Deloitte.

UCSB

# Actuarial Science as Data Science

*A vision for actuarial science in the 21st century*

California Actuarial Student Congress
UC – Santa Barbara
April 8, 2016

James Guszcza, PhD, FCAS, FSA
Deloitte Consulting | US
jguszcza@deloitte.com

# James Guszcza – US Chief Data Scientist, Deloitte Consulting

Deloitte.

**Deloitte Consulting LLP**
350 South Grand Avenue
Los Angeles, CA 90071

**James Guszcza, PhD, FCAS**
Chief Data Scientist
Deloitte Consulting | US

Tel: +1 310 883 4042
jguszcza@deloitte.com

Member of
**Deloitte Touche Tohmatsu**

James Guszcza is the Chief Data Scientist of Deloitte Consulting in the United States, as well as a member of Deloitte's Advanced Analytics and Modeling practice.

Jim has applied statistical and machine learning methods to such diverse business problems as healthcare utilization, customer and employee retention, talent management, customer segmentation, insurance pricing and underwriting, credit scoring, child support enforcement, patient safety, claims management, and fraud detection. He has also spearheaded Deloitte's use of behavioral nudge tactics to more effectively act on model indications.

A frequent author and conference speaker, Jim designed and teaches hands-on business analytics training seminars for both the Casualty Actuarial Society and the Society of Actuaries.

Jim is a former professor at the University of Wisconsin-Madison business school, and he holds a PhD in the Philosophy of Science from The University of Chicago. Jim is a Fellow of the Casualty Actuarial Society, and on its board of directors.

# Agenda

Actuarial science and data science

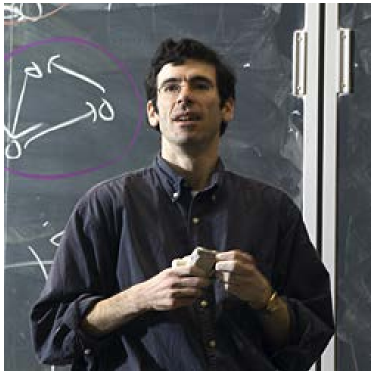
Why analytics is everywhere

A small note about big data

A new mindset for data science

# Data science and actuarial science

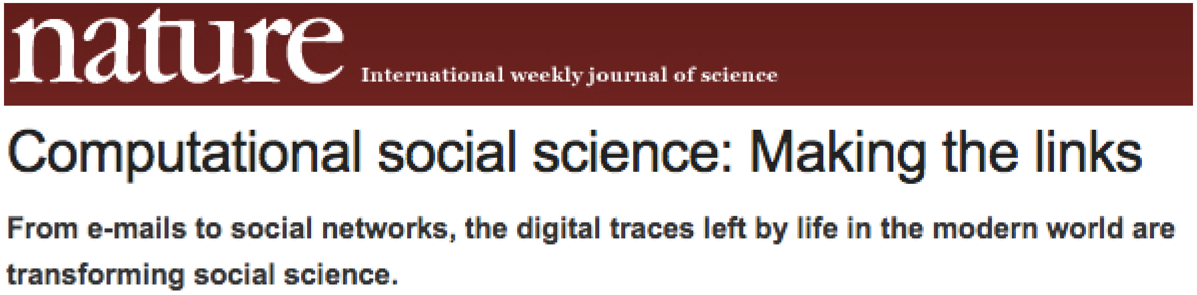
# “The potential to transform everything”



*“The term itself is vague, but it is getting at something that is real...*

*Big Data is a tagline for **a process that has the potential to transform everything**.”*

*— Jon Kleinberg, Cornell University*



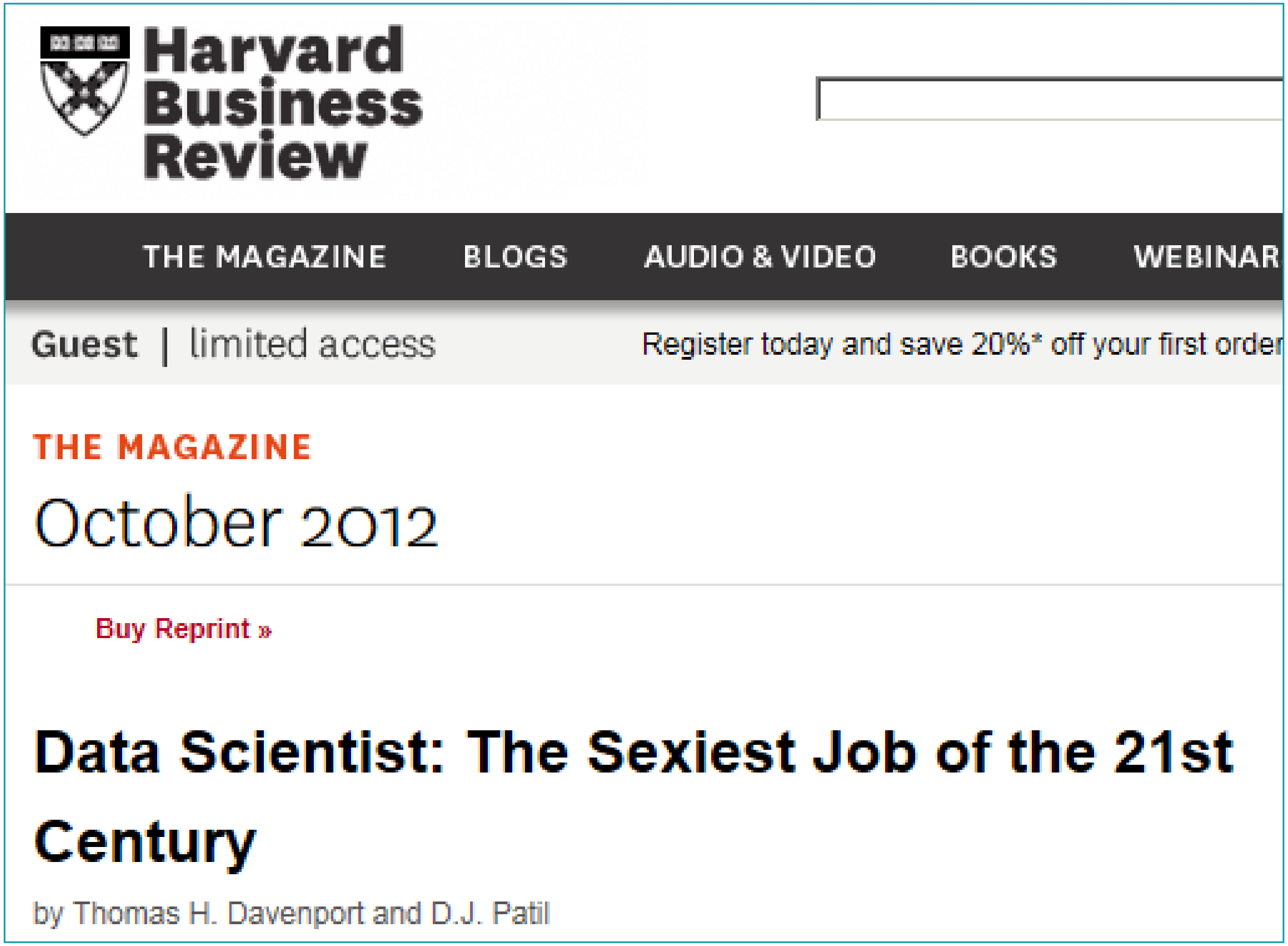
nature International weekly journal of science

Computational social science: Making the links

**From e-mails to social networks, the digital traces left by life in the modern world are transforming social science.**

# Glamorous models

(No, I’m not making this up)



Harvard Business Review

THE MAGAZINE BLOGS AUDIO & VIDEO BOOKS WEBINAR

Guest | limited access

Register today and save 20%* off your first order

THE MAGAZINE

October 2012

Buy Reprint »

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

# At the center of it all:  data science

Or:  “The Collision between Statistics and Computation”

*An intuitive definition of “data science”…*

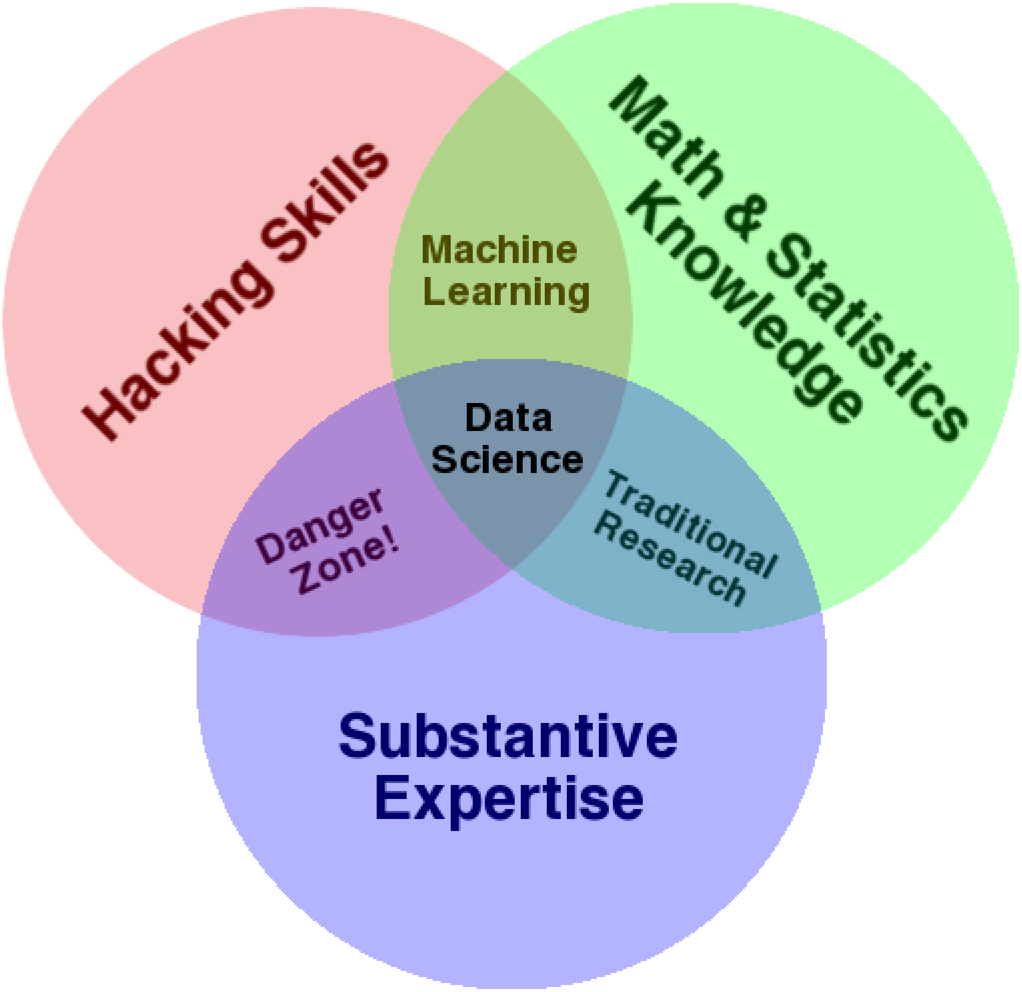


Image borrowed from Drew Conway’s blog
http://www.dataists.com/2010/09/the-data-science-venn-diagram

# At the center of it all: data science

Or: “The Collision between Statistics and Computation”

*Is the actuarial profession here?*

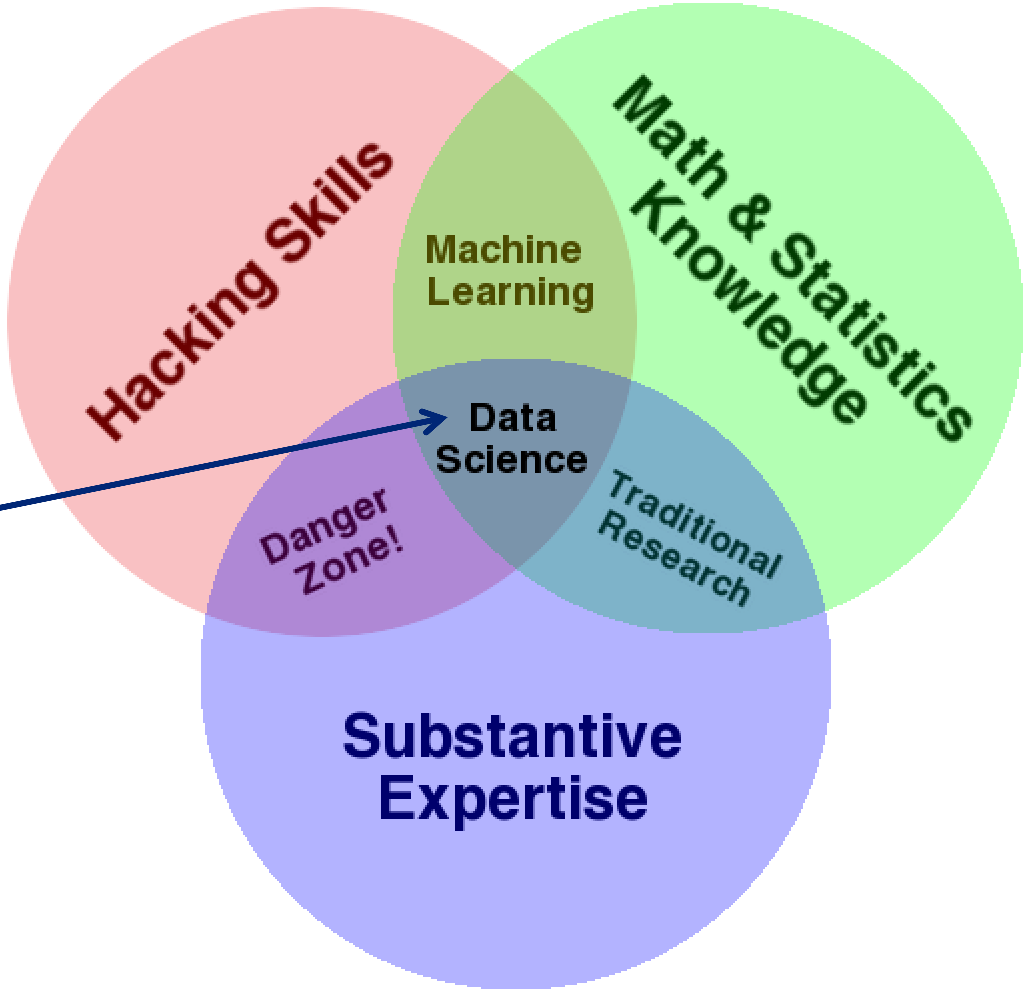


Image borrowed from Drew Conway’s blog
http://www.dataists.com/2010/09/the-data-science-venn-diagram

# At the center of it all:  data science

Or:  “The Collision between Statistics and Computation”

*Or are we here?*

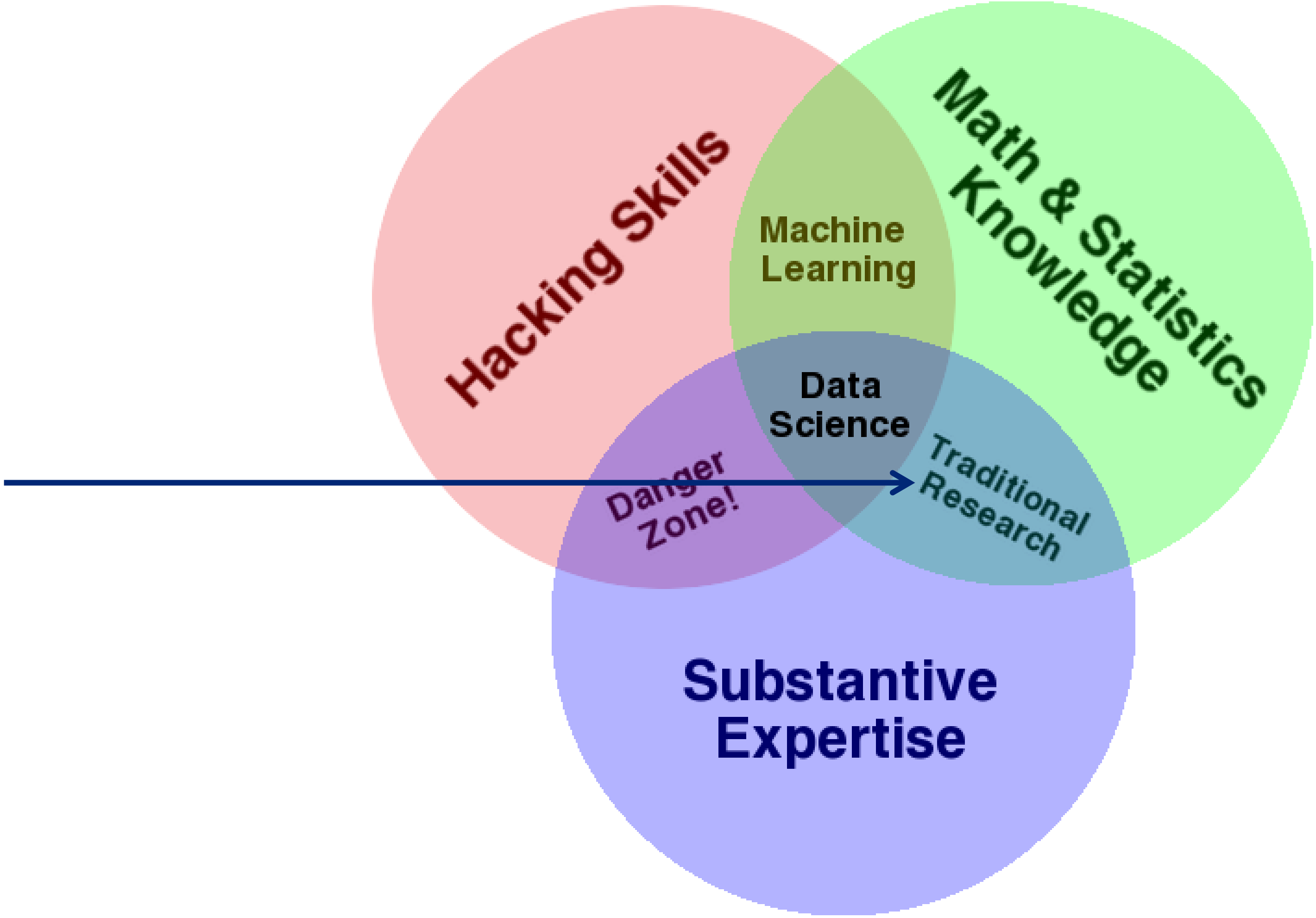


Image borrowed from Drew Conway’s blog
http://www.dataists.com/2010/09/the-data-science-venn-diagram

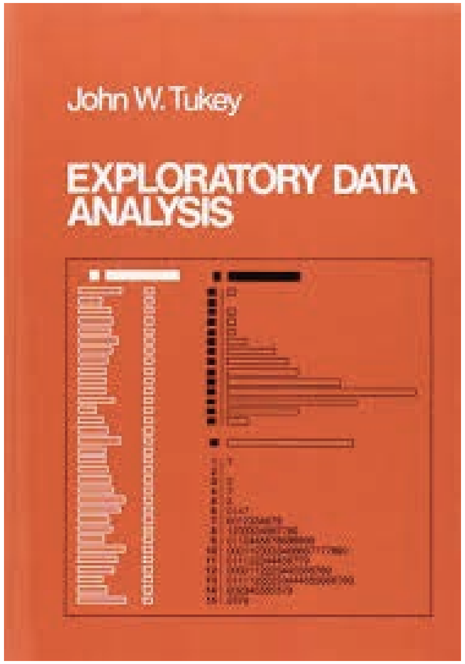
# 50 years of data science

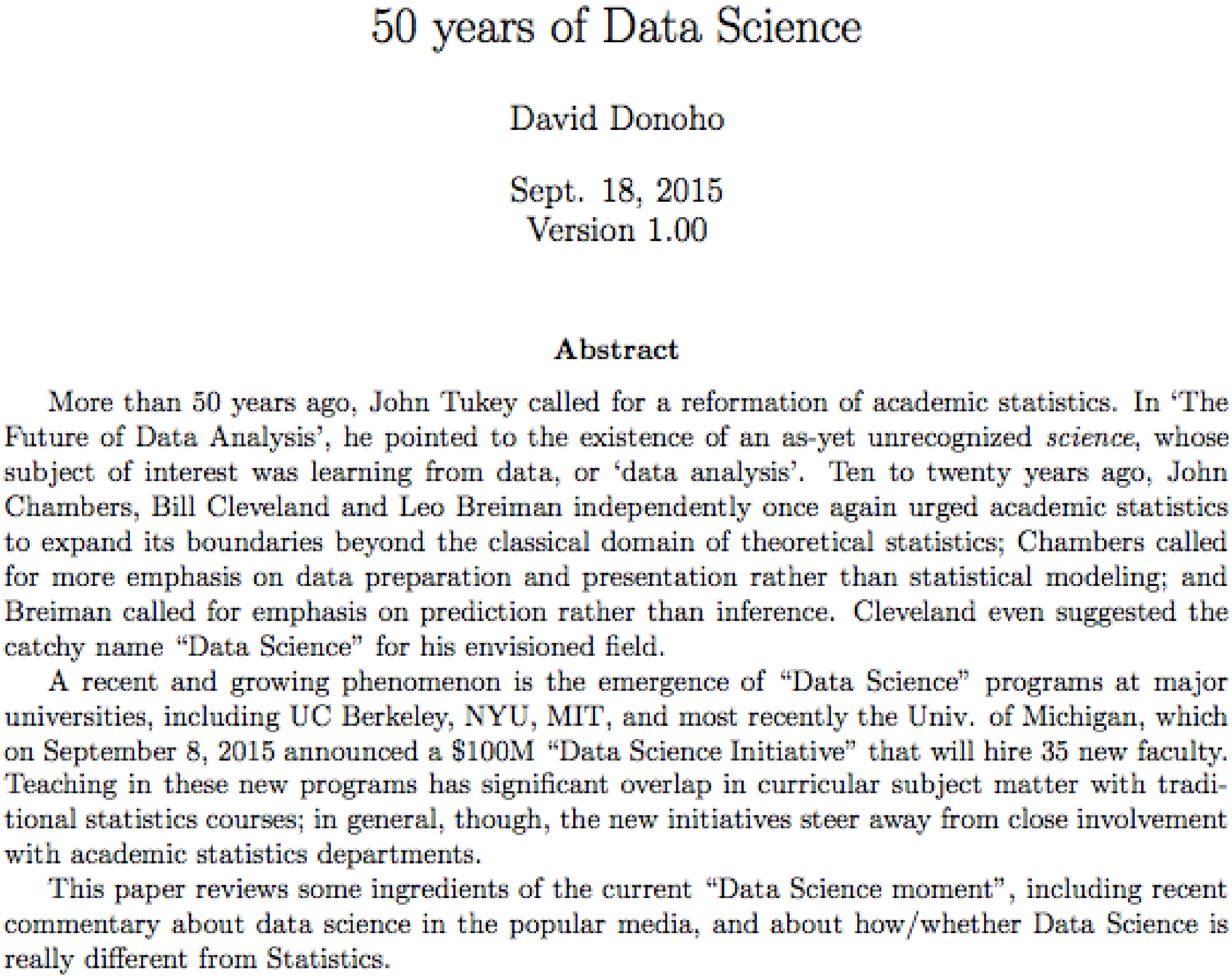
“For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt…

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.”

-- John Tukey Princeton/Bell Labs

“The Future of Data Analysis” (1962)





# 50 years of Data Science

David Donoho

Sept. 18, 2015
Version 1.00

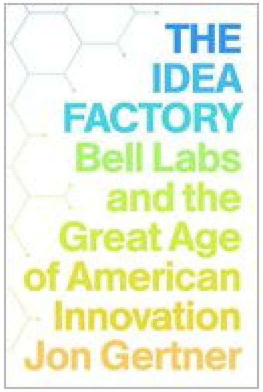
## Abstract

More than 50 years ago, John Tukey called for a reformation of academic statistics. In ‘The Future of Data Analysis’, he pointed to the existence of an as-yet unrecognized *science*, whose subject of interest was learning from data, or ‘data analysis’. Ten to twenty years ago, John Chambers, Bill Cleveland and Leo Breiman independently once again urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics; Chambers called for more emphasis on data preparation and presentation rather than statistical modeling; and Breiman called for emphasis on prediction rather than inference. Cleveland even suggested the catchy name “Data Science” for his envisioned field.

A recent and growing phenomenon is the emergence of “Data Science” programs at major universities, including UC Berkeley, NYU, MIT, and most recently the Univ. of Michigan, which on September 8, 2015 announced a $100M “Data Science Initiative” that will hire 35 new faculty. Teaching in these new programs has significant overlap in curricular subject matter with traditional statistics courses; in general, though, the new initiatives steer away from close involvement with academic statistics departments.

This paper reviews some ingredients of the current “Data Science moment”, including recent commentary about data science in the popular media, and about how/whether Data Science is really different from Statistics.

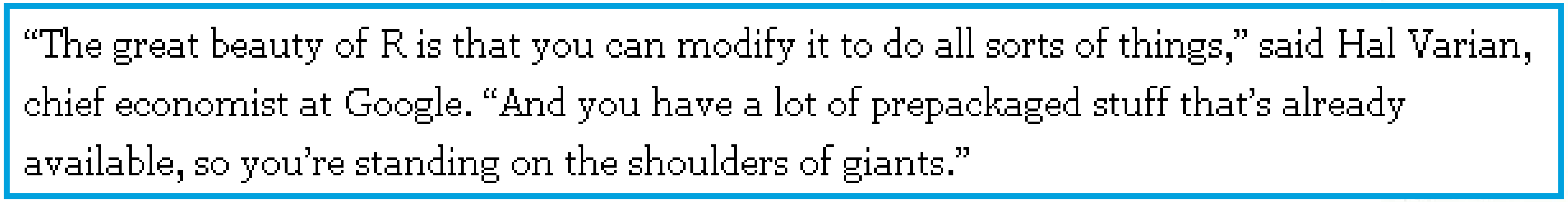
# Just one letter, young man


THE IDEA FACTORY
Bell Labs and the Great Age of American Innovation
Jon Gertner

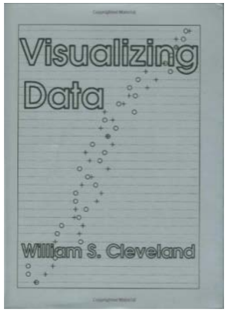
*R is an open-source, object-oriented statistical programming language. In the past decade, it has become the global lingua franca of statistics.*

History:

- Original S language developed by John Chambers at Bell labs
- R is an open-source implementation of the S language
- Developed by Robert Gentlemen and Ross Ihaka at U Auckland

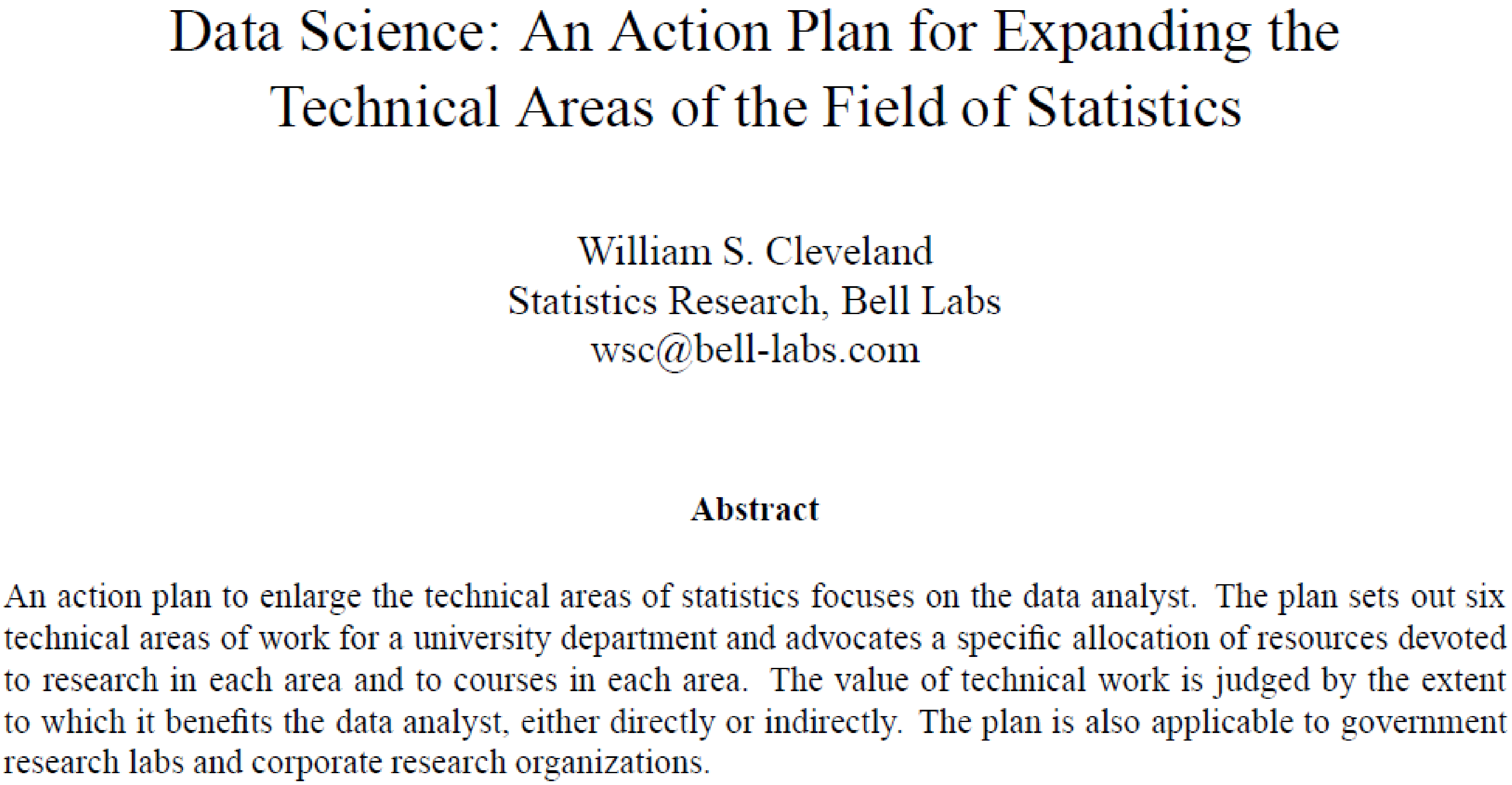

"The great beauty of R is that you can modify it to do all sorts of things," said Hal Varian, chief economist at Google. "And you have a lot of prepackaged stuff that's already available, so you're standing on the shoulders of giants."

# The origin of “data science”



Visualizing Data

William S. Cleveland



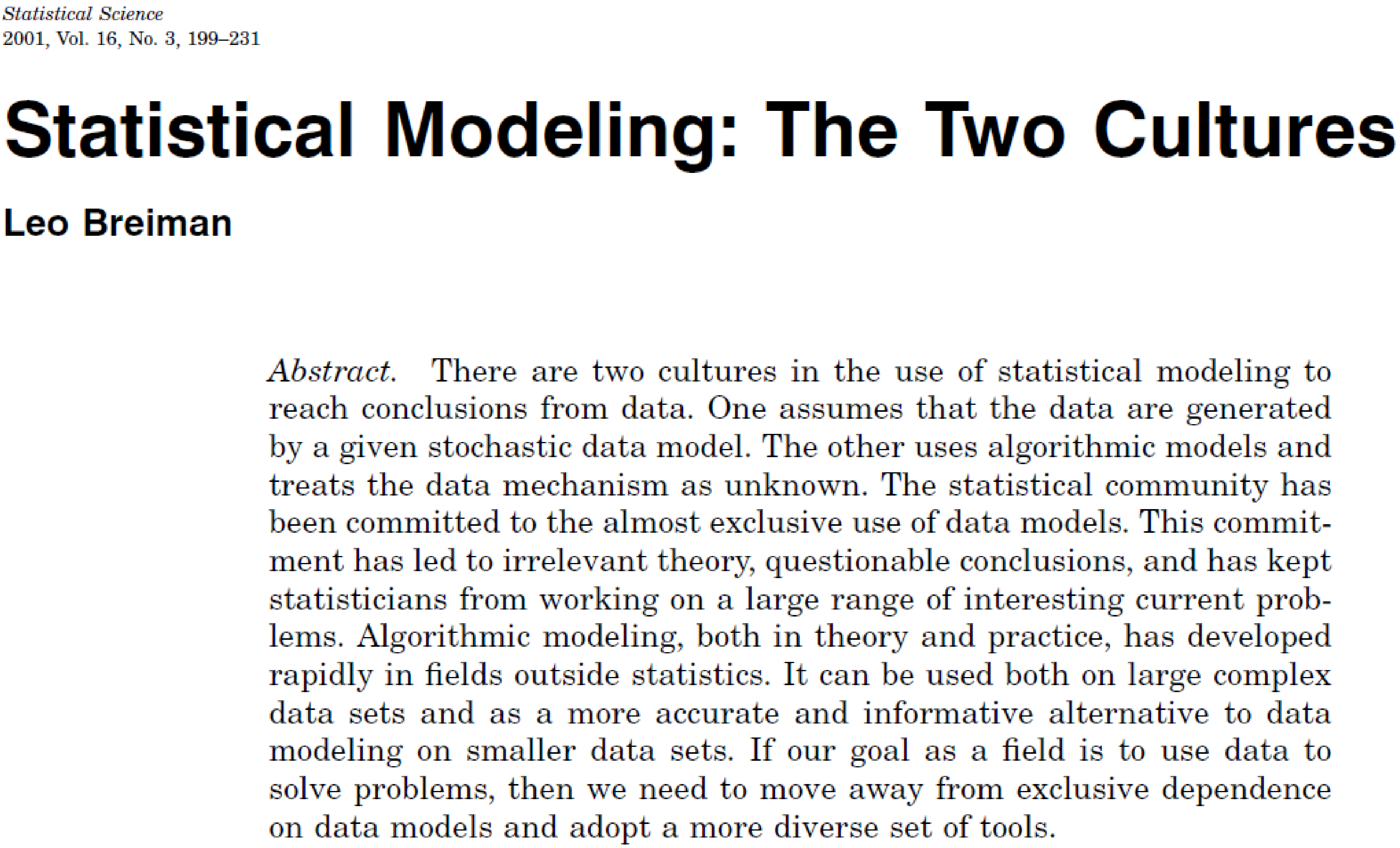
## Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

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.

# A provocative – and prescient – essay

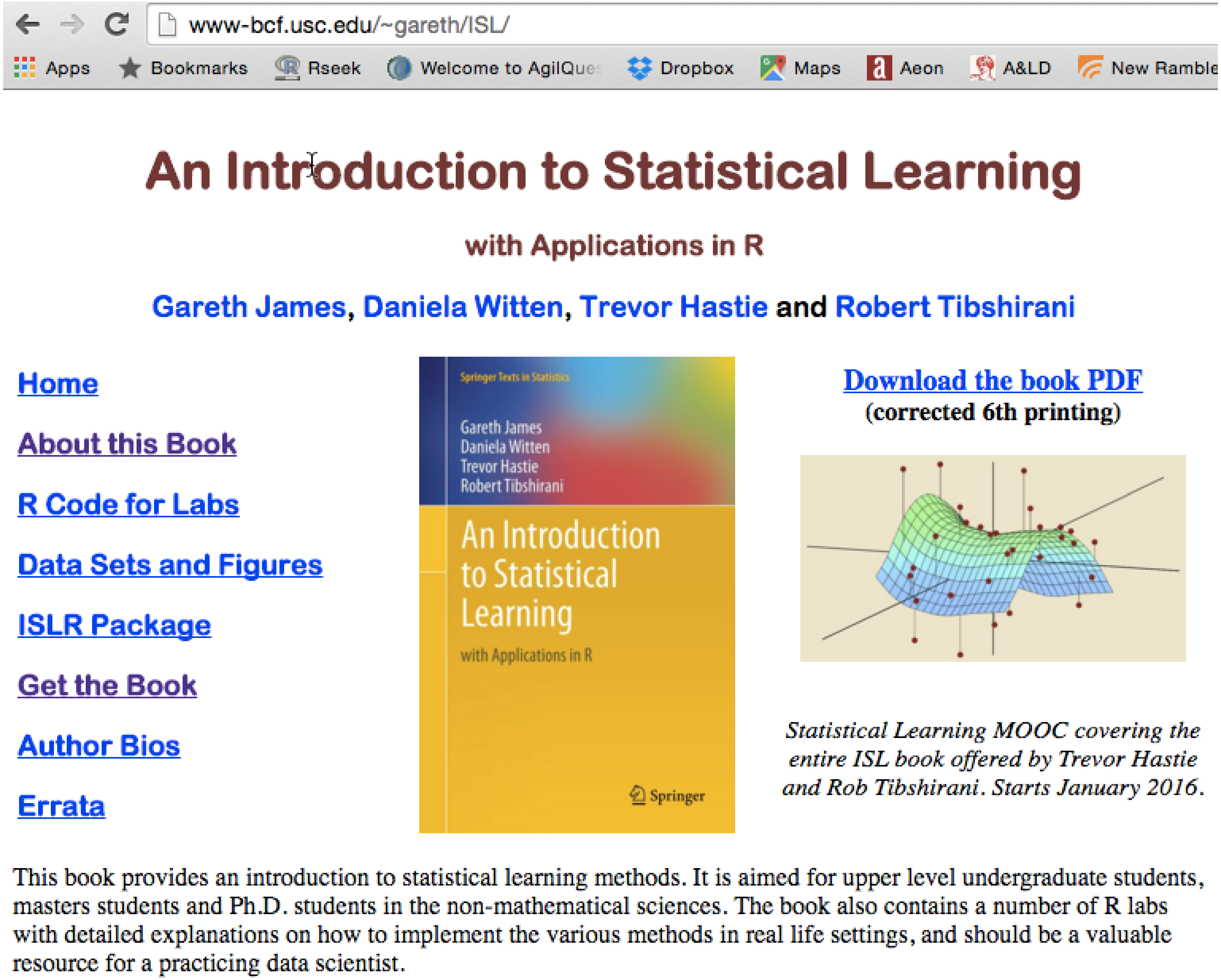


*Statistical Science*
2001, Vol. 16, No. 3, 199–231

## Statistical Modeling: The Two Cultures

**Leo Breiman**

*Abstract.* There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.



# An Introduction to Statistical Learning

## with Applications in R

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

- Home
- About this Book
- R Code for Labs
- Data Sets and Figures
- ISLR Package
- Get the Book
- Author Bios
- Errata

**Download the book PDF**
**(corrected 6th printing)**

*Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani. Starts January 2016.*

This book provides an introduction to statistical learning methods. It is aimed for upper level undergraduate students, masters students and Ph.D. students in the non-mathematical sciences. The book also contains a number of R labs with detailed explanations on how to implement the various methods in real life settings, and should be a valuable resource for a practicing data scientist.

# The culture of data science

"The best thing about being a statistician is that you get to play in everyone's back yard."

-- John Tukey<br>
Princeton/Bell Labs

# The culture of data science

"The best thing about being a statistician is that you get to play in everyone's back yard."

-- John Tukey
Princeton/Bell Labs

"The dominant trait among data scientists is an intense curiosity… This often entails the associative thinking that characterizes the most creative scientists in any field."

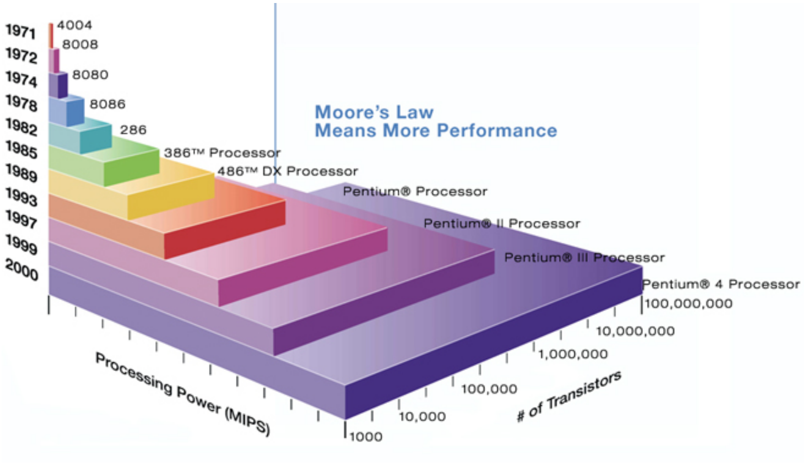
-- D.J. Patil

# Why analytics is everywhere

# The technological answer (“Moore, Moore, Moore”)

## *Technology (Moore’s Law)*

- Cost of storage and computing power has decreased exponentially



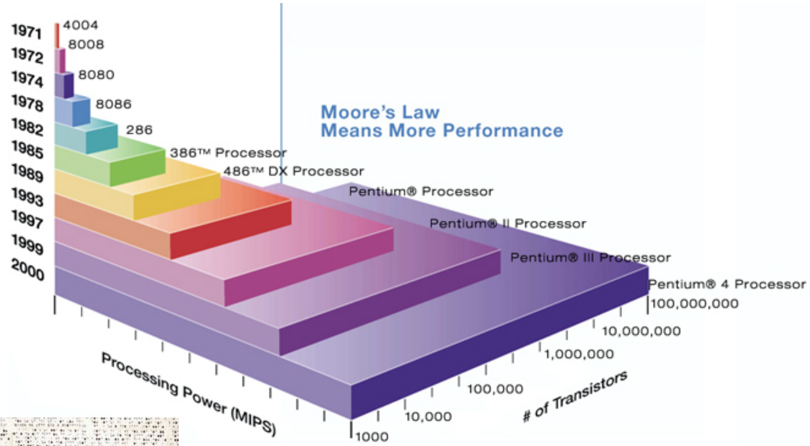
# The technological answer (“Moore, Moore, Moore”)

## *Technology (Moore’s Law)*

- Cost of storage and computing power has decreased exponentially

## *Data*

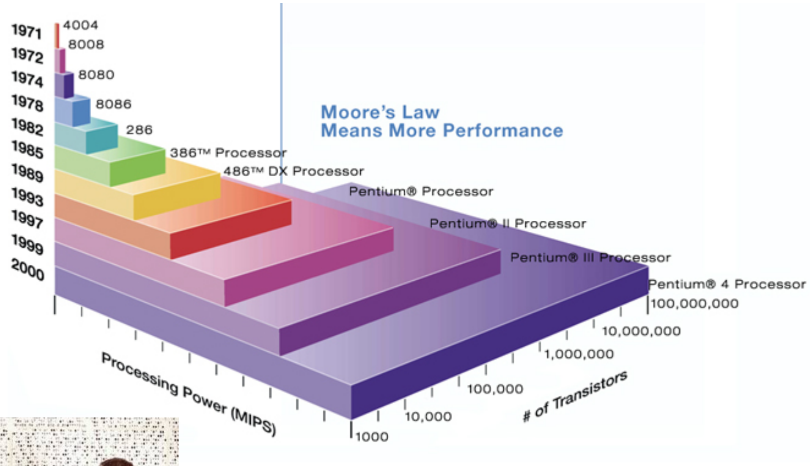
- It’s everywhere
- Mobile devices, the internet of things, cloud computing, …



# The technological answer (“Moore, Moore, Moore”)

## *Technology (Moore’s Law)*

- Cost of storage and computing power has decreased exponentially

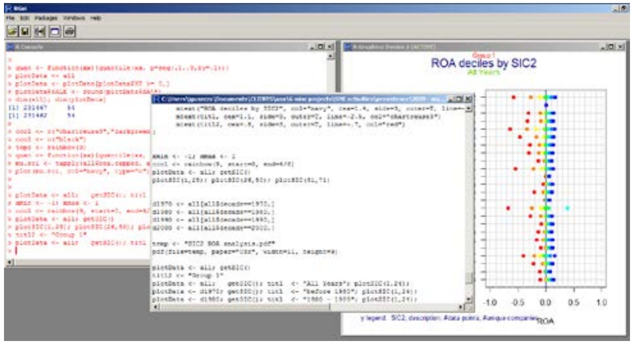


## *Data*

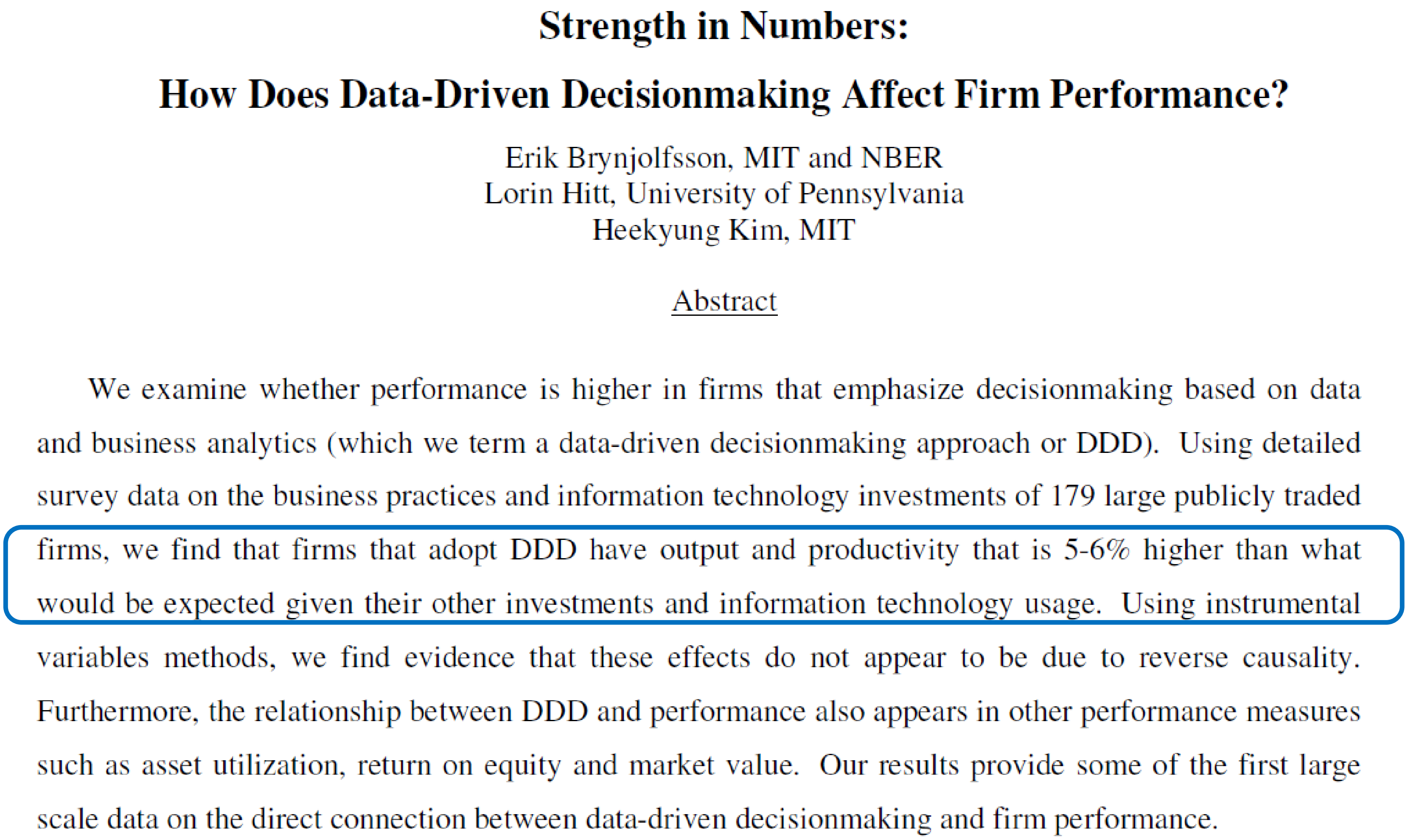
- It’s everywhere
- Mobile devices, the internet of things, cloud computing, …

## *Software and algorithms*

- Great analytic ideas keep coming from statistics, economics, machine learning, marketing, …
- Free tools like R, Python



# The business answer



## Strength in Numbers:

## How Does Data-Driven Decisionmaking Affect Firm Performance?

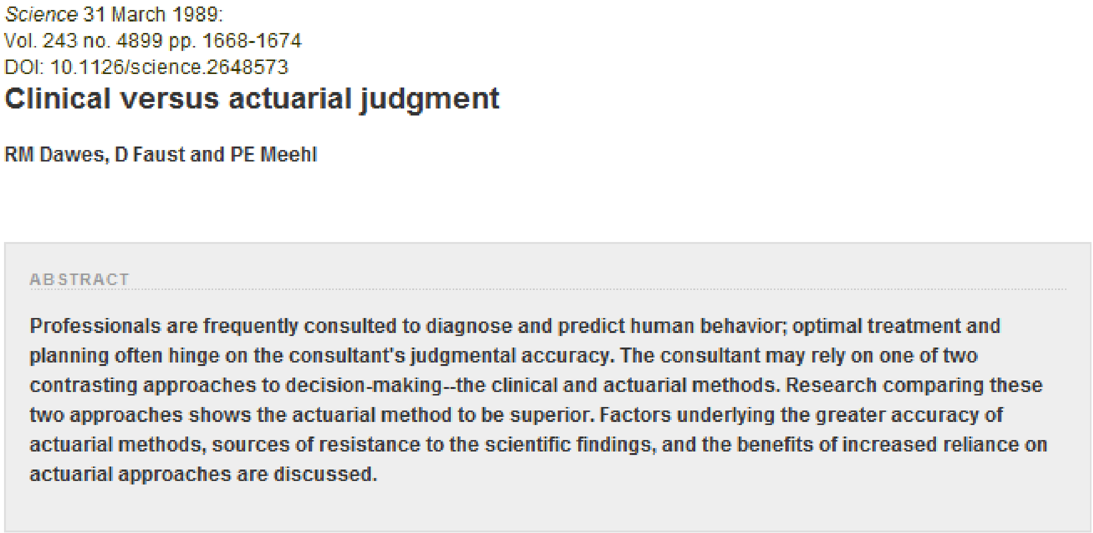
Erik Brynjolfsson, MIT and NBER
Lorin Hitt, University of Pennsylvania
Heekyung Kim, MIT

Abstract

We examine whether performance is higher in firms that emphasize decisionmaking based on data and business analytics (which we term a data-driven decisionmaking approach or DDD). Using detailed survey data on the business practices and information technology investments of 179 large publicly traded firms, we find that firms that adopt DDD have output and productivity that is 5-6% higher than what would be expected given their other investments and information technology usage. Using instrumental variables methods, we find evidence that these effects do not appear to be due to reverse causality. Furthermore, the relationship between DDD and performance also appears in other performance measures such as asset utilization, return on equity and market value. Our results provide some of the first large scale data on the direct connection between data-driven decisionmaking and firm performance.

# “Clinical Versus Actuarial Judgment:  the Motion Picture”

> “Human judges are not merely worse than optimal regression equations; they are worse than almost any regression equation.”
>
> – Richard Nisbett and Lee Ross



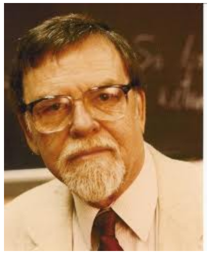
*Science* 31 March 1989:
Vol. 243 no. 4899 pp. 1668-1674
DOI: 10.1126/science.2648573

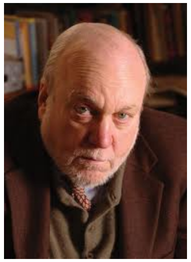
## Clinical versus actuarial judgment

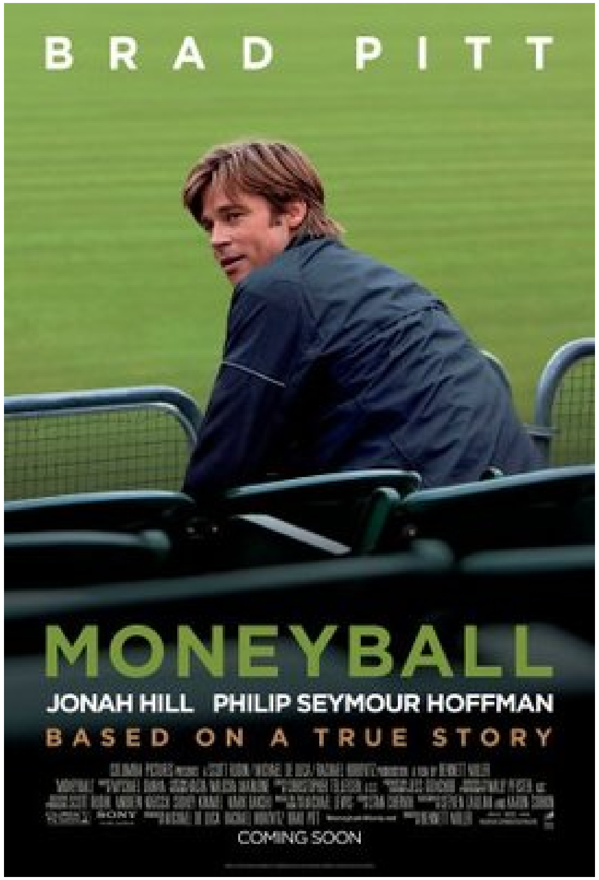
RM Dawes, D Faust and PE Meehl

ABSTRACT

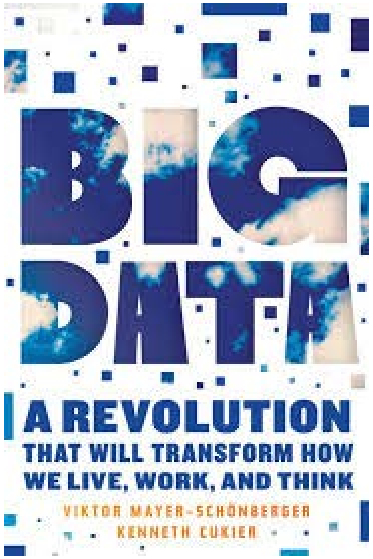
Professionals are frequently consulted to diagnose and predict human behavior; optimal treatment and planning often hinge on the consultant's judgmental accuracy. The consultant may rely on one of two contrasting approaches to decision-making--the clinical and actuarial methods. Research comparing these two approaches shows the actuarial method to be superior. Factors underlying the greater accuracy of actuarial methods, sources of resistance to the scientific findings, and the benefits of increased reliance on actuarial approaches are discussed.







# The city of New York does actuarial science



BIG DATA

A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK

VIKTOR MAYER-SCHÖNBERGER
KENNETH CUKIER

Big Data in the Big Apple

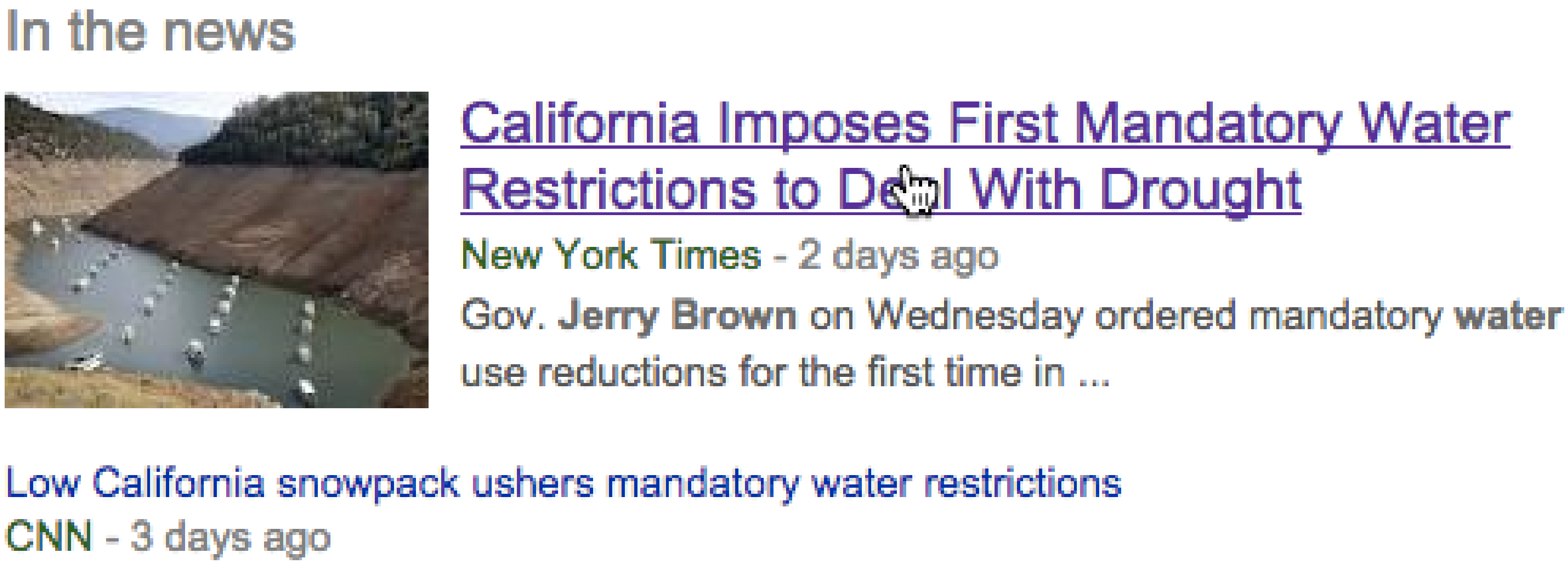
How New York's first "director of analytics" revolutionized the city's building inspections.

407

By Viktor Schönberger and Kenneth Cukier

A new way to figure out which old buildings are most at risk

# Dry idea



In the news

California Imposes First Mandatory Water Restrictions to Deal With Drought

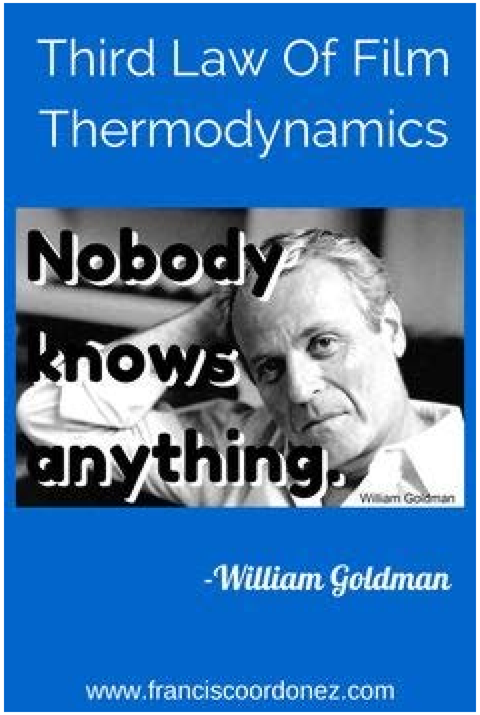
New York Times - 2 days ago

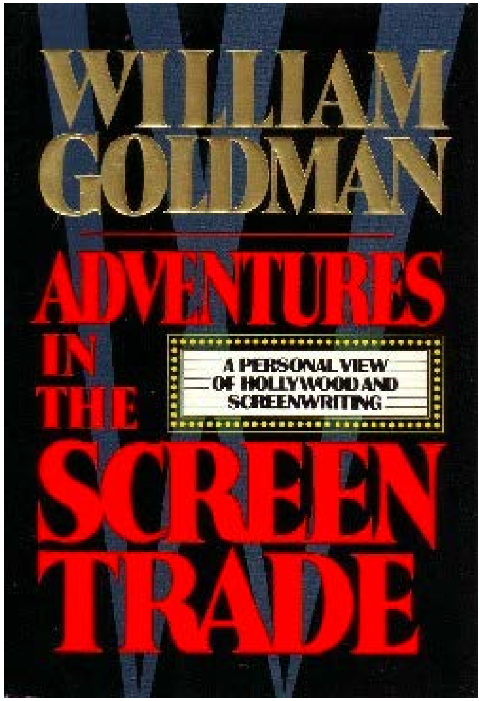
Gov. **Jerry Brown** on Wednesday ordered mandatory **water** use reductions for the first time in ...

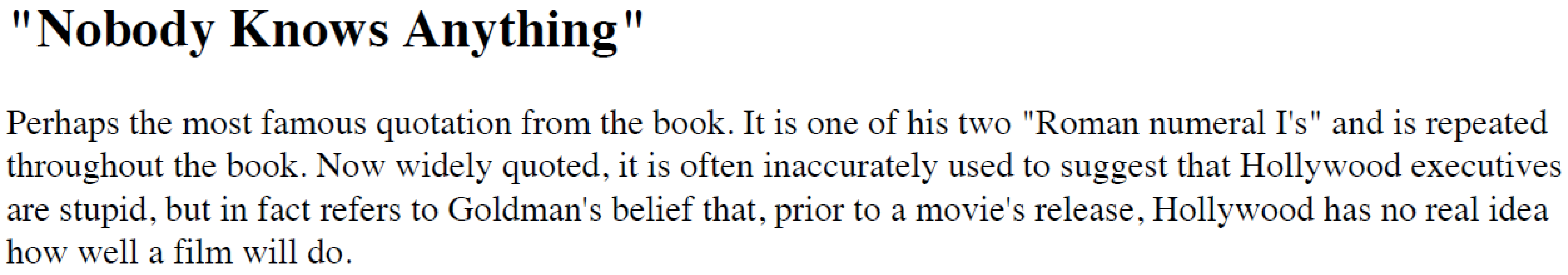
Low California snowpack ushers mandatory water restrictions

CNN - 3 days ago

# In Hollywood, “Nobody Knows Anything”




WILLIAM GOLDMAN
ADVENTURES IN THE SCREEN TRADE
A PERSONAL VIEW OF HOLLYWOOD AND SCREENWRITING



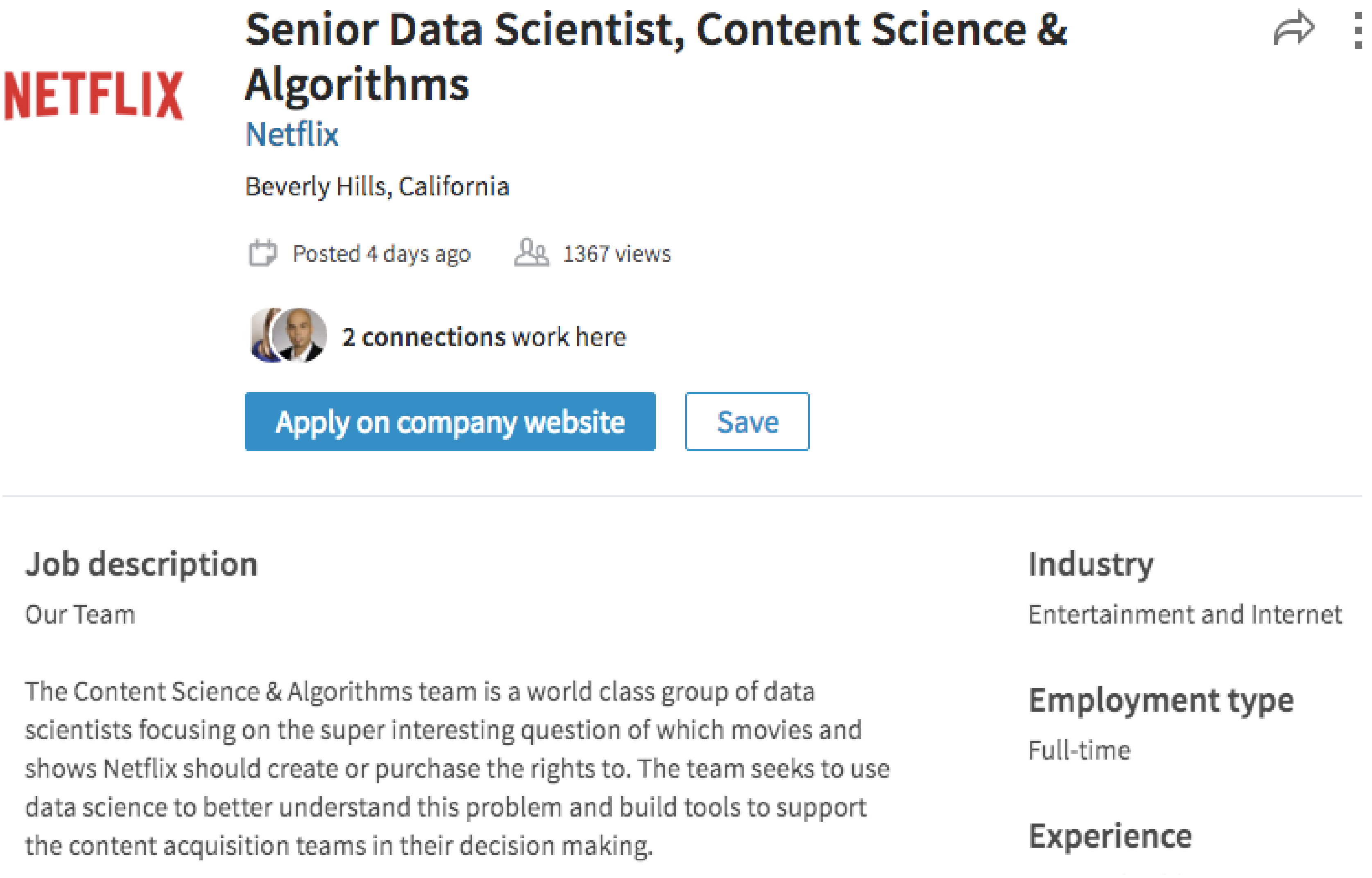
## "Nobody Knows Anything"

Perhaps the most famous quotation from the book. It is one of his two "Roman numeral I's" and is repeated throughout the book. Now widely quoted, it is often inaccurately used to suggest that Hollywood executives are stupid, but in fact refers to Goldman's belief that, prior to a movie's release, Hollywood has no real idea how well a film will do.

# (except Netflix)

*Better viewing through “datafication”*

# … but I really want to direct …



Senior Data Scientist, Content Science & Algorithms

Netflix

Beverly Hills, California

Posted 4 days ago 1367 views

2 connections work here

Job description

Our Team

The Content Science & Algorithms team is a world class group of data scientists focusing on the super interesting question of which movies and shows Netflix should create or purchase the rights to. The team seeks to use data science to better understand this problem and build tools to support the content acquisition teams in their decision making.

Industry

Entertainment and Internet

Employment type

Full-time

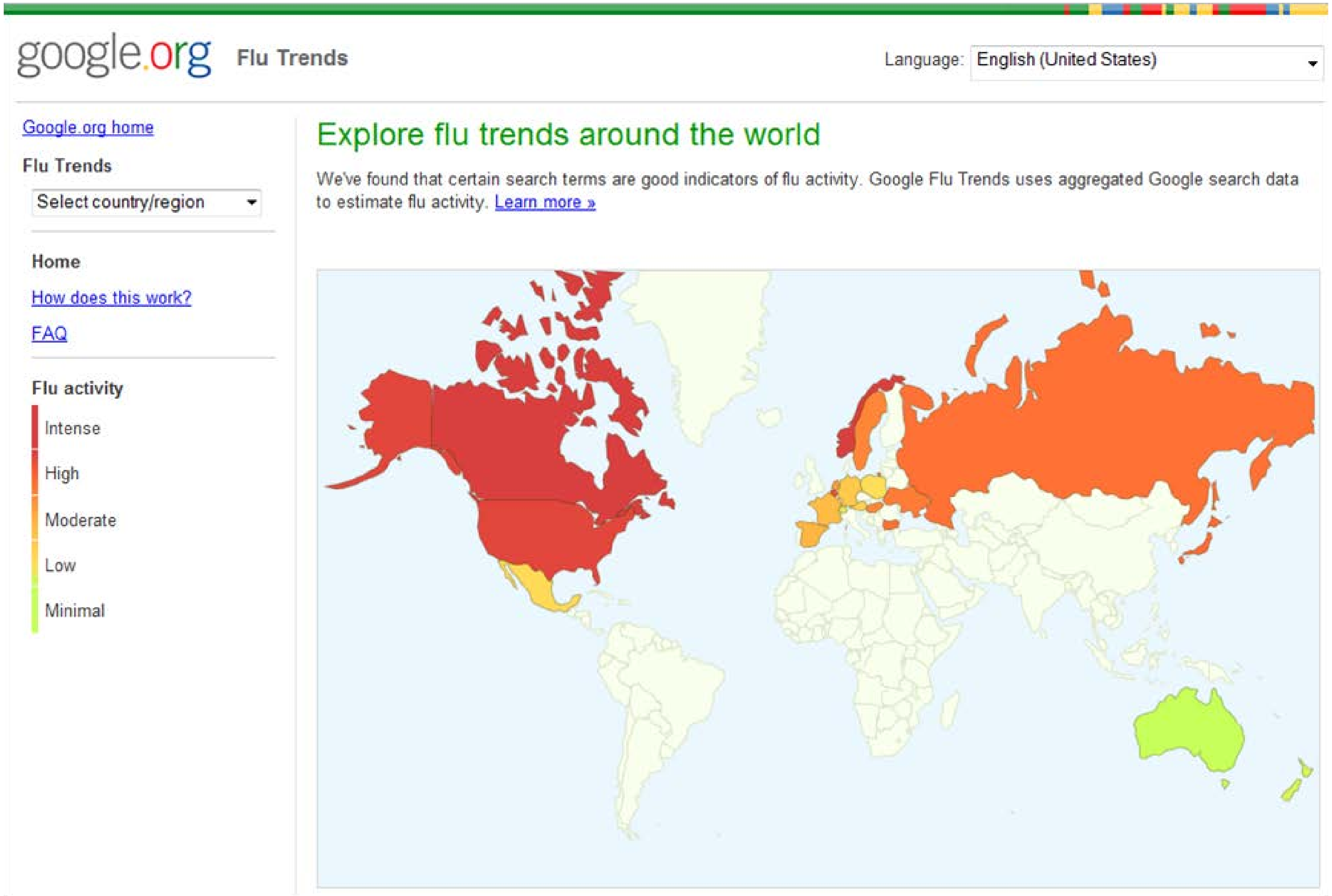
Experience

# Marketing science meets political science

Harper Reed - The CTO of the 2012 Obama re-election campaign
(pictured with unidentified staffer)

# Big data and behavioral data

# The pulse of the nation



google.org Flu Trends

Language: English (United States)

Google.org home

Flu Trends

Select country/region

Home

How does this work?

FAQ

Flu activity

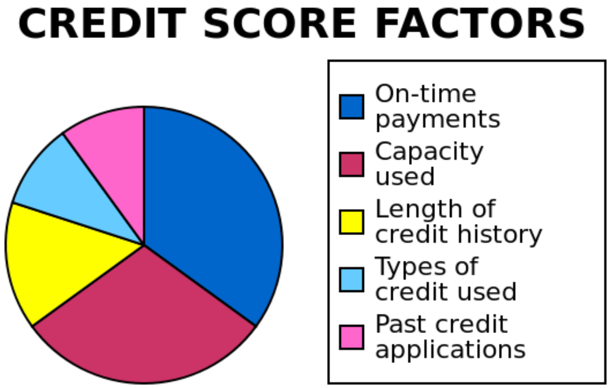
- 🟥 Intense
- 🟥 High
- 🟧 Moderate
- 🟨 Low
- 🟩 Minimal

## Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

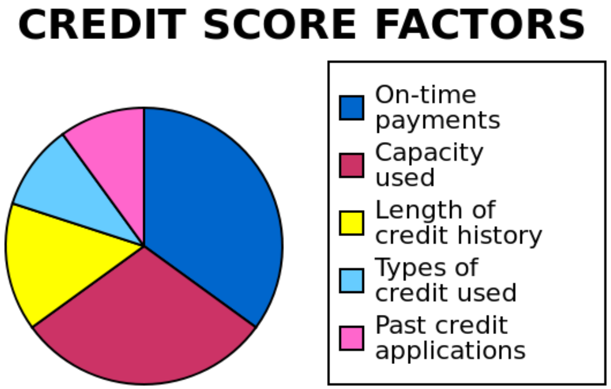
*Google data repurposed to track flu hot-spots*

# An early example of business analytics

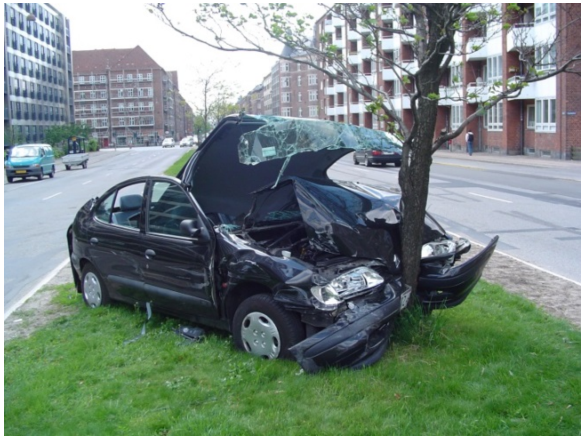


(This we know)

# A more striking correlation



(!)



# More food for thought

(!!)

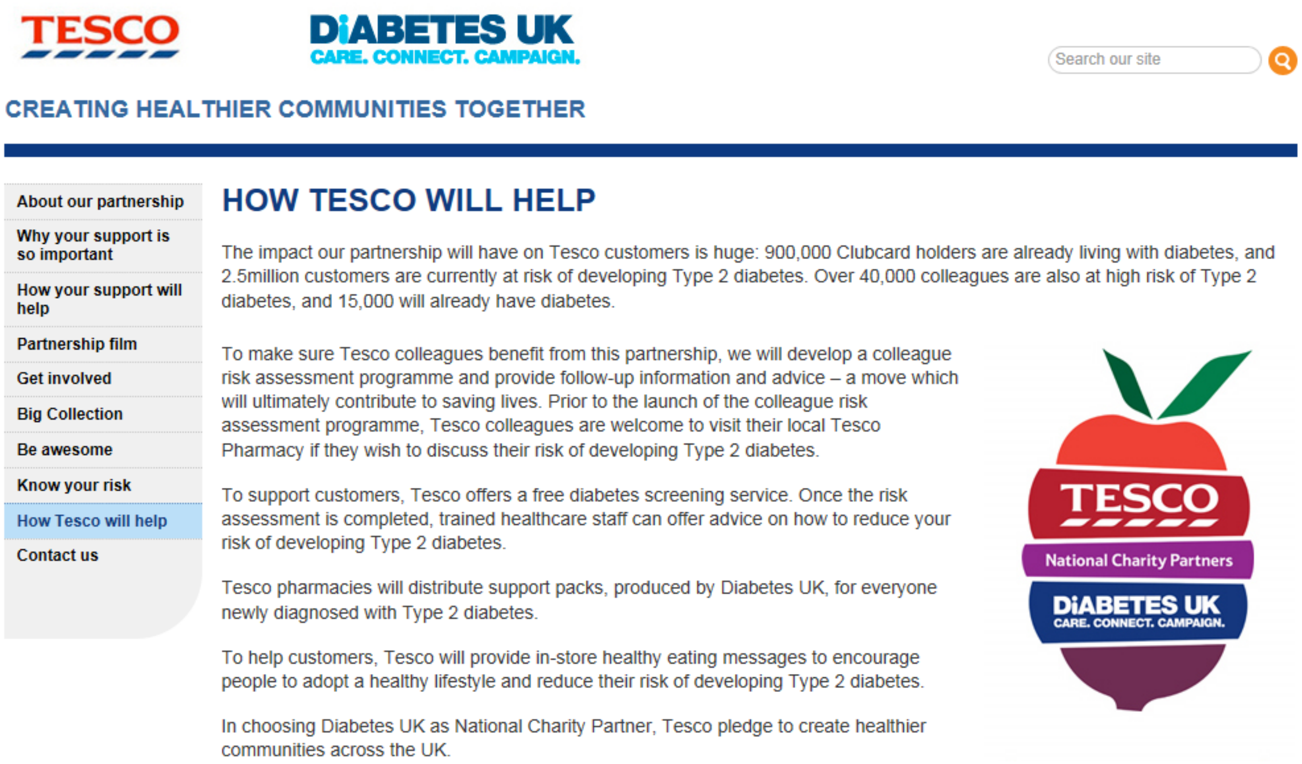
# And also…

☹

# Behavioral data predicts behavioral disease states

*Tesco has investigated the use of clubcard data to predict which of its customers is at highest risk of diabetes*

http://www.diabetes.org.uk/Tesco/How-Tesco-will-help/



CREATING HEALTHIER COMMUNITIES TOGETHER

- About our partnership
- Why your support is so important
- How your support will help
- Partnership film
- Get involved
- Big Collection
- Be awesome
- Know your risk
- How Tesco will help
- Contact us

## HOW TESCO WILL HELP

The impact our partnership will have on Tesco customers is huge: 900,000 Clubcard holders are already living with diabetes, and 2.5million customers are currently at risk of developing Type 2 diabetes. Over 40,000 colleagues are also at high risk of Type 2 diabetes, and 15,000 will already have diabetes.

To make sure Tesco colleagues benefit from this partnership, we will develop a colleague risk assessment programme and provide follow-up information and advice – a move which will ultimately contribute to saving lives. Prior to the launch of the colleague risk assessment programme, Tesco colleagues are welcome to visit their local Tesco Pharmacy if they wish to discuss their risk of developing Type 2 diabetes.

To support customers, Tesco offers a free diabetes screening service. Once the risk assessment is completed, trained healthcare staff can offer advice on how to reduce your risk of developing Type 2 diabetes.

Tesco pharmacies will distribute support packs, produced by Diabetes UK, for everyone newly diagnosed with Type 2 diabetes.

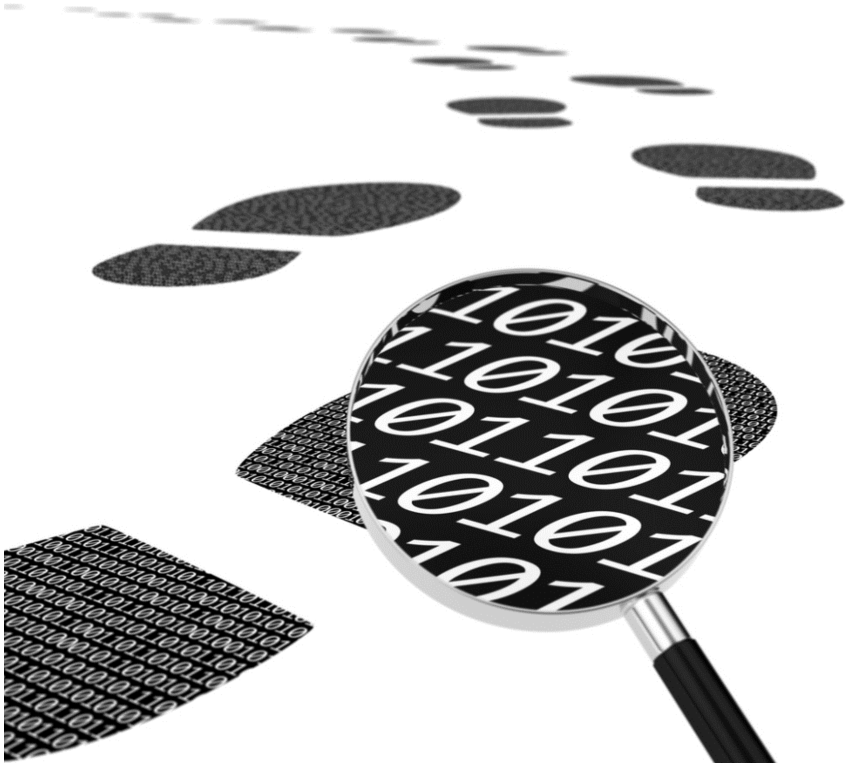
To help customers, Tesco will provide in-store healthy eating messages to encourage people to adopt a healthy lifestyle and reduce their risk of developing Type 2 diabetes.

In choosing Diabetes UK as National Charity Partner, Tesco pledge to create healthier communities across the UK.

# We are part of the “internet of things”

➔ In our daily lives, we increasingly leave behind digital traces of:

- How we drive
- What we buy
- What we eat
- What we watch, read
- What / how we opine
- Where we travel
- Whom we know / networks
- How we socialize
- How we surf the web



*The resulting data can be a major source of operational improvements and business innovation…*

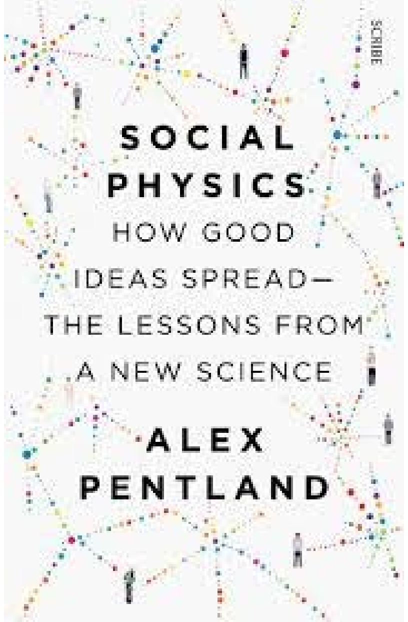
*… and societal change…*

# Why big data is a big deal

“I believe that the power of Big Data is that it is information about people's behavior instead of information about their beliefs… It's not about the things you post [online] … which is what most people think about, and it's not data from internal company processes and RFIDs. This sort of Big Data comes from things like location data off of your cell phone or credit card, it's the little **data breadcrumbs** that you leave behind you as you move around in the world.

…those breadcrumbs tell… the story of your life... Big data is increasingly about real behavior, and by analyzing this sort of data, scientists can tell an enormous amount about you. They can tell whether you are the sort of person who will pay back loans. They can tell you if you're likely to get diabetes”

—Sandy Pentland, MIT Media Lab
“Reinventing Society in the Wake of Big Data”
edge.org conversation



# Like, you know

Researchers at the Cambridge University Psychometrics Centre built predictive models of personal details based purely on social network “Likes” of a sample of 58,000 people.

| | |
|---|---|
| • Relationship status, substance abuse | 65-73% accurate |
| • Political leanings (democrat vs Republican) | 85% accurate |
| • Religion (Christian vs Muslim) | 82% accurate |
| • Male sexual orientation | 88% accurate |
| • Ethnicity (African-American vs Caucasian) | 95% accurate |

“Observation of Likes alone was nearly is roughly as informative as using an individual’s actual personality test score.”

“Similar predictions could be made from all manner of digital data, with this kind of secondary ‘inference’ made with remarkable accuracy”

-- “Digital Records Could Expose Intimate Details and Personality Traits of Millions”
University of Cambridge Research News

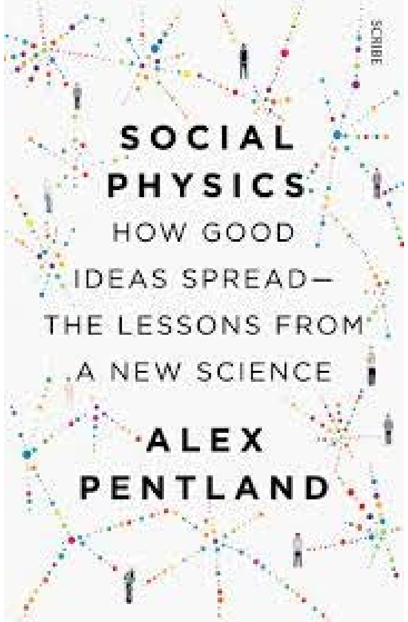
http://www.cam.ac.uk/research/news/digital-records-could-expose-intimate-details-and-personality-traits-of-millions

# A new mindset for data science

# Why big data it is a societal issue

“Since this data is mostly about people, there are enormous issues about privacy, data ownership, and data control. You can imagine using Big Data to make a world that is incredibly invasive, incredibly 'Big Brother'… **George Orwell was not nearly creative enough when he wrote *1984*.**”

—Sandy Pentland, MIT Media Lab
“Reinventing Society in the Wake of Big Data”
edge.org conversation

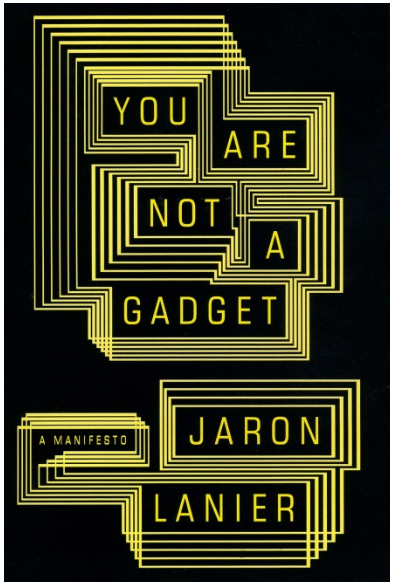


# Who owns the future?

“Information wants to be free.” – (Silicon valley mantra)

“A huge problem is the use of cloud computing to support the fantasy that information is alive in its own right, and that the activities or expressions of individual people are nothing but one form of computing resource, targeted for aggregation.  This is, unfortunately, an approximate statement of the latest ideology that has taken hold of the cloud.”

– Jaron Lanier (virtual reality pioneer and human reality advocate)



# Can we do better by doing good?

*"The best minds of my generation are thinking about how to make people click ads… that sucks."*

-- Jeff Hammerbacher
Cloudera Founder

*"For all the damage that misapplied data can do, data used correctly is a powerful **positive force**."*

-- Cathy O'Neil, mathbabe.org
"On Being a Data Skeptic"

# What are companies for?

"There is one and only one social responsibility of business – to use its resources and engage in activities designed to increase its profits… so long as it engages on open and free competition without deception or fraud."

-- **Milton Friedman**

"The Social Responsibility of Business is to Increase its Profits"

"The only valid purpose of a firm is to create a customer."

-- **Peter Drucker**

*Management: Tasks, Responsibilities, Practices*

# What are companies for?

“There needs to be a completely new approach to how we operate as business leaders, one that clearly puts people at the centre of all we do.”

-- **B Team Leadership Statement**

Davos, January 22, 2014

# Copies available in the lobby

For the full story, go to:

http://dupress.com/articles/dr14-personalized-and-personal/

Deloitte Review

ISSUE 14 | 2014

Beyond Yahoo
The "virtual" versus "campus" debate

A Design for Life
An hour with Alain de Botton

3D Opportunity
Additive manufacturing paths to performance, innovation, and growth

The Personalized and the Personal
Socially responsible innovation through big data