

Analytics and Visualization of Big Data

Fadel M. Megahed

Lecture 01: Course Motivation, Structure and Overview



AUBURN UNIVERSITY

SAMUEL GINN
COLLEGE OF ENGINEERING

Department of Industrial and Systems Engineering

Spring 13

What is Analytics?

2

- It is the discovery of knowledge from data.
- Uses statistics, computer programming and OR.
- Insights are often communicated via visual approaches.
- Used to describe, predict, and improve business performance.

Data \neq Knowledge (or Information)



Source: <https://www.unboundid.com/blog/2012/10/31/big-data-analytics-meets-the-identity-economy-in-amsterdam/>



- Questions: How do you determine the best shooter in the NBA?
- Possible Metrics:
 - Number of field goals (FG) made
 - Limitation: If you take a lot of shots, you will make more.
 - Field goal percentage (FG%)
 - Limitation: Does not take into account where you are taking the shot from. Therefore, centers and forwards will skew this statistic!!
 - Enhanced Field goal percentages (eFG%)
 - Limitation: Not all 2 and 3 point shots are made equal!!

Source: Goldsberry, K. (2012) "CourtVision: New Visual and Spatial Analytics for the NBA". *MIT Sloan Sports Analytics Conference*.
http://www.sloansportsconference.com/wp-content/uploads/2012/02/Goldsberry_Sloan_Submission.pdf



A Visual Analytics Example (Cont.)

4

Data: Using game data sets for every NBA game played between 2006 and 2011, we compiled a spatial field goal database that included Cartesian coordinates (x,y) for every field goal attempted in this 5-year period. This data set includes player name, shot location, and shot outcome for over 700,000 field goal attempts. We mapped the shot data atop a base map of a NBA basketball court (Figure 1). Although a regulation NBA court is 4,700 ft², (50ft x 94ft), almost all (>98%) field goal attempts occur within a 1,284 ft² area in between the baseline and a relatively thin buffer around the 3-point arc; we call this area the “scoring area.” We divided the scoring area into a grid consisting of 1,284 unique “shooting cells,” each 1 ft² (Figure 1). To quantify shooting range, we applied spatial analyses to evaluate shooting performance across the grid and within each shooting cell.

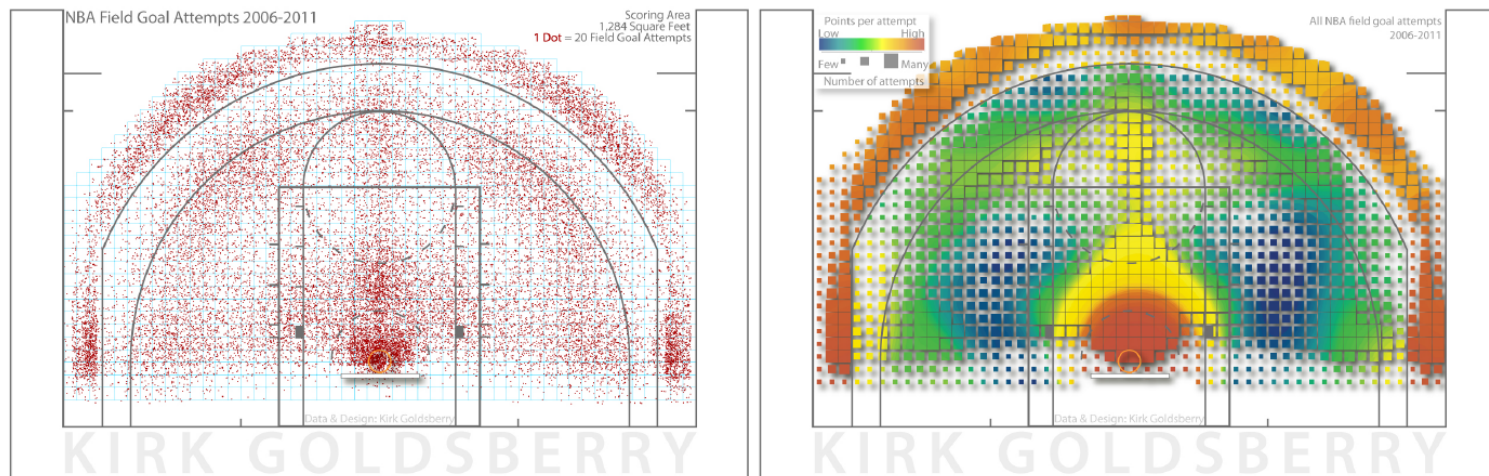


Figure 1: Our composite shot maps from 2006-2011 NBA game data. The map on the left summarizes the density of all field goal attempts during the study period. The map on the right reveals league-wide tendencies in both shot attempts and points per attempt. Larger squares indicate areas where many field goals were attempted; smaller squares indicate fewer attempts. The color of the squares is determined by a spectral color scheme and indicates the average points per attempt for each location. Orange areas indicate areas where more points result from an average attempt, and blue areas indicate fewer points per attempt.

Source: Goldsberry, K. (2012) “CourtVision: New Visual and Spatial Analytics for the NBA”. *MIT Sloan Sports Analytics Conference*.
http://www.sloansportsconference.com/wp-content/uploads/2012/02/Goldsberry_Sloan_Submission.pdf

- Using this visual representations and some devised metrics that account for the knowledge gained by such representations, Prof. Goldsberry was able to answer this question.
- Very succinctly, his answer was based on:
 - The Spread Metric
 - The Range Metric
 - Visual representations similar to the previous figure
- To know his answer and rationale, please read his paper 😊

Source: Goldsberry, K. (2012) "CourtVision: New Visual and Spatial Analytics for the NBA". *MIT Sloan Sports Analytics Conference*.
http://www.sloansportsconference.com/wp-content/uploads/2012/02/Goldsberry_Sloan_Submission.pdf



What is Big Data?

6

- Big data is a popular term that is used to describe the large, diverse, complex and/or longitudinal datasets generated from a variety of instruments, sensors and/or computer-based transactions.¹

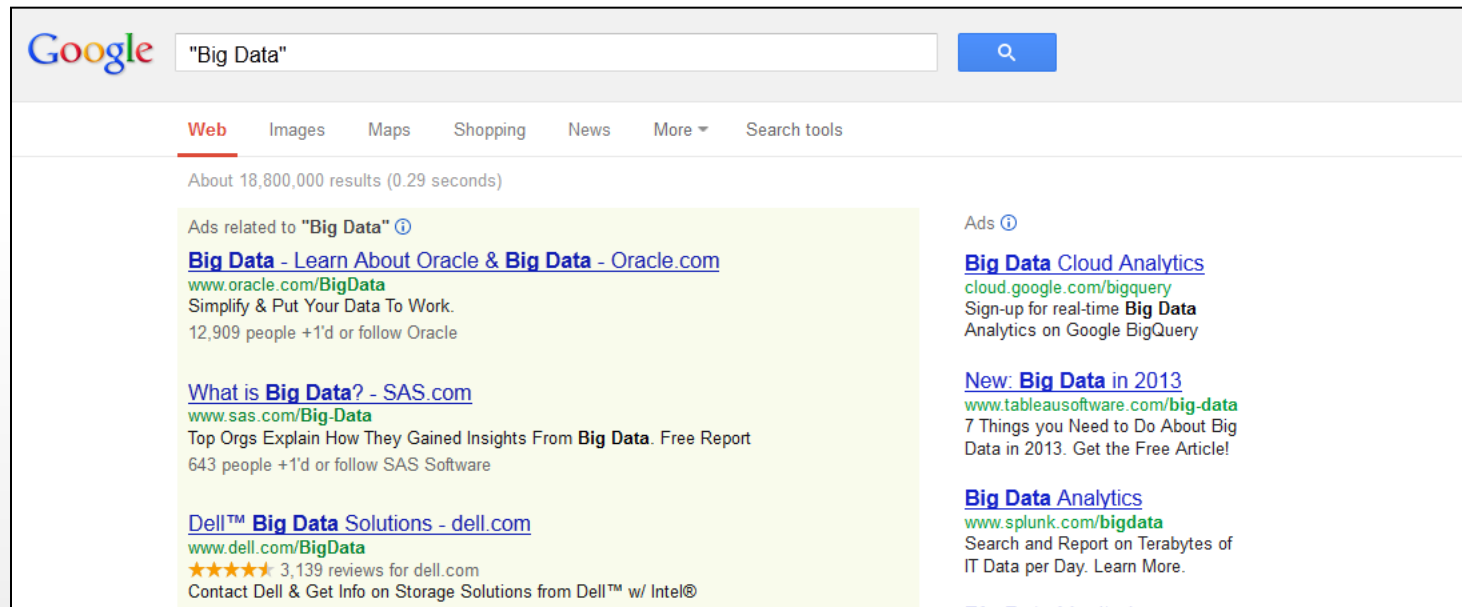


¹ National Science Foundation. Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767. Last accessed 1/9/2013.

What is Big Data? (Cont.)

7

- Big data is a popular term that is used to describe the large, diverse, complex and/or longitudinal datasets generated from a variety of instruments, sensors and/or computer-based transactions.¹



¹ National Science Foundation. Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767. Last accessed 1/9/2013.

What is Big Data? (Cont.)

8

- Big data is a popular term that is used to describe the large, diverse, complex and/or longitudinal datasets generated from a variety of instruments, sensors and/or computer-based transactions.¹

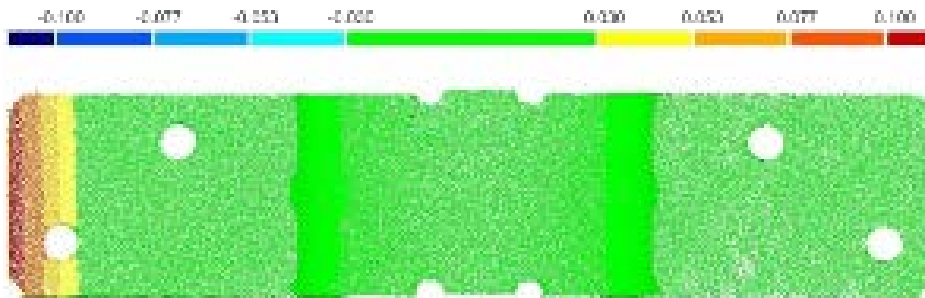


¹ National Science Foundation. Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)
http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767. Last accessed 1/9/2013.

What is Big Data (Cont.)?

9

- The term big data refers not only to the size or volume of data, but also to the variety of data and the velocity or speed of data accrual.¹



World Bank @WorldBank 16h
#BigData needs human insight. New blog says Numbers Are Never Enough. bit.ly/UHKL43
Retweeted 25 times
Expand

Mike Dauber @dauber 28m
Scott Adams has certainly realized that #BigData is great fodder for Dilbert - love that he added in-memory dilbert.com/strips/comic/2...
Expand

Computerworld @Computerworld 33m
RT @johnkwaters: Linguist @GeoffNunberg's word of the year: #bigdata. Obvious? The case he makes is worth a listen ow.ly/gubvz
View summary

InformationOnDemand @ibm_iod 1h
How can we build security for #bigdata environments in the new year? Kim Madia shares insights: ibm.co/Ta02Ts
Expand

Jonathan Hirsch @JonathanHirsch 1h
Theme from Borisy: merge genomics and clinical data, ask questions. Others on panel say we're far from this vision. #bigdata #jpm13
Expand

InformIT @InformIT 1h
#BigData fans, here's an overview of #Hadoop's architecture and how to build a MapReduce application ow.ly/gF4QZ
Expand

Linda Avey @lindaavey 2h
"Don't call it big data, call it good science" Lon Cardon, GSK.
#BigData #JPM13
Expand

¹ National Science Foundation. Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767. Last accessed 1/9/2013.

- 10,000 payment card transactions are made every second around the world.²
- Walmart handles more than 1 million customer transactions an hour.³
- 340 million tweets are sent per day. That's nearly 4,000 tweets per second.⁴
- Facebook has more than 901 million active users generating social interaction data.⁵

*The information in this slide is based on the examples highlighted by SAS in <http://www.sas.com/big-data/>. Please visit their webpage for more information on Big Data and to find the references for the statements in this slide.



The Possible Future of Engineering & Modeling

11

Source: <https://www.youtube.com/watch?v=mlaXbGuMV00>



- Analyze millions of SKUs to determine optimal prices that maximize profit and clear inventory.
- Mine customer data for insights that drive new strategies for customer acquisition and retention.
- Analyze data from social media to detect new market trends and changes in demand.
- Detect fraudulent behavior in credit cards.
- Others??

Source: <http://www.sas.com/big-data/>



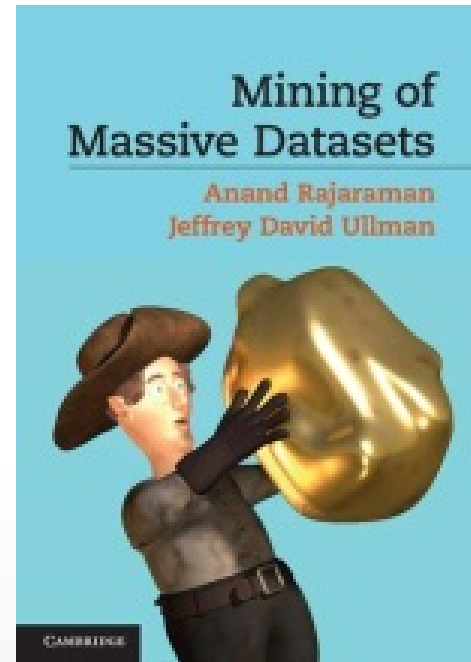
Class Objective, Design and Structure



- Explain the basics behind the hardware and software needed for “big data” analytics.
- Analyze high-dimensional data.
- Describe the components of successful search engines.
- Mine the web using structured and unstructured data.
- Train algorithms that can be used to extract new knowledge from data.
- Develop visualizations that makes the data “sing” 😊.



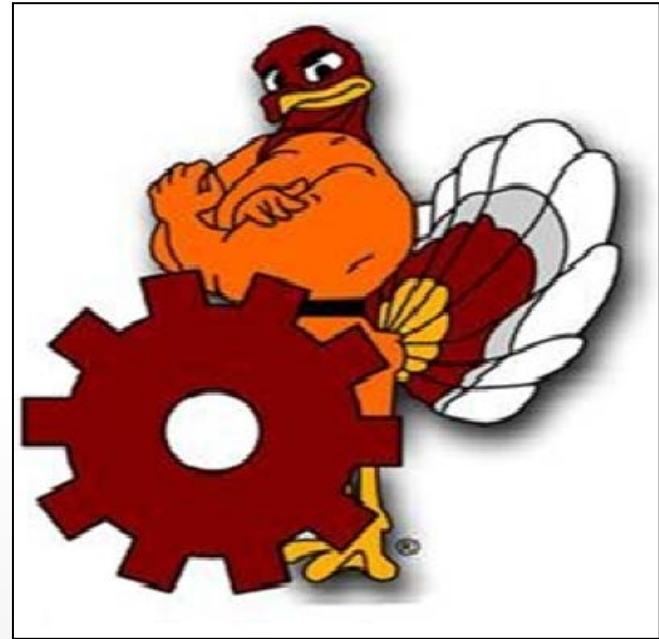
- **Title:** Mining of Massive Datasets
- **Author:** A. Rajaraman, J. Leskovec, J.D. Ullman
- **Web Link:** Click [here](#).
- **Publication Date:** 2012



Meet the “Big Data” Team

16

- Name:
 - Fadel Megahed
- From:
 - Cairo, Egypt
- Education:
 - PhD, Industrial & Systems Engineering (Virginia Tech)
 - MS, Industrial & Systems Engineering (Virginia Tech)
 - BS, Mechanical Engineering (The American University in Cairo)



Meet the “Big Data” Team

17

- Name:
 - Xinyu Que
- From:
 - Wuxi, China
- Education:
 - MS, Computer Science (U. Connecticut)
 - MS, Computer Science (Chinese Academy of Sci.)
 - BS, Computer Science (Jiangnan University)



Meet the “Big Data” Team

18

- Name:
 - Patrick Almas
- From:
 - Florida, USA
- Education:
 - MS Candidate, Industrial & Systems Engineering (Auburn University)
 - BS, Industrial and Systems Engineering (University of Miami)



Meet the “Big Data” Team

19

- Name:
- Where are you from?
- Why are you taking this class? What are you hoping to get out of it?
- What data you plan to mine (manufacturing, health-care, cyber-security, sports, etc.)?



* I would really appreciate if you can write this down and give it to me on paper (or via email for the INSY 7976 students)

- “(...) you will select what work you complete and thus, have a strong control over your learning outcomes from the class. I will also promote the concept of a **learning community**, where the participants will be encouraged to share/develop resources that are relevant to the material covered in class. *After all an industrial engineering data-mining class that does not efficiently and effectively utilize its resources in generating knowledge should not exist!!*”



What are the Assignments that you will Choose from?

21

- In this course, you select what work you complete bearing these 3 simple rules in mind:
 - At least 70 percent of the total points possible for each individual assignment must be earned; otherwise, no points will be recorded for the assignment.
 - Once the due date for an assignment has passed, that assignment cannot be completed.
 - With the exceptions of exams, you are allowed/encouraged to ask your colleagues questions as long as you use Piazza for this purpose. To access our Piazza page, please click [here](#).

Question: Should we use Piazza as our class document repository?



What are the Assignments that you will Choose from?

22

- Homework:
 - Help you gauge your understanding of the material
 - ~ One HW Set every 2 chapters (6 total)
 - Recommendation: Read the text and review class slides before exploring the homework
 - Each homework is worth 10 points (i.e. you can get up to 60 total)


Disclaimer: The information in this slide is aimed at providing you with an overview. Please read the syllabus for more details.



What are the Assignments that you will Choose from?


23

- DM Competitions*: (≤ 2 , possible 50 pts each, click [here](#))




Read more about how Kaggle works

Check out [How it Works](#). Ask a question on the [Kaggle Forum](#).



Explore the competitions





Download some [active competition](#) data files and a sample entry. Or practice on a completed competition.



Meet the community



Visit the [forums](#) for each competition to discuss methods & results.

Host a competition for...



-  **Analytics**
Get the world's best predictive model.
-  **Data Exploration**
Find the diamonds in your data.
-  **Recruitment**
Uncover objectively brilliant candidates.
-  **Education**
Free, powerful classroom competitions.

Calling all data-driven startups. "Let the crowd be your cofounder"
[Kaggle Startup Program now open!](#)


GE Quests

	GE Flight Quest Think you can change the future of flight?	36 days next deadline 90 teams \$250,000
	GE Hospital Quest Think it's possible to make hospital visits hassle-free? GE does.	38 days next deadline 100 teams \$100,000

Featured Competitions

	Traveling Santa Problem Solve ye olde traveling salesman problem to help Santa Claus deliver his presents	8.0 days next deadline 300 teams \$3,000
	Heritage Health Prize Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)	2 months next deadline 1496 teams \$3 million

Research Competitions

	Visualize the State of Public Education in Colorado Using 3 years of school grading data supplied by the Colorado Department of Education and R-Squared Research, visually uncover	9.0 days next deadline \$5,000
---	--	-----------------------------------

On the Forums

Congratulations Anthony !!

Tuning model parameters

Predicting wine quality

Random Forest

Unable to download data file for GE

GEF submissions pending?

On the Blog

GE Hospital Quest: Milestone ...

1st Place: Observing Dark Wor...

A Bayesian approach to Observ...

Newsletter: Startups, Santa, ...

Let the Crowd be Your Cofound...

Newsletter: GE Industrial Int...

7 2 6 9 9 participants

1 9 6 9 8 6 entries



What are the Assignments that you will Choose from?

24

- Project*: (You can only do one of the two options)
 - A data mining project* (100 points possible)
 - A big data project (200 points possible)
- These two projects should be thought about in 3 Phases:
 - Project Identification
 - Data Manipulation
 - Presenting your results (Grad Students must document their work in a conference/journal-style paper, see example in the Syllabus)



What are the Assignments that you will Choose from?

25

- Animated Visualization: (50 points possible)
 - We provided an example of visual analytics in sports
 - Another example is highlighted in this video

Source: http://www.ted.com/talks/hans_rosling_at_state.html



What are the Assignments that you will Choose from?

26

- Class Blog: (260 points possible, click [here](#))
 - Bloggers:
 - At least 16 blogs in 8 weeks (zero credit if less than 16)
 - At least four of them should be tutorials
 - Purpose of the non-tutorials: is to bring to the attention of class (and the world) interesting readings, news stories, and visual-analytic tools relevant to DM/Big Data.
 - Participants:
 - Utilize the information gained from the blog and apply it to a different dataset.
 - Provide ~250 word comment (with screenshots, figures, and/or tables) showing what you have learned.
 - At least 8 posts (zero credit if less than 8)



What are the Assignments that you will Choose from?

27

- Exams ($\times 2$); each worth 80 points
 - Somewhat similar to the homework
 - There will be questions based on the blog
 - The exams are 75 mins each



The points total for each assignment group is provided below:

Homework, 6 total at	
10 points per homework	60 points
DM competitions, 2 total at	
50 points per competition	100 points
DM Project	100 points
Big Data Project	200 points
Conference-Style Paper	100 points [Mandatory for Grads doing a Project]
Visualization Assignment	50 points
Blog – Blogger, 20 max (16 min)	
10 points per post	200 points
Blog – Participant, 8 max (8 min)	
7.5 points per post	60 points
Exam I	80 points
Exam II	80 points
<hr/>	
930 points TOTAL [Assuming Big Data Project]	

Letter grades will be assigned using the following scale:

Letter Grade	Points
A (UG)	515 and above
A (G)	535 and above
B (UG)	465-514
B (G)	485-534
C (UG)	400-464
C (G)	420-484
F (UG)	399 and below
F (G)	419 and above

*Please note that I will round your grades to the *nearest* integer at the *end* of the course.



- Course calendar is the last page in the syllabus
- I expect that each student in this class will develop an expertise in one or multiple areas related to the topics that we discuss in class.
- You totally control your learning in this class.
 - Zig Ziglar, a motivational speaker/author once said *“people often say that motivation doesn't last. Well, neither does bathing - that's why we recommend it daily.”*
 - There are a ton of supplementary data that is available online
- Start learning about big data (or DM) today ☺



- UC Berkley class on Big Data and Twitter,
<http://www.youtube.com/playlist?list=PLE8C1256A28C1487F>
- The Linguistic Data consortium, <http://ldc.upenn.edu/>
- U Michigan compilation of several datasets,
<http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/subject.jsp>
- Wikipedia dumps, <http://dumps.wikimedia.org/enwiki/>
- Stanford collection of social media datasets,
<http://snap.stanford.edu/news.html>
- The global terrorism database, <http://www.start.umd.edu/gtd/about/> (Need to request data using a contact form)



- Public datasets on AWS,
http://aws.amazon.com/datasets?_encoding=UTF8&jiveRedirect=1
- Some suggested data mining projects with their datasets, <http://cms.uhd.edu/faculty/chenp/class/4319/project/index.html>
- Large health data sets, <http://www.ehdp.com/vitalnet/datasets.htm>
- Google Ngrams, <http://books.google.com/ngrams/>
- Analyzing twitter data webtool, <http://www.tweetcharts.com/>
- A great list of ted talks on using large datasets,
http://www.ted.com/playlists/56/making_sense_of_too_much_data.html.

