DATA SCIENCE AND R

JARED HESSE
What is the difference between Data Science, Data Analysis, Big Data, Data Analytics, Data Mining and Machine Learning?

**Data Science**
- Deals with structured and unstructured data

**Data Analysis**
- Human activities aimed at gaining some insight on a dataset
- Everything that relates to data cleaning, preparation and analysis

**Data Analytics**
- Automating insights into a dataset and suppressing the usage of queries and data aggregation procedures
- Can represent various dependencies between input variables, but also can use Data Mining techniques and tools to discover hidden patterns in the dataset under analysis
- Analyst can use some Data Analytic tools to obtain desired results, but in principle, Data Analytics can be performed without social data processing

**Data Mining**
- Uses the predictive force of machine learning by applying various machine learning algorithms to Big Data

**Big Data**
- Huge data volumes that cannot be processed effectively with traditional applications
- Begins with raw data that is not aggregated and it is often impossible to store such data in the memory of a single computer

**Machine Learning**
- Artificial Intelligence techniques that are broadly used in Data Mining
- Uses a training dataset to build a model that can predict values of target variables

Source: onthe.io
What is Dirty Data?

• Incomplete or missing entries
• Misspelled entries
• Anomalous entries
  • E.g., an alphanumeric character in an otherwise numeric “ID” column
• Outlier entries
• Inconsistent (redundant) entries
• Duplicate entries
• Just plain confusing entries:

<table>
<thead>
<tr>
<th>price_per_unit</th>
<th>purchased_quantity</th>
<th>spend</th>
<th>verified_purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>250</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>800</td>
<td>2</td>
<td>400</td>
</tr>
<tr>
<td>4</td>
<td>700</td>
<td>5</td>
<td>3500</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Solutions:

• Imputation
• Deletion
• Inference (Expert and Learned)
Why R?

- **Power and Versatility**
  - R can tackle nearly any analysis or task imaginable, especially in the context of data science: data cleansing, report creation, machine learning algorithms, time series prediction, statistical analysis, multivariate analysis, survival analysis, etc.

- **Scalability**
  - Integration with distributed computing software such as Apache Spark is raising the limit of datasets that R can tackle in-memory.
  - The size of the data that R can handle is only limited by the amount of RAM that a machine has – my work frequently has me working with datasets of ~1,000,000 observations.

- **Package Library**
  - Over 20,000+ packages freely available that cover any topic, analysis, or obstacle – when in doubt you can create your own packages as well.

- **Visualization**
  - The simplicity and modularity of “ggplot2” has made it the de facto standard for visualizing data.

- **Support**
  - The R community is humongous and continues to grow – there are dozens of R-centric websites and freely available resources for tutorials, questions, and development.

- **Sugary Syntax and R as a “Functional” Language**
  ```r
  # total cholesterol values
  lipid_results %>% filter(Type=="Total") %>% select(patient_id, value) %>%
  mutate(value_Fixed = parse_number(value)) %>%
  group_by(patient_id) %>%
  mutate(Total_Cholesterol = max(value_Fixed)) %>%
  ungroup() %>%
  filter(is.na(Total_Cholesterol)) %>%
  select(patient_id, Total_Cholesterol) %>%
  mutate(patient_id = as.character(patient_id)) %>%
  distinct() -> total_lipid_out
  
  # Split into usable format
  split_fun <- function(df)[
    split <- str_count(df[,1])
    names <- unlist(str_split(colnames(df),";"))
    x <- names(df)
    toparuna_(df, x, sep=";", into = names)
  ]
  split_fun(anchors) -> anchors
  split_fun(exclusions) -> exclusions
  split_fun(icd_exclusions) -> icd_exclusions
  
  # LOGISTIC REGRESSION ON VAN WALVAREN SCORES ALOE
  library(caret)
  van_Walvaren %>%
  select(vanwalvaren_score, patient_dead) %>%
  mutate(patient_dead = as.factor(patient_dead)) -> vars
  Train <- createDataPartition(vars$patient_dead, p=0.6, list=FALSE)
  testing <- vars[-Train,]
  training <- vars[Train,]
  mod.fit <- train(patient_dead ~ vanwalvaren_score, data-training, method="glm", family="binomial")
  pred <- predict(mod.fit, newdata-testing)
  accuracy <- table(pred, testing$patient_dead)
  sum(diag(accuracy))/sum(accuracy) # accuracy
  confusionMatrix(data=pred, testing$patient_dead) # final confusion matrix
  ```
What is Healthcare Data Like? (Spoiler: FILTHY)

Structured

- Healthcare claims data are great examples of structured data.
- Usually comes from a database (SQL, MongoDB, Hadoop, etc.).
- Ordered – contains primary keys (ID columns) that link data across database.
- Organized – this is a true rarity.

Unstructured/Semi-Structured

- Messy, frequently dirtier than structured data.
- Comes from a variety of different sources.
- Not ordered, or very little organization.

"The patient came in complaining of chest pain, shortness of breath, and lingering headaches...smokes 2 packs a day...family history of heart disease...has been experiencing similar symptoms for the past 12 hours..."

<table>
<thead>
<tr>
<th>ID</th>
<th>Claim Date</th>
<th>Cost</th>
<th>Diagnosis</th>
<th>Procedure</th>
<th>Place of Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>2/5/2013</td>
<td>500</td>
<td>5716</td>
<td>NA</td>
<td>Home</td>
</tr>
<tr>
<td>456</td>
<td>5/8/2012</td>
<td>2000</td>
<td>K743</td>
<td>NA</td>
<td>Office</td>
</tr>
<tr>
<td>785</td>
<td>9/12/2016</td>
<td>800</td>
<td>110</td>
<td>NA</td>
<td>Hospital</td>
</tr>
<tr>
<td>101</td>
<td>10/30/2015</td>
<td>6000</td>
<td>111</td>
<td>50008</td>
<td>Office</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Test</th>
<th>Result</th>
<th>Encounter Number</th>
<th>Units</th>
<th>Test Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>Jared's Test of Coolness</td>
<td>5</td>
<td>1</td>
<td>coolness/ml</td>
<td>0005</td>
</tr>
<tr>
<td>123</td>
<td>COOLNESS ML JARED</td>
<td>NEG</td>
<td>2</td>
<td>NA</td>
<td>0005</td>
</tr>
<tr>
<td>123</td>
<td>Jared's Test of Coolness</td>
<td>NEG</td>
<td>3</td>
<td>coolness/ml</td>
<td>NA</td>
</tr>
<tr>
<td>123</td>
<td>COOLNESS TEST</td>
<td>SEE DOCTOR</td>
<td>0</td>
<td>NA</td>
<td>0005</td>
</tr>
<tr>
<td>123</td>
<td>TEST OF COOLNESS - JRD</td>
<td>300</td>
<td>5</td>
<td>coolness/ml</td>
<td>NA</td>
</tr>
<tr>
<td>123</td>
<td>xxxxxxxx</td>
<td>4</td>
<td>5</td>
<td>coolness/ml</td>
<td>NA</td>
</tr>
</tbody>
</table>
Real World Efforts with Healthcare Data

• “Precision Medicine”
  • Of the last 10,000 Asian women near age 50 who were treated for the same tumor, what medications were used, was surgery or radiation necessary, and what were the outcomes?

• Predictive Analytics of Outcomes
  • What kind of patient $x$ is most/least likely to $y$?
    • $y$ = be readmitted after surgery, die within the next year, cost more money to the hospital, improve within three months...

• Forecasting
  • How much medical supplies of $x$ will I need based on the last $y$ years of data that I have?
  • Is $x$ disease on the rise for my hospitals? What will its incidence look like in the future?

• Understanding Rare Diseases
  • How many patients have $x$ disease in $y$ geographic location that fit $z$ criteria?
  • Through a patients’ history, how can we understand the progression of $x$ disease?

• Diagnosis and Treatment
  • Diagnosis is recognizing patterns – deep machine learning is just that:
    • “Lumiata has developed predictive analytics tools that can discover accurate insights and make predictions related to symptoms, diagnoses, procedures, and medications for individual patients or patient groups.”
  • Treatments based off of “learning” from millions of past patients:
    • What hip replacement works best for female patients 80 years and older?

• Real-World Evidence
  • Can we complement the “gold-standard” of evidence (clinical trials) with large-scale real world evidence found in data?
A Day in the Life

- **Raw Dataset Size:**
  - Anywhere from 1 – 10 TB in size
  - Comes in .csv form. Imagine a book that is 1 TB in size – it would be 85,899,345 pages long!

- **Analytical Dataset Size:**
  - Usually “long” (many rows) as it is “wide” (numerous columns). ~300 MB – 9 GB.
  - Task-dependent – can range from 100 rows, to 37,000 rows x 50 cols to 94,000,000 rows x 7 cols, etc.

- **Process:**
  - **Cleaning** – source dependent, anywhere from 0 – 4 hours per analysis. Always iterative.
  - **Visualizing** – understand the story the data is trying to tell 1 – 1.5 hours usually
  - **Analyzing** – anywhere from hours to days:
    - Frequencies/Statistics/Counts/Filters
    - Machine learning (usually not applicable, data analysis is not always “sexy”)
    - Novel work
    - Geolocation
  - **Sanity Checking** – this is where coworkers come in handy!
  - **Reporting** – 0.5-2 hours to write-up a finished analysis and deliver it
A Day In the Life – Data Science is Learning
## Average Salary for Skill: R

<table>
<thead>
<tr>
<th>Skill Type</th>
<th>Median Salary</th>
<th>25K</th>
<th>50K</th>
<th>100K</th>
<th>150K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientist</td>
<td>$88,375</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>612 salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Analyst</td>
<td>$62,013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>595 salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistician</td>
<td>$72,477</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>146 salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Scientist, IT</td>
<td>$88,212</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>138 salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senior Data Scientist</td>
<td>$122,808</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>97 salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biostatistician</td>
<td>$70,942</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>94 salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical Analyst</td>
<td>$62,163</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79 salaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Country:** United States  |  **Currency:** USD  |  **Updated:** 8 Feb 2017  |  **Individuals Reporting:** 2,964

![Data Science Skills Diagram](https://via.placeholder.com/200)
Selected R Packages and Readings

• **Recommended Packages:**
  • The “tidyverse” – data manipulation from A-Z:
    • dplyr, purr, forcats, broom, lubridate, tidyr, readr, stringr, etc.
  • data.table – extremely fast data manipulation and ingestion for large datasets
  • caret – implementation of 1000+ algorithms in R

• **Recommended Readings:**
  • R for Data Science (free interactive e-book)
    • [http://r4ds.had.co.nz/](http://r4ds.had.co.nz/)
    • A MUST-READ for anyone at any skill level using R – not just for data science.
    • “The finest introduction to R ever written” – Jared Hesse.
  • Advanced R (free interactive e-book)
    • [http://adv-r.had.co.nz/](http://adv-r.had.co.nz/)
  • R Cookbook
  • Introduction to Statistical Learning/The Elements of Statistical Learning

• **Websites**
  • r-bloggers.com
  • rweekly.org
  • stackoverflow.com/tags/r