Alexander James Gaspar Simoes  
B.S. Northeastern University 2010

Submitted to the  
Program in Media Arts and Sciences,  
School of Architecture and Planning,  
in partial fulfillment of the requirements for the degree of  
Master of Science in Media Arts and Sciences  
at the Massachusetts Institute of Technology  

June 2012

© Massachusetts Institute of Technology 2010. All rights reserved.

Author

Alexander Simoes  
May 11th 2012

Certified

César A. Hidalgo  
ABC Career Development Professor  
MIT Media Arts and Sciences

Accepted

Mitchel Resnick  
Department Head  
Media Arts and Sciences
The Observatory: Designing Data-Driven Decision making tools

Alexander James Gaspar Simoes
B.S. Northeastern University 2010

Submitted to the
Program in Media Arts and Sciences,
School of Architecture and Planning,
On May 11th 2012 in partial fulfillment
of the requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

Creative usages of graphics to encode information date back to at least the beginnings of the industrial revolution. It is also around that same time that a gap between the wealthiest of nations and least begins to develop. How can we use techniques of visualization complimented with vast amounts of data to provide a lens by which we may understand economic development?

The amount of computing power and data available at our finger tips is increasing everyday. This thesis will introduce The Observatory as a tool that combines big data with interactive visualizations as a means for discerning the patterns found in economic development over the past 50 years. The tool draws on influences from other interactive visualizations tools as well as theory and literature from the field of complexity economics. The impact of this tool has already begun to emerge with its proliferation online and usage by experts in the field of development economics.

Advisor

César A. Hidalgo
ABC Career Development Professor
MIT Media Lab
The Observatory: Designing Data-Driven Decision making tools

Alexander Simoes
B.S. Northeastern University 2010

Submitted to the
Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Master of Science
at the Massachusetts Institute of Technology

Advisor

César A. Hidalgo
ABC Career Development Professor
MIT Media Lab

Reader

Ricardo Hausmann
Professor of the Practice of Economic Development
Center for International Development at Harvard University

Reader

Catherine Havasi
Research Scientist
MIT Media Lab
Acknowledgments

César Hidalgo my advisor and mentor for all of his advice and guidance throughout this entire process. He has been an inspiration both in and outside the walls of academia.

Ricardo Hausmann and Catherine Havasi, my readers, for their suggestions and feedback on earlier drafts. They are brilliant minds whose work has been of great motivation for this thesis.

The rest of the Macro Connections team, Phil Salesses, Shahar Rosen for helping me with many of the higher level concepts. Also, the incredible undergraduate researchers (UROPs) - Anurag Kashyap, Crystal Noel, for helping me wrestle data, write efficient algorithms and learn a bit about being an MIT undergrad.
Contents

1. An ark for the flood
   Introduction

2. Beautiful pictures
   History of visualization

3. Closing the gap
   A new view of economics

4. The 1s and 0s
   Data

5. The Observatory
   The tool

6. Impact

7. Conclusion

8. References
1. An ark for the flood

Introduction

Since the industrial revolution, we have seen a large gap develop between the wealthiest of nations and the least (Maddison, 2007) (University of Pennsylvania, 2011). The problem of economic development is a complex matter with traps along the way making it common for countries to get stuck in the pursuit thereof (Hausmann & Hidalgo, The network structure of economic output, 2011). The reason for this is that economic growth comes from the development of products with similar capabilities to the ones that a country already possesses. Therefore, we can think of a country’s economic development as traversing through a network of products connected by the capabilities required to produce them. If a country tries to reach toward a product too far away and thus out of reach from the space they occupy in this network, then they can fall short, stuck in a trap. Is it possible to build tools to enable policy makers to avoid these traps and understand their country’s inherent productive capacities? GPS devices enable users to drive on roads they’ve never travelled by displaying the route towards their destination and so too could data visualizations steer countries towards a path of economic prosperity.

We are currently at an interesting crossroads in which technology has made it easier than ever to store vast amounts of information for relatively cheap cost. Current estimates have set the world’s total amount of information at more than a zettabyte or 1 trillion gigabytes, and forecast that the volume of which will increase at a rate of more than double every two years (IDC, 2011). Just as a matter of example, Facebook alone has stated that it holds 30 petabytes of data, a significant chuck of which consisting of user uploaded photos (Facebook, 2011).

Thus far we have mainly focused on digitizing the world around us—whether that be scanning the books in our libraries to create the world’s first digital collection of books, freed from the confines of physical buildings (Google, 2005), or preserving multimedia works from the inevitable decay of physical mediums (Library of Congress Digital Preservation, 2011). Many new and revealing insights have come of such efforts including the Google n-gram viewer, a queryable database and visualization of over 1 trillion words from published books (Google, 2006). Figure 1 shows the frequency of the words map, chart, diagram and visualization over time giving insight as to when these terms were first used and how their popularity has risen and fallen over the years.
Figure 1 – The Google n-gram viewer extracts important insights about our culture based on the frequency of words found in published books.

Yet as we progress we will need tools built not only for the preservation of historic media, but those to query, curate and maintain control over the information which is being created (Lesk, 1997). For example, on Twitter alone, a site known for its concern with brevity, 6.25 gigabytes of new data are being created everyday (250 million tweets per day with 100 characters) (MSNBC, 2011). Thus, the time to create tools to combat this inevitable data deluge is nigh. Tools created to leverage the growing trend of big data, without the abuse of aggregate measures, could begin to help tackle such complex problems as understanding economic development. They could lead the way towards better understanding and thus more informed decision-making.

This thesis introduces The Observatory of Economic Complexity as such a tool. A guiding motivation here is the usage of a particular big dataset in uncovering answers to a difficult problem. How can we use observational data for this purpose? What does it mean to build tools with big data in mind? Unfortunately there is no silver bullet in using massive quantities of data to make predictions, but there are best practices to be learned. In section 4 I define the term big data and show how The Observatory makes use of it.

The dataset used in this thesis is bilateral world trade data starting in the year 1962. This dataset, described in much further detail below, is a collection of around 200 countries, 2,015 products (from 2 different classifications) spanning 48 years. Here we are using the products that a given country produces (exports) as an analogue for the productive capacity they hold. The burgeoning field of Complexity Economics, discussed further in section 3, shows the importance of exports to a country's economic growth. Here, I further this claim by showing how the specific graphical display of this data can enable users to better understand the implications of this research, providing the use-case of the process of creating The Observatory of Economic Complexity.
Graphical displays of data have been around for quite some time with most scholars dating the earliest examples to the 18th century (Friendly, 2009). Here I show how the combination of interactive visualizations with big data can be used as a means for understanding economic development. The Observatory is a tool built as a website hosting a suite of data visualizations meant to encourage understanding and the predictive nature of economic development. In order to give context, I start with a brief but important history of visual representations of data, highlighting some of the major advances in the field.

In presenting The Observatory (section 5) I will also highlight some best practices to be used as a general framework in dealing with big data and show how the mechanisms behind this tool can be generalized to fit other similar datasets to facilitate understanding in other areas of interest beyond economic development. This is not to say that the tool provides a cookie-cutter like solution but that the process and steps taken to build it could be extrapolated and used in other domains.

2. Beautiful pictures

History of Visualization

An historic view of visualization

There exists a long history of using graphical representations of data to promote change or advocate for a given topic. William Playfair, who is often cited as the father of statistical graphics introduced the following types of diagrams (all of which are now considered ubiquitous): the line graph, bar chart and pie chart. Figure 2 shows a line graph depicting the balance of trade between England and both Norway and Denmark as a function of time, published in 1786 (Friendly, 2009). Shading the space between the two lines a different color when the trade balance is positive versus negative helps the viewer to immediately understand the impact of the chart. It is interesting to note that as a visualization pioneer, Playfair started with economic data, specifically trade data, a very similar dataset to that being shown by The Observatory.
Another early example of graphical statistics is the work of French Engineer Charles Minard in his now famous map of Napoleon’s march to Moscow and back, created in 1869, shown in Figure 3a. It is tempting at first glance to simply view this image as a static diagram depicting the relative path taken by Napoleon’s army (Friendly, 2009). Yet upon closer inspection we see that Minard has visually encoded much more data than a traditional map would show. The thickness of the line corresponds to the number of soldiers present at any given location and time displaying as a byproduct deaths and desertions. A simple and immediately understandable visual statistic serving to humanize and add context to the data represented by the graphics. In Figure 3b we see the exact same graphic generated using digital technologies (Google maps and Protovis – a javascript library for data visualization) (Bostock M., 2010). This shows how the principles of successful data visualization remain the same while technologies continue to update. For this reason, we should try to leverage technology for what it can add as benefit to the well-established visualization techniques that have been pioneered over the years. The Observatory, as shown in this thesis, takes a similar stance in utilizing modern technologies to allow for interaction and also leverage computing power to dynamically generate thousands of different combinations of visualizations at a click of a button. Thus, for the first time, users can engage and shape their view by slicing the data – allowing for a rich user experience.
British physician John Snow is well known for his contribution to this field as well – pioneering the technique of spatial analysis. By coupling accurate geographic locational data of Cholera outbreaks he helped propel the germ theory of infectious diseases (Friendly, 2009). Until this time, medicine had relied on the ancient theory of humors or temperament to explain illness. For example, one could either have an excess or deficit of one of the four humors (earth, water, air or fire) and as such, need an adjustment. A technique known as blood letting would require patients to use devices (often leaches) to drain blood from their body to stabilize their humors (Seigworth, 1980). In Figure 4 we see Snow’s map showing the
locations of infections in 1854, mostly located nearby a local water pump, providing evidence that Cholera was an infectious disease. Later the pump shown in the map would be shut down. Thus, this graphic is another example of how visual language can persuade decision makers towards actionable results.

Lastly we have the work of Florence Nightingale advocating for improved sanitary standards in care given to soldiers during the Crimean war of 1853 - 1856. To make her point she compiled the now famous Nightingale Rose diagram (Figure 5) showing the different causes of death among combatants (blue is disease, red is wounds, and black is uncategorized). The graphic uses color to display the classification of each cause and the size of each segment (corresponding to the month of the year) to denote the total number of deaths. The fact
that so many deaths were preventable is immediately apparent when viewing this diagram (Friendly, 2009) (Munzner, 2009). One criticism to this particular display of information is the propensity of the human eye to misread the size of each category of data shown in the wedges. Since the area of a wedge, geometrically speaking, is larger as it gets further from the center, it is easy to introduce a bias in the information being communicated. Thus, the same graphic displayed with a different ordering of categories could take on an entirely different meaning.

Figure 5 – Nightingale’s graphical representation of causes of death during the Crimean War

A visual grammar
It was not until the mid 20th century that the term graphic design first began to enter the lexicon (Figure 6). When this did occur, a formal language began to establish itself around the visual encoding of data as graphics and gave birth to the field of information design. Seminal to this movement was Jacques Bertin, a French cartographer and author of the highly influential Sémiologie Graphique (Munzner, 2009). In Figure 7 we see a chart depicting the different forms of visual language generalized and abstracted to show how their variance can distinguish different dimensions of data.
The work of formalizing the language of embedding information in graphics was progressed further by another influential political scientist turned statistician, Edward Tufte. He is well known for many of his publications as well as bringing to light many core concepts illustrating best practices, including the notion that as much of the ink (or pixels) of a visualization as possible should have meaning. This concept is known as the data to ink ratio (Mulrow, 2002). Tufte has long lamented the proliferation of such out-of-the-box solutions as
Microsoft’s PowerPoint blaming it for lowering the standard of communication with an audience. He has even gone so far as partially blaming it for the engineering oversights of the failed Columbia Shuttle mission (Tufte, The Cognitive Style of PowerPoint, Pitching Out Corrupts Within). An underlying concept of this thesis, shared by Tufte in many of his writings, is that the full details of any given set of data are necessary to represent the complexities of the world around us.

**Put into practice**

The importance of information or data visualization cannot be underestimated as it can lead to very powerful changes. A fascinating example of this is the Tidy Street project, executed by an organization called CHANGE (a collaboration between Nottingham University and The Open University). Residents of Tidy street, a short road in Brighton UK were asked to participate in a study in which their daily electrical consumption was shared and displayed in chalk on their street as shown in Figure 8. Over the course of the study, roughly a few months, a remarkable thing happened – the total electrical consumption actually went down by 15%, not exactly a trivial amount (The Guardian, 2011). This illustrates the fact that by simply displaying information in a manner that is concise and easy to understand, change can be provoked.

Thaler and Sunstein make a similar point about nudging people towards better decisions on topics ranging from healthcare to investments. One key to the advice they give, is to design systems that give feedback. “Well designed systems tell people when they are doing well and when they are making mistakes” (Thaler & Sunstein, 2008). One could clearly see how information visualization can fall into this category of providing a “nudge” in the correct direction towards anything from minimizing environmental impact to taking steps towards economic development. This is another driving force in this thesis, showing how actionable advice can come out of well-designed tools via the graphical representation of data from the world around us.
Ambient displays, a technology invented by the Tangible Media Group at the MIT Media Lab, are another example of the use of visual technology designed for the purpose of influencing behavior. Using variance in color, sound, airflow or other peripheral perception, ambient displays encode information, yet remain in the background of a given space. To denote meaning ambient devices such as The Orb (Figure 9) provide the viewer with a spectrum of color as an actionable display correlating to a particular interest (stock prices, energy usage, availability of conference rooms) (Ishii & Ullmer, 1997).

Although there have been many pioneering advances in the field of graphical representations of data over the years, many of them fall short when dealing with large datasets. Charts such as pie graphs, bar charts, line graphs, etc., are great for getting a general idea of a dataset.
but tend to fall short in properly expressing the complexities of the world we live in. For this reason we need to look at the raw or disaggregated data to gain insight into higher level meaning. Thus, a custom solution is often required. This thesis will look at a specific use case – that of shedding light on the problem of economic development.

**Computers– the birth of an interactive medium**

No longer are the beautiful pictures of print design trapped in a static medium. With computers we can click, hover, drag and gesture our way around the “page” to provide new views of our data. Many brilliant minds have begun to leverage this medium to show new types of visualizations that couple the computer’s ability to process huge amounts of data with the input of users to interact or take action on these results.

One such example of this is the tree map, created by Ben Shneiderman, a computer scientist from the University of Maryland. Similar to a pie chart in which area is representational of the value of it’s corresponding item, this visualization uses rectangles for a given data point instead of wedges which (as shown in Figure 5) can be misleading. Each item is then hierarchically nested within a parent category with an area equal to the sum of its parts (Shneiderman, Tree visualization with tree-maps: 2-d space-filling approach, 1992). One particular use case of this, as shown in Figure 10, is that of a file system. Although mapping the values of each folder’s size is a trivial task to a computer, deciding which files are superfluous is best left to a human. ( Wouldn’t want to learn one day that a file is missing all of a sudden because our computer decided it need to be deleted!) Here is where the true power of data-driven decision making tools shines through. The user is able to interact by slicing, or skinning the data as he or she sees fit to aid their final decision of deciding which files to delete. Such a task becomes infinitely more difficult when one is forced to use a textual, command line representation of the same data.
Martin Watterberg, a pioneering data scientist and visualization guru further improved the tree map layout with his SmartMoney Map of the Market webpage shown in Figure 11. Here we see a tree map displaying the size of each stock with color corresponding to percentage of loss or gain for the day (Wattenberg, 1999). By minimizing the aspect ratio of each element (making them more square) the user is able to compare the sizes more easily, unlike a pie chart's wedges which get more difficult to compare as more items are added. This particular use case of displaying stock prices is the perfect example to illustrate how interactivity can benefit the user by tailoring the information displayed to that of the user’s need. Since the data is clustered hierarchically, we see (as an overview) only the names of the top-level categories and on mouse over a contextual display of the specific stock of interest. This falls perfectly in line with Shneiderman’s Visual Information Seeking Mantra of “Overview first, zoom and filter, then details-on-demand” (Shneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations, 1996).
Moving into the 21st century, with computer and internet access becoming more and more ubiquitous, we reach a new era of data visualization. Now, with the ease of sharing brought on by the increased connectedness of the web, we are able to both share our own data and harness the capabilities of users across the globe to create our data. Place Pulse is an example of such a project which was created by Phil Salesses of the Macro Connections group at the MIT Media Lab. By harnessing the interactive nature of the internet along with crowdsourcing, he has employed an army's worth of users to generate data for him and any other interested parties to use for the application of understanding urban aesthetics. In Figure 12 we see how users are shown two images and asked to evaluate them based on a given question such as “which do you think is safer?” By aggregating votes of enough users, the tool becomes useful in telling which areas of a city are considered more or less desirable, as shown in Figure 13, thus yielding actionable advice in the form of personalized data (Salesses, DeVincenzi, Martino, & Hidalgo, 2011). Quite similar to the John Snow example mentioned earlier (figure 5), the results from the Place Pulse project can be visually encoded on a geographic map. The only difference being that we did not need to walk from door to door to get this data!
Figure 12 – Place Pulse, a project leveraging user participation to gain insight into urban aesthetics.

2 Voronoi Perception Maps of New York City

Figure 13 – Results from the Place Pulse project visually encoded on a map.

Open Mind Common Sense is a website that has been created for the purpose of collecting implicit knowledge about the world around us that often goes unsaid. For example, the English sentence “I went to see a movie last night” leaves out the fact that the film was shown in a theater. This information, considered obvious to most native speakers, is assumed and as such only used for clarification or distinction e.g. “for the 30th anniversary of George Lucas’s Star Wars, they will be showing it in theaters.” This assumed information is useful to computers, especially for the purpose of artificial intelligence. Since this data often
cannot be found in sentences written or spoken, the Open Mind Common Sense site allows human users to contribute via prompts or free text (Havasi, Speer, & Alonso, 2007). Thus this site is another example of crowdsourcing, allowing users connected online to contribute basic knowledge towards the goal of larger project.

Networks
There exists a long history of the use of networks to visualize relationships in data, specifically social relationships. Early scribbles of the tree of life drawn as a network are found in Charles Darwin's notebooks (Figure 14a). Linton Freeman, in his well-documented account on the different advances made in graphical styles and encodings, dates the earliest hand-drawn social networks to Jacob Moreno in 1932, as shown in Figure 14b (Freeman, 2000). Freeman also shows examples of other graphical enhancements, many of which are now ubiquitous, such as the use of color, size and spatial positioning to embed information in nodes/links. Delving deeper into this history, he notes different layouts that have been experimented with, such as a radial layouts used to emphasize the specific number of degrees away from the center that a node is.

As with all visualizations, certain styles may serve to better inform the user of different aspects of the data set. For instance, computers are able to computationally generate layouts with useful spatial encodings for discovering clusters of nodes. This is useful for many reasons, including the ability to discover communities in a network and to experiment with different layout algorithms to find the best fit for a given dataset. Many software packages now exist to help users generate networks from their own data, some stemming from specific use cases (Cytoscape for biomolecular interactions (Shannon, et al., 2003)) and
others for the sole purpose of developing a tool to generate networks, such as Gephi (Bastian, Heymann, & Jacomy, 2009), Pajek (Batagelj & Mrvar, 1997), Ubigraph (Veldhuizen, 2008) and Ucinet (Borgatti & Freeman, 2002).

There is also much work that has gone into creating specific network visualizations for the task that they be more suitable to a specific domain. Such digital domains as email and social networks have allowed users to connect and share in ways previously not possible. Yet many of the most popular websites for these tasks, such as Facebook, Twitter and Google’s Gmail rely on outdated user interfaces – paged results focusing heavily on temporal ordering like those shown in Figure 15. I will use the term typewriter paradigm as a reference to these interfaces, harkening back to a time before graphical user interfaces, when using a typewriter forced results to be printed in the order they were first input. Getting stuck in this typewriter paradigm of paged lists binds the representation of our data, potentially obscuring the most important features and patterns. The PostHistory and Social Network Fragments project is one such solution to this problem. Using email archives as a data source Viégas et al discovered the usefulness of having both a traditional node-link network view along with a heatmap encoded calendar view stressing the temporal aspect of the data to reveal social patterns (Viégas, Fernanda B., Nguyen, Potter, & Donath, 2004). Vizster, a network visualization for the now defunct Friendster social networking site does the same. Created as an end-user visualization for the exploration and understanding of one’s own social connectedness, it helped users learn about themselves and the network around them (Heer & boyd, 2005).

Figure 15 – Three popular sites for online communication (Facebook, Twitter and Gmail) stuck in the typewriter paradigm of using paged lists for ordering data

Computational layouts of networks are great for exploration and learning about a data source but are less useful for comparisons. This is due to the fact that the specific x, y position of any node is not meaningful and would thus change the next time the layout is generated. Many layout algorithms fail to give precedence to aesthetic heuristics like minimizing link overlaps and spacing clusters apart to show distinctions in communities. Often times, for a cleaner layout, users creating a network may first rely on an algorithm for initial node/link placement and then make minor adjustments by hand. The Product Space, a
network of products traded on the world market, is an example of such a hybrid, as shown in Figure 16 (Hausmann R., et al., 2011). Hausmann et al use this visualization as a sort of map for economic development; highlighting which products a country exports in this network as means of comparison with both itself (over time) and other countries. Here it is in the viewer's best interest for the positions of nodes/links to remain static.

![Figure 16 - The Product Space, a network of relatedness between products traded in the global market](image)

Until recently, to show the path taken by different countries towards economic development in the Product Space, one would need to regenerate the network with specific products turned on or off corresponding to that country’s exports (shown graphically by highlighting nodes). As part of this thesis I will show the process of transforming this network into a dynamic data-driven tool useful to policy makers.

Visual representations of networks are not always necessary to provide comprehension. Networks can also act as a system structure that can lead to new insights and help encode connections in information. ConceptNet is an example of such a network. Nodes are concepts like nouns or short phrases and edges are the relationships between them. These
edges are predicates that attempt to describe how one concept relates to another. The data for this project, as mentioned earlier, comes from the crowdsourcing application of The Open Mind Common Sense Project. To illustrate how this network connects concepts, let us take the word “keyboard” which may have a relationship of “UsedFor” to “typing” and also a relationship of “IsA” to “musical instrument”. These are distinctions we may understand as humans but fail to encode in the machines we create. Natural language is highly irregular and often times based on implicit or contextual assumptions. The ConceptNet connects such information so that it can be understood and made useful in a variety of applications including artificial intelligence (Havasi, Speer, & Alonso, 2007).

Visualization Libraries

There are many toolkits available for producing visualizations, some more successful than others. A common goal that they all share is that of allowing users with varying degrees of computer programming knowledge to produce graphical representations of their data. There are two polar ends of this spectrum with full-fledged software solutions (such as Microsoft Excel) at one end and base level programming languages that expose graphical primitives, to enable algorithmic drawing, on the other. Yet, the most interesting cases seem to be somewhere in between.

Many Eyes is a website that allows users to upload their own data and generate pre-built visualizations of their choosing (Viegas, Wattenberg, van Ham, Kriss, & McKeon, 2007). One requirement of the use of this service is that any data uploaded be made public, allowing other users to view all the different visualizations and underlying data used to create them. A benefit to this solution is the fact that no programming experience is required. Similar to Microsoft Excel in this regard, anyone can bring their data and choose from a selection of partially customizable visualizations to find the best way of communicating the story of their dataset. One of the problems with this approach is that users are forced to select from a set of visualizations and styles that may or may not suit their data.

Tableau is another example of software intended for a non-programming audience. Borne from research on visually representing databases at Stanford University, Tableau is now both a desktop and online software service (Hanrahan, 2006). Users are able to apply different visualizations to their raw data based on parameters of their choosing. They may then publish this creation, share it on a website or use it for their own internal purpose. This is the closest one can get to a custom designed data visualization without needing to actually design or program it themselves. In Figure 17 we are able to see a screen shot of the software’s interface (Tableau Software, 2003).
The processing programming language and IDE (integrated development environment) is an example of a visualization tool closer to the programming end of the spectrum. Processing is a wrapper around the Java programming language that exposes a set of graphical and representational primitives to aid the user in creating custom graphical displays of their data (Reas & Fry, 2003). There is no prebuilt bar graph or line chart “type” in Processing, these visualizations would need to be created via shapes and text. Built as a tool for designers, Processing affords non-programmers the power to generate dynamic, algorithm driven software with a more gradual learning curve than most programming languages.

D3 or Data-Driven Documents is another similar visualization toolkit that allows users to attach data to create custom graphical elements. The distinction and beauty of D3 being the fact that it is written entirely in JavaScript and works in all modern web browsers. The central philosophy of this visualization toolkit is that data structures are attached to graphical elements (SVG or HTML styled with CSS). Another difference between D3 and other drawing libraries is that D3 does not try to introduce a new set of graphical primitives but instead relies on a well known graphical standard, SVG (scalable vector graphics) and creates a wrapper around that (Bostock, Ogievetsky, & Heer, 2011). The Observatory, as shown in the
technical details, makes use of this library and embraces its expressiveness as a tool to allow custom visualizations tailored to the specific needs of the dataset used.

**Economic Tools**

It is worth noting other interactive visualizations used specifically for the display of economic data, since this is the domain of The Observatory. Arguably the most successful tool is that of Hans Rosling, called Gapminder. With the broad goal of creating a “fact based world view”, Gapminder uses an interactive X,Y plot that shows the change in any two development indicators over time as shown in Figure 18a. The user may select which indicators they would like to view (for example GDP per capita and Life Expectancy) to reveal the underlying relationship between the values for these countries (represented as colored dots) (Rosling, 2008). Though the visualizations themselves lack novelty, the tool has done quite a bit to show the power of interactive graphics in communicating powerful ideas such as the rise in CO₂ emissions as a result of increased income per capita. Google has since bought the underlying technology used by Gapminder to create their own Public Data Explorer (Google, 2011). Shown in Figure 18b, it looks very similar but allows users to view different public datasets and even upload their own (similar to the Many Eyes project and Tableau software).

![Figure 18 - a comparison of the Gapminder software and Google’s Public Data Explorer that is based on the same underlying technology](image)

There are also a few visualization tools developed by the World Bank to help users understand many of their indicators along with a public API (Application Programming Interface) for other developers to use with their data. Many of the World Bank tools such as the Mapping for Results and WITS (World Integrated Trade Solution) websites are quite crude and rely heavily on single view layouts (usually geographic projections). It is clear that a gap exists in the market for a useful data driven tool for understanding economic development.
3. Closing the gap
A new view of economics

It is hard to put into words a concept so broad and so illusive as economic development. An entirely naive and colloquial definition would be something along the lines of – you either got it, or you ain’t. But that wouldn’t be telling much of the story, as countries like Brazil, China and India would be complete anomalies. Yet depending on the lens through which we view economic development, they actually tend to fall perfectly in line.

The power and harm of aggregation
As humans, we love facts, statistics, rankings or other ways by which we can easily make comparisons. It helps us make quick decisions. For instance, if I need to get across town and a display board tells me the next train will not arrive for over an hour, I would most likely seek an alternate route, possibly choosing to take a cab. As another example, while driving I make mental notes of the gas prices to help me decide where to fill up (stations along the highway are always the most expensive!) One last illustrative example is that of choosing the best colleges to apply to. I would try to find the highest ranked schools that would accept someone with my GPA and SAT scores. These are all situations in which single numbers aggregate and hide complex meanings to help us make easier or faster decisions. By using the US News rankings I am saved from having to interview students from a random set of colleges to make my decision since this data has been aggregated and parsed for me.

Yet as useful as aggregation can be, it’s not all sunshine and roses. As any statistician will tell you, numbers can often be misleading. Moreover the way in which we choose to display data can particularly affect the conclusions an audience reaches. This phenomenon, initially noted by Edward Simpson’s 1951 paper The Interpretation of Interaction in Contingency Tables has been thusly coined Simpson’s Paradox (Simpson, 1951). A well known example of this (noted in most introductory statistics textbooks) is a case of claimed gender bias in the admittance of graduate students to the University of California, Berkeley. As shown in Table 1, there appears to be a higher percentage of male graduate students admitted. Yet, when disaggregated by department (Table 2) we see that in fact more departments had a female bias in their admissions. The reason for the discrepancy is due to the fact that the departments in which there were more men than women accepted were the ones in which many more women than men applied, thus skewing the results when aggregated. Table 2 shows the departmental breakdown of the percentage of each gender admitted. The percentages of women accepted for each department are highlighted.
<table>
<thead>
<tr>
<th>Applicants</th>
<th>Admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>8442</td>
</tr>
<tr>
<td>Women</td>
<td>4321</td>
</tr>
</tbody>
</table>

Table 1 – Gender composition of the percentage of admitted students to the University of California, Berkeley’s graduate programs

<table>
<thead>
<tr>
<th>Department</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>Accepted</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>62%</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63%</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37%</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33%</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28%</td>
</tr>
<tr>
<td>F</td>
<td>272</td>
<td>6%</td>
</tr>
</tbody>
</table>

Table 2 – Percentage of accepted students broken down by department

Macro Economics is a field where aggregations are extremely common, one long standing example being the Gross Domestic Product. Developed as a means for summarizing the economic pulse of a nation in a single statistic, the GDP was born. The use of such an abstraction allows the public to know whether a country’s economy is growing or shrinking, yet it does not do much to educate them on the inner workings of this process. Simon Kuznets, the creator of this measure, was among the first to offer criticisms. In his Nobel Prize acceptance speech he begins by stating that the measure does little to provide the public with a full view or explanation of the economy (Kuznets, 1971).

Another measure that attempts to correct for the biases of GDP is the Human Development Index introduced by the United Nations Development Program in their 1990 Human Development Report. The HDI also delivers a single numerical value corresponding to a combined index of the health, education and wealth of a country. The problem with this index is that it makes the same mistake as GDP, in that it tries to summarize a country’s quality of life by using aggregate measures. One may conclude that it is a better statistic than GDP due to the fact that it takes into account other factors beyond a country’s financial records, but it is still a ways off from capturing the true essence of economic development.

It’s what you do that matters
As previously stated, the traditional approach to viewing a country’s productive capacity is to look at aggregate measures such as its wealth or human capital or with what natural resources it may be endowed. This theory would make sense as a model, or if we were to try and shape the world according to our intuition i.e. countries would export their abundant goods and import the rest, leading towards a path of specialization (Heckscher & Ohlin,
1991). Yet this is not what trade data says; instead, the picture painted is one of diversification over time leading towards economic growth (Hausmann, Hwang, & Rodrik, What you export matters, 2006). It has been shown that taking a disaggregated view of a given country’s economy or looking at its Economic Complexity Index are much higher predictors of growth (Hausmann R., et al., 2011). This is not to say that other, more traditional measures of economic development such as education or governance indicators do not matter, but that productive knowledge depends on many of these other indicators and as such embeds much of their data within, thus reflecting it.

Take Figure 19, which shows the exports of Rwanda over a 40 year period. Around 1995 we see that they are cut in half, undoing a decade worth of growth due to a massive civil strife better known as the Rawandan Genocide. Clearly the effects of governance are reflected in this country’s export production.

Another note to make when taking a disaggregated look at a country’s exports is that not all products should be considered equal. There exists a nested nature when looking at world trade i.e. there are certain products that many countries export and some that only a few
export (Hidalgo & Hausmann, 2009). What can the ubiquity of products tell us about them and their country of origin? For one, we may note that the less ubiquitous products are highly specialized and are produced by the most diversified economies. Thus we reach the conclusion that to produce these less ubiquitous products it takes many inputs and embedded knowledge from the countries that are producing them.

For example, to have the capacity to produce and export a given pharmaceutical there exists the implicit embedded knowledge required to produce this drug in compliance with safety regulations and intellectual property law. We can think of the knowledge or contribution of each individual involved in this value chain as a personbyte. There are also the personbytes of the logistics required to coordinate and produce these drugs, such as the assembly lines and packaging of the final product. Let us not also forget about the shipping routes required to move these products, since we are talking about exports, the countries that produce any particular product need also to be able to transport it to its destination. We cannot discount these requirements – roads, shipping routes, access to ports, airports, etc.

The problem now lies in how this intangible knowledge can be measured. Much of the said knowledge is difficult to accumulate since it cannot be found in textbooks or learned in a classroom. It comes from years of experience and the interactions of people in human networks. Yet by looking at world trade data and seeing which products a country exports we can infer (based on the other countries that export them) how complex or difficult they are to produce. It is through this intuition that we gain insight into the usefulness of the product space as a decision making tool. Figure 20 shows a magnified segment of the product space with each node representing a product exported on the world market (in the SITC4 product classification). Node size indicates world trade volume (for that product), color corresponds to community e.g. blue is machinery, purple is chemical (the full list is shown in Appendix a). The exports of a country can then be shown in full opacity, such as those of Brazil in 1992 revealed in Figure 20. The most likely products for a given country to begin producing would be those that are close by, meaning the nodes shown in half opacity connected to nodes of full opacity.
The theory behind the metaphor of nearby and faraway products corresponds to the point mentioned in the introduction that there are traps in the path towards economic development. Such traps come from the difficulty of gaining the productive knowledge required to begin producing new products. It is hard for an industry to exist without the necessary productive knowledge, yet it makes little sense to begin accumulating such knowledge if there is no industry that would require it. For this reason it is more probable that a country would exploit the productive knowledge they already have and expand it in a similar direction to gain the knowledge required for another similar product.

By using the knowledge that what a country exports matters and that economic growth comes from the development of adjacent industries, we can see the usefulness of a tool to empower individuals with the means of facilitating change. Such an individual could be a policymaker looking to incentivize an industry to spur economic growth, or an entrepreneur seeking her next venture (or better yet next venue). After what has been stated in previous sections, it would not make sense to say there is one solution to solve this problem. Therefore this thesis will show a particular tool built with these criteria in mind. The following sections will outline what The Observatory of Economic Complexity is by discussing the data, visual languages explored and technical details required to produce it.

Figure 20 – Brazil's exports in 1992 shown in the SITC4 Product Space taken from The Observatory
4. The 1s and 0s

Long ago was the age of loading values into an Excel spreadsheet to produce a pre-built chart to understand a particular set of data’s contents. There are many new challenges brought about by the increased demand (both computationally and theoretically) of big data. Thus, these techniques of yesteryear are rendered obsolete.

Big Data

When used throughout the course of this thesis, the term “big data” will refer to datasets that are large in the following three aspects:

1. **Size**
   This is usually the meaning of big data, referring to the literal storage capacity required for a given dataset. Here, this would mean that the size of the dataset would be that of or approaching the size of memory for the workstation being used to process the data.

2. **Resolution**
   The dataset must also be big in resolution, meaning that the data is as disaggregate as possible, providing the highest level of granularity. For example, when looking at trade data, comparing total exports of a set of countries can only show overall trends from year to year. Yet, when viewing these numbers as the sum of the products that add up to them allows for further analysis and research. Portugal and Kazakhstan have similar total export amounts but the variance in the products they export is highly askew.

3. **Scope**
   The last criteria and probably most abstract is that of being big in scope. In general terms, this would mean that the use of such a dataset could have great significance or value, beyond its original intent. Again, using trade data as a qualifying example, though the intent of the UN in aggregating it is purely a statistical one, The Observatory uses it in a way that shows its significance with regard to economic development.

The aforementioned criteria are highly dependent on the dataset being evaluated. Some may fall on different ends of the spectrum e.g. being smaller in size and larger in resolution.

There are also many problems that arise when dealing with big data. Certainly not meant to be an exhaustive list, the following are difficulties one may face:
• Size of data makes loading files into memory impossible
• Dirty data becomes harder to spot (too costly time-wise to look over entire dataset with just the human eye)
• Trends or relationships may go unnoticed with flat files
• Searching or querying can be costly using lazy “touch-all” methods

Since we are dealing with many new problems, we need new tools to better equip ourselves in dealing with them. In this light I intend to evaluate The Observatory, as both an example of a tool for dealing with big data and as a practical tool for decision makers.

Data Repositories
Access to data is something that can be useful for many different reasons including research initiatives, reducing redundancy (no reason for the same data to be collected more than once) and providing a means for civic engagement. The problem is that much of the data available comes at a cost or with certain stipulations regarding its use. Transparency in data collection methods and data repositories allows errors to be uncovered as well as spurring innovation in such topics as data mining, new collection methods and most of all, consistency in reporting formats. Yet many open data initiatives have found little success.

Two relevant examples of data repositories and initiatives with entirely different results are the World Bank and UN COMTRADE. Both in the business of keeping records and statistics from world governments, the World Bank gets many things right that COMTRADE does not. The first being access. Over 8,000 indicators are available from World Bank datasets via API and online queryable website (The World Bank, 2010). The COMTRADE site, on the other hand, requires downloading their dataset in batch queries limited to 50,000 entries (UN COMTRADE, 2010).

Bipartite
In terms of graph theory the data used by The Observatory will be in the form of a bipartite graph. A bipartite graph is one that contains two disparate sets of elements with edges connecting elements from each of the two sets together. Figure 21 is an example of such a graph.
Since the data used in The Observatory is bilateral (importers and exporters), there is a third column in our bipartite graph. We can think of trade flow as an input output chain with products. Looking at Figure 22 we see a country (element from group A) exports a product (element from group B) to another country that is now considered the importer (a different element from group A). Due to the fact that trade data is symmetric, we will not say that the data is a tripartite graph but instead a mirrored bipartite graph. Saying that USA exports cars to Japan is the same as saying Japan imports cars from USA. An important distinction to be made here is that USA may export cars to Japan and also import cars from Japan but these will not be the same value.
Here we are able to see that the exporters and importers are comprised of the same elements but with different connections or edges to the products. Lastly, I will show another example of city industry data represented as a bipartite graph (Figure 23). This illustrates the fact that The Observatory could be used with many different types of data, essentially any dataset represented as a bipartite graph.
Trade data

The data used in The Observatory comes from several different sources each with their own particular details. In general terms, there are four distinct variables that govern the expression of this data. They are: exporters (countries), importers (countries), products and years, which I will refer to from here forward as \( C_i C_e P Y \). As a byproduct of the bilateral nature of this data, there is also a fifth variable, the trade flow, indicating if the data being shown are exports or imports. I will abbreviate this as \( TF \).

The data source with the most extensive coverage temporally is a dataset from UN COMTRADE cleaned and complimented with country customs data by Robert Feenstra, available from the Center For International Data at the University of California, Davis (Feenstra, Lipsey, Deng, Ma, & Mo, 2005). This dataset uses the SITC (Standard International Trade Classification) revision 2 classification of products disaggregated to the four-digit level, with 774 products in total. This source includes data for the years 1962 – 2000.

Since the Feenstra dataset has not been updated to include more recent data, The Observatory uses data directly from the UN COMTRADE data repository. The Macro Connections team at MIT Media Lab and Center for International Development at Harvard University has developed their own algorithm to clean this data source. For the sake of comparing apples to apples, this data also uses the SITC revision 2 product classification.

Figure 23 – City data disaggregated by industry shown as a bipartite graph
As mentioned, the SITC data was used for comparison sake, to allow users to view data for an entire 47 year period (1962 – 2009) with a consistent product classification. This technique does not come without a price. In Figure 24 we see the Netherland’s exports in the year 2009 in both the SITC4 and HS4 classifications. The SITC4 data (shown left) reports 27% of their exports as unclassified. Since the SITC product classification is no longer used by reporting countries, there exists a discrepancy in translation from the Harmonized System (the newly accepted reporting standard) to SITC. For this reason, The Observatory also includes a third source of data using the more up to date HS classification compiled by Centre d'Etudes Prospectives et d'Informations Internationales (CEPII) for the years 1995 – 2010. Though this data is reported with six digits of resolution, the version used on The Observatory aggregates the data to use four digits. I will address the reason for disregarding this finer resolution in the section of this thesis about The Observatory.

![Figure 24 - A tree map created using The Observatory, of Netherland's exports in 2009 in two different product classifications](image)

5. The Observatory

The tool

Borrowing its name from a place whose sole purpose is for gazing at the night sky, The Observatory of Economic Complexity serves an analogous purpose. When one looks up at the sky on a clear night in an urban area they might be lucky to see 30 stars. More remote areas, with less light pollution would obviously increase this number, yet there will always be the basic limitation of both distance and the physical ability of the human eye as impediments to getting any fine grained images to study and research. For this reason, the telescope was created. Gallileo aided by the use of a telescope was able to prove a
hypothesis by Copernicus that we all take for granted today, which is that the Earth does in fact rotate around the sun. By selecting smaller portions of the night sky to view in higher detail or resolution, telescopes provide a more accurate and disaggregate view of the extra terrestrial world above us.

The Observatory of Economic Complexity allows the user the same privilege as a literal observatory except instead of celestial views it provides those of the world economy. As shown in the previous section, there are many different ways that this data can be sliced and showing each one is paramount to the purpose of The Observatory and the success of any other tool for understanding big datasets.

**Views**

Slicing the data into different views allows the user the ability to answer different questions. The following table shows all of the different ways one may choose to view the data:

<table>
<thead>
<tr>
<th>Name</th>
<th>Question</th>
<th>Parameters / Combinations</th>
<th>Shown</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Country trade</td>
<td>What products does a particular country trade?</td>
<td>Trade Flow, Country, Year, Products</td>
<td></td>
</tr>
<tr>
<td>2 Trade Partners</td>
<td>What countries does a particular country trade with?</td>
<td>Trade Flow, Country, Year, Countries</td>
<td></td>
</tr>
<tr>
<td>3 Product trade</td>
<td>What countries trade a particular product?</td>
<td>Trade Flow, Product, Year, Countries</td>
<td></td>
</tr>
<tr>
<td>4 Trade between</td>
<td>What does a particular country trade with another country?</td>
<td>Trade Flow, Country (origin), Country (destination), Year, Products</td>
<td></td>
</tr>
<tr>
<td>5 Country trade by</td>
<td>Where does a particular country trade a particular product?</td>
<td>Trade Flow, Country, Product, Year, Countries</td>
<td></td>
</tr>
<tr>
<td>product</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 - The different ways data may be sliced on The Observatory

**Apps**

Each app represents a different view of the data. In Figure 25 you will see the four different types of apps available on the observatory site [a] tree maps, [b] stacked area charts, [c] the product space and [d] maps. Each of these apps can be used to further show different slices of the data shown in table 3.
Tree Maps

Based on the original tree map algorithm by Ben Shneiderman and improved upon by Martin Wattenberg’s “Map of the Market” we use this tiling algorithm to generate a rectangle that when taken as a whole represents 100% of either a particular country's trade in a given year or of a particular product's traders in a given year. Figure 24 shows a tree map representing The Netherlands exports in the year 2009. Here, each rectangle represents a product that The Netherlands exports. As with all tree maps the area of each rectangle maps to a value in 2-dimensional space, in this example the value is the export share as a percentage of The Netherlands total in 2009. Moreover, each rectangle’s color represents its element’s category (a full listing of which is shown in Appendix a). For products, this is based on the communities found in the Product Space and for countries this is the geographic region in which they are located. At a glance one is able to see the largest elements that make up the total trade of whatever view is being shown. For example, in Figure 24 we see unclassified transaction as the largest export for The Netherlands in 2009. As a point of comparison in Figure 26 we show the tree map of exports for two different countries – Brazil and China. Both of which have undergone rapid development in recent years but show entirely different export baskets. Brazil appears to have a much more agriculture and natural resource based
productive capacity, while China relies mostly on machinery and garments (as denoted by the blue and green sections).

Figure 26 – Tree maps of Brazil’s & China’s exports, as seen in The Observatory, shows two countries undergoing rapid development yet clearly with different product structures.

Stacked Area Charts
The stacked area charts used in The Observatory are unique in their ability to show the same dataset we have been working with but showing those same values for multiple years. One may view data as either shares of a total or in nominal values of US dollars. We are able to plot here how a country’s trade diversification changes over time (what products they are exporting or importing) or how a product’s traders change over time (which countries export or import a given product). Let us use the example of Rwanda from Figure 19. We are able to see their exports over a 40 year period (from 1965 – 2005), showing their changes over time. Once again, it is immediately apparent that something had gone awry around 1990 corresponding to the Rwandan genocide. Much of the knowledge that can be learned from simply gleaning at these visualizations would not be possible without spending hours poring over spreadsheets of data.
Product Space
The Product Space visualizations, as mentioned earlier, are based on the research of (Hidalgo, Klinger, Barabási, & Hausmann, 2007). Each node in this network represents a product and links connect products that tend to be exported by the same countries. The visual language used by these graphics shows products being exported by a country with full opacity. The location of a country in the Product Space is highly predictive of the products that they may begin to export in the future. The product space is therefore a predictive tool that can help guide industrial policy. As shown in Figure 27 we see how China occupies the Product Space in the years 1985, 1990, 1995 and 2009. Based on this visualization we are able to see how China’s development changes from exporting products mainly concentrated in the periphery of the Product Space in 1985 to those in the more densely connected center. Specifically the electronics cluster in the top left (cyan nodes) shows remarkable growth as we see in the last frame of Figure 27 almost every node is highlighted.

Maps
The last of the four visualizations on The Observatory is a collection of maps showing where a given product is traded (exported, imported, net exported or net imported). Here we are able to see geographically which countries or particular regions of the world are the major traders of a selected product. In regard to the other visualizations, the maps are most likely the least useful for analysis. A classic issue when showing data tied to geographic location is
that smaller countries appear to be underrepresented and vice-versa. In Figure 28 we see a map showing the countries that export fresh cut flowers vis-à-vis a tree map of the exact same data. In the tree map it is strikingly apparent that The Netherlands has a dominant position in this market, yet on the map, due to the small geographic size of The Netherlands, this information is obscured.

![Exporters of fresh cut flowers (2009)](image)

**Figure 28** – A geographic map vis-à-vis a tree map, both from The Observatory, showing the exact same data

**Technical Details**

The site is currently running at atlas.media.mit.edu as a public website with access to the entire web. The following details discuss how the site manages the data – held in a database and transforms it into online friendly interactive visualizations.

**Database**

On the server there is a MySQL relational database that holds all of the values for 48 years worth of bilateral trade data in two different product classifications (SITC4 and HS4) totaling 11.8 GB of data. The primitives of this schema (graphically represented in Figure 29) are countries and products (the two groups of the bipartite graph) with a third variable, year, which allows for higher temporal resolution of the data. These primitives are then aggregated into both country-product-year and country-country-product-year (CPY and CCPY respectively) tables which house the values for these unique rows. A sample entry for the CCPY table, showing butter trade between New Zealand and Australia, would look similar to that shown in Table 4. The CPY table shows the same data as CCPY except that it is aggregated to show the sums of bilateral partners as total values for every country-product-year combination. An easy way to think of this is that the CPY table contains an implicit destination of the world for every entry. A Generalized example of such a row, showing the total value of butter traded between New Zealand and the world is shown in Table 5.
Figure 29 – Database schema used to house bilateral trade data, the heart of The Observatory

<table>
<thead>
<tr>
<th>Origin</th>
<th>Destination</th>
<th>Product</th>
<th>Year</th>
<th>Export Value</th>
<th>Import Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Zealand</td>
<td>Australia</td>
<td>Butter</td>
<td>2009</td>
<td>$8,338,000</td>
<td>$982,000</td>
</tr>
</tbody>
</table>

Table 4 – A sample row from the CCPY table

<table>
<thead>
<tr>
<th>Country</th>
<th>Product</th>
<th>Year</th>
<th>Export Value</th>
<th>Import Value</th>
<th>RCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Zealand</td>
<td>Butter</td>
<td>2009</td>
<td>$23,661,000</td>
<td>$5,312,000</td>
<td>8.4352222</td>
</tr>
</tbody>
</table>

Table 5 – A sample row from the CPY table

Server-side
The technology used to access the requested data from the database lives on the server. The language used to script this process is Python, since the server-side technology will also need...
to serve webpages and route requests, the Django web framework is used. Django applications facilitate the use of the Model-View-Controller design pattern, shown in Figure 30, as a means of bringing structure to users' web applications. The model will allow access to the data by creating an object-relational-mapping for each table in the database. The view provides the styling of the front-end of the application (what is sent to the browser), fusing data from the database with static mark-up. The controller handles the business logic, routing the incoming request to the proper function, which in turn decides which models and views to combine and serve to the user via an HTTP response. The decoupling of these roles allows the developer to structure their code in such a way that they may modify certain aspects of their application independently.

Figure 30 - A diagram describing the model-view-controller design pattern, used by The Observatory via the Django web framework
Front-end

Thus far, the technologies mentioned only reside on the server and remain unknown to any users operating the site. Front-end web development is the opposite, any and all code written for the browser will be shown graphically to the users and made available as source to any curious minds aware of the “view-source” ability of their browser. The three technologies are HTML (for markup), CSS (for aesthetic styling) and JavaScript (for logic).

A fourth, now native in-browser technology, SVG (scalable vector graphics) is also used in The Observatory for the purpose of building interactive visualizations. Much like HTML, SVG is a mark-up language consisting of graphical primitives in lieu of HTML’s formatting tags. Figure 31 shows the mark-up (top) required to produce the SVG image (bottom). The process of turning raw data (as retrieved from the database) into SVG graphics is facilitated through an intermediary logic, written in JavaScript and residing on the front-end. Paramount to this transformation is the D3 JavaScript library, which acts as a graphical abstraction allowing the developer to attach data points to specific elements and map them to their values. Figure 32, written in JavaScript using the D3 library, shows an example of attaching data to SVG elements (rectangles) that in turn dictates their size and position on the screen.

```xml
<svg width="100%" height="100%">
  <rect fill="red" stroke="black" x="15" y="15" width="100" height="50"/>
  <circle fill="yellow" stroke="black" cx="150" cy="15" r="20"/>
</svg>
```

Figure 31 - SVG markup (top) shown alongside its rendered graphical equivalent (bottom)
Interface

The Observatory of Economic Complexity would not be much of a tool if attention weren’t paid to the way in which users could access its information. Figure 33 shows a screen shot from the “explore” page, used to control which visualization is being viewed. The header (a) functions to show the current question being asked, along with the site wide navigation and current language selection (top right). The left-hand side is the app pane, a list of all the different visualizations on the site. Here they are grouped by parent category, only showing the current selection’s group. Drilling down into each of these top-level categories will show the respective views available for that app. The middle of the page is the canvas, reserved for displaying the current visualization. Since all of the visualizations are interactive, on mouseover, text is shown along the bottom pertaining to the current node selected. There is also a key just above the timeline, which shows either product categories or country regions depending on the view. Along the very bottom, is the timeline, which allows the user to scrub through the different years of data available. The data for all years is loaded when the user first builds the visualization, therefore switching between years does not require another call to the server and can be rendered immediately. Lastly, furthest on the right is the control pane. These are the different variables that may be set pertaining to the current view along with links to export the visualization as a static image or text.

```
// Create SVG container and insert into the page
var svg = d3.select('#my_id').append('svg')

// attach array of integers to SVG element
var square = svg.selectAll('rect').data([1, 6, 10, 12]);

// newly created elements enter the stage
var enter = square.enter().append('rect');

// style elements
enter.attr('fill', 'red');
enter.attr('width', String);
enter.attr('height', String);
enter.attr('y', 20);

// set the x position based on binded data
enter.attr('x', function(d) {
  return Math.pow(d, 2);
});
```

---

Figure 32 – JavaScript code using the D3 library to produce graphics bound to an array of data
Use Case: Flowers in the Netherlands

Let us now step through the example of a user interested in the flower trade of the Netherlands. Upon entering the site, the user may choose to explore the global trade of flowers on the world market. Shown in Figure 34a, we see this is around 50% as of 2010. Now the user may use the timeline to display previous years data and see how this has changed. Figure 34b shows that in 1995 the Netherlands was exporting 60%, proving that trade has declined in only 15 years. Next, to answer the question of who is actually buying flowers, the user may look into the importers of fresh cut flowers. Here, in Figure 34c we see the major importers are Germany, UK, USA, France, Russia, Japan and the Netherlands. So where would the Netherlands be buying these flowers from? Here the user would uncover the fact that Kenya and other countries in East Africa are the dominant countries, as shown in Figure 34d. Looking at Figure 34e we see that in Germany, the Netherlands is the top exporter while in USA and Japan, the Netherlands is being edged out by developing countries like Ecuador, Columbia and China (Figure 34f - g). And lastly in Figure 34h we see that in Russia, Kenya is starting to show up as an exporter. This user is able to get a much more in depth analysis of how the flower market is changing by using this disaggregate set of data allowing them to reach a conclusion that the Netherlands is losing dominance in the market due to developing countries bypassing them and exporting directly to market. This is just one particular use case...
case of a single country and single product, yet there are thousands of combinations or views of this data to be explored.

Figure 34 – A sample use case of using The Observatory to uncover changes in the Dutch flower market over a 15 year period.

API

Though the casual users may be satisfied to poke around and explore The Observatory via its web interface, advanced users may wish to use just the visualizations for their own purposes. Much like YouTube or other sites hosting rich media content, The Observatory supports the ability to embed all of its visualizations via an API. On each unique page in the “explore” section, there exists the equivalent API code, allowing users to embed the visualizations on their own site. Different options exist for each of the visualizations, such as language of text, layout and interactivity. The full documentation can be found at atlas.media.mit.edu/api.
6. Impact

Facilitating Decision Making
Using comparisons to understand differences in world trade enables users to visually understand the economic make-up of a country. Specifically by using the product space to visually encode the network of products traded on the global market and overlaying a given country’s exports we are able to see near-by products.

Through the lifecycle of The Observatory I have conducted many informal user studies and consulted with academics and experts in the field of economic development. Below are some of the insights gained.

Online usage
As of April 30th 2012, the number of total users to the website was just over 83,000, with about 174,000 total page views. The majority of users have come from the developed world (USA, Spain, United Kingdom, Mexico and Germany are the top 5) but there are also users coming from further reaches of the globe such as Iran, Ghana, Fiji, Rwanda and Paraguay to name a few. A common trend shown in the website’s analytics is that the site will be posted to an online social news aggregator such as Reddit in the United States or Menéame in Spain causing a spike in traffic as shown in Figure 35. That being said, there is still a steady stream of visitors to the site everyday, usually around 200 visitors.

Real world usage
The Observatory of Economic Complexity has been exhibited in the lobby of the MIT Media Lab as shown in Figure 36. Since this was the first time that the site had been on display for a wide audience there were many things learned that in turn informed later iterations. Since users were interacting with the site via touch-screen we realized the first enhancement that was needed was for larger buttons. This also proved beneficial for users of mobile devices
and tablets. Another missing aspect from the original designs were the keys signally what the different colors represented in each of the visualizations. Lastly, in this first version there was no way to easily switch between years for any given view. For this reason a timeline was implemented to allow users to easily move a slider back and forth to change the year or simply press play to see an animation.

Figure 36 – An exhibit featuring The Observatory in the lobby of the MIT Media Lab

The Global Empowerment Meeting is an event held annually at Harvard University hosted by the Center for International Development. It is a gathering of policy makers, academics and business leaders for the purpose of developing new strategies to accelerate growth and unlock potential in developing countries. The most recent GEM, held in October of 2011 featured exercises in which participants were asked to develop strategies for a given country using The Observatory. The exercise proved successful in showing experts in the field of development economics how to use the site as well as informing future changes and enhancements to the user interface. One particularly insightful request was that for showing not only what a country exports or imports but also their net exports and net imports. This was a simple improvement that did not require any other data to be added to the site yet was not something that I, as a developer, would have thought to implement.

There are also two other exhibitions that have taken place featuring The Observatory. The first was at the United Nations Conference on Trade and Development (UNCTAD). Here an earlier version of the site was shown alongside posters of the Product Space. The other exhibit for which I was personally in attendance was the 2010 Global Manufacturing Competitiveness summit in Washington DC. As shown in Figure 37, three flat screen monitors were set up with iPads on podiums as controllers. Users were able to select the different visualizations on the iPad and see it displayed on the screen in front of them. This way the exhibit was interactive, yet also allowed onlookers to learn from the exhibit.
Visualization is a technique that is not new and in fact has been around as early as the industrial revolution. Humans are visual thinkers, equipped with the ability to easily discern patterns in graphics. It is this reason that a successful means of communicating data is by encoding it graphically. That being said, a problem begins to arise when attempting to represent the world around us by means of visualization as it is intrinsically at odds with the complexity of our reality. It is this reason that custom solutions need to be created taking into account the nuances of a particular domain's problems. Edward Tufte writes “Simplicity of reading derives from the context of detailed and complex information, properly arranged. A most unconventional design strategy is revealed: to clarify, add detail.” He is noting the fact that our world cannot be fully expressed by stripping it of its details, yet he posits that a better understanding can come from its reorganization (Tufte, Envisioning Information, 1990).

Big data presents one way in which we may save these complex details about the world around us. Many new and fascinating ways to look at problems that may have previously relied on expert analysis or theory to answer, can come from using large, disaggregated datasets. One such domain is development economics. By looking at actual data collected over an historic period we can begin to see trends and understand how patterns of economic growth emerge. Using data for research does not come without its caveats and should not be considered a magic bullet. Datasets are not always complete and may contain inherent biases in the direction of benefit for those that are reporting it. For these reasons we should
continue to push for openness and transparency in how data is collected. We will also
continue to need experts to help contextualize the conclusions that big data may help us
attain. It may be the case that insights gained from using actual data will help inform theories
for the future. It is important that tools be created to make use of big data, as methods for
collecting and disseminating it will be more ubiquitous as time goes on.

As shown in this thesis, The Observatory is a tool built for the purpose of understanding
economic development by leveraging techniques for dealing with big data and providing an
interactive platform that may be shared and used by the public. The site includes 47 years
worth of trade data in two different product classifications and over 200 countries. Users
may choose from a selection of over 100,000 different views or slices of the data shedding
light upon different aspects and patterns otherwise hidden. The Observatory is a tool built
for the specific purpose of using data to understand a problem, in this case economic
development, though the techniques outlined in this thesis could be applied to many
different domains with similar datasets used. Hopefully future development on The
Observatory will allow the site to live on for years to come with updated data and interface
enhancements to reveal new and exciting patterns allowing policy makers the ability to look
into the crystal ball of economic development.
8. References


Tuft, E. *The Cognitive Style of PowerPoint, Pitching Out Corrupts Within*.


9. Appendix

a. List of Product Communities

<table>
<thead>
<tr>
<th>Community</th>
<th>Color</th>
<th>Icon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machinery</td>
<td></td>
<td><img src="gear.png" alt="Icon" /></td>
</tr>
<tr>
<td>Electronics</td>
<td></td>
<td><img src="laptop.png" alt="Icon" /></td>
</tr>
<tr>
<td>Aircraft</td>
<td></td>
<td><img src="plane.png" alt="Icon" /></td>
</tr>
<tr>
<td>Boilers</td>
<td></td>
<td><img src="fire.png" alt="Icon" /></td>
</tr>
<tr>
<td>Ships</td>
<td></td>
<td><img src="anchor.png" alt="Icon" /></td>
</tr>
<tr>
<td>Metal Products</td>
<td></td>
<td><img src="hammer.png" alt="Icon" /></td>
</tr>
<tr>
<td>Construction materials and equipment</td>
<td></td>
<td><img src="house.png" alt="Icon" /></td>
</tr>
<tr>
<td>Home and office products</td>
<td></td>
<td><img src="desk.png" alt="Icon" /></td>
</tr>
<tr>
<td>Pulp and paper</td>
<td></td>
<td><img src="tree.png" alt="Icon" /></td>
</tr>
<tr>
<td>Beer, spirits and cigarettes</td>
<td></td>
<td><img src="beer.png" alt="Icon" /></td>
</tr>
<tr>
<td>Food processing</td>
<td></td>
<td><img src="bottle.png" alt="Icon" /></td>
</tr>
<tr>
<td>Petrochemicals</td>
<td></td>
<td><img src="petrol.png" alt="Icon" /></td>
</tr>
<tr>
<td>Inorganic salts and acids</td>
<td></td>
<td><img src="lab.png" alt="Icon" /></td>
</tr>
<tr>
<td>Other chemicals</td>
<td></td>
<td><img src="lab.png" alt="Icon" /></td>
</tr>
<tr>
<td>Agrochemicals</td>
<td></td>
<td><img src="farm.png" alt="Icon" /></td>
</tr>
<tr>
<td>Chemicals and health related products</td>
<td></td>
<td><img src="medicine.png" alt="Icon" /></td>
</tr>
<tr>
<td>Coal</td>
<td></td>
<td><img src="mine.png" alt="Icon" /></td>
</tr>
<tr>
<td>Mining</td>
<td></td>
<td><img src="hammer.png" alt="Icon" /></td>
</tr>
<tr>
<td>Oil</td>
<td></td>
<td><img src="hammer.png" alt="Icon" /></td>
</tr>
<tr>
<td>Precious stones</td>
<td></td>
<td><img src="diamond.png" alt="Icon" /></td>
</tr>
<tr>
<td>Textile &amp; fabrics</td>
<td></td>
<td><img src="textile.png" alt="Icon" /></td>
</tr>
<tr>
<td>Garments</td>
<td>![Garments Icon]</td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>Cereals and vegetable oils</td>
<td>![Cereals Icon]</td>
<td></td>
</tr>
<tr>
<td>Cotton, rice, soy beans and others</td>
<td>![Cotton Icon]</td>
<td></td>
</tr>
<tr>
<td>Tropical treecrops and flowers</td>
<td>![Tropical Icon]</td>
<td></td>
</tr>
<tr>
<td>Tobacco</td>
<td>![Tobacco Icon]</td>
<td></td>
</tr>
<tr>
<td>Fruit</td>
<td>![Fruit Icon]</td>
<td></td>
</tr>
<tr>
<td>Miscellaneous agriculture</td>
<td>![Miscellaneous Icon]</td>
<td></td>
</tr>
<tr>
<td>Fish &amp; seafood</td>
<td>![Fish Icon]</td>
<td></td>
</tr>
<tr>
<td>Meat &amp; eggs</td>
<td>![Meat Icon]</td>
<td></td>
</tr>
<tr>
<td>Animal fibers</td>
<td>![Animal Icon]</td>
<td></td>
</tr>
<tr>
<td>Milk &amp; cheese</td>
<td>![Milk Icon]</td>
<td></td>
</tr>
<tr>
<td>Leather</td>
<td>![Leather Icon]</td>
<td></td>
</tr>
<tr>
<td>Not classified</td>
<td>![Not Classified Icon]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HS4 Community</th>
<th>Color</th>
<th>Icon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal &amp; Animal Products</td>
<td>![Animal Icon]</td>
<td></td>
</tr>
<tr>
<td>Vegetable Products</td>
<td>![Vegetable Icon]</td>
<td></td>
</tr>
<tr>
<td>Foodstuffs</td>
<td>![Foodstuffs Icon]</td>
<td></td>
</tr>
<tr>
<td>Mineral Products</td>
<td>![Mineral Icon]</td>
<td></td>
</tr>
<tr>
<td>Chemical &amp; Allied Industries</td>
<td>![Chemical Icon]</td>
<td></td>
</tr>
<tr>
<td>Plastics / Rubbers</td>
<td>![Plastics Icon]</td>
<td></td>
</tr>
<tr>
<td>Raw Hides, Skins, Leather, &amp; Furs</td>
<td>![Raw Hides Icon]</td>
<td></td>
</tr>
<tr>
<td>Wood &amp; Wood Products</td>
<td>![Wood Icon]</td>
<td></td>
</tr>
<tr>
<td>Textiles</td>
<td>![Textiles Icon]</td>
<td></td>
</tr>
<tr>
<td>Footwear / Headgear</td>
<td>![Footwear Icon]</td>
<td></td>
</tr>
</tbody>
</table>
Stone / Glass
Metals
Machinery / Electrical
Transportation
Miscellaneous
Service