Data Science for OpenFDA

Story
Slides
Slide 1 Cluster Analysis of Pharmaceutical Adverse Events
Slide 2 Manufacturers by All Drug Classes
Slide 3 Manufacturers by All Adverse Events
Slide 1 Using OpenFDA for Big Data Analysis
Slide 2 Market Demand / Necessity
Slide 3 About Big Data Lens
Slide 4 NLP = Natural Language Processing
Slide 5 Models and Predictions
Slide 6 Data Fusion to Know a Patient
Slide 7 OpenFDA Queries 1
Slide 8 OpenFDA Queries 2
Slide 9 Important OpenFDA data types 1
Slide 10 Important OpenFDA data types 2
Slide 11 More OpenFDA data types
Slide 12 Who Died?
Slide 13 EndPoint Reference
Slide 14 First OpenFDA App
Slide 15 Data Fusion for Prescription Drugs
Slide 16 Place Use for NLP
Slide 17 Big Data Use Cases
Slide 18 Big Data Design Principals
Slide 19 Great Graphics Start Here
Slide 20 Contact Information

Spotfire Dashboard
Spotfire Hierarchical Clustering and Treemaps
  EPC by Maker: Clustering
  EPC by Maker: Treemap
  Death by Maker: Clustering
Death by Maker: Treemap
Reactions by Maker: Clustering
Reactions by Maker: Treemap
What is the Hierarchical Clustering Tool?
Overview of Hierarchical Clustering Theory
Distance Measures Overview
Clustering Methods Overview
Hierarchical Clustering References

Emails
HHS VizRisk Challenge
July 31, 2014
July 21, 2014
July 21, 2014
July 21, 2014
July 21, 2014
July 21, 2014
July 21, 2014
July 10, 2014

FDA Voice Blog: FDA Leverages Big Data Via Cloud Computing
OpenFDA: Innovative Initiative Opens Door to Wealth of FDA’s Publicly Available Data

Big Data Lens
ABOUT
DEMO
PRODUCTS
TRIPWIRE (COMING)
SNAP - SOCIAL NEWS ANALYSIS PLATFORM
TECHNOLOGY SEQUENCE ANALYSIS (TSA)
TREND IMPACT ANALYSIS (TIA)
SERVICES
A MODERN RECIPE FOR BIG DATA...
Content is everywhere. We collect it for you.
Then we turn content into data
Data becomes understanding when it becomes a model.
Models with high explanatory power are used to predict.
An API (application programming interface) gives you access to your daily prediction.
EXPERIENCE
BLOG
OLD HAT - TRADITIONAL RELATIONAL DATABASES AND HADOOP/MapReduce
FIRST MOVER ADVANTAGE IN YOUR PREDICTIVE MODEL
ON TRUSTING BIG DATA(2) AND BEING “PREDICATE FREE”
TWO VIRTUOUS MODELS - THE POWER OF DEDUCTION AND INDUCTION
OF FOOTBALL & POLITICS (IN BRAZIL)
WHAT BIG DATA PATENT ANALYSIS LOOKS LIKE
TECHNOLOGY READINESS LEVEL: EXCELLENT BIG DATA METHODOLOGY
WHY PREDICTIVE ANALYTICS GO WRONG
HAND WRITTEN NOTES IN ELECTRONIC HEALTH RECORDS
WHAT DOES HUMAN COMPREHENSION MEAN?
MACHINE LEARNING TOOLS COMPARED
SEARCH AND DESTROY
2014 WITH THE GO LANGUAGE
NOT LONG FROM NOW ...
BIG DATA FROM TWITTER
CORRELATION & CAUSALITY
COOL OLD TECHNIQUES (PART 2)
COOL OLD TECHNIQUES
BIG DATA IDEAS THAT OPPOSE
MICROSOFT GETTING INTO BIG DATA ... OR THEY ALWAYS WERE
A NEED FOR TRUST IN BIG DATA
WHY NOT MAKE YOUR OWN BIG DATA?
CAUTION - MESSY LANDSCAPE AHEAD
POLITICS & POLL OF POLLS = BIG DATA
WHAT IS MORE IMPORTANT DATA OR ANALYTICS?
SOLD OUT - YIKES NYC STRATA / HADOOP CONF.
DATA EMANATION
BIG DATA SPENDING TO DOUBLE BY 2016
BIG DATA GEEKS - HIDING IN PLAIN SITE
BI OR BIG DATA? - START BY THINKING BACKWARDS
STARTING SMALL WITH BIG DATA
WHAT BIG DATA LOOKS LIKE
IF BABOONS CAN DO IT ....
BIG DATA - OLD BECOMES NEW AGAIN (WITH SOME TWISTS)

CONTACT
SURVEY
1. Are you actively involved in predictive analytics in your company?
2. As the starting step in your predictive analytics work do you ...?
3. Thinking about data sources for your predictive analytics work which of these do you include?

http://semanticommunity.info/Data_Science/Data_Science_for_OpenFDA
Updated: Sat, 19 Sep 2015 06:44:30 GMT
Powered by mindtouch™
4. What is your preferred way of storing the data?
5. Do you conduct a pre-processing stage on the data after acquiring it?
6. What one thing do you do more often before modeling and predicting using the acquired data?
7. Which Big Data Predictive Analytics Suite would you use to model your data?
8. Which of the following measures of fit do you use and value most highly for a predictive model?
9. Which Predictive Analytics techniques do you use and value most highly?
10. If you like, tell us about yourself

---

### Story

**Welcome to openFDA!**

There are a few things we need to make sure you’re aware of before you play around with the API.

This API is not for clinical or production use. While we make every effort to ensure that data is accurate, you should assume all results are unvalidated. The data within the openFDA API is Public Domain and released under Creative Commons CC0.

This API is in beta. Do not assume it will stay online. As we continue work, we may need to change things around, rework queries, or move to a new URL.

You will see this message once a week throughout the beta period.

**The openFDA team**

**END OF SPLASH SCREEN**

Brook Aker developed the slides below with the sample URLs and queries which I checked.

I wanted some spreadsheet data to work with for visualizations. This is why I do not like APIs.

MORE TO FOLLOW

### Slides

Slides
Slide 1 Cluster Analysis of Pharmaceutical Adverse Events

Cluster Analysis of Pharmaceutical Adverse Events

From OpenFDA data
7/21/14

Slide 2 Manufacturers by All Drug Classes

Manufacturers by All Drug Classes

Group distinguished by abnormal number of adverse events for the products they make — includes companies Abbvie and Pfizer

Group above average for the number of product adverse events — includes private third-party companies CVS, Kroger, Wal-Mart, Publix

Other manufacturers not troubling in the number of adverse events
Slide 2 Market Demand / Necessity

Market Demand / Necessity

+ "Given that those who will experience adverse events sometimes number in the thousands, automated algorithms to reliably identify such individuals and generate alerts to patients, payers & prescribers may be the only feasible means to conduct proactive safety enforcement on the necessary scale."

+ "Given the wide application of expert systems in other public health and safety contexts, it seems likely that safety analysis would gain in effectiveness by adopting automated procedures in analyzing and reporting their data."

Slide 3 About Big Data Lens

About Big Data Lens

+ Custom Algorithms for Big Data problems
+ Specializing in Natural Language Processing, Predictive Analytics & Machine Learning
+ Creates API/App based big data algorithms to be used in decision making

http://semanticommunity.info/Data_Science/Data_Science_for_OpenFDA
Updated: Sat, 19 Sep 2015 06:44:30 GMT
Powered by mindtouch™
Slide 4 NLP = Natural Language Processing

NLP = Natural Language Processing

- **The roles they play:** nouns, verbs, adjectives, adverbs, etc.
- **The logic they infer:** words in certain order denote intended direction – "John Smith prescribed Wellbutrin to Robert Jones" vs. "Robert Jones was prescribed Wellbutrin by John Smith". NLP understands John Smith is the prescriber and Robert Jones is the patient in both cases.
- **The larger concept:** whole sentences or paragraphs are categorized as an aid to clarity, e.g. "prescription" vs. "side effect"
- **The definition behind words:** "script" can mean prescription or it can mean a snippet of software code. Using the words around it NLP determines the correct definition.
- **The patterns they reveal:** proper names, places, addresses, phone numbers, etc. reveal themselves by the patterns in which they are written.

Slide 5 Models and Predictions

Models and Predictions

- **Big Data Modeling Techniques Used:**
  - Support Vector Machines
  - High Dimensional Discriminant Analysis
  - Linear Discriminant Analysis
  - Logit / Probit / Discrete Choices Models
  - Neural Networks
  - Bagging / Boosting Models
  - Random Forest

- Sequential Ensemble Methods produce highest fit
Slide 6 Data Fusion to Know a Patient

Slide 7 OpenFDA Queries 1

https://api.fda.gov/drug/event.json?...meddrapt.exact

Slide 8 OpenFDA Queries 2

https://api.fda.gov/drug/event.json?...meddrapt.exact
Slide 9 Important OpenFDA data types 1

Important OpenFDA data types

+ What the drug is supposed to fix:
  + Pharmacologic Class (EPC) - pharm_class_epc

+ How the drug works:
  + Mechanism of Action (MOA) - pharm_class_moa

+ What the drug affects:
  + Physiologic Effect (PE) - pharm_class_pe

+ What is in the drug:
  + Chemical Structure (CS) - pharm_class_cs

Slide 10 Important OpenFDA data types 2

https://api.fda.gov/drug/event.json?...e+Inhibitor%22
Slide 11 More OpenFDA data types

More OpenFDA data types

- **How serious is the reaction**: serious (1 for Yes, 2 for No)
  - "serious": "1",
  - "seriousnesscongenitalanomaly": "1",
  - "seriousnessdeath": "1",
  - "seriousnessdisabling": "1",
  - "seriousnesshospitalization": "1",
  - "seriousnesslife-threatening": "1",
  - "seriousnessother": "1"
- **What is the drug indicated for**: drugindication
- **Circumstances for taking drug**: patient.drug.drugadditional

Slide 12 Who Died?

https://api.fda.gov/drug/event.json?...eactionoutcome

http://semanticommunity.info/Data_Science/Data_Science_for_OpenFDA
Updated: Sat, 19 Sep 2015 06:44:30 GMT
Powered by mindtouch"
Who Died?

- What is the result of the reaction: patient.reaction.reactionoutcome
- **628 people** have died taking anti-depressants / anti-anxiety drugs in last 10 years.

---

Slide 13 EndPoint Reference

[https://open.fda.gov/drug/event/reference/](https://open.fda.gov/drug/event/reference/)

---

Slide 14 First OpenFDA App

[http://searchopenfda.socialhealthinsights.com](http://searchopenfda.socialhealthinsights.com)
First OpenFDA App

http://searchopenfda.socialhealthinsights.com

Updated: Sat, 19 Sep 2015 06:44:30 GMT

13
Slide 16 Place Use for NLP

Place Use for NLP

- Doctors Notes in EHR (provides watch words, prognosis)
- Structured Product Language - "the label", "package insert" (only place for detailed indication, contra-indications, warnings, etc.)
- Both are always written long form text

Slide 17 Big Data Use Cases

Big Data Use Cases

- Adverse Effects Risk Model
- Active Ingredients - Effectiveness Model
- Advanced Drug-Drug Interaction Model
- Importation Risk Model
- Adherence Model
- Gaps in Care Model

http://semanticommunity.info/Data_Science/Data_Science_for_OpenFDA
Updated: Sat, 19 Sep 2015 06:44:30 GMT
Powered by mindtouch
Slide 18 Big Data Design Principals

**Big Data Design Principals**

1. Catalog as many API sources of data as you can
2. Pick sources with common “folding points”
3. Research what is novel – e.g. pick your outcome
4. Sample small to test model(s)
5. Plan for Scale
6. Dress for Success (e.g. great graphics + good GUIs)

Slide 19 Great Graphics Start Here


Slide 20 Contact Information

Brooke Aker  
[baker@bigdatalens.com](mailto:baker@bigdatalens.com)  
860-614-2411  
[http://www.bigdatalens.com](http://www.bigdatalens.com)
Spotfire Dashboard

For Internet Explorer Users and Those Wanting Full Screen Display Use: Web Player Get Spotfire for iPad App

Media, iframe, embed and object tags are not supported inside of a PDF.

Spotfire Hierarchical Clustering and Treemaps

EPC by Maker: Clustering
Reactions by Maker: Clustering

Reactions by Maker: Treemap

What is the Hierarchical Clustering Tool?

The Hierarchical Clustering tool groups rows and/or columns in a data table and arranges them in a heat map visualization with a dendrogram (a tree graph) based on the distance or similarity between them. When using the hierarchical clustering tool, the input is a data table, and the result is a heat map with dendrograms. You can also initiate hierarchical clustering on an existing heat map from the Dendrograms page of the Heat Map Properties. See How to Use the Heat Map to learn more.

To perform a clustering with the Hierarchical Clustering tool:

Select Tools > Hierarchical Clustering....

Response: The Hierarchical Clustering dialog is displayed.

If the analysis contains more than one data table, select a Data table to perform the clustering calculation on.

Click Select Columns....

Response: The Select Columns dialog is displayed.
Select the columns you want to include in the clustering, and then click OK to close the dialog.

Select the Cluster rows check box if you want to create a row dendrogram.

Click the Settings... button to open the Edit Clustering Settings dialog.

Select a Clustering method.

Comment: For more information on clustering methods, see Clustering Methods Overview.

Select a Distance measure.

Comment: For more information on distance measures, see Distance Measures Overview. Distances exceeding $3.40282e+038$ cannot be represented.

Select Ordering weight to use in the clustering calculation.

Comment: For more information see Ordering Weight.

Select an Empty value replacement Method from the drop-down list.

Comment: The available replacement methods are described in Details on Edit Clustering Settings.

Select a Normalization Method to use in the clustering calculation.

Comment: For more information, see Normalizing Columns.

Click OK.

Select the Cluster columns check box if you want to create a column dendrogram.

Go through steps 6 to 12 to define settings for the column dendrogram.

Click OK.

Response: The hierarchical clustering calculation is performed, and a heat map visualization with the specified dendrograms is created. A cluster column is also added to the data table and made available in the filters panel.

Comment: See Dendrograms and Clustering to learn more about dendrograms and cluster columns.

Overview of Hierarchical Clustering Theory

Hierarchical clustering arranges items in a hierarchy with a treelike structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree-structured graph called a dendrogram. In Spotfire, hierarchical clustering and dendrograms are strongly connected to heat map visualizations. You can cluster both rows and columns in the heat map. Row dendrograms show the distance or similarity between rows, and which nodes each
row belongs to as a result of clustering. Column dendrograms show the distance or similarity between the variables (the selected cell value columns). The example below shows a heat map with a row dendrogram.
You can perform hierarchical clustering in two different ways: by using the Hierarchical Clustering tool, or by performing hierarchical clustering on an existing heat map visualization. If you use the Hierarchical clustering tool, a heat map with a dendrogram will be created. To learn more about heat maps and dendrograms, see What is a Heat Map? and Dendrograms and Clustering.

Algorithm

The algorithm used for hierarchical clustering in Spotfire is a hierarchical agglomerative method. For row clustering, the cluster analysis begins with each row placed in a separate cluster. Then the distance between all possible combinations of two rows is calculated using a selected distance measure. The two most similar clusters are then grouped together and form a new cluster. In subsequent steps, the distance between the new cluster and all remaining clusters is recalculated using a selected clustering method. The number of clusters is thereby reduced by one in each iteration step. Eventually, all rows are grouped into one large cluster. The order of the rows in a dendrogram are defined by the selected ordering weight. The cluster analysis works the same way for column clustering.

Note: Only numeric columns will be included when clustering.

Distance Measures Overview

The following measures can be used to calculate the distance or similarity between rows or columns:

Correlation

Cosine Correlation

Tanimoto Coefficient
Euclidean Distance

City Block Distance

Square Euclidean Distance

Half Square Euclidean Distance

The term dimension is used in all distance measures. The concept of dimension is simple if we are describing the physical position of a point in three dimensional space when the positions on the x, y and z axes refer to the different dimensions of the point. However, the data in a dimension can be of any type. If, for example, you describe a group of people by their height, their age and their nationality, then this is also a three dimensional system. For a row (or column), the number of dimensions is equal to the number of variables in the row (or column).

Note: The result from a cluster calculation will be presented either as the similarity between the clustered rows or columns, or as the distance between them. Euclidean distance, City block distance, Square Euclidean distance, and Half square Euclidean distance will present the distance between the rows or columns. The results from Correlation, Cosine correlation, and Tanimoto coefficient, on the other hand, are presented as similarity between the rows or columns.

Note: When used in clustering, the similarity measures Correlation, Cosine correlation, and Tanimoto coefficient may be transformed so that they are always greater than or equal to zero (using 1 – similarity value).

---

**Clustering Methods Overview**

Hierarchical clustering starts by calculating the distance between all possible combinations of two rows or columns using a selected distance measure. These calculated distances are then used to derive the distance between all clusters that are formed from the rows or columns during the clustering. You can select one of the following clustering methods:

* UPGMA
* WPGMA
* Single Linkage
* Complete Linkage
* Ward’s Method

---

**Hierarchical Clustering References**

Hierarchical clustering


General information about clustering


See also:

Overview of Hierarchical Clustering Theory

---

**Emails**

**HHS VizRisk Challenge**

*July 21st, 2014*

oncchallenges@capconcorp.com

The U.S. Dept of Health and Human Services (HHS) invites you to participate in:

The **HHS VizRisk Challenge** is the first-ever government hackathon that brings together students, designers, coders, healthcare enthusiasts, and government officials to produce visualizations of behavioral health data to inform personal and policy decisions.

**Open for submissions July 28th - October 28th**

All skillsets welcome!

Whether you're an experienced programmer or are familiar with just Adobe software and Microsoft Office, we have suggested projects for you! Final electronic project submissions can include pdfs, web-based tools, or anything that shows the data in an intuitive and insightful way. We will provide mentors with technical, health-related, and statistical backgrounds for those without experience.
Why hack for health?

• Compete for $15,000 in prizes  
• Reveal insights in data trends from NIH, CDC, Medicare/Medicaid, and more!  
• Use technology to shape health decisions on a personal and population level

Sign up today


Facebook | directors@hhsvizrisk.org | #HHSVizRisk

STAY CONNECTED

July 31, 2014

This one took a awhile to figure out. Shows the level of severity for adverse events from low to high (anomaly, disability, hospitalization, life threat, death) along with the general category of serious which acts as a total of the subcategories but is almost always slight higher then the addition of the subcategories. That is because some adverse effects are considered serious but don’t fall into one of the 5 subcategories.

I did my best to de-duplicate the data. Also where a company’s name was very similar but the reported numbers were not I took the higher total under the assumption the higher reporting represented a more common usage of the company name. Adding them together seemed like I might be double counting. Spreadsheet

July 21, 2014

This one is number of adverse events by manufacturers & the type of adverse events. Adverse Events come in levels. Obviously DEATH being the worst. Found a duplicate company. this version is corrected Spreadsheet
July 21, 2014

Here is another diagram along the same lines - just to show good separation among the groups. Not sure what happened to group 4 - must be hiding behind another group. Anyway we are going in the right direction

July 21, 2014

So here is one thing you can do with the data I gave you this morning. As attached. Cluster analysis to see which companies group together in the total number of adverse events for all products they make. My interpretation is that out of 700+ companies you have;

- about 10 companies who are “extreme” in adverse events drawing into question their manufacturing standards, formulations, ingredients, or safety / contamination procedures.
- about 10 companies who are “troubling” in the same regard.
- about 20 companies who are “above average” and may require increased monitoring. This group includes non-traditional manufacturers like retail drug locations (e.g. drug store chains)

This passes the smell test to me ~ 6% of companies may present troubling aspects of their operations that introduce undue adverse events into pharmaceutical products. While troubling, a higher percentage would make me wonder if the data, reporting or analysis were off. That or our regulatory and oversight regime was pretty lousy.

More to come

July 21, 2014

Yes - numbers are the total number of reported adverse events arrayed by drug class and manufacturer. Details within these may include the package label (always a written description of what the drug is for, not for, etc.)
July 21, 2014

Here is the first spreadsheet - Manufacturers by Established Pharmacological Class (EPC) where numbers are the reported adverse events. I worked through all the Manufacturers starting with A and added columns where the company name was very similar. So either I finish that process B-Z or that might be a wrong assumption. In any case gives you something to look at for now.

Will apply some machine learning to this data set to see what it shows us as a test. We can clean the rest of the data if the results look promising and then repeat.

July 21, 2014

Happy to get this for you. Just can't get all of it. It is too much and they put a governor on how the number of calls to their API per day. Would take 2 weeks to download the whole thing with the governor in place.

One thing I have almost worked out and can give you probably later today are 3 spreadsheets of the following:

1. The total number of adverse events arrayed by class of drug and manufacturer. The idea here is to test whether any single manufacturer stands out as having more adverse events than other manufacturers for all classes of drugs or a single class of drugs. In looking at a simple tally for example you can see Mylan Pharmaceuticals has far and away more adverse events for their products compared to others. The question is why?

2. The type of adverse reaction by class of drug. Question of cost benefit - is the cure worse than the disease?

3. The type of adverse reaction by manufacturer. Some reactions are not important. Others result in death. Important to find out what is killing people!

Would this type of initial analysis get HHS attention?

July 10, 2014

OK you got it. How much (many records) are you looking for? All 3.6 Million!

Thank you and I do not know beforehand until I look at all the data rows and columns if possible

I can write the routine to spit out data into CSV (spreadsheet) format for you if you want. Do you know what you want to concentrate on?

FDA Voice Blog: FDA Leverages Big Data Via Cloud Computing

Source: Email June 19, 2014
FDA Leverages Big Data Via Cloud Computing

By: Taha A. Kass-Hout, M.D., M.S., FDA’s Chief Health Informatics Officer and Director of FDA’s Office of Informatics and Technology Innovation.

Last year, I worked with a group of colleagues throughout the Food and Drug Administration (FDA) on a project that is critical for the agency’s future: the modernization of our information technology platforms to prepare for the influx of “Big Data”—the enormous data sets we receive daily from manufacturers, health care providers, regulatory bodies, scientists and others.

These data sets are not only larger than ever before, they are also arriving more frequently than ever and varying enormously in format, and quality.

This year alone, we expect to receive somewhere between 1.5 and 2 million submissions through our eSubmission Gateway – and some submissions can now be as large as a Terabyte (one trillion bytes) in size. This is the very definition of a big data.
But, at FDA, we view it as an opportunity and a challenge. To meet both, we are building an innovative technology environment that can handle vast amounts of data and provide powerful tools to identify and extract the information we need to collect, store and analyze.

A key example is our recent leveraging of cloud computing.

"Cloud computing" is, basically, computing on demand. Think of how you use water, or electricity, at the same time as do your neighbors and millions of others. You pay only for what you use, and service is always guaranteed. You don't need to wait till your neighbor is done to use the washer or dryer because there is only enough electrical capacity to handle one person at a time.

The same is true of cloud computing, which stores data on the Internet, rather than on the hard drive or drives of computers. In essence, it gives us the ongoing, simultaneous capacity to collect, control and analyze enormous data sets.

For example, FDA, partnering with state and local health organizations, identifies thousands of foodborne pathogen contaminants every year. We sequence, store and analyze this data to understand, locate, and contain life-threatening outbreaks. Again, cloud computing aids us in this effort.
Finally, FDA has some of the world’s most valuable data stores about human health and medicine. Through OpenFDA, our newest IT program, we are making some of these existing publicly available data sets more easily accessible to the public and to our regulatory stakeholders in a structured, computer readable format that will make it possible for technology specialists, such as mobile application creators, web developers, data visualization artists and researchers to quickly search, query, or pull massive amounts of public information instantaneously and directly from FDA datasets on an as needed basis. OpenFDA is beginning with an initial pilot program involving the millions of reports of drug adverse events and medication errors that have been submitted to the FDA from 2004 to 2013 and will later be expanded to include the agency’s databases on product recalls and product labeling.

OpenFDA promotes data sharing, data access, and transparency in our regulatory and safety processes, and spurs innovative ideas for mining the data and promoting the public health.

Big data is important to the way we carry out regulatory science, which is the science of developing new tools and approaches to assess the safety, efficacy, quality, and performance of FDA-regulated products. Through innovative methods such as cloud computing, we are taking advantage of this flood tide of new information to continue to protect and promote the public health.

Taha A. Kass-Hout, M.D., M.S., is FDA’s Chief Health Informatics Officer and Director of FDA’s Office of Informatics and Technology Innovation.

OpenFDA: Innovative Initiative Opens Door to Wealth of FDA’s Publicly Available Data


Posted on June 2, 2014 by FDA Voice

By: Taha A. Kass-Hout, M.D., M.S.

Today, I am pleased to announce the launch of openFDA, a new initiative from our Office of Informatics and Technology Innovation (OITI). OpenFDA is specifically designed to make it easier for web developers, researchers, and the public to access and use the many large, important, health data sets collected by the agency.
These publicly available data sets, once successfully integrated and analyzed, can provide knowledge and insights that cannot be gained from any other single source.

Consider the 3 million plus reports of drug adverse reactions or medication errors submitted to FAERS, the FDA Adverse Event Reporting System (previously AERS), since 2004.

Researchers, scientists, software developers, and other technically-focused individuals in both the private and public sectors have always been invited to mine that publicly available data set – and others – to educate consumers, which in turn can further our regulatory or scientific missions, and ultimately, save lives.

But obtaining this information hasn’t always been easy.

In the past, these vast datasets could be difficult for industry to access and to use. Pharmaceutical companies, for example, send hundreds of Freedom of Information Act (FOIA) requests to FDA every year because that has been one of the ways they could get this data. Other methods called for downloading large amounts of files encoded in a variety of formats or not fully documented, or using a website to point-and-click and browse through a database – all slow and labor-intensive processes.
OpenFDA will make our publicly available data accessible in a structured, computer-readable format. It provides a “search-based” Application Programming Interface – the set of requirements that govern how one software application can talk to another – that makes it possible to find both structured and unstructured content online.

Software developers can now build their own applications (such as a mobile phone app or an interactive website) that can quickly search, query or pull massive amounts of public information instantaneously and directly from FDA datasets in real time on an “as-needed” basis. Additionally, with this approach, applications can be built on one common platform that is free and open to use. Publicly available data provided through openFDA are in the public domain with a CC0 Public Domain Dedication.

Drug adverse events is the first dataset – with reports submitted from 2004 through 2013 available now.

Using this data, a mobile developer could create a search app for a smart phone, for example, which a consumer could then use to determine whether anyone else has experienced the same adverse event they did after taking a certain drug.

As we focus on making existing public data more easily accessible, and providing appropriate documentation and examples to developers, it’s important to note that we will not release any data that could be used to identify individuals or reveal other private information.

OpenFDA uses cutting-edge technologies deployed on FDA’s new Public Cloud Computing infrastructure enabled by OITI, and will serve as a pilot for how FDA can interact internally and with external stakeholders, spur innovation, and develop or use novel applications securely and efficiently. As we move forward with the early stages of openFDA, we will be listening closely to the public, researchers, industry and all other users for their feedback on how to make openFDA even more useful in promoting and protecting the public health.

Taha A. Kass-Hout, M.D., M.S., is FDA’s Chief Health Informatics Officer and Director of FDA’s Office of Informatics and Technology Innovation.
WELCOME

CUSTOM BIG DATA ALGORITHMS AND PREDICTIVE ANALYTICS

Bringing together human-like comprehension of web pages, advanced analysis, and cloud based processing for first time, Big Data Lens creates high scale, high speed algorithms that connect, conclude and *predict what happens next*.

Let us show you what we can do.

ABOUT

BIG DATA MEANS A WHOLE NEW WEB

FINALLY, ALL OF THE WEB IS A DATABASE YOU CAN USE...

Bringing together human-like comprehension of web pages, advanced analysis, and cloud based processing for first time, Big Data Lens creates high scale, high speed algorithms that connect, conclude and *predict what happens next*.

We are a group of technologists, intelligence specialists, statisticians, & mathematicians who are passionate about the practice of Big Data. For us, methods and technologies have come together for the first time that solve some of the thorniest problems in information science and ones we have been wrestling with for years. We get excited everyday that we get to apply these new techniques in creative and innovative ways on your behalf.

And a stylized version of what we are all about.
Here is a demo application we call SNAP (Social News Analysis Platform). Just click on the image below to get there or click here. Try it with a search like you would in Google by adding some search terms. Then slide the bar below to ask for the number of days of news to examine. Then click search.

You will get a wealth of analytical outputs that tell you the 5Ws and H ... who, what, where, when, why and how. Note the inclusion of social media, the links from within social media and more. Take that Google!!

Some users have reported getting no results. Try a simpler or broader query. This demo application is running on a select group of sites and not the whole internet so very specific or complicated queries can sometimes turn up empty. A production system would not have such limitations.
PRODUCTS

TRIPWIRE (COMING)

SNAP - SOCIAL NEWS ANALYSIS PLATFORM

A very Big Data tool to search for, analyze and display large quantities of web based, written information. SNAP is used to find the important companies, persons, organizations, the actions they take, the sentiment reported about them or their products, and the relationships between them. What you want from SNAP is always a matter of discussion and customization. See our demo here.

TECHNOLOGY SEQUENCE ANALYSIS (TSA)

This Big Data technique looks to predict the path of technology development in the medium to long range. Often used by strategists, researchers, and policy makers, TSA looks at the probability and timing as they contribute to choices for a finished technology good. The TSA model is run as a simulation with a range of outcomes.

Because every good developed faces a number of alternative choices over time the TSA is used to answer questions like "what is the most likely technology choices that will be made, how long will it take to get there and what capabilities with the final good have?" Alternatively users can ask "How fast can we get to the finished good and what tradeoffs must we make to get there?" Finally, in complex TSA networks it is often asked "Which technology input is holding us up more than any other?" The simulation answers these questions and contributes to thoughtful planning and resource allocation.

This simplistic TSA is for an advanced battery charger. It relies on 2 sub-components - a system to provide fuel and a system to generate energy from the fuel. The fuel system in turn could be accomplished through compressed gas or a solid fuel. Likewise energy generation could be done through combustion or through chemical processes. Real TSA's are far more complex and thus a simulation shows optimal, shortest and most impactful pathways.
TREND IMPACT ANALYSIS (TIA)

This Big Data technique looks at disruptive events often missing from market forecasts but are typically the culprit for proving most experts wrong as the future unfolds. Like TSA, the TIA is run a simulation and a range of market forecast outcomes are the result.

From there a user can ask questions of the simulation like "In 2018 what are the chances market demand be above 300 million units?" The simulation might answer 82%. This a strong result in which to make resource allocation and competitive plans. It is also a far cry from research reports which give you a single point forecast and no confidence measure in the prognostication.

![Analysis Results](image)

The TIA forecast shown here, for example so a far greater upside relative to the historical or 'expert' forecast. Thorough planning in terms of product development, market sizing, supply chain logistics and competition can rely on this richer look at how markets are likely to development over time and gain an advantage.

SERVICES

A MODERN RECIPE FOR BIG DATA...

Below are the typical steps we follow to create an API access point for you to move from raw Web Information to a constant flow of customized, actionable intelligence. Its the modern recipe for Big Data.

Content is everywhere. We collect it for you.

We connect to it when its a data feed. We hook into it when its social media. We crawl it when its a website. Just tell us the type of information you want or rely on now. We'll take it from there, and get you a lot more.
Then we turn content into data

- something you don't see elsewhere. Our technology moves across what has been written and understands it like you do, turning a news stories, blog or a tweet into data that can be used for analysis. Statistical Natural Language Processing (NLP) is how we tap into the company actions, consumer sentiment, & political movement reported everyday.

Data becomes understanding when it becomes a model.

We work with you to focus on what you need to measure. Maybe it's the timing of a new drug launch, who will win an election, or the price of energy. Then we apply machine learning techniques - testing the things that influence the drug launch, election or energy prices until we arrive at a stable and useable model.

Models with high explanatory power are used to predict.

Like a stock index represents a summation of the days events we do the same and provide you with a daily prediction on that drug launch, election outcome or change in energy price. Tracing back through daily events is always available. But an index sure beats trying to read it all for yourself and juggle it around in your head.
An API (application programming interface) gives you access to your daily prediction.

Like a website address but with some parameters you can adjust you receive back a 1 week, 1 month or 6 month view on the drug launch, election outcome or energy prices. But more than that an API is a means for you to use the prediction index data in other programs or models proprietary to you.

EXPERIENCE

At Big Data Lens we’ve gained a lot of experience in how to find data, meld data, make data flow, analyze it, model it, predict it and get the machine to learn from for constant updates. Unlike the big vendors, we use as much open source material, tools and methods as apply to the problem at hand. That keeps the cost low but the impact for you is the same - actionable insight that drops straight to the bottom line.

The industries and applications below are just a representative sample. Call us to discuss these or others you have in mind.

BLOG

OLD HAT - TRADITIONAL RELATIONAL DATABASES AND HADOOP/MAPREDUCE

July 3, 2014

I attended the Federal Big Data Working Group Meetup last Monday night. They had 3 speakers from MIT and their Big Data / Data Base Group speaking. Prof(s). Sam Madden and Mike Stonebraker are excellent at laying out the case, the architecture and uses for Big Data and Predictive Analytics.

Prof. Stonebraker is now well known for his controversial views on the "Big Elephant" database vendors like Oracle and IBM. He contends they are legacy, row based systems running on millions of lines of 30 year old code, and therefore very very slow. He should know since he has helped build every successive wave of database systems since the beginning. He is a big proponent of column based systems and now more recently of array based systems - like SciDB (which we recommend).
He also believes the time has past for Hadoop and MapReduce as the central architectural components to Big Data practice.

I did not record his talk but found a slightly older version that looks like just about the same thing here.

Enjoy.

**FIRST MOVER ADVANTAGE IN YOUR PREDICTIVE MODEL**

July 2, 2014

The New York Times posted this useful article yesterday about the power and peril of predictive analytics. The author is not quite right when he states "If Netflix can predict which movies I like, surely they can use the same analytics to create better TV shows. But it doesn't work that way. Being good at prediction often does not mean being better at creation."

Any good predictive algorithm can find the features of movies I like as part of a recommendation engine. And yes it cannot create the next movie I will like but it surely can inform writers about making sure the features I like are in the next movie they create.

And the author is certainly not right in his contention *"For example, we may find the number of employees formatting their résumés is a good predictor of a company's bankruptcy. But stopping this behavior hardly seems like a fruitful strategy for fending off creditors."* Here he assumes understanding the outcome of a system is the same as changing it. Not anymore so than when your doctor tells you to take aspirin for the headache you have due to a virus. The aspiring relieves a symptom but not the disease.

But the author is dead on with this statement *"Rarity and novelty often contribute to interestingness — or at the least to drawing attention. But once an algorithm finds those things that draw attention and starts exploiting them, their value erodes. When few people do something, it catches the eye; when everyone does it, it is ho-hum."*

In short, if you are the first to discover signal from what looked like noise then use it to your advantage. But recognize that once everyone else notices how well you are doing by manipulating the system they will copy you. The entire system will change and your advantage will disappear.

But all is not lost. You simply have to build your model again and find a new first mover advantage. Don't rest on your laurels.

**ON TRUSTING BIG DATA(2) AND BEING “PREDICATE FREE”**

July 1, 2014

There seems to be a divide between Big Data sources when it comes to trust. Big Data efforts based on internal sources (e.g. log data, transactions, etc.) are trusted. But apparently Big Data efforts based on external sources in the form of social media etc. is not. See this IBM study for more details.
But this may be due to one of two things. The first is simple experience with the external data. Yes it comes from the wild and wholly world of the internet but like the internal data that does not mean there isn't gold in there.

Which leads to the second point - technique. There is a wild array of techniques you can use to analyze your Big Data. Finding the right size, shape and analytical technique in order to find that gold is not easy. It takes trial and error and great skill in interpreting the results.

Like all marketing efforts, Big Data is now suffering from "the gloss over effect". We are given the impression that Big Data value is as simple as hooking a pipe of Big Data into a magic analytical box and out pops the gold you were told is hiding there. Sorry - not that easy.

Which is where the Predicate Free expression comes in. Big Data is sometimes associated with the idea that you can learn a lot from a system simply by examining it output. "If there are fumes coming from a tailpipe it must be a car", or "if you smell food when you are near a factory it must be a bakery" kind of thinking. The work is Predicate Free meaning you don't need to be bothered by hypothesis, testing and proving. A system observed over time will always be the same system observed.

This level of trust can end up being misplaced. Lots of exogenous shocks to systems change them forever. Many times those exogenous shocks have never been seen before and so are not accounted for in any Big Data model. A re-estimate is in order at that point and being Predicate Free trips you up.

So the realism that sets in after giving your Big Data this initial try might also contribute to the skepticism. Big Data should be thought of more like a science project - embarking on a journey of discovery. It takes many dead ends and blind alleys before you reach the promised land. But it is there.

**TWO VIRTUOUS MODELS - THE POWER OF DEDUCTION AND INDUCTION**

June 25, 2014

At a Big Data conference last summer I sat puzzled when listening to a speech about collecting and processing expense spreadsheets. "How is this Big Data practice?", I wondered. "Looks and smells like more traditional Business Intelligence to me", I thought. There was no analysis, only reporting facts from a very small set of data relative to all the data floating around that large enterprise.

It's bothered me every since. We throw around the words 'Big Data' and slap it onto old practices to make them seem new and exciting. This waters down Big Data and confuses those who are new to it.

But there is a proper way to understand the roles older Business Intelligence, statistical modeling, data warehousing efforts, etc. and true Big Data practices fit together.

They fit together because Business Intelligence (and it's cousins) is **deductive** or hypothesis based while Big Data is **inductive** or observational and each informs the other. Here is how.

Deductive models work like the diagram below. Start with a hypothesis about how a system, market, consumer or patient acts. Then collect data to represent the stimulus. Traditionally, the amount of data collected was small since it rarely already existed, had to be generated with surveys, or perhaps imputed through analogies. Finally statistical
methods established enough causality to arrive at enough truth to represent the system. So deductive models are forward running – they end up representing a system heretofore not observed.

Inductive models work the opposite way. They start by observing a system already in place and one that is putting out data as a by-product of its operation. Like exhaust from a car tailpipe, many modern systems are digitally based and thus put out "data exhaust".

And thus Big Data is called Big since the collection of exhaust can be huge. Your cell phone alone broadcasts application usage, physical movement, URL consumption, talk time, geographic location, speed, distance, etc. It is the same with your computer, your car and soon your home.

With inductive models you arrive at some level of system understanding but not truth since the data you have collected is the outcome of the system not the input. On the other hand the deductive model is never the truth since it suffers from small and incomplete data.

Now put these two together to arrive at a virtuous support of one another. Inductive models reveal possible dynamics in a system because of the size of the data and because that data is output based. This can be feed back to the deductive system so it may be improved and thus getting closer to truth. Likewise the improved deductive model points the way for new types of data to look for in and use the inductive model. This is a virtuous cycle to be exploited.

This is often a disagreement in Big Data circles about one type of model over the other, or how Big Data spoils the decades of work of statisticians. This need not be the case. Both models can and should live to together. They are both better for it.

**OF FOOTBALL & POLITICS (IN BRAZIL)**

June 18, 2014

Brazil is the 5th biggest country in the world with 202 million people. Its a fantastic place for food, culture, business and sport. The World Cup of Football (soccer to us Americans) is running there now for a month. The summer Olympics are next summer. No rest for the weary in between, Presidential elections begin in earnest in July and end in October.

Like the US, keeping tabs on the public mood when it comes to politics is a tough task. So many people, a vast territory, and therefore many opinions. Surveys are costly and take time. By the time the traditional survey is complete it may be obsolete.

Digital measure of mood, sentiment, desires, suggestions etc. however is done in real time. This Big Data analysis from social media forms a supplement to traditional surveys. But it can also be done from publications. And then sorted by state, time or other dimensions.

We pulled this together to keeps tabs on the presidential elections in Brazil. The result an application that makes the allocation of electioneering resources quick and responsive. A couple of the hour-by-hour mood trackers by parties and candidates are shown below.

Now imagine doing the same to track how customers feel about your products, services, brands, and then being as fast a World Cup footballer in changing direction, get around a competitor and scoring a goal.
WHAT BIG DATA PATENT ANALYSIS LOOKS LIKE

June 13, 2014

Every company must seek to know something of the opportunities and threats that can impact their operations. The risk of being caught by a surprise is just too great. This includes understanding government regulation, technology advances and competitors. It's these last two where patent analysis comes to the fore.

A company reveals a lot about itself, its technologies and plans for the future in terms of product or service in a patent. If the company wishes to protect the Intellectual Property (IP) of its inventions they form and submit a patent application. This becomes a matter of public record.

Many large companies churn out hundreds of patents per year. Collecting and comparing large groups of patents between companies becomes a daunting challenge when each can run to hundreds of pages of dense technical and legal prose.

Below is a small movie showing a quick way to begin analyzing such a mountain of data. It starts by using a specially tuned version of a Natural Language Processing engine to look for the claims, novelties, authors, prior art, etc. resident in each application. The construction of a co-occurrence matrix of several dimensions lends itself to the modern equivalent of "descriptive statistics" about what is in the patents and how they compare between two companies.

To protect the confidentiality of customers the example in the movie uses false data

Note how simple sorting once the dimensions of interest are built reveal similarities and differences between patents, or patent groups. This is vital to understanding areas of concentration from a competitive point of view and shows where technological investment occurs. Similarities also point out where you will likely be in legal conflict in getting a patent approved. Differences on the other hand show uniqueness and potential areas of advantage both legally and in the marketplace.

Beyond this simple examination more sophisticated patent modeling work can include areas like the following:

• **Patent Vectors**: Analysis of claims, methods, prior art etc. are modeled into vectors or directional motion to help understand speed and aim of the technological advance of a company.

• **Patent Licensing**: Patents can represent up to 85% of a company’s value. Use of the vectors and modeling other industry participants or even competitors can be used to identify new candidates for out-licensing, or cross licensing, or suppliers for in-licensing and open innovation.

• **Patent Litigation**: Patent co-occurrence matrices and vector models can help locate other patents that may invalidate prior art and avoid lengthy litigation.

• **M&A Due Diligence**: Vector models and graph database connection analysis of patents can establish chain-of-title to determine if the assets you are buying or selling are encumbered by unforeseen entanglements and are worth the price.

• **Competitive Intelligence**: These models look at competitors’ IP strategy, portfolio strength, strongest patents, filing trends, and trademarks. This gives a concise view of where opportunities and threats lie.
• **Innovation & Patentability**: Using the vectors to uncover areas of new and unique novelties is the goal in this kind of model. In conjunction with the Competitive Intelligence from about, you form your own IP strategy.

**TECHNOLOGY READINESS LEVEL: EXCELLENT BIG DATA METHODOLOGY**

June 4, 2014

In the 1980’s NASA developed an assessment tool for grading technologies and their readiness for application or market which they appropriately enough named, Technology Readiness Level (TRL). See the reference on this [method here](http://semanticommunity.info/Data_Science/Data_Science_for_OpenFDA).

At Big Data Lens we have pulled this technique into the modern Big Data era and embedded it into our core semantic processing platform. There we use it to hunt down and score new technologies and products for a customers desired uses.

In the old days the technology in question would go through an exhaustive technical review in labs, in the field and finally into composite reports. Now we turn the entire internet on, grabbing and measuring patents, news reports, social media and a ton more stuff to arrive at a TRL level automatically. This means scale. So if you don't know who is making what and how advanced or applicable it is you can use the internet as a database to find out.

**WHY PREDICTIVE ANALYTICS GO WRONG**

May 30, 2014

You collected the sample data, you cleaned it, organized it and feed it to R or your favorite estimator. You applied 23 different models and found the best fit. You tested with new data and confirm your model choice with a review of the confusion matrix. And viola your predictions are sound and strong. Ready to go live. Nice work.

But a week later larger errors are creeping into your predictions. Your customers are complaining. What went wrong? New conditions cropped up in the data - that's what went wrong. In other words, the world changed but your model did not. It's the bane of all good data scientists .... data is not static and dull, changes not observed when you estimated your model come into being and throw a monkey wrench into things. So how to deal with it?

Add probability! You can think forward to the kind of events that *may occur* and include them in your algorithm by specifying them in terms of probability, timing and impact. Then when estimating the model you can do so in a simulation fashion to see how, when and where one or more such events you worried might come true impact the predictions. Because you have, in effect, practiced the future, your algorithms and thus predictions have become much more resilient.

Adding this type of thinking and technique to your predictive analysis bag of tricks helps get you out of the quandary of constantly monitoring, re-estimating and re-deploying. You'll be a stronger data scientist for it.

**HAND WRITTEN NOTES IN ELECTRONIC HEALTH RECORDS**

May 20, 2014

Lots of folks are now in the game of slicing and dicing Electronic Health Records (EHR) or Electronic Medical Records (EMR). There is much to be learned about diagnoses, outcomes, and populations and more when we take the fielded or
structured portions of the EHR and apply Big Data techniques and Machine Learning models. But what of the handwritten notes doctors, PAs, and nurses make in the record? What can that tell us? How does it influence our thinking about the diagnoses, outcomes and populations?

In short, those notes provide the "color" that fielded and rigid forms cannot. The EHR wasn't designed to capture everything so sometimes something the medical professional sees goes into the notes. If the medical professional expresses "worry" or some other emotion in the notes they are in effect taking a first step towards predicting an outcome before it happens. This is the analytical side of the medical professional brain at work - tallying all the indicators from the exam (or multiple exams and tests) into a written expression.

Analyzing those written notes turns out to be a real goldmine for medicine. Utilizing good Natural Language Processing (NLP) to know what those notes say, what they mean and then mixing those results with the fielded information on the rest of the EHR opens new avenues of understanding and prediction. Heart disease is difficult to predict. But recent work in mining both the notes and EHR suggest a prediction rate of > 80% for heart failure.

What boxes are ticked and what choices from menus isn't the whole story. What medical professionals write down does complete the story however. NLP is the way to bring these two sets of information together. Making medicine both preventative and curative is the result.

WHAT DOES HUMAN COMPREHENSION MEAN?

May 15, 2014

When we consider how to make better Natural Language Processing (NLP) technology we are always thinking in terms of human comprehension as the benchmark. When I read something how do I make sense of it? What does my brain do and how fast does it do it? What kind of processing power does it take to do that?

We came across some interesting facts to answer that last question. It is widely agreed that the processing power equivalent of the human brain would be 100 million million-instructions-per-second (MIPS) in speed and 100 million megabytes (MB) in memory. No wonder you need 8 hours of sleep per night to clear out all the leftover junk from processing that much information all day long.

And everyone loves to talk about IBM's Watson winning Jeopardy. But as this article reminds us "Watson needed 16 terabytes of memory, 90 powerful servers, a total of 2880 processor cores, and mind-boggling quantities of electrical power just to wrap its big computery head around the concept of wordplay". So that was no chip on a board that beat Ken Jennings.

But we are headed in the right direction with the work-a-day solutions we use here at Big Data Lens - smarter algorithms, more efficient code, faster processing using cloud infrastructure and more. We may not get to the magic "100 and 100" benchmark even in our lifetime but for each new leap there are excellent new vistas to explore and lots of new insight to be gained.

MACHINE LEARNING TOOLS COMPARED

May 10, 2014
At Big Data Lens we use a lot of tools to estimate Machine Learning models. Namely, we have employed R, Weka, Orange and an online service called BigML. We won't get into the technical details here from a comparison perspective. Just the the good, bad and ugly of when and how to use one tool over the other.

**BigML** is an online service for uploading data and creating Machine Learning models in a fairly automated way. It is fantastically easy to use, very graphically driven and works quickly to get you answers. The number of estimation techniques is limited though they seem to be adding to this all the time. What we really like about this is the ability to have the model spit itself out directly in code that can be deployed right away. Saves a lot of time. Small data sets and models are free. When you need something fast and you don't need to mull over a dozen estimation techniques BigML works great. We like these guys so give them a try.

**Orange** is a slick desktop app that you can workflow your data through a number of steps and estimation techniques. It uses a widget and connector construction on the GUI to move you from raw data to finished model. The number of estimation techniques here is greater than with BigML. Estimation reporting is good but for the true statistician in you may find it frustrating to compare models beyond summary measurements. Orange includes modules for bioinformatics making it unique and rich for the life science crowd. It also include a text mining module but as experts in this field we can tell you to avoid this part of Orange. So overall a bit richer and a bit more sophisticated than BigML and appropriate for quick comparisons across estimation techniques.

**Weka** is a MOOC and university based project. Now in it's 3rd generation, It includes far more estimation techniques that our first 2 tools. But for all its power in estimation techniques it lacks in the user interface department. In out testing and use stability sometimes seemed not up to par even though our data sets were relatively small. But if you like a community approach with lots of plug-ins and don't care about the GUI then Weka can be a good choice. Additionally, for the stats-maven in you Weka will report all the dirty laundry you want.

**R** is like Weka only better in our estimation. The community seems bigger (some will probably argue this), the product seems more stable, and the available IDEs to ride on top of R are good and there is more than one choice. We use RStudio for example. But for us the most important differentiation is what seems like the endless estimation techniques you can find and deploy within R. While writing this blog we checked the number of packages on the CRAN mirror you can install inside R at 5,527. Checking the last 3 days suggested > 10 new or updated packages each day. For difficult problems you would be hard pressed not to find an available estimation package to try. Now its true the graphics require further plug-ins, you have to write your own (sometimes) lengthy scripts and read somewhat verbose logs to do cross-comparisons of model outputs. But the control and variety in solving the toughest Machine Learning problems make R hard to beat. Now they just need a better logo.

**SEARCH AND DESTROY**

May 7, 2014

We added a demo to our site today. Finally. It takes a long time to agree and then build a demo application that highlights the core of what you are and what you do. Plus we have to satisfy customers in the meantime.

But today we launched SNAP - Social News Analysis Platform. See it under the new demo tab. There you add one or more search terms just like any search engine. Using the slide bar you can ask for more or less results by the number of days backwards to look. Click Search and SNAP not just fetches what you are looking for but analyzes it.

[http://semanticommunity.info/Data_Science/Data_Science_for_OpenFDA](http://semanticommunity.info/Data_Science/Data_Science_for_OpenFDA)  
Updated: Sat, 19 Sep 2015 06:44:30 GMT  
Powered by [MindTouch](http://www.mindtouch.us)
This is what a Whole New Web looks like. What do 1,267 look like? What do that many web pages, news stories, social media contributions tell you? How are things related? Where are things happening in total? What are people talking about categorically? Who is society are things affecting? And how to do they feel about?

All of that was always present in the way in which you used the web. But now you can see it. And use it. So go ahead. Search and destroy the boundaries that made the web difficult to use. Go ahead and use SNAP.

Then tell us what you think. We'd love to hear from you and make improvements.

2014 WITH THE GO LANGUAGE

January 9, 2014

Usually we don't write much about the underlying technology that makes Big Data Lens work. We decided to change that here in 2014. We doing it for several reasons.

One is the contribution sharing experiences about the use of different technology makes to the tech community as a whole. We benefit from others experience and so we should make similar contributions to others. Its just part of the tech ethos.

Another reason is we just get excited about certain technology when it works well. The Go Language is a case in point. Now we know not to many readers get excited about about a programming language so we'll keep it high level.

First, the Go Language http://golang.org/ was started by Google. We don't know for sure but assume some of Google is built on this language. So it must be good. Turns out it is. As the website states;

"Go is expressive, concise, clean, and efficient. Its concurrency mechanisms make it easy to write programs that get the most out of multicore and networked machines, while its novel type system enables flexible and modular program construction. Go compiles quickly to machine code yet has the convenience of garbage collection and the power of runtime reflection. It's a fast, statically typed, compiled language that feels like a dynamically typed, interpreted language."

So that is a mouthful but what it really means is the this language is FAST. We have written our core Natural Language Processing (NLP) in Java, Kotlin and now Go. Go gives us about a 30% speed advantage over the other compiled languages. We no longer measure our processing in milliseconds (a thousandth of a second) - we measure only in microseconds (a millionth of a second).

Fast processing is the key to making Web Services work on a economic and customer service level. A 30% time savings on a small batch of processing is nice. But a 30% savings on Big Data scale processing is huge and means less cloud costs, less time customers wait for results, and savings that can be passed on in lower prices.

We'll detail other aspects of the Go language over the course of the year. But there is one more aspect worth mentioning - the icon. Every really good programming language has a some icon to it up. Used to be very techy kinds of things to show the seriousness of the effort. In later years we have switched to friendly looking animals or robots. Go has this happy looking gofer ready to chew through the toughest problems.
NOT LONG FROM NOW ...

October 14, 2013

Big Data holds a lot of promise. Some is current. Some far off. At Big Data Lens we look at the middle distance.

We don't believe slapping the a label on small sets of data using traditional techniques makes it Big Data. I am reminded of this when thinking about a presentation I sat through at a Big Data conference this past summer listening to a big company finance guy crow about his wondrous automatic manipulation of expense spreadsheets. This he called Big Data. It is not. The presenter who followed, however, had it right. His definition stands. Big Data is about doing what could NOT be done before. That means the volume of data processed (in a reasonable amount of time) using techniques never applied before.

So the middle distance is not about processing expense spreadsheets that could have be done with a small python script on 10 minutes work. Nor is Big Data about curing cancer in the short term - though it certainly may play a part in this. in the long term.

The middle distance is about using the whole internet in new ways, to find products that protect soldiers from risk that have not been applied before as we do with the Department of Defense. Or finding technologies to reduce risk and improve planning as we do for the Army. Or even finding and analyzing disparate sources of data for compliance and optimization as we do for a healthcare company.

In the slightly longer middle distance Big Data is about using the internet to understand competitors patterns of behavior and predict their next moves. Or to understand market dynamics and customer groups in ways unknown before to streamline operations. Or to allow social media to inform product design in ways that obviate the need for expensive and costly surveys, research and so on.

The middle distance for Big Data is about being practical. The middle distance for Big Data is about getting the most for this new science without having to wait a long long time. The middle distance is the sweet spot for Big Data.

BIG DATA FROM TWITTER

May 10, 2013

I both like and dislike this article from Smithsonian.com today about an analysis of Twitter data.

I like the analysis because it adheres to the right principal about Big Data - namely that simply collecting the data without analysis does nothing for you. So the study authors did do a lot of good analysis on place, language, time of day, distance between tweets and retweets and so on. Done properly, it gives an interesting picture of the way humans communicate - early and often - as the saying goes for voting.

I like even better the added data the investigators brought into the examination. They studied the time of the tweet so they could know if natural or artificial light was present during the tweet itself. Subtracting these and mapping it shows either the lack of rural electrification or that Twitter is a banned political substance in some parts of the world (e.g. Iran and China). The great thing the investigators did here is the use of data, properly analyzed, tells you a lot about other
things you were not expecting to see or even asking questions about. Something that is self-evident and powerful once it is revealed. That is the essence of good Big Data.

What I don't like is what they could have done with the Twitter data. Certainly I care about the novel insights they can glean from the data. But in the end they never tried to look at what people were saying, what did they mean - what fears and pain, what successes and joys are all those millions and millions of people talking about? That's how we would have approached this endeavor here at Big Data Lens.

Maybe it's just dribble. But like the discovery of politics and energy you can probably be sure there is something deep and powerful in all that human communication waiting to be found.

**CORRELATION & CAUSALITY**

April 17, 2013

Yesterday we caught the columnist David Brooks' piece on Big Data. You can find it here. He is right to point out that efforts by Big Data proponents to break the long-standing axiom that "correlation does not equal causality" in data science is misguided.

This phrase "Correlation does not equal causality" was never meant to downplay or white wash any of the analytical techniques that can be misused to establish causality. Rather it was meant to make sure data scientists added the narrative to the statistics.

The narrative is the explanation of why the correlation appears. 20 years ago diapers and beer being bought at the same time in the convenience store was more than a correlation. The narrative that went with it - young fathers sent out into the dark cold night to fetch might as well as well pick up some new family stress relief along the way - launched an entire new field of Information Technology called Business Intelligence. Without the narrative its just statistics, and might well be a spurious correlation and little understanding in it. The beer and diapers would have stayed in separate isles.

Policy makers in government change course when they hear a story backed up by numbers. Business leaders reallocate resources when they hear a story backed up by numbers. If you are part of Big Data do the numbers, build the machine learning algorithm, but tell the story also. Else your efforts will be far less impactful than they should be.

**COOL OLD TECHNIQUES (PART 2)**

April 15, 2013

Big Data is breathing some new life into a another analytical technique we dusted off recently. It's about understanding the development of technology over time into finished products.

Every company who dreams up a new product is faced with choices along the way in terms of inputs. Buy or build, buy this one or that one, buy that one but modify it and so on. Each input choice made influences the timing and finished product features. Input A is better for the finished product but might take longer to complete holding everything up so perhaps the lesser input B is sufficient.
Big Data comes in because it helps inform the choices, capabilities and timing of the inputs into the finished new product. The analytical technique then is one of specifying each input in terms of the probability of completion and capabilities by a certain time. Now you have a network you can simulate. It looks (in a simplified form) something like this;

Inputs are specified by probabilities of time to completion and capabilities. Combining inputs into sub-systems which are then also specified by probabilities of time to completion and capabilities. Note also that the choice of one input over another are present. Or in some cases the condition of needing 2 inputs to be complete instead of a choice one over another is also present.

These networks can get very big very fast. In the past, this technique might be used to model a future enemy weapons system. A sub or a plane is fiendishly complex so simulation techniques are used to estimate likely outcomes, outcomes with the shortest timeframe, outcomes with the most robust capabilities and so on. But also in the past the networks were so complex and time consuming to estimate they were rarely kept up to date. But Big Data comes into play now by keeping such networks alive and useful - both changes to the inputs as well as the faster machine learning re-estimations that make the simulations useful.

**COOL OLD TECHNIQUES**

April 11, 2013

Recently we've been looking back at some old analytical techniques to see if they apply in a big data era. Some do and some don't. Those that do share some common traits ... they are insightful in ways that stand the test of time, and they are just plain useful.

One such technique looks at the kind of things that throw off every forecast - things that have not yet happened and so were never built into a model or equation. There are lots of things like this. In the old days you would seek out subject matter experts and gather a range of opinions on things that could bugger up the future. Now we plug and play big data techniques to survey and account for things that have not yet happened.

Gather enough of these and you then must simulate the outcome to see how things combine to make multiple futures. Hedging our bets you say? No No. Its how you use the range of outcomes that makes it powerful.

Planning is always about managing uncertainty and that is exactly what a range of outcomes gives us. Ask "how likely is it the market demand for our goods will be at level X?" and you get the answer 73%. That is certainly a better answer than a point forecast you know is wrong before you have even heard the answer.

Smart companies use this old technique in the big data era to be more thorough in their planning, risk assessments and resource allocations. Governments use this technique to measure the likelihood a foe will acquire a capability or advance a technology.

In the end this old technique lives again. In fact it gets better with more data, accounting for futures as big data collection and analysis techniques inform them where we used to rely on only a small number of subject matter experts. Welcome to Big Data.
BIG DATA IDEAS THAT OPPOSE

November 18, 2012

“The more data that you have, the better the model that you will be able to build. Sure, the algorithm is important, but whoever has the most data will win.”—Gil Elbaz, Factual

“With any Big Data project, if you don’t spend time thinking about analysis, you’re wasting your money. You must have a structured idea of what you want to get out of unstructured data”—Ron Gill, NetSuite

So lets see if I understand. With Big Data any lack of precision can be beat by adding more data to the algo. Pour more into the sausage grinder and you get better sausage. Simple right?

But wait. Our second quote tells us to carefully plan and understand the nature of Big Data so you know what algo to apply to it. Or all is lost.

So here are two Big Data ideas that oppose each other. In reality you need a bit of both. Or as the the saying goes this is a case of each idea being necessary but not sufficient on its own. But take them together and you do have the makings of good Big Data practice.

In any project, many iterations are needed to discover the inflection point where more data generates a diminishing return in terms of a quality algorithmic outcome. You must seek it, no matter how large the data as input turns out to be, since to do otherwise sacrifices accuracy.

But likewise, not having a theoretic underpinning - that structured plan to handle the unstructured data - is a fools errand. Too many Big Data projects rely on simply adding more data to goose the algo's precision. Spurious is the term typically used when you can observe high correlation values but not explain them logically. Without this "story-telling" understanding of the algorithm you are on shaky ground.

Use both these ideas as good practice for good Big Data - everyday.

MICROSOFT GETTING INTO BIG DATA ... OR THEY ALWAYS WERE

October 30, 2012

News items like this one from the New York Times are great for detailing how companies large and small are getting into Big Data. So now Microsoft is getting into the game as well. Or is it? Seems to us they always were. Even if they never called it Big Data.

What is striking is how Microsoft portrays the use of Big Data - as a power behind their current products (e.g. Excel or Outlook email) and not for stand-alone use. Their chief data scientist is portrayed as working on "incremental" steps for the last 20 years. Really? And 100 of their PhD's come together to brainstorm new ways to use Big Data and still it looks like it all hidden inside the current crop of products.
Plenty of cash, bright minds and no deadlines sounds like a recipe of middle of the road Big Data nothingness to us. Nothing focuses the mind like a bootstrap, small budget, gotta-get-it-done yesterday Big Data project. That's where we have found the most success. You should too.

**A NEED FOR TRUST IN BIG DATA**

October 26, 2012

There seems to be a divide between Big Data sources when it comes to trust. Big Data efforts based on internal sources (e.g. log data, transactions, etc.) are trusted. But apparently Big Data efforts based on external sources in the form of social media etc is not. See this IBM study for more details.

But this may be due to one of two things. The first is simple experience with the external data. Yes it comes from the wild and wholly world of the internet but like the internal data that does not mean there isn't gold in there.

Which leads to the second point - technique. There is a wild array of techniques you can use to analyze your Big Data. Finding the right size, shape and analytical technique in order to find that gold is not easy. It takes trial and error and great skill in interpreting the results.

Like all marketing efforts, Big Data is now suffering from "the gloss over effect". We are given the impression that Big Data value is as simple as hooking a pipe of Big Data into a magic analytical box and out pops the gold you were told is hiding there. Sorry - not that easy.

So the realism that sets in after giving your Big Data this initial try might also contribute to the skepticism. Big Data should be thought of more like a science project - embarking on a journey of discovery. It takes many dead ends and blind alleys before you reach the promised land. But it is there.

**WHY NOT MAKE YOUR OWN BIG DATA?**

October 25, 2012

Lots of studies look at the sources of Big Data. The info-graphic from IBM shown here suggests that the majority of Big Data comes from 6 sources. You might call these sources "the usual suspects". But to our way of thinking what percentage of all the worlds information do these 6 sources represent? Not much is the answer. Think 80/20. 80% of the worlds information is in a written (e.g. words) form. None of the 6 sources listed here are words - well social media is but limited in its length.

So where is the rest of the worlds information and does it constitute Big Data? It is the long form written news, blogs (like this one), journals, books and so on. And yes that too is Big Data. Or it can be once it is processed.

That is where Statistical Natural Language Processing (sNLP) comes in. Transforming the written word into machine understood data and viola you have Big Data too. Now mix and mosh it with the transactional, log data, and so as listed above and you've got a rich stew from which you can model and predict.
Competitors actions - they're in the text. Scientific discoveries - they're in the text, not in log files and transactional data. Include the written stuff by using sNLP and you open a whole new world of Big Data for your use.

So if the above chart looks pretty limiting in terms of what you are going to get out of Big Data because what goes in is pretty limited think again and start generating your own Big Data - and profit by it.

**CAUTION - MESSY LANDSCAPE AHEAD**

October 21, 2012

A landscape is a fashionable thing in the tech world these days. It is an attempt to put all important players and their contribution to the ecosystem into a single picture. We found a couple of these recently online for Big Data. One of these is below.

We cannot for the life of us figure out if a landscape like this is needed because a complicated Big Data ecosystem needs some ‘splaining or if such a complicated landscape is meant to impress. The cynic in us says both.

Our read on landscapes like this are two fold. One it says the industry is not mature. So caution lots of consolidation ahead. But, two, much more importantly, if you are a newbie to Big Data this picture is going to scare the hell out of you. You'll retreat to the safety of your BI intelligence tools (which you could argue don't belong on any Big Data landscape) and never give Big Data a try. To complex. Methods aren't yet formed. Too much trial and error and no real ROI yet. The boss will give any budget to buy from this spaghetti western. So a landscape really doesn't advance Big Data into the mainstream.

At Big Data Lens we take a different approach. We try to provide an end to end service that takes the worry and hassle out of data design, data acquisition, data massage, data analytics, data prediction and in the end the production of VALUE from Big Data.

Ignore scary landscapes. Stick to simple steps with simple tools. That is how you get value from Big Data.

**POLITICS & POLL OF POLLS = BIG DATA**

October 19, 2012

Even if you are not a fan of The New York Times Nate Silver's Blog called FiveThirtyEight - Political Calculus is an excellent example of Big Data at work. It contains all the elements you should expect from any Big Data effort. Lots of input in this case are the various polls conducted around the country. Then lots of analysis about the data. Including easy to understand charts and graphs that tell a story. But also the written explanation - pointing out things you might miss in a graph or detailing things a graph can never display.

So whatever your political leanings wade into this and see what Big Data looks like from a polished professional effort.

**WHAT IS MORE IMPORTANT DATA OR ANALYTICS?**

October 17, 2012
Big Data is such a catchall phrase. Besides the on-going arguments about the definition of Big Data we'd like to inject a different debate. Namely, what is and also what should be the balance between the data side of Big Data and analytics side of Big Data.

So here is a crude measure to figure out what is the current balance between data and analytics. The search index trend over the last 18 months or so between the search terms "Big Data" and "Big Data Analytics". That picture is shown here, where the blue line is the trend line on "Big Data" and the red line is the trend line on "Big Data Analytics".

Two things pop out. One is Big Data started before anything thought was given to the analytics side of the equation. Second, the analytics side of Big Data is not gathering as much speed as Big Data broadly.

This is troubling. Why? Because in the end Big Data should not be about the acquisition, manipulation and storage of just more data. You have to know why you need to collect it, what you intend to do with, and what question the data should help answer. And that has everything to do with the kind of analytics you are going to subject the data too.

The real work of Big Data comes before you ever start collecting one kilobyte of data. More on that next time.

**SOLD OUT - YIKES NYC STRATA / HADOOP CONF.**

October 17, 2012

Only in its second year, the O'Reilly Strata / Hadoop Conference is sold out? Really? Says a lot about either the popularity or the myth of Big Data. We choose to believe the crush of attendees is about the popularity and importance of Big Data and not people trying add to the mythology of it all.

Then if you check the Data Business Meetup happening at Bloomberg headquarters during DataWeek NYC you find 390 attendees who want to hear from the all start VC casting call. No spots left to attend this either.

Lots of Big Data talk and walk next week in NYC. We'll be making the rounds, look for us !!

**DATA EMANATION**

October 15, 2012

From our devices, our checkins, our emails and interactions - even our bodies - we emanate data. A lot of it. Get your hands on enough of it and let a machine learn from it, dispensing with theories or preconceived notions, and thats what Big Data is.

It's why Big Data is different from Business Intelligence. It's why Big Data is different than the last IT buzzword and why it has staying power. Because the data is not going away. It's only going to grow and move faster. We needed new tools, new thinking and new velocity to utilize this raw resource from the internet to gain insight, make customers happy, avoid mistakes, beat the competition, discover new medicine and make government work better. So we got Big Data.

Long live Big Data!
BIG DATA SPENDING TO DOUBLE BY 2016

October 15, 2012

Gartner does a bang-up job racking and stacking the number to size up a market. Here they have done this by looking at all aspects of spending on big data. This includes the continued need to understand customers, the use of social media to do so and buying the big data expertise even as they build those capabilities in-house.

Rest of the article is here.

BIG DATA GEEKS - HIDING IN PLAIN SITE

October 15, 2012

Our guess is you have not been handed a business card that declares someone's title as "Big Data Scientist" or "VP of Big Data" just yet. No worries. Professionals who wield power over data have long been among us - just not perhaps in the most glamorous of roles. Up till now.

Turns out the new rock stars are those who can whiz through mountains of data. And typically they are called actuaries, economists, mathematicians or to pull a really old-ish title out of the bin - operations research. You gotta love this interactive graph from McKinsey that shows how many Big Data pros there are already and where they work. Snatch them up if you can!

Find the rest of the report here

BI OR BIG DATA? - START BY THINKING BACKWARDS

October 13, 2012

In business there seems to be a debate on where BI ends and Big Data begins. Here is an easy way to figure it out. Start by thinking backwards. This means think first about the decision that needs to be made, the nature of it, the data that may exist to support it, where the data is, and how you might analyze that data in order to arrive at an answer. Looking at these attributes will help you understand where BI ends and Big Data begins.

Is the decision to be made new? Not always true but generally BI is about updating old data, models and answers or perhaps taking a small tangent on old data and models to seek new answers. Big Data is more about solving something that eluded previous analysis.

Lot of BI is inwardly focused - running the trucks on time and so forth. Big Data looks to focus on the outward. Things like regulation, competition, advances in science & technology and so on. Which leads to the 3rd question to ask - does data already exist that you can feed into your analysis? For BI the data is mostly lying around inside the company. For Big Data the data is the fire hose of social media, lots of news feeds and / or a collection of API feeds you have never looked at before.
An intuitive understanding of what to do to the data once you have it is another hallmark of BI vs Big Data. BI is about traditional well understood analysis. Big Data is about machine learning - discovering what the underlying patterns, connections and models are without deciding that ahead of time.

The biggest bugaboo is whether the data is structured or unstructured. As a general rule for BI this means the data is numerical where the unstructured stuff is messy text - written information that conveys essential ideas in a million different ways.

Answer the questions in the chart and you'll know how to proceed.

**STARTING SMALL WITH BIG DATA**

October 11, 2012

An October 3rd blog post in Harvard Business Review - found here - provides sound advice on succeeding in Big Data by starting small. Don't over shoot, try all the steps needed in a Big Data project at a small scale before making the full investment in going Big Data scale. All safe and correct as in a traditional IT project.

But there are several missing ideas here that are unique to Big Data. One is the nature of the data itself. The nature of Big Data is .... well Big. Google Trends is a fascinating Big Data project. It comes out of Google's never ending collection of search query strings. But had Google tried to start small on the presentation of their data none of it would have made sense. They had to wait until a sufficient amount of data piled up. Only then did Google realize they had something interesting to compute and show to anyone.

For example the chart included here shows the interest level of the iPhone and Android over the last 6 years. iPhone is the blue line and Android the red. The various peaks in the iPhone search index are, as you might surmise, around the release of successive versions of the phone.

Analytically the story of the level of interest and the competition between iPhone and Android doesn't become clear until many data points have been collected and enough time has passed to see the trend.

The literature on machine learning also suggests that more data is better. In fact in a study on the efficacy of various algorithms this 2011 paper shows that an "ensemble" use of 4 or more learning algorithms begins to approach a 90% human accuracy threshold. This shows the power of not starting small but going big with data as well as going big with the analytical approaches to surround it produce human-like comprehension. And after all what confidence would we have in Big Data unless it was telling us something we would conclude on our own if we had the time to absorb it all.

**WHAT BIG DATA LOOKS LIKE**

October 9, 2012

Many so-called experts have trouble telling you what Big Data is. There has yet to emerge a standard definition. Some try to describe it as as being three-dimensional as in 3Vs - meaning data that has a Volume so large it cannot be processed by traditional database or processing tools, or the Velocity of the data (what is added or subtracted from a dataset) that again traditional tools could not keep up, and the final V being Variety of data that traditional systems have
trouble handling. More recently analysts have added a fourth V for Virtual - to distinguish data that is online as opposed to data an organization already has captured or owns.

No matter what definition you subscribe too we thought looking at Big Data in action - at least in a raw form) can bring to life these abstractions. Take a look at the short video below of the real time tweets about iPhones on the left and Android on the right. Then ask yourself if this flow of data meets the 3-4 Vs described above. Sure looks like Big Data to us.

Then leave us a note on how you might guess we would capture and process this Big Data into something useful and impactful. Since in the end, after watch the data flow by your eyes for 38 seconds you start to go cross-eyed. Big Data Lens is here to prevent that.

**IF BABOONS CAN DO IT ....**

October 8, 2012

Here is a fascinating study on how we acquire the first step in forming language. The researchers could get Baboons to recognize real words from just non-sense strings of letters with 75% accuracy after roughly 10,000 attempts. This is pattern recognition and is fundamental to taking the next step which is to associate words with meaning followed finally by the ability to distinguish one meaning over another given a certain context.

And a here is a video of the baboons correctly making those choices over at National Geographic

it is this kind of clever patter recognition that forms the basis of our statistical natural language procession machine learning algorithms. This approach is at the heart of Big Data Lens effort to make all the web a useable - understandable database.

**BIG DATA - OLD BECOMES NEW AGAIN (WITH SOME TWISTS)**

September 29, 2012

So what is old? Some of the techniques are old but they still rock. Regression is the most powerful analytical technique ever invented. It forms the basis for all other modeling techniques including machine learning. Why? Because of the mathematical foundation on which it is built. Namely - calculus and the great insight that levels or scalers or measurements don't matter. Only the way they change matters. Its the first derivative - remember that? It's what happens AT THE MARGIN that is really important.

You hear this everywhere. - in Economics, Finance, Science, Engineering. What happens to the bridge's structural integrity if you add 1 more pound on top of it? What happens to the efficacy of a new drug if you one more ml of an ingredient. And so on.

And what is new? Taking the effective cool old stuff and applying it in ways and at a scale no ever imagined. Huge quantities of input data are just a start. Then processing that data with the cool old techniques up on self-expanding cloud services to get a job done that less than a generation ago took most of NASA to accomplish. Now you do it from your desktop and a couple of interfaces. And the coolest of all? Letting the machine learn and show that it can do so by finding patterns and making fits of data that would have taken a modeler months of work, if ever.
So welcome to Big Data. Maybe it’s geeky but they outcomes are cool and getting better all the time.

CONTACT

LEAVE US A NOTE ON WHAT YOU WANT BIG DATA TO DO FOR YOU. WE'LL BE SURE TO GET RIGHT BACK TO YOU TO DISCUSS YOUR PROJECT.

SURVEY

Here is a short 10-question survey that will jog your thinking about the best practices around Predictive Analytics. Those who leave their name and email will get the full results. Otherwise looks for some reporting on this in our blog section towards the middle of July 2014.

1. Are you actively involved in predictive analytics in your company?

☐

☐ Are you actively involved in predictive analytics in your company? Yes

☐ No

☐ We’re thinking about it and trying to figure out how to do it.

☐ Other (please specify)

2. As the starting step in your predictive analytics work do you ...?

☐

☐ As the starting step in your predictive analytics work do you ...? Start with a hypothesis of what you are seeking to prove?

☐ Meet with executive sponsors to understand their needs?

☐ Seek the opinions of Subject Matter Experts (SMEs) to know their analytical thinking?

☐ Other (please specify)
3. Thinking about data sources for your predictive analytics work which of these do you include?

- Use data from internal sources
- Use data from external sources
- Use data that is a proxy for the real data you want but cannot find or access
- Use only fielded / structured data that is known, curated, and reliable.
- Include unstructured data (e.g. text data) from web sources even if it requires additional processing steps (e.g. Natural Language Processing - NLP)?
- Other (please specify)

4. What is your preferred way of storing the data?

- What is your preferred way of storing the data?

5. Do you conduct a pre-processing stage on the data after acquiring it?

- Do you conduct a pre-processing stage on the data after acquiring it? Yes, we remove any record that is incomplete
- Yes, we employ inference techniques to fill in records that contain any missing information
- Yes, we meld and convert data to the format needed for the analytical work to come
- Yes, we find and remove duplicate records
- Yes, we find and remove records that conflict with one another
- Yes, we calculate and add data based on what has been collected to enhance each record
No, we let the data "speak" for itself

Other (please specify)

6. What one thing do you do more often before modeling and predicting using the acquired data?

What one thing do you do more often before modeling and predicting using the acquired data?

7. Which Big Data Predictive Analytics Suite would you use to model your data?

 Which Big Data Predictive Analytics Suite would you use to model your data? Orange

Weka

R

SAS

SPSS

Combination of the above

Other (please specify)

8. Which of the following measures of fit do you use and value most highly for a predictive model?

<table>
<thead>
<tr>
<th>R-squared or similar</th>
<th>Highest Value</th>
<th>Moderate Value</th>
<th>Low Value</th>
<th>No Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Highest Value</td>
<td>Moderate Value</td>
<td>Low Value</td>
<td>No Value</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------</td>
<td>----------------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>Confusion Matrix</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Review of Graphical Display</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

9. Which Predictive Analytics techniques do you use and value most highly?

<table>
<thead>
<tr>
<th>Technique</th>
<th>Highest Value</th>
<th>Moderate Value</th>
<th>Low Value</th>
<th>No Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Discriminant Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multinomial Logit Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probit Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bagging</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boosting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Forests</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural Networks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support Vector Machine (SVM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choice Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensemble Methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
10. If you like, tell us about yourself

If you like, tell us about yourself  Name

Company

Country

Email Address

Done

Powered by SurveyMonkey
Check out our sample surveys and create your own now!

NEXT