Informatics for Real Estate: Urban Technology Databases

by

Ryan Michael Stroud

B.S., Architectural Engineering with Honors, 2010

California Polytechnic State University, San Luis Obispo

Submitted to the Program in Real Estate Development in Conjunction with the Center for Real Estate in Partial Fulfillment of the Requirements for the Degree of Master of Science in Real Estate Development

at the

Massachusetts Institute of Technology

September, 2017

©2017 Ryan Michael Stroud. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature of Author ________________

Certified by ________________

Accepted by ________________

Signature redacted

Center for Real Estate

July 28, 2017

Dr. Andrea M. Chegut
Research Scientist, Center for Real Estate
Thesis Supervisor

Signature redacted

Professor Albert Saiz
Daniel Rose Associate Professor of Urban Economics and Real Estate,
Department of Urban Studies and Center for Real Estate

ARCHIVES

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEP 13 2017

LIBRARIES
Informatics for Real Estate: Urban Technology Databases

by

Ryan Michael Stroud

Submitted to the Program in Real Estate Development in Conjunction with the Center for Real Estate on July 28, 2017 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Real Estate Development

ABSTRACT

Big Data Analytics is a term that represents an entire spectrum of analytical applications utilizing significant quantities of data, ranging from optimization at one end of the spectrum, to gaining new insights at the other end of the spectrum. This thesis focuses on the latter, leveraging private, public, and manually developed databases at the MIT School of Architecture and Planning’s Center for Real Estate’s Real Estate Innovation Lab (REIL) to observe, dissect, and ultimately improve our collective understanding of the current state of urban technology databases. The thesis seeks to explore how companies are providing data within the realm of the built environment, through a study of the information products that they offer.

To preserve the confidentiality of the original commercial databases and limit the scope of the investigation, the dataset for this study contains only the data fields from 31 unique databases provided by 14 commercial real estate data aggregators. In essence, the dataset for this thesis is a database of databases, stripped of their numerical information and focused on a study of the variation in data.

For analysis this employs computational, statistical, and graphical methods to interpret the information provided by the commercial real estate data aggregators. With an increasingly digital future ahead, this thesis proposes a general framework for examining numerous databases and their respective approaches to the built environment. This thesis also explores the merits of specific processes and presentation methods that translate an immense and disparate array of information into user-friendly analytical tools.

Thesis Supervisor: Dr. Andrea M. Chegut
Title: Research Scientist, Center for Real Estate
Acknowledgements

This thesis, and likewise this degree, would not have been possible without the significant contributions of many incredible individuals that I’m lucky to have in my life.

To my beautiful wife, Kerry, thank you for your unwavering support and pushing me to pursue my goals. My gratitude for you extends beyond the number of words that I could put on a page.

To my parents, Dave and Anne, thank you for making countless sacrifices so that I could always pursue my education, by any means necessary. Thank you for the unbelievable amount of guidance and support you’ve shown me over the years, teaching me to always be inquisitive, challenging me to make a meaningful contribution to our world while I’m here, and leading by example.

To my parents-in-law, Bob and Kathy, thank you for being an extraordinary exception to the idea that you shouldn’t get along with your in-laws. Thank you for your unending kindness and support, from helping us cross continents to offering us a roof when we didn’t have one.

To my sister, Stephanie, thank you for always being the bright place in my life, no matter the physical distance. And to Justin and Parker, thank you for always reminding me how to laugh and enjoy.

To my most loyal companion, Phoebe, thank you for helping me take life a little less seriously.

To all my teachers, professors, mentors, coaches, critics, and friends, thank you for making such a positive impact on my life and challenging me to go the distance.
Table of Contents

Section 1 – Introduction ........................................................................................................... 5

Section 2 – Literature Review.................................................................................................. 7
  The Rise of Big Data ................................................................................................................. 9
  Data Visualization ..................................................................................................................... 14
  Classification ............................................................................................................................ 17

Section 3 – Data and Methodology ......................................................................................... 19
  Methodology ............................................................................................................................ 23

Section 4 – Results .................................................................................................................... 32
  Word Cloud .............................................................................................................................. 33
  Bar Graph ................................................................................................................................. 35
  Donut Chart .............................................................................................................................. 37
  Scatter Plot ............................................................................................................................... 39
  Radar Chart ............................................................................................................................. 40
  Network Diagram .................................................................................................................... 48
  Chord Diagram ........................................................................................................................ 53

Section 5 – Conclusions ........................................................................................................... 60

Works Cited .............................................................................................................................. 63
Section 1 – Introduction

One of the current characteristics of data science with respect to commercial real estate and urban technology databases is the lack of a well-established scientific process, let alone consensus on the definition of what constitutes “Big Data.” Academic papers and publications on the subject of data science with respect to commercial real estate have become increasingly frequently within the past decade, though the content of many remains relatively introductory. Additionally, those publications that have sought additional depth, and attempted a rigorous analysis pertaining to data science and commercial real estate, have been relatively limited in breadth. This is not a criticism on the body of work, simply an observation that this is a nascent field of abundant opportunity.

This thesis aims to contribute to the field of data science with respect to commercial real estate by conducting one of the first large-scale comparative studies of commercial real estate and urban technology databases. Within the thesis, the author explores the following questions: what can we know about a property, building, or asset? What types of knowledge are available, and where is that knowledge coming from? Entertaining the metaphor that the pursuit of knowledge is an expedition, this thesis intends to equip fellow explorers by taking inventory of our available supplies (data), and uniting various fragments of information from the world’s leading cartographers (urban technology databases) to chart the known world.

To identify the current distribution of data, the author employed computational, statistical, and graphical methods to interpret data from commercial real estate information providers. Sources of data included CrediFi, CompStak, Green Building Information Gateway (GBIG), Gensler, GeoTel, JLL, NCREIF, New York City Department of Finance, Real Capital Analytics (RCA), REmeter, S&P Global Platts, WiredRE, and WiredScore, as well as privately curated data by MIT’s Real Estate Innovation Lab.

We begin the study in Section 2 with a review of literature surrounding the pertinent fields of data science, complexity and networking theory, taxonomy, and informational visualization to generate the conceptual framework around which the study is performed. In Section 3, we then summarize the process for creating a usable amalgamation of databases that was critical to producing easily-interpreted results,
followed by a consolidated summary of results. In Section 4, the results of this thesis demonstrate that an incredible amount of information about the built environment is being gathered, recorded, monitored, analyzed and distributed. Commercial real estate data aggregators are pushing the limits of what we can quantify about a property, building, or asset.

Finally, in Section 5 the conclusion seeks to interpret the results obtained from the study, and speak to the implications for our field in light of this study. The inception of data science has been incredibly interdisciplinary, and the future of the field is likely to be just as collaborative, as those with deep technical knowledge yet diverse backgrounds join forces to perform meaningful analyses with Big Data. The commercial real estate industry is moving from an era of analogue analysis to an era of highly informed, highly quantified, digital analysis. What could previously only be put into words about a property, building, or asset can now be quantified and analyzed in increasing levels of detail; the role and importance of data to the commercial real estate industry is only likely to increase.
Section 2 – Literature Review

Compounding interest is a concept commonly associated with the field of finance, yet with the exponential growth of digital technology throughout the 20th century, compounding interest seems to also describe the state of enthusiasm surrounding increasingly technical disciplines. Hal Varian, currently Google’s chief economist and an emeritus professor at UC Berkeley, has famously emphasized, “I keep saying the sexy job in the next ten years will be statisticians. People think I’m joking, but who would’ve guessed that computer engineers would’ve been the sexy job of the 1990’s? The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that’s going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it” (McKinsey).

Big Data forms the foundation for advanced analytics, deep learning, artificial intelligence, automation, and each of their respective economic opportunities. Adam Berenzweig, co-founder and CTO of Clarifai (a startup focused on automated image recognition), stated at a Stanford Graduate School of Business panel, “we can't really talk about deep learning without talking about data. In some sense, deep learning is nothing more than what happened when machine learning hit big data” (Jurvetson, "Deep Learning: Intelligence from Big Data"). The wholesale aggregation of data, paired with the computational and analytical techniques honed over thousands of years (via mankind’s integration of myriad scientific disciplines), has unlocked opportunities that were previously unimaginable. Erik Brynjolfsson and Andrew McAfee, Director and Co-Director of the MIT Initiative on the Digital Economy at the MIT Sloan School of Management, hypothesized in their 2014 New York Times best-seller The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies that digital technologies are spurring an economic revolution on par with the Industrial Revolution. Brynjolfsson and McAfee wrote, “one of the main reasons we cite digitization as a main force shaping the second machine age is that digitization
increases understanding. It does this by making huge amounts of data readily accessible, and data are the lifeblood of science” (Brynjolfsson and McAfee, 67).

The Big Data revolution within the field of commercial real estate was recently detailed in the March/April 2017 issue of REIT magazine. “A revolution is coming in real estate investment, according to MIT professor David Geltner. ‘Rules of thumb’ and ‘conventional wisdom’ are out, he says. Empirical data and analytics are in.” According to Geltner, “This is an industry that has always needed to use a lot of numbers. But it is an industry that never had real data. This could change now because we do now have much more and better data, and getting more and better all the time. And we now have computational power that can make use of this data. There is the potential for a major cultural shift in how the real estate investment industry does business” (NAREIT).

In an ongoing research project initiated in 2017 at MIT’s Real Estate Innovation Lab, researchers identified over 1,644 technology startups disrupting the field of commercial real estate. Within these 1,644 technology startups, approximately 688 of these are related to data services and/or data analytics (based on their company descriptions), and 313 of these have been reviewed and verified by MIT researchers at the time of writing as data-driven startups in the sector. The impressive quantity of data-driven startups related to commercial real estate is one of many indicators that the future of our industry will be increasingly digital. As JLL stated regarding the importance of Big Data Analytics, and future technologists and data scientists in the field of commercial real estate, “it’s not just about . . . raw data on productivity, energy consumption and building efficiencies, it’s also about mining the information to get the most out of it. As such, the commercial real estate sector is now going head to head with the tech sector to compete for the best people” (JLL Staff).
The Rise of Big Data

Big Data – the term has become ubiquitous in contemporary society, from business to academia to advertisements, much akin to the proliferation of electronic devices and technology that have permeated nearly every aspect of our daily lives. The synchronous rise of Big Data and the proliferation of electronic devices is not merely coincidence, as the increasing number of electronic devices throughout society record and transmit ever increasing amounts of data. With continual advancements in technology, the ability to store and utilize this data has increased exponentially. Gordon Moore, co-founder and chairman emeritus of Intel Corporation, made a famous prediction in a 1965 article that, “the complexity for minimum component costs has increased at a rate of roughly a factor of two per year . . . Certainly over the short term this rate can be expected to continue, if not to increase” (Moore 115). Moore’s predictions regarding the amount of computing power you could purchase for one dollar has held up “astonishingly well for over four decades . . . and today it’s common to use eighteen months as the doubling period for general computing power” (Brynjolfsson and McAfee, 41). According to experts in the field, the amount of data processed between 2012 and 2014 surpasses the estimate of all cumulative data processing conducted throughout the prior 3,000 years (Smolan).

The aggregation of data forms the basis for analytics that provide meaningful insights to numerous fields of study and practical applications. Although technological advancements have increased the complexity of data analytics in recent years, humans have been deriving insights from recorded observations (data) for millennia. An article published by the World Economic Forum cites a Neolithic Era artifact known as the Ishango bone as “one of the earliest pieces of prehistoric data storage” (Marr), notwithstanding the ongoing discussion between scientists regarding the exact utilization of the tally marks engraved in this artifact since its excavation in 1960 (de Heinzelin 106). While some who studied the Ishango bone interpreted the tally marks as a numerical system (Gillings 138), and others interpreted the tally marks as the representation of a lunar calendar (Marshack), this artifact stands as a significant milestone in the general development of methodical observations.
One of the most notable milestones in the history of Big Data came in 1890 with the United States census. Herman Hollerith, the son of German immigrants in New York state, graduated from the Columbia University School of Mines in 1879. Hollerith was apparently “not interested in ‘practical mining work,’ according to his children” but more interested in the technical engineering curriculum (Blodgett and Schultz 221). Ironically, Hollerith would go down in history for his breakthrough technological achievements in mining – not necessarily for minerals, but data. Hollerith’s abilities in the classroom at Columbia impressed his professor, William Petit Trowbridge, who asked Hollerith to become his assistant. Together they worked as statisticians for the 1880 census, and Hollerith noted the potential for improvements in the methods of conducting the census (Census History Staff). Taking inspiration from automated weaving machines, type distributing machines, and his conversations with Dr. John Shaw Billings (the Director of Vital Statistics for the 1880 census), Hollerith left Washington for the Massachusetts Institute of Technology in 1882 as an instructor in mechanical engineering, where he spent a year “developing his ideas and early equipment for recording and tabulating census data” (Blodgett and Schultz 222).

On September 23, 1884, Herman Hollerith filed for United States Patent 395,782, entitled the “Art of Compiling Statistics.” In the 1887 revision of this patent, Hollerith describes “the art of compiling statistics which consists of first preparing a series of separate record cards, each card representing an individual or subject” (Hollerith, Art of Compiling Statistics). This was soon accompanied by United States Patent 395,783, entitled the “Apparatus for Compiling Statistics.” He would use his prototypes to conduct and tabulate mortality statistics in Maryland, New Jersey, and New York from 1887 to 1889 (Blodgett and Schultz 222). Hollerith’s patented design utilized a punch card system, with various holes in precise locations on the punch cards representing unique statistics, which could then be sorted and read electronically by a press containing pins that would register the perforations from the punch cards (see fig. 1 and fig. 2). Hollerith’s electric tabulating machines demonstrated dramatic increases in efficiency, with improved statistics and lower processing costs.
Hollerith’s success resonated with the Census Office, and Hollerith won the contract for the 1890 United States census, in addition to subsequent censuses for multiple nations. Where the 1880 United States census had required approximately eight years to complete, the time required to tabulate the 1890 United States census was reduced to only one year. Railroad companies also took note, and licensed the Hollerith machines for calculating fare information. Hollerith founded the Tabulating Machine Company in 1896, and was joined by Thomas J. Watson in 1918 in the company that would eventually change its name to the International Business Machines Corporation in 1924 (Census History Staff).
Figure 2. Diagram of an Apparatus for Compiling Statistics from Herman Hollerith, “Art of Compiling Statistics;”
Data Visualization

Manuel Lima, an accomplished thought leader on information visualization, traces the origins and evolution of information visualization through time in his inquisitive book, *Visual Complexity: Mapping Patterns of Information*. In his aforementioned work, Lima emphasizes the interconnected nature of classification and taxonomy to the field of information visualization. A prime example of this important connection is the tree diagram, with its roots reaching back to Porphyry of Greece (234 - ca. 305 CE) who built upon the foundational work of Aristotle (384 - 322 BCE). Porphyry’s publication *Isagoge* reframed the classification system introduced by Aristotle’s *Categories* into “a hierarchical, finite structure of classification, in what becomes known as the tree of Porphyry” (Lima 27). The tree of Porphyry introduced the concept of visualizing Aristotle’s *Categories* in a two-dimensional graphic, with references to a tree’s trunk. Porphyry’s graphic concept was “frequently represented in medieval and Renaissance works on logic and set the stage for theological and philosophical developments by scholars throughout the ages. It was also, as far as we know, the earliest metaphorical tree of knowledge” (Lima 28).

Ramon Llull (1232 – 1315), an original renaissance man born on Mallorca just before the dawn of the European Renaissance (which is commonly accepted as beginning in the 14th century), produced some of the most well-known early works on illustrative knowledge representation. While a majority of medieval tree diagrams had focused on genealogy and religious history, Llull’s 1296 publication * Arbor scientiae* illustrates a beautiful collection of 16 trees representing the various scientific domains, introduced by a primary tree [see fig. 3] called the arbor scientiae (Lima 31). Translated as the “tree of science,” the illustrations eloquently make the connection between scientific classification systems and information visualization.
While Ramon Llull’s trees of science greatly influenced historic philosophers including Francis Bacon and René Descartes, Manuel Lima asserts that Llull’s “most recognized contribution to European thinking was the pursuit of an ‘organic and unitary corpus of knowledge and a systematic classification of reality,’” which included a series of diagrams, symbolic notations, and mechanical apparatuses” that later influenced computational theory and Gottfried Leibniz (Lima 33). These diagrams, symbolic notations, and mechanical apparatuses that Lima refers to are detailed in Llull’s work, *Ars magna, generalis et ultima* (the “Ultimate General Art”), published in 1305 [see fig. 4]. Prime examples include the Prima Figura which traces all the possible combinations between the dignities (the nine absolute principles – or divine dignities – which communicate their nature to each other and spread throughout creation), and the Secunda Figura which serves to connect the relative principles with triples of definitions (Eco 58).
These intricate diagrams, symbolic notations, and mechanical apparatuses detailed in Llull’s work, *Ars magna, generalis et ultima*, also exemplify the gradual shift from the top-down, centralized organizational structures (such as the hierarchical tree diagrams) to alternate models with the capacity to represent increased interconnection and complexity. This shift to increasingly complex diagrammatic structures echoed throughout the European Renaissance, as innovators and philosophers pioneered new layouts to index and catalog the expanding encyclopedia of human knowledge.
Manuel Lima artfully details society’s ever-increasing need for the ability to describe compounding complexity in the following excerpt from *Visual Complexity: Mapping Patterns of Information*:

The complex connectedness of modern times requires new tools of analysis and exploration, but above all, it demands a new way of thinking. It demands a pluralistic understanding of the world that is able to envision the wider structural plan and at the same time examine the intricate mesh of connections among its smallest elements. It ultimately calls for a holistic systems approach; it calls for network thinking. (45)

**Classification**

Matching the gradual shift in visualizations to accommodate compounding complexity, the field of classification evolved organically from hierarchical structures of taxonomy to social, collaborative indexing. The term *folksonomy* was coined by Thomas Vander Wal in 2004 (Lima 62), an information architect also credited with initiating the term *infocloud* in 2004 (Vander Wal). Folksonomy is a linguistic blend of the words folk and taxonomy, representing the concept of social classification and social tagging that arose (circa the early 2000’s) from the open-source nature of websites that inherently promoted user collaboration. As opposed to the top-down, rigid format of taxonomy (a word with Greek roots indicating order, arrangement, law, and science), folksonomy describes a bottom-up, democratic process where objects/data are assigned a variety of tags by the user community that can be digitally searched, sorted, and grouped based on the tags applied.

To digital natives, “those generations that grow up with this new technology … [that] spent their entire lives surrounded by and using computers, videogames, digital music players, video cams, cell phones, and all the other toys and tools of the digital age” (Prensky), the concept of folksonomy and collaborative indexing is relatively apparent and straightforward given the capabilities of modern personal computers and the internet. To those digital natives born during/after the 1980’s who matured alongside the maturation of the personal computer, the concept of informally tagging digital media (e.g., digital music tracks in iTunes)
for rapidly sorting and filtering their digital collections seems quite obvious, and having to adhere to an inflexible (standardized) system of indexing could seem unnecessarily oppressive and archaic.

Yet to digital immigrants, “those of us who were not born into the digital world but have, at some later point in our lives, become fascinated by and adopted many or most aspects of the new technology” (Prensky), the utility of a standardized system of rigid taxonomy seems quite obvious, coming from an environment where rapid digital sorting and filtering was simply not available. It may be difficult for some digital natives to fathom, but there was an era before the existence of any “large-scale hypertextual web search engine” (Brin and Page); prior to the proliferation of personal computers, rigid taxonomy was a necessary mechanism for physically locating items within any form of archive, without the modern ability to instantaneously summon information at will.
Section 3 – Data and Methodology

The dataset utilized for this thesis is a derivation of numerous commercial databases entrusted to MIT’s Real Estate Innovation Lab (REIL), a research lab for the built environment within the MIT School of Architecture and Planning and Center for Real Estate that links innovation, design, and economic impact. In order to preserve the confidentiality of the original commercial databases and limit the scope of the thesis investigation, the dataset utilized for this thesis contains only the data fields from 31 unique databases provided by 14 commercial real estate data aggregators. In essence, the dataset utilized for this thesis is the meta-data (the database of databases), completely stripped of their numerical information and focused exclusively on the variation in data fields.

Studying solely the data fields provided by the 14 commercial real estate data aggregators offers the added benefit of eliminating sources of error from the parent databases, sidestepping the discussion surrounding the accuracy of each data point and its respective method of acquisition, while preserving the overall goal of comparing the information products on offer. In total, there are 2,568 data fields from the 31 commercial real estate databases studied, and 1,743 unique data fields after consolidating (and documenting the) duplicates. These duplicates refer to those data fields with a direct character-for-character title match, yet with such a large number of data fields there are many with similar meanings (despite the lack of a direct character-for-character title match) that could potentially be further consolidated. As such, the author embarked on an exhaustive process of manually cleaning the data set, applying a standardized system of Category Tags to each of the 2,568 data fields from the 31 commercial real estate databases.

To execute a meaningful study on the data fields from the 31 commercial real estate databases, and explore any connections between the 14 commercial real estate data aggregators, the author first conducted a thorough process of manually cleaning the data set. All of the 2,568 data fields were consolidated into a single spreadsheet titled “Master Sort” with one column and 2,569 rows (including a single row for titles). The first column was titled “Name I (Original)”. A second column was started to the right, titled “Category Tag”. Leveraging the Data Dictionaries that were provided by a few of the 14 commercial real estate data
aggregators, as well as some inference for fields that the Data Dictionaries did not necessarily cover, the author applied a standardized system of Category Tags to each of the original 2,568 data fields.

In total, 903 unique Category Tags were generated to describe the original 2,568 data fields. Undoubtedly, 903 Category Tags is still a large number of categories (representing 35.2% of the original quantity of data fields), but there are reasons for not oversimplifying the categorization in the first pass through the data fields. Foremost, the 31 commercial real estate databases all pertain to the field of real property, yet they represent information generated for quite disparate fields and customers. Some of the 31 commercial real estate databases are developed for (perhaps the most common assumption) the field of real estate brokers, buyers, sellers, lessors, and investors. Further, some of the 31 commercial real estate databases are generated for parties involved in the development and construction of real property, as well as urban planning and government. Additionally, some of the 31 commercial real estate databases are geared towards the corporate users, tenants, and insurers of real property. On the other hand, some of the 31 commercial real estate databases are developed for specialty niches such as telecommunications, data centers, as well as power generation and transmission organizations. Consequently, there are a handful of Category Tags that were frequently utilized (such as Address, Area, and ID Number) and a vast quantity of infrequent Category Tags that are exceedingly specific to the target field and customer (such as Air Flow at 100% Load, Column Spacing, Cost of Ponds, Easements, Fly Ash Reinjection, Green Measures, Hotel ADR, Job Count, Loan DSCR, Mercury Emission Control, Sold in Crisis, Trophy Building, Type of Foundation, and Walkscore).

The histogram charting the frequency of Category Tags illustrates a distribution with a very long tail of unique, single utilization categories (see fig. 5). While some may argue that 903 Category Tags does not represent sufficient consolidation of the 2,568 data fields, the author maintains that the unique, single utilization categories represent immense utility to their target users. Oversimplifying the categorization process in the first pass through the data fields would eliminate important and unique information, and
misrepresent the diversity of the 31 commercial real estate databases, as information diversity would be as significant a finding as commonality.

Figure 5. Histogram, with the 903 unique Category Tags on the x-axis, and frequency on the y-axis. Source: Author.

Once the Category Tags had been established in the “Master Sort” sheet, the original 2,568 data fields from the sheet titled “Raw Data” were separated into 14 different sheets, separated by their source companies (the 14 commercial real estate data aggregators). The 14 different sheets were titled “DB Group 1” to “DB Group 14” representing their respective source companies. To further clarify this step, the original data fields from the 31 commercial real estate databases (on the sheet titled “Raw Data”) were copied from each of their original horizontal rows, and pasted (transposed to a vertical column) into 14 separate sheets. This way, data from each of the 14 source companies (the 14 commercial real estate data aggregators) were isolated in their own separate sheets.

As there were 31 commercial real estate databases from 14 source companies, some of the 14 separate “DB Group” sheets included data fields from multiple databases. In these cases, where one source company provided multiple databases, the data fields from each of their databases were consolidated into one column. Consistently, this first column (in each of the 14 separate “DB Group” sheets) was titled “Name 1 (Original).” This consolidation methodology was utilized to meet the goal of comparing the information products on offer from the 14 source companies (the 14 commercial real estate data aggregators). While it
would also be academically interesting to study each of the 31 commercial real estate databases independently, the primary objective was to explore how companies are providing data within the realm of the built environment through a study of the information products that they offer.

After the data fields from each of the 14 source companies were isolated in their own separate sheets, a VLOOKUP function in Excel was used to match the original data fields to their respective Category Tags. To further clarify this step, for each of the “DB Group 1” through “DB Group 14” sheets, the data fields in the “Name 1 (Original)” columns were matched to their new Category Tags by using VLOOKUP to reference the sheet titled “Master Sort” (with all 2,568 data fields and their 903 matching Category Tags). An example of the formula utilized is as follows (example for cell B30): 

`=VLOOKUP(A30,'Master-Sort'!$A$2:$B$2569, 2, FALSE)`. This essentially indexed the data fields to the master list of Category Tags (see fig. 6).

![Figure 6](image-url)  
Figure 6. Excerpt from the sheet titled “DB Group 1” representing the data separated by the 14 source companies. Source: Author.
Methodology

Once the data fields had been matched to their Category Tags, the entire column with the results titled “VLOOKUP (Category Tags)” was copied and pasted in a new column to the right using Paste Special – Values. This new column was titled “Category Tag (Static)” to represent the fact that the Paste Special – Values function created a static result from the dynamic VLOOKUP function (see fig. 7). The author was concerned that relying on a chain of dynamic, nested functions could eventually lead to erroneous outputs as the Excel sheets became increasingly complex and were increasingly manipulated for various scenarios. This step was precautionary, and applied a static version of the Category Tag to each of the data fields in the “Name 1 (Original)” columns, so that the cell values would not change as the Excel sheets were further sorted and manipulated.

Figure 7. Excerpt from the sheet titled “DB Group 1” showing the data fields matched to their Category Tags. Source: Author.
The next step in processing the data focused on calculating the frequency of Category Tags. The frequencies of the 903 individual Category Tags within the “Master Sort” sheet were tallied, representing the frequency of each Category Tag across all databases. The “Master Sort” sheet was first organized by sorting the Category Tags alphabetically, from A to Z. Next, a column titled “Frequency (Auto)” used the following formula to calculate the frequency of the adjacent Category Tag (example for cell C94):

`=COUNTIF($B$2:$B$9999,B94)`.

The entire column with the results was then copied and pasted in a new column to the right using Paste Special – Values, to create a static result from the dynamic function (see the above paragraph for author’s reasoning on the static vs. dynamic results). A column titled “Frequency Count (Auto)” used a different formula to calculate a running tally of the adjacent Category Tag (example for cell E95):

`=IF(B94=B95, E94+1, 1)`.

The entire column with the results was then copied and pasted in a new column to the right using Paste Special – Values, to create a static result from the dynamic function (see fig. 8).

Figure 8. Excerpt from the sheet titled “Master Sort” representing the Category Tag frequency counting process, and identification of duplicates for the 903 Category Tags. Source: Author.
Continuing with the frequency of Category Tags within the “Master Sort” sheet, a column titled “Max? (Auto)” used the following formula to flag the maximum value in the “Frequency Count (Auto)” column (example for cell G115): =IF(E115=C115, "MAX", "" ). The entire column with the results was then copied and pasted in a new column to the right using Paste Special – Values, to create a static result from the dynamic function. Next, the duplicate Category Tags would have to be identified and eliminated to calculate the total number of unique Category Tags.

While Excel has an automated feature that deletes duplicate entries within a certain column, this automated feature doesn’t allow much user control for which one of the duplicate items gets deleted. Therefore, the author devised a simple logic function to identify duplicate Category Tags, and flag for deletion only those duplicate Category Tags that were not associated with the maximum frequency tally and were not singular/unique Category Tags. Again, this approach is quite cautious, yet the author conducted all steps with care to preserve the integrity of the data under investigation (relying on a chain of dynamic functions could eventually lead to erroneous outputs as the Excel sheets were often sorted/filtered for various scenarios). A column titled “Delete? (Auto)” used a simple formula to identify duplicate Category Tags (example for cell I94): =IF(C94>1, IF(G94="MAX", "Save", "Delete"), "Unique-Singular"). The entire column with the results was then copied and pasted in a new column to the right using Paste Special – Values, to create a static result from the dynamic function.

At this point, duplicate Category Tags could be carefully identified and temporarily eliminated (via column filters for those rows marked as “Delete”) to calculate the total number of unique Category Tags (see fig. 9). Note that, unlike the automated Excel feature for deleting duplicate entries, this approach utilizing column filters was temporary and reversible. Ultimately, 903 unique Category Tags were identified (with 338 Category Tags having frequencies of two or greater).
<table>
<thead>
<tr>
<th>Name 1 [Original]</th>
<th>Category Tag</th>
<th>Frequency (Static)</th>
<th>Frequency Count (Static)</th>
<th>Max? (Static)</th>
<th>Delete? (Static)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active way</td>
<td>Active</td>
<td>1</td>
<td>1</td>
<td>Unique-Singular</td>
<td></td>
</tr>
<tr>
<td>Additional Rent Free</td>
<td>Additional Rent Free</td>
<td>3</td>
<td>3 MAX</td>
<td>Save</td>
<td></td>
</tr>
<tr>
<td>STREET_ADDRESS</td>
<td>Address</td>
<td>37</td>
<td>37 MAX</td>
<td>Save</td>
<td></td>
</tr>
<tr>
<td>Is Affordable Housing</td>
<td>Affordable Housing</td>
<td>1</td>
<td>1</td>
<td>Unique-Singular</td>
<td></td>
</tr>
<tr>
<td>AIR FLOW, UNI PCT_LOAD</td>
<td>AIR FLOW, UNI PCT_LOAD</td>
<td>1</td>
<td>1</td>
<td>Unique-Singular</td>
<td></td>
</tr>
<tr>
<td>AIRPERMIT LIMITS</td>
<td>Amenity</td>
<td>1</td>
<td>1</td>
<td>Unique-Singular</td>
<td></td>
</tr>
<tr>
<td>amenities date</td>
<td>Amenities</td>
<td>4</td>
<td>4 MAX</td>
<td>Save</td>
<td></td>
</tr>
<tr>
<td>First Responder System present</td>
<td>Amenities First Responder System</td>
<td>2</td>
<td>2 MAX</td>
<td>Save</td>
<td></td>
</tr>
<tr>
<td>_backup generator</td>
<td>Amenities</td>
<td>1</td>
<td>1</td>
<td>Unique-Singular</td>
<td></td>
</tr>
<tr>
<td>Theatre, 1K</td>
<td>Amenities</td>
<td>1</td>
<td>1</td>
<td>Unique-Singular</td>
<td></td>
</tr>
<tr>
<td>ups onsite</td>
<td>Amenities</td>
<td>1</td>
<td>1</td>
<td>Unique-Singular</td>
<td></td>
</tr>
<tr>
<td>Anchor GLA</td>
<td>Anchor Leased Area</td>
<td>1</td>
<td>1</td>
<td>Unique-Singular</td>
<td></td>
</tr>
<tr>
<td>Anchor Tenants</td>
<td>Anchor Tenants</td>
<td>1</td>
<td>1</td>
<td>Unique-Singular</td>
<td></td>
</tr>
<tr>
<td>Annual Taxes</td>
<td>Annual Taxes</td>
<td>3</td>
<td>3 MAX</td>
<td>Save</td>
<td></td>
</tr>
</tbody>
</table>

Figure 9. Excerpt from the sheet titled “Master Sort” representing the Category Tag frequency results, and the temporary elimination (via filtering) of Category Tag duplicates. Source: Author.

After the frequency of Category Tags within the “Master Sort” sheet had been determined, a “Priority List” of the overall most frequent Category Tags (across all databases) could be assembled. While certain database groups (the 14 sheets titled “DB Group 1” to “DB Group 14” representing the respective source companies) would undoubtedly have their own internal frequencies of Category Tags, establishing the overall frequencies of Category Tags across all databases (a separate sheet titled the “Priority List”) would provide a useful benchmark for the study. It would also serve as an interesting comparison to see how the internal frequencies of Category Tags in the 14 database groups (the sheets titled “DB Group 1” to “DB Group 14”) compared to the overall frequencies of Category Tags on the “Priority List” (see fig. 10).
After establishing the "Priority List" representing the most frequent Category Tags across all databases, the 14 separate sheets titled "DB Group 1" to "DB Group 14" (representing the 14 respective source companies) were reviewed, to match each of their Category Tags to the values established in the "Priority List" using a VLOOKUP function (see fig. 11). An example of the formula utilized is as follows (example for cell D27): =VLOOKUP(C27,'Priority List'!$A$2:$B$904, 2, FALSE). Likewise, the internal frequencies of Category Tags within each of the 14 separate sheets titled "DB Group 1" to "DB Group 14" were calculated for comparison. An example of the formula utilized is as follows (example for cell F27):
As always, the columns with dynamic results were then copied and pasted into new adjacent columns using Paste Special – Values, to create static results columns from the dynamic functions. Moreover, it became apparent in creating/applying the “Priority List” of Category Tags that many of the high-priority Category Tags were similar in nature; perhaps a second (bigger picture) round of grouping the data could shed additional light on the study.

Figure 11. Excerpt from the sheet titled “DB Group 1” showing the Category Tags matched to their Priority List numbers, and the “DB Group 1” internal frequencies of Category Tags. Source: Author.

Finally, a second pass through the data fields and their Category Tags was conducted, to further condense the Category Tags into collections of similar subjects. To further clarify this step, the author sought to identify those Category Tags with overarching themes (e.g., address and city) and consolidate the thematic groups by their general topics (e.g., location information). Consequently, a new sheet titled “Master Sort Two” was created (by making a copy of the first “Master Sort” sheet) with a new classification column titled “Topic” (see fig. 12). This new classification column titled “Topic” was hierarchically
situated one level above the “Category Tags” and hierarchically situated two levels above the original data fields in “Name 1 (Original)” column.

Figure 12. Excerpt from the sheet titled “Master Sort Two” representing the classification column titled “Topic” to identify those Category Tags with overarching themes. Source: Author.

Ultimately, 579 unique Topics were identified, with 171 Topics having frequencies of two or greater. A process for flagging and filtering out duplicates was conducted on the new sheet titled “Master Sort Two” (as discussed earlier in this section), with respect to the new classification column titled “Topic”. The histogram charting the frequency of Topics also illustrates a distribution with a very long tail of unique,
single utilization categories (see fig. 13). Likewise, another sheet titled “Topic List” with the overall most frequent Topics (across all databases) was assembled (see fig. 14).

Figure 13. Histogram, with 579 Topics on the x-axis, and frequency on the y-axis. Source: Author.
<table>
<thead>
<tr>
<th></th>
<th>Topic</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Location</td>
<td>202</td>
<td>202</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>3</td>
<td>ISP</td>
<td>132</td>
<td>132</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>4</td>
<td>Buyer</td>
<td>100</td>
<td>100</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>5</td>
<td>Seller</td>
<td>100</td>
<td>100</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>6</td>
<td>Loan Terms</td>
<td>96</td>
<td>96</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>7</td>
<td>Market</td>
<td>90</td>
<td>90</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>8</td>
<td>Date</td>
<td>88</td>
<td>88</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>9</td>
<td>Price</td>
<td>67</td>
<td>67</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>10</td>
<td>Rent Terms</td>
<td>54</td>
<td>54</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>11</td>
<td>ID</td>
<td>52</td>
<td>52</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>12</td>
<td>Broker</td>
<td>36</td>
<td>36</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>13</td>
<td>Area Building</td>
<td>33</td>
<td>33</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>14</td>
<td>Tenant</td>
<td>31</td>
<td>31</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>15</td>
<td>Owner</td>
<td>30</td>
<td>30</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>16</td>
<td>Product Type</td>
<td>28</td>
<td>28</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>17</td>
<td>Units</td>
<td>26</td>
<td>26</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>18</td>
<td>Lender</td>
<td>25</td>
<td>25</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>19</td>
<td>Risk Score</td>
<td>24</td>
<td>24</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>20</td>
<td>Transaction Type</td>
<td>22</td>
<td>22</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>21</td>
<td>Cap Rate</td>
<td>21</td>
<td>21</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>22</td>
<td>Increases Financial</td>
<td>20</td>
<td>20</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>23</td>
<td>Op Ex</td>
<td>20</td>
<td>20</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>24</td>
<td>Property Type</td>
<td>20</td>
<td>20</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>25</td>
<td>Stories</td>
<td>20</td>
<td>20</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>26</td>
<td>CMBS</td>
<td>18</td>
<td>18</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>27</td>
<td>Lease Term</td>
<td>18</td>
<td>18</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>28</td>
<td>Primary Contact</td>
<td>18</td>
<td>18</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>29</td>
<td>Appraisal Terms</td>
<td>17</td>
<td>17</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>30</td>
<td>Price per Area</td>
<td>17</td>
<td>17</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>31</td>
<td>Year Built</td>
<td>17</td>
<td>17</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>32</td>
<td>Amenities</td>
<td>16</td>
<td>16</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>33</td>
<td>Market Financials</td>
<td>16</td>
<td>16</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>34</td>
<td>Project</td>
<td>16</td>
<td>16</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>35</td>
<td>Status</td>
<td>16</td>
<td>16</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>36</td>
<td>Area Land</td>
<td>13</td>
<td>13</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>37</td>
<td>Comments</td>
<td>13</td>
<td>13</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>38</td>
<td>CapEx</td>
<td>12</td>
<td>12</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>39</td>
<td>Green</td>
<td>12</td>
<td>12</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>40</td>
<td>Parking</td>
<td>12</td>
<td>12</td>
<td>Max</td>
<td>Save</td>
</tr>
<tr>
<td>41</td>
<td>NOI</td>
<td>11</td>
<td>11</td>
<td>Max</td>
<td>Save</td>
</tr>
</tbody>
</table>

Figure 14. Excerpt from the sheet titled “Topic List” representing the most frequent Topics across all databases.

Source: Author.
Section 4 – Results

All charts, graphs, diagrams, and illustrations in the Results section were created for this thesis by the author. Data was carefully formatted in Microsoft Excel into various arrangements that were compatible with the data visualization tools in Microsoft Power BI, in order to visually represent various statistics on the Category Tags and Topics.
**Word Cloud**

The word cloud is a visualization tool that represents the frequency of a particular word within a dataset by illustrating the size of the word as proportional to its frequency (see fig. 15 and fig. 16). Therefore, words that appear more frequently within the dataset are shown in larger font sizes, while words that appear less frequently within the dataset are shown in smaller font sizes, and highly infrequent words in the dataset are omitted. The word cloud is a quick-and-dirty tool that provides an overview of the most common words in the dataset.

![Word Cloud Image](image)

Figure 15. Word Cloud of Category Tags, with size proportional to overall frequency. Source: Author.
Figure 16. Word Cloud of Topics, with size proportional to overall frequency. Source: Author.
Bar Graph

This humble graph uses bars to visually represent discrete quantities of certain variables. In the case of this thesis, the bar graph is useful for visualizing the frequencies of Category Tags (see fig. 17) and Topics (see fig. 18). For visual clarity, only those Category Tags and Topics with frequencies greater than or equal to 12 were shown in the bar graphs, as there was limited space available (especially for 903 Category Tags and 579 Topics).

Figure 17. Bar Graph of Category Tags, filtered for Category Tags with frequency ≥12 for clarity. Source: Author.
Figure 18. Bar Graph of Topics, filtered for Category Tags with frequency ≥12 for clarity. Source: Author.
Donut Chart

The donut chart is a relative of the pie chart, illustrating the relationship of parts to a whole.

Figure 19. Donut Chart of all Category Tags. Source: Author.

Figure 20. Donut Chart of all Topics. Source: Author.
Figure 21. Donut Chart of Category Tags, filtered for Category Tags with frequency ≥12. Source: Author.

Figure 22. Donut Chart of Topics, filtered for Category Tags with frequency ≥21. Source: Author.
Scatter Plot

Much like the concept of a weighted arithmetic mean, where some data points contribute more to the final average than others (as compared to an ordinary arithmetic mean), this scatter plot (see fig. 23) attempts to illustrate the weighting of various Category Tags in terms of their Topic frequency. In other words, a particular Category Tag could have been relatively infrequent as a Category Tag (and appear low on the y-axis), yet be included in a Topic that was highly frequent (which would shift the Category Tag further out on the x-axis). An example of this is “Region” which was relatively infrequent as a Category Tag, yet frequent as a Topic (since those related to location were the most common).

Figure 23. Category Tags plotted by their frequency along the y-axis, and plotted by their Category Tag frequency multiplied by their respective Topic frequency along the x-axis. Source: Author.
Radar Chart

The radar chart (also known as a spider chart, web chart, or polar chart) plots the frequency of multiple quantitative variables along axes originating from the same point. Each axis represents a separate variable, and all axes are equi-angular about the origin. The frequency of the variables are plotted as radii extending from the origin, with lengths proportional to their respective frequencies. Radar charts are useful for identifying outliers in a single dataset, and useful for viewing commonality (overlap) when multiple datasets are overlaid on the same radar chart (after a standardized chart layout is achieved). The following charts (see fig. 24 through fig. 31) illustrate the author’s process of finding a standardized layout that best represents a single dataset (for Category Tags, and again for Topics).

Figure 24. Radar Chart, showing Category Tags with frequencies ≥6. Source: Author.
Figure 25. Radar Chart, showing Category Tags with frequencies ≥9. Source: Author.
Figure 26. Radar Chart, showing Category Tags with frequencies ≥12. Source: Author.
Figure 27. Radar Chart, showing Category Tags with frequencies ≥12 and ≤29. Source: Author.
Figure 28. Radar Chart, showing Topics with frequencies ≥6. Source: Author.
Figure 29. Radar Chart, showing Topics with frequencies $\geq 12$ and $\leq 200$. Source: Author.
Figure 30. Radar Chart, showing Topics with square root of frequencies ≥2. Source: Author.
Figure 31. Radar Chart, showing Topics with square root of frequencies $\geq 2$ and $\leq 14$. Source: Author.
Network Diagram

The intention of the network diagram is to show how things are interconnected, and highlight relationships. This type of diagram is best suited to datasets that contain information flows, from a source to a destination (e.g., internet traffic and clicks from one webpage to another). As the dataset utilized for this thesis did not necessarily have an inherent flow of information, a pseudo flow was created by arranging items within each database based on their overall frequency. For example, the Category Tags within each database were arranged based on the “Priority List” spreadsheet hierarchy (recall fig. 11), with a pseudo flow of information from the most frequent Category Tag in the database to the least frequent Category Tag in the database. This methodology was repeated for the various Topics within each database, with a pseudo flow of information from the most frequent Topic in the database to the least frequent Topic in the database. This process created a connected string (representing the flow) of information for each of the 14 database groups, and the results were plotted as a network diagram, where common items (Category Tags or Topics) between databases would share a node (see fig. 32 through fig. 35). This type of diagram highlights commonality between databases as information intersects at nodes (with the size of nodes representing the frequency of overlap), and also highlights database independence as unique information flows off into long strings of single nodes.
Figure 32. Network Diagram for the 14 database groups, with nodes representing Category Tags. Source: Author.
Figure 33. Network Diagram for the 14 database groups, with nodes representing Topics. Source: Author.
Figure 34. Network Diagram for the 14 database groups, with nodes representing Topics, filtered to omit the single-utilization Topics for additional clarity. Source: Author.
Figure 35. Network Diagram for the 14 database groups, with nodes representing Topics, filtered to omit the single-utilization Topics for additional clarity. Source: Author.
**Chord Diagram**

The chord diagram is an elegant way to illustrate interconnection between things, and visually represent what they share in common. Like a donut chart, the outer ring shows the proportionality between various items through the relative sizes of each piece (of the outer ring). The arcing connections that travel across the circular diagram from one item to another represent the connections, with the size of the arcing connection representing the strength (or in this case frequency) of the connection between items. For this thesis, careful sorting and filtering of data was necessary to prevent over-cluttering of the diagrams (as demonstrated in fig. 36).

![Chord Diagram](image)

Figure 36. Chord Diagram for the 14 database groups, representing their connections to all of the 903 Category Tags.

Source: Author.
Figure 37. Chord Diagram for database groups to Topics, filtered to show only those data points with internal database frequencies >1, and filtered to represent only those Topics with frequency >100. Source: Author.
Figure 38. Chord Diagram for database groups to Topics, filtered to show only those data points with internal database frequencies $>1$, and filtered to represent only those Topics with frequency $>90$. Source: Author.
Figure 39. Chord Diagram for database groups to Topics, filtered to show only those data points with internal database frequencies >1, and filtered to represent only those Topics with frequency >60. Source: Author.
Figure 40. Chord Diagram for database groups to Topics, filtered to show only those data points with internal database frequencies >1, and filtered to represent only those Topics with frequency >40. Source: Author.
Figure 41. Chord Diagram for database groups to Topics, filtered to show only those data points with internal database frequencies >1, and filtered to represent only those Topics with frequency >40 and <200. Source: Author.
Figure 42. Chord Diagram for database groups to Topics, filtered to represent only those Topics with frequency = 1.

Source: Author.
Section 5 – Conclusions

To yield meaningful insights from Big Data, several distinct disciplines have developed independently over the course of human history, converging only within the last few decades to forge an entirely new field known as Data Science. This interdisciplinary field was conceived atop countless layers of discovery and innovation, by numerous individuals who dedicated their lives to incrementally advancing the fundamental knowledge necessary to arrive at the threshold of Data Science, and unlock the door to immense analytical potential.

Figure 43. Timeline by Mamatha Upadhyaya showing the development of Data Science through several distinct fields, from *A Brief History of Data Science*, Capgemini, 11 July 2014. Web. 14 May 2017.
It is with a mindfulness of the interdisciplinary nature of the field, and a humility for the monumental foundations laid by the pioneering forefathers, that our generation must approach such a transformative tool of enormous power. Much like its inception, the future of the field is likely to be just as interdisciplinary, as those with deep technical knowledge yet diverse backgrounds collaborate to perform meaningful analyses with Big Data. "For example, we know of a data scientist studying a fraud problem who realized that it was analogous to a type of DNA sequencing problem. By bringing together those disparate worlds, he and his team were able to craft a solution that dramatically reduced fraud losses" (Davenport and Patil).

The results of this thesis demonstrate that an incredible amount of data is available to describe the built environment. After sorting through the 2,568 data fields and labeling the information with 903 unique Category Tags, there were only 338 Category Tags with overall frequencies greater than or equal to two (37.4% of the total Category Tags). After doing a second pass through the 2,568 data fields and labeling the information with 579 Topics, there were only 171 Topics with overall frequencies greater than or equal to two (29.5% of the total Topics). This reflects distributions with very long tails, as seen in the histograms for the Category Tags (recall fig. 5) and the Topics (recall fig. 13). In other words, there were numerous single utilization Category Tags, and numerous single utilization Topics (recall fig. 42). This indicates that these 14 commercial real estate data aggregators are pushing the limits of what we can know about a property, building, or asset.

The known world, in terms of data on the built environment, is expanding. With more data covering more aspects of the built environment, our quantitative analytical processes can only improve. To reiterate the words of MIT’s Prof. David Geltner, “we do now have much more and better data, and getting more and better all the time. And we now have computational power that can make use of this data. There is the potential for a major cultural shift in how the real estate investment industry does business” (NAREIT).

Nonetheless, this thesis is inherently limited as only the data fields from the various commercial real estate data aggregators were studied. Therefore, this thesis cannot evaluate the accuracy of the data
from various providers, the data collection methods of the various providers, or speak to the analytical methods required to perform such evaluations. This thesis simply aims to explore the following: what can we know about a property, building, or asset? To this end, the thesis accomplished the goal of exploring the types of data that are available for the built environment, and who is disseminating that knowledge. In a nascent field where large-scale comparative studies of commercial real estate and urban technology databases have not yet been attempted, this thesis made a meaningful contribution by exploring the limits of what we currently know.

With an increasingly digital future ahead, this thesis also provides a general framework for creating a usable amalgamation of databases, and highlights the merits of specific tools and presentation methods that translate an immense and disparate array of information into user-friendly analytical tools. The database manipulation methods covered in the Data and Methodology section, as well as the charts, graphs, diagrams, and illustrations in the Results section may serve others attempting similar comparative studies. This thesis presents a process that can be repeated as the cumulative data on the built environment expands.

“Data scientists’ most basic, universal skill is the ability to write code. This may be less true in five years’ time, when many more people will have the title ‘data scientist’ on their business cards. More enduring will be the need for data scientists to communicate in language that all their stakeholders understand – and to demonstrate the special skills involved in storytelling with data, whether verbally, visually, or – ideally – both” (Davenport and Patil).
Works Cited


