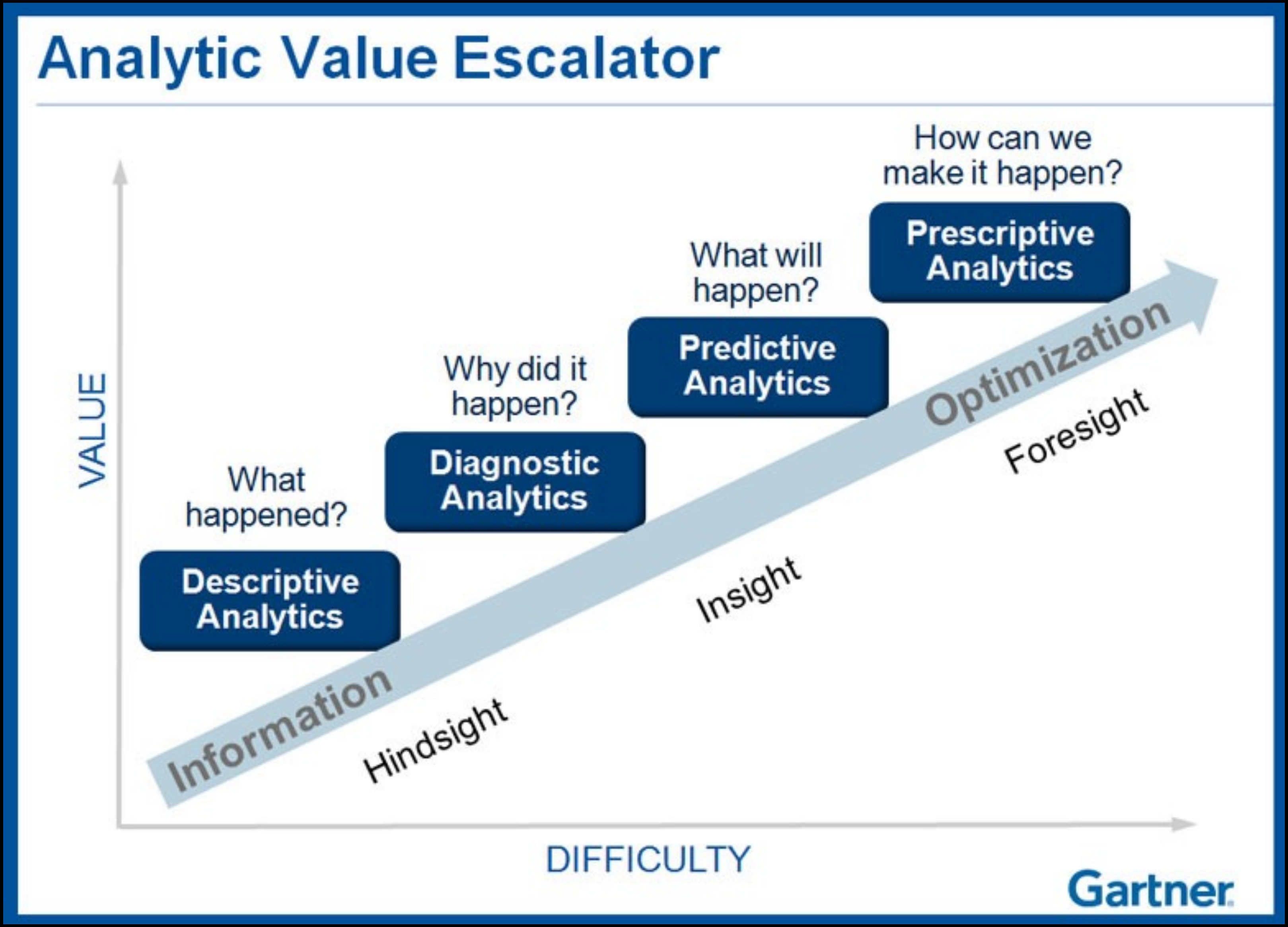


```
Remote Address: 54.243.85.177:80
Request URL: http://et.nytimes.com/?subject=page&url=http%3A%2F%2Fwww.nytimes.com%2F2014%2F10%2F30%2Fsports%2Fbaseball%2Fmadison-bumgarner-sf-giants-world-series-2014-rises-to-the-momer
aws-drop.html%3Faction%3Dclick%26pgtype%3DHomepage%26region%3DCColumn%26module%3DMostViewed%26version%3DFull%26src%3Dmv%26WT.nav%3DMostViewed&assetUrl=http%3A%2F%2Fwww.nytimes.com%2F20
%2F30%2Fsports%2Fbaseball%2Fmadison-bumgarner-sf-giants-world-series-2014-rises-to-the-moment-and-jaws-drop.html&referrer=http%3A%2F%2Fwww.nytimes.com%2F&totalTime=60152&nyt4totpg=2&ev
stamp=1414698392800&numPages=2&totalChars=8509&viewedChars=701&impressions=%5B%7B%22module%22%3A%22Ribbon%22%2C%22version%22%3A%22context%22%2C%22region%22%3A%22Header%22%2C%22eventNam
%22impression%22%2C%22action%22%3A%22impression%22%2C%22contentCollection%22%3A%22Baseball%22%2C%22pgtype%22%3A%22article%22%2C%22eventtimestamp%22%3A1414698389957%7D%5D&sourceApp=nyt-
back=NYTD.EventTracker.cb1&datumId=ed6e9e11f5a7c97b&agentId=ab774d534473809fd329aba34679069c
Request Method: GET
```

"...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”

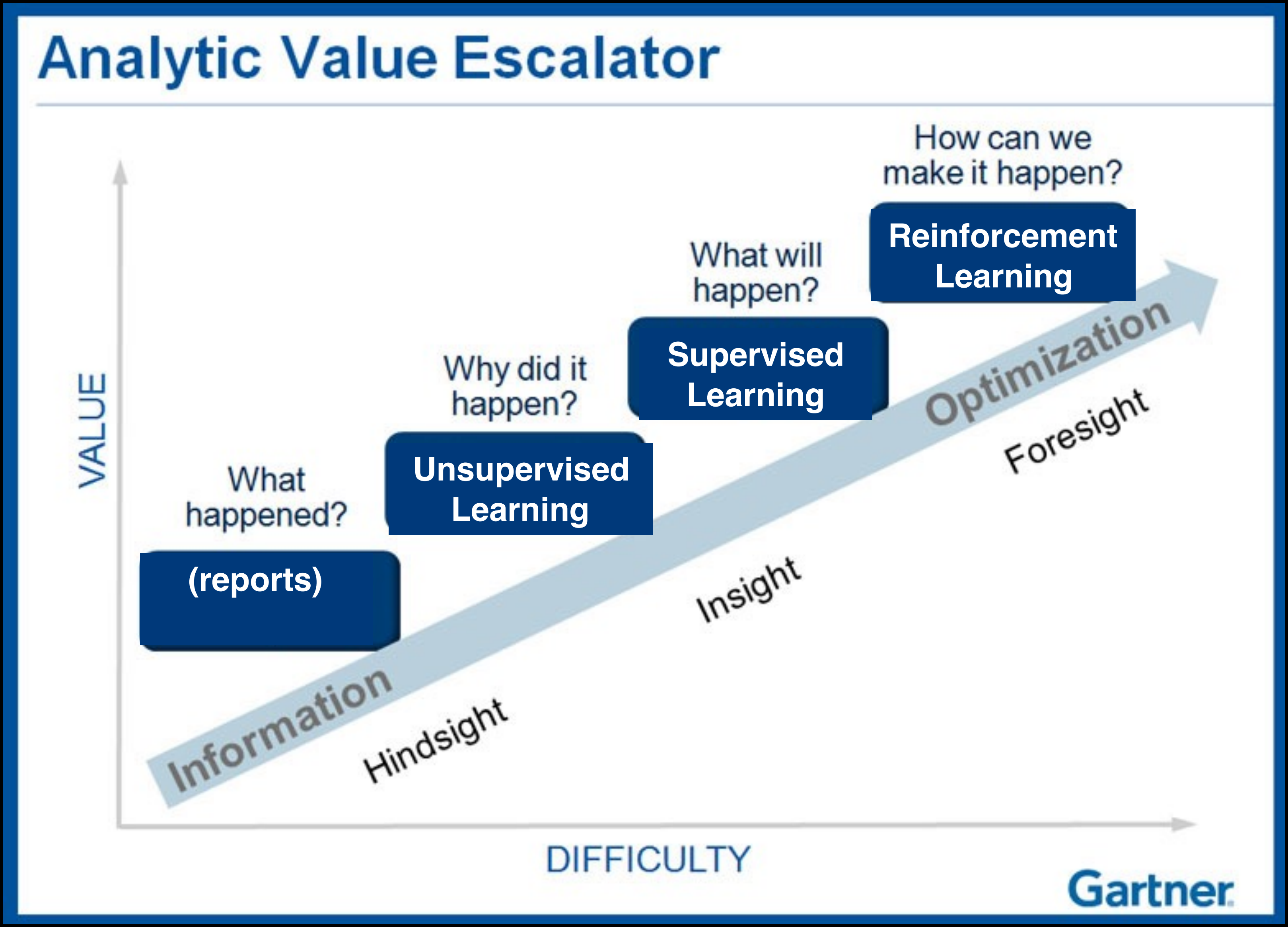
```
522%255D%257D;
05.1413490762.
573021; _dyus_
t*3f9eb=5466cf
2%7D; NYT-SSO=
00000NAI00000C
XiB0K1WS800PNb
BNqwVbNUTm52P0
g5rHN0MmcbYPmw
l.4.71.56.0.0&
igf=i.0&iru=i.
540-11d4e3f9d7

ge&region=CCol

egion=CColumn&
```

```
totalChars: 8509
viewedChars: 701
impressions: [{"module":"Ribbon","version":"context","region":"Header","eventName":"impression","action":"impression","contentCollection":"Baseball","pgtype":"article","eventtimestamp"
98389957}]
sourceApp: nyt-v5
callback: NYTD.EventTracker.cb1
datumId: ed6e9e11f5a7c97b
agentId: ab774d534473809fd329aba34679069c
▾Response Headers    view source
Connection: keep-alive
Content-Type: application/javascript
Date: Thu, 30 Oct 2014 19:47:29 GMT
Set-Cookie: nyt-a=ab774d534473809fd329aba34679069c; Domain=.nytimes.com; Path=/; Expires=Fri, 30 Oct 2015 19:47:29 GMT
transfer-encoding: chunked
```



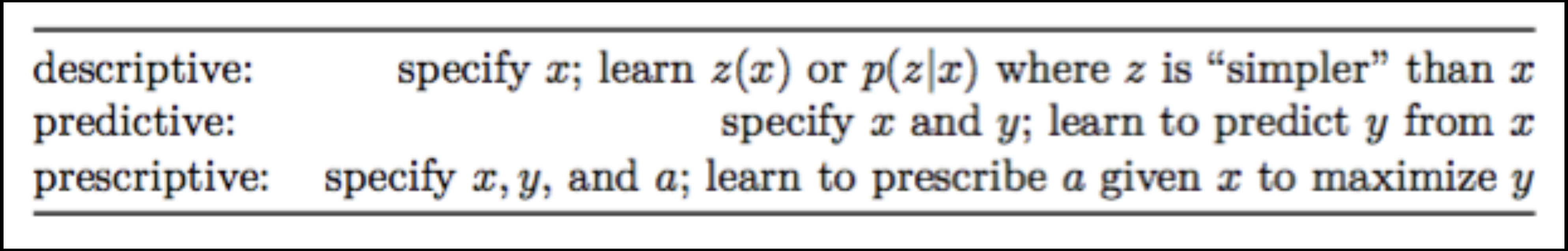
Analytic Value Escalator

h/t michael littman



h/t michael littman

descriptive, predictive, prescriptive modeling

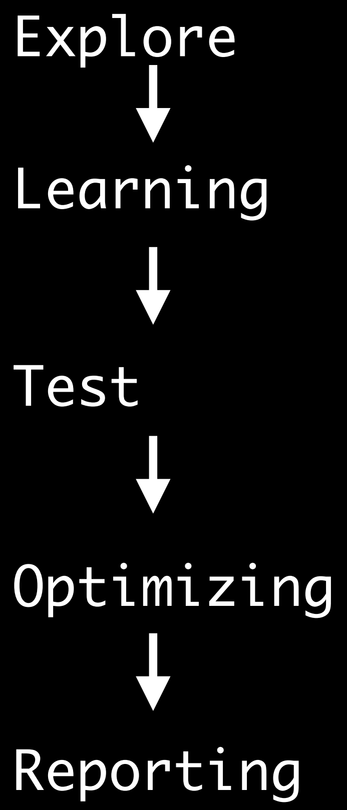


descriptive:	specify x; learn $z(x)$ or $p(z\|x)$ where z is “simpler” than x
predictive:	specify x and y; learn to predict y from x
prescriptive:	specify x, y, and a; learn to prescribe a given x to maximize y

descriptive:

predictive:

prescriptive:



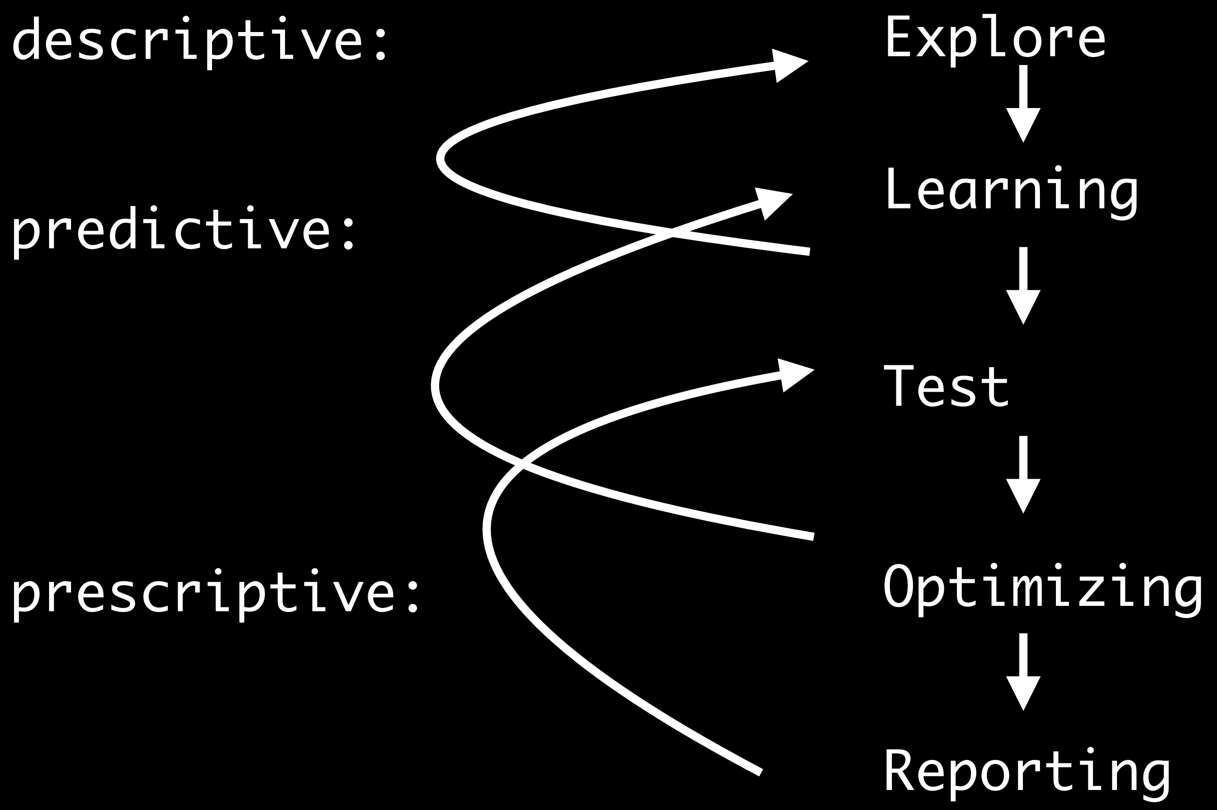
Explore

Learning

Test

Optimizing

Reporting



descriptive:

predictive:

prescriptive:

Explore

Learning

Test

Optimizing

Reporting

data science: ideas

data skills

data science and…

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

cf. “data scientists at work”, ch 1

nota bene!

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

data skills

data science

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

cf. “data scientists at work”, ch 1

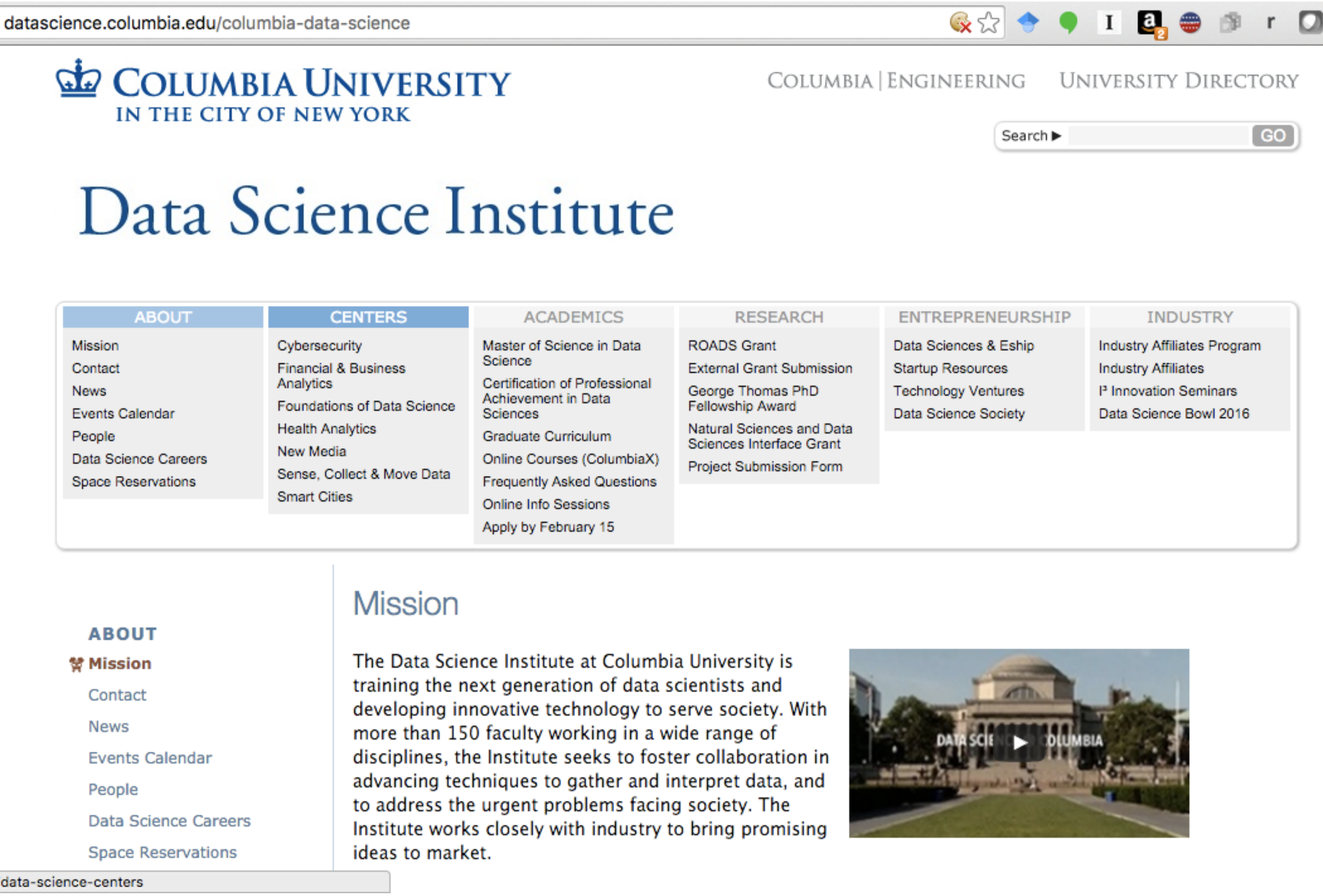
* or at least @ NYT, 2016

data science

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

</talk>

<appendices>



Data Science Institute

ABOUT
- Mission
- Contact
- News
- Events Calendar
- People
- Data Science Careers
- Space Reservations

CENTERS
- Cybersecurity
- Financial & Business Analytics
- Foundations of Data Science
- Health Analytics
- New Media
- Sense, Collect & Move Data
- Smart Cities

ACADEMICS
- Master of Science in Data Science
- Certification of Professional Achievement in Data Sciences
- Graduate Curriculum
- Online Courses (ColumbiaX)
- Frequently Asked Questions
- Online Info Sessions
- Apply by February 15

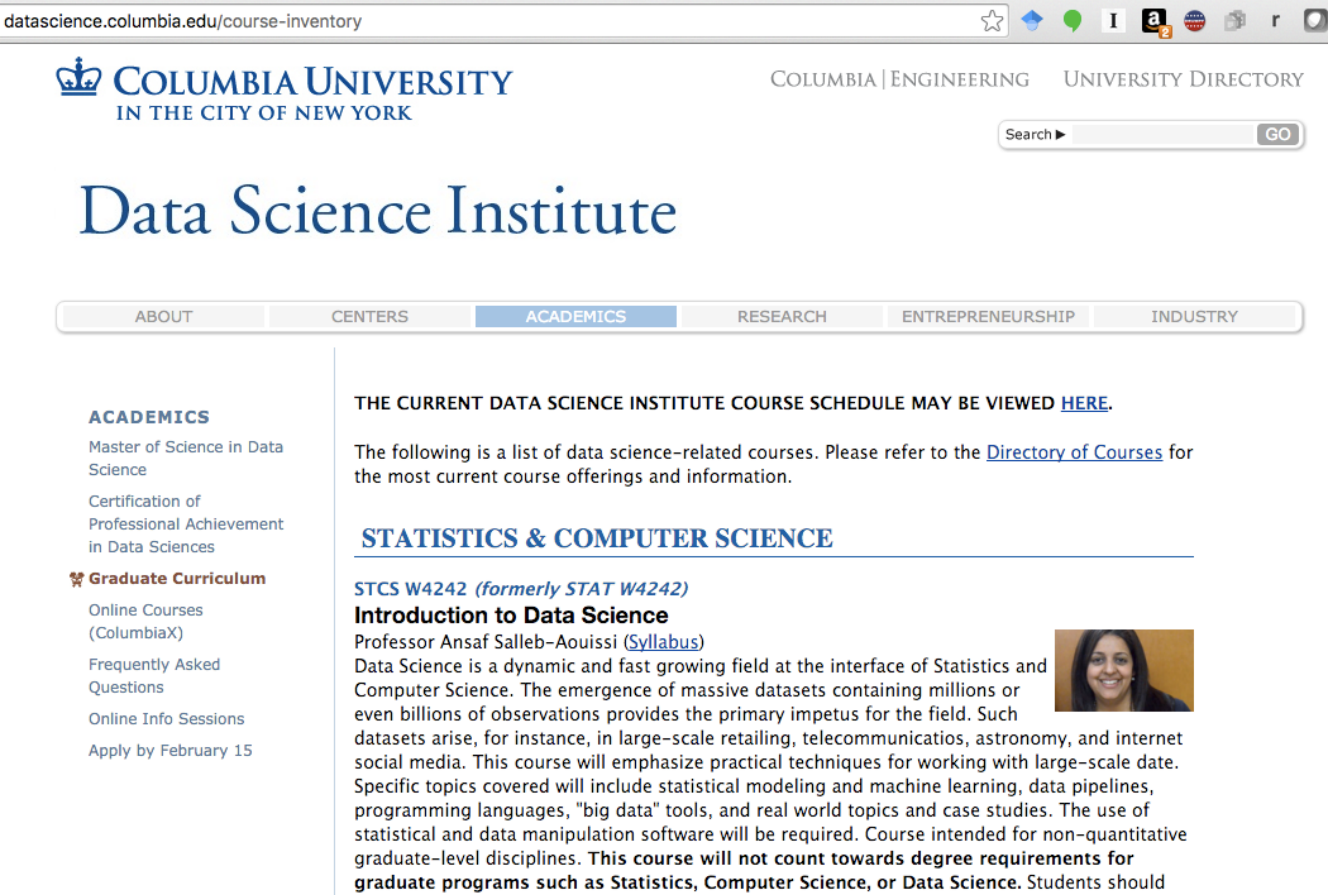
RESEARCH
- ROADS Grant
- External Grant Submission
- George Thomas PhD Fellowship Award
- Natural Sciences and Data Sciences Interface Grant
- Project Submission Form

ENTREPRENEURSHIP
- Data Sciences & Eship
- Startup Resources
- Technology Ventures
- Data Science Society

INDUSTRY
- Industry Affiliates Program
- Industry Affiliates
- I³ Innovation Seminars
- Data Science Bowl 2016

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.

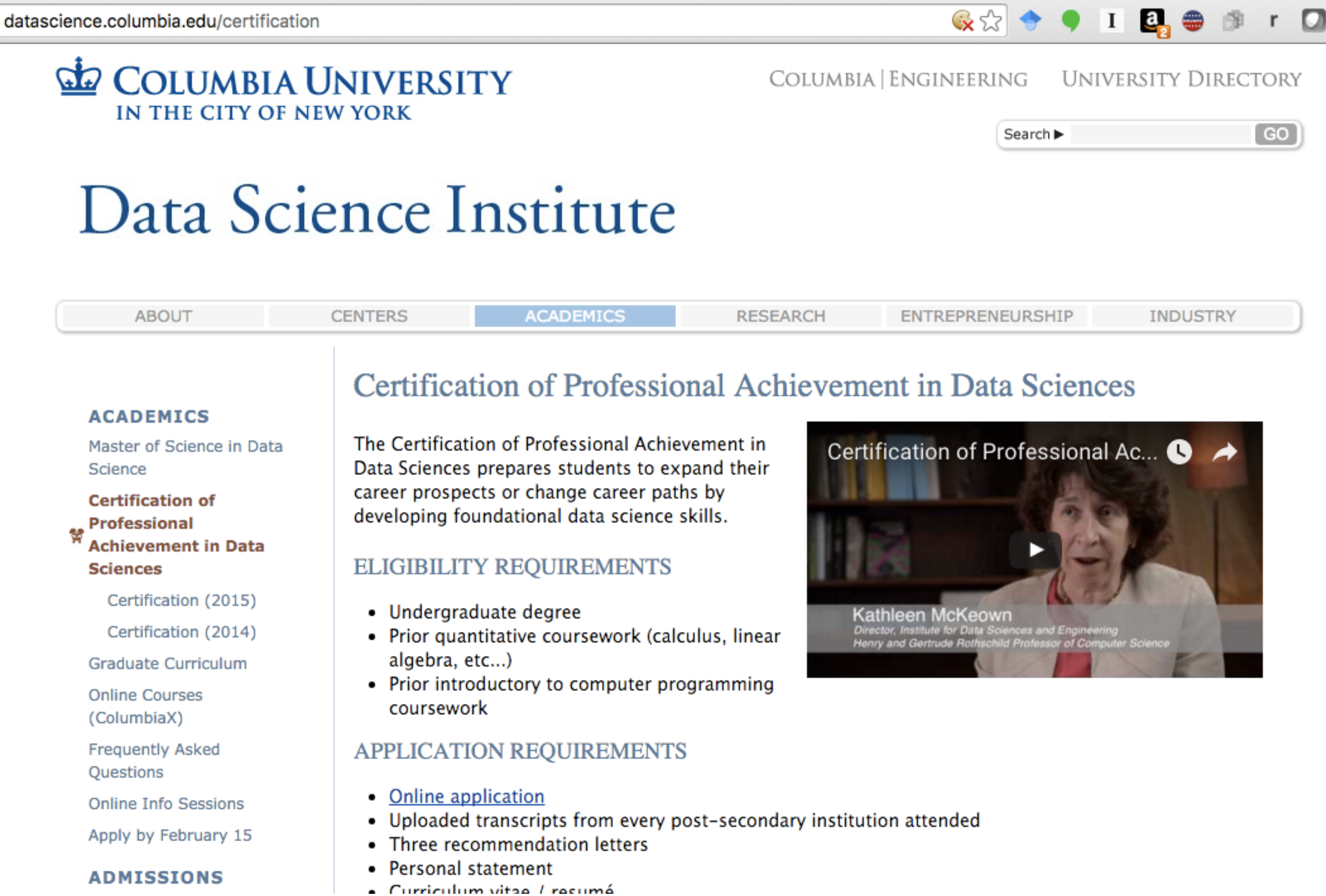
STATISTICS & COMPUTER SCIENCE

STCS W4242 *(formerly STAT W4242)*

Introduction to Data Science

Professor Ansaf 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, telecommunicatios, astronomy, and internet social media. This course will emphasize practical techniques for working with large-scale date. 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

ACADEMICS

- Master of Science in Data Science
- **Certification of Professional Achievement in Data Sciences**
  - Certification (2015)
  - Certification (2014)
- Graduate Curriculum
- Online Courses (ColumbiaX)
- Frequently Asked Questions
- Online Info Sessions
- Apply by February 15

ADMISSIONS

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.

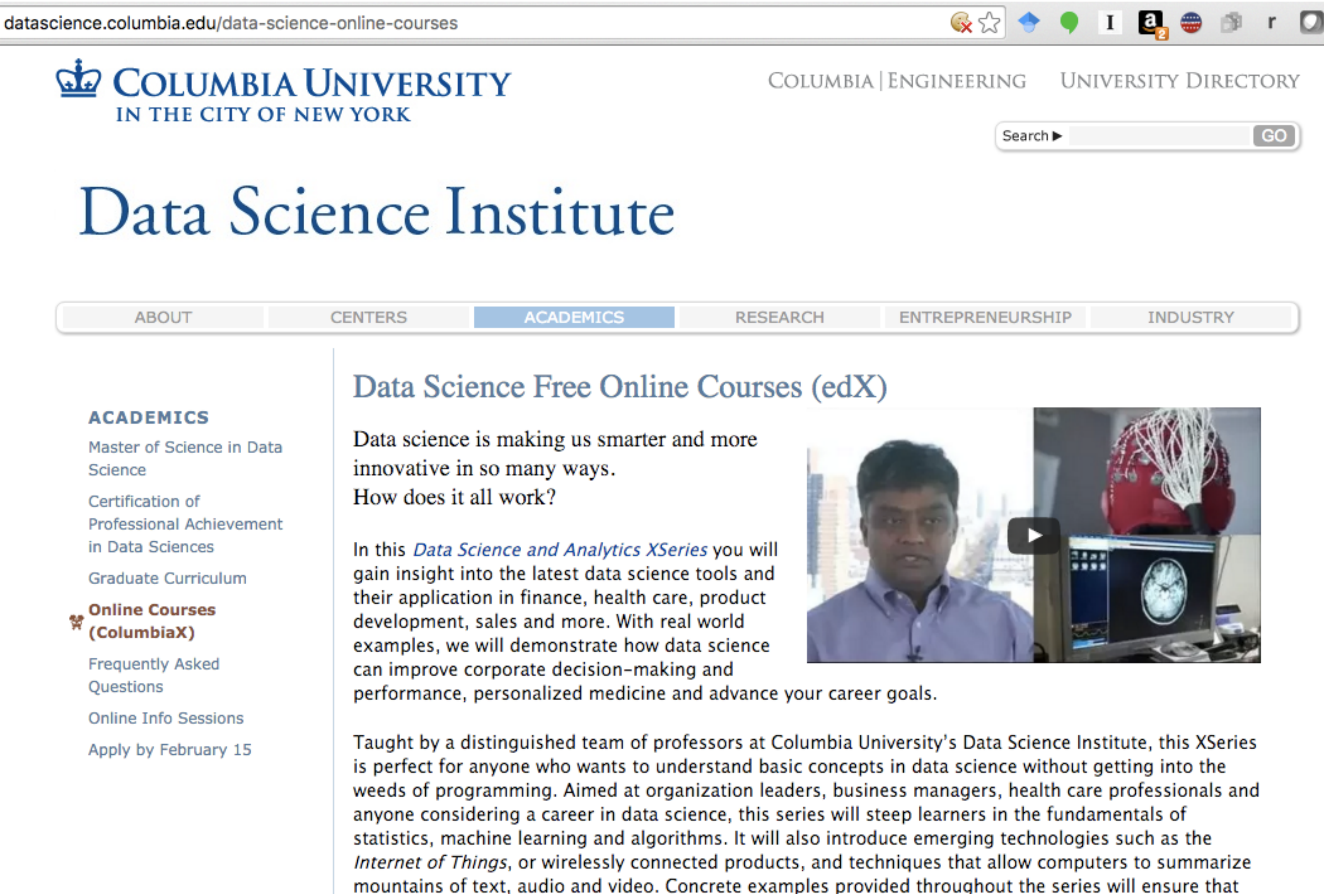
Certification of Professional Ac...
Kathleen McKeown
Director, Institute for Data Sciences and Engineering
Henry and Gertrude Rothschild Professor of Computer Science

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 / resumé



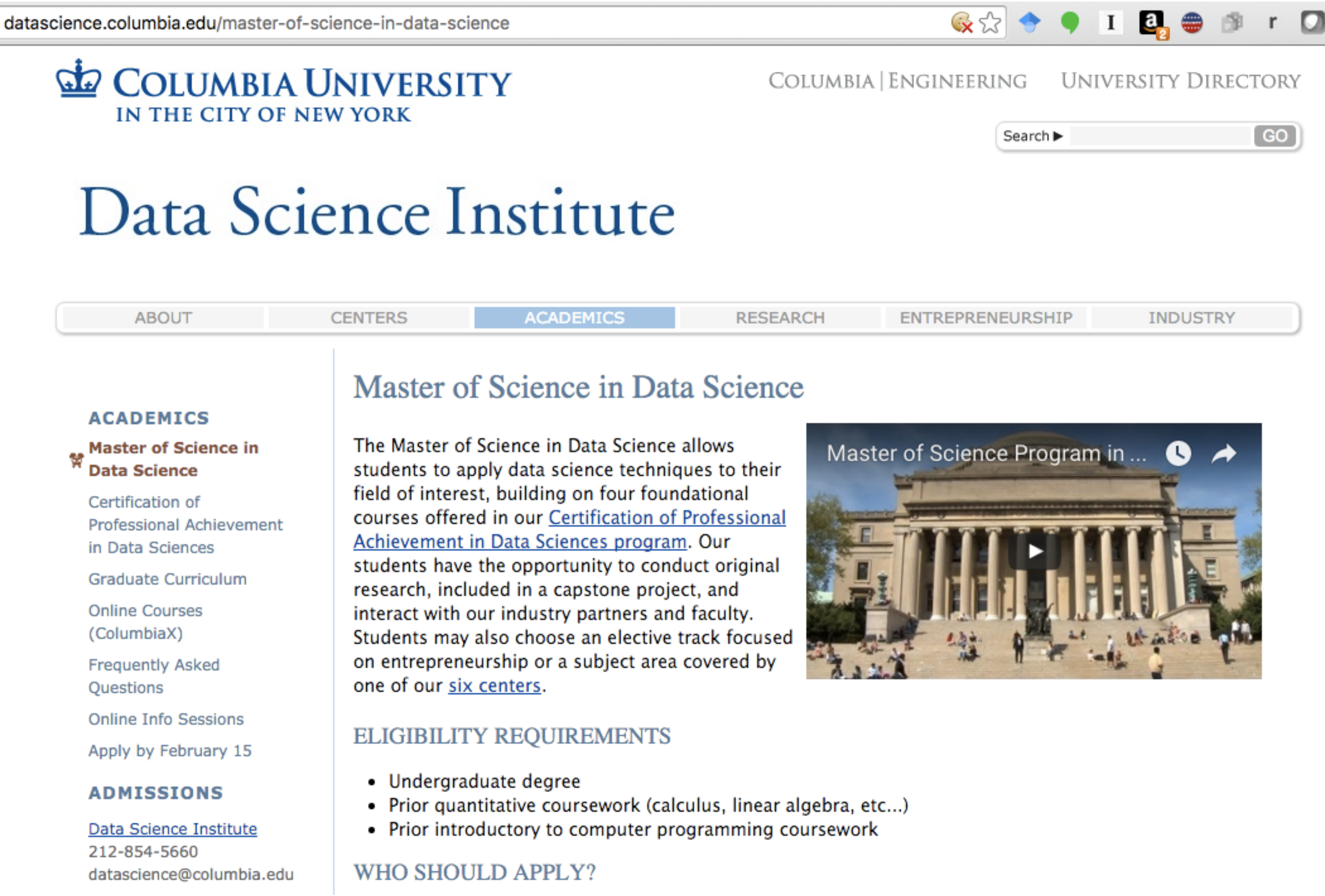
Data Science Institute

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?

ACADEMICS

Master of Science in Data Science

Certification of Professional Achievement in Data Sciences

Graduate Curriculum

Online Courses (ColumbiaX)

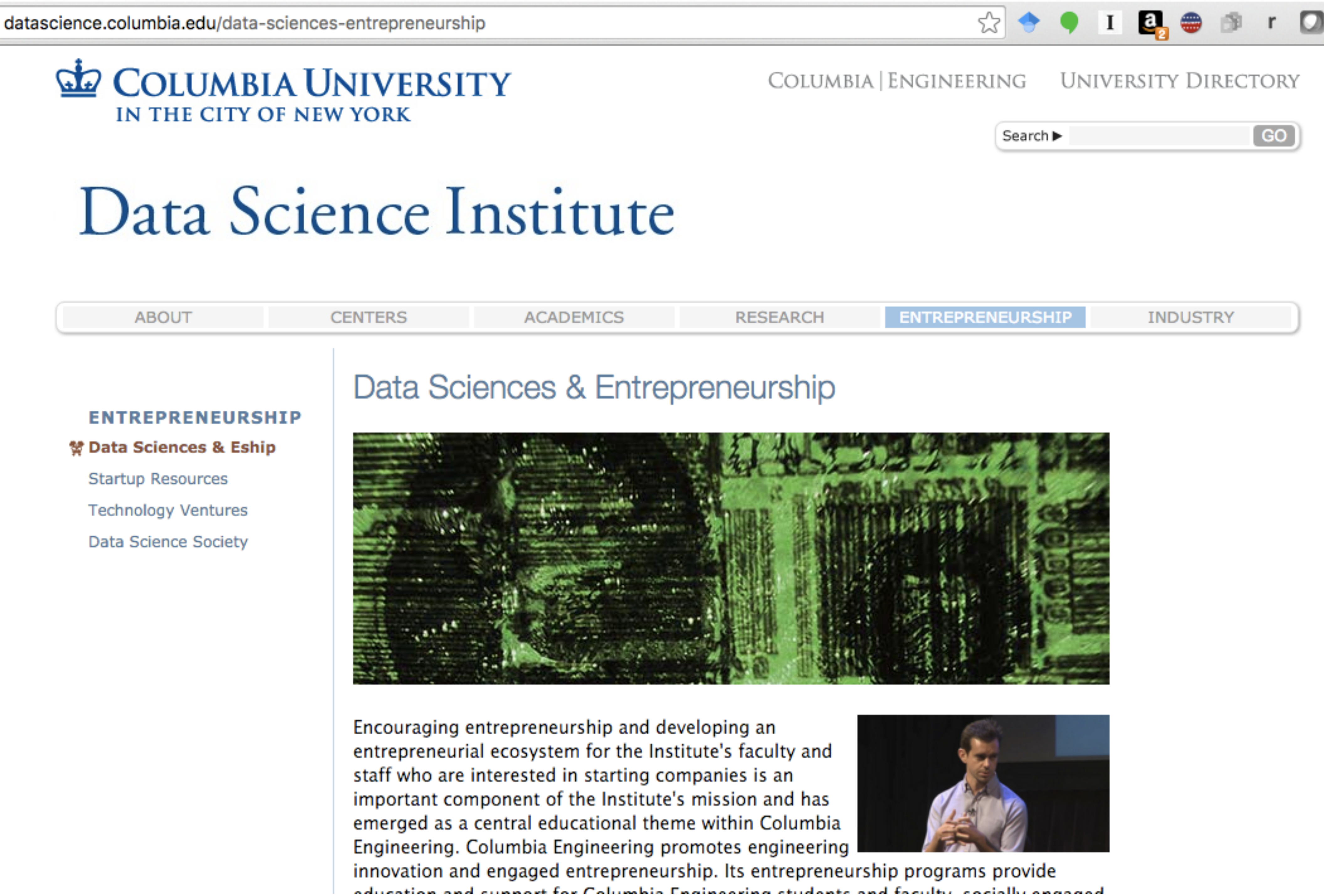
Frequently Asked Questions

Online Info Sessions

Apply by February 15

ADMISSIONS

Data Science Institute
212-854-5660
datascience@columbia.edu

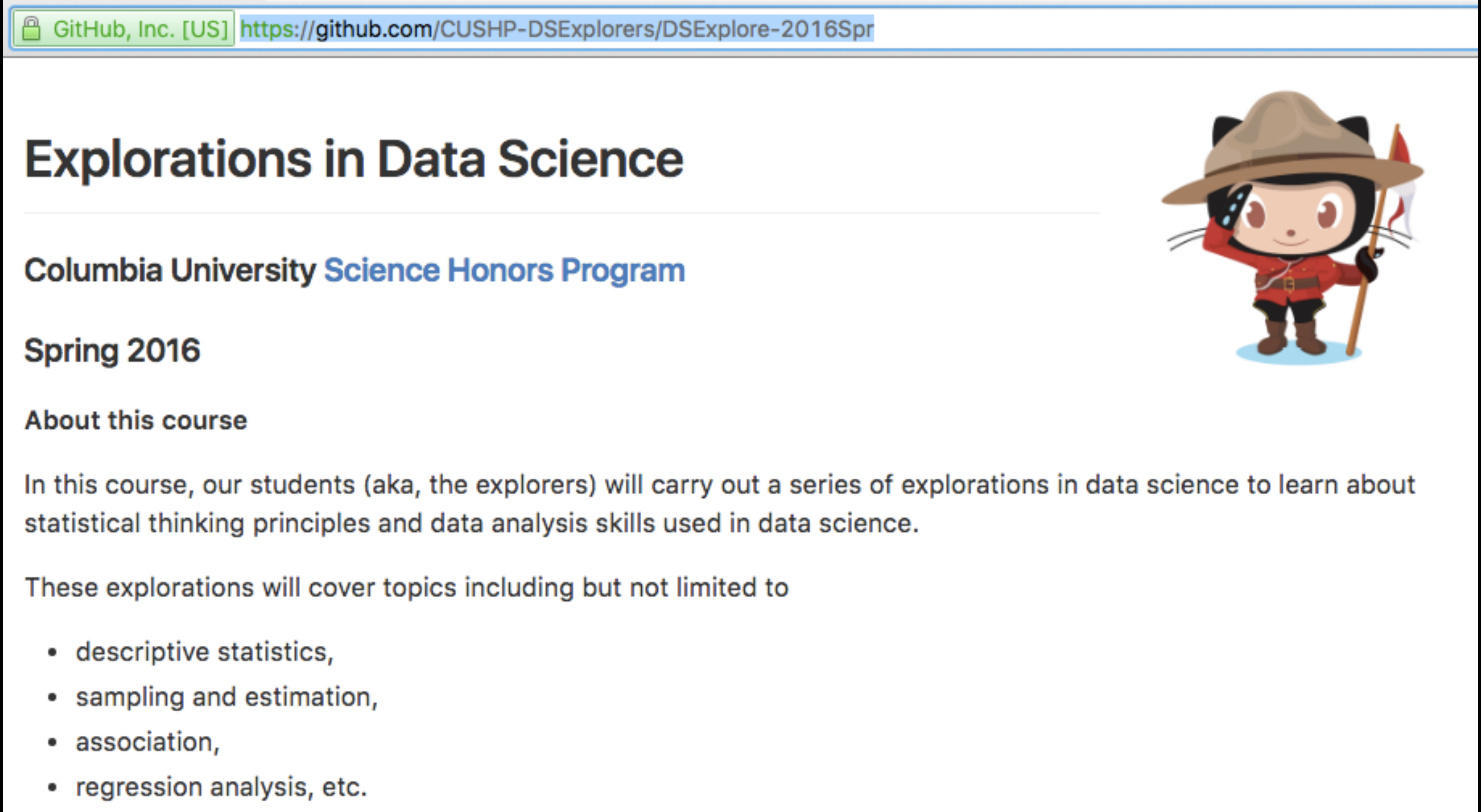


datascience.columbia.edu/data-sciences-entrepreneurship

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

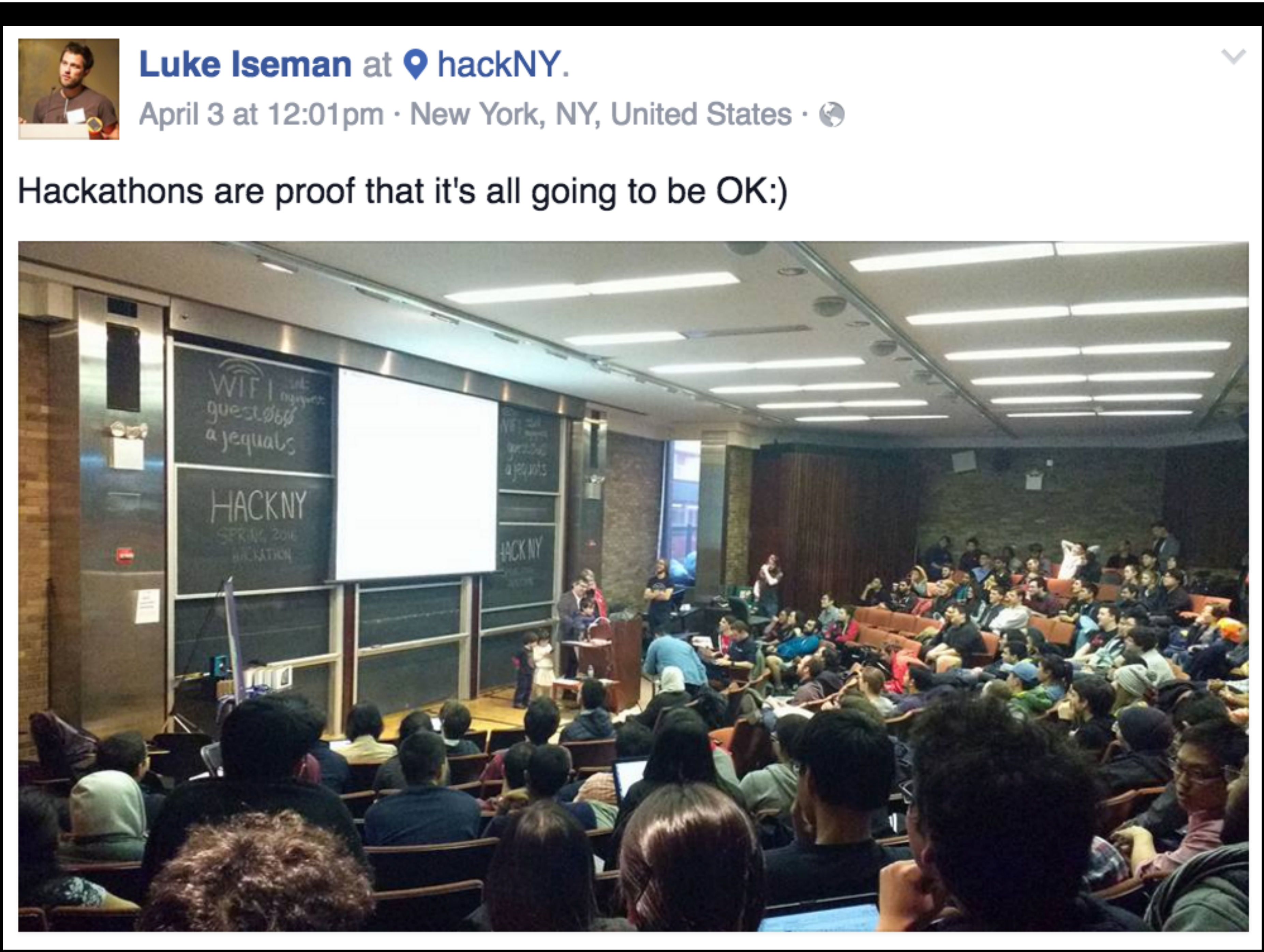
In this course, our students (aka, the explorers) will carry out a series of explorations in data science to learn about statistical thinking principles and data analysis skills used in data science.

These explorations will cover topics including but not limited to

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

high school!

role of higher ed
in complementary/experiential education?



Luke Iseman at hackNY.

April 3 at 12:01pm · New York, NY, United States ·

Hackathons are proof that it's all going to be OK:)

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