Data Science in Finance

Algorithms, Analytics, Applications, Big Data

JOIM

September 2014

The Art of Data Science

— *"All models are wrong, but some are useful."*
George E. P. Box and N.R. Draper in "Empirical Model Building and Response Surfaces," John Wiley & Sons, New York, 1987.

[1] The term "data scientist" was coined by D.J. Patil. He was the Chief Scientist for LinkedIn. In 2011 Forbes placed him second in their Data Scientist List, just behind Larry Page of Google.

McKinsey estimates a data scientist shortage of at least a quarter million ….

Big Data – the Human Side

- Google's Eric Schmidt: Until 2003 just 5 exabytes of human data; Now: we generate 5 exabytes every 2 days!

Multiples of bytes V·T·E

Decimal			Binary		
1000^4	TB	terabyte	1024^4	TiB	tebibyte
1000^5	PB	petabyte	1024^5	PiB	pebibyte
1000^6	EB	**exabyte**	1024^6	EiB	exbibyte
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte
1000^8	YB	yottabyte	1024^8	YiB	yobibyte

Orders of magnitude of data

We leave behind "digital exhaust" – phone calls, text, emails, browser history, GPS data, cookies. Used through satellites, sensors, RFID, GPS, phone, fitbit, etc.

The average person processes more data in one day than a person did in a lifetime in the 1500s.

Data and Business Transformation

- Companies are using medical data and claims data to offer incentivized health programs to employees. Caesar's Entertainment Corp. analyzed data for 65,000 employees and found substantial cost savings.

- Zynga Inc, famous for its game Farmville, accumulates 25 terabytes of data every day and analyzes it to make choices about new game features.

- UPS installed sensors to collect data on speed and location of its vans, which combined with GPS information, reduced fuel usage in 2011 by 8.4 million gallons, and shaved 85 million miles off its routes.

- In a seminal paper, Halevy, Norvig and Pereira (2009), argue that even simple theories and models, with big data, have the potential to do better than complex models with less data. ("Big Data: What's Your Plan?" McKinsey Quarterly, March 2013.)

Big Data – The Human Side

- Why Progressive +
- Ways to Save −
  - Discounts
  - Comparison Rates
  - Name Your Price® Tool
  - Snapshot®
  - Bundle
- Coverages +

Take **Snapshot**® for a test drive, then decide if you want to switch

Preview what you'd save

Zip Code

Sign Up Today

Test drive Snapshot—and all the benefits of personalizing your insurance—in just 30 days.

PLUG IN Preview your projected discount after plugging in the Snapshot device, and watch as it evolves based on your latest driving.

GET UPDATES See how often you slam on the brakes, how many miles you drive each day, and more.

SHARE Discover the best driver in your household ... and shout it out to your social networks!

SEE SAVINGS! See your personalized Progressive rate and decide if you want to switch your insurance.

Find out what makes Snapshot so different-and, dare we say, fun?

Watch Preview

PREDICTIVE ANALYTICS

TEXT ANALYTICS

BIG DATA

NETWORK ANALYTICS

NEWS ANALYTICS

CONTRIBUTED ARTICLES

Data Science and Prediction

By Vasant Dhar

Communications of the ACM, Vol. 56 No. 12, Pages 64-73
10.1145/2500499

Comments (2)

VIEW AS: SHARE:

Use of the term "data science" is increasingly common, as is "big data." But what does it mean? Is there something unique about it? What skills do "data scientists" need to be productive in a world deluged by data? What are the implications for scientific inquiry? Here, I address these questions from the perspective of predictive modeling.

Back to Top

Key Insights

- **Data science is the study of the generalizable extraction of knowledge from data.**
- **A common epistemic requirement in assessing whether new knowledge is actionable for decision making is its predictive power, not just its ability to**

The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

ARTICLE

Measuring Readability in Financial Disclosures

TIM LOUGHRAN and BILL MCDONALD†

Article first published online: 18 JUL 2014
DOI: 10.1111/jofi.12162

Issue

The Journal of Finance
Volume 69, Issue 4, pages 1643–1671, August 2014

Am score

Additional Information (Show All)

How to Cite | Author Information | Publication History

Abstract | Article | References | Supporting Information | Cited By

View Full Article with Supporting Information (HTML) | Enhanced Article (HTML) | Get PDF (272K)

ABSTRACT

Defining and measuring readability in the context of financial disclosures becomes important with the increasing use of textual analysis and the Securities and Exchange Commission's plain English initiative. We propose defining readability as the effective communication of valuation-relevant information. The Fog Index—the most commonly applied readability measure—is shown to be poorly specified in financial applications. Of Fog's two components, one is misspecified and the other is difficult to measure. We report that 10-K document file size provides a simple readability proxy that outperforms the Fog Index, does not require document parsing, facilitates replication, and is correlated with alternative readability constructs.

Faculty & Research

Haas Home
Faculty & Research
Directory (Alphabetical)
Directory (By Group)
Faculty Photos
Executive Fellows
Nobel Laureates
Research Awards
Teaching Awards
Research News
Library
California Management Review
Open Academic Positions

Faculty and Executive Leadership Directory

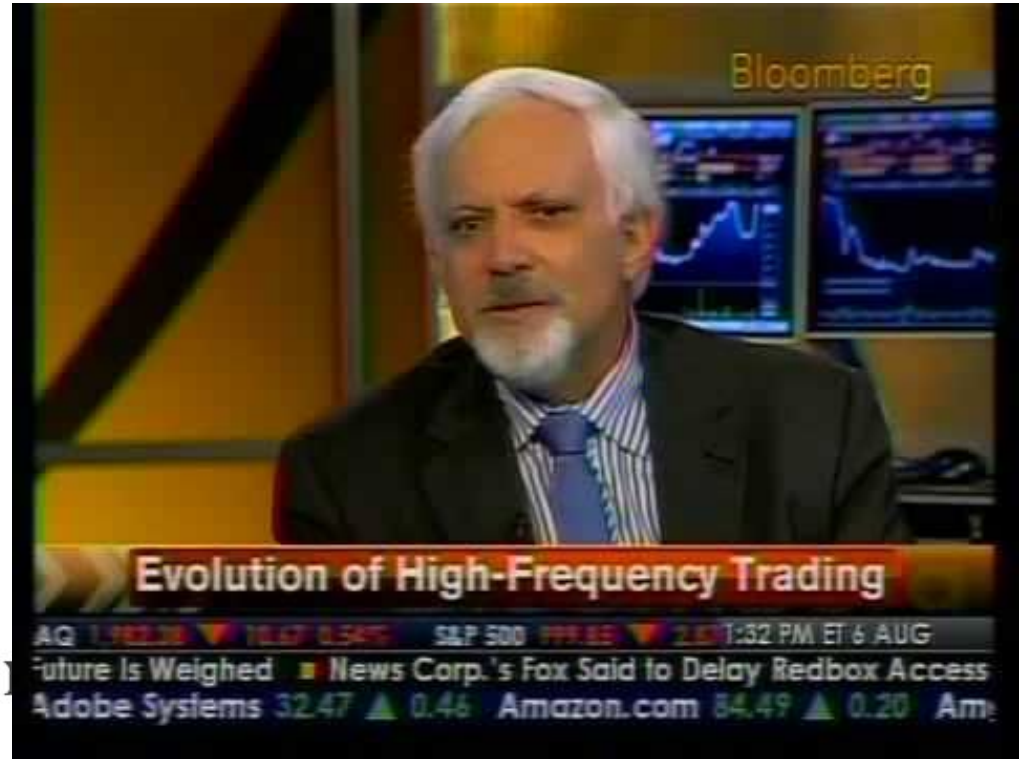
Nancy E. Wallace

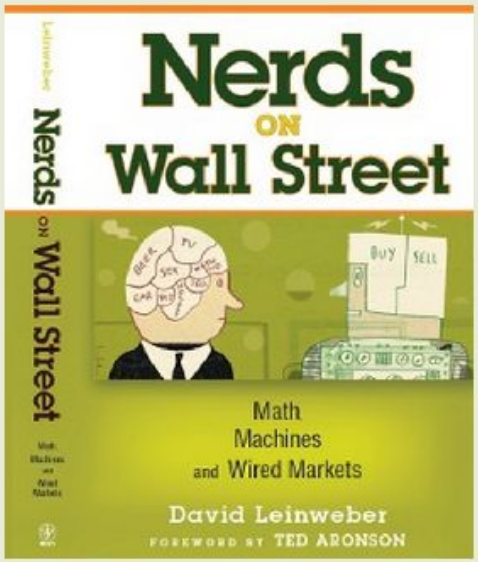
Lisle and Roslyn Payne Chair in Real Estate Capital Markets
Professor and Chair of the Real Estate Group
Co-Chair, Fisher Center for Real Estate and Urban Economics
Haas Real Estate Group
510-642-4732
Email: click on the envelope icon below for full email address
Academic Status: On duty
Office Hours: Tuesdays 2-3:30 p.m. F626
Curriculum Vitae (in PDF format, Acrobat Reader required)

Securitization Networks and the Endogenous Evolution of Financial Norms in Mortgage Markets

Nerds on Wall Street

Computational Finance





Dr. Leinweber has been a financial speaker throughout his thirty-year investment management career. He gives custom financial presentations to organizations that want to hear from one of the best financial speakers on the speaking circuit today.

The Impact of Technology on Wall Street and Investing

Event Driven Trading and the "New News"

The Economist

Banks and big data

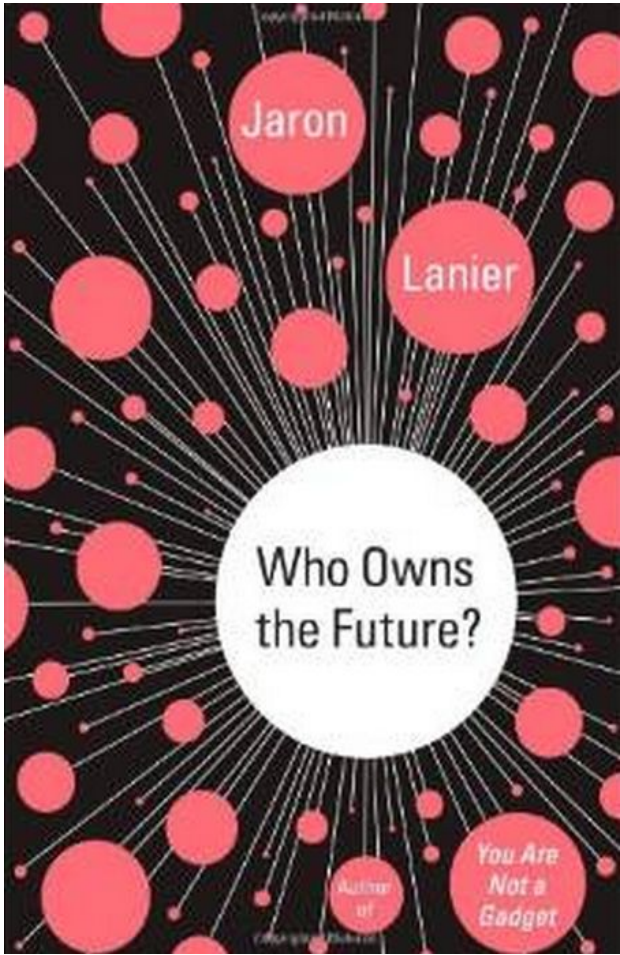
Shopping at the bank

It is harder to make money from banking. How about marketing?

Oct 27th 2012 | from the print edition

"A leader in this field is Cardlytics, a private American company founded in 2008. It has developed technology to analyze transaction data held by banks and to use this information to sell targeted advertisements to retailers and others. A supermarket might, for example, be interested in customers who spend $100 or more a month at rival grocers but who have not entered its own stores for six months. It might then offer these people a 20% discount on their next shopping trip at its stores."

Cardlytics inserts an advertisement into these customers' online bank statements, ideally under a relevant transaction such as a payment to a rival retailer. Customers accept the discount online by clicking a box. This connects the discount coupon to their debit card, and the discount is automatically rebated to their account when they shop at the store within a certain period of time).



IBM's Watson Is Learning Its Way To Saving Lives

By Jon Gertner

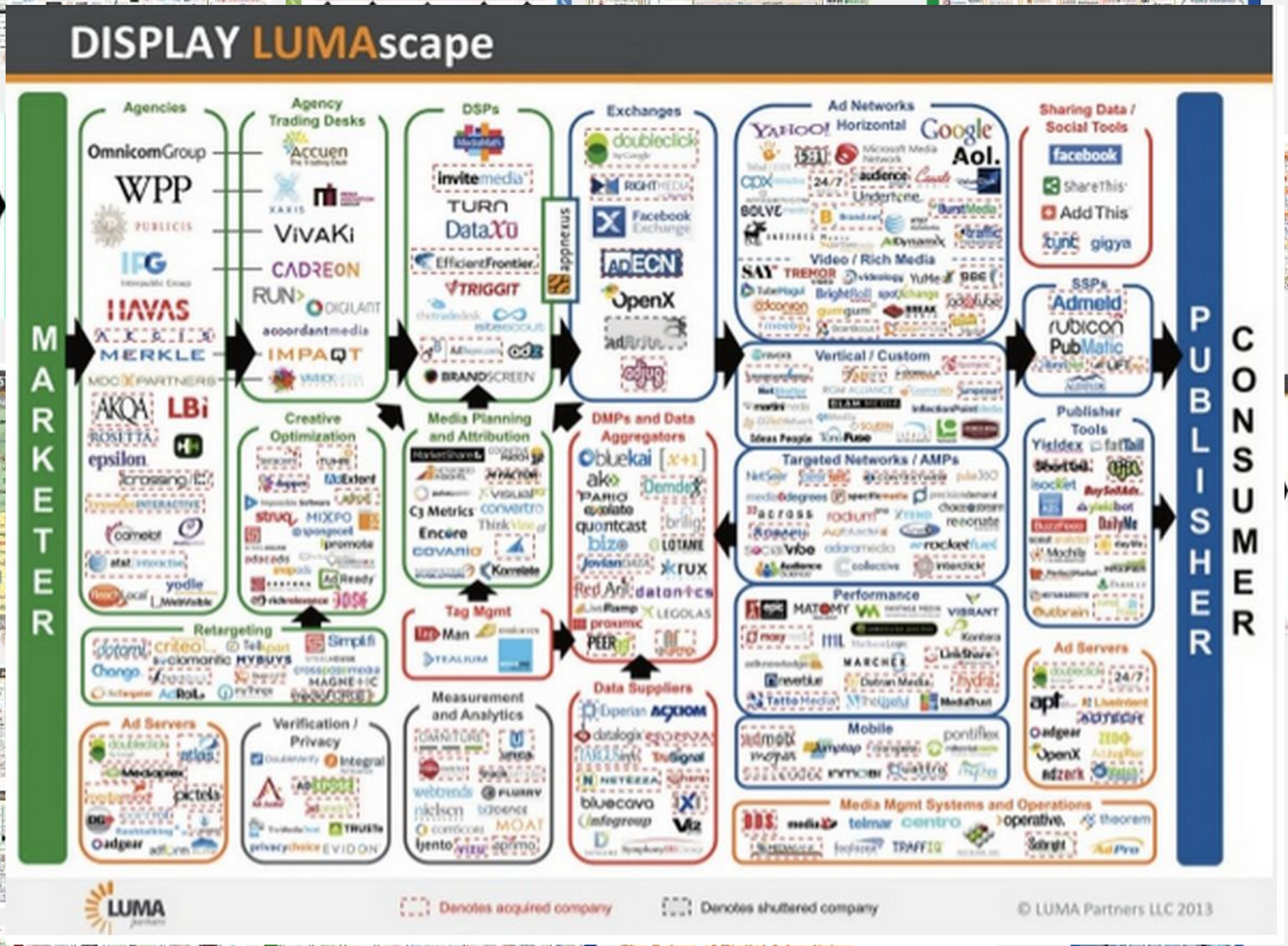
|

October 15, 2012

IBM expects $16BN in data science revenue by 2015
The “Big Server” race….

A few years ago, IBM’s new computer was a game-playing curiosity. Now Watson is poised to change the way human beings make decisions about medicine, finance, and work.

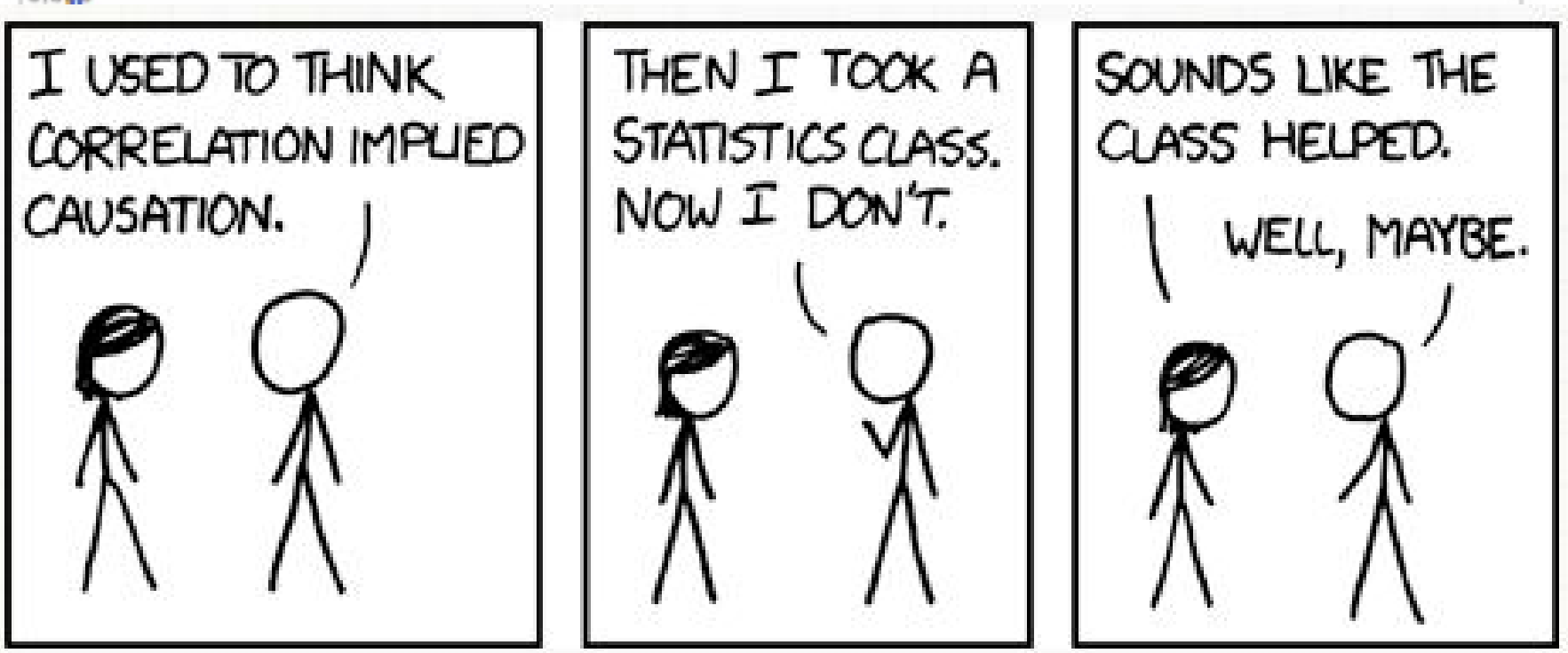
90% of the world’s information was created in the last two years,
80-90% of that information is in unstructured text.



The Future of Digital Advertising

Google Trends (Flu)

Correlation vs Causality



Machine Learning

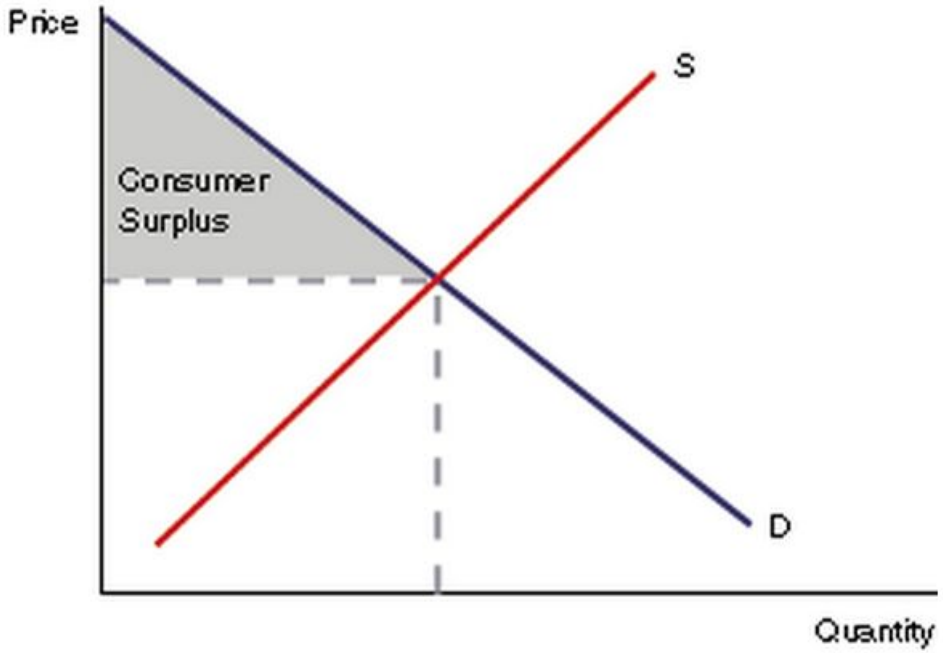
- How systems learn from data
- Spam filters (see Paul Graham's article, "A Plan for Spam): http://www.paulgraham.com/spam.html
- Neural nets for credit card approvals
- Vinny Bruzzese, known as the "mad scientist of Hollywood" who uses machine learning to predict movie revenues. ("Solving Equation of a Hit Film Script, With Data," New York Times, May 5, 2013.) Complement machine learning with judgment and interviews.

The Dark Side 1: Big Data, Big Errors

- Too much data leads to misleadingly significant relationships. Correlations may be overstated.
- Too many columns (p) and too few rows of data(n).
- Explore the data fully, don't stop when a significant result occurs, which is a problem with big data.
- Nassim Taleb describes these issues elegantly - "I am not saying there is no information in big data. There is plenty of information. The problem – the central issue – is that the needle comes in an increasingly larger haystack." ("Beware the Big Errors of Big Data" Wired, February 2013.)

Dark Side 2: Privacy

- Profiling (e.g., Groupon)
- Price discrimination
- Security vs Efficiency



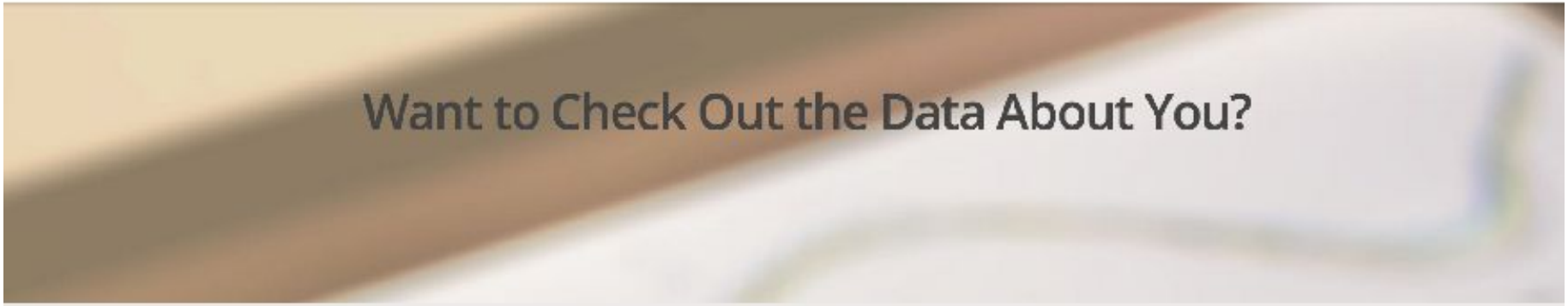
How do companies get data about me and what do they do with it?

Marketing data is collected about us every day by many entities. Most businesses collect some data about you to support their marketing efforts. It can come from commercial brands gathering information on their clients. These practices are typically described in their privacy policy. Data can come from publicly available sources such as city or state records or census data, or it can come from companies like Acxiom who collect and aggregate it from surveys, registrations, purchases, postings, etc. For organizations to make relevant offers to you, they need data to identify products and services you might be interested in.

LEARN MORE

What types of data do companies use about me?

Companies use lots of different kinds of data about you such as your age, whether you are married, single or divorced, what kind of vehicle you drive, whether you own your home or rent, etc. Generally speaking, data about you falls into two categories: core data and derived or modeled insights. Core data includes things like your address, phone number, age, etc. Derived and modeled insights are the result of analytical processes that use your core data to infer things about you such as whether or not you like sports cars or enjoy cooking. Both types of information are valuable to companies because it's often the only way that they can understand which products or services



Want to Check Out the Data About You?

See and Edit Marketing Data about You.

CLICK HERE

acxiom.

Acxiom is one of many companies that supplies businesses with marketing data to ensure that you are receiving offers you might be interested in. Want to learn more about Acxiom in particular?

Click here.

About Ads - the Digital Advertising Alliance's (DAA) Self-Regulatory Program for Online Behavioral Advertising.

OPT-OUT

If you've read the facts and have decided that you'd prefer not to receive targeted ads or offers using Acxiom marketing data, click here to opt-out from the use of Acxiom's marketing data. This will not reduce the number of ads and offers you receive, it just means that some of them may be less relevant to you.

Click here to opt-out.

RESOURCES

Industry Involvement
AboutTheData.com Privacy Policy
Acxiom Global Consumer Privacy
Acxiom Online Privacy
Acxiom.com

PREDICTIVE ANALYTICS

TEXT ANALYTICS

DATA SCIENCE

NETWORK ANALYTICS

NEWS ANALYTICS