A Pricing Model for Data Markets

Judd Heckman, UC Berkeley School of Information
Elizabeth Peters, UC Berkeley School of Information
Nikhil Gopinath Kurup, UC Berkeley School of Information
Erin Boehmer, UC Berkeley School of Information
Milad Davaloo, UC Berkeley School of Information

Abstract
As companies and organizations explore the booming frontier of data, they are operating in data markets that are largely unregulated. One of the foremost challenges within these emerging markets is establishing an accepted methodology for assessing the value of datasets. Current data pricing strategies are often driven by the seller, with little visibility into the cost of collection, cleansing, and packaging to the buyer. This asymmetry of information results in a lack of pricing transparency, hurting the seller, who is unable to price optimally in the market, and hurting the buyer, who cannot strategically assess pricing options across data service providers. A more structured data market with a standardized pricing model would improve the transaction experience for all parties. In this paper, we describe a potential dataset valuation model and the impact such a model could have on data markets. We also explore how the model would influence certain practices, such as by adding proprietary datasets as assets on corporate balance sheets, as well as how it could contribute to the formation of a futures market for data.
Keywords: data valuation, data markets, pricing model, machine learning, regression
Citation: Heckman, J.R., Boehmer, E.L., Peters, E.H., Davaloo, M., Kurup, N.G. (2015). A Pricing Model for Data Markets. In *iConference 2015 Proceedings*.

Acknowledgements: Steven Weber, UC Berkeley School of Information
Research Data: In case you want to publish research data please contact the editor.
Contact: jheckman@ischool.berkeley.edu, erin@ischool.berkeley.edu, elizabeth.peters@ischool.berkeley.edu, dmilad@ischool.berkeley.edu, nikhil@ischool.berkeley.edu

1 Introduction

Data science presents the potential for building smarter systems and growing knowledge bases, inspiring widespread interest and optimism in the field (Anderson, 2008). Yet the current market for selling data among interested parties is failing. An efficient data market would facilitate more efficient and impactful advancements, allowing groups to strategically purchase available datasets to avoid the labor costs and skill requirements necessary for data curation, confidently understand how a dataset was collected and cleaned, and reliably generate a profit from sought-after datasets. Instead, the existing data market is largely ad hoc, with data trading through informal partnerships or private agreements. One of the foremost challenges within this emerging market is the lack of an accepted methodology for assessing the value of datasets. A myriad of players are putting value on data in order to generate a profit, each using a custom model for data valuation. These players include corporations that collect data on their products or services, organizations that gather data from targeted populations, and third-party data aggregators that provide datasets derived from various sources. Therefore, the markets that exist today tend to be vertical within an industry and narrow in scope, with little connection to data value at an overall level.

2 Background and Literature Review

In the current data marketplace, little transparency exists between buyers and sellers regarding how data has been collected and manipulated prior to sale and how it will be used post-sale. This is in part a competitive strategy for companies, but it can hinder the market, as demonstrated by Akerlof, Spence, and Stiglitz (1970) in their analyses of markets with

asymmetric information. This lack of transparency leads parties involved in the transaction to be misinformed and results in what Akerlof termed a “market for lemons.” This market asymmetry develops when sellers have more information about a product than buyers, allowing sellers to charge more for goods than is justified by their quality. This lowers the value of all goods in the market, as buyers are less willing to take the risk of making a purchase because of their fear of paying an unfair price for a sub-standard good. The market may dissolve as sellers of quality products are discouraged from selling their goods in a market inundated by inferior products. Thus, we chose to study data valuation in order to propose a rigorous and transparent pricing model to enhance the data market and lessen the likelihood of the “lemon” market asymmetry.

If a standard model for data pricing existed – one that considered many aspects of value such as the age of the data, the reliability of the sample, and other factors – sellers would be able to price optimally in the market and buyers could make appropriate comparisons across data service providers to get a fair price. In this sense, a more efficient existing data market is ironically the black market, where stolen credit card information is dynamically valued according to openly-available metadata such as the card’s credit limit and the “freshness” of the data (or equivalently, the likelihood that the card owner has cancelled the compromised card) (Leger, 2014). If the data market adopted some of these valuation strategies and standardized a pricing model, the transaction experience for all parties would improve drastically and facilitate more efficient and effective data science.

It is important to recognize previous research in this area. Moody and Walsh addressed the subject of asset valuation of information in their 1999 paper. Their premise viewed data as a raw material, information systems as the manufacturer, and information as the end product requiring valuation. This paper addresses valuation of data itself rather than focusing on the even more abstract concept of information. This should prove to be more useful as the distinction between ‘information’ and ‘data’ often lies in its use rather than its inherent properties.

3 Research Justification

The potential impact of constructing a functional pricing model can be realized by examining how this problem is similar to a pricing issue that evolved in traditional financial markets. Options contracts, which grant the buyer the right to buy or sell an underlying asset at a specified strike price before the expiration date, have long been used in some form or another. However, the model developed by Black and Scholes gave legitimacy to the then newly-formed Chicago Board Options Exchange, a guaranteed clearing house for trading options with standardized terms (1973). Black-Scholes defines a stochastic partial differential equation which calculates the theoretical price of an option over time. The model incorporates various factors including the current value, returns, and volatility of the underlying asset; the strike price and time to expiration of the contract; and the prevailing risk-free interest rate. As either volatility or the time to expiration decreases, the value of the option depreciates. While not entirely analogous, data valuation has similarities to this pricing model as it is determined by a complex interaction of multiple factors, which could include both a concept of volatility and time decay. A generalizable, scientifically rigorous approach to pricing data would likewise help to legitimize and standardize the future marketplace for data.

Assuming a grand pricing model can be constructed, a futures market for data could also be established. A futures market, typically associated with commodities such as corn or cattle, trades contracts that specify the quantities and price of an underlying asset at which the asset will be bought at a designated time in the future. This is already happening with patents. Exchanges are in the process of being formed, such as the Intellectual Property Exchange International (IPXI), which garnered investments from Royal Philips Electronics as well as the Chicago Board Options Exchange (Sachdev, 2013).

Futures contracts allow for greater speculation in an asset class, as they add the dimension of time to the expected move in price. A new class of financial analysts and market speculators specializing in data could help guide the market to produce the appropriate data by sending market signals about their understanding of the future valuation of data. For example, if a futures market had already been in place, speculators may have anticipated the incredible expansion of transportation services companies such as Uber and Lyft. They would have then invested in futures contracts on data believed to be valuable for those companies, such as aggregations of public transit data or neighborhood demographic data. This investment would in turn send a signal to companies that have the capacity to produce such data to focus their efforts there instead of towards a less valuable avenue. This would result in greater competition and availability in the marketplace for such data, allowing innovative companies to grow faster and more efficiently due to higher availability, quantity, and quality of their needed data.

Hedging is a financial tactic that can be accomplished through futures markets by making it possible to buy or sell a financial instrument that acts as insurance against an adverse movement of the underlying asset. Imagine a company that is heavily reliant upon sentiment analyses of Twitter data to assess the pulse of its customer base and drive future product decisions. If for some reason the Twitter data suddenly changes in nature so that it no longer provides accurate or timely information, the company can be protected if it has hedged its “long” position on Twitter with a futures contract that represents the short position.

Another impetus for this research is the growing need to value data as a corporate asset. Mayer-Schonberger and Kenneth (2013) describe the IPO of Facebook in 2012 as a prime example of this need. The company's reported traditional assets before the IPO was $6.3 billion. However, its initial stock price gave it a total valuation of $104 billion. This gap represents intangible assets, which for Facebook is predominately data. The difference between Facebook’s traditional and intangible assets explains, in part, why investors had such a troublesome time establishing its market value once trading began.

Assessing intangible value is not a new challenge for business. The use and valuation of patents has been undergoing its own transformation in recent years. Kevin G. Rivette and David Kline (2000) documented this revolution, describing the circumstances under which companies recognized untapped licensing revenues and developed valuation models for intellectual property. They reported how IBM went from receiving $30 million in patent licensing royalties in 1990 to a $1 billion per year in 2000, representing one-ninth of IBM's yearly pre-tax profits. This strategic importance of patents has placed patent valuation in the spotlight and emphasized the difficulty of the task. Existing valuation approaches for intangibles like patents and data include cost-based methods, which attempt to determine the expense of generating or replacing an asset, and market-based methods, which rely on previous market transactions of comparable assets (European IPR Helpdesk, 2013). Both methods are unsatisfying because they do not directly assess the value of the asset itself

and are subject to externalities such as market fluctuations. It is necessary to develop a model that more directly assesses the intrinsic value of data.

There is merit to a data market for researchers and government municipalities, in addition to pure corporate settings. A recent project conducted at the Institute for Transportation studies at the University of California – Berkeley with California PATH the went through a Request for Proposal (RFP) process to purchase existing traffic and transportation data from the commercial sector, rather than collecting data themselves. This RFP "was testing the waters of the probe data market and, as such, required a balance of scientific rigor and simplicity—specific enough to get the data we wanted but not so complex that it discouraged vendors from responding" (Bayen, Sharafsaleh, & Patire, 2013). At the time of the study and in our research, the idea of collaborating with commercial entities in such a way seems to be a novel idea. The results of this report show that there is a valid need for such a market and one does in fact exist with little standardization. Solidifying a method to assess data quality and pricing could allow for easier collaboration between commercial entities and researchers, with potentially great long term benefit.

Another potential use for such a market comes in addressing the use and sale of data (often Personally Identifiable Information) between parties. Ari Gesher of Palantir suggests that one method to reduce the transfer of data beyond its intended purpose is taxation (Big Data: Values and Governance, Panel 3, April 1 2014). He poses if a company wants to sell the data it has stored on its users to another 3rd party for a purpose the user may have not agreed to, the sale of that data should be taxed. In order for a tax to be assessed the data must have a proper valuation, which is where having a standardized pricing model and data market would come in to effect. The use of the market accompanied by taxation as Gesher proposes would improve transparency and privacy by incorporating an audit trail of data sale and add mechanisms for protection (as well as additional tax revenue for governments).

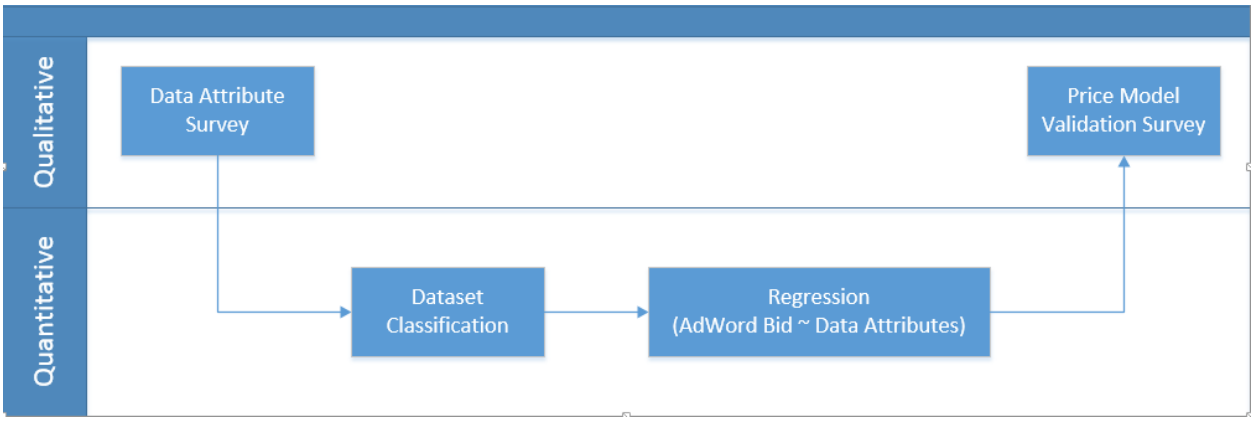
Devising a data market that addresses these challenges would prove valuable to companies currently involved in the buying and selling of data, and could fundamentally change business strategies for previously uninvolved groups. Consider, for example, non-profits that exist within the limits of narrow funding and are often restricted by the demands of funders. In exchange for funding, a non-profit might be asked by 50 funders for the same data in different formats (NTEN, 2012). Meanwhile, the organizations also collect metrics and external data to inform their programs and to make business decisions, exchanging data such as donor lists with other groups on an ad hoc basis (Poderis, 2008). Consider instead a business strategy where non-profits collect datasets needed to support their programs, using a rigorous data market valuation model to optimize the asking price of the asset after collection. The non-profit newsroom ProPublica announced a similar strategy in 2014 with its ProPublica Data Store, where data prepared and cleaned by the investigative journalists at ProPublica are sold at varying price tags to academics, journalists, and corporations (Ellis, 2014). A data market based on a rigorous pricing model would both assure non-profits an audience of buyers and an accurate depiction of a dataset's intrinsic value. Thus non-profits would be less reliant on the demands of various funders and could focus efforts again on the mission.

4 Research Design

The high-level design of our research includes a combination of qualitative and quantitative methods. A survey was conducted in order to determine how experts in the field view

various attributes of data. This data will be used to construct the machine learning regression models that will help establish a correlation between data attributes and the price of a given dataset. A final survey will be conducted to validate the results of the model against real world data valuation practices.

Figure 1: Overview of Research Process



4.1 Attribute Selection

There are a number of data characteristics that can affect the value of data. With the ultimate goal of identifying a model that can be used to price data in the open market, we examined how other digital assets are traded. This included the pricing strategy for digital media (audio, images, videos), licensing fees for intellectual property assets and patents, pricing variables used for software-as-a-service products, and techniques from software engineering for estimation and pricing. Based on this examination, we identified a set of candidate parameters, broken down into three main categories, which could help determine the value of data. These are:

- Value-based parameters (value of data to the consumer):
  - The value of the data in terms of saving in time, effort, or money
  - The ROI for the customer (or a profit share arrangement with the customer based on the profit derived from the acquired data)
  - Risk exposure – Data cleansed of personally identifiable information and privacy violations could be priced higher
  - Data exclusivity – Whether the data is provided on an exclusive basis, non-exclusive basis, or some combination of the two would influence price
  - Level of ownership – Is the customer buying (implying transfer of ownership), leasing (allowing use for a fixed time) or licensing (allowing limited use for a specific purpose)?
- Qualitative parameters (attributes or meta-attributes of the dataset):
  - Age of the data
  - Credibility of the data
  - Accuracy of the data elements
  - Quality of the data – missing fields for certain rows, incorrect types, data precision, etc.
  - Format and level of structure of the data – plain text, streaming data, tabular datasets, etc.
- Fixed and marginal cost parameters (directly measurable cost):

- Cost of collecting the data
- Cost of data storage, bandwidth, and other operational costs
- Cost of data-as-a-service offerings – add-on services to process the data, computing resources for the data, analytic reports, or aggregation on the data
- Delivery cadence – one-time, batch, or continuous basis

The operational value of data, which is the cost of producing the data by the seller, can be easily determined from the marginal cost of generating, storing, and sharing the data. This is the minimum price that a seller can command to cover the total cost of generating and delivering the data. Currently, the market value of data is mostly determined through value-based parameters, which are difficult to quantify and model. While it is possible to use the value-based parameters to command a premium price for the data, it will become necessary to move to a set of parameters that can be measured and modeled.

4.2 Qualitative Assessment

In order to gauge the importance of various qualitative parameters that we identified above, we created a short survey and distributed it across several data-oriented groups on LinkedIn. We also reached out to a number of our colleagues involved in the data science community. We collected responses from 12 individuals over the course of a 7-day span.

The first two questions asked the respondents to put themselves in the shoes of a data buyer and then a data seller, and assign a level of importance ("Very Important", "Somewhat Important", "Does not matter", "Don't know / Can't say") to a number of predefined data characteristics. We also asked questions about additional services expected with the purchase of data, level of ownership, and if the respondents had any other characteristics in mind that would add or take away from the value of a dataset. Finally, we asked the respondents to optionally identify themselves with their current role or job title and by indicating whether they had ever been involved in buying or selling data. A summary of the results of the survey is provided in the appendix.

A paired T-test comparing mean ratings showed that "Age of data" was the only characteristic rated higher by buyers than sellers. Given that this is a statistically significant difference ($p = 0.005$), it could imply that buyers overestimate the importance of the age of data, or that sellers underestimate how important data freshness is to buyers. Two other characteristics ("cost of collecting data" and "exclusivity of access to data") have significantly higher mean ratings for sellers compared to buyers ($p<0.05$). "Delivery cadence" also shows a marginally significant difference between buyers and sellers ($p<0.1$). The other characteristics do not have significant differences in their mean rating.

The fact that "cost of collecting data" has a significant impact for sellers, but not for buyers, causes concern for rising information asymmetry in the data market. Generally, sellers have more knowledge about their process of collecting, cleansing, and packaging data, while buyers have little or no understanding or appreciation of these processes. Furthermore, with the increasing availability of large-scale datasets, data providers may be encouraged to pass off subpar data as higher quality data, and buyers may have little opportunity to fully assess the data for quality and lack the mechanisms to avoid and expose those sellers. The data market is prone to fall prey to a "market for lemons" scenario, which further legitimizes the need for a standardized valuation approach.

While this preliminary survey helps to shed light on how participants in the data market value the different aspects of data, more robust qualitative research, that includes interviews and a larger yet more targeted survey, is necessary to gain substantive knowledge. This activity is underway, and the results will be accessible at the following URL when complete:

http://groups.ischool.berkeley.edu/datamarkets/

4.3 Proposed Model

Creating a universal model for all data types would be a monumental task, and data sources may require different pricing models both based on the type of data and the potential avenues for which it is used. One potential approach includes using a classification algorithm on the attributes of the data itself prior to applying a pricing model. This could be done with various clustering or supervised learning techniques. Once a dataset has been classified, the appropriate pricing model could then be applied.

The development of the model itself would require further exploration of the objective, independent variables that could have a relationship on data value, some of which have been outlined already. Additionally, an appropriate number of sample datasets with their prices and attributes would have to be collected as inputs to the model, ideally ranging from large to small datasets, spanning multiple industries and utilizations. A general linear model is demonstrated below, but other options may exist.

$$\begin{aligned} &Estimated\ Value\ of\ Data \\ &\quad = Fixed\ Cost + \beta_1 * Age\ of\ Data + \beta_2 * Periodicity\ of\ Data + \beta_3 \\ &\quad * Volume\ of\ data + \beta_4 * Accuracy\ of\ Data + \cdots \end{aligned}$$

Estimating the value of data to build out this model is a difficult task; many datasets may not have public prices associated with them, and even when available, the current valuations are likely subjective. One available proxy for value could be the Google AdWords suggested bid prices for specific query terms. In this scenario, the value of datasets would be tied directly to specific search topics that have an automated price set by the market using a sealed bid auction system. Google provides advertisers with average monthly searches, competition rating, and suggested bid price for the potential search terms an ad can be associated with. These values provide a guide to public demand and market price for a search term (WordStream, 2014). If a search term could be identified that is highly-related to a specific set of data and minimally associated with anything else, then it could be used as a proxy for the dataset in assessing its value. Figure 2 shows how Google AdWords assigns values to search terms related to 2010 United States Census data.

Figure 2: Example AdWord output

Keyword (by relevance)	Avg. monthly searches	Competition	Suggested bid	Ad impr. share	Add to plan
census data 2010	390	Low	$1.10	0%	»
census 2010 data	320	Low	$1.45	0%	»
2010 census data	2,900	Low	$1.82	0%	»
2010 census data by zip code	210	Low	$2.76	0%	»

This data can then be used as the expected predicted values to train a machine learning regression model that determines how relevant attributes of the data contribute to that value. It would be challenging to choose an appropriate dataset that meets the experiment parameters and additional tests would be necessary to ensure target keywords are strongly correlated with that dataset and only that dataset. It also requires subsequent studies to be designed and conducted if a model is desired that actually estimates the dataset's value, as this model would estimate the proxy value as determined by Google's suggested bid for the associated search term.

Once the appropriate classification system and valuation model are created, it would be imperative to buttress the external validity of the model by surveying people or organizations that take part in the data marketplace as it exists today to check the relationships of the variables and the prediction of the model. Data buyers and data owners would be the primary players. Data owners would include both those that inherently use and sell data already, but also less obvious organizations that produce and use data internally without intention of distribution. Other potential contributors include third party data brokers, data platform developers, and investors.

5 Conclusion

A dynamic, standardized pricing model would revolutionize the existing data market, facilitating transparent transactions and making data science more efficient. Because the value of a dataset is dependent on numerous variables, such as the age and the quality of the data, the model will take time to develop, test, and train. However, if successful, this model will be extremely beneficial to any group that produces or uses data.

Ultimately, assuming a grand pricing model can be constructed, a futures market for data could be established. Subsequently, a new class of financial analysts and market speculators specializing in data would help guide the market to produce the appropriate data by sending market signals about their understanding of the future valuation of data. This would result in greater competition, allowing innovative companies to grow faster and more efficiently due to higher availability, quantity, and quality of their needed data.

The potential for an efficient data market with an appropriate valuation model is nascent and exciting. The amount of data in our world is growing exponentially, as is the number of people and organizations seeking to understand and build upon the value of data. This hypothesized pricing model and initial research is a rudimentary step toward a full-fledged market structure and data valuation algorithm. But ultimately, in the data-driven world that we live in today, a robust data valuation model would help make data transactions more nimble and transparent, resulting in enormous benefit to all involved.

6 References

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics* , 488-500.

Anderson, C. (2008, June 23). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.* Retrieved from Wired Magazine: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Bayen, A. M., Sharafsaleh, M. A., & Patire, A. D. (2013). *Hybrid Traffic Data Collection Roadmap: Pilot Procurement of Third-Party Traffic Data.* Berkeley, CA: Partners for Advanced Transportation Technology.

Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy* , 637-654.

Ellis, J. (2014, March 4). *ProPublica opens up shop with a new site to sell custom datasets*. Retrieved from Nieman Labs: http://www.niemanlab.org/2014/03/propublica-opens-up-shop-with-a-new-site-to-sell-custom-datasets/

European IPR Helpdesk. (2013, June). *Fact Sheet: Intellectual Property Valuation.* Retrieved from IPR Helpdesk: http://www.iprhelpdesk.eu/sites/default/files/newsdocuments/Fact_Sheet_IP_Valuation_0.pdf

Leger, D. L. (2014, September 4). *How stolen credit cards fenced on the Dark Web.* Retrieved from USA Today: http://www.cnbc.com/id/101971708

Mayer-Schonberger, V., & Kenneth, C. (2013). *Big Data: A Revolution that will Transform How We Live, Work, and Think.* New York: Houghton Mifflin Haracourt.

Moody, D., & Walsh, P. (1999). Measuring The Value Of Information: An Asset Valuation Approach. *Seventh European Conference on Information Systems* (pp. http://wwwinfo.deis.unical.it/zumpano/2004-2005/PSI/lezione2/ValueOfInformation.pdf). Frederiksberg, Dennmark: Copenhagen Business School.

NTEN. (2012). *The State of Nonprofit Data.* Portland, OR: NTEN.

Poderis, T. (2008). *Donor List Selling or Exchange with Other Non-Profit Organizations*. Retrieved from Raise-Funds.com: http://www.raise-funds.com/2008/donor-list-selling-or-exchange-with-other-non-profit-organizations/

Rivette, K. G., & Kline, D. (2000). *Rembrandt's in the Attic.* Boston: Harvard Business School Press.

Sachdev, A. (2013, July 7). New exchange is formed for trading patent rights. *Chicago Tribune* , pp. http://articles.chicagotribune.com/2013-07-07/business/ct-biz-0707-patent-exchange-20130707_1_patent-system-patent-violations-james-malackowski.

Wikipedia. (2014, May 1). *Black-Scholes Model*. Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Black%E2%80%93Scholes_model

Wikipedia. (2014, June 15). *Patent Valuation*. Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Patent_valuation

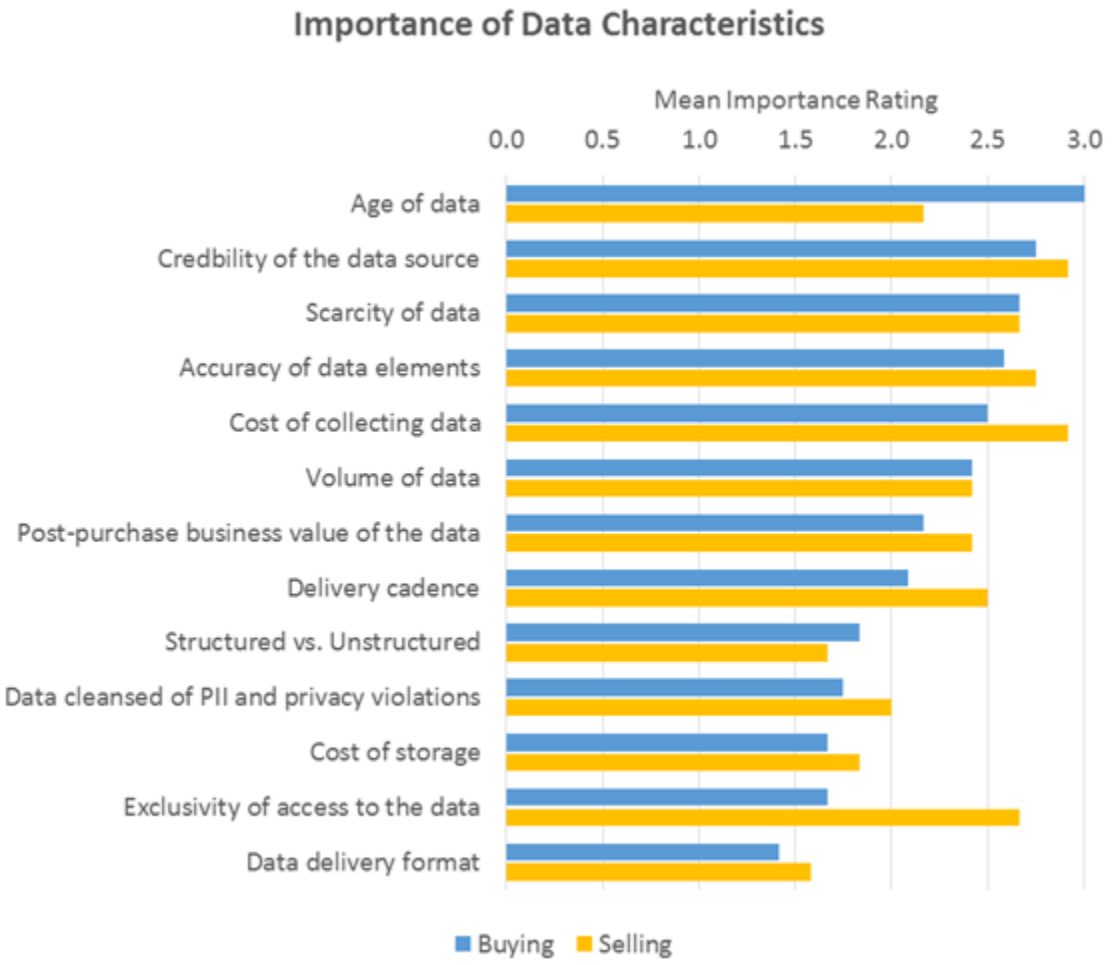
WordStream. (2014, September 4). *What Is Google AdWords? How the AdWords Auction Works*. Retrieved from Wordstream: http://www.wordstream.com/articles/what-is-google-adwords

7 Table of Figures

Figure 1: Overview of Research Process 3

Figure 2: Example AdWord Output 6

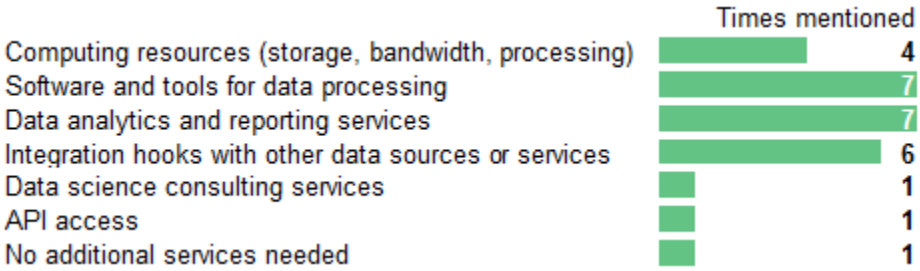
8 Appendix



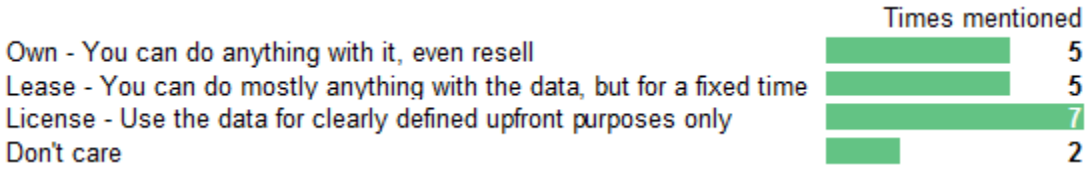
Mean Importance Rating in the chart above is based on the following assigned values:

Very Important	3
Somewhat Important	2
Does not matter	1
Don't know/Can't say	0

What additional services would you expect with your data purchase for additional cost? (Allowed to choose multiple options)



What level of ownership would you require for the datasets that you buy? (Allowed to choose multiple options)



What else would add to (or take away from) the value of a dataset in your opinion?

- "Knowing where the original data was sourced from, compiled."
- "Data Quality Metrics… for instance most data that's published in real time has a lot of noise with it. But if the vendor were to give me a sense of what part is noise vs. what is trustworthy based on the metrics they have collected historically, I would then use those parameters in my judgment. In most Financial Institutions, most users of data know when to trust a Bloomberg feed vs a Reuters feed and for what type of data."
- "The amount of time needed from purchase to delivery of dataset. How long does it take to collect the data and is it still of value by the time it is ready for delivery."
- "Good/complete metadata / data dictionary. Consistent format. Missing data / censoring. Aggregated too much. For international data, encoding/translation/transliteration can be a pain and a source of errors."
- "Related to volume, but very different in some cases, is coverage. Often you're buying data to line it up with something you already have, so I don't care if they have 100MM records, I care more about how many records match those already in my database."

Have you ever been involved in selling or buying data?

- "Yes, I used to work for a job board so we have sourced resume data for our database."
- "Yes. Government procurement, and OEM pricing data is 'public knowledge' in the aviation industry, but is difficult to gather for all the many variations of aircraft part numbers. We found a source and resell the data, yet our competitors are able to obtain more records than us, causing frustration to our customers, and lowers the value of our service."
- "Yes, collected stock market data on new issues of corporate securities, municipal issues, and mergers and acquisitions that was sold monthly to Wall Street investment banks.... Also collected data on municipal issues for major debt rating age