Welcome!

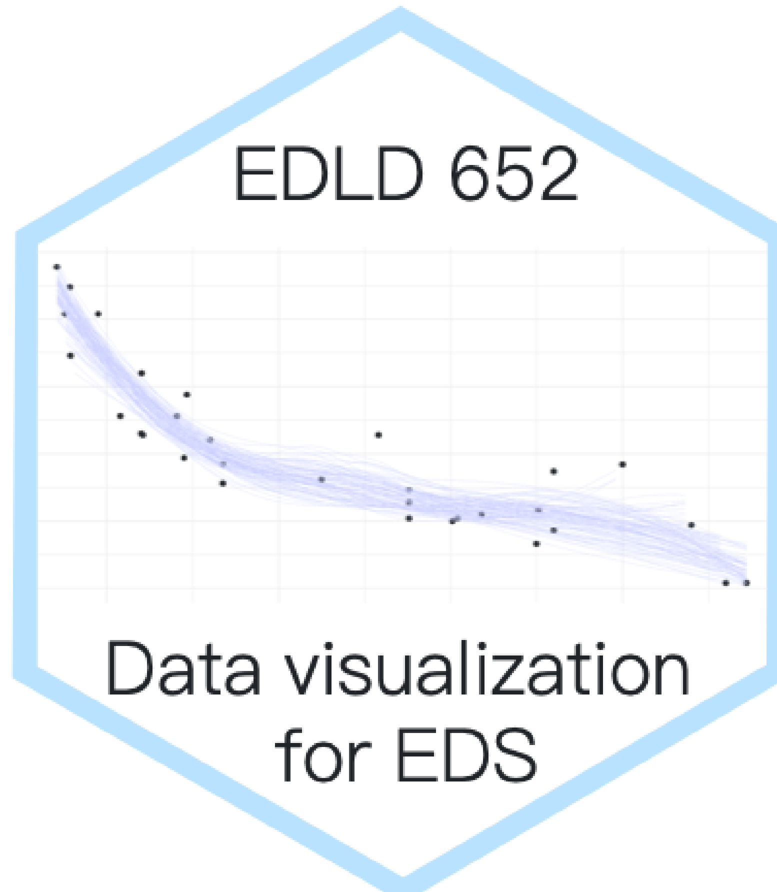
An overview of the course

Daniel Anderson

Week 1, Class 1

Agenda

- Getting on the same page
- Syllabus
- A little bit of git/GitHub



whoami

- Research Assistant Professor: Behavioral Research and Teaching
- Dad (two daughters: 8 and 6)
- Pronouns: he/him/his
- Primary areas of interest: 💗💗 R💗💗, computational research, achievement gaps, systemic inequities, and variance between educational institutions

whoisyou?

- Introduce yourself
- Why are you here?
- What pronouns would you like us to use for you for this class?
- What was one thing you did not related to academic work over winter break?

A few class policies

- Be kind
- Be understanding and have patience, with others and yourself
- Help others whenever possible

Truly the most important part of this class. Important not just in terms of decency, but also in your learning, and most importantly, for equity.

A bit more specific

Normally I would have information here about welcoming kids into class.

Because we're virtual, that part is both easier and harder.

If you need to not attend class, or a portion of class, for any reason, that is fine.

Ideally you would let me know ahead of time. But we're in the middle of a pandemic and life is cray. Please try to contact me beforehand. If this isn't possible, please check in with me after.

Last intro thing

- I'm here for you
- We won't have specific office hours, but know I'm always willing to meet
- This course, like all in the sequence, can be difficult. Don't suffer in silence. Don't do this alone.

Syllabus

Course Website(s)

website repo

Home Schedule Assignments Syllabus Tags Class rep

Data Visualization

for educational data science

Welcome to the second course in the Educational Data Science Specialization taught at the University of Orego
the College of Education. This course will be taught through R, a free and open-source statistical computing en
and will provide students with the foundational principles and practice of data visualization, particularly as appli
scientific and technical data. We will have weekly lectures, covering a wide variety of topics including human pe
color theory, and principles of visual design. We will also cover mediums for communication across diverse audi
an emphasis on different web applications. Weekly hands-on laboratory sessions provide students the opportu
the lecture material into practice.

Materials

- Nearly everything will be distributed through the repo and through the website.
- Please clone the repo now, if you haven't already. Pull each week for the most recent changes.
- We'll use Canvas for grading, and that is essentially it.

R Markdown notes

- These slides were produced with **{xaringan}**, an R Markdown variant. I encourage you to try it out and use it for your final project presentation.

- The website was also produced with R Markdown (sort of)

  - It's a **{blogdown}** website with some custom CSS and Hugo shortcodes

- This course is not just about data viz, but also mediums for communication. This includes websites and data dashboards among other possibilities.

My assumptions about you

I assume you

- Understand the R package ecosystem (how to find, install, load, and learn about them)

- Can read "flat" (i.e., rectangular) datasets into R

  - I don't care what you use, but you should be using RStudio Projects & the {here} package
  - See Jenny Bryan's blog post for why.

- Can perform basic data wrangling and transformations in R, using the tidyverse
    - Leverage appropriate functions for introductory data science tasks (pipeline)
    - "clean up" the dataset using scripts and reproducible workflows
- Use version control with R via git and GitHub
- Use R Markdown to create reproducible dynamic reports

Learning objectives

- Transform data in a variety of ways to create effective data visualizations
- Understand and fluently apply different types of data joins
- Understand best practices in data visualization
- Customize ggplot2 graphics by reordering factors, creating themes, and applying ggthemes
- Create an online data visualization portfolio using distill and/or flexdashboards to demonstrate key learning

Examples

Below are some links to final projects from students who took this class last year.

Dashboards

- Alexis Adams-Clark
- Brendan Cullen
- Ouafaa Hmaddi

Blog post

- Teresa Chen
- Ashley Miller
- Karlena Ochoa

Weekly learning objectives

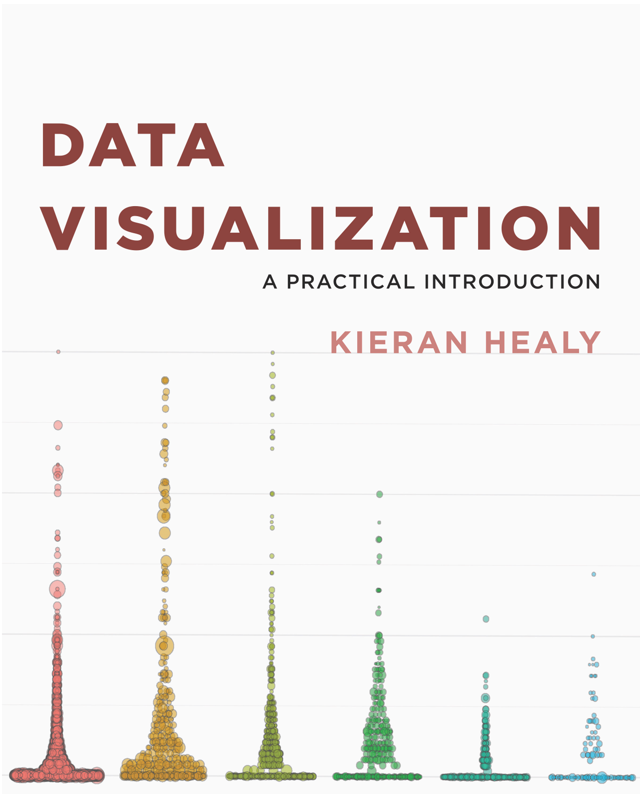
Provide you a frame for what you should be working to learn for that specific week.

This week's objectives

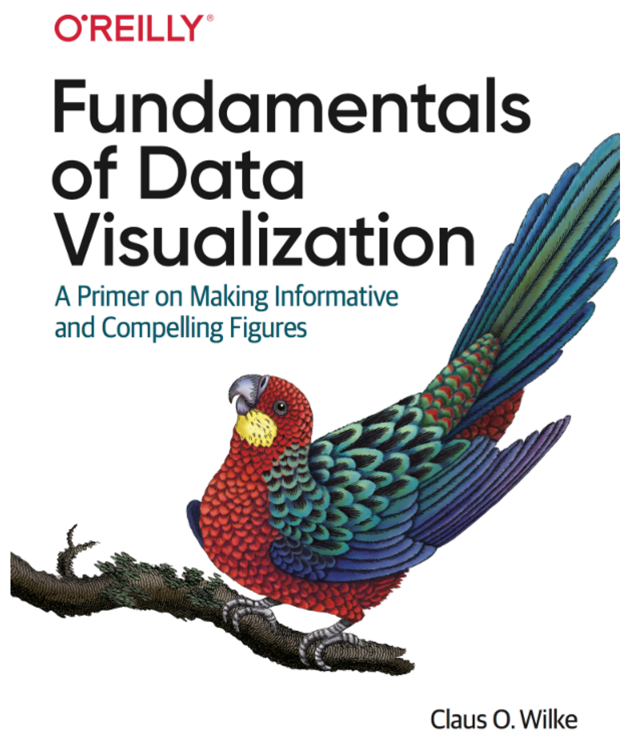
- Understand the requirements of the course
- Understand the requirements of the final project
- Be ready to go with *git* and GitHub

Required Textbooks (free)

Healy

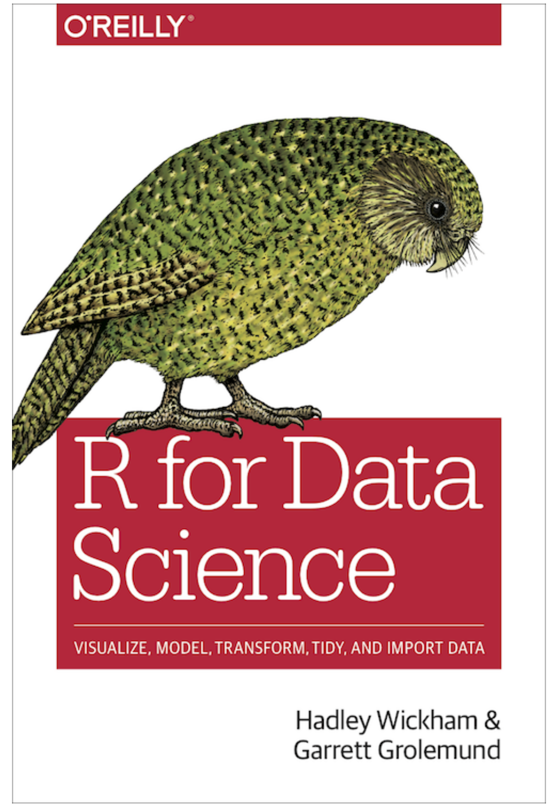


Wilke



Other books (also free)

Bryan



Wickham & Grolemund

Another resource

See the current draft here. Please read Chapter 8 before next class.

Social Data Science with R

Daniel Anderson

Brendan Cullen

Ouafaa Hmaddi

2020-12-24

Preface

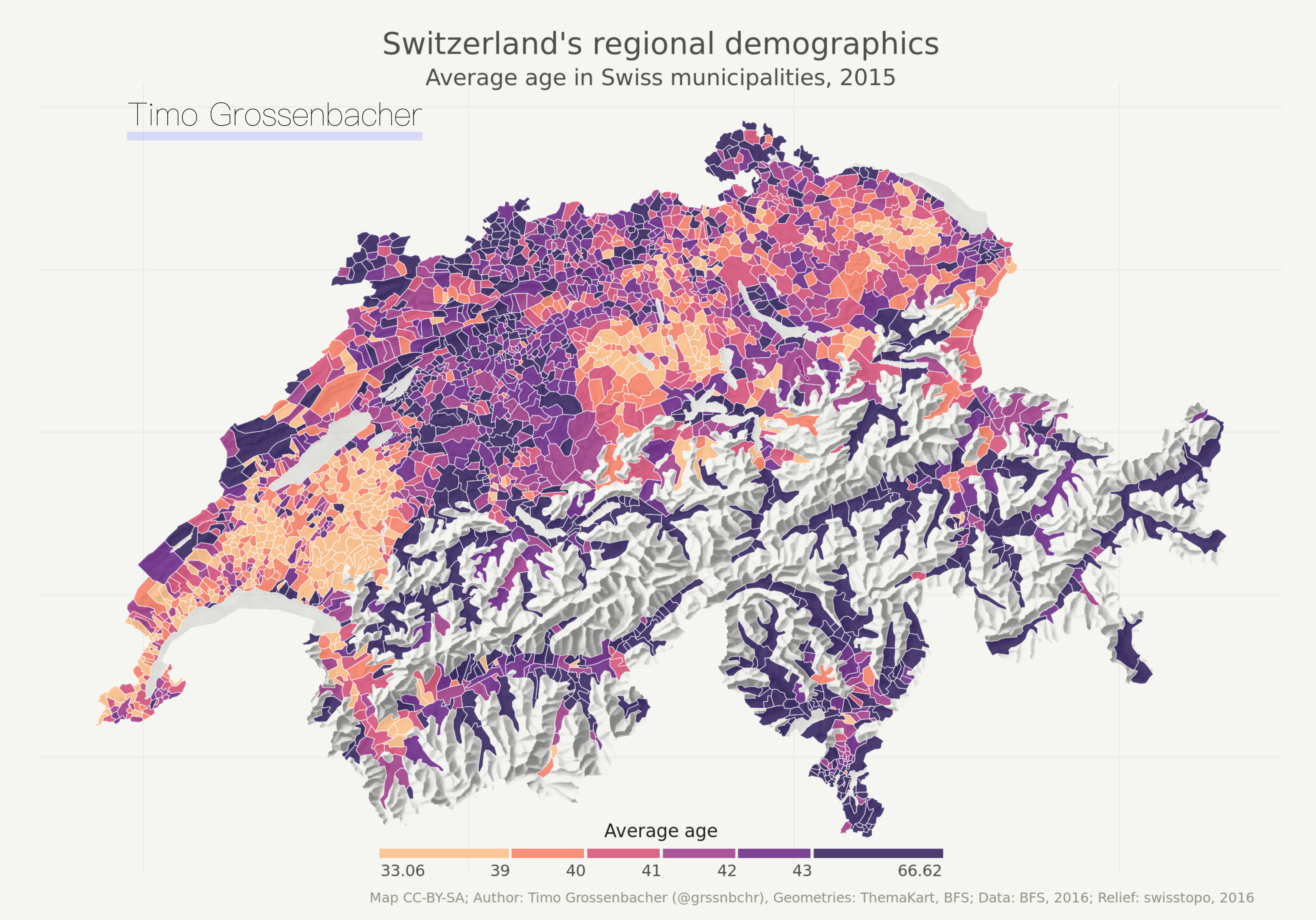
Extra credit opportunity

10 points: Deep dive into a topic not covered by the course

Some options

- Geographic data (we'll have an intro, but there's a ton here and we won't really do it justice)
- Network data
- Text data
- DAGs
- Flow data (e.g., alluvial diagrams)
- Relational data (SQL & friends)
- Interactive plots
- Animated plots

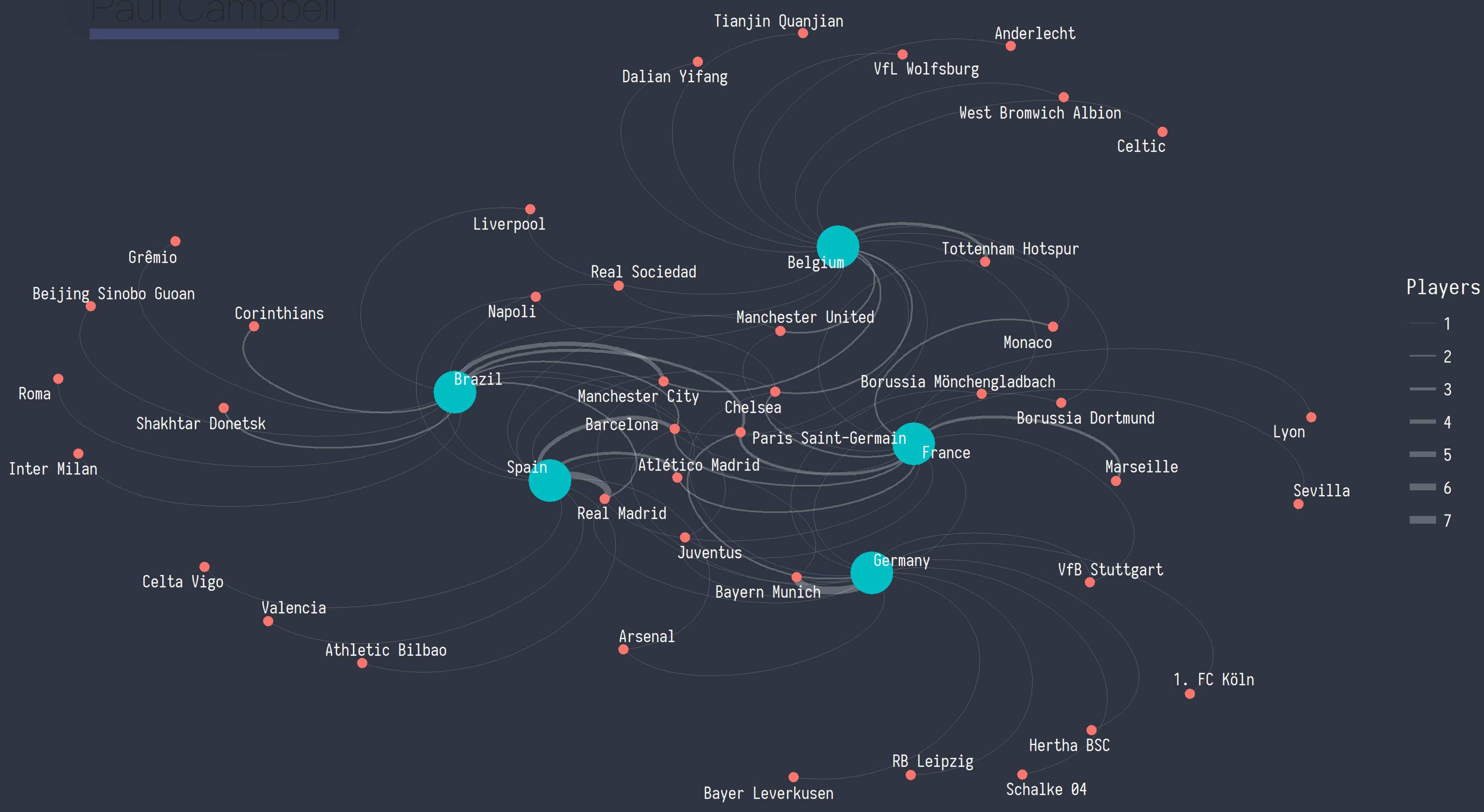
Some examples



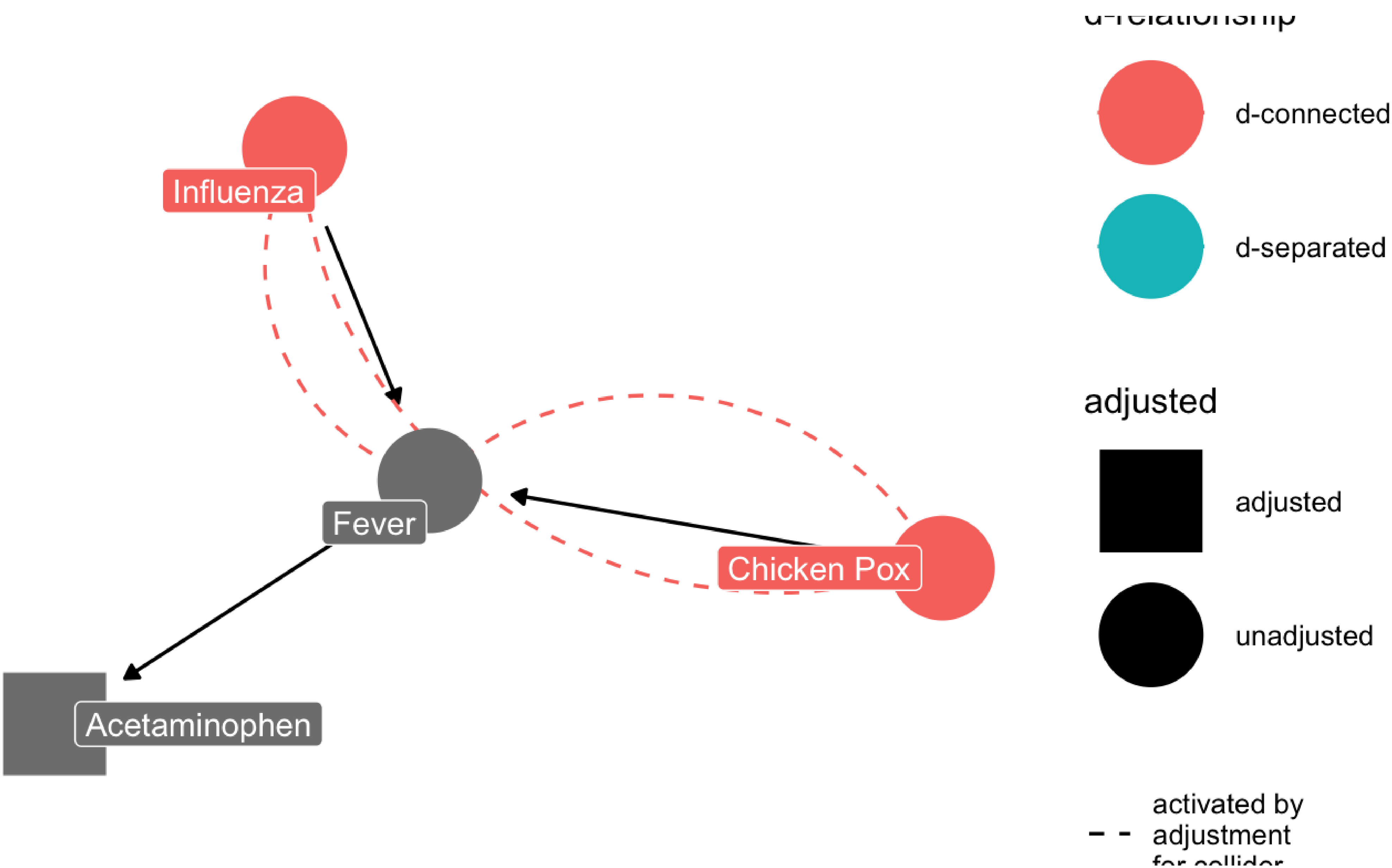
Switzerland's regional demographics
Average age in Swiss municipalities, 2015
Timo Grossenbacher
Average age
33.06
39
40
41
42
43
66.62
Map CC-BY-SA; Author: Timo Grossenbacher (@grssnbchr), Geometries: ThemaKart, BFS; Data: BFS, 2016; Relief: swisstopo, 2016

World Cup 2018 | Club Country Network

Belgium, Brazil, France, Germany, Spain

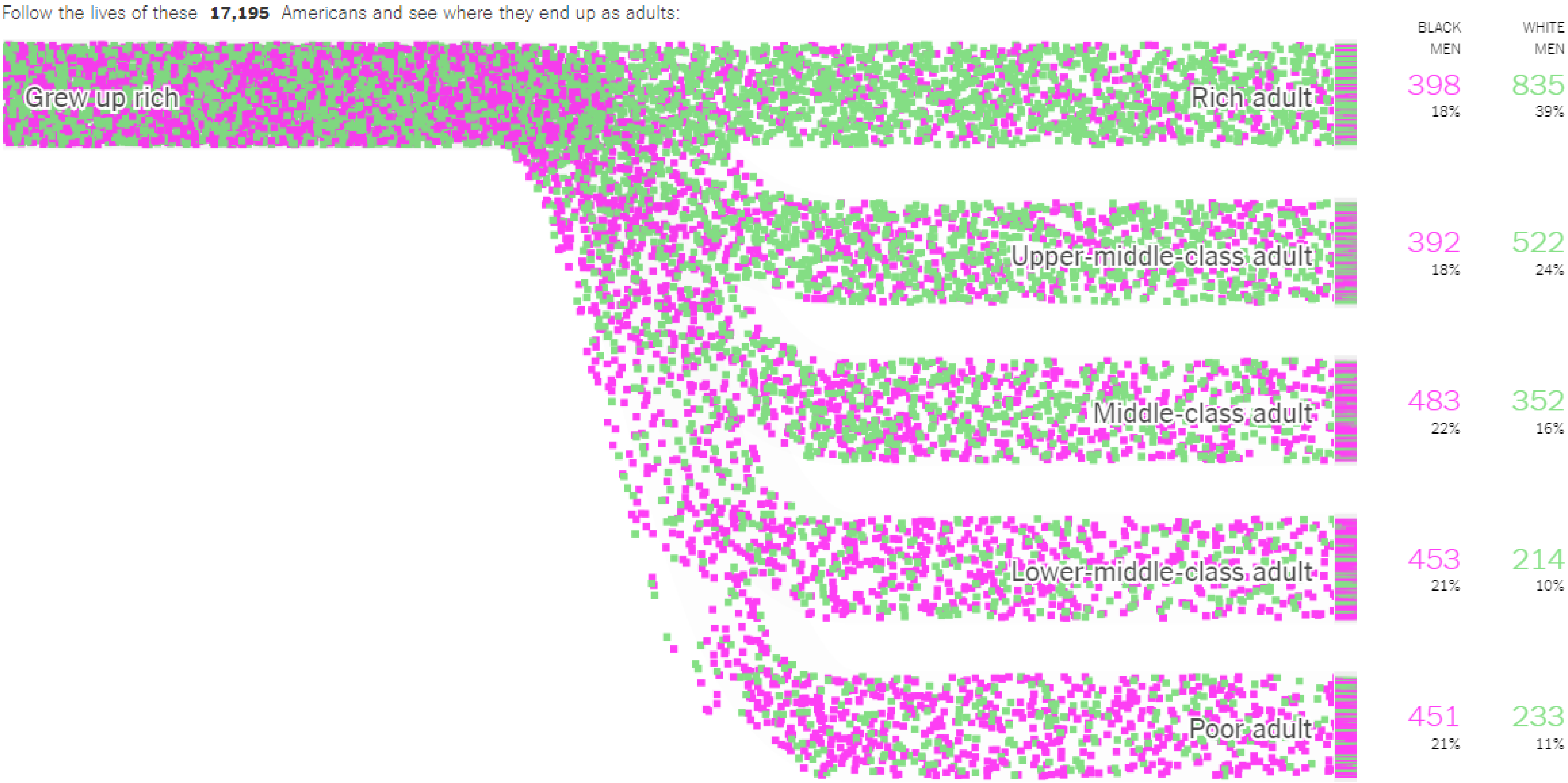


@paulcampbell91 | Source: Wikipedia

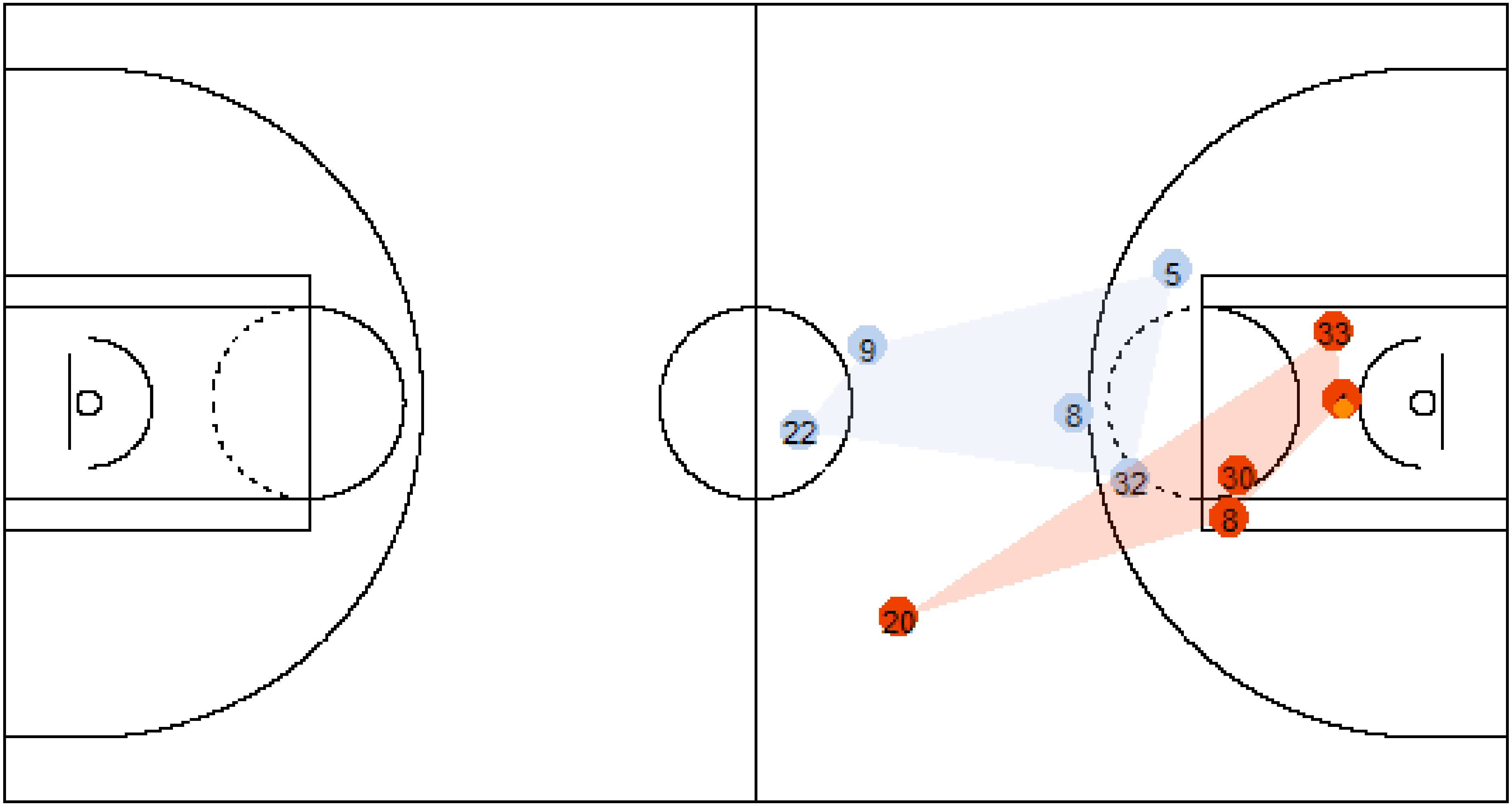


ggdag via Malcolm Barrett

Black and white boys raised in wealthy families



Patrick Honner via NYT


9
22
5
8
32
33
30
8
20

Labs

See the assignments page of the website.

10 points each (30 points total; 15%)

1. Distributions & GitHub collabo
2. Visual perception & plot reproducing
3. Color

Homework

20 points each (40 points; 20%)

- Basically the same as the labs, but scored correct/incorrect, and no in–class time devoted to them.

- Okay to work on collaboratively – I actively encourage you to do so as long as you're using a shared repo

- **Homework 1**: Creating new visuals while utilizing different types of joins

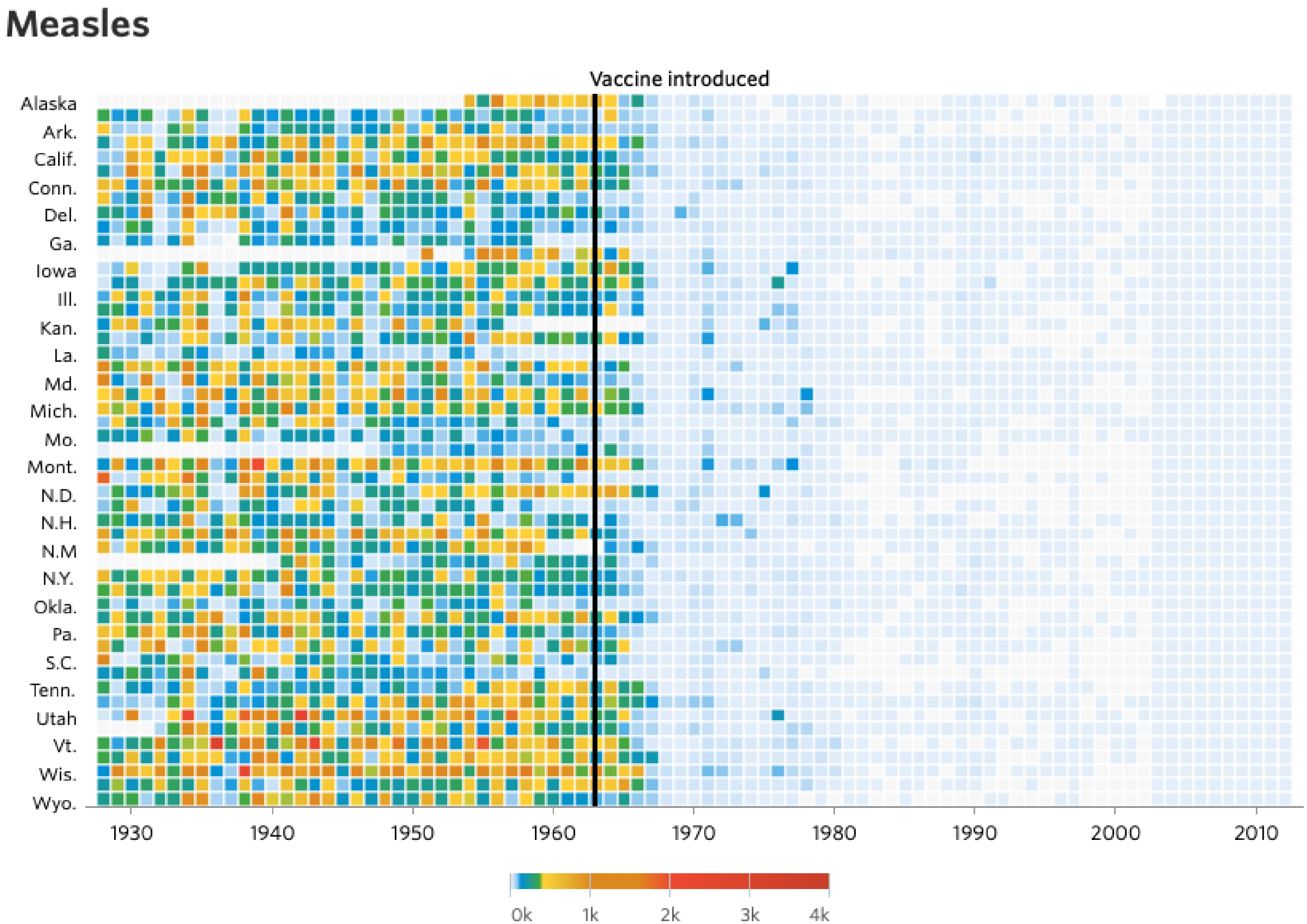
- **Homework 2**: Visualizing uncertainty, tables, and plot refinement

Quick note on reproducibility

A great blog post by Rafael Irizarry shows how almost any plot you see in popular media can be reproduced in R with ggplot.

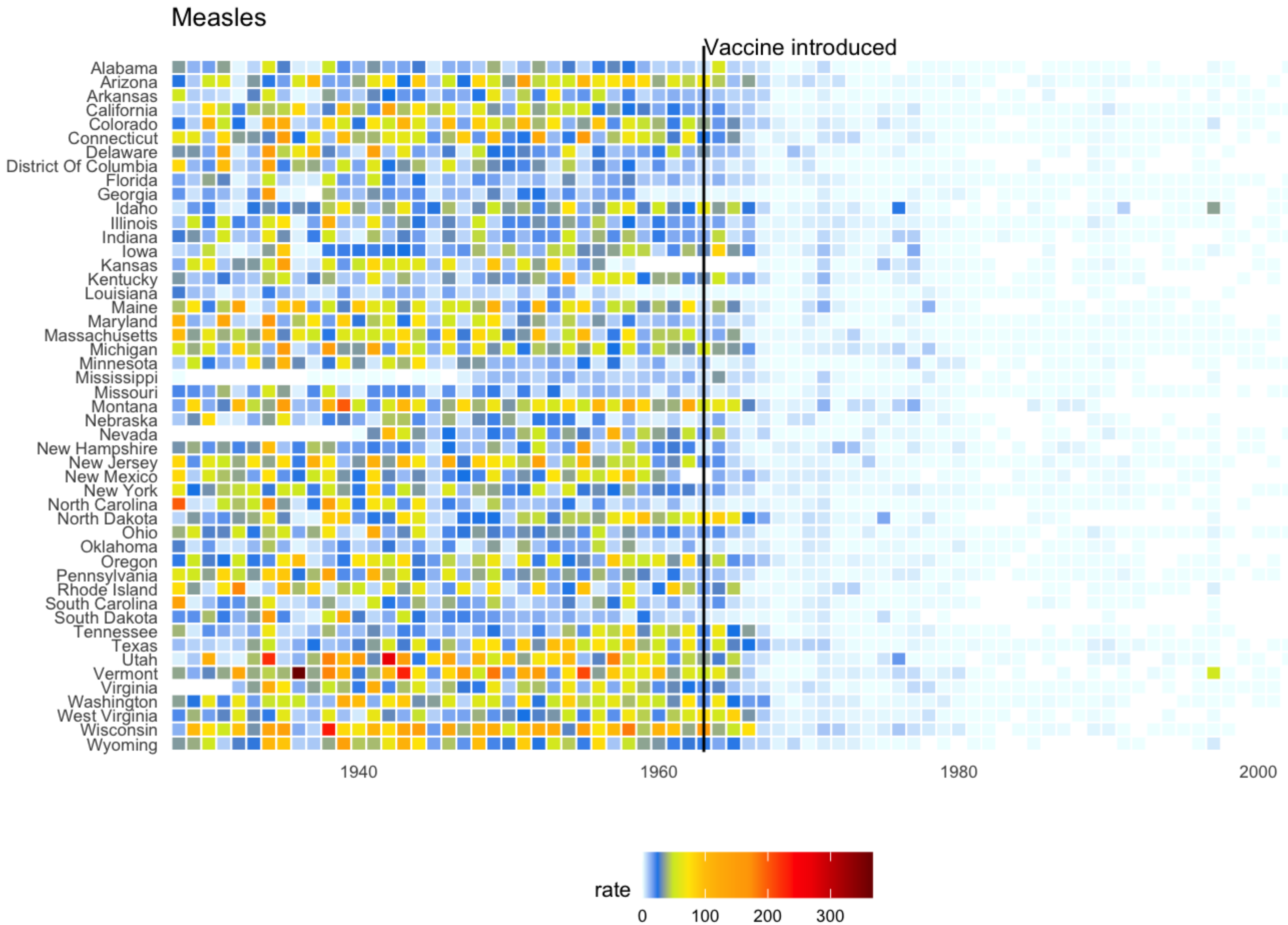
For example

WSJ Version



Note: CDC data from 2003-2012 comes from its Summary of Notifiable Diseases, which publishes yearly rather than weekly and counts confirmed cases as opposed to provisional ones.

ggplot reproduction



Data viz "in the wild" presentations

Everyone will present for ~5 minutes – order randomly assigned (coming up next)

- Find two data viz examples intended for two different audiences
- Discuss the following
    - What's trying to be communicated
    - How effective do you judge it? Why?
    - At least 1 area of strength
    - At least 1 area for (potential) improvement

Presentation order

Date	Presenter
2021–01–11	Kay
2021–01–11	Wanjia
2021–01–13	Joe
2021–01–13	Janette
2021–01–20	Kavya
2021–01–20	Meg
2021–01–25	Anisha
2021–01–25	Rachael
2021–01–27	Zach
2021–01–27	Tess
2021–02–01	Chris

Date	Presenter
2021–02–01	Vinita
2021–02–03	Shijing
2021–02–03	David
2021–02–10	Raleigh
2021–02–10	Maggie
2021–02–15	Ann–Marie
2021–02–15	Murat
2021–02–17	Sarah Don
2021–02–22	Hyeonjin
2021–02–24	Anwesha
2021–03–01	Makayla
2021–03–03	Sarah Dim

I will email this out as well.

Final Project

120 points total (60%)

Six parts

- Proposal (10 points): Due 1/27/21
- Draft (15 points): Due 2/24/21
- Peer review (15 points): Assigned, 2/24/21; Due 3/3/21
- Presentation (20 points): 3/8/21 and 3/10/21 (Week 10)
- Product (60 points): Due 11:59:59 PM, 3/17/21

Product

Four components:

- A web-deployed portfolio showcasing your #dataviz skills.
  - distill (what I'll lecture on), R Markdown, or blogdown website
  - Technical document with pagedown or bookdown
  - Scientific poster with pagedown
  - flexdashboard

- At least three finalized data displays, with each accompanied by a strong narrative/story, as well as the history of how the visualization changed over time.
- Housed on GitHub
  - Fully reproducible
- Deployed through GitHub pages (or netlify or similar)

Proposal

Four components:

- Description of the data source (**must** be publicly available)
- Preliminary ideas of different viz
- Identification of the intended audience for each viz
- The intended message to be communicated for each viz

Draft

- Expected to still be a work in progress
  - Data visualizations should be largely complete
- Deployment not expected
- Provided to your peers so they can learn from you as much as you can learn from their feedback

Peer Review

- We are all professionals here. It is imperative we act like it.
- Understand the purpose of the exercise.
- Zero tolerance policy for inappropriate comments
- Should be vigorously encouraging

Utilizing GitHub

You'll be assigned three proposals to review (5 points each)

- Fork their repo, embed comments & suggest changes to their code, submit a PR

Presentation

Order randomly assigned. Basically a chance to share what you created!

- Discuss what is trying to be communicated
- Share the final products
- Discuss the progression along the way and why specific changes were made

Grading

Points

200 points total

- 3 labs at 10 points each (30 points; 15%)
- 2 homework assignments at 20 points each (40 points; 20%)
- five–minute data visualization "in the wild" presentation (10 points; 5%)
- Final Project (120 points; 60%)
  - Proposal (10 points; 5%)
  - Draft (15 points; 7.5%)
  - Peer review (25 points; 12.5%)
  - Presentation (10 points; 10%)
  - Product (60 points; 30%)

Grading

Lower percent	Lower point range	Grade	Upper point range	Upper percent
0.97	(194 pts)	A+		
0.93	(186 pts)	A	(194 pts)	0.97
0.90	(180 pts)	A–	(186 pts)	0.93
0.87	(174 pts)	B+	(180 pts)	0.90
0.83	(166 pts)	B	(174 pts)	0.87
0.80	(160 pts)	B–	(166 pts)	0.83
0.77	(154 pts)	C+	(160 pts)	0.80
0.73	(146 pts)	C	(154 pts)	0.77
0.70	(140 pts)	C–	(146 pts)	0.73
		F	(140 pts)	0.70


COMMUNICATION

GitHub

Full lecture on Wednesday

My goal: To make you the most prepared cohort with GitHub to date!

Demo

- The gitkraken GUI
- Creating a GitHub repo
- Sharing access (or creating an organization)
- Cloning the repo
- stage, commit, push
- pull
- branching
- forking and issues

Next time

Collaborating with GitHub