The Data Science Landscape: Algorithms, Analytics, and Applications

Sanjiv R. Das
Santa Clara University

What is Data Science?

Three Economic Stages

A. The Producer Economy
B. The Consumer Economy
C. Creator Economy:
    i. We produce and consume
    ii. The currency is information

(Paul Saffo, Stanford)

- Data + Algorithms (ML) + Talent
- Talent = Business + Math/Stat + Code + Communication

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.

Data Science

- Converting (big) data into (intelligent) decisions.
- Predicting behavior (analytics): shopping, dating, voting.
- Cloud computing.
- McKinsey Global Institute: projected that the United States needed 140,000 to 190,000 more workers with “deep analytical” expertise and 1.5 million more data literate managers, whether retrained or hired.

The Downside

- String of financial failures.
- Models may be spurious: Emanuel Derman – “Models. Behaving. Badly” (book)
- Claudia Perlich, chief scientist at Media6Degrees, an online ad targeting startup in New York, puts the problem this way: “You can fool yourself with data like you can’t with anything else. I fear a Big Data bubble.”
- It’s not always easy to ask the right questions.
- Bigger data does not mean better data. Models are needed.

Big Data

- Users upload an hour of video data to YouTube every second.
- 87% of the U.S. population has heard of Twitter, and 7% use it. More than 400 million tweets a day!
- In contrast, 88% of the population has heard of Facebook, and 41% use it.
- USC's Martin Hilbert calculated that more than 300 exabytes of data storage was being used in 2007, an exabyte being one billion gigabytes, i.e., 10^{18} bytes, and 260 of binary usage.

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				
Value	Metric		Value	JEDEC		IEC	
1000	kB	kilobyte	1024	KB	kilobyte	KiB	kibibyte
1000^2	MB	megabyte	1024^2	MB	megabyte	MiB	mebibyte
1000^3	GB	gigabyte	1024^3	GB	gigabyte	GiB	gibibyte
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

Big Data – The Human Side

- We leave behind “digital exhaust” – phone calls, text, emails, browser history, GPS data, cookies. Used through satellites, sensors, RFID, GPS, etc.
- The average person processes more data in one day than a person did in a lifetime in the 1500s.
- Data transparency: causes human movements through Twitter, unexpected basis for human revolutions.

Big Data – The Human Side

- Babycenter.com:
  - 1 in 3 children born in the US already have a digital identity (e.g. sonogram).
  - 92% have one by age 2.
  - Average digital birth occurs at age 6 months.
- Insurance: Progressive now offers Snapshot, a phone app that tracks cars location, acceleration, braking, distance traveled, and within parameters, offers 15% discounts.

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

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.
- McKinsey argues that a successful data analytics plan contains three elements: interlinked data inputs, analytics models, and decision-support tools. ("How Big Data is Changing the Whole Equation for Business," Wall Street Journal March 8, 2013.)
- 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.)

The Economist

Banks and big data

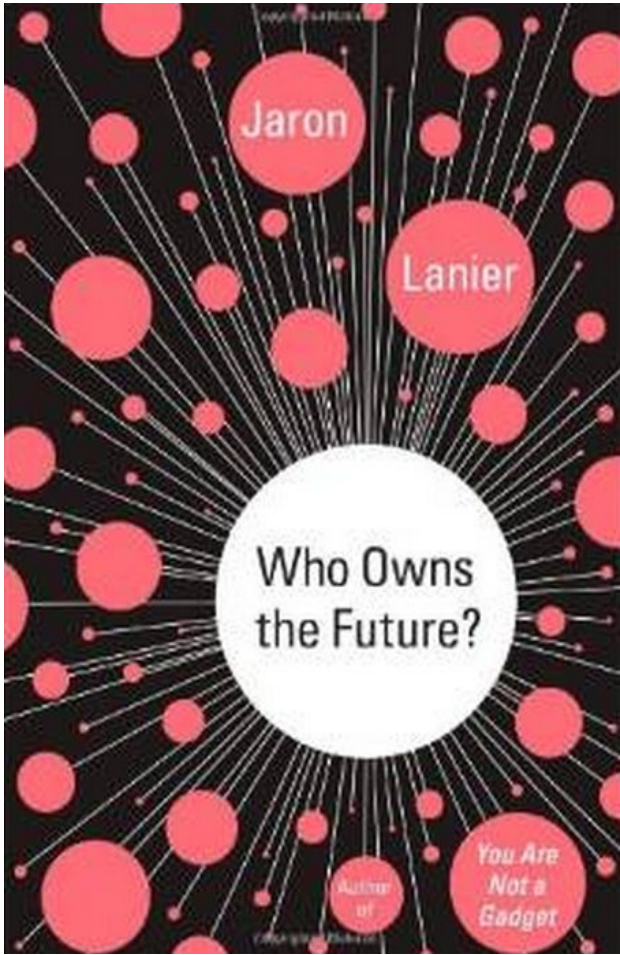
Shopping at the bank

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."

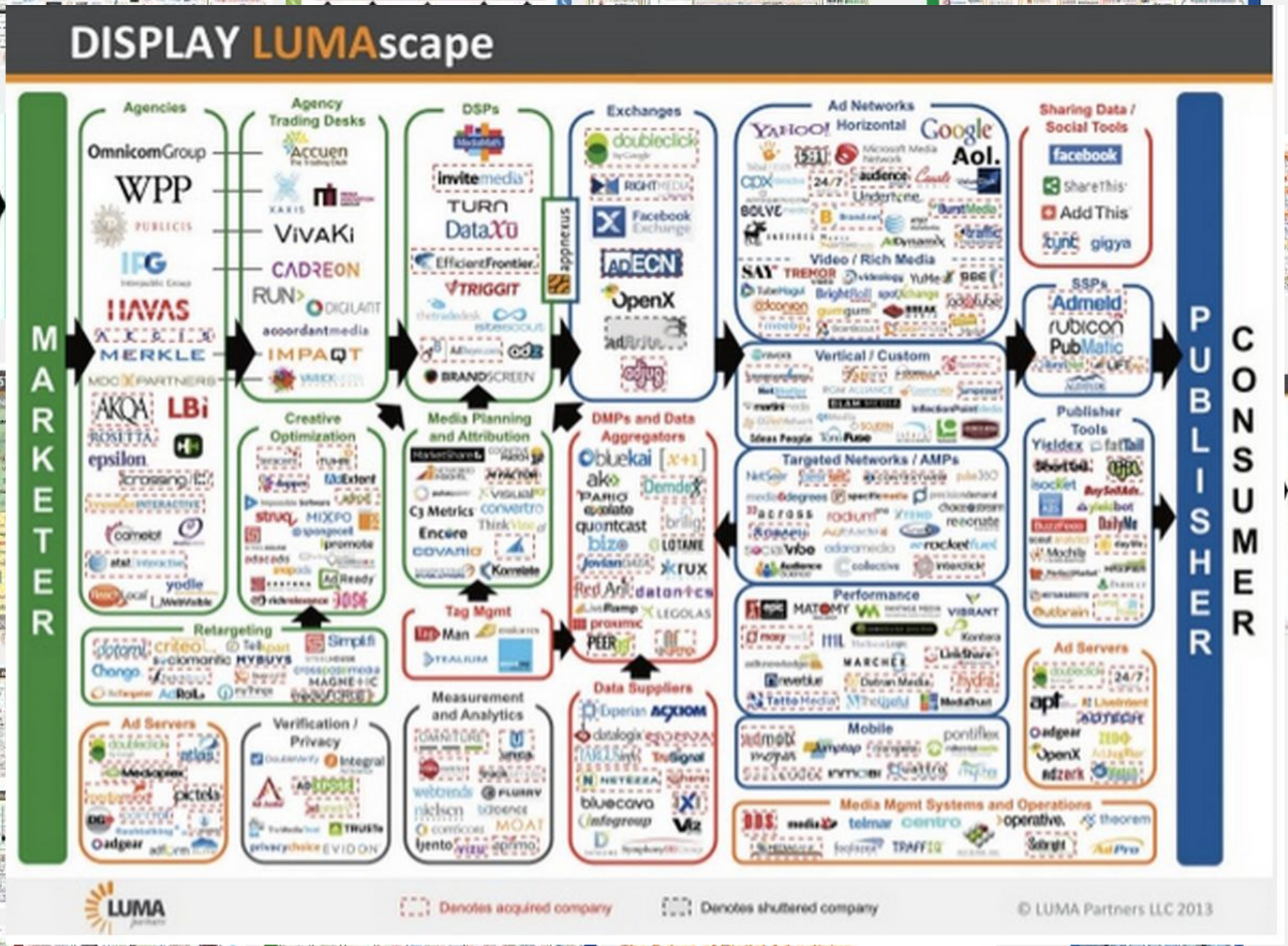
Cardlytics inserts an advertisement into these customers' online bank statements, ideally under a relevant transaction such as a payment to a rival retailer. Customers accept the discount online by clicking a box. This connects the discount coupon to their debit card, and the discount is automatically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



High Potential Revenue

The potential revenues are alluring to banks, which are struggling to maintain profitability in the face of low interest rates (which depress their margins) and consumer legislation that has outlawed or reduced many of the fees they are able to charge. Cardlytics, which recently struck an agreement with Bank of America and is now working with 327 banks that reach 78m households in America, says it is able to charge retailers an average advertising fee worth some 10% of the price of the purchase made by the customer, which is then split with the bank. It also claims mouthwatering rates of "click through". Some 15-20% of customers accept offers, of which about one third are actually redeemed. By contrast click-through rates in most other online-advertising media are less than 2%, and in some cases as low as 0.2%.

The debate: privacy vs efficiency?



The Future of Digital Advertising

The Consequences of Machine Intelligence

By Moshe Y. Vardi

If machines are capable of doing almost any work humans can do, what will humans do?

> *We are facing the prospect of being completely out-competed by our own creations.*

A more thoughtful answer is that technology has been destroying jobs since the start of the Industrial Revolution, yet new jobs are continually created. The AI Revolution, however, is different than the Industrial Revolution. In the 19th century machines competed with human brawn. Now machines are competing with human brain. Robots combine brain and brawn. We are facing the prospect of being completely out-competed by our own creations. Another typical answer is that if machines will do all of our work, then we will be free to pursue leisure activities. The economist John Maynard Keynes addressed this issue already in 1930, when he wrote, "The increase of technical efficiency has been taking place faster than we can deal with the problem of labour absorption." Keynes imagined 2030 as a time in which most people worked only 15 hours a week, and would occupy themselves mostly with leisure activities.

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

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.

"You can't manage what you don't measure."

Data-driven decisions are better decisions—it's as simple as that. Using big data enables managers to decide on the basis of evidence rather than intuition. For that reason it has the potential to revolutionize management.

Companies that were born digital, such as Google and Amazon, are already masters of big data. But the potential to gain competitive advantage from it may be even greater for other companies.

The managerial challenges, however, are very real. Senior decision makers have to embrace evidence-based decision making. Their companies need to hire scientists who can find patterns in data and translate them into useful business information. And whole organizations need to redefine their understanding of "judgment."

Volume. As of 2012, about 2.5 exabytes of data are created each day, and that number is doubling every 40 months or so. More data cross the internet every second than were stored in the entire internet just 20 years ago. This gives companies an opportunity to work with many petabyes of data in a single data set—and not just from the internet. For instance, it is estimated that Walmart collects more than 2.5 petabytes of data every hour from its customer transactions. A petabyte is one quadrillion bytes, or the equivalent of about 20 million filing cabinets' worth of text. An exabyte is 1,000 times that amount, or one billion gigabytes.

Velocity. For many applications, the speed of data creation is even more important than the volume. Real-time or nearly real-time information makes it possible for a company to be much more agile than its competitors. For instance, our colleague Alex "Sandy" Pentland and his group at the MIT Media Lab used location data from mobile phones to infer how many people were in Macy's parking lots on Black Friday—the start of the Christmas shopping season in the United States. This made it possible to estimate the retailer's sales on that critical day even before Macy's itself had recorded those sales. Rapid insights like that can provide an obvious competitive advantage to Wall Street analysts and Main Street managers.

Gartner group (2006)

Variety. Big data takes the form of messages, updates, and images posted to social networks; readings from sensors; GPS signals from cell phones, and more. Many of the most important sources of big data are relatively new. The huge amounts of information from social networks, for example, are only as old as the networks themselves; Facebook was launched in 2004, Twitter in 2006. The same holds for smartphones and the other mobile devices that now provide enormous streams of data tied to people, activities, and locations. Because these devices are ubiquitous, it's easy to forget that the iPhone was unveiled only five years ago, and the iPad in 2010. Thus the structured databases that stored most corporate information until recently are ill suited to storing and processing big data. At the same time, the steadily declining costs of all the elements of computing—storage, memory, processing, bandwidth, and so on—mean that previously expensive data-intensive approaches are quickly becoming economical.

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

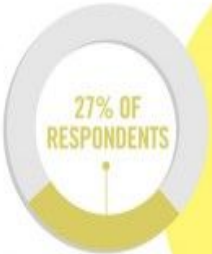
– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Improved Airline ETAs

Minutes matter in airports. So does accurate information about flight arrival times: If a plane lands before the ground staff is ready for it, the passengers and crew are effectively trapped, and if it shows up later than expected, the staff sits idle, driving up costs. So when a major U.S. airline learned from an internal study that about 10% of the flights into its major hub had at least a 10-minute gap between the estimated time of arrival and the actual arrival time—and 30% had a gap of at least five minutes—it decided to take action.

After switching to RightETA, the airline virtually eliminated gaps between estimated and actual arrival times. PASSUR believes that enabling an airline to know when its planes are going to land and plan accordingly is worth several million dollars a year at each airport. It's a simple formula: Using big data leads to better predictions, and better predictions yield better decisions.

Who is a Data Scientist?

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

The shortage of data scientists is becoming a serious constraint in some sectors.

Idea in Brief

A new role is fast gaining prominence in organizations: that of the data scientist. Data scientists are the people who understand how to fish out answers to important business questions from today's tsunami of unstructured information. As companies rush to capitalize on the potential of big data, the largest constraint many face is the scarcity of this special talent.

SKILLS

Data scientists' most basic, universal skill is the ability to write code. This may be less true in five years' time, when many more people will have the title "data scientist" on their business cards. More enduring will be the need for data scientists to communicate in language that all their stakeholders understand—and to demonstrate the special skills involved in storytelling with data, whether verbally, visually, or—ideally—both.

TRAITS

But we would say the dominant trait among data scientists is an intense curiosity—a desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested. This often entails the associative thinking that characterizes the most creative scientists in any field. For example, we know of a data scientist studying a fraud problem who realized that it was analogous to a type of DNA sequencing problem. By bringing together those disparate worlds, he and his team were able to craft a solution that dramatically reduced fraud losses.

Data scientists want to build things, not just give advice. One describes being a consultant as "the dead zone."

How Do You Become a Data Scientist?

Skills: Programming, Math, Stats

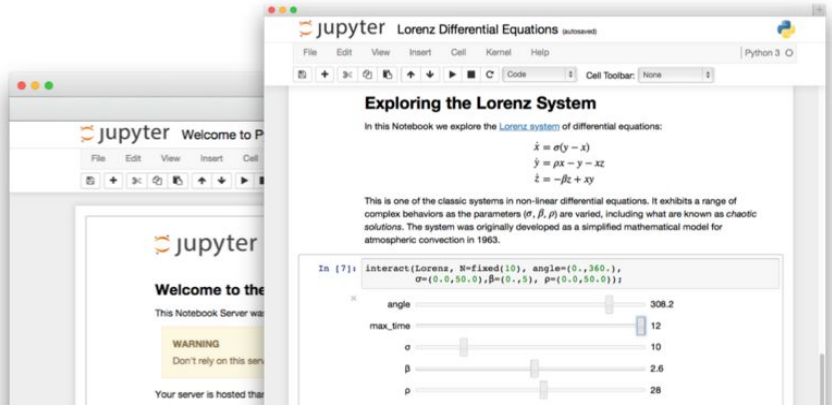
- Python
- R
- Database Management (SQL, noSQL)
- Visualization (Tableau)
- Cloud computing: Hadoop, Spark, AWS, etc.
- Econometrics
- Machine Learning
  - Supervised vs Unsupervised learning
  - Regression vs Classification techniques
- Business skills
- Advanced skills, e.g., Deep Learning: TensorFlow, Caffe, Torch, Theano.

jupyter

Open source, interactive data science and scientific computing across over 40 programming languages.

Jupyter Notebook

The Jupyter Notebook is a web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include:



The R Project for Statistical Computing

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Mailing Lists
Bug Tracking
Development Site
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Documentation

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 3.2.4 (Very Secure Dishes) prerelease versions** will appear starting Monday 2016-02-29. Final release is scheduled for Thursday 2016-03-10.
- **R version 3.3.0 (Supposedly Educational) prerelease versions** will appear starting Monday 2016-03-14. Final release is scheduled for Thursday 2016-04-14.
- The **R Logo** is available for download in high-resolution PNG or SVG formats.
- **useR! 2016**, will take place at Stanford University, CA, USA, June 27 - June 30, 2016.
- **The R Journal Volume 7/2** is available.
- **R version 3.2.3 (Wooden Christmas-Tree)** has been released on 2015-12-10.
- **R version 3.1.3 (Smooth Sidewalk)** has been released on 2015-03-09.

Tableau is business intelligence software that helps people see and understand their data.

FAST ANALYTICS

Connect and visualize your data in minutes. Tableau is 10 to 100x faster than existing solutions.

EASE OF USE

Anyone can analyze data with Tableau's intuitive drag & drop products. No programming, just insight.

TensorFlow ™

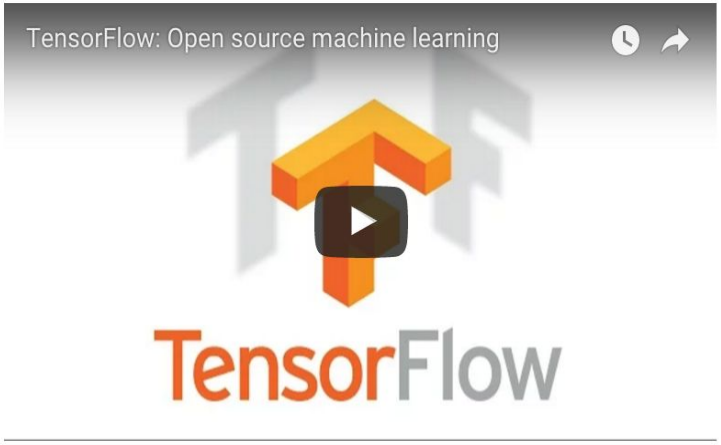
GET STARTED TUTORIALS HOW TO API RESOURCES ABOUT

TensorFlow is an Open Source Software Library for Machine Intelligence

GET STARTED

About TensorFlow

TensorFlow™ is an open source software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them. The flexible architecture allows you to deploy computation to one or more CPUs or GPUs in a desktop, server, or mobile device with a single API. TensorFlow was originally developed by researchers and engineers working on the Google Brain Team within Google's Machine Intelligence research organization for the purposes of conducting machine learning and deep neural networks research, but the system is general enough to be applicable in a wide variety of other domains as well.



https://www.youtube.com/watch?v=oZikw5k_2FM&feature=youtu.be

Meetup Groups

- http://www.meetup.com/R-Users/
- http://www.meetup.com/BAyPIGgies/
- http://www.meetup.com/SF-Bay-ACM/
- http://www.meetup.com/SF-Bay-Areas-Big-Data-Think-Tank/

Resources

Beginnners Introduction to coding in R:
http://www.computerworld.com/article/2497143/business-intelligence/business-intelligence-beginner-s-guide-to-r-introduction.html

My free book on data science may be downloaded here: http://algo.scu.edu/~sanjivdas/WebBook/DSA_Book.pdf
This is a work in progress, so keep the link as I keep adding more to this book from time to time. I also have lots of cleanup to finish it, but you can use the code to try out analyses as you go. The book uses the R programming language which you should find useful and a solid complement to Python.

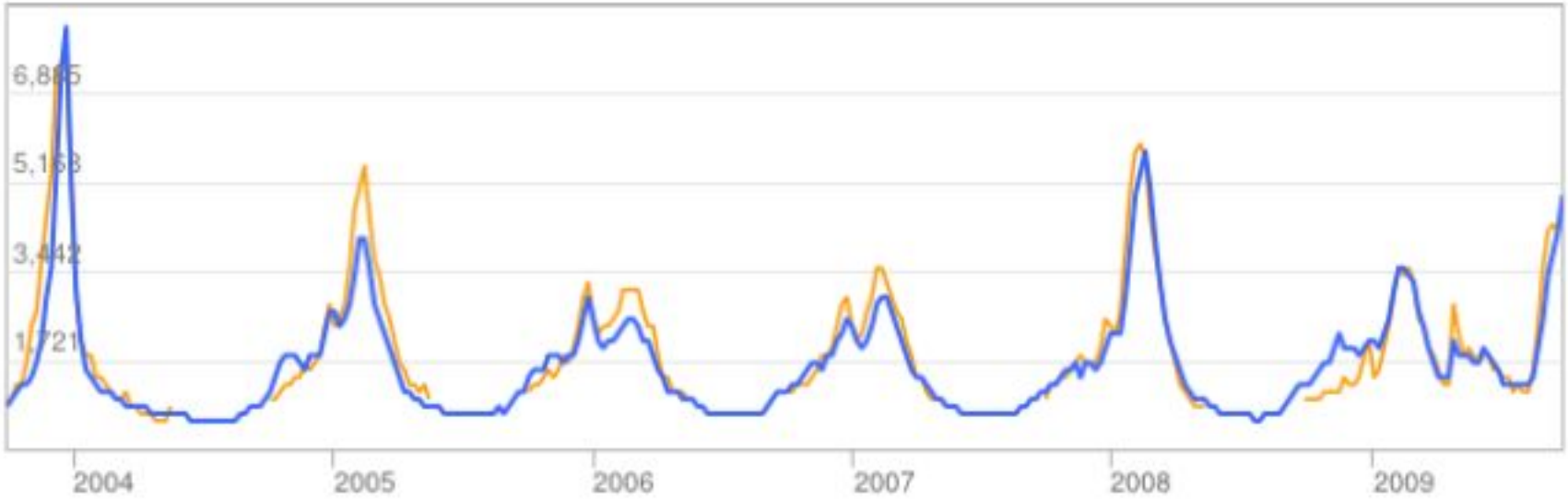
The best Python book for data science is by Wes McKinney:
http://www.cin.ufpe.br/~embat/Python%20for%20Data%20Analysis.pdf

You might also want to take the online course on Data Science offered by Coursera:
https://www.coursera.org/specializations/jhu-data-science

Good R tutorial: http://www.cyclismo.org/tutorial/R/

What is Different About Data Science?
Core Ideas

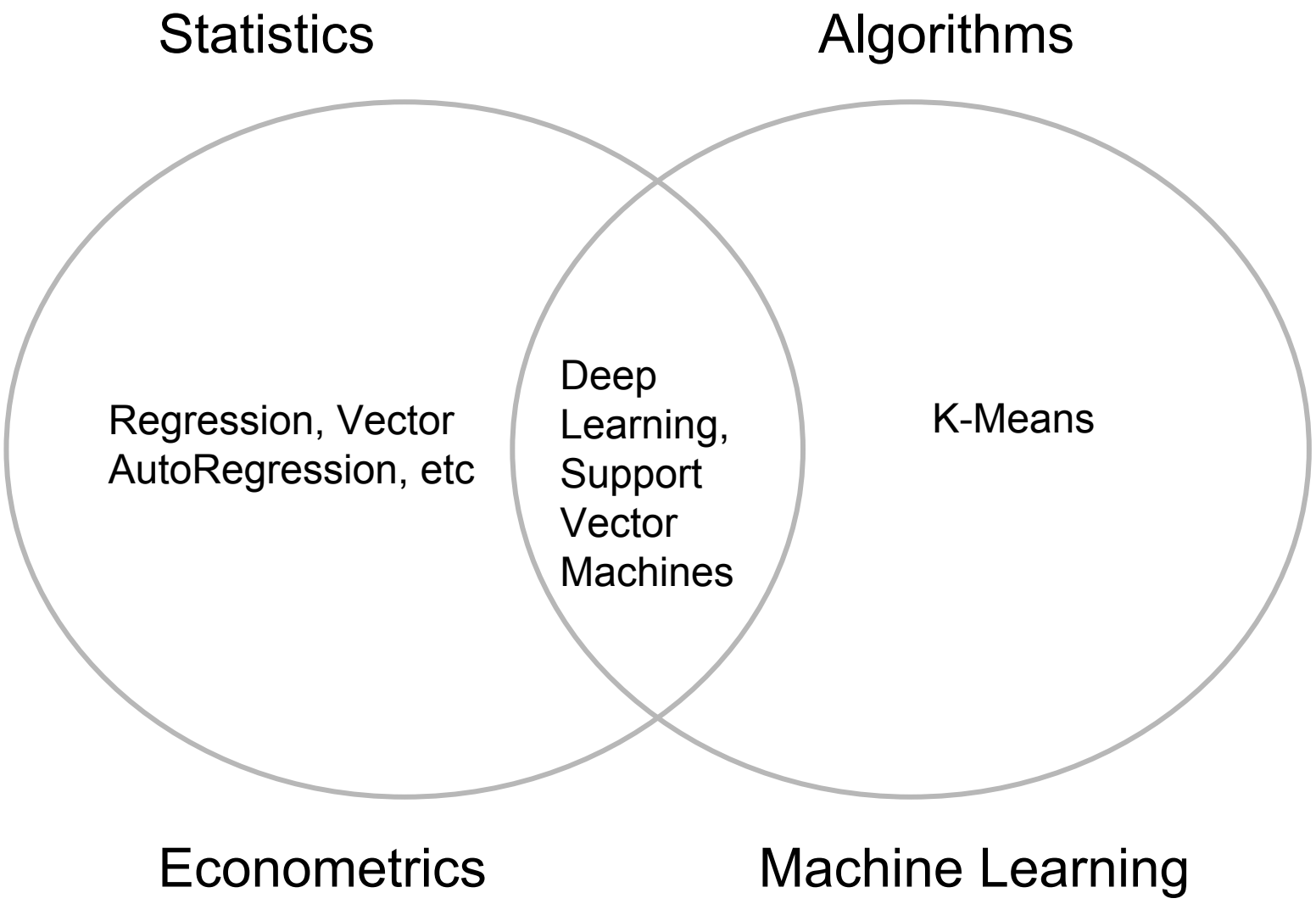
Google Flu Trends



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.


Statistics
Algorithms
Regression, Vector AutoRegression, etc
Deep Learning, Support Vector Machines
K-Means
Econometrics
Machine Learning

Elements of Data Analytics

1. Machine Learning: (a) Supervised; (b) Unsupervised; (c) Reinforcement; (d) Evolutionary.
2. Decision making has a time horizon: (a) snapshot, static; (b) dynamic.
3. Frequentist vs Bayesian approaches.
4. Big data $\rightarrow$ dimension reduction.
5. Output: (a) predictive densities; (b) sample paths (scenarios).
6. Forecasts vs Predictions ("predictive analytics").

Supervised Learning

Modeling input-output pairs (x_i, y_i), fitting a function $y = f(x)$ using training data.

- Regressions
- Autoregressive systems
- Vector autoregressions
- Logit/Probit
- Bayes classifier
- Support vector machines
- Word count classifier
- Adjective-Adverb classifier
- Vector-distance classifier
- Confusion matrix
- Discriminant analysis
- Neural networks
- Prediction trees.

Unsupervised Learning

Uses only the x_is and finds relationships across these.

- Vector-distance, and other distance metrics, like Euclidian, Manhattan, etc.
- Factor analysis, PCA
- Centrality
- Communities
- Strongly connected components
- Shortest paths
- Cluster analysis

Innovation and Experimentation

- New concepts and fresh algorithms
- Mass-scale experiments
- A/B Testing: Google does more than 7000 such experiments in a year. (“The A/B Test: Inside the Technology that’s Changing the Rules of Business,” by Brian Christian, Wired, April 2012.)

Other Topics in Data Science

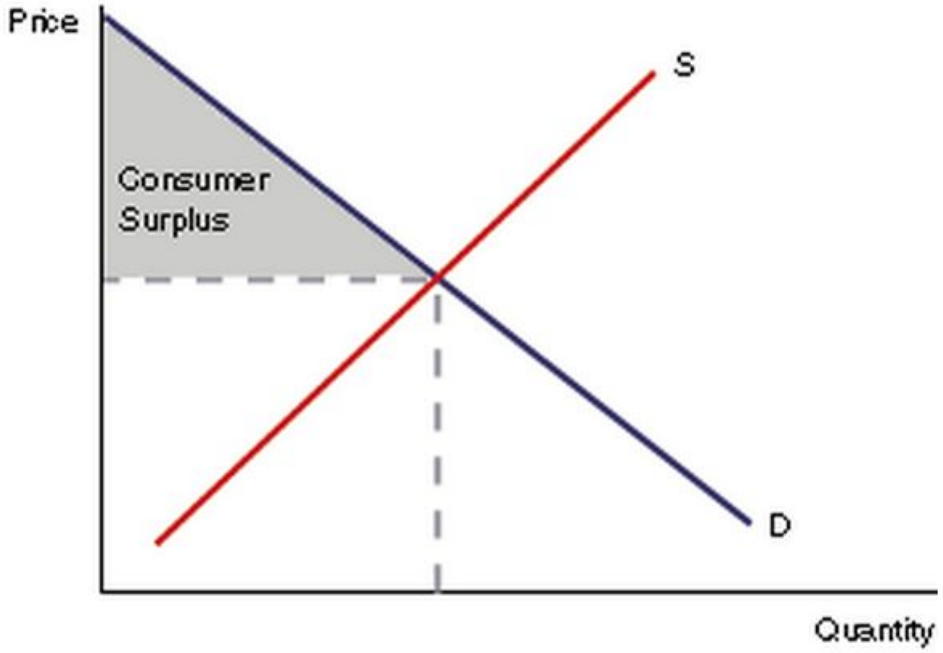
- Optimization routines in R
- Panel data analysis
- Genetic algorithms
- Conjoint analysis
- Golden Ratio and its applications
- Over-fitting analysis
- Kalman filters
- Hidden Markov Models
- Monte Carlo Markov Chain (MCMC)
- Factor-based asset allocation
- Risk Metrics
- Extreme value distributions and normal inverse Gaussians
- Copula models
- Power laws
- Splines
- Monte Carlo Simulation

The Dark Side 1: Big Data, Big Errors

- Too much data leads to misleadingly significant relationships. Correlations may be overstated.
- Too many columns and too few rows of data.
- Explore the data fully, don't stop when a significant result occurs, which is easy to get with big data.
- Beware confirmation bias.
- 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

- Social interaction vs solitude and privacy. Technology makes this trade off steeper.
- Profiling (e.g., Groupon)
- Price discrimination
- Security vs Efficiency
- 1921 dystopian novel "We" by Yevgeny Zamyatin where every building was made of glass so the govt could always watch you.



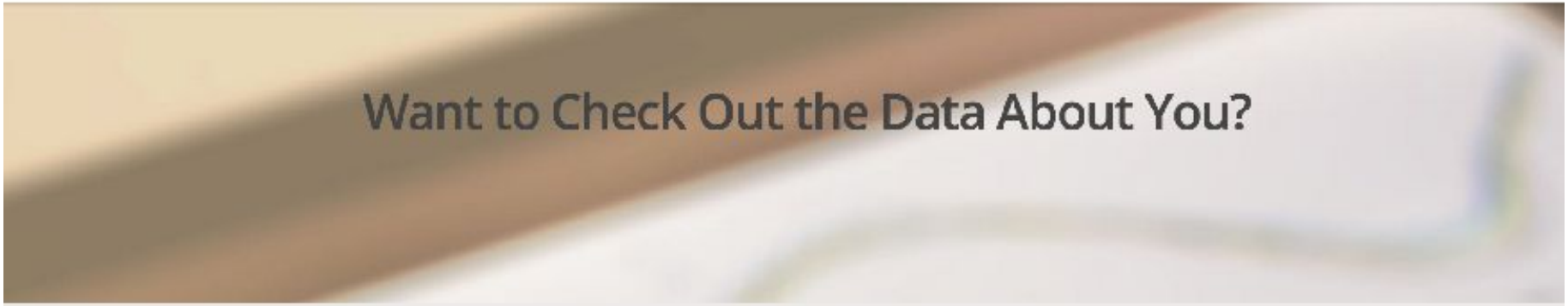
How do companies get data about me and what do they do with it?

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

Theories, Models, Intuition, Causality, Prediction, Correlation

- *Theories* are statements of how the world should be or is, and are derived from axioms that are assumptions about the world, or precedent theories.
- *Models* are implementations of theory, and in data science are often algorithms based on theories that are run on data.
- The results of running a model lead to *intuition*, i.e., a deeper understanding of the world based on theory, model, and data.

Unreasonable Effectiveness of Big Data

- Chris Anderson: "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." Wired, v16(7), 23rd June, 2008.

Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of big data, more isn't just more. More is different.

Models > Big Data

- Thomas Davenport writes in his foreword to Seigel (2013) that models are key, and should not be increasingly eschewed with increasing data:

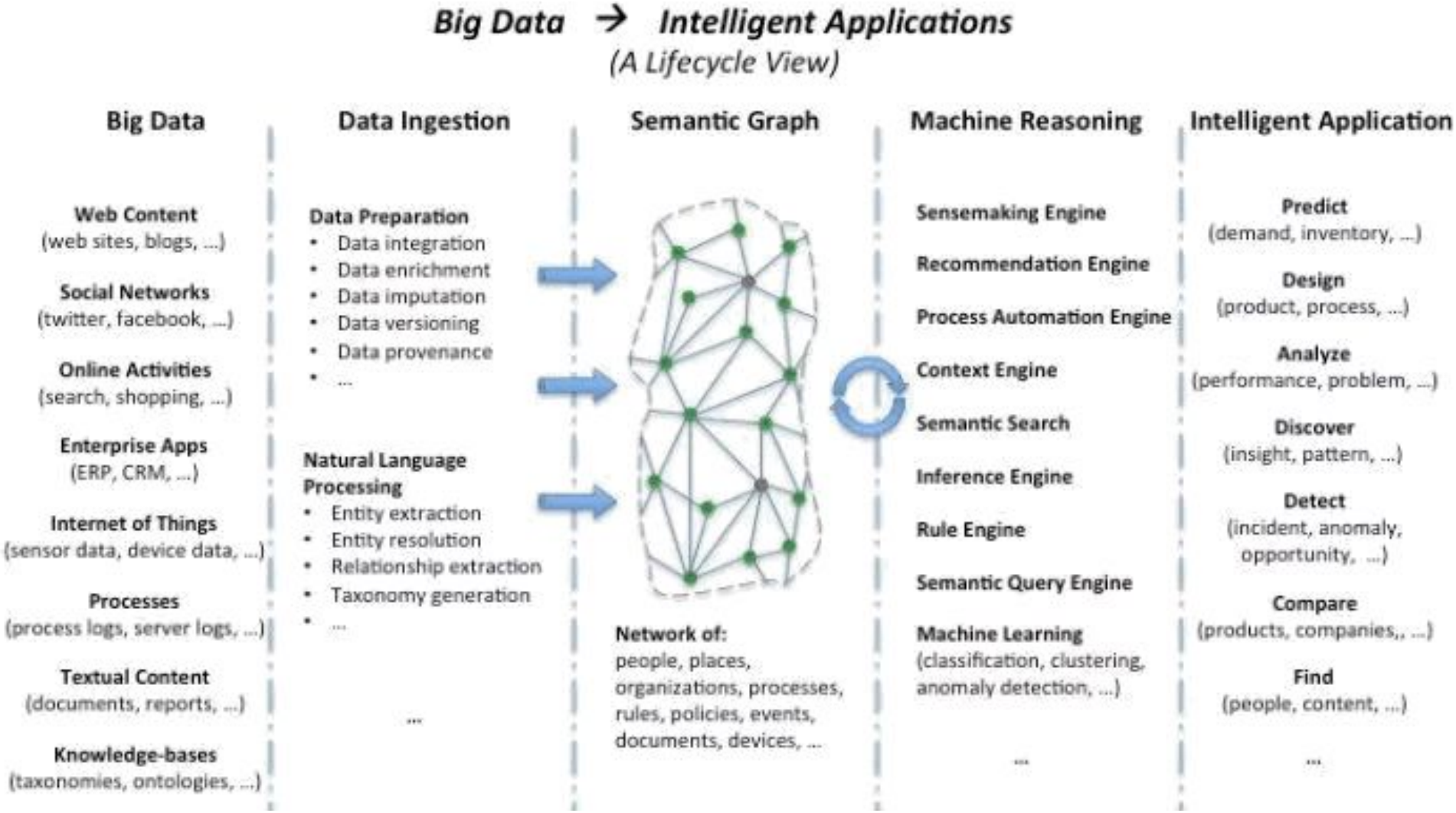
> *But the point of predictive analytics is not the relative size or unruliness of your data, but what you do with it. I have found that "big data often means small math," and many big data practitioners are content just to use their data to create some appealing visual analytics. That's not nearly as valuable as creating a predictive model.*

Causality (Granger)

Once we have established intuition for the results of a model, it remains to be seen whether the relationships we observe are causal, predictive, or merely correlational. Theory may be causal and tested as such. Granger (1969) causality is often stated in mathematical form for two stationary[17] time series of data as follows. X is said to Granger cause Y if in the following equation system,

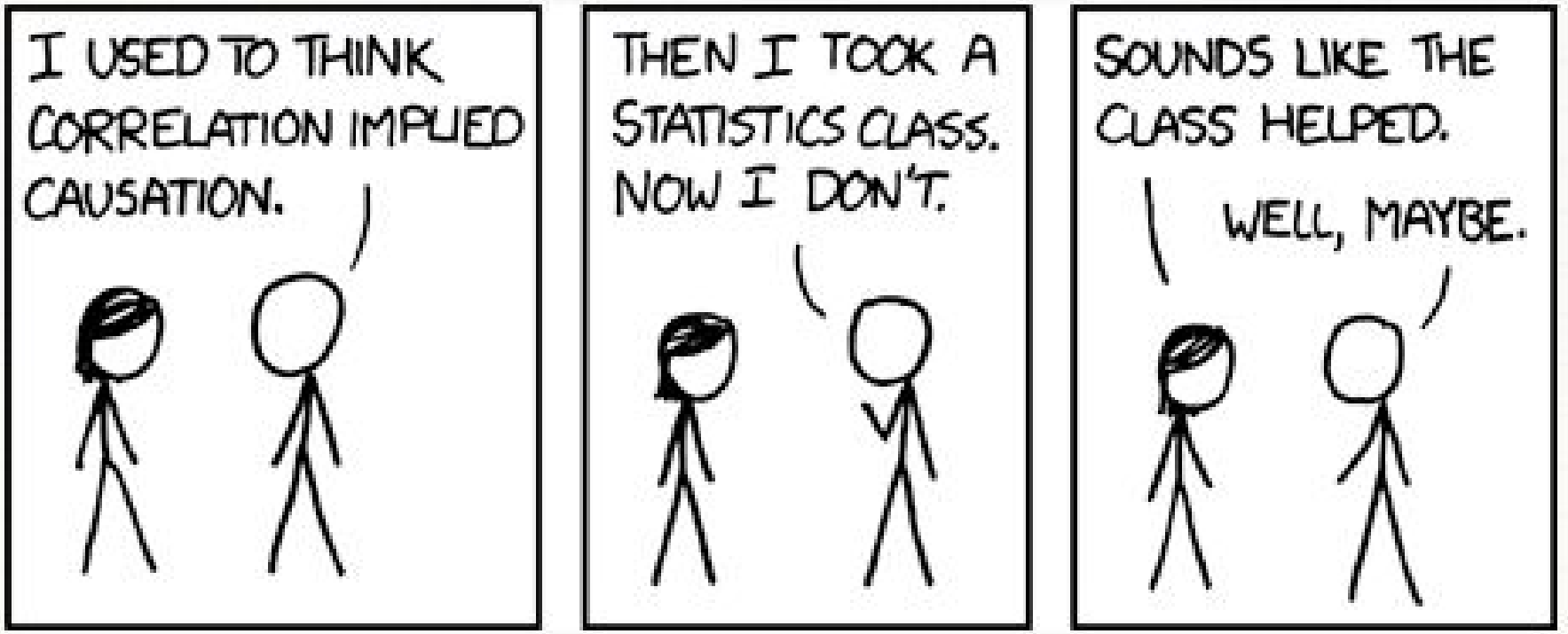
$$\begin{aligned} Y(t) &= a_1 + b_1 Y(t-1) + c_1 X(t-1) + e_1 \\ X(t) &= a_2 + b_2 Y(t-1) + c_2 X(t-1) + e_2 \end{aligned}$$

the coefficient c_1 is significant and b_2 is not significant. Hence, X causes Y, but not vice versa. Causality is a hard property to establish, even with theoretical foundation, as the causal effect has to be well-entrenched in the data.



From Cirrus Shakeri, **"From Big Data to Intelligent Applications,"** post, January 2015
https://www.linkedin.com/pulse/from-big-data-intelligent-applications-cirrus-shakeri

http://www.odbms.org/blog/2016/02/a-grand-tour-of-big-data-interview-with-alan-morrison/


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.