Introduction to Data Science

CS 5665
Utah State University
Department of Computer Science
Instructor: Prof. Kyumin Lee

Announcements

Practicums

- https://usu.instructure.com/courses/431417/discussion_topics/1385921
- 22 topics were posted! Let me know which topic you want to present by next Tuesday
- Suggest additional topics!
- We plan to begin the first practicum week after next week

In the news…

http://www.nytimes.com/interactive/2012/06/11/sports/basketball/nba-shot-analysis.html

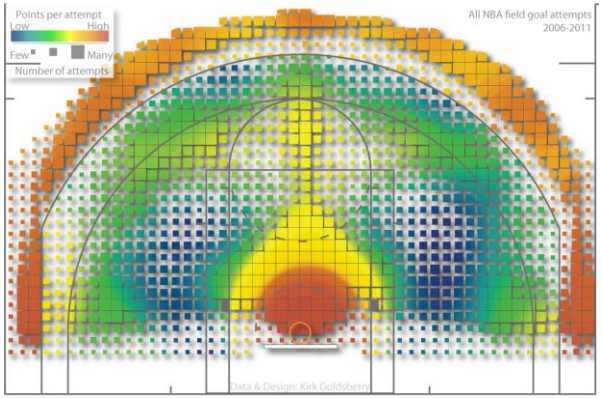
CourtVision: New Visual and Spatial Analytics for the NBA

Kirk Goldsberry, Ph.D.
Harvard University,
1730 Cambridge St, Cambridge, MA, 02138
Email: kgoldsberry@fas.harvard.edu

Abstract

This paper investigates spatial and visual analytics as means to enhance basketball expertise. We introduce CourtVision, a new ensemble of analytical techniques designed to quantify, visualize, and communicate spatial aspects of NBA performance with unprecedented precision and clarity. We propose a new way to quantify the shooting range of NBA players and present original methods that measure, chart, and reveal differences in NBA players' shooting abilities. We conduct a case study, which applies these methods to 1) inspect spatially aware shot site performances for every player in the NBA, and 2) to determine which players exhibit the most potent spatial shooting behaviors. We present evidence that Steve Nash and Ray Allen have the best shooting range in the NBA. We conclude by suggesting that visual and spatial analysis represent vital new methodologies for NBA analysts.

www.sloansportsconference.com/wp-content/uploads/2012/02/Goldsberry_Sloan_Submission.pdf



- Data:
  - Every NBA game 2006--2011
  - (x,y) coordinate for every shot
  - (Player Name, Shot location, Shot outcome) ==> 700,000 tuples
  - Divided court into 1284 "shooting cells", each 1 square foot

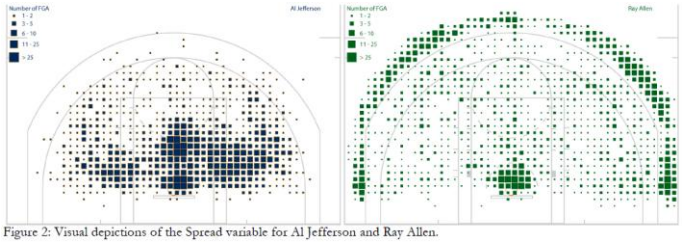


Figure 2: Visual depictions of the Spread variable for Al Jefferson and Ray Allen.

Other articles

http://digital.cs.usu.edu/~kyumin/cs5665/schedule.htm

Data Science…

- Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured, which is a continuation of data analysis fields such as data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD)

https://en.wikipedia.org/wiki/Data_science

Is this Science?

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

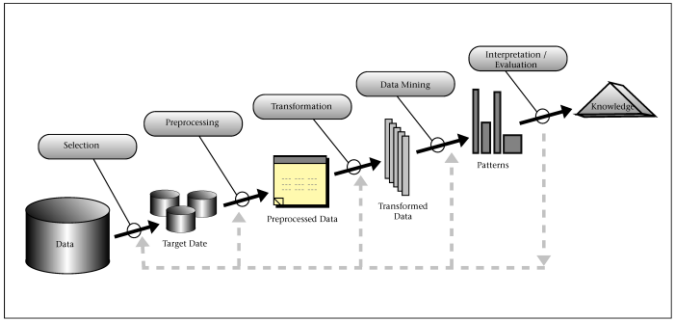
The key word in "Data Science" is not Data, it is Science

- Data science is only useful when the data are used to answer a question.
- It is much, much easier to say "My data are bigger than yours" or to say, "I can code in Hadoop, can you?" than to say, "I have this really hard question, can I answer it with my data?"
- The issue is that the hype around big data/data science will flame out if data science is only about "data" and not about "science". The long term impact of data science will be measured by the scientific questions we can answer with the data.

http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/

Origins of Data Science

Fayyad (1996)



An Overview of the Steps That Compose the KDD Process.

“Business intelligence”

- 1958: “the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal”
- 1989: “concepts and methods to improve business decision making by using fact-based support systems”

https://en.wikipedia.org/wiki/Business_intelligence

The Data Science Process

Jeff Hammerbacher

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

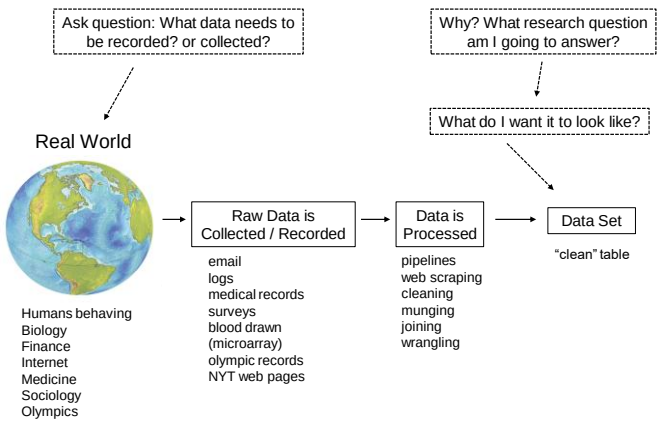
dataists

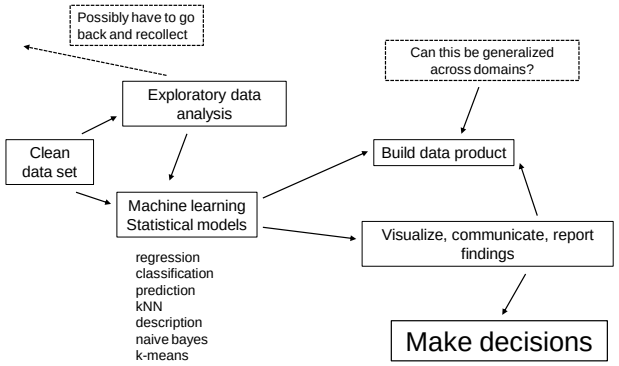
1. Obtain
2. Scrub
3. Explore
4. Model
5. Interpret

Jim Gray

1. Capture
2. Curate
3. Communicate

This Class





Are these examples of Data Science?

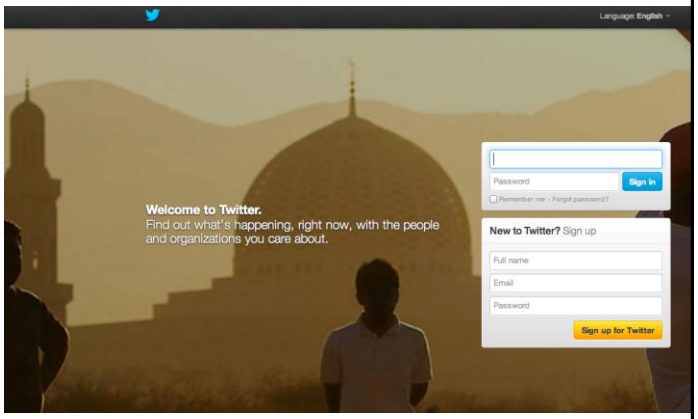
How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

306 comments, 167 called-out + Comment Now + Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Target has got you in its aim

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest II number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources.



SMARTER THAN YOU THINK

Rock-Paper-Scissors: You vs. the Computer

RECOMMEND TWITTER LINKEDIN SIGN IN TO E-MAIL SHARE

Computers mimic human reasoning by building on simple rules and statistical averages. Test your strategy against the computer in this rock-paper-scissors game illustrating basic artificial intelligence. Choose from two different modes: novice, where the computer learns to play from scratch, and veteran, where the computer pits over 200,000 rounds of previous experience against you.

Note: A truly random game of rock-paper-scissors would result in a statistical tie with each player winning, tying and losing one-third of the time. However, people are not truly random and thus can be studied and analyzed. While this computer won't win all rounds, over time it can exploit a person's tendencies and patterns to gain an advantage over its opponent.

HUMAN | WINS 0 | TIES 0 | WINS 0 | COMPUTER

VS.

Choose your opponent:

Novice

Play against a computer that has no previous experience and learns to play based solely on your tendencies.

Veteran

Play against a computer that uses data gathered from thousands of games of rock-paper-scissors versus other people.

Use Google Analytics to Analyze Holiday Traffic and Sales

By David A. Utter
EcommerceBytes.com
January 14, 2013

Tweet 8 | Like 4 | Share | +1 +2

The 2012 holiday shopping season has drawn to a close. With luck, our ecommerce audience enjoyed the fruits of their labors, especially with US shoppers spending $42.3 billion online. According to comScore, that was a 14 percent year over year increase for the November-December holiday shopping season.

Those shoppers brought along something else to ecommerce sites besides open wallets and full shopping carts. They first brought traffic to these sites, to browse and click and perhaps to buy. Consumers delivered a lot of data to online seller server logs, leaving ecommerce pros to sift through it with tools like Google Analytics.

Traffic represents an interesting area to review. Interpreting it can show how engaged visitors are with site content, how well they convert based on one's expectations, and how well the site performed at deriving ecommerce effectiveness.

For example, Google suggested some points to consider when reviewing the Ecommerce tab in Google Analytics. This shows the revenue and transactions associated with the site's traffic sources. They suggest a couple of possible scenarios one might encounter here:

- A lot of traffic but little revenue from a source (are you running the wrong ads, or not carrying products relevant to those visitors?)
- A limited number of visits from a source, but with a high number of transactions and a high average value per transaction (indicating a potentially lucrative market that you haven't fully addressed).

Google's dominance in search and advertising likely means ecommerce pros made at least a minimal effort to reach customers via AdWords, Google's paid search program. But AdWords may not have performed to one's expectations. Google

Stories

People using Facebook in extraordinary ways Like 347k

Mapping the World's Friendships

Learn which countries share the closest friendship connections on Facebook.
Methodology

Show Continent Search

2. Canada
4. United Kingdom
United States of America
3. Dominican Republic
1. Mexico
5. Australia

A Closer Look: United States & Mexico 1/5

As of 2009, approximately 30% of the foreign-born population in the U.S. was from Mexico, the most by far compared to the immigration rates of any other single country. Source

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 »

Download world flu activity data - Animated flu trends for Google Earth - Compare flu trends across regions in Public Data Explorer

The top emoticons

It's no big surprise, but the popular emoticons dominate the usage patterns. Of the **96,269,892** Tweets that contained emoti
20 smileys accounted for 90% of all occurences. Here they are:

	Emoticon	Usage	Percent	Notes
#1	:)	32,115,789	33.360%	*Happy face*
#2	:D	10,595,385	11.006%	*Laugh*
#3	:(	7,613,014	7.908%	*Sad face*
#4	;)	7,238,295	7.519%	*Wink*
#5	:-)	4,254,708	4.420%	*Happy face (with nose)*
#6	:P	3,588,863	3.728%	*Tongue out*
#7	=)	3,564,080	3.702%	*Happy face*
#8	(:	2,720,383	2.826%	*Happy face (mirror)*
#9	;-)	2,085,015	2.166%	*Wink (with nose)*
#10	:/	1,840,827	1.912%	*Uneasy, undecided, skeptical, annoyed?*
#11	XD	1,795,792	1.865%	*Big grin*
#12	=D	1,434,004	1.490%	*Laugh*
#13	:o	1,077,124	1.119%	*Shock, Yawn*
#14	=]	1,055,517	1.096%	*Happy face*
#15	D:	1,048,320	1.089%	*Grin (mirror)*

Data Science: The Context

Goal of Data Science

- Discovery of patterns and models that are:
  - Valid: hold on new data with some certainty
  - Useful: should be possible to act on the item
  - Unexpected: non-obvious to the system
  - Understandable: humans should be able to interpret the pattern

Two Major Tasks

- **Predictive** Methods (supervised learning methods)
  - Use some variables to predict unknown or future values of other variables
- **Descriptive** Methods (unsupervised learning methods)
  - Find human-interpretable patterns that describe the data
    - e.g., categorize customers by their product preferences (clustering) or understand relations (association)

Meaningfulness of Answers

- A big data mining risk is that you will "discover" patterns that are meaningless
- **Bonferroni's principle** (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

Example: Rhine Paradox

- Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception (ESP).
- He devised (something like) an experiment where subjects were asked to guess 10 hidden cards – red or blue.
- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right!

Example: Rhine Paradox

- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- What did he **conclude**?
- He concluded that you shouldn't tell people they have ESP; it causes them to lose it.