Päivi Ala-Risku
M.S. in Data Journalism, 2018-2019
Columbia University Graduate School of Journalism

Final report

I spent a year studying at Columbia University, New York, and completed a master's degree in Data Journalism in May 2019. In twelve months we learned how to code with Python, how to gather and analyze massive amounts of data, how to visualize the data and how to write long form stories with data.

This is a brief overview how the one year program was structured:

Summer 2018:

1) Data Foundations: Learning Python coding language, Pandas, web scraping, using command line.
2) Databases: Learning about SQL databases, scraping, creating our own project.
3) Data Studio: Creating and visualizing data projects on a weekly basis.
4) Reporting: Reporting on New York courts and writing on deadline and long form stories. (Interesting but maybe not that useful for me as I already have been working as a journalist for several years. The same applies to other courses offering basic journalistic skills: of course one can always learn more but as the schedule was so hectic I decided to put more effort on the coding courses)

Fall 2018:

1) Data & Computation and Innovation. Basically a sequel for the summer foundations course. We learned more about data visualization and learned how to code with JavaScript and D3 to create scrollytelling stories and interactive graphics.
2) Writing with Data. Writing exercises related to data: e.g. How to write about some statistics without using bigger numbers than ten.
3) Invesigative Techniques with Data. Learning about different kinds of databases and ways to find data. Mostly US related, but also some international data.
4) Law. Media laws in the United States.
5) Ethics. Journalists visited us and discussed about ethical consideration they deal with.
6) Master's project. Our master's project was a long form piece (5000 words) with a major data component. I decided to analyze nutrition content of different meals provided by several food companies. I also interviewed experts in nutrition and customers using the products.

Spring 2019:

1) Algorithms. Learning the basics about statistics and algorithms. How to report, investigate and write stories about algorithms: what to ask, what to pay attention to?
2) Computational Journalism. A bit repetitive course: basics in Python, Pandas, scraping. Group projects and hackaton. I also learned how to use optimization program.
3) Food Writing. Not a data related course but a course we could freely choose. I chose food writing and wrote many different kind of food stories published at www.nytable.com.
4) History. The history of Journalism in US and also other parts of world.
5) Master's Project. Continuing working on master's project with the help of bi-weekly meeting with our project instructor.

The year was full of work: I learned more than I could've imagined but I also slept less than ever in my life. I spent weekdays at university from nine to five and in the evening, after the kids had fallen asleep I started doing homework for 3-4 hours a day, on average. Altogether, I would estimate I worked at least 60-70 hours a week.

Next, I'll go through what I learned about the trend in data journalism in the U.S. currently

Data Journalism in the U.S. now

1. Scrollytelling

Scrollytelling is all the rage in the U.S. right now. Scrollytelling means online stories where graphics and illustrations change when user scrolls their screen. No clicking or other input from reader is required. The reasoning behind this goes: "the journalist should do the job of analyzing data and present just the most important things to the reader and not to burden them with too many choices". And yes, scrollytelling stories are often informative, entertaining and just plain awesome. Still, I think, it's sometimes even more informative and entertaining, sometimes even obligatory, to make the reader click or offer some form of input to serve the reader better. For many data sets the journalist cannot simply know and dictate what's the most important and valuable piece of information for every single reader. So in my opinion, the journalists job is not only analyze and present the bigger picture but also, when possible, make the data easily available also for the reader to explore what interests them most.

(Example of scrollytelling:)
Could Trump Really Deport Millions of Unauthorized Immigrants?
https://www.nytimes.com/interactive/2016/11/29/us/trump-unauthorized-immigrants.html

(Example of allowing the reader explore the data:)
Vain 11 prosenttia synnyttää ilman kipulääkkeitä - Lue sinulle räätälöity juttu siitä, miten lapsesi tulisi maailmaan. https://dynamic.hs.fi/2017/synnytyssairaalat/

2. Text analysis

Analyzing massive amount of texts with the help of computer is very popular among data journalists in the U.S. For example, Los Angeles Times created an algorithm that analyzed thousands of police reports and found out that serious crimes were often labeled wrongly as minor offences.

Text analysis is, of course, much easier in English language than it is in Finnish. Different language libraries for coders are abundant in English where as in Finnish language there simply is not that many tools available. Inflections in Finnish language add to the challenges of text analysis in Finland. Then again: the amount of text data in Finland can sometimes be smaller which makes it easier to be read completely by humans i.e. journalist.

(Here's an example of large scale computer assisted text analysis:)
LAPD underreported serious assaults, skewing crime stats for 8 years
https://www.latimes.com/local/cityhall/la-me-crime-stats-20151015-story.html

(Here's an example of story where journalists read all defamation cases in Finland, but as there have been only 867 cases it was relatively easy to read them all manually:)

Solvaukset syyttäjän pöydällä
https://www.hs.fi/sunnuntai/art-2000006028892.html

3. Holding algorithms accountable:

Journalists are covering algorithms a lot in the U.S. They analyze how the algoritgms work, are they working as they should and what does it mean for an algorithm to work in a way they work: journalists are trying to hold algorithms accountable.

For example ProPublica investigated a software that is used in courts to do risk assessments of criminals. The system doesn't ask for race but yet ProPublica found out that it's biased against blacks. Or is it? Washington Post then wrote an article arguing that it depends on what you're looking for. And the most important lesson is that they were both right: with the current demographics it is impossible to create an algorithm that would be neutral in all ways. So the lesson is: computing, algorithms and coding is never completely objective and neutral: people make the decisions what to optimize for and there can always be intentional and unintentional biases.

Examples:
“Machine Bias. There’s software used across the country to predict future criminals. And it’s biased against blacks.“
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

“A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear.”
https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.1e6ed6e00c30

4. Scraping

Scraping, i.e. collecting data from a web page/site with the help of some code or program, is a popular way to collect data for journalistic projects in the U.S. For example, New York Times scraped data about runners from a runners’ site to investigate whether a specific type of Nike shoes really makes you faster (the answer is, surprisingly, that they actually seem to make you faster).

The problem is, however, that laws regulating scraping are murky and unclear making large scale scraping project often impossible. But there are ways to prevent possible problems. For example, ProPublica has run a continuously updating story about Facebook ads. The data is collected by readers who install a small browser plugin that records the ads they see on Facebook, and then send that information to ProPublica. That way, ProPublica is not scraping it by themselves, and the individual readers are free to do whatever they wish with their data, including donating it to ProPublica.

Examples:
Facebook Political Ad Collector. How Political Advertisers Target You
https://projects.propublica.org/facebook-ads/?lang=en-US

Nike Says Its $250 Running Shoes Will Make You Run Much Faster. What if That’s Actually True?
https://www.nytimes.com/interactive/2018/07/18/upshot/nike-vaporfly-shoe-strava.html

5. Specific themes: Privacy & Twitter & Popular Culture

Some themes appear to be more popular than others among data journalists in the U.S. For example, there are lots of stories about privacy in the online world, Twitter and all kinds of popular culture themes such as music and football.

Examples:
(Great example of text analysis, scrollytelling and privacy:)
We Read 150 Privacy Policies. They Were an Incomprehensible Disaster.
https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html
(Great example of reporting related to Twitter. A pulitzer prize winner and one of its writer is Mark Hansen, who was my professor in Computational Journalism class.)
The Follower Factory:
https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html
(Great example of scrolly telling and data journalism covering popular culture)
The unlikely odds of making it big
https://pudding.cool/2017/01/making-it-big/

Closing words

The year I spent at Columbia University was one of the best years I have ever had. I'm grateful for the Helsingin Sanomat Foundations for making it possible and for my spouse and my kids for being the best companions I could ever hope for in our great adventure in New York.