Predictive Analytics

Chapter One

Introduction to Predictive Analytics and Related Disciplines

Please note that you are not allowed to share these slides outside this course including public websites like coursehero and similar ones

Anasse Bari, Ph.D.

Outline

- Defining Predictive Analytics.
- Defining Data Science and Big Data.
- Introducing Skills needed for Predictive Analytics and Data Science.
- Highlighting Use-cases around Predictive Analytics.
- Introducing the Lifecycle of Data Analytics' Projects.
- Explaining Predictive Analytics problems and their relationship to Data Clustering, Data Classification, Link Analysis and Recommender Systems.
- Introducing Supervised and Unsupervised Learning.
- Defining Statistics, Machine Learning, Data Mining, and Business Intelligence.
- Introducing (briefly) Hadoop and MapReduce. *(There will be a separate chapter on both topics)*

Defining Predictive Analytics

- "Predictive analytics is a bright light bulb powered by your data -- Predictive Analytics is the art and science of using data to make better informed decisions. Predictive analytics helps you uncover hidden patterns and relationships in your data that can help you predict with greater confidence what may happen in the future, and provide you with valuable, actionable insights for your organization…."
  Bari et al, Predictive Analytics for Dummies 2017
- Predictive Analytics is the technology that learns from experience.

  Experience = Data = Recording of Events

  "Data is a deal killer at cocktail parties" Eric Siegel
- A technology that uncovers relationships and patterns within large volumes of data that can be used to predict behavior and events.
- Predictive analytics is a **branch of *data science*** that is forward looking; it uses past events to anticipate future outcomes.

8 Expert Answers to "What is Predictive Analytics?"

RADIUS Products Solutions Why Radius B2B Marketing Resources Company

So, we put together this list of 8 definitions, gathered from industry experts and other business resources.

So, what exactly is 'predictive analytics'?

1) Forrester, "Predictive Analytics Can Infuse Your Applications With an 'Unfair Advantage," 2015

"Techniques, tools, and technologies that use data to find models– models that can anticipate outcomes with a significant probability of accuracy."

2) Gartner, IT Glossary

"Predictive analytics describes any approach to data mining with four attributes:

- An emphasis on prediction (rather than description, classification or clustering)
- Rapid analysis measured in hours or days (rather than the stereotypical months of traditional data mining)
- An emphasis on the business relevance of the resulting insights (no ivory tower analyses)
- (increasingly) An emphasis on ease of use, thus making the tools accessible to business users."

3) Predictive Analytics For Dummies by Anasse Bari et al, 2014

"Predictive Analytics is the art and science of using data to make better informed decisions. Predictive analytics helps you uncover hidden patterns and relationships in your data that can help you predict with greater confidence what may happen in the future, and provide you with valuable, actionable insights for your organization...."

Read more at:
https://radius.com/2015/11/11/8-expert-answers-to-what-is-predictive-analytics/

Data Science
What is it?

Anasse Bari, Ph.D.

Defining Data Science

- We will consider the difference between *Unreal Data Science* and Real Data Science.
- Expanding on the R programming example is NOT real Data Science

  What is R? and *Why Knowing just R does not really make you a Data Scientist?*

  - R is an open source statistical programming language and environment that is at least ***20 years*** old -It is the successor of S+.
  - R was and *still* limited to in-memory data processing and has been very popular in the statistical community.
  - R was extended to other tech such as RHadoop. (R+Hadoop) to bypass its limitations (in-memory limitation)

References:
https://www.r-project.org/
Developing Analytic Talent: Becoming a Data Scientist, Dr. Vincent Granville, 2014 Johns and Wiley
https://cran.r-project.org/web/packages/.../Ch_introduction_to_R.pdf

Defining Data Science

- Hadoop and MapReduce

Hadoop is an *implementation* of MapReduce like Java is an *implementation* of object oriented programming.

- NoSQL

NoSQL means not only SQL. It used to describe database or data management systems that support new more efficient ways to access data. (NoSQL RDMS examples: MongoDB, MarkLogic..)

- How about Graph Databases?

Graph databases that rely on the concepts of nodes and edges to manage and access data like special data. Examples?

References:
- https://www.r-project.org/
- Developing Analytic Talent: Becoming a Data Scientist, Dr. Vincent Granville, 2014 Johns and Wiley
- https://cran.r-project.org/web/packages/.../Ch_introduction_to_R.pdf

Defining Data Science

- Hadoop is perceived as a new technology that was inspired by Google.
- The reality is:
  - Distributed architectures happen to exist before Google, MapReduce..
  - Hash Joins (a type of no sql joins). HW#0
  - Today modern RDMBS offers hash joins instead of SQL joins.

What did we learn?

Hadoop, MapReduce, NoSQL,and Python (Which are Perl successors) and have their roots in systems and techniques that started to be developed decades ago and have matured over the last 10 years.

Data Science is MORE THAN THAT!

You can be a real data scientist and have none of these skills.

References:
- https://www.r-project.org/
- Developing Analytic Talent: Becoming a Data Scientist, Dr. Vincent Granville, 2014 Johns and Wiley
- https://cran.r-project.org/web/packages/.../Ch_introduction_to_R.pdf

Defining Data Science

- To be a real data scientist you need to have the following array of skills:
  - Business acumen
  - Real Big Data Expertise (for instance, you can easily process 50million row data set in 2 hours)
  - Ability to Sense data
  - A distrust of models
  - Knowledge of the curse of big data
  - Ability to communicate and understand problems management is trying to solve.
  - Ability to correctly assess lift or ROI on the salary paid to you
  - Ability to quickly identify a simple, robust and scalable solution to a problem.
  - Ability to convince and drive management in the right direction, sometime against its will, for the benefit of the company it uses and shareholders.

References:
https://www.infoq.com/news/2016/09/data-needed-data-science
Developing Analytic Talent: Becoming a Data Scientist, Dr. Vincent Granville, 2014 Johns and Wiley
Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data 1st Edition by EMC Education Services, 2014

Defining Data Science

- To be a real data scientist you need to have the following array of skills:
  - A real passion for analytics
  - Real applied experience with success stories
  - Data architecture knowledge
  - Data gathering and cleaning skills
  - Computational complexity basic – how to develop robust, efficient, scalable and portable architectures
  - Good Knowledge of Algorithms

  ….

Defining Data Science

A data scientist is also a strategist – they can develop ***actionable insights*** that make ***business impact***. This requires ***creativity*** to develop analytics solutions based on ***business constraint*s** and ***limitations***.

Reference:

- Developing Analytic Talent: Becoming a Data Scientist, Dr. Vincent Granville, 2014 Johns and Wiley

Defining Data Science

Basic Skills needed to be a data scientist?

Mathematical Skills:

- Algebra, including basic matrix theory
- Calculus, logarithm, exponential and power functions. Differential equations, integrals.

Basic Statistics:

- Fundamentals of statistics and probability, familiarity with random variables probability, mean variance, percentiles, cross validation, goodness of fit.

Defining Data Science

Basic Skills needed to be a data scientist?

Technical Skills:

- Includes R, Python (or Perl), Java, Excel, SQL, visualization tools, FTP, basic unix commands (sort grep, head, tail cran jobs..)
- Understanding of distributed systems (data transfers between hard disk and memory and over the net..)
- Basic Knowledge of web crawlers to access unstructured data found on the net.
- Knowledge of Major Analytics Platforms: Weka, R, Revolution Analytics, RapidMiner, StatSoft, IBM Watson, Analytics APIs..

So .. what is *Data Science*?

Defining Data Science

First, let us review the definition of science

What is Science?

"Science is the pursuit and application of knowledge and understanding of the natural and social world following a systematic methodology based on evidence"

The Science Council's definition of science - http://www.sciencecouncil.org/definition

Defining Data Science

Second, now that we defined science, let us review the definition of data

What is Data?

Facts that you can draw conclusions from.

"Information in raw or unorganized form (such as alphabets, numbers, or symbols) that refer to, or represent, conditions, ideas, or objects.

Data is limitless and present everywhere in the universe."

http://www.businessdictionary.com/definition/data.html#ixzz3ySbUZG5Z

http://schoolofdata.org/handbook/courses/what-is-data/

Defining Data Science

Second, now that we defined science, let us review the definition of data

See more at: http://schoolofdata.org/handbook/courses/what-is-data/#sthash.MPc6jSIa.dpuf

What is Data?

Qualitative data is everything that refers to the quality of something: A description of colors, texture and feel of an object , a description of experiences, and interview are all qualitative data.

Quantitative data is data that refers to a number. E.g. the number of golf balls, the size, the price, a score on a test etc.

However there are also other categories that you will most likely encounter:

Categorical data puts the item you are describing into a category: In our example the condition "used" would be categorical (with categories such as "new", "used" ,"broken" etc.)

Discrete data is numerical data that has gaps in it: e.g. the count of golf balls. There can only be whole numbers of golf ball (there is no such thing as 0.3 golf balls). Other examples are scores in tests (where you receive e.g. 7/10) or shoe sizes.

Continuous data is numerical data with a continuous range: e.g. size of the golfballs can be any value (e.q. 10.53mm or 10.54mm but also 10.536mm), or the size of your foot (as opposed to your shoe size, which is discrete): In continuous data, all values are possible with no gaps in between.

Defining Data Science

More Data Types:

- Structured Data
- Unstructured Data
- Semi-Structured Data

Unstructured Data

Data for Humans

A plain sentence – "we have 5 white used golf balls with a diameter of 43mm at 50 cents each" – might be easy to understand for a human, but for a computer this is hard to understand. The above sentence is what we call unstructured data. Unstructured has no fixed underlying structure – the sentence could easily be changed and it's not clear which word refers to what exactly. Likewise, PDFs and scanned images may contain information which is pleasing to the human-eye as it is laid-out nicely, but they are not *machine-readable*.

http://schoolofdata.org/handbook/courses/what-is-data/

Structured Data

Data for Computers

Computers are inherently different from humans. It can be exceptionally hard to make computers extract information from certain sources. Some tasks that humans find easy are still difficult to automate with computers. For example, interpreting text that is presented as an image is still a challenge for a computer. If you want your computer to process and analyse your data, it has to be able to read and process the data. This means it needs to be structured and in a *machine-readable* form.

One of the most commonly used formats for exchanging data is CSV. CSV stands for comma separated values. The same thing expressed as CSV can look something like:

```
“quantity”, “color”, “condition”, “item”, “category”, “diameter (mm)”, “price per u
nit (AUD)”
5,”white”,”used”,”ball”,”golf”,43,0.5
```

This is way simpler for your computer to understand and can be read directly by spreadsheet software. Note that words have quotes around them: This distinguishes them as text (string values in computer speak) – whereas numbers do not have quotes. It is worth mentioning that there are many more formats out there that are structured and machine readable.

http://schoolofdata.org/handbook/courses/what-is-data/

Defining Data Science (Cont’d)

- Semi-Structured Data

What is Data Science?

Data Science *is the art and science for analyzing Big Data for the purpose of extracting insights and forward-insights in order to create new opportunities for organizations and individuals to derive new value and create competitive advantage form their most valuable asset: DATA.*

References:
Developing Analytic Talent: Becoming a Data Scientist, Dr. Vincent Granville, 2014 Johns and Wiley
Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data 1st Edition by EMC Education Services, 2014
Predictive Analytics for Dummies, A.Bari et. al 2014

What is Data Science?

Data Science *is broader than predictive analytics, data mining, statistics and machine learning, it encapsulates data integration, data gathering, data molding, data mining, data visualization, data architecture and analytics evaluations with metrics to measure ROI.*

References:
Developing Analytic Talent: Becoming a Data Scientist, Dr. Vincent Granville, 2014 Johns and Wiley
Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data 1st Edition by EMC Education Services, 2014
Predictive Analytics for Dummies, A.Bari et. al 2016

Big Data

What is it?

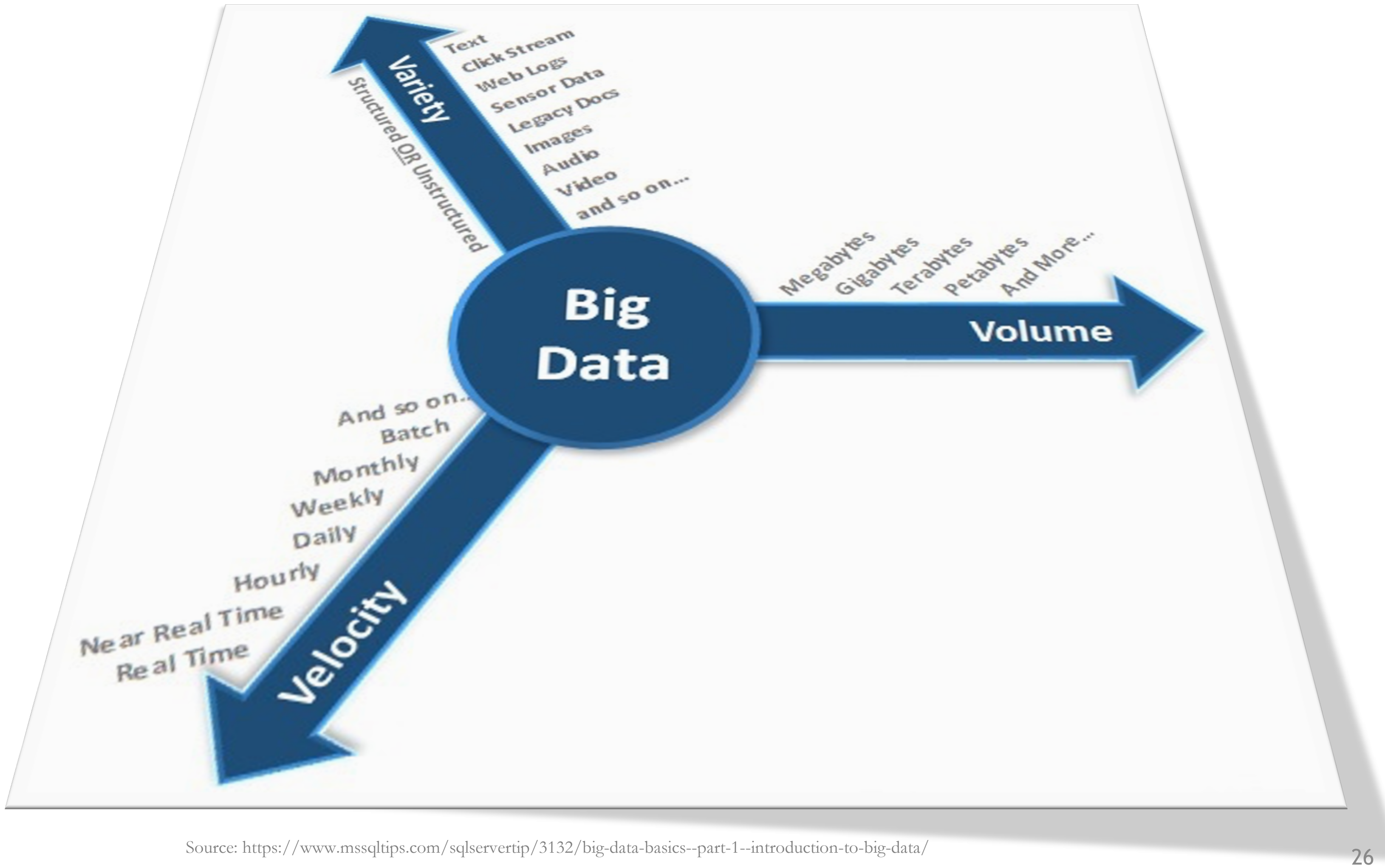
What is Big Data?

Are 10GBs of Data can be considered as Big Data?

Is 1TB of Data can be considered as Big Data?

Is Big Data has to be Big in Volume?

How Big the Data should be to be called Big Data?



Source: https://www.mssqltips.com/sqlservertip/3132/big-data-basics--part-1--introduction-to-big-data/

What is big data?

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone.

This data comes from everywhere:

sensors used to gather climate information,
posts to social media sites,
digital pictures and videos,
purchase transaction records,
and cell phone GPS signals to name a few.

This data is "big data."

Source: IBM Labs

Volume

Enterprises are awash with ever-growing data of all types, easily amassing terabytes—even petabytes—of information.

Turn 12 terabytes of Tweets created each day into improved product sentiment analysis

Convert 350 billion annual meter readings to better predict power consumption

Source: IBM Labs

Velocity

Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

Scrutinize 5 million trade events created each day to identify potential fraud

Analyze 500 million daily call detail records in real-time to predict customer churn faster

Source: IBM Labs

Variety

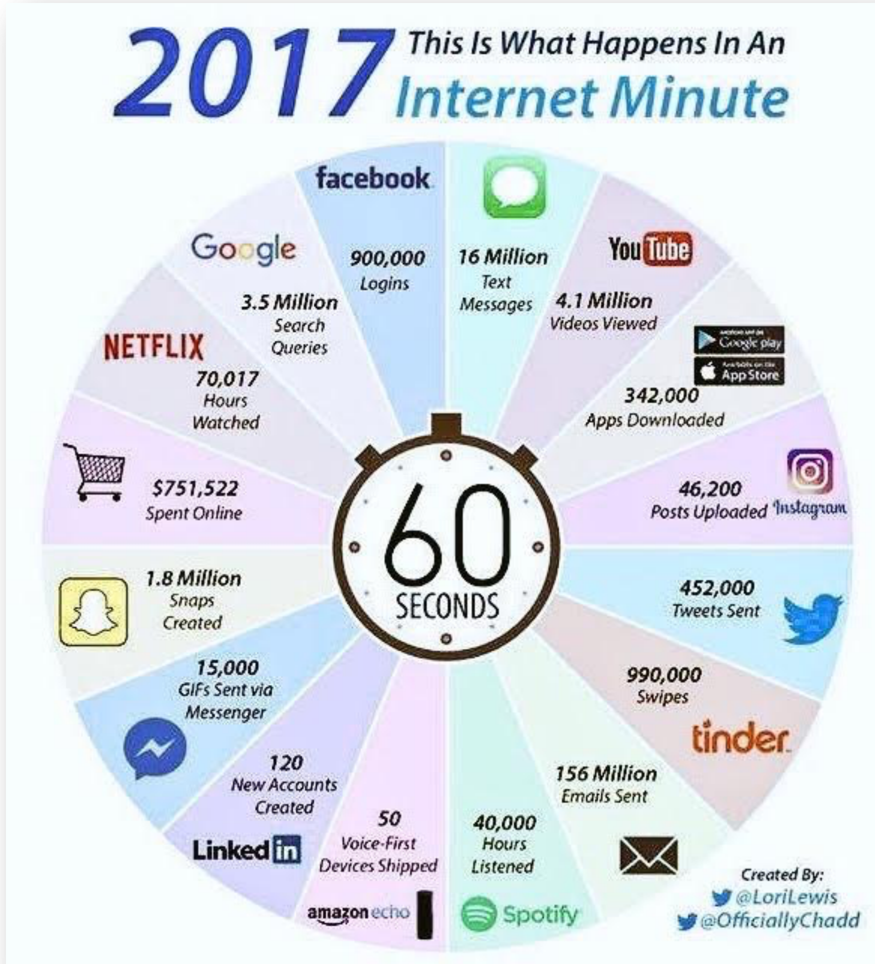
- Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. **New insights are found when analyzing these data types together.**

    - Monitor 100s of live video feeds from surveillance cameras to target points of interest

    - Exploit the 80% data growth in images, video and documents to improve customer satisfaction

Source: IBM Labs

Source: GoGlobe.com


2017 This Is What Happens In An Internet Minute
facebook
900,000 Logins
16 Million Text Messages
YouTube
4.1 Million Videos Viewed
Google play
App Store
342,000 Apps Downloaded
46,200 Posts Uploaded Instagram
452,000 Tweets Sent
990,000 Swipes
tinder
156 Million Emails Sent
40,000 Hours Listened
Spotify
50 Voice-First Devices Shipped
amazon echo
120 New Accounts Created
LinkedIn
15,000 GIFs Sent via Messenger
1.8 Million Snaps Created
$751,522 Spent Online
NETFLIX
70,017 Hours Watched
Google
3.5 Million Search Queries
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd

Main Types of (Big)Data: Structured and unstructured data

Structured Information

- Information that is stored in databases
- Well-formed documents
    - XML

Unstructured information

- Web pages , presentations, documents (PDF, Doc..)
- Emails, images, videos
- Blogs
- Log files

Semi-structured

Use cases at a Glance

Recent Predictive Analytics Use-cases

- Bollen, Johan, Huina Mao, and Xiaojun Zeng. "**Twitter mood predicts the stock market.**" Journal of Computational Science 2.1 (2011): 1-8.

Abstract—**Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making.** Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. We analyze the text content of daily Twitter feeds by two mood tracking tools, namely Opinion Finder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. **A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis is that public mood states,as measured by the Opinion Finder and GPOMS mood time series, are predictive of changes in DJIA closing values.** Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6%

Recent Predictive Analytics Use-cases

- Bollen, Johan, Huina Mao, and Xiaojun Zeng. **"Twitter mood predicts the stock market."** Journal of Computational Science 2.1 (2011): 1-8.

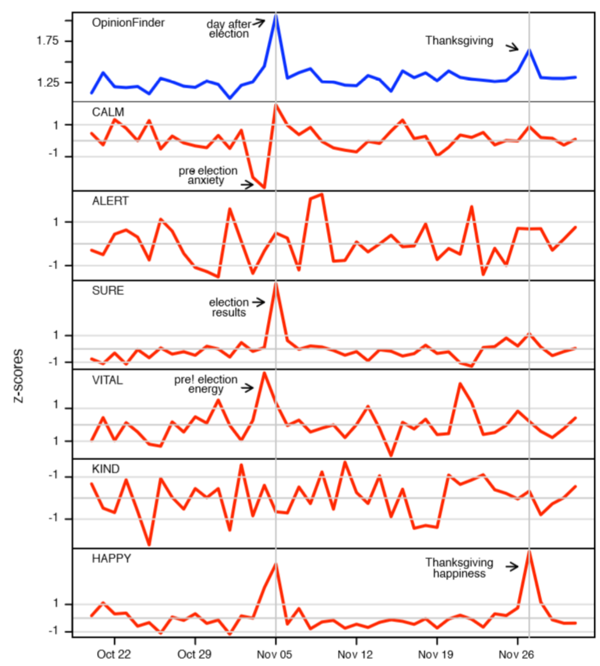


Fig. 2. Tracking public mood states from tweets posted between October 2008 to December 2008 shows public responses to presidential election and thanksgiving.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. **"Twitter mood predicts the stock market."** Journal of Computational Science 2.1 (2011): 1-8.

Dow Jones Industrial Average (DJIA)

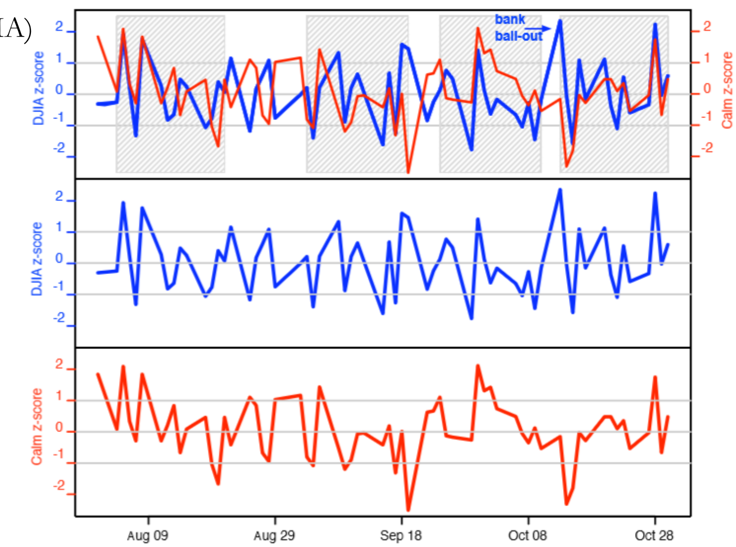


Fig. 3. A panel of three graphs. The top graph shows the overlap of the day-to-day difference of DJIA values (blue: $\mathbb{Z}_{D_t}$) with the GPOMS' Calm time series (red: $\mathbb{Z}_{X_t}$) that has been lagged by 3 days. Where the two graphs overlap the Calm time series predict changes in the DJIA closing values that occur 3 days later. Areas of significant congruence are marked by gray areas. The middle and bottom graphs show the separate DJIA and GPOMS' Calm time series.

Twitter Mood and it Correlation with the Stock Market using ***Granger Causality***

"The basic "Granger Causality" definition is quite simple.

Suppose that we have three terms, X_t, Y_t, and W_t, and that we first attempt to forecast X_{t+1} using past terms of Y_t and W_t. We then try to forecast X_{t+1} using past terms of X_t, Y_t, and W_t. If the second forecast is found to be more successful, according to standard cost functions, then the past of Y appears to contain information helping in forecasting X_{t+1} that is not in past X_t or W_t. … Thus, Y_t would "Granger cause" X_{t+1} if (a) Y_t occurs before X_{t+1} ; and (b) it contains information useful in forecasting X_{t+1} that is not found in a group of other appropriate variables."

- Clive Granger, 2003 Nobel Laureate in Economics.

Granger Causality

- Attempt to get a better forecast using *other times series* to forecast a specific time series.
- Many times series move simultaneously (e.g. NY stock market can be affected by the fluctuation (irregular rising or falling) of the London stock market)
- Many times series are highly correlated (especially in the financial domain e.g. fund manager managing several asset classes)
- First: Correlation Test - perform a test to the correlation between time series (e.g Pearson Correlation) a correlation higher than 0.5 >> is promising)
- ***Correlation does NOT mean Causation (therefore the use of Granger Causality)***
- Predicting Y(t) using Y(t-1) and X(t-1) (X and Y are correlated time series)

More Use-cases

- **" The value of data analytics in mergers and acquisitions"** Brian Gentile Mon 18 May 2015. The Stack

"…Whilst the acquisition is going through the buying process, predictive data analytics techniques can also be used to see ***how the market will likely respond after a deal is made.*** Ultimately, such specific analysis offers the security required to justify such an important activity..." Source: Brian Gentile Mon 18 May 2015

More Use-cases

- "**How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did**" Kashmir Hill FORBES, 2012
  "..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…" source: Kashmir Hill FORBES, 2012

More Predictive Analytics Use-cases

Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, Takeshi Sakaki et. At The University of Tokyo, 2010

Abstract -- We investigate the **real-time interaction of events** such as earthquakes in Twitter and propose an algorithm to monitor tweets and to detect a target event. To detect a target event, we devise a classifier of tweets **based on features such as the keywords in a tweet, the number of words, and their context**. Subsequently, we produce a probabilistic spatiotemporal model for the target event that can find the center and the trajectory of the event location. We consider each **Twitter user as a sensor** and apply Kalman filtering and particle filtering, which are widely used for location estimation in ubiquitous/pervasive computing.

WWW 2010 • Full Paper

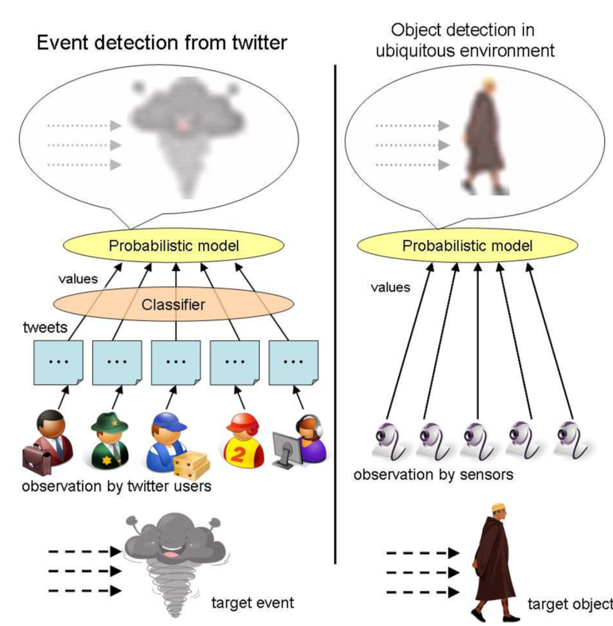


Figure 3: Correspondence between event detection from Twitter and object detection in a ubiquitous

IEEE SPECTRUM

Follow on:

To Predict Flu Outbreak, Just Google It

By Charles Choi
Posted 29 Jan 2015 | 12:00 GMT

Photo: iStockphoto

By combining insights from Google Flu Trends with data from the CDC, scientists now say they can predict the spread of flu a week into the future in the United States.

Each year, 250,000 to 500,000 people die of influenza worldwide, with 3,00 to 50,000 of those fatalities happening in the United States. These deaths are largely preventable by using flu shots, but the CDC must have up-to-date knowledge about where influenza is happening to make sure these vaccines get to where they are needed.

The CDC continuously monitors both the number of doctor visits attributed t flu-like illness as well as the number of patient samples that test positive for influenza. However, it can take a long time to collect and analyze all this activity, resulting in data that is typically up to two weeks out of date once it's

Each year, 250,000 to 500,000 people die of influenza worldwide, with 3,000 to 50,000 of those fatalities happening in the United States. These deaths are largely preventable by using flu shots, but the CDC must have up-to-date knowledge about where influenza is happening to make sure these vaccines get to where they are needed.

The CDC continuously monitors both the number of doctor visits attributed to flu-like illness as well as the number of patient samples that test positive for influenza. However, it can take a long time to collect and analyze all this activity, resulting in data that is typically up to two weeks out of date once it's made available.

Recently Google unveiled Google Flu Trends as a way to predict flu levels in real-time — two weeks earlier than the CDC — by analyzing how often people Google search terms related to influenza. However, while Google Flu Trends is promising, this "big data" approach has made dramatic errors — for example, it predicted double the number of doctors' visits from the flu in 2013 than really happened. This is because people can search Google for information about influenza when they do not actually have the flu if they are enticed by factors such as increased media attention related to illness.

Now researchers at the University of California, San Diego, the combo of Google and CDC data could do better than predict U.S. flu levels in real-time — it could forecast a week into the future.

The scientists used CDC data to determine which U.S. regions experienced influenza outbreaks at similar times in the past. The flu is best at spreading between these areas due to factors such as geographic proximity. This information helped correct exaggerations in Google's estimates and also shed light on future influenza levels by revealing which the virus might spread.

"Big data does not always work the best in a vacuum," said study lead author Michael Davidson, a data scientist at the University of California, San Diego. "By combining big data with traditional sources of data, we can often do better than by relying on big data alone."

In the future, scientists might combine other sources of flu data, such as Wikipedia page visits, with Google Flu Trends and CDC data for even more accurate and timely estimates, Davidson said. He and his colleagues detailed their findings online Jan. 29 in the journal *Scientific Reports*.

http://spectrum.ieee.org/tech-talk/computing/it/google-cdc-partner-to-predict-flu-spread

However..


DAVID LAZER AND RYAN KENNEDY SCIENCE 10.01.15 7:00 AM

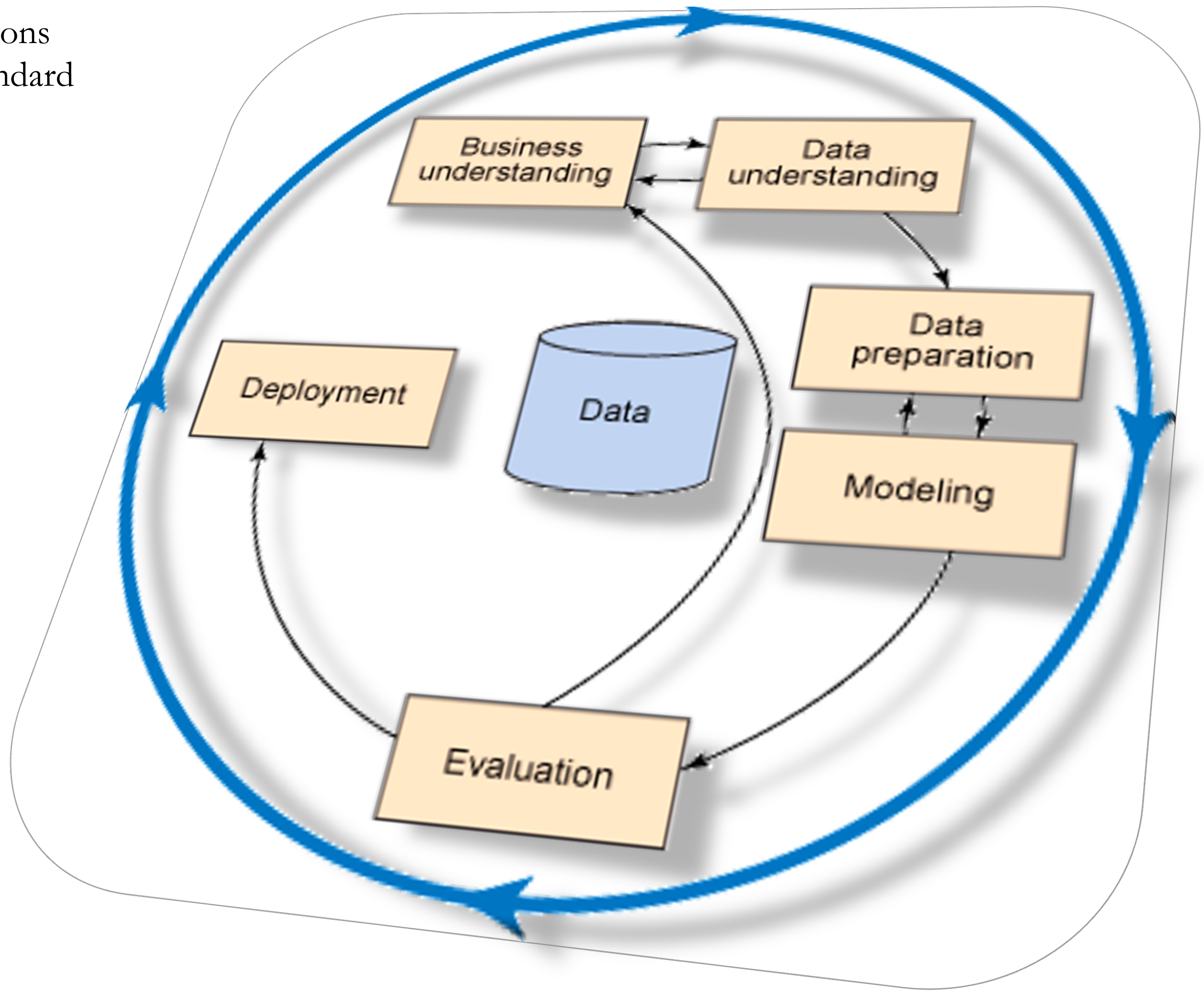
WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS

[Read More at: http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/](http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/)

Data Analytics Project' Lifecycle

it is different from the software engineering lifecycle

Most of the Data Analytics Lifecycle definitions has been inspired for the Cross Industry Standard Process (CRISP-DM)



Cross Industry Standard Process for Data Mining (CRISP-DM) is a data mining process model that describes commonly used approaches that expert data miners use to tackle problems.

Source:
http://www.ibm.com/developerworks/bpm/library/techarticles/1407_chandran/index.html

Business Understanding (also sometime it revers to the discovery phase) involves determining and defining business objectives in business terms, translating these to data mining goals and making a project assessment and plan.

Reference: http://www.ibm.com/developerworks/bpm/library/techarticles/1407_chandran/index.html

Data Understanding involves collecting initial data, describing the data in terms of amount, type and quality of data, exploring data using available tools and verifying data quality.

Reference: http://www.ibm.com/developerworks/bpm/library/techarticles/1407_chandran/index.html

Data Preparation is an important and time-consuming part of data mining which can take up 50–80% of the project's time and effort. It involves selecting data to include, cleaning data to improve data quality, constructing new data that may be required, integrating multiple data sets, and formatting data.

Model Planning and Building involves selecting suitable modeling techniques, generating test designs to validate the model, building predictive models and assessing these models. Case of Predictive Analytics: A predictive model is a mathematical function that predicts the value of some output variables based on the mapping between input variables. Historical data is used to train the model to arrive at the most suitable modeling technique. For example, a predictive model might predict the risk of developing a certain disease based on patient details. Some commonly used modeling techniques are as follows:

- Regression analysis that analyzes the relationship between the response or dependent variable and a set of independent or predictor variables.
- Decision trees that help explore possible outcomes for various options.
- Cluster analysis that groups objects into clusters to look for patterns.
- Association techniques that discover relationships between variables in large databases.

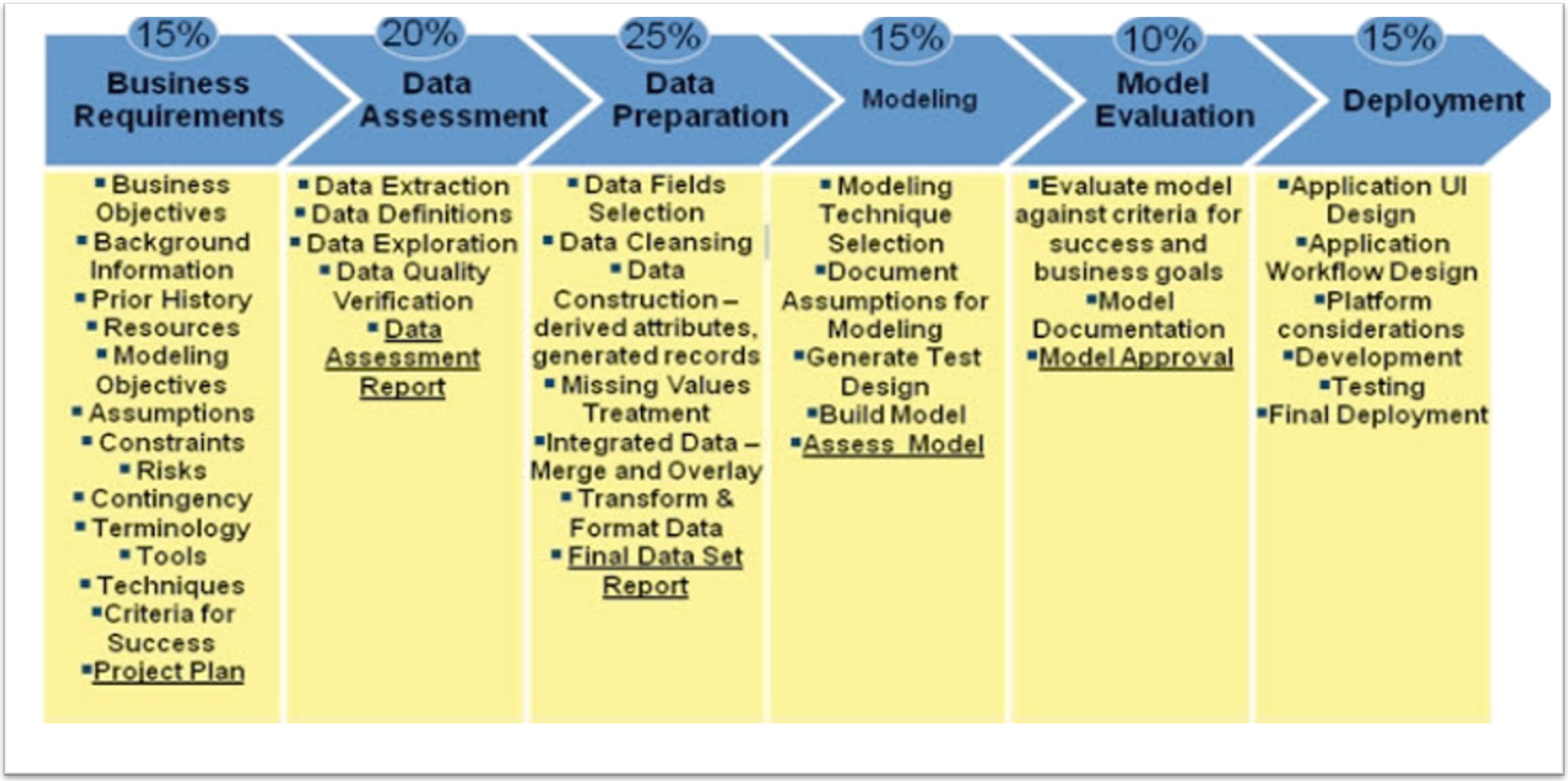
Source: http://www.ibm.com/developerworks/bpm/library/techarticles/1407_chandran/index.html

Evaluation involves evaluating the results against the business success criteria defined at the beginning of the project.

Reference: http://www.ibm.com/developerworks/bpm/library/techarticles/1407_chandran/index.html

Deployment involves consolidating the findings, determining what might be deployed and planning the monitoring and maintenance required to keep the model relevant.

Reference: http://www.ibm.com/developerworks/bpm/library/techarticles/1407_chandran/index.html

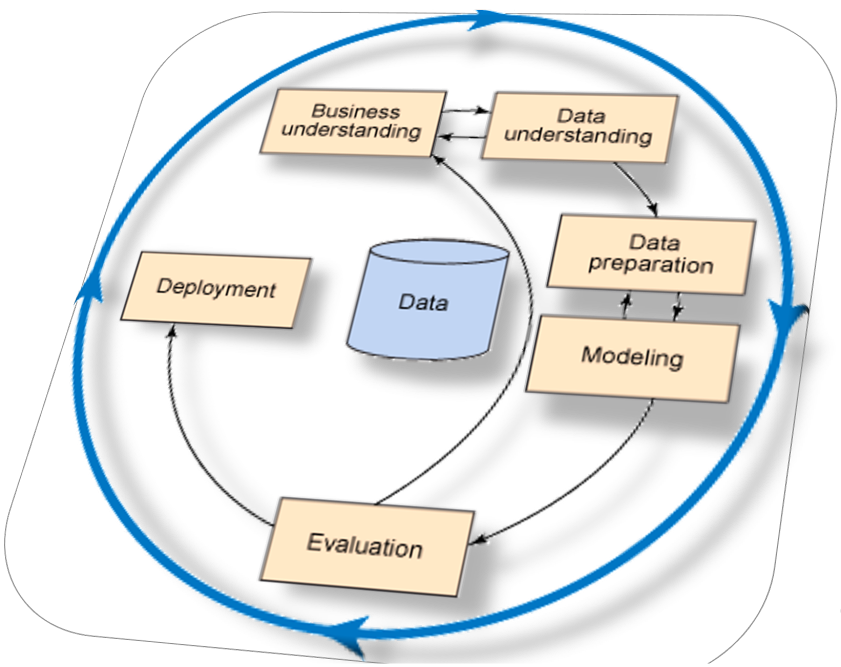


Source: http://www.ibm.com/developerworks/bpm/library/techarticles/1407_chandran/index.html

Roles (?) : Data scientist, Project Manager, BI analyst…

Doing Analytics is very Scientific

Analytics Skills = a * Business/Domain Knowledge/Creativity) +b * Statistics + c* Data mining algorithms + d* Machine Learning Algorithms + e* computer programming + f* Big Data Analytics tools and paradigms (Hadoop, MapReduce, NoSQL…) + g* Data Visualization + h* Linear Algebra + i* SQL (Knowledge of RDMBS, SQL…) + …



Scientific methodology includes the following:

Hypothesis

Objective observation: Measurement and data (possibly although not necessarily using mathematics as a tool)

Evidence

Experiment and/or observation as benchmarks for testing hypotheses

Induction: reasoning to establish general rules or conclusions drawn from facts or examples

Repetition

Critical analysis

Verification and testing: critical exposure to scrutiny, peer review and assessment

Assigned Readings and references are attached to this chapter

Reference: http://www.ibm.com/developerworks/bpm/library/techarticles/1407_chandran/index.html