Chapter One

Introduction to Predictive Analytics and Related Disciplines
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’ Project.
- 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 2016

- 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?”

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

• One way to understand what something is, is to decipher a clear picture of what is NOT.

• Defining and Understanding Data Science is No exception.

• Let us start by investigating what Data Science is NOT.

• The term has been much abused and a lot of hype arounds big data and data science.
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

Copyrights © Anasse Bari
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 50 million 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:
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 constraints* and *limitations*.

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
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.

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 conclusion 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

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.g. 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.

See more at: [http://schoolofdata.org/handbook/courses/what-is-data/#sthash.MPc6jSIa.dpuf](http://schoolofdata.org/handbook/courses/what-is-data/#sthash.MPc6jSIa.dpuf)
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/
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 unit (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

- 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 from their most valuable asset: DATA.

Data Science is brother 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 2014
Skills for Predictive Analytics and Data Science

**Data Analytics Skills** = a * (Business/Domain Knowledge/Creativity/Passion for Analytics) + 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 RDBMS, SQL…) + …

*a,b,c,d,e… are heuristics ranging between 0 and 1 and are adjustable depending on the project’s requirements.*
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?
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
Please watch:
https://www.youtube.com/watch?v=iFyGuvyesw4
Two Main Type 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
Use cases at a Glance

might be interesting to inspire you to get ideas for your semester long project..
Recent Predictive Analytics Use-cases


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


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

Fig. 3. A panel of three graphs. The top graph shows the overlap of the day-to-day difference of DJIA values (blue: $Z_{D_t}$) with the GPOMS’ Calm time series (red: $Z_{X_t}$) 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.
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.

Figure 3: Correspondence between event detection from Twitter and object detection in a ubiquitous environment.
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 3,000 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 excited 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..

Read More at: 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.

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.

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.

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.

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

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

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 RDBMS, SQL…) + …

Scientific methodology includes the following:

**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

---

Copyrights © Anasse Bari

[http://www.sciencecouncil.org/definition](http://www.sciencecouncil.org/definition)
Introduction to Predictive Analytics Problems

Anasse Bari, Ph.D.
Consider Data Clustering  *(In-class discussion)*

“Pattern Recognition and Machine Learning” By Chris Bishop
Data Analytics

• Data is very large, diverse, and rapidly changing (volume, velocity and variety)

• We aim to make sense of it, understand it and make decisions based on it.

• We want to discover knowledge within it (Knowledge Discovery)

• We enlist the help of major disciplines such as:
  • Statistics, Machine Learning, Data Visualization, Data Mining, Business Intelligence, Predictive Analytics and Data Science.
Why *Data Analytics* now?

- Much more operational data is being created and captured
  - Enterprise software
    - ERP: Enterprise Resource Planning
    - CRM: Customer Relationship Management
    - SCM: Software Configuration Management
- Much more unstructured data is being captured and stored
  - Web clicks
  - Facebook
  - Instagram
  - Twitter
Business Need

- Consider loans and credit cards applications approvals

- Banks do NOT know their customers. The only knowledge a bank has is their information stored in the computer and maybe other external data about their customers (?)

- Credit agencies and banks collect a lot of customers’ behavioral data from many sources.

- This information is used to predict the chances of a customer paying back a loan.
Business Need

• Need to make sense of business data

• Need to better serve customers by learning from past interactions

• The intelligence hidden in data could give you **the edge in a competitive business world**
Decision Making

- Management are interested in summary data to identify trends, challenges, patterns, profits strategies and opportunities.

- Management look into the data for *decision-support*

- Data driven decisions
  - Consistent and bias-free decisions
..and The Future is Wide Open

• Predictive analytics is implemented in many areas of business and all of them have potential for substantial growth.

• Especially important: ample room for growth using predictive analytics as a strategic asset.
Current Database Solution are designed for structured data.

Optimized to **answer known questions quickly**

**Schemas** dictate form/context

**Difficult to adapt** to new data types and new questions

**Expensive** at petabyte scale

Source: IBM Labs

Copyrights @ Anasse Bari

Copyrights @ Anasse Bari
Hadoop

- An operating system for (1) storage and (2) processing large data
- An Ecosystem
- Low cost, reliable scale-out architecture
- Distributed computing
- Proven success in many large businesses and organizations
Hadoop’s Ecosystem

- **Apache Hive**: SQL-like language and metadata repository
- **Apache Pig**: High-level language for expressing data analysis programs
- **Apache HBase**: The Hadoop database. Random, real-time read/write access
- **Apache Zookeeper**: Highly reliable distributed coordination service
- **Oozie**: Server-based workflow engine for Hadoop activities
- **Hue**: Browser-based desktop interface for interacting with Hadoop
- **Flume**: Distributed service for collecting and aggregating log and event data
- **Sqoop**: Integrating Hadoop with RDBMS
- **Apache Whirr**: Library for running Hadoop in the cloud

https://hadoop.apache.org/
## Big data Architecture at a Glance

<table>
<thead>
<tr>
<th>Function</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manage &amp; store huge volume of any data</td>
<td>Hadoop File System, MapReduce</td>
</tr>
<tr>
<td>Structure and control data</td>
<td>Data Warehousing</td>
</tr>
<tr>
<td>Manage streaming data</td>
<td>Stream Computing</td>
</tr>
<tr>
<td>Analyze unstructured data</td>
<td>Text Analytics Engine</td>
</tr>
<tr>
<td>Integrate and govern all data sources</td>
<td>Integration, Data Quality, Security, Lifecycle Management, MDM</td>
</tr>
</tbody>
</table>

https://hadoop.apache.org/
Big data scenario

- The processing is hosted in distributed servers -- Cloud(?)
- Data is stored in distributed storage
- Data is stored and indexed in high performance schema free databases
- Distributed programming model (Map Reduce)
- Performing Analytics and semantic processing on the data

Value of big data

Drive incremental revenue
• Predict customer behavior across all channels
• Understand and monetize customer behavior

Improve operational effectiveness
• Machines/sensors: predict failures, network attacks
• Financial risk management: reduce fraud, increase security

Reduce data warehouse cost
• Integrate new data sources without increased database cost
• Provide online access to ‘dark data’

Major PA’s Related Disciplines

Let’s define Machine Learning, Data Mining, Statistics…

Anasse Bari, Ph.D.
Defining Statistics

- A branch of mathematics

- Aim at analyzing whether the data is significant / meaningful

- Statistics is able to determine the probability that similar results would occur if the same conditions occurs

- Descriptive statistics are methods for organizing and summarizing data.
  - Graphical techniques (tables or graphs)
  - Numerical techniques (mean and median)

- Inferential statistics are methods for using sample data to make general estimates, predictions, or decisions (inferences) about populations.
Defining Machine learning

- Machines can learn from input data and past experience.

- Widely used within the domain of Artificial Intelligence.

- Builds adaptive systems to improve their efficiency.
Defining Machine learning

- Computer science asks “How can we build systems that solve problems?”

- Statistics asks “What can be inferred from data plus a set of modeling assumptions? With what reliability?”

- Machine learning “incorporates additional questions to the problem solving, inferences, learning and about what computational architectures and algorithms can be used to most effectively capture, store, index, retrieve and merge these data, how multiple learning subtasks can be orchestrated in a larger system”

Defining Data mining

- Efficient automated discovery of previously unknown patterns in large volumes of data.

- Also known as Knowledge Discovery in Databases (KDD)

- Businesses are mostly interested in discovering past patterns to predict future behaviour.
Defining Data mining

- KDD is the “non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. Fayad

https://mitpress.mit.edu/books/advances-knowledge-discovery-and-data-mining

- **Valid**: the patterns hold in general

- **Novel**: we did not know the pattern beforehand

- **Useful**: we can devise actions from the patterns

- **Understandable**: we can interpret the patterns

What is Business Intelligence?

- Business intelligence enables the business to make intelligent, fact-based decisions.

Source: ISACA, https://www.isaca.org
Evolution of Business Intelligence

- **1990’s**
  - BI Reporting
  - OLAP & Data Warehouse
  - Business Objects, SAS, Informatica, Cognos, other SQL Reporting Tools

- **2000’s**
  - Interactive Business Intelligence
  - In-memory RDBMS
  - QlikView, Tableau, HANA

- **2010’s**
  - Big Data
  - Batch Processing & Distributed Data Store
  - Hadoop/Spark; HBase/Cassandra
  - Real Time
  - Graph Databases
Predictive Analytics questions

- How can you use historical data to make informed decisions about future events?

- What are the attributes that have predictive power of certain behavior?

- What anomalies may exist within the data?

- Is there any correlation or relationship among the various attributes in your data?

- Is there a trend that is underway that can be detected in your data?
Comparison

- Statistics
  - more theory-based
  - more focused on testing hypotheses

- Machine learning
  - more heuristics
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics

- Data mining and knowledge discovery
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results

- Predictive analytics
  - uses data mining techniques
  - makes a prediction along with a score of probable outcome

- Data Science
  - All of the above with consider of Big Data challenges
  - Deriving actionable insights
Overlap

- They help us organize and interpret the data

- There is a lot of overlap
  - Distinction can be fuzzy and very unclear
Techniques
Data mining methods

- **Supervised Learning**: learning from labelled data
  - Classification, regression, prediction.
  - The training data is labeled with the class of the observations
  - New data is classified based on the training set

- **Unsupervised Learning**: learning from unlabelled data
  - Clustering, visualisation, dimensionality reduction
  - The class labels of training data is unknown
  - The aim is to find clusters in the data

- **Associations**
  - Market basket analysis
Supervised learning

Source: http://www.astroml.org/sklearn_tutorial/general_concepts.html
Unsupervised learning

http://www.astroml.org/sklearn_tutorial/general_concepts.html
Analytics Techniques

- Association: A & B occur frequently
- Link analysis: finding relationships
- Classification: predicting an item class
- Regression: predicting a continuous value
- Clustering: finding groupings in data
- Summarization: describing groups in data
- Trend & deviation: finding changes
- Visualization: to facilitate discovery
Mining Associations

- Association analysis involves discovery of relationships or correlations among a set of items.

- Discovering that personal loans are repaid with 80% confidence when the person owns her/his home.

- Discovering that 80% of customers who buy cereal and milk also buy bread and 5% of customers buy all these products together.

- Finding association rules is just the beginning in a data mining effort.

- Challenge: any large dataset can lead to a very large number of association rules.
- Many of these rules are uninteresting, trivial, or redundant.
Classification

• A set of training objects each with a number of attribute values are given to the classifier.

• The classifier formulates rules for each class in the training set.

• The rules are subsequently applied to classify new objects and determine which class the new objects belong to.

• Classification is used for predicting class labels of new data objects.
Classification Process (1): Model Construction

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>YEARS</th>
<th>TENURED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike</td>
<td>Assistant Prof</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>Mary</td>
<td>Assistant Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Bill</td>
<td>Professor</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>Jim</td>
<td>Associate Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Dave</td>
<td>Assistant Prof</td>
<td>6</td>
<td>no</td>
</tr>
<tr>
<td>Anne</td>
<td>Associate Prof</td>
<td>3</td>
<td>no</td>
</tr>
</tbody>
</table>

IF rank = ‘professor’ OR years > 6 THEN tenured = ‘yes’

Source: www.cs.bu.edu/fac/gkollios/ada01/.../Classification1.ppt
Classification Process (2): Use the Model in Prediction

Source: www.cs.bu.edu/fac/gkollios/ad
a01/.../Classification1.pp
Data Clustering

• It aims to build clusters such that each cluster is similar within itself but is dissimilar to others.

• Clustering does not rely on class-labeled data objects.

• Find groupings within the dataset
Problems suitable for analytics

- Require knowledge-based decisions
- Have a changing environment
- Have sub-optimal current methods
- Have accessible, sufficient, and relevant data
- Provides high payoff for the right decisions!
Examples of Data Science Applications

- Recommendations
  - LinkedIn connection
  - Amazon.com product
  - Netflix.com product
Sample applications

- Spam email detection
- Credit scoring
- Fraud detection
- Medical diagnosis
- Market basket analysis
Machine learning applications

- Speech and hand-writing recognition
  - Speech to text

- Playing games
  - Learning to play chess

- Autonomous control
  - Learning to drive an autonomous vehicle
  - Robot control

- Mining bioinformatics data
  - Sequence motifs, alignment

- Identifying new astronomical structures
  - Learning regularities in a very large database of image data
Machine learning applications

- The Google search engine uses numerous machine learning techniques
  - Autosuggest and type ahead
  - Grouping together top news stories from various sources
  - Analyzing data from over 3 billion web pages to improve search results
  - Analyzing which search results are most often followed, i.e. which results are most relevant
Data mining applications

- Web data mining
- Search engines
- Data warehouse
- Sequential patterns
- Time-series analysis
Retail analytics

- Market basket analytics
- Text analytics
- Customer segmentation
- Tailored product assortments
- Inventory forecasting
Sensor Data Analytics

- Reduce equipment downtime
- Extend lifetime of equipment
- Lower operational cost
  - Through smart scheduling and servicing
  - Prioritizing maintenance activities
  - Selecting equipment vendors
- Increase service availability and quality
- Customer satisfaction
Delivering value

- Sensors are sending real-time oceanographic data that scientists use to create detailed visualizations of sonar, tidal, current and other activity such as temperature and quality of the water.

- As high volumes of data are collected over time, the scope to use them to forecast future ocean/weather events is invaluable.

Source: http://www.telegraph.co.uk/sponsored/sport/rugby-trytracker/10630542/water-ibm-analytics.html
Recommender Systems

Introduction
Recommender System: Amazon.com
Collaborative Filtering

- Goal: predict what movies/books/… a person may be interested in, on the basis of
  - Past preferences of the person
  - Other people with similar past preferences
  - The preferences of such people for a new movie/book/…

- Amazon recommendations are an instance of collaborative filtering, where users collaborate in the task of filtering information to find information of interest

Source:
Twitter.com
www.cs.kent.edu/~jin/Cloud12Spring/BigData.pptx

Copyrights @ Anasse Bari
End of Chapter One:
Introduction to Predictive Analytics
Questions