Institute of Mathematical Statistics

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability

IMS Presidential Address

at the ASC-IMS meeting in Sydney, July, 2014

Let Us Own Data Science*

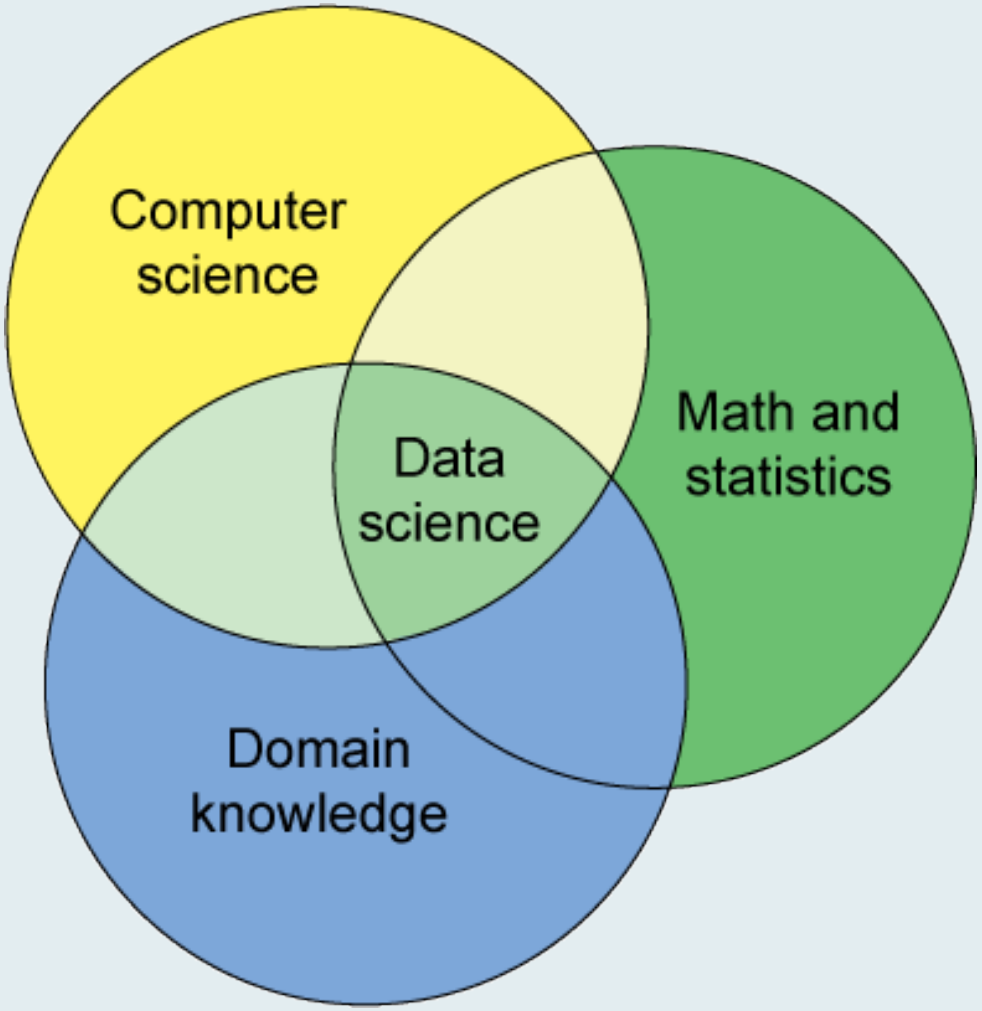
Bin Yu, IMS President (Statistician and Data Scientist)

statistics.berkeley.edu/~binyu

Statistics and EECS, University of California-Berkeley

*"Own" here does not exclude other owners of data science.

Data science is all the rage



http://www.ibm.com/developerworks/jp/opensource/library/os-datascience/figure1.png

We gather here: SSAI and IMS people

http://upload.wikimedia.org/wikipedia/commons/8/80/Sydney_Opera_house_3.jpg

Turn the clock back 67 years to 1947 Sydney

http://www.davidmoorephotography.com.au/media/images/100photographs/onetoten/s1361-2.jpg

Beginning of SSAI in Sydney, 1947

The Statistical Society of New South Wales was formed in the spring of 1947 in Sydney to "(further) the study and application of statistical methods in all branches of knowledge and human activity."

http://sydney.edu.au/senate/images/unihistory3/smith31.gif

SSNSW First President: Helen Newton Turner

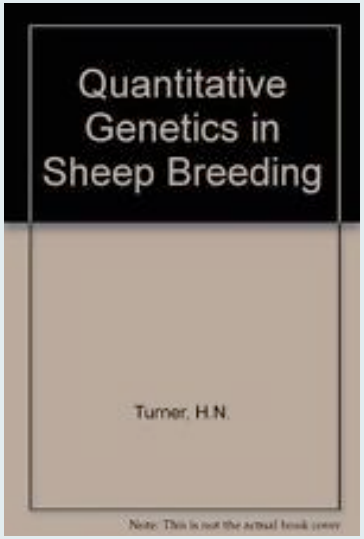
A statistician and a geneticist

(1908-1995)

Helen Newton Turner, the animal geneticist who spent most of her professional life working for the improvement of the wool industry, died this week [late November1995-Ed.]. Not many scientists in Australia have contributed as much and as directly to the growth and well-being of a major Australian industry as she did. Not that she would have agreed with this description; she was a very modest woman and would have said, 'It wasn't me, it was my team'.

http://sydney.edu.au/senate/documents/Students/Turner_obituary.pdf

Helen Turner: a scientist and a publicist



- The first (contribution) was as an experimental scientist, introducing objective measurement methods into breeding.
- The second was as a communicator and publicist of the new methods.
- The third was as an educator of postgraduate students and Department of Agriculture staff.

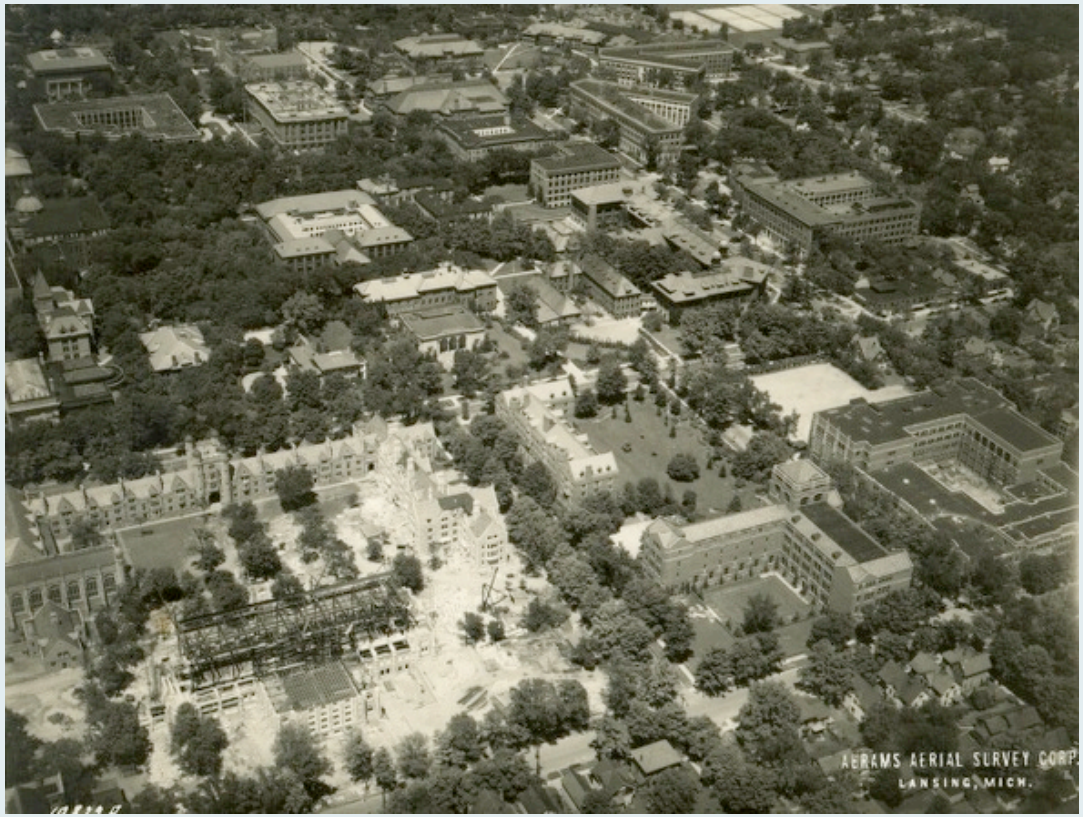
http://sydney.edu.au/senate/documents/Students/Turner_obituary.pdf

Turn the clock back 17 more years to 1930 Ann Arbor

http://www.annarbor.com/assets_c/2012/11/klines_1930s-thumb-300x228-128075.jpg

Beginning of IMS in Ann Arbor, 1930

The Annals of Mathematical Statistics was started at Univ of Michigan



http://www.annarbor.com/assets_c/2012/06/1930-31aerialview-thumb-590x445-114160.jpg

Annals First Editor: Harry C. Carver

A mathematical statistician and an aerial navigation expert

(1890-1977)

Decoration for Exceptional Service by US Air Force

When he was ready to retire at 70, he set up criteria for average temperature, total rainfall, number of days of sunshine, etc., and then conducted a systematic search of the U.S. weather records to find the winning location.

His decision was to choose Santa Barbara, California, and he moved.

Quote from Carver in 1943

"… logarithms must be considered now as a tool of the past. Present-day commercial institutions almost without exception use computing machines rather than logarithms in the conduct of their business in the interest of efficiency in both time and labor: moreover progressive schools now have installations of these machines that enable their students to work more problems in less time than formerly..."

Preface, "An introduction to Air Navigation" (1943)

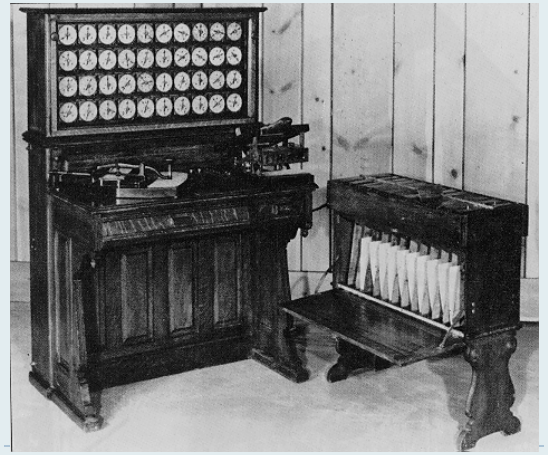
by H. C. Carver, published by Edwards Brothers, Inc.

Carver: early "machine learner"

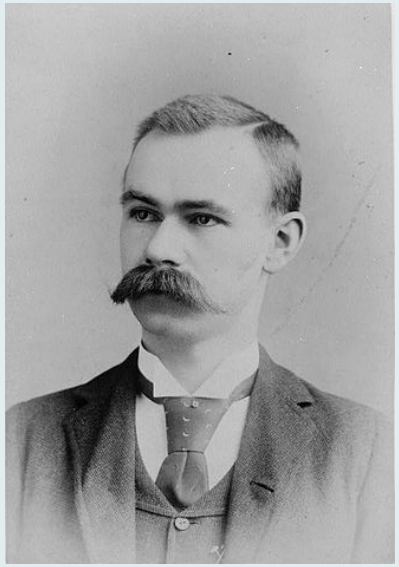
Air navigation is the process of planning, recording, and controlling the movement of a craft from one place to another.

Carver got the first Hollerith tabulating machine at U-M (later IBM).

His approach to air navigation could be viewed with a modern lens as on-line "machine learning" -- fast estimation based on instantly measured data through optimization (with possibly uncertainty measure)



Inventor of Hollerith Machine: Herman Hollerith



A statistician and an inventor

(1860-1929)

Founder of the Tabulating Machine Company

that later merged to become IBM

Hollerith is widely regarded as the father of modern machine data processing. With his invention of the punched card evaluating machine the beginning of the era of automatic data processing systems was marked. His draft of this concept dominated the computing landscape for nearly a century.

-- Wikipedia

Turner + Carver = "Data Scientist"

Putting all the traits of Turner and Carver together, we get a good portrait of data scientist:

1. Statistics (S)
2. Domain (science) knowledge (D)
3. Computing (C)
4. Collaboration ("team work") (C)
5. Communication (to outsiders) (C)

Data Science = SDC^3

W. Cochran (1953): Sampling Techniques

A statistician
(1909 – 1980)

S. S. Wilks Medal of ASA

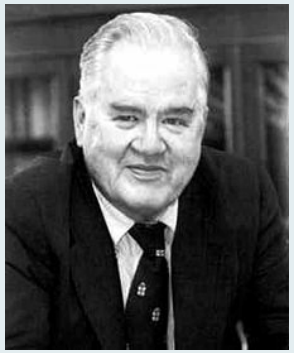
Set up (bio)stat depts of
Iowa-State, NC-State, Johns Hopkins, and Harvard

http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Cochran.html

"Our knowledge, our attitudes, and our actions are based to a very large extent upon samples. This is equally true in everyday life and in scientific research."

"But when the material is far from uniform, as is often the case, the method by which the sample is obtained is critical, and the study of techniques that ensure a trustworthy sample becomes important." -- Introduction in Sampling Techniques

J.W. Tukey (1962): Future of data analysis



A mathematician
(1915 – 2000)

U.S. Medal of Science
IEEE Medal for co-invention of FFT

It will still be true that there will be aspects of data analysis well called technology, but there will also be the hallmarks of stimulating science: intellectual adventure, demanding calls upon insight, and a need to find out "how things really are" by investigation and the confrontation of insights with experience.

Tukey's definition of "data science"?

Tukey's Principles of Statistical Practice summarized by A.D. Gordon (in my words)

- Usefulness and limitation of theory
- Importance of robustness
- Importance of massive empirical experience of a method for guiding its use
- Importance of data's influence on methods chosen
- Rejection of the role of "police"
- Resistance to once-for-all solutions and over-unification of statistics
- Iterative nature of data analysis
- Importance of computing
- Training of statisticians

-- Wikipedia

Statisticians do a big part of the job of a data scientist

No existing discipline does more of the job of a data scientist

To fortify our position in DS, we should focus on

Critical thinking enables Statistics + Domain knowledge

Computing

(parallel computation, memory and communication dominate scalability)

Leadership, interpersonal, and communication

abilities enable collaboration + communication with outside

For the twitter generation

Think or sink

Compute or concede

Lead or lose

We do the job so let us

call ourselves data scientists!

What is the big deal about a name?

Words do mean things.

The game of branding

Fields evolve, including Statistics and Computer Science

1. Statistics:

   Same term with diff. meanings to diff. people, and changing meanings over time

   "statistics" covers a vast range of activities so not very informative.

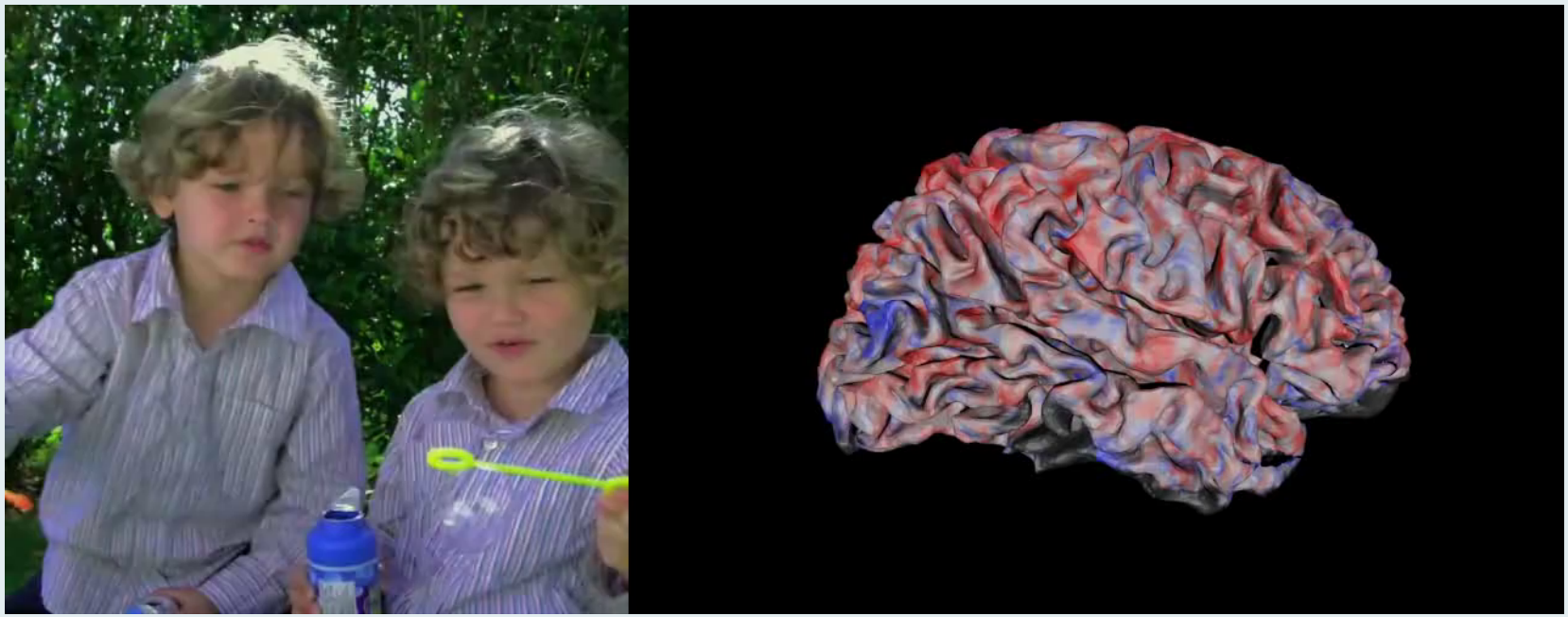
2. Computer Science:

   New names for new developments over time

   "AI, Data Mining, Machine Learning, Deep Learning"…

Imagine a 18 years old J. choosing between statistics and data science

J's brain is excited by new and bored with old as shown below: the video stimuli (left) and corresponding fMRI signals (right) used in Nishimoto, Vu, Naselaris, Benjamini, Yu and Gallant (2011). *Current Biology*.



"Statistics" is a multi-valued function

"Statistics" = first statistics course, AP statistics class,…

Statistics is the study of the collection, organization, analysis, interpretation and presentation of data. It deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments. When analyzing data, it is possible to use one of two statistics methodologies: descriptive statistics or inferential statistics.

-- Wikipedia

Most popular by Google with “Statistics”

Statistics - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Statistics**
Statistics is the study of the collection, organization, analysis, interpretation and presentation of data. It deals with all aspects of data including the planning of ...
Data - Mathematical statistics - Descriptive statistics - Statistical inference

Demographics of Melbourne - Wikipedia, the free ...
en.wikipedia.org/wiki/Demographics_of_Melbourne
1 Demographic **statistics**; 2 Demographic history. 2.1 European settlement and Gold Rush immigration; 2.2 Post-war immigration. 3 Socioeconomics; 4 Foreign ...

News for **statistics**

Bureau of **Statistics** in dark on foreign investment in ...
Sydney Morning Herald - 7 hours ago
Senior Bureau of **Statistics** officials have admitted they read trade magazines and newspapers to keep on top of changes in the amount of ...

Statistics reveal loss of Mathew Stokes a major blow for ...
The Age - 14 hours ago

Facebook releases diversity **statistics**
USA TODAY - 1 hour ago

More news for **statistics**

Australian Bureau of **Statistics**

Statistics

Field Of Study

Statistics is the study of the collection, organization, analysis, interpretation and presentation of data. It deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments. Wikipedia

Related topics

Probability distributions are a fundamental concept in statistics. itl.nist.gov
Explore: Probability distribution

In probability theory and statistics, **variance** measures how far a set of numbers is spread out. Wikipedia
Explore: Variance

Regression analysis is a statistical tool for the investigation of relationships between variables. law.uchicago.edu
Explore: Regression analysis

Feedback

How about "Statistician"? -- Wikipedia

A statistician is someone who works with theoretical or applied statistics. The profession exists in both the private and public sectors. It is common to combine statistical knowledge with expertise in other subjects.

It doesn't speak to an outsider… Read on…

… Typical work includes collaborating with scientists, providing mathematical modeling, simulations, designing randomized experiments and randomized sampling plans, analyzing experimental or survey results, and forecasting future events (such as sales of a product).

More jargons…

Image of "statisticians" for an outsider

We "don't deal with risk, with uncertainty… we're too absolute, we do p-values, confidence intervals, definite things like that."

We "raise arcane concerns about mathematical methods."

and "I had no interest in very experienced statisticians." …

"I wasn't even thinking about what model I was going to us. I wanted actionable insight, and that was all I cared about."

-- Terence's Stuff, IMS Bulletin 2014 June Issue

Then J. looks at "Data science"

"Data science" is about data and science!

It is NEW…

Don't know what it is, but let's find out…

Data science is the study of the generalizable extraction of knowledge from data, yet the key word is *science*.

-- Wikipedia

More from google "data science search"

"Data scientists are inquisitive: exploring, asking questions, doing "what if" analysis, questioning existing assumptions and processes. Armed with data and analytical results, a top-tier data scientist will then communicate informed conclusions and recommendations across an organization's leadership structure."

http://www-01.ibm.com/software/data/infosphere/data-scientist/

Sounds like an applied statistician's description?

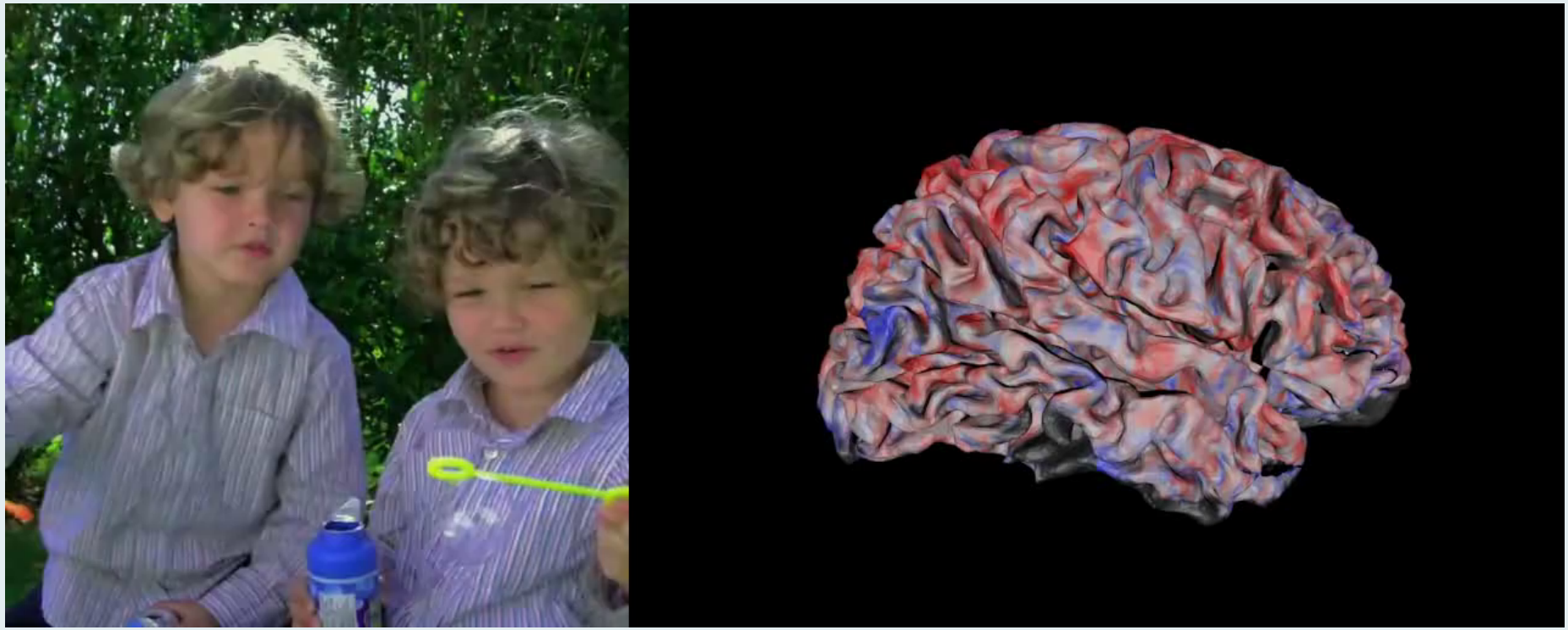
How many statisticians in academic departments relate to this description as their job description and for what percentage?

Many of our students go to industry and need these skills…

Imagine a 18 years old J. choosing between statistics and data science

J's brain is excited by new and bored with old.

Nishimoto et al (2011). Current Biology



What would J. do?

Statistics or Data Science?

The power of a good new name

- A data science MA program attracts a lot more applicants than a statistics MA program.
- Moore-Sloan foundations established three data science environ. centers
- Moore foundation is giving DDD investigator awards
- Universities are looking into data science FTEs
- Industry is advertising for many DS-positions…

Claiming "data science" as our own in 1998 by Jeff Wu

Quotes from Jeff Wu's inaugural lecture of his Carver (!) Chair Professorship at Univ of Michigan in 1998.

Statistics = Data Science ?

C. F. Jeff Wu

University of Michigan, Ann Arbor

A proposal:

"Statistics" ⟶ "Data Science"

"Statisticians" ⟶ "Data Scientists"

Quotes from Wu's slides (cont)

- Several good names have been taken up: computer science, information science, material science, cognitive science

- "Data Science" is likely the remaining good name reserved for us

L. Breiman (2001): Statistical modeling: the two cultures

A probabilist, and statistician, machine learner
(1928 – 2005)

CART, Bagging, Random Forests

"If our goal as field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a diverse set of tools."

Data Science: Inevitable (re)-merging of computational and statistical thinking

We have to own data science, because

- Domain problems don't differentiate computation and statistics
- Data science is the new accepted term to deal with a modern data problem in its entirety.

Gains for statistics community are

- attracting talent and resource, and
- securing jobs for our majors, MAs and PhDs.

How to own data science?

As Turner, Carver, Hollerithm,

Cochran and Tukey did,

work on real problems,

relevant methodology/theory

will follow naturally.

What are the problems of today?

Genomics

Neuroscience

Astronomy

Nano-science

…

Personalized medicine/healthcare

…

Computational social science

…

Industry

…

McKinsey Report (2011) Big data: The next frontier for innovation, competition, and productivity

…

6. There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.

7. Several issues will have to be addressed to capture the full potential of big data. Policies related to privacy, security, intellectual property, and even liability will need to be addressed in a big data world.

The Statistics Tent is Big Enough

Suggestions for Actions

At individual level, we could

- Put "data scientist" next to "statistician" on your website and resume, if your job is partly data science
- Get on interesting and/or important projects
- Upgrade computing, interpersonal and leadership skills
- Scale up our algorithms and carry out relevant research (analysis of parallel and/or random algorithms)

Our public image needs much improvement

- 5-min elevator talk (everyone!) in layman language
- pubic speech
- interview with media
- blog, twitter
- youtube video
- website with accessible descriptions of work

Update statistics and data science wikipedia pages
by you and/or students in your class

As a community, we should

- Write vision statement and/or white paper to persuade upper admin and funding agencies to provide resource (money and positions) (recall Cochran started 4 stat/biostat departments!)

- Reform statistical curriculum (intro to advanced courses) –

  MOOC or on-line (multi-media) courses that integrate the essence of statistics and computing principles, with interesting modern data problems as examples.

IMS Data Science Projects

1. An IMS-MSR Data Science Conference in 2015

   "Foundations of data science:

   Synergies between statistics and machine learning"

   Organization committee:

   Co-chairs: David Dunson, Rafa Irizarry, and Sham Kakade

   A. Braverman, S. Dumais, A. Munk, M. Wainwright

2. IMS membership drive to attract young people, especially young machine learners

One IMS procedure change

In 2015,

Named and Medallion Lecture nominations

will be

open to the community

http://imstat.org/awards/lectures/nominations.htm

Let us own data science.

Think or sink

Compute or concede

Lead or lose

More references (a biased selection)

Lindsay, B. G., Kettenring, J., and Siegmund, D. O. (2004), "A Report on the Future of Statistics," Statistical Science, 19, 387–413.

Yu, B. (2007). "Embracing Statistical Challenges of the Information Technology Age." Technometrics, 49, 237—248

Computing Community Consortium (CCC) (2012). White papers. "Challenges and Opportunities with Big Data," "From Data to Knowledge to Action: A Global Enabler for the 21st Century," and "Big-Data Computing."

Jordan et al (2013). US NRC report on Massive Data.

Rudin et al (2014). ASA white paper. Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society