Working as a *data science* intern
At Intelent Inc.

Rishab Srivastava*

August 12, 2017

Abstract

I interned as a Data Science intern at Intelent Inc, a big-data analytics consultancy firm based out of Princeton, New Jersey. Intelent takes on projects from fields such as pharmaceutical, insurance, government regulatory agencies and other data-driven industries.

Founded in 2012, by experts with over two decades of experience, Intelent provides organizations a platform for integrating and analyzing data, deriving actionable insights and driving measurable impact. Intelent provides business value and enables adaptable and scalable solutions for our clients using the right mix of people, process, proven and exploratory technologies.[1]

Intelent is a subsidiary of TAKE Solutions Ltd., a listed company on Bombay Stock Exchange. TAKE[2] is a global technology solutions and service provider of Life Sciences, Supply Chain Management & Enterprise Solutions.

Figure 1: Intelent Inc, a big-data analytics consultancy firm.
www.intelent.com

*Rishab is a Computer Science and Economics double major at the University of California, Berkeley.

[1] http://www.intelent.com/about-intelent

[2] http://www.takesolutions.com/

Contents

1 My internship experience

1.1 Official job description

Officially, as a *data science* intern, I was responsible for working with Intelent engineers to design and build advanced analytics tool and solutions across multiple industry use cases. My job entailed working on advanced data management, reporting and visualization techniques supporting end-to-end business processes.

1.2 Workflow

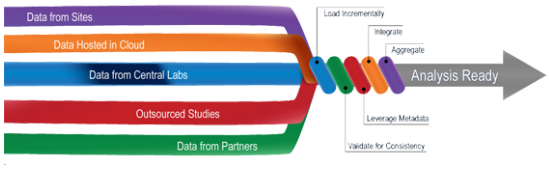

Figure 2: Data preparation and analysis workflow at Intelent
www.intelent.com/data-integration

The advanced analytics/data science process begins after preliminary analytics. According to my course *DS100: Principles and Techniques of Data*

Science, is preliminary inspection of a sample of the data is termed as EDA: Exploratory Data Analysis.

At Intelent, a typical project team consists of a project head, one big data engineer and a off-shore few data analysts. A request for proposal (RFP) for a project came from a pharmaceutical company who wanted data integration/data analytics services for a use-case such as clinical trials.

As an intern, I was handed the data after the big data engineer was done extracting it and standardizing it. Usually, the data was either scraped from the Internet using a java library like jsoup or it was obtained from the client, which were usually clincal research organizations (CRO) i.e. third-party organizations which conduct clinical trials on behalf of pharmaceutical companies such as Navitas Life Sciences, Novartis and Cipla. I analyzed datasets from the Open Government Food and Drug Administration website[3] to get a better understanding of the pharmaceutical industry domain.

1.3 Research work

Under the my supervisor and principal analyst of the company, Mr. Suresh Selvarangan, I was also assigned some research work in the field of biostatistics, FDA safety reporting standards and regulations, and *electronic Clinical Report Form* (eCRF) software.

Intelent, in association with its partner pharmaceutical company, *Navitas Life Sciences* have developed a product called **ThoughtSphere**[4], which is a cloud-based Risk Analytics platform that enables adaptive/risk-based monitoring for clinical research studies.

Figure 3: ThoughtSphere, a centralized risk-based monitoring platform for clinical trials

www.thoughtsphere.com

2 Industry vs Academia

2.1 Format of the data

As soon as I received the data, I realized it is not in any format I am familiar with such as comma separated values (CSV), eXtensible Markup Language (XML) or JavaScript Object Notation (JSON).

The pharmaceutical domain works on the **SAS (Statistical Analysis System)**, which is essentially a modified standardized version of the XML format,

[3] https://open.fda.gov/
[4] http://www.thoughtsphere.com/

as required by Food and Drug Administration (FDA) to analyze the drug and its effects before it goes to market.

2.2 Understanding the data

I imported the sas7bdat python package into the jupyter notebook to convert the SAS data into a **Pandas dataframe** which can be analyzed on **Jupyter**. As a sample, I looked at the Adverse Events (AE) dataset. Each row/tuple represented a single adverse event. The dataframe contained 34 columns, each one of them with a different term.

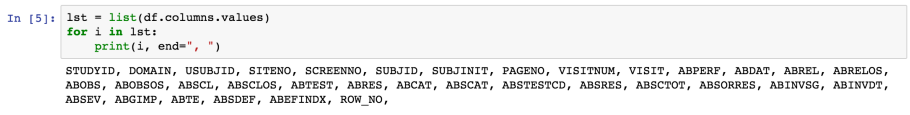

Figure 4: The column headers for the data

This made me realize how important it is to understand the **domain/discipline** that the data belongs to. These terms belong to a model called **Study Data Tabulation Model**[5], which defines a standard structure for human clinical trial (study) data tabulations.

2.3 Irregularities in the data

2.3.1 Transforming the data

During my Data Science course, I remember Professor Hellerstein stating that real-world data is far from perfect and most data scientists spend more than half their time integrating, cleansing and transforming the data without any actual analysis. As a result, data wrangling or assessing and transforming the raw data is essential[6].

While working with real-world data, I didn't realize how incomplete the data might be. To my surprise, more than 40% of the tuples of few columns were filled with empty Strings or NaN values. Hence, **trimming** or **winsorizing**[7] is important.

2.3.2 Granularity

The data could be analyzed in a variety of ways depending on which column is used as the primary key - STUDYID (Study ID), SUBJID (Subject ID), SITENO (Site number) or VISITNO (Visit number).

[5]https://www.cdisc.org/standards/foundational/sdtm

[6]https://drive.google.com/file/d/0B2k285AK-3KEbE5GQ3BCZXI1V28/view

[7]https://stats.stackexchange.com/questions/90443/what-are-the-relative-merits-of-winsorizing-vs-trimming-data

I decided to analyze the data using SUBJID to recognize long-term trends in the patients and how they react to the stimuli, how the drug causes adverse events, and how is the event dealt with.

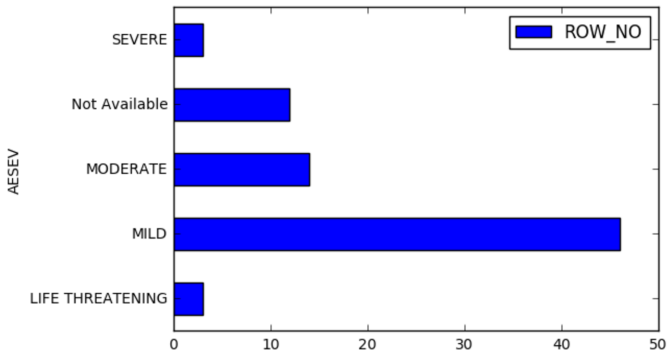

Figure 5: Distribution of the severity of the adverse event

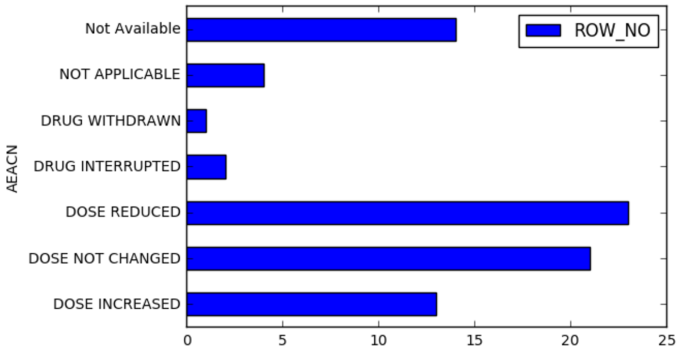

Figure 6: Measures taken to counter the adverse event caused by the drug

3 Conclusion

I am not at liberty to discuss the conclusions derived from the data as the microscopic trends and inferences derived from the data are bound by confidentiality.

In conclusion, I would like to thank my supervisor Mr. Suresh Selvarangan as well as my colleagues Mr. Naveen Prasath and Mr. Kalyan Gopalakrishnan for making this internship a remarkable learning experience for me.

These two months have made me realize how my courses at Berkeley are perfectly geared to achieve my future career goals in the field of Big Data Analytics. I will continue taking courses in Computer Science, Statistics and Math to further strengthen my technical skills which can be utilized in the industry.

Over the course of this internship, I have picked up a variety of technical and soft skills as well as understood a lot about how industry *Data Analytics and Consulting* is different from its academic counterpart. I want to continue this interest I have generated in Big Data by taking on more academic as well as experiential pursuits in this extremely engaging field.