

VENTURES

Final Process Book

<http://didjeridou.github.io/cs171-pr-ventures/>

Guillaume Laliberté

CSCI E-171: Visualization

Prof. Alexander Lex

Samuel Gratzl

May 2015

Ventures - Final Process Book

Website: <http://didjeridou.github.io/cs171-pr-ventures/>

YouTube video tour: <https://youtu.be/dX-Ueq1er00>

Github repository: <https://github.com/didjeridou/cs171-pr-ventures>

Definitions (from Oxford Pocket Dictionary)	3
1. Background and Motivation	3
2. Project Objectives	3
3. Data	4
3.1 Source: Crunchbase	4
3.2 Cleaning	4
4. Visualization	5
First Iteration (late March)	5
Second and third Iteration (April 3rd)	6
Third Iteration (mid April)	6
Third Iteration V2 (April 17th)	6
Fourth iteration (April 25th)	7
5. Code structure	10
6. Evaluation	11
6.1 Answers	11
6.2 Improvements	12
6.3 Take-home	12
Startup-wise	12
Visualization-wise	12
6.4 Conclusion	12

Definitions (from Oxford Pocket Dictionary)

Venture

A risky or daring journey or undertaking

Startup

A newly established business typically aiming for fast, exponential growth.

Venture capital (also referred to as "venture")

Capital invested in a project in which there is a substantial element of risk, typically a new or expanding business.

Round (funding round, seed, Series, IPO)

Investment, by which a business or other enterprise raises money to fund operations, expansion, a capital project, an acquisition, or some other business purpose.

1. Background and Motivation

When is the best time to fundraise and what are some important patterns to be aware of?

Where is my startup most likely to get funded?

What are my chances of actually succeeding?

As a young entrepreneur, I have always been interested in these questions and in general how people create companies from scratch. I have myself founded a few companies, and as I am getting better at it, I try to understand how funding and financing plays a role in a company's success.

This is why I decided to create a visualization that would capture various dimensions of fundraising in the startup world. To me this project is not just school work: it is a tool I intend to use to comprehend how the startup scene is evolving, and so I can use that data to my advantage.

2. Project Objectives

The objective of this project is to provide answers the questions above.

In other words, the goal is to uncover patterns in how industries and trends about what kind of companies get funded, and uncover correlations between industry, capital and success.

To answer these questions, I need to:

- Identify the relevant datasets (Series A, IPOs, VC, acquisitions, etc.)
- Answer questions about how, where and when companies get funded
- Tell a compelling story about the startup ecosystem

3. Data

3.1 Source: Crunchbase

I initially selected data from the *National Venture Capital Association* (available through *Quandl*), but these dataset led me to an uninteresting visualization.

Instead, for the final iteration, I got access to *Crunchbase*, a database of (most) startups and their relevant funding information, meaning that I knew precisely how startups in America and Canada raised money, from who, when, and where.

However, with great dataset comes great cleaning process.

3.2 Cleaning

This was a very interesting challenge.

Crunchbase ended up giving me an access key to download their entire database. As much as this was incredibly valuable, I was just as destabilized when I got a 100mb Excel file. There were so many entries that Numbers (Apple's spreadsheet app) could not even display half of the rows—it's limited to 65,000 rows.

In order to circumvent that limitation, I used a few tricks:

I was able to export one Excel sheet at the time to CSV files. From there, I was able to use D3 to weed out entries that I would not use. For example, because I focus on North American startup environment, I could delete all entries that were not from a Canada-based or US-based startup.

From there, I needed a quick and simple way to visualize the data to understand what to do with it. Having the data in a table was now possible, so I then went back to Numbers to see what the information looked like (much easier to read than a CSV file). I quickly identified a few other entries I could delete, like funding entries without a round, startups with under \$100K in funding and more.

Even with that, loading the data into D3 was slow on my computer, and there was a ton of information that was processed even though I did not need it. For example, daily data points made every visualization lag and were not useful. Perhaps grouping those entries in monthly data points would be interesting!

Going back to the drawing board, I was able to narrow down the precise datasets I needed, complete one last filtering with D3 and exported the datasets to JSON files. The basic chart I had set up, the browser and Numbers were now all running smoothly, and I had all the information I needed.

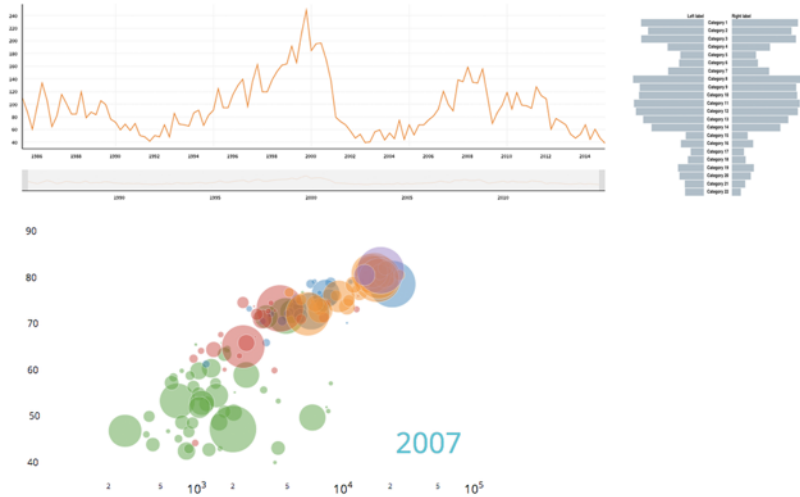
4. Visualization

In this section I present the various iterations of my design. There were a lot of changes from the first to the final.

FIRST ITERATION (LATE MARCH)

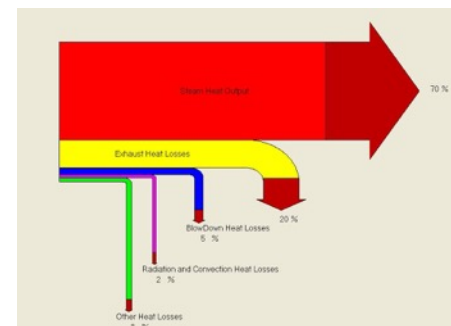
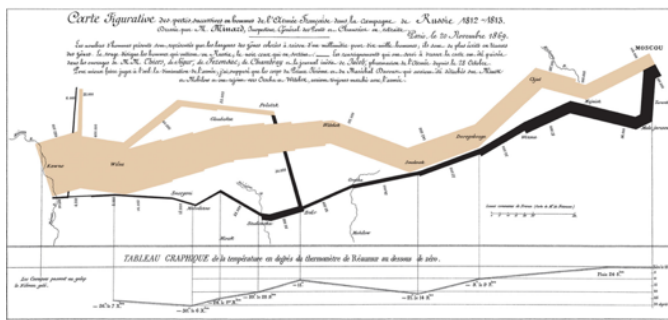
There are a few element useful to identify patterns. It seems like one line chart or stacked area chart could be the centre piece where most of the interaction happens. It allows the use to compare a few datasets (with filters) and to select a time period (brush).

A bubble chart will also be a good solution to compare the institutional investors and startups in terms of deals count or size.



One thing I would really like to achieve is a Sankey diagram: it could be useful to visualize how many companies go bankrupt at each round of funding (from early stage to IPO). This will be an optional feature because of time constraints, but it will be the first I complete if I have enough time.

There are a few other graphs that I think are really inspiring and which could be used in this project. The first one is a classic, one of my favourite graphs featured in-class. I think it could be really useful in showing how startups "die" over time¹. Similar graphs are often used to illustrate energy consumption²:

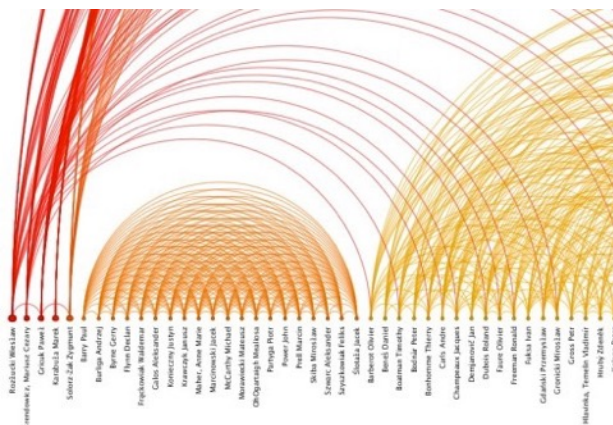


¹ <http://tomtunguz.com/images/sankey-minard.png>

² http://3.bp.blogspot.com/_XswVxxXMFWE/TJdHL-IK-ol/AAAAAAAAABg/WIbe3ZvHzqE/s1600/SankeyDiagram.bmp

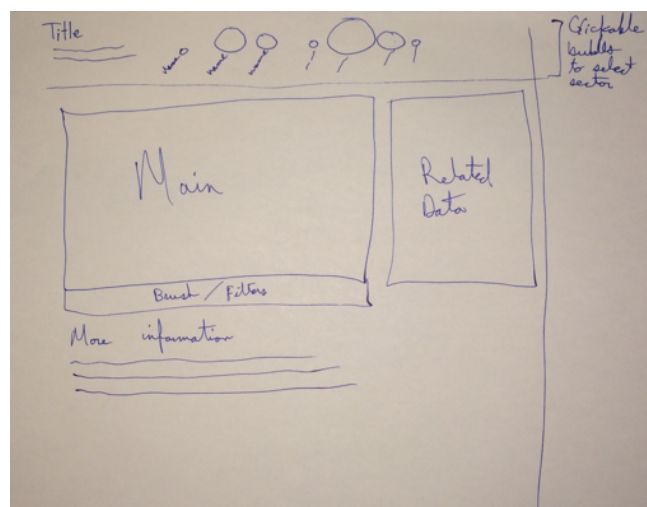
SECOND AND THIRD ITERATION (APRIL 3RD)

If I can find the data needed, another interesting alternative would be to illustrate the relationships between startups and investment firms. This would allow the user to quickly identify how firms rely on each other. It could be done with and arc connection graph:

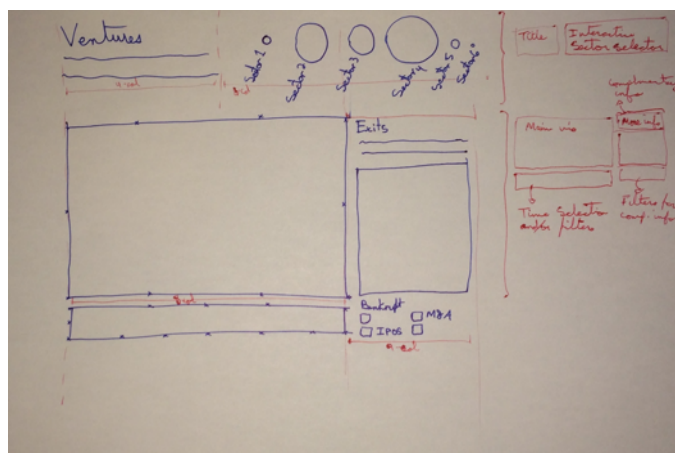
**THIRD ITERATION (MID APRIL)**

At this point, I realized (by myself and with helpful feedback) that the layout of the first option could be more interactive and that I could make it more relevant by further defining the datasets, and that the second option, the connection graph that would illustrate how investment firms are connected to deal, would end up very cluttered.

Before moving forward, I decided to refine the layout to try to get to something better.

**THIRD ITERATION V2 (APRIL 17TH)**

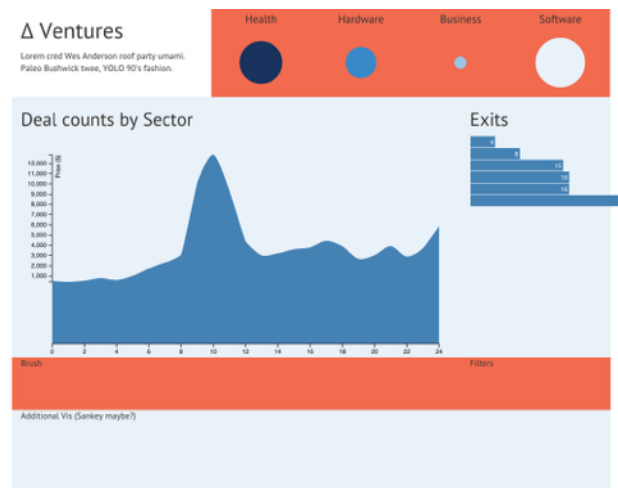
For the V2 of the third iteration, I refined the layout and started planning how this could work with the Bootstrap layout system.



IMPLEMENTATION

At this point I really thought I had something interesting to work with. I moved forward and started implementing the various graphs.

Unfortunately, there was something missing: I was not inspired, and the result was really not up to my expectations.



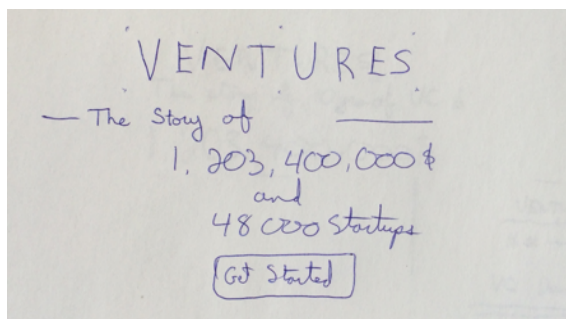
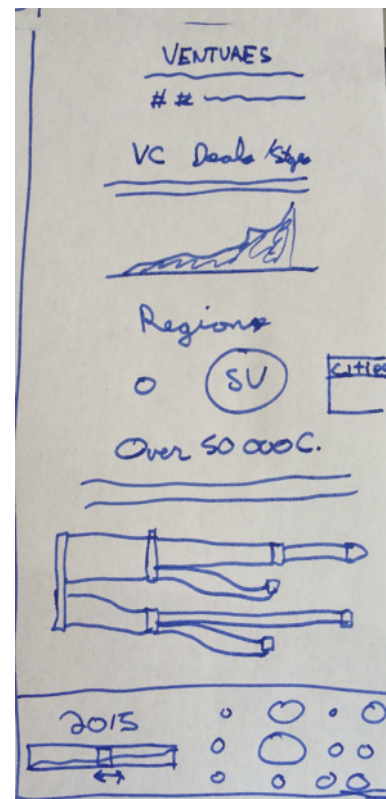
FOURTH ITERATION (APRIL 25TH)

After speaking with my CS171 advisor, I had two realizations.

- 1) I was limited by my dataset. It was too narrow, and did not provide a good diversity of information.
- 2) My vision for this project was not right. I was building a tool to navigate the data instead of uncovering insights about it. As mentioned in class, the project should feel like a story, something you would want to explore and that will give the user key information about a world that may be alien to him.

From there, I decided to radically change my approach. I started exploring concepts on how to tell a story. Twenty storytelling websites later, I identified a few patterns:

- Good online stories are vertical
- The different sections are clearly identified
- Each section is about only one thing (answers one question)
- The user must have a clear understanding of why there is a story.



"AHA MOMENT"

For me, this was the "Aha" moment. I started focusing exclusively on how I could tell a story with the data. At this point I had my access to Crunchbase, and started exploring the data (see the "Data" section of this process book).

VENTURES

The Story of 40K Startups and One Trillion Dollars

GET STARTED



In 15 years, over \$220 billion was invested in startups. That money is now worth well over \$1 trillion

CrunchBase

Visualize data from CrunchBase, the definitive database of the startup ecosystem



Learn when and where investments are made, and how you can optimize your fundraising

The first step I took was to clearly define the story I wanted to tell and the questions I needed answered.

The rest suddenly seemed much easier to design. I figured out a way to divide the story of my data into four chapters, such that each question I have could be answered properly.

It now was clear to me that one graph, no matter how interactive, could probably not answer every question.

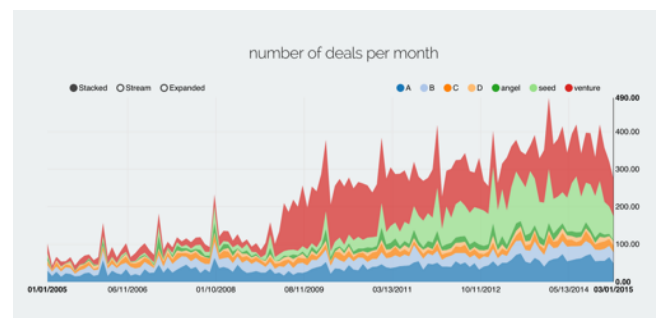
I drafted each new view to answer one of the questions I had, and added for each one a "*Findings*" section to highlight what I discovered using these charts.

When is the best time to fundraise and what are some important patterns to be aware of?

Volume, growth, growth by type of investment and minimums or maximums are very easy to identify when presented in a proper area chart. I chose to build a chart with three different layouts (stacked, stream and expanded) such that different patterns can pop out more easily.

The stacked chart makes it easy to see the growth over time of the number of deals.

The stream chart is not as good as the stacked chart for many reasons, but it reveals peaks and peak-patterns very efficiently.



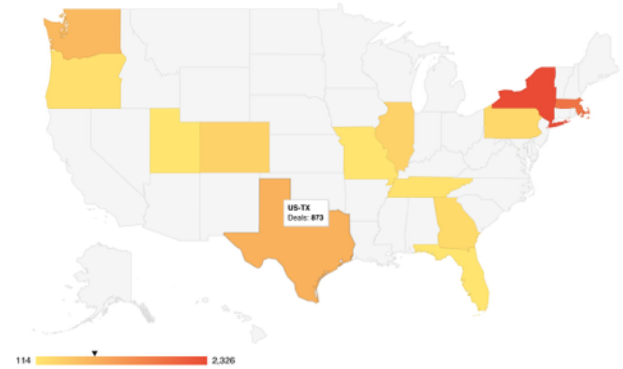
Finally, the expanded chart allows the user to compare the growth of different types of deals relative to each other in a more efficient way than the previous charts.

A complementary bar chart was added to provide some context about the size the different types of investment.

Where is my startup most likely to get funded?

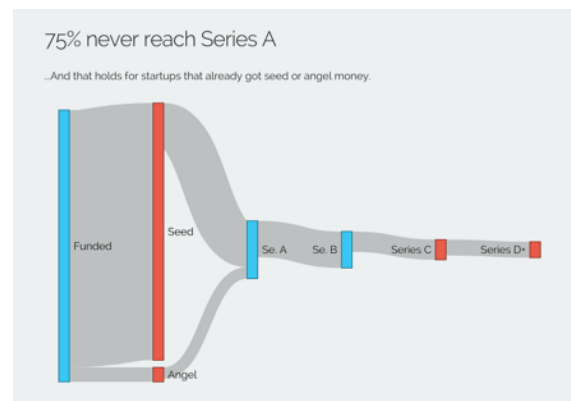
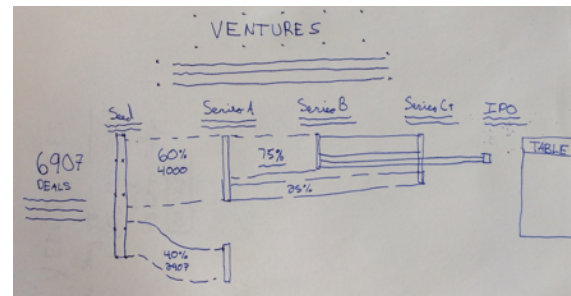
For this one, an interactive map seemed like a simple and very intuitive solution.

Because the goal is to identify startup hubs, I decided to use a color scale with an interactive mouseover tooltip.



What are my chances of actually succeeding?

Finally, I decided to move forward with a Sankey graph and a complementary table to illustrate how few startups get to advanced funding rounds.



But 62% who gets to Series A have a much better chance

And the 5-6% who reach Series D+ win big

	Raised by	Round	Amount	Sector	Region
1	Groupon	D	\$950mm	Retail	IL
2	Cloudera	F	\$900mm	Software	CA
3	Magic Leap	B	\$542mm	Video	FL
4	Lyft	E	\$530mm	Software	CA
5	Airbnb	D	\$475mm	Tourism	CA

5. Code structure

/Ventures	
index.html	Main page
README.md	
/_docs	Contains documents, process books
...	
/css	Custom CSS and grid css
complement.css	
normalize.css	
nv.d3.min.css	
style.css	
/data	Contains the CSV and JSON data
50cities.json	These JSON contain the data for their
dealsize.json	respective visualizations.
months.json	'rounds_month.csv' is an example of an
most_funded.json	intermediary step between Crunchbase
rounds_month.csv	and the JSON data for D3
sankey.json	
states.json	
/images	Contains images for the project
...	
/js	
datascript.js	Place for temp. scripts, data cleaning
geomap_vis.js	Script for chapter two map vis
sankey_vis.js	Script for the Sankey vis in chapter 3
script.js	Main JS, contains fcts for next steps
table_vis.js	Script for the table vis in Ch. 4
/libs	Contains libraries
...	

6. Evaluation

The final step to create a compelling visualization was to put the pieces together. The result is quite interesting, and helped me answer the questions I had and learn more about startup funding:

<http://didjeridou.github.io/cs171-pr-ventures/>

6.1 Answers

Each visualization answered my questions, and all the detailed findings are available on the website, but here are the highlights:

When is the best time to fundraise and what are some important patterns to be aware of?

- January is the easiest time to close funding deals
- Since the 2008 crisis, there are 2 to 3 times more institutional investors funding startups
- More companies get seed money, but around the same number as before raise the next round

Where is my startup most likely to get funded?

- California; More deals are made in that state than in all other states
- Massachusetts and New York are the runners up
- For those preferring the south, Texas is the place

What are my chances of actually succeeding?

- Only 5-6% reach advanced funding rounds (Series C, D+)
- Even startups getting seed money are likely going to fail or never make it to the next round
- Those who raise the most money do so at different stages, in different industries and in different states (but mostly California)

6.2 Improvements

After starting over several times and drastically changing my approach, I really think the result is a compelling visualization. It is interesting to navigate, tells an informative story, and make use of a variety of visualizations that answer specific questions.

There is, however, a few things I would improve if I had more time:

- I would add **interactivity between the graphs and the findings**, such that when the user click on a finding, the graphs or charts switch to a specific view.
- **The colours** and order of elements is an important detail and I had trouble with the colours in the Sankey chart. If I had more time, I would properly unify the colours, using one predefined and meaningful colour palette.
- Finally, I would like to add the possibility to compare the North American data to the **Asian data**. More specifically, I am interested in understanding how the startup ecosystem works in **India, China, Japan and Singapore**.

6.3 Take-home

STARTUP-WISE

- A startup's chances of getting funded are better in January and December
- Getting funded does not mean a startup will succeed
- Consider moving to California
- Startup that reach Series A are three times as likely to succeed

VISUALIZATION-WISE

- Iterating often is good. It helps getting to a more thoughtful solution
- A good dataset is key; it is truly inspiring
- It is all about telling a good story

6.4 Conclusion

This project was really challenging, but I learned a lot. What we learn in class really becomes meaningful when we get the chance to work on a project of our own. I am really proud of this final project, and I will in fact keep working on it to add the previously mentioned improvements.

Thank you to the CS171 staff for this really fascinating class. I will never look at any visualization with the same eyes!