

Lecture 2 | Motivation

Max Pellert

IS 616: Large Scale Data Analysis and Visualization

Data Visualization?

Why do you need that?

Visual communication is important in all areas: industry, science, ...

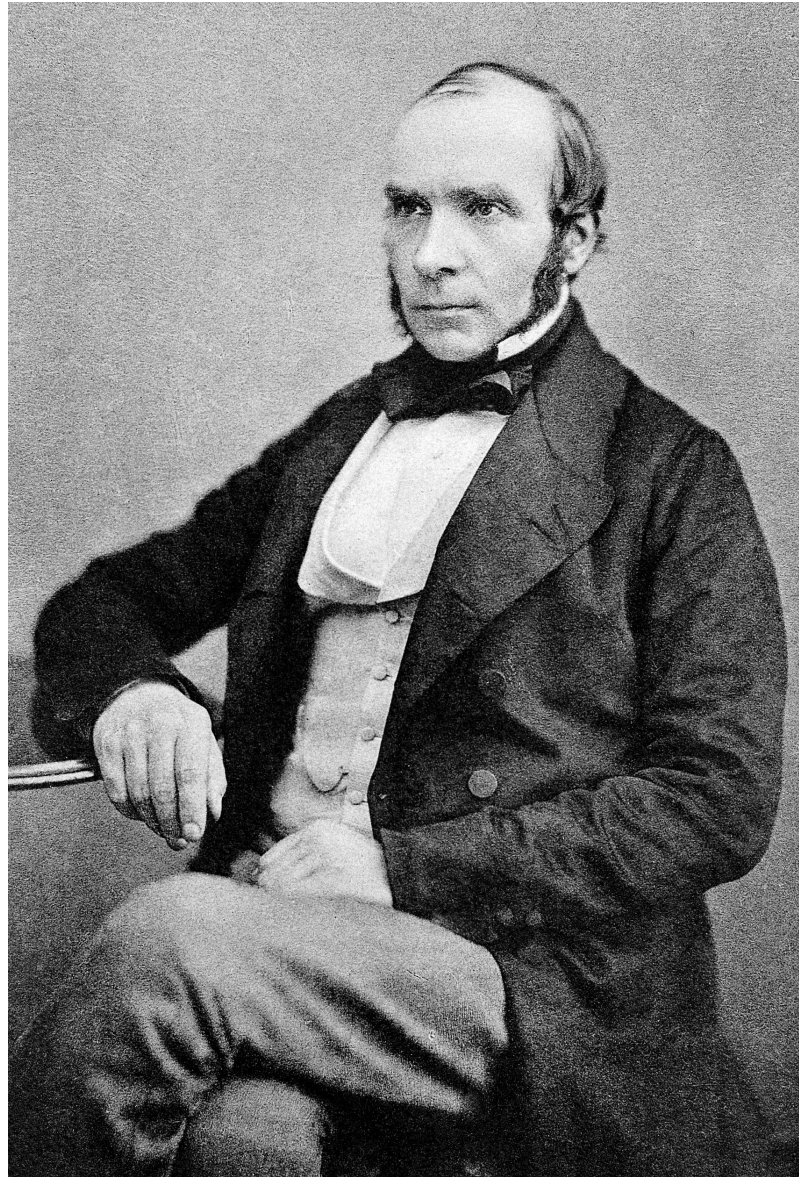
But also for yourself, to learn and to remember



In the context of science especially: to get new insights

Especially important for very large data sets

But keep in mind: pure eyeballing can also mislead (connected to the problem of induction, “reasoning after the facts”)

A tale of epidemics...



John Snow (15 March 1813 – 16 June 1858^[1]) was an English physician and a leader in the development of [anaesthesia](#) and [medical hygiene](#). He is considered one of the founders of modern [epidemiology](#), in part because of his work in tracing the source of a [cholera outbreak in Soho, London, in 1854](#), 
 Snow's findings inspired the adoption of anaesthesia as well as fundamental changes in the water and [waste systems of London](#), which led to similar changes in other cities, and a significant improvement in general [public health](#) around the world.^[2]

The Cholera Epidemic in London, 1854

Cholera broke out in the Broad Street area of central London on the evening of August 31, 1854

Causes and possible interventions were unclear

“Miasma theory” (“pollution”) held that “bad air” or “night air” was responsible

Miasma could emanate from rotting organic matter, for example from burying grounds of plague victims from two centuries earlier

The Cholera Epidemic in London, 1854

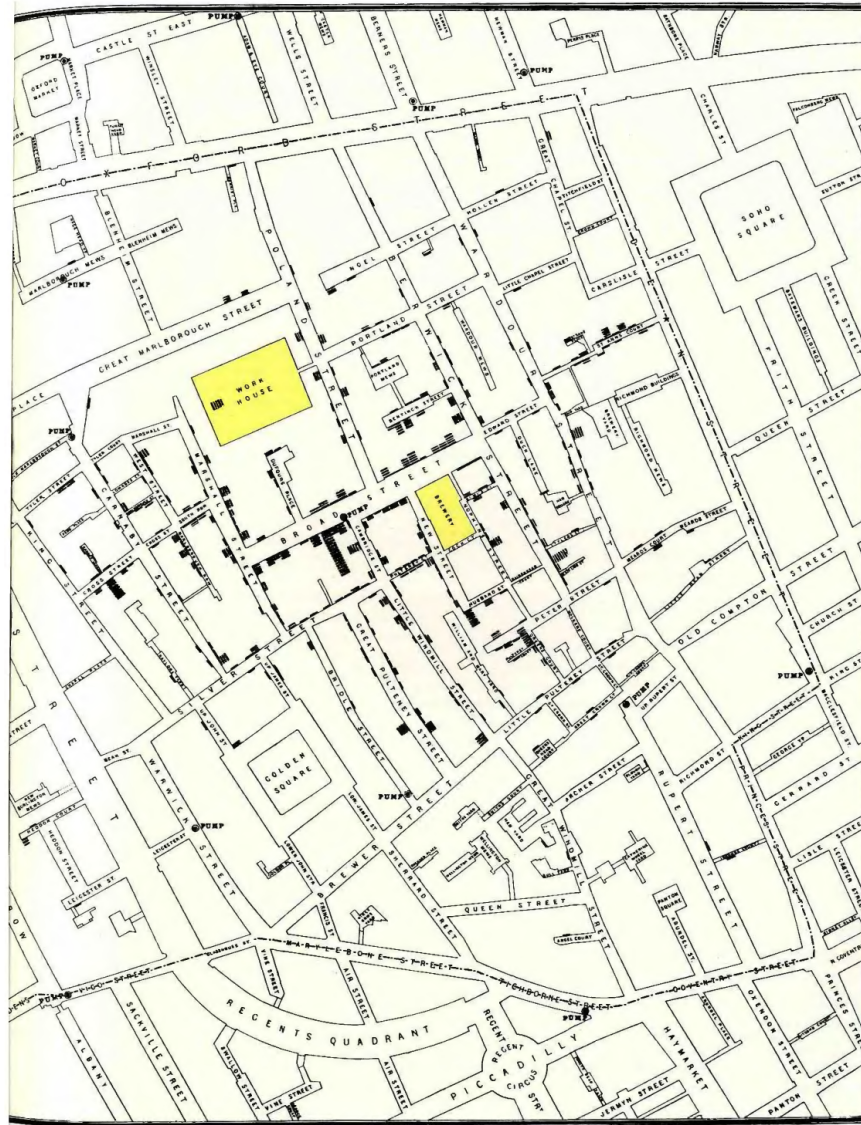
John Snow who investigated earlier epidemics had a different theory of the causes

He wasn't successful in verifying his suspicions directly

So he tried an indirect strategy to find the causes: **Data Visualization**

He obtained a list of 83 deaths from cholera (including the addresses of the victims)

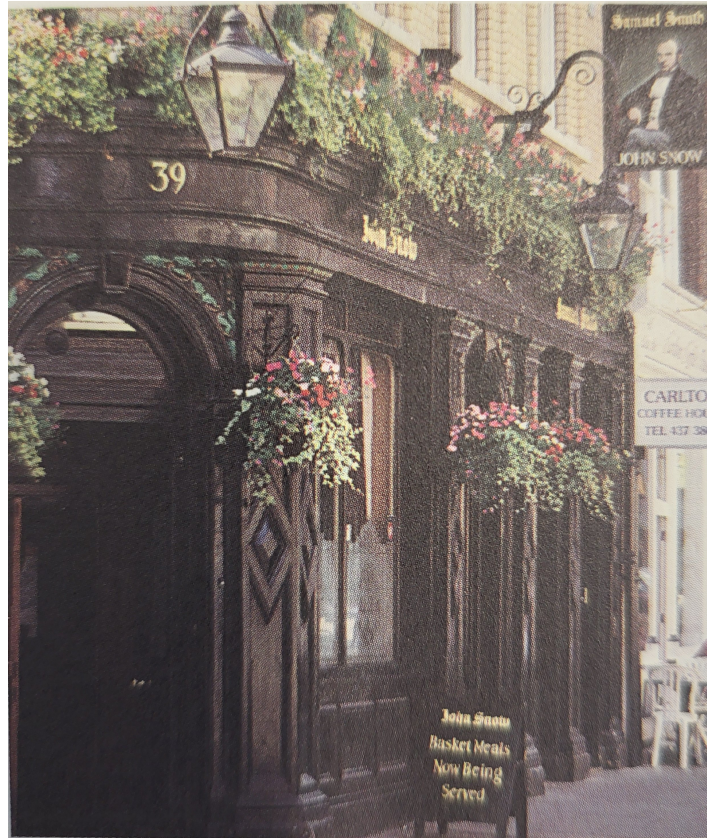
And plotted them on the map of the part of London that was affected by Cholera

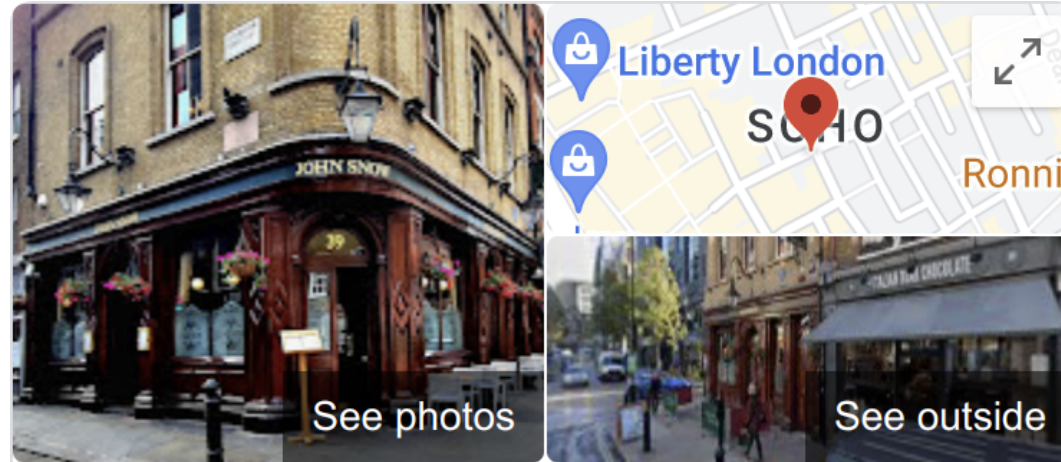




Intervention

John Snow had the handle of the pump removed





John Snow

Website

Directions

Save

4,2 ★★★★★ 1.577 Google reviews ⓘ

€ · Pub

Dark-wood saloon bar serving Yorkshire ales, named after doctor who traced London cholera outbreak.

Causes

The epidemic soon ended

Revolutionized our understanding of transmission processes:
germ theory of disease

In 1886: discovery of the bacterium *vibrio cholerae*

What actually made the water impure and dangerous?

Industrial Revolution: rapid urbanization but no
infrastructure

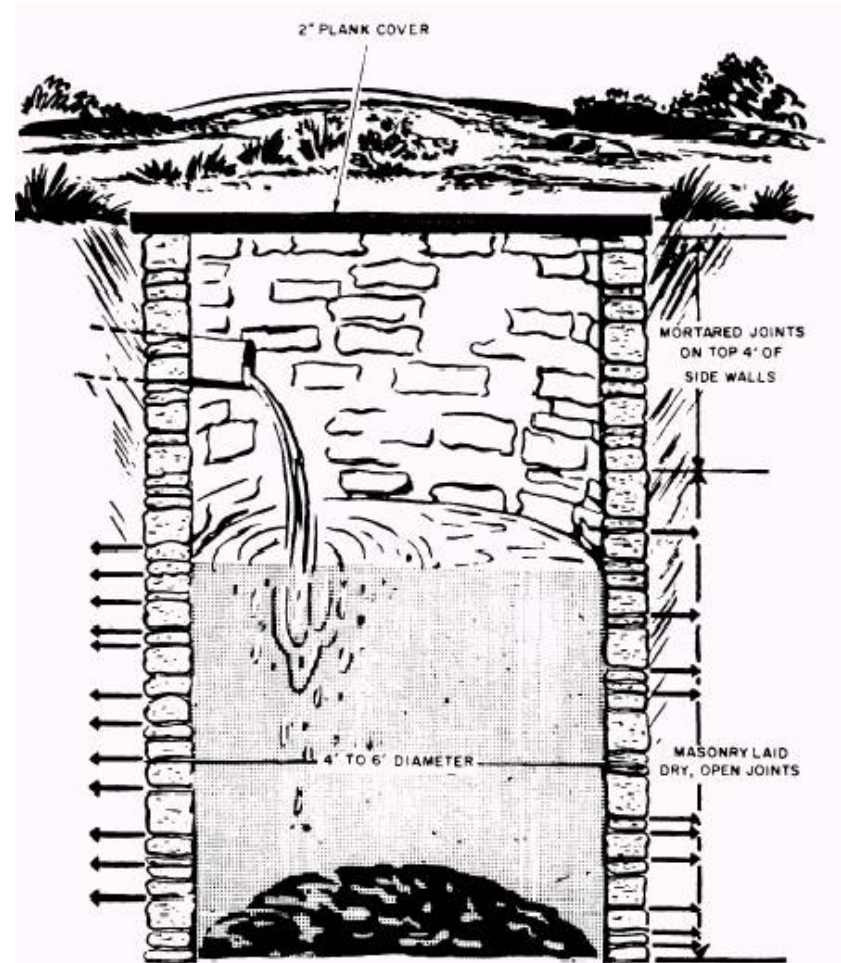


“Nightmen”

London Labour and the London Poor is a work of **Victorian** journalism by **Henry Mayhew**. In the 1840s, he observed, documented and described the state of working people in **London** for a series of articles in a newspaper, the ***Morning Chronicle***, which were later compiled into book form.

<https://www.gutenberg.org/files/60440/60440-h/60440-h.htm>





“Leaching cesspools” (Illustration)

What makes this investigation so strong?

Providing context, with the right graphic display

From a one-dimensional temporal ordering into a two-dimensional spatial comparison

Quantitative comparisons: Why did no workers at the brewery so close to the pump die?

They are allowed to drink a daily quantity of beer. The owner of the brewery believes “they do not drink water at all”

What makes this investigation so strong?

Considering alternative explanations and contrary cases

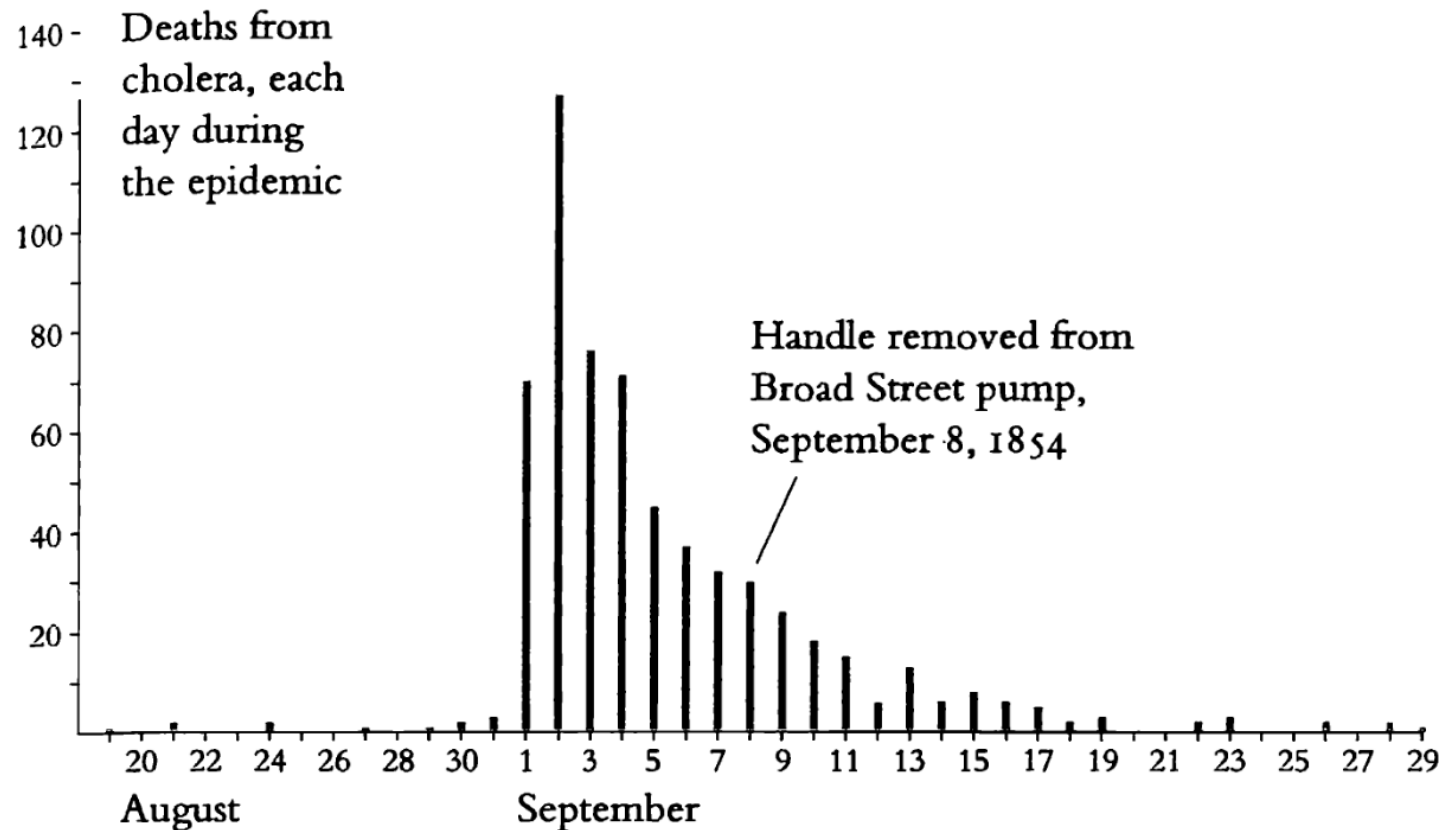
Seemingly unconnected cases of cholera in other areas reveal connections: a cabinet-maker works near the pump, a girl goes to school close-by

Assessment of possible errors in the numbers reported in graphics

“An area of the map may be free of cases merely because it is not populated” -> whole area very densely populated

A Note

Evidence of the effect of the intervention actually not that clear cut:



You could also aggregate the data differently, to artificially boost the story:

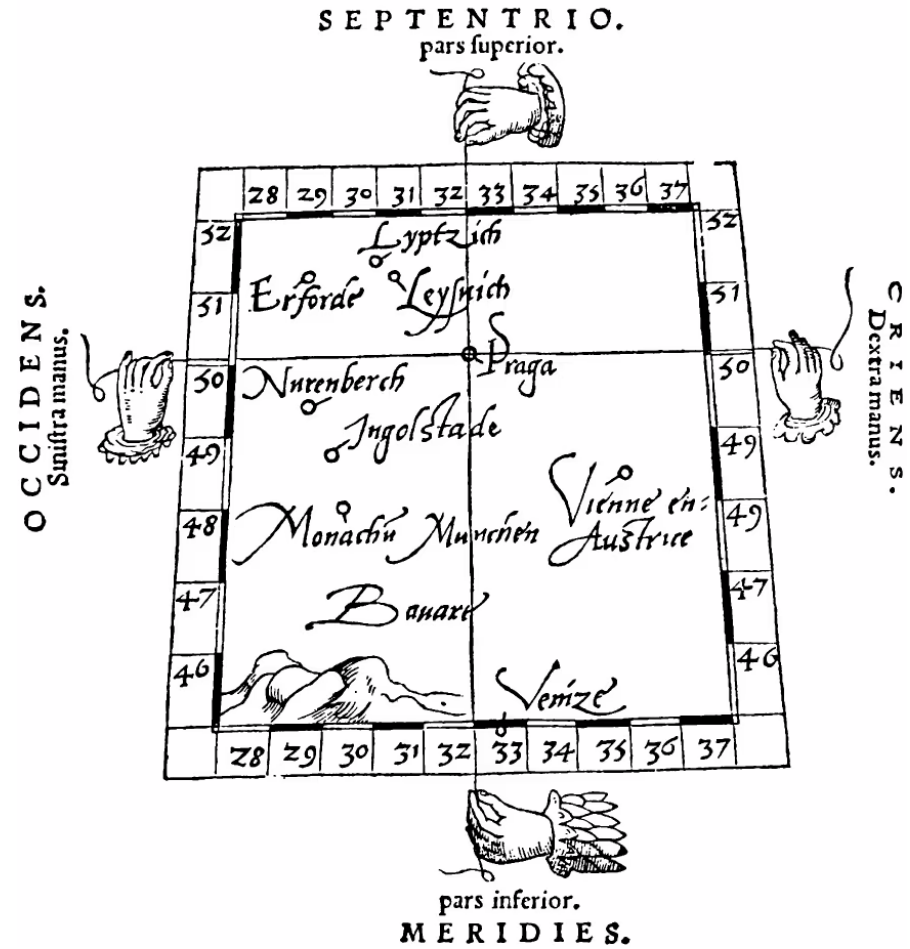


(Tufte calls this “chart-junk” as we will see later in the course)

John Snow (15 March 1813 – 16 June 1858^[1]) was an English physician and a leader in the development of [anaesthesia](#) and [medical hygiene](#). He is considered one of the founders of modern [epidemiology](#), in part because of his work in tracing the source of a [cholera outbreak in Soho, London, in 1854](#), which he curtailed by removing the handle of a water pump.^[*citation needed*] Snow's findings inspired the adoption of anaesthesia as well as fundamental changes in the water and [waste systems of London](#), which led to similar changes in other cities, and a significant improvement in general [public health](#) around the world.^[2]

Ecce formulam, usum, atque

structuram Tabularum Ptolomæi, cum quibusdam locis, in quibus studiosus Geographiæ se satis exercere potest.



Enriching visual displays

From a visualization point of view, John Snow actually used a very simple mechanism

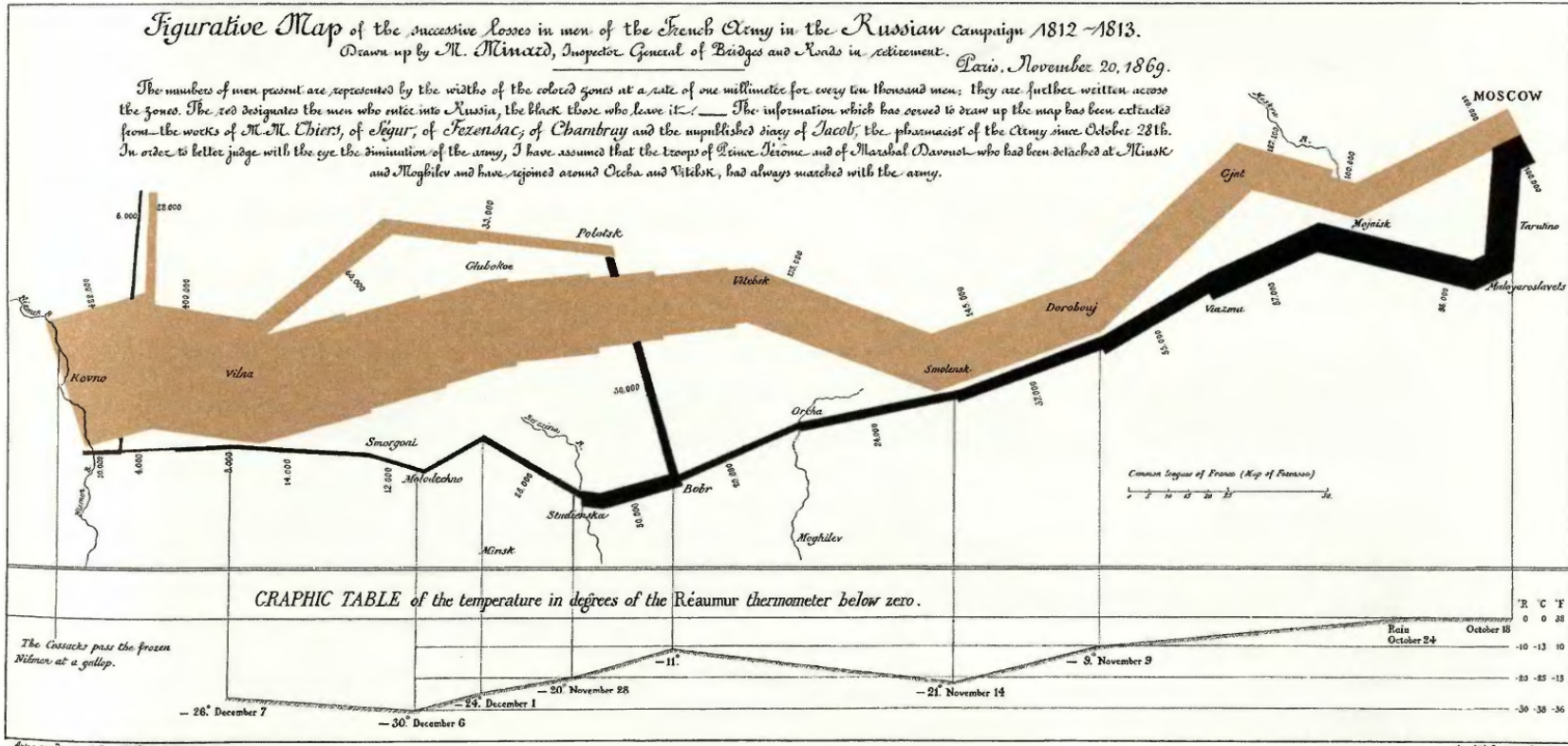
Marking deaths on a map

Going beyond that, graphics can really excel at condensing and bringing much disparate information together to make it comparable

Figurative Map of the successive losses in men of the French Army in the Russian Campaign 1812-1813.

Drawn up by M. Minard, Inspector General of Bridges and Roads in retirement. Paris, November 20, 1869.

The numbers of men present are represented by the widths of the colored zones at a rate of one millimetre for every ten thousand men; they are further written across the zones. The red designates the men who enter into Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M. M. Chiers, of Ségur, of Fezensac, of Chambray and the unpublished diary of Jacobi, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Louis Jérôme and of Marshal Davoust who had been detached at Minsk and Moghilev and have rejoined around Oecha and Vitelsk, had always marched with the army.



Avant par Regnier, à Paris, 54 Mars 59 69 à Paris.

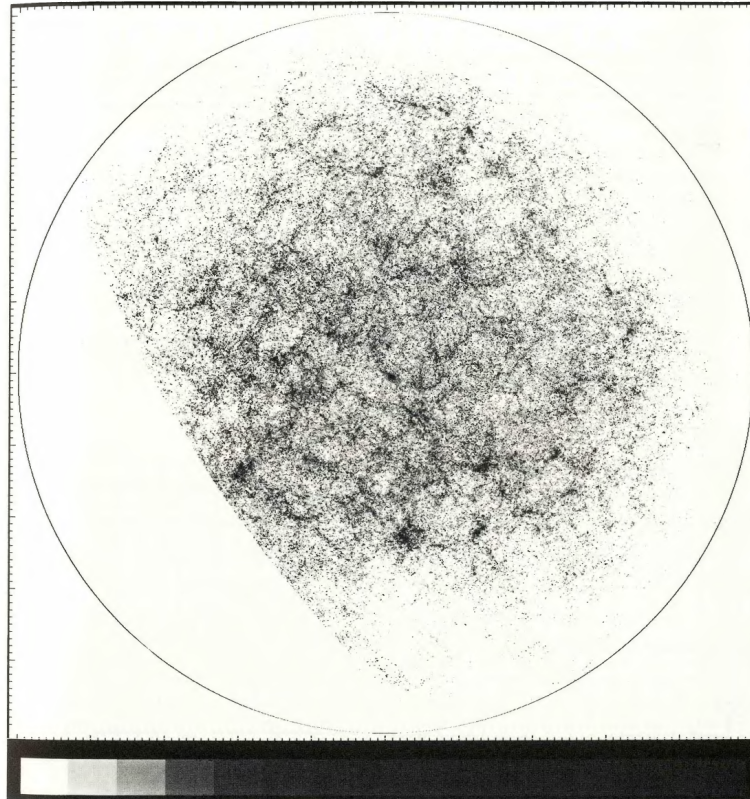
Imp. Lit. Régim. Bourd.

The first is the classic of Charles Joseph Minard (1781-1870), the French engineer, which shows the terrible fate of Napoleon's army in Russia. Described by E. J. Marey as seeming to defy the pen of the historian by its brutal eloquence,¹² this combination of data map and time-series, drawn in 1869, portrays a sequence of devastating losses suffered in Napoleon's Russian campaign of 1812. Beginning at left on the Polish-Russian border near the Niemen River, the thick tan flow-line shows the size of the Grand Army (422,000) as it invaded Russia in June 1812. The width of this band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then sacked and deserted, with 100,000 men. The path of Napoleon's retreat from Moscow is depicted by the darker, lower band, which is linked to a temperature scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. *Six* variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. At upper right we see Minard's French original, which was printed as a two-color lithograph in the form of a small poster. And at lower right, our English translation.

It may well be the best statistical graphic ever drawn.

What about today's "large-scale" data?

Often even more powerful in uncovering hidden phenomena!



The Follower Factory

A very good example of data journalism

Put the spotlight on identity theft and fake accounts in social media

<https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html>

The New York Times



The Follower Factory

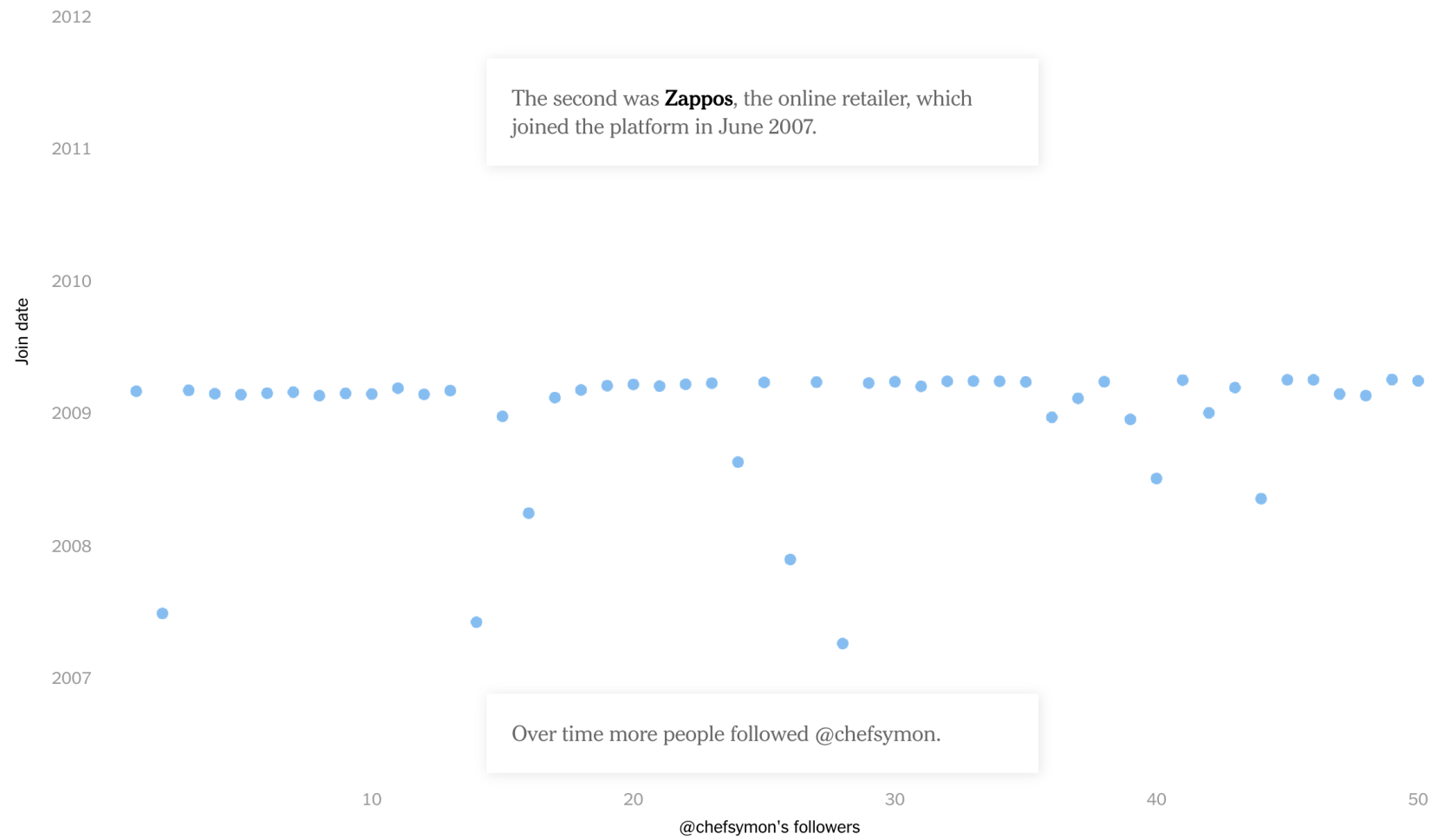
Everyone wants to be popular online.
Some even pay for it.
Inside social media's black market.

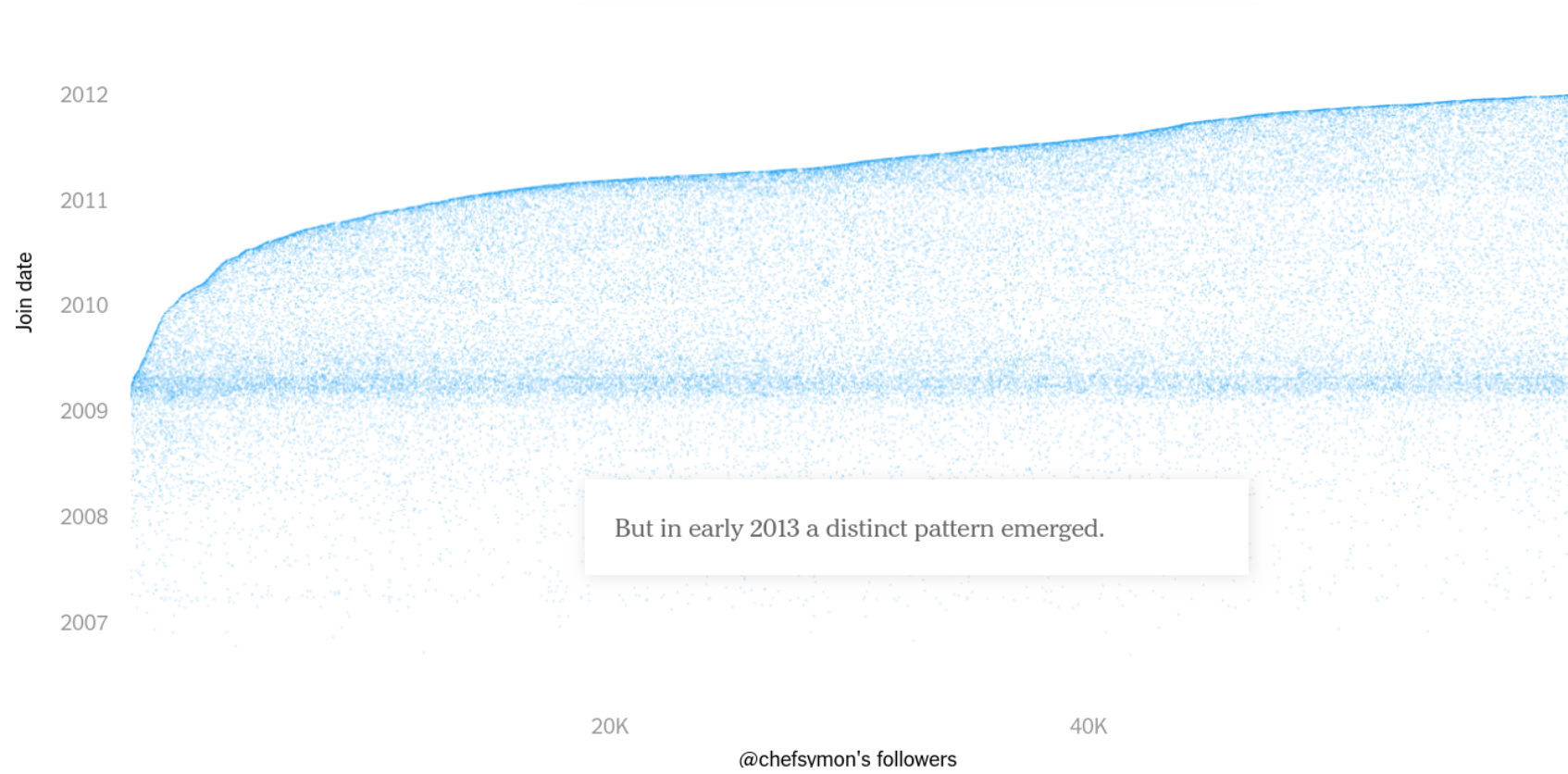
By NICHOLAS CONFESSORE, GABRIEL J.X. DANCE,
RICHARD HARRIS and MARK HANSEN

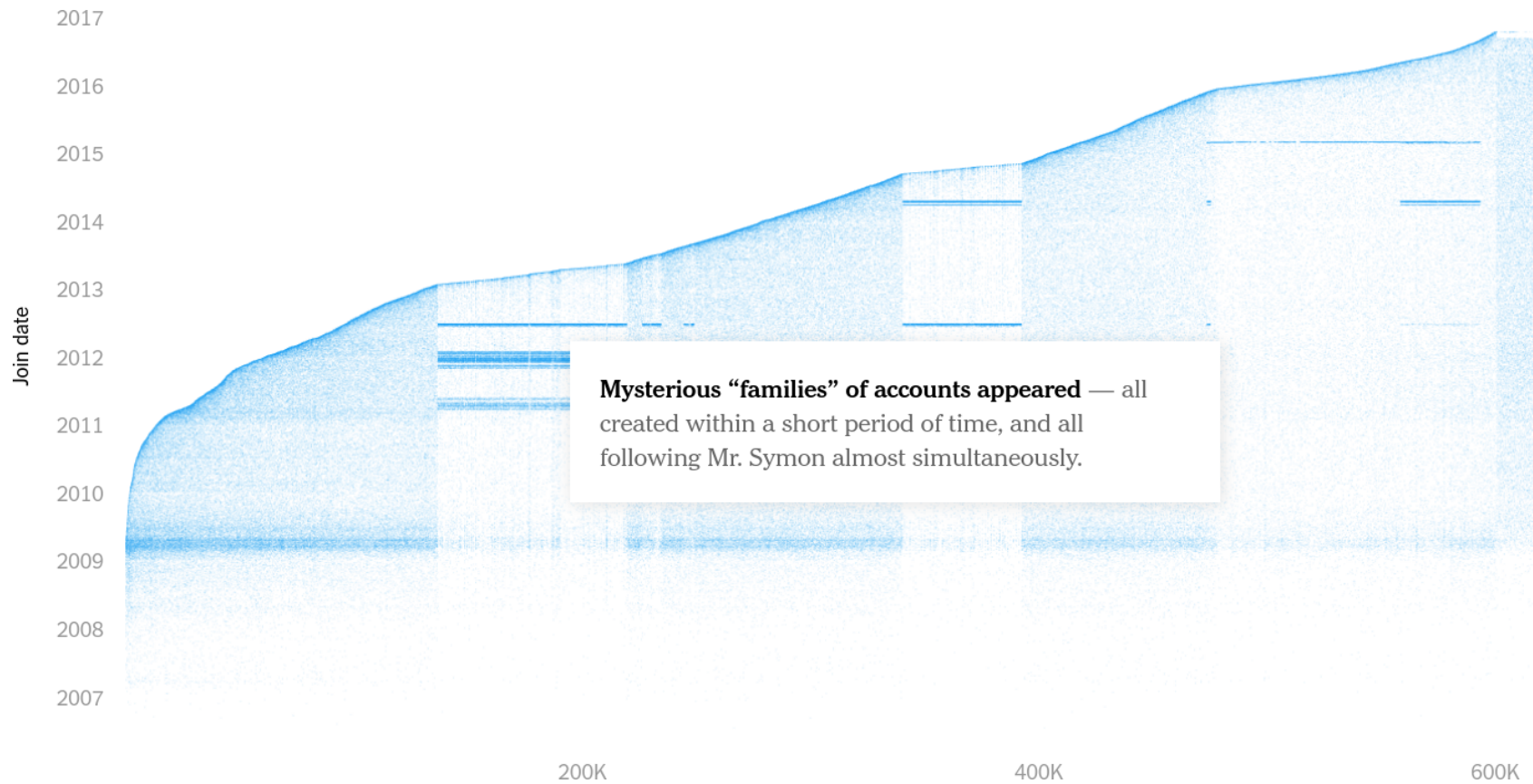
JAN. 27, 2018

[Leer en español](#)

Kathy Ireland, model and entrepreneur







Mysterious “families” of accounts appeared — all created within a short period of time, and all following Mr. Symon almost simultaneously.

The Follower Factory

Leaders | Technology

The data deluge

Businesses, governments and society are only starting to tap its vast potential



<https://www.economist.com/leaders/2010/02/25/the-data-deluge>

The Follower Factory

Paradoxically, visualization techniques can sometimes profit from “too much data”

Allows you to see through the deluge sometimes

But you have to be careful, visualizations of big data can mislead as well as visualization of small data

We will see examples of that in the course

DS **r/datascience** • 2 days ago
by Every-Eggplant9205

Join



R vs Python - detailed examples from proficient bilingual programmers

As an academic, R was a priority for me to learn over Python. Years later, I always see people saying "Python is a general-purpose language and R is for stats", but I've never come across a single programming task that couldn't be completed with extraordinary efficiency in R. I've used R for everything from big data analysis (tens to hundreds of GBs of raw data), machine learning, data visualization, modeling, bioinformatics, building interactive applications, making professional reports, etc.

Is there any truth to the dogmatic saying that "Python is better than R for general purpose data science"? It certainly doesn't appear that way on my end, but **I would love some specifics for how Python beats R in certain categories as motivation to learn the language.** For example, if R is a statistical language and machine learning is rooted in statistics, how could Python possibly be any better for that?

↑ 454 ↓

💬 137

🔗 Share

🚩 Report

https://www.reddit.com/r/datascience/comments/16dk5b6/r_vs_python_detailed_examples_from_proficient/

From my experience **Python excels** (vs R) when you move to writing production-grade code:

- in my experience base Python (dicts, lists, iterating strings letter by letter) are much faster than base types in R
- better OOP system than R's set of S3/S4/R6
- function decorators
- context managers
- asynchronous i/o
- type hinting and checking (R has a package typing that has something along these lines but nowhere to the level what Python has in terms of say Pydantic and mypy)
- far more elaborate set of linting tools, e.g. black and flake8 trump anything in R
- new versions and features coming far more quickly than R
- data orchestration/automation tools that work out of the box, e.g. Airflow, Prefect (stupid easy learning curve, slap few decorators and you have your workflow)
- version pinning, e.g. pyenv, poetry, basically reproducible workflows
- massive community support, unlike R, Python doesn't rely on one company (Posit) and bunch of academics to keep it alive.
- FAANG companies have interest in developing not only Python packages but language itself, even more so with Global Interpreter Lock removal
- web scraping, interfacing with various APIs even as common as AWS is a lot smoother in Python
- PySpark >>> SparkR/sparklyr
- PyPI >>> CRAN (CRAN submission is like a bad joke from stone age, CRAN doesn't support Linux binaries(!!!))

R excels in maybe lower number of other places, typically statistical tools, specific-domain support (e.g. bioinformatics/comp bio) and exploratory data analysis, but in things it is better it is just so *good*:

- the number of stats packages is far beyond anything in Python
- the number of bioinformatics packages is FAR beyond Python (especially on Bioconductor)
- tidyverse (dplyr/tidyr especially) destroys every single thing I tried in Python, pandas here looks like a bad joke in comparison
- delayed evaluation, especially in function arguments, results in some crazy things you can do wrt metaprogramming (e.g. package rlang is incredible, allows you to easily take the user provided code apart, supplement it, then just evaluate it in whatever environment you want... which I am sure breaks bunch of good coding practices but damn is it useful)
- data.table syntax way cleaner than polars (again thanks to clever implementation of tidy evaluation and R-specific features)
- Python's plotnine is good, but ggplot2 is still king - the number of additional gg* packages allows you to make some incredible visualizations that are very hard to do in Python
- super-fluid integration with RMarkdown (although now Quarto is embracing Python so this point may be moot)
- even though renv is a little buggy in my experience, RStudio/Posit Package Manager is fantastic
- RStudio under very active development and IDE for exploratory work is in some specific ways better than anything for Python including VSCode (e.g. it recognizes data.frame/data.table/tibble contexts and column names and previews are available via tabbing)

Acknowledgements

https://www.gutenberg.org/files/60440/60440-h/60440-h.htm#i_image23

<https://www.reddit.com/user/Useful-Possibility80/>