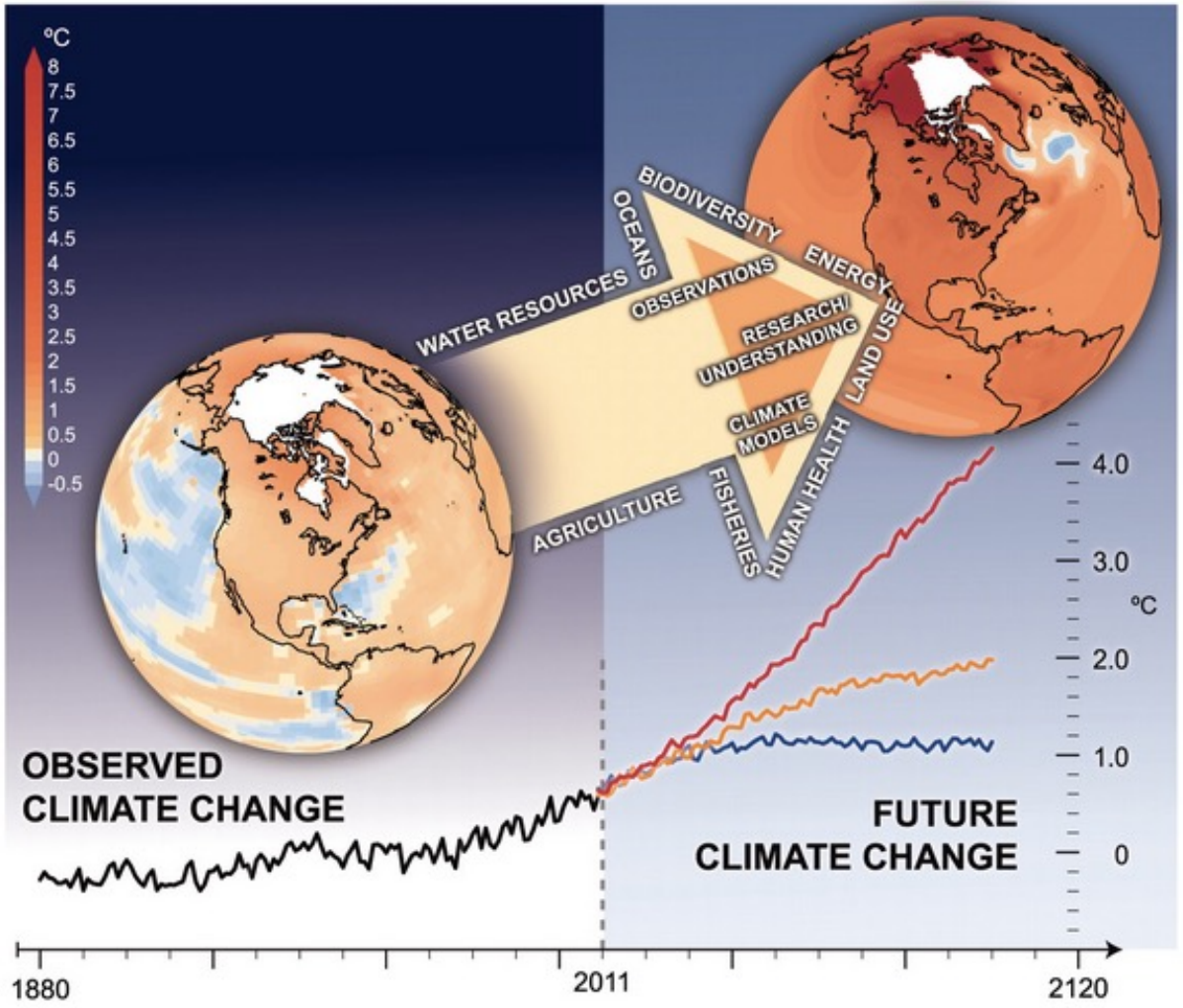
ENVS 410/510: Data analysis and visualization

Lauren Hallett

Global Change



Overpeck+ (2011) doi:10.1126/science.1197869

Remote sensors

credit: NASA

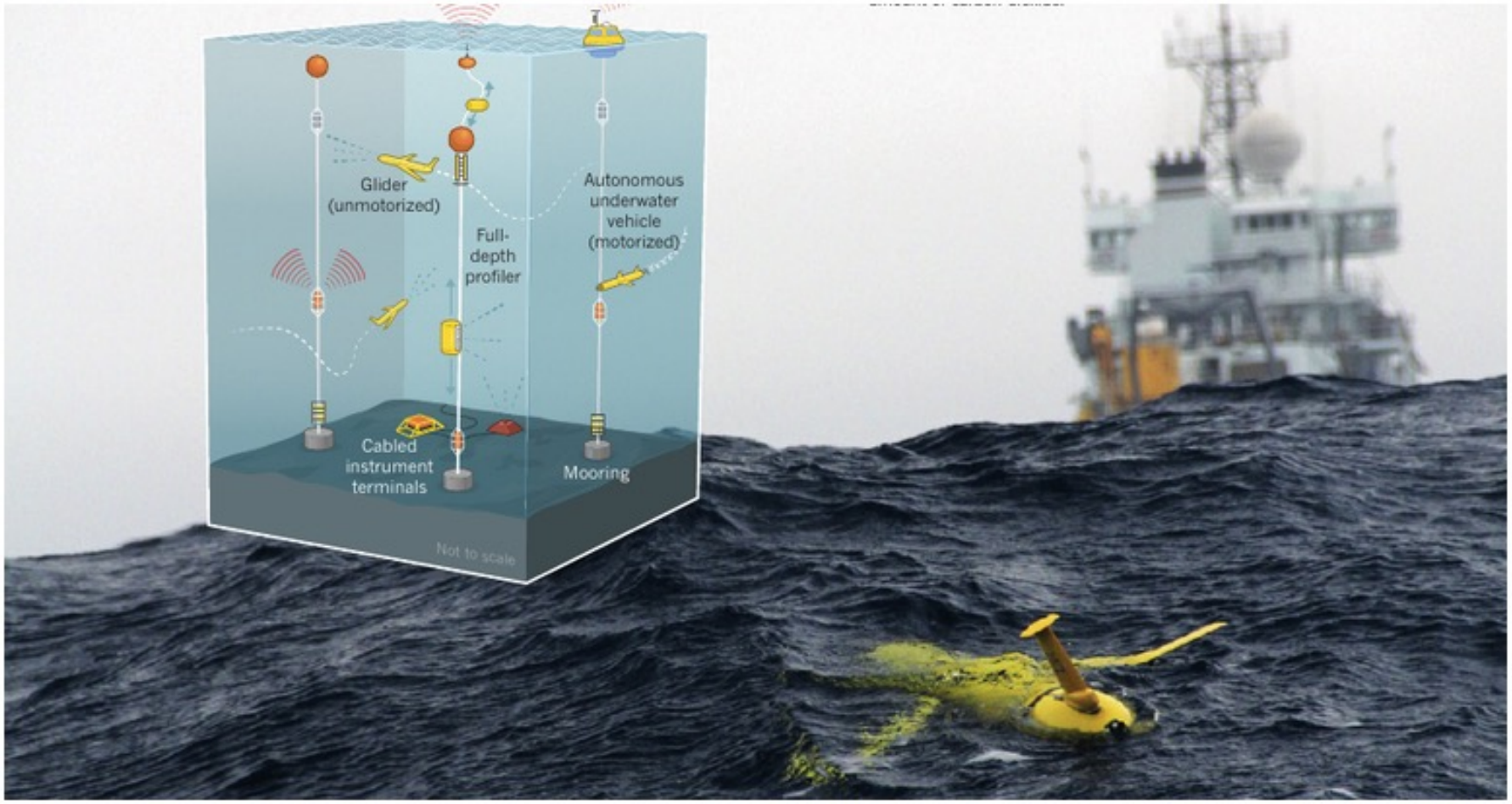
micro sensors

credit: NSF

NEON

credit: Hopkin (2006) doi:10.1038/444420a

001



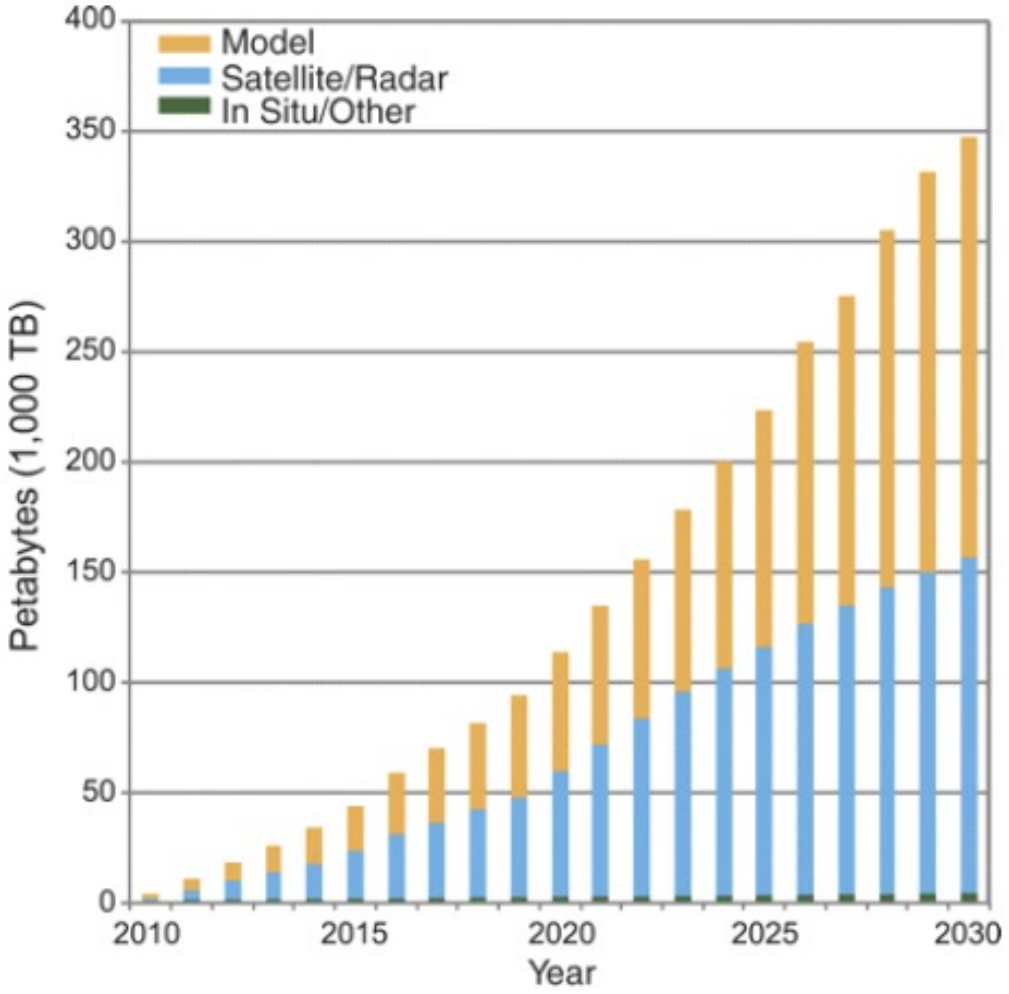
credit: Witze (2013) doi:10.1038/501480a

Computer simulations

credit: NSF Cyverse / Jetstream

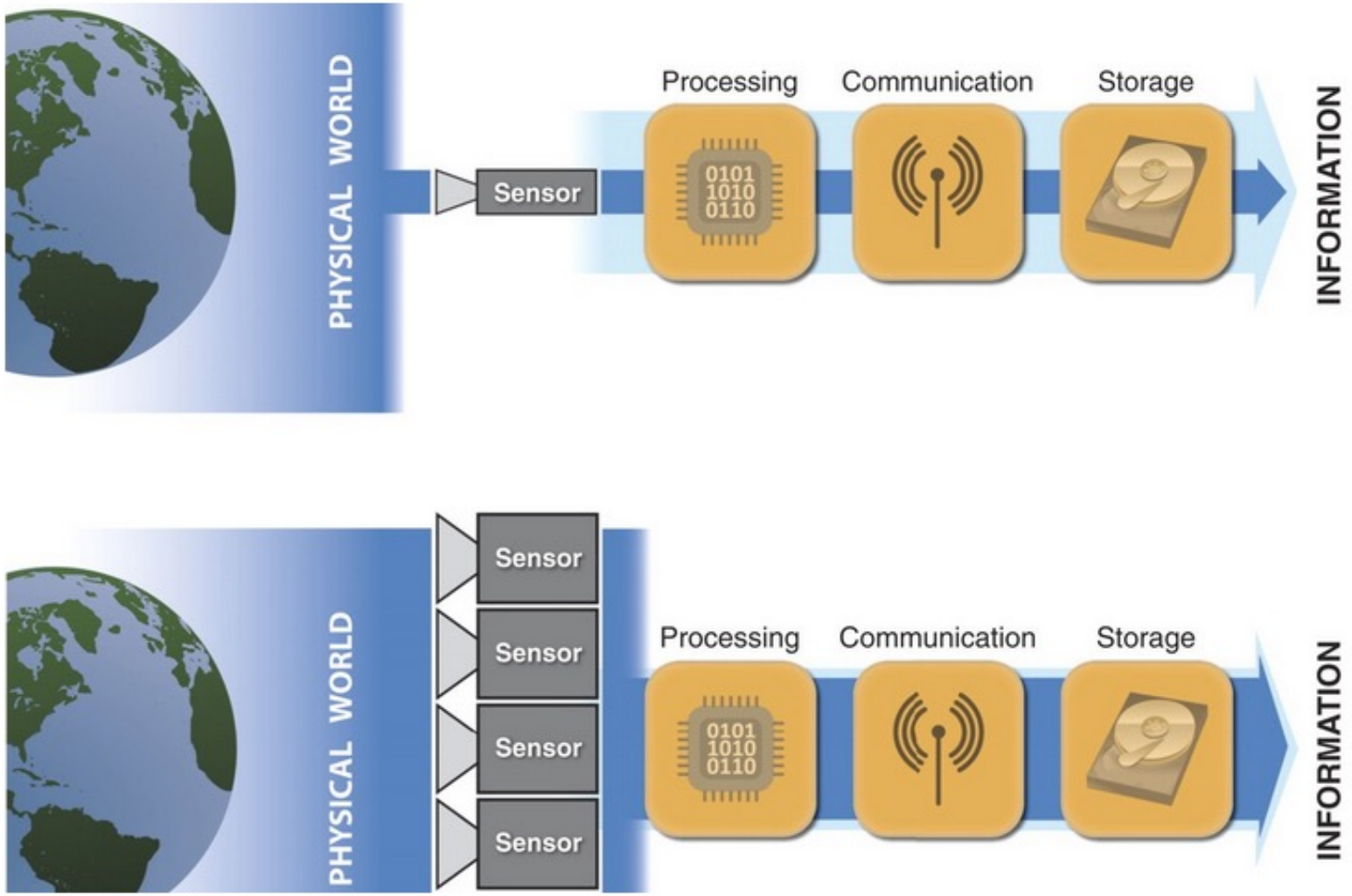
Field-based study

Growth of climate data by type



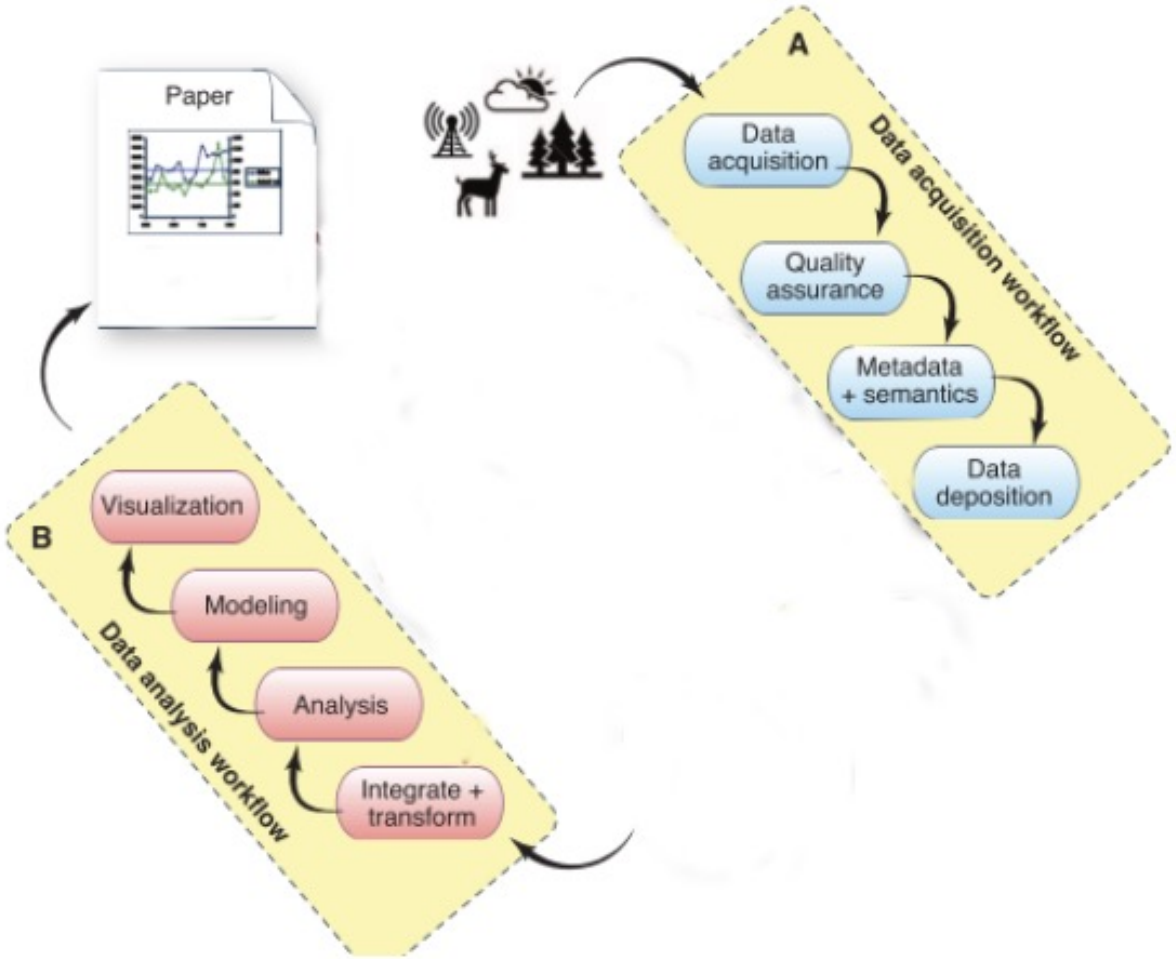
Overpeck+ (2011) doi:10.1126/science.1197869

Engineering bottlenecks



Baraniuk (2011) doi:10.1126/science.1197448

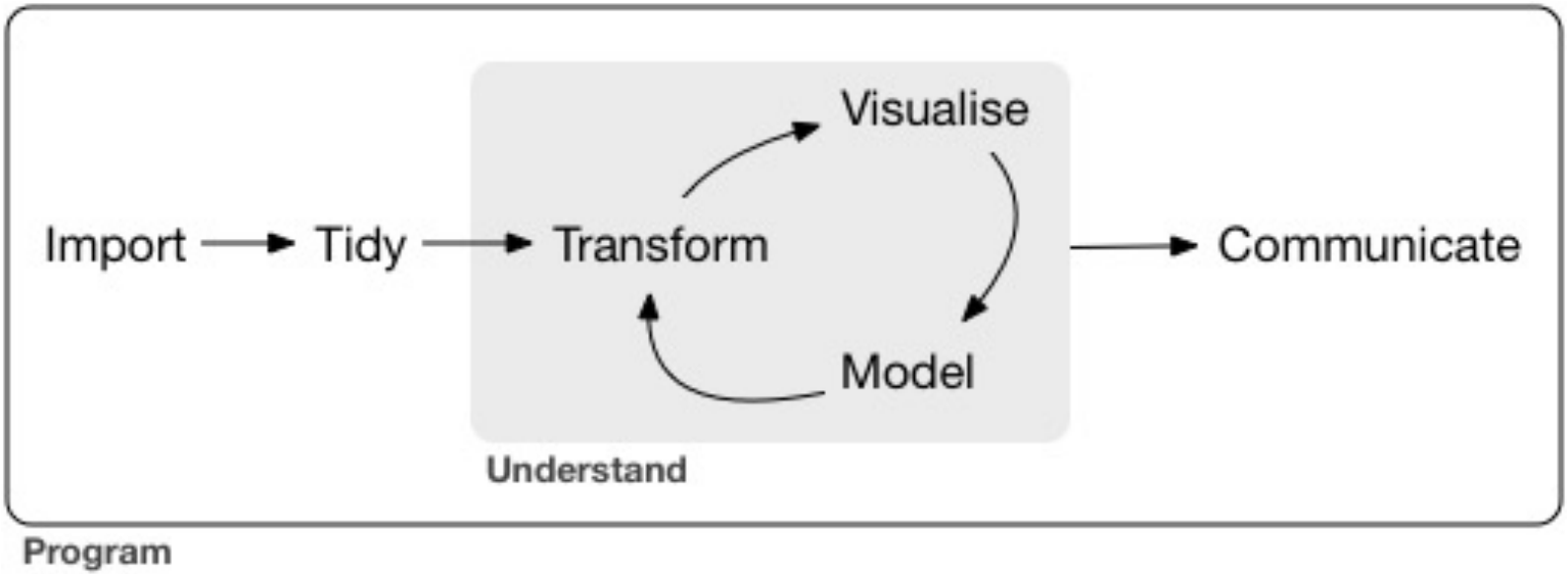
Science bottlenecks



adapted from Reichman+ (2011) doi:10.1126/science.1197962

Our focus is on these
Science bottlenecks

Course content: data life cycle



Wickham and Grolemund R for Data Science

Data cleaning

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.
Peter DaSilva for The New York Times

http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?partner=rss&emc=rss&smid=tw-nytimesscience&_r=0

Analysis and modeling

- Formally the 80%
- Computational developments have made this much easier
- And led to an explosion in new methods and tools

Science bottlenecks

Different ways for data to be “big”

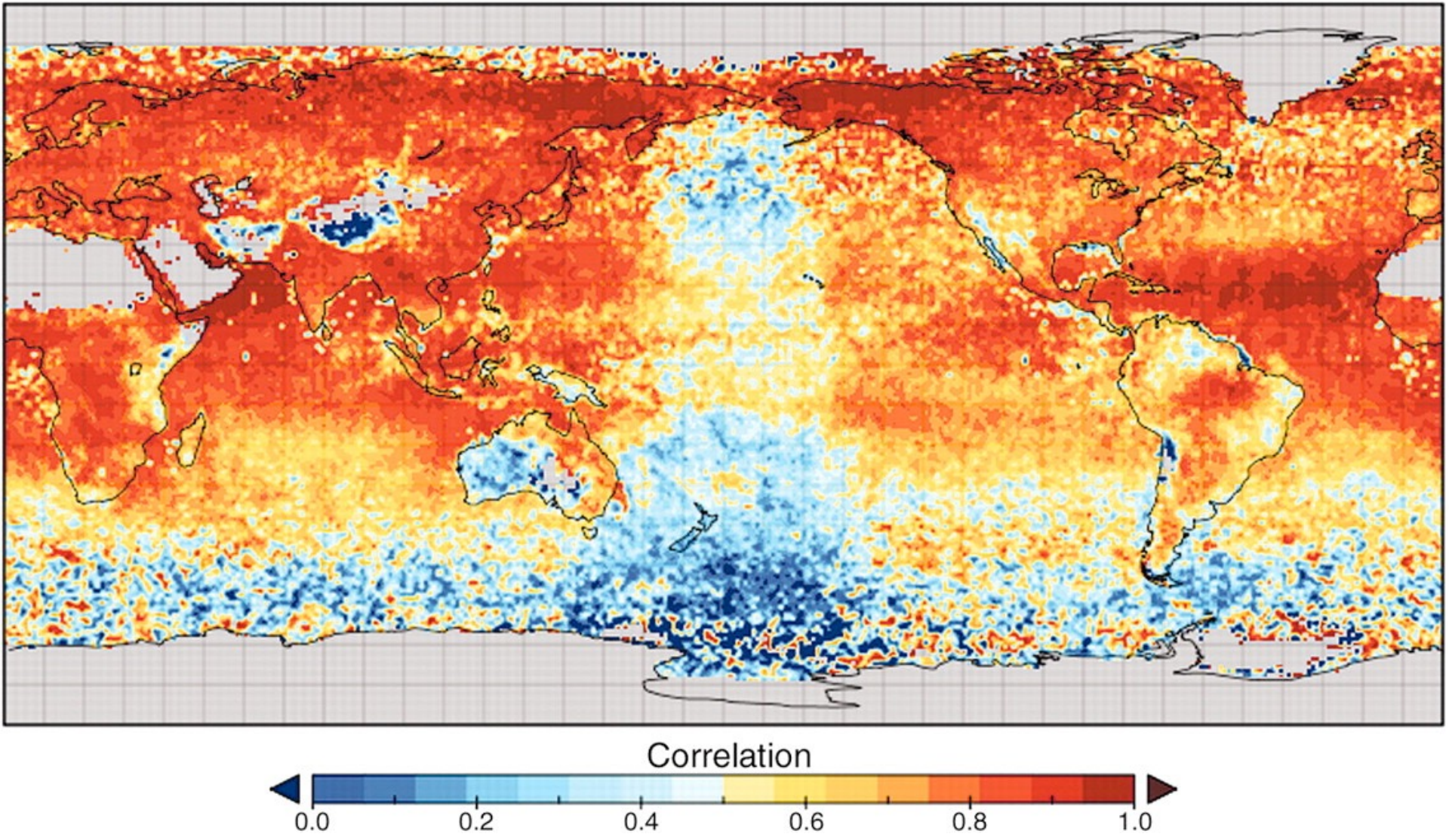
- *Volume*
- *Variety*
- *Velocity*

Volume

Data visualization

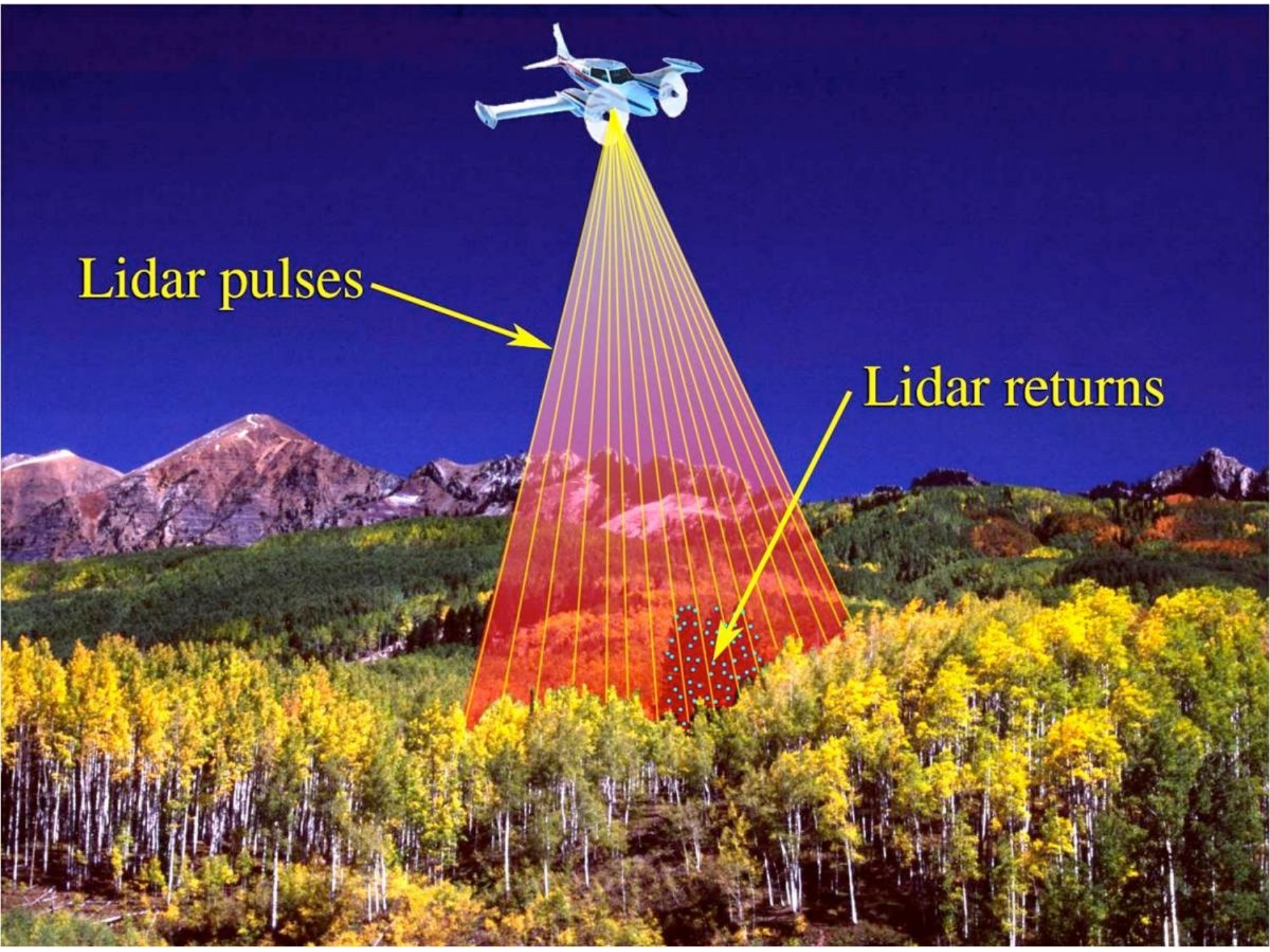
- Today the visualization component has become a bottleneck
- To often visualization becomes only an end-product rather than an exploration tool
- A problem because visualization is a quick way to spot errors

Fox & Hendler (2001): doi:10.1126/science.1197654

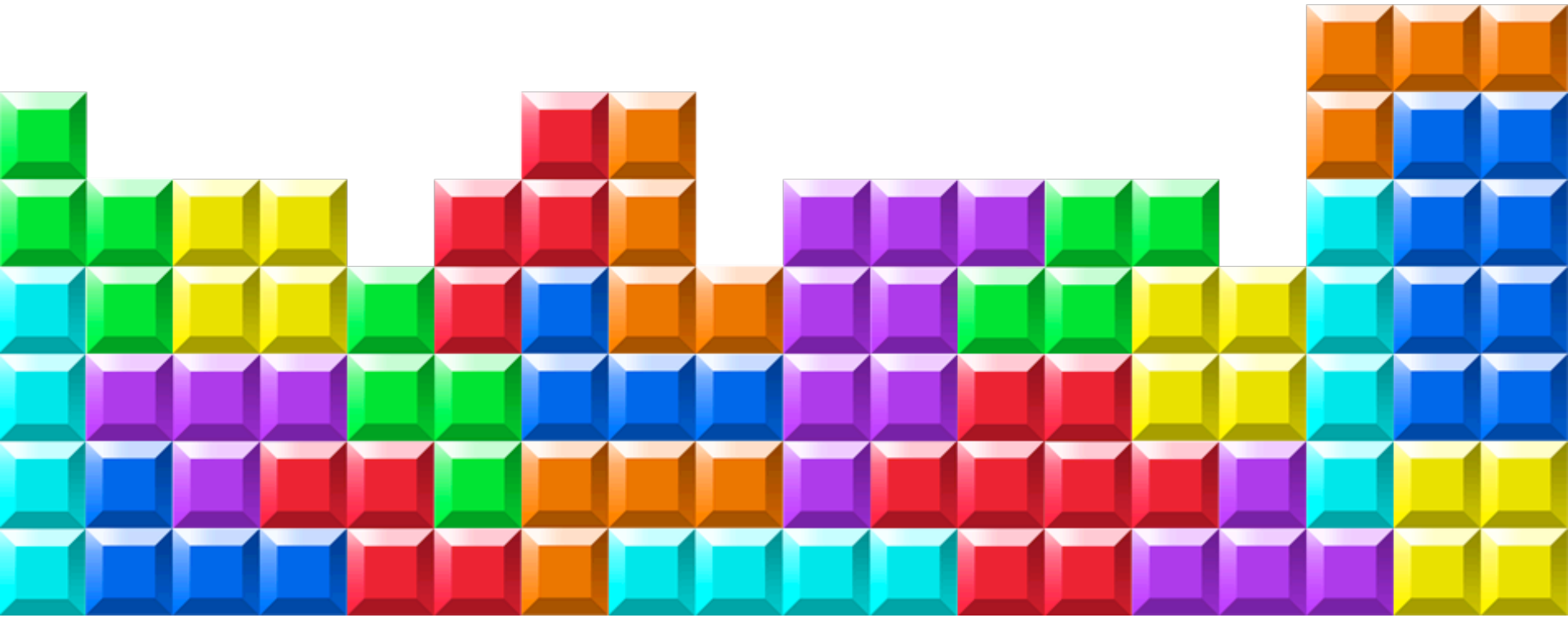


Fox & Hendler (2001): doi:10.1126/science.1197654

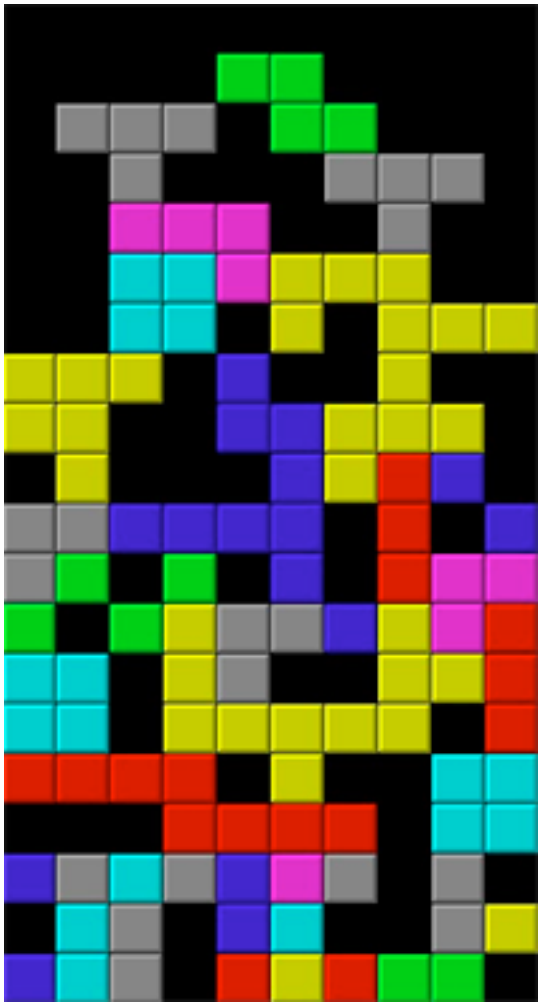
Variety


Lidar pulses
Lidar returns

Vertically integrated data repositories

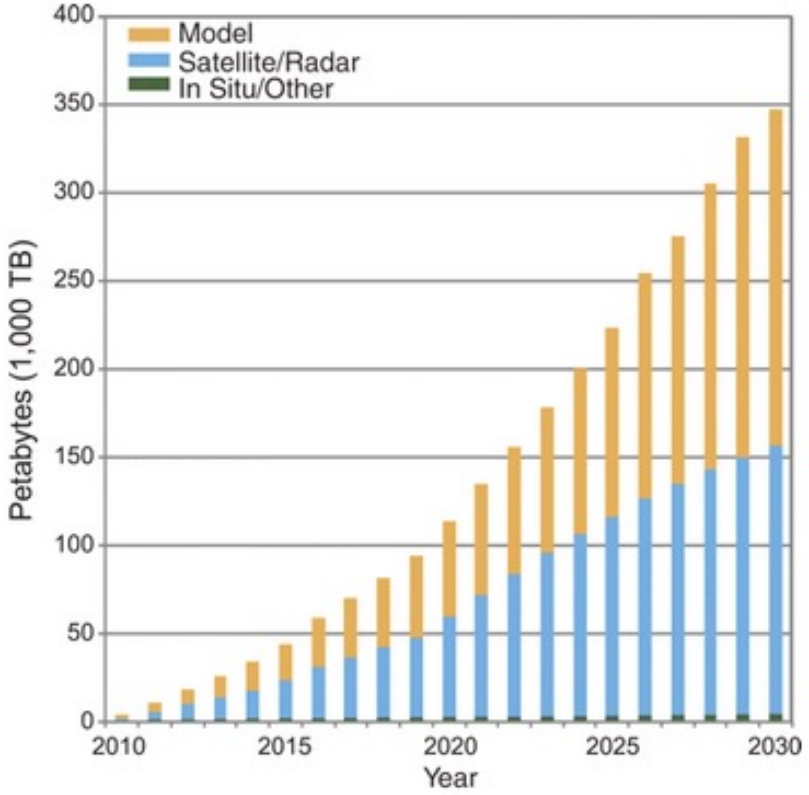


Limits to vertical integration



Velocity

Velocity: the need for Reproducibility



Overpeck+ (2011) doi:10.1126/science.1197869

Most data are yet to come

Example: Synthetic analysis of biodiversity loss

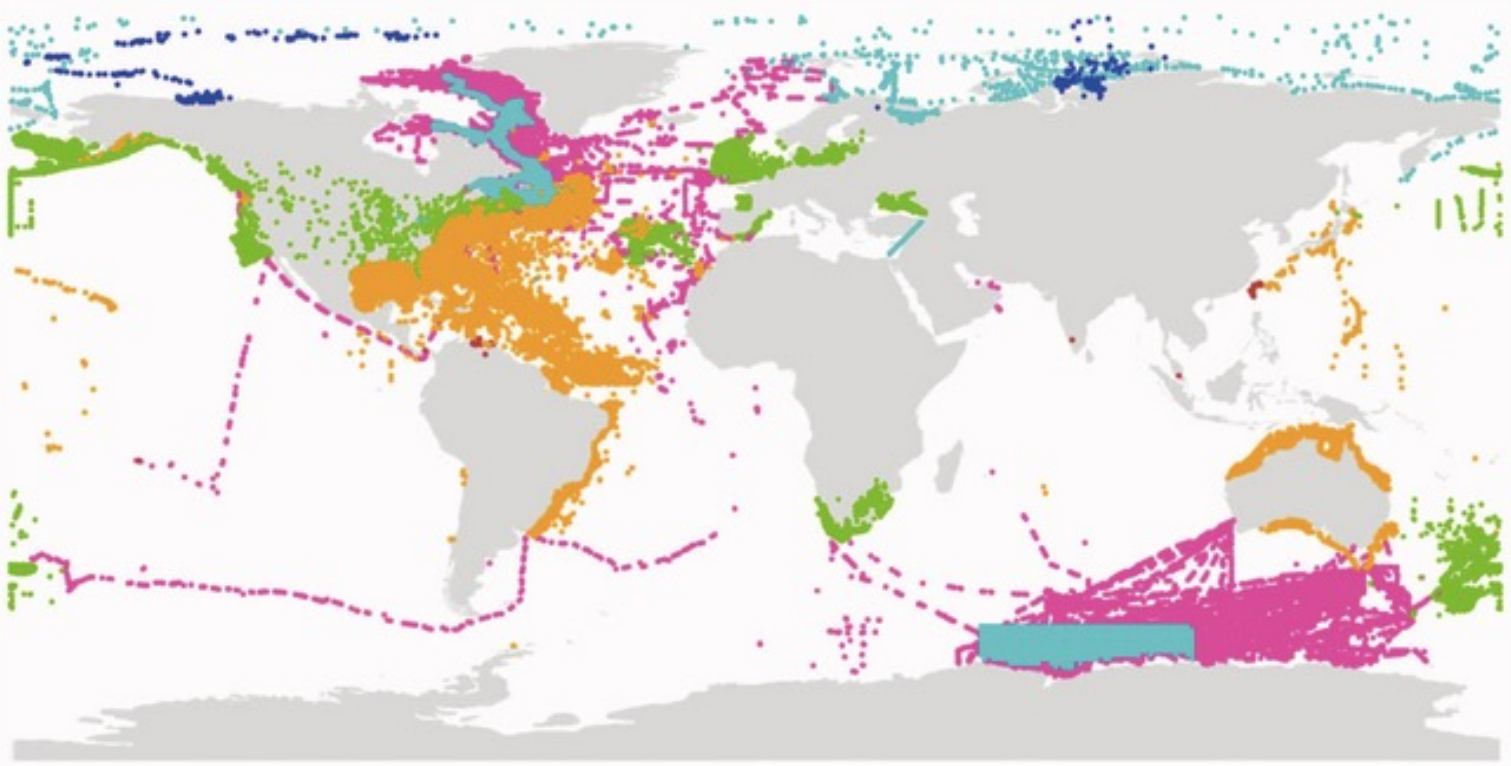


Fig. 1. Distribution of the survey sites included in our analysis. Data sets are color-coded to reflect their climatic region: pink, global; royal blue, polar; turquoise, polar-temperate; green, temperate; gold, temperate-tropical; red, tropical. See table S1 for details and sources of the data sets.

Sythesizes over 140 data sets.

Finds no evidence for systematic loss

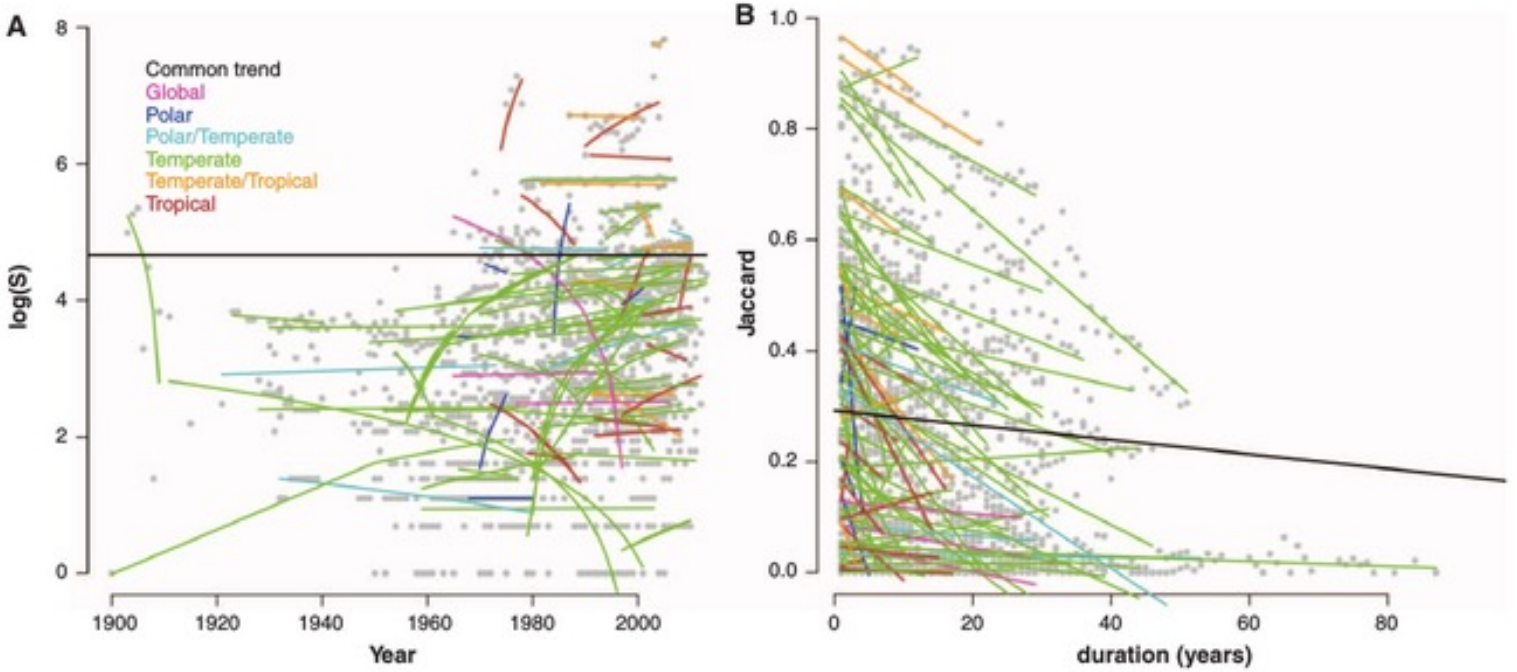
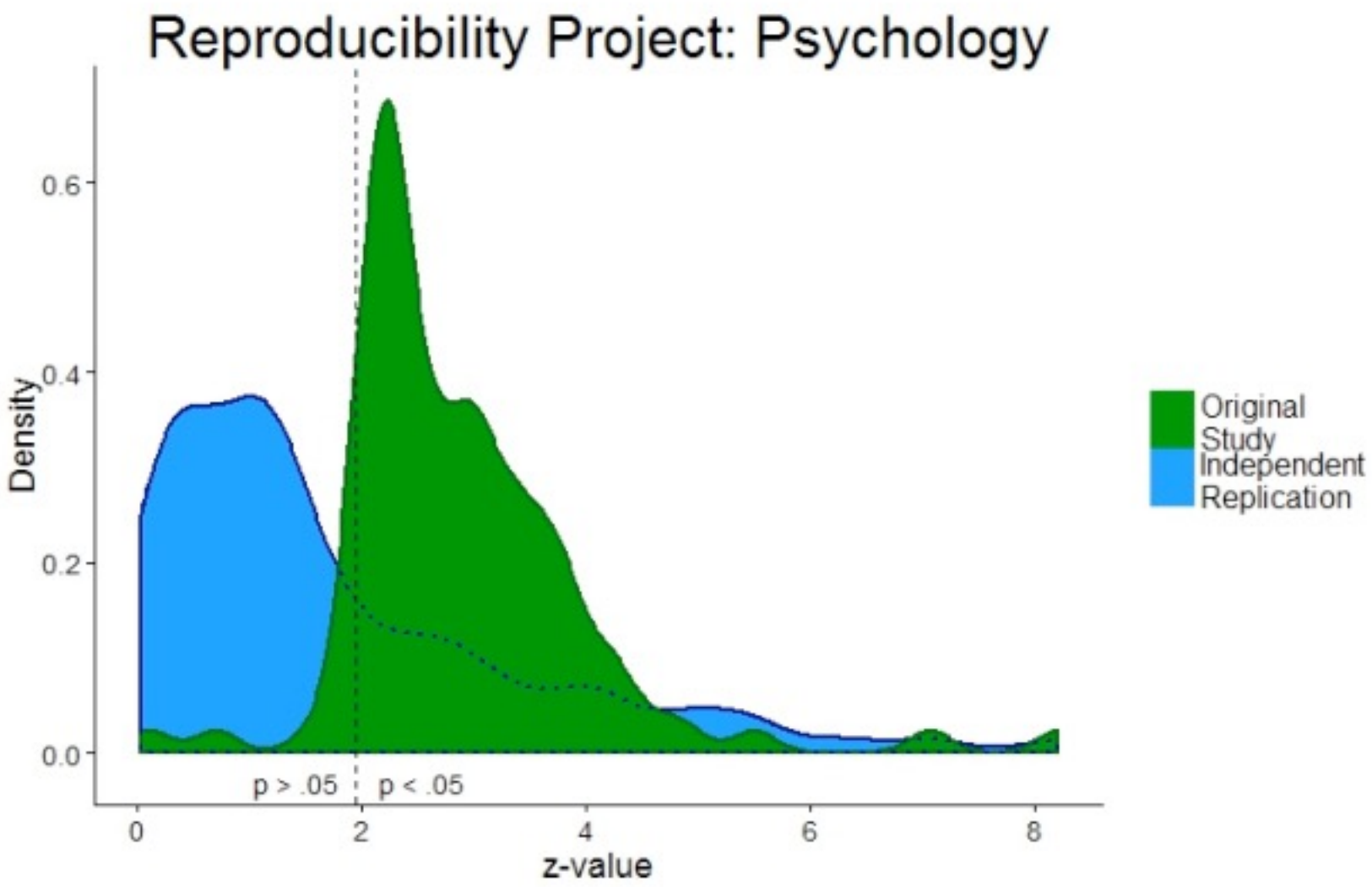


Fig. 2. Temporal change in α diversity and temporal β diversity. (**A** and **B**) Temporal change in species richness (A) and species composition (B) as measured by Jaccard similarity between each sample and the first sample in the time series. Data points are represented by gray circles and models fitted by solid lines. The black line corresponds to a model in which a single slope, but different intercepts, were fitted to all the time series, and is represented here with the mean intercept. The colored lines correspond to a model where each time series had a different slope and intercept. Color coding corresponds to Fig. 1. Figure S10 presents a similar analysis for a different approach to rarefaction.

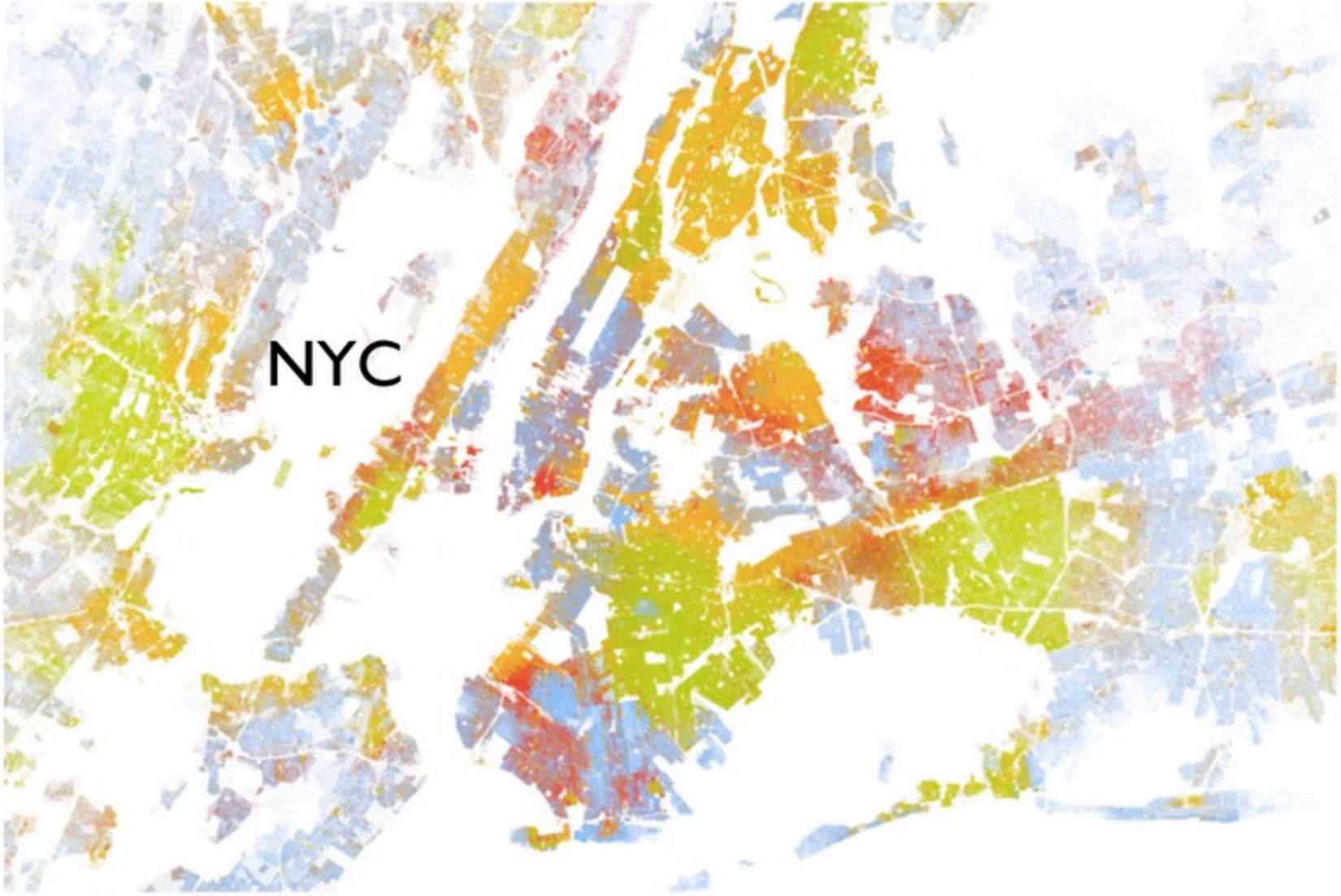
How easy would it be to update this to reflect new data?

Reproducibility: Important for other reasons as well


Reproducibility Project: Psychology
Density
0.6
0.4
0.2
0.0
p > .05
p < .05
0
2
4
6
8
z-value
Original Study
Independent Replication

The Best Map Ever Made of America's Racial Segregation

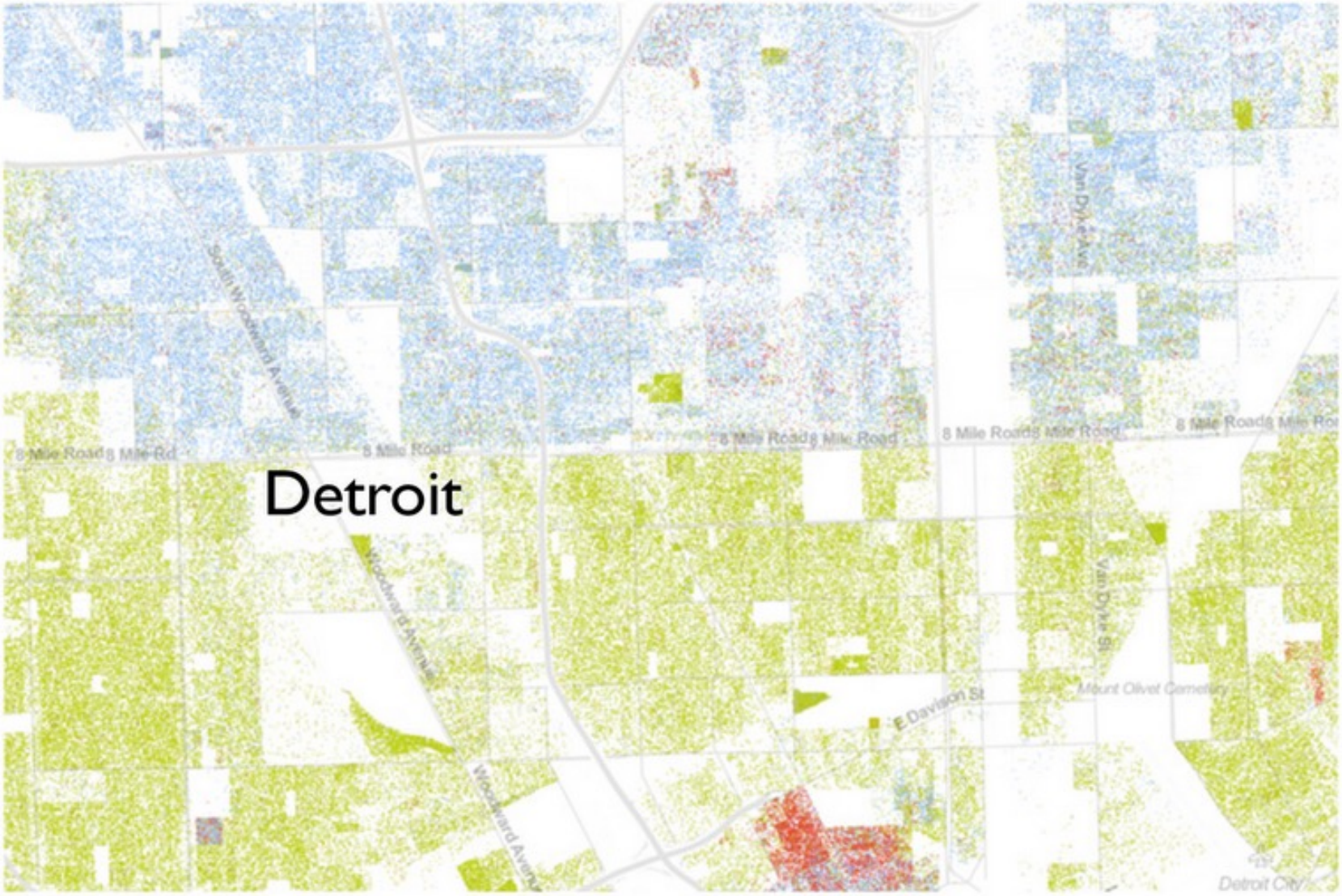
http://www.wired.com/design/2013/08/how-segregated-is-your-city-this-eye-opening-map-shows-you/?viewall=true



This map, created by Dustin Cable at University of Virginia's Weldon Cooper Center for Public Service, is the most comprehensive representation of racial distribution in America ever made. Here: New York City. *Image: Dustin Cable* **White: blue dots; African American: green dots; Asian: red; Latino: orange; all others: brown**

The Best Map Ever Made of America's Racial Segregation

http://www.wired.com/design/2013/08/how-segregated-is-your-city-this-eye-opening-map-shows-you/?viewall=true



In Detroit, among the most segregated cities in America, 8 Mile Road serves as a sharp dividing line. *Image: Dustin Cable* **White: blue dots; African American: green dots; Asian: red; Latino: orange; all others: brown**

This is the most comprehensive map of race in America ever created.

White people are shown with blue dots; African-Americans with green; Asians with red; and Latinos with orange, with all other race categories from the Census represented by brown. Since the dots are smaller than pixels at most zoom levels, Cable assigned shades of color based on the multiple dots therein. From a distance, for example, certain neighborhoods will look purple, but zooming-in reveals a finer-grained breakdown of red and blue–or, really, black and white.

"There are a lot of moving parts in this process, so this can cause different shades of color to appear at different zoom levels in really dense areas, like you see in NYC," Cable explains. "I played around with dot size and transparency for a while and settled on the current scheme as being adequate." You can read more about Cable's methodology here, but it comes down to this: When you're dealing with 300 million dots at varying levels of zoom, getting the presentation just right is as much an art as a science.

http://www.coopercenter.org/demographics/Racial-Dot-Map

Cool result is accompanied by explanation of how it was done

Methodology

Python was used to read the 50 state shapefiles (with the merged SF1 data). The GDAL and Shapely libraries were used to read the data and create the point objects. The code retrieves the population data for each census block, creates the appropriate number of geographic points randomly distributed within each census block, and outputs the point information to a database file. The resulting file has x-y coordinates for each point, a quadkey reference to the Google Maps tile system, and a categorical variable for race. The final database file has 308,745,538 observations and is about 21 GB in size. The processing time was about five hours for the entire nation.

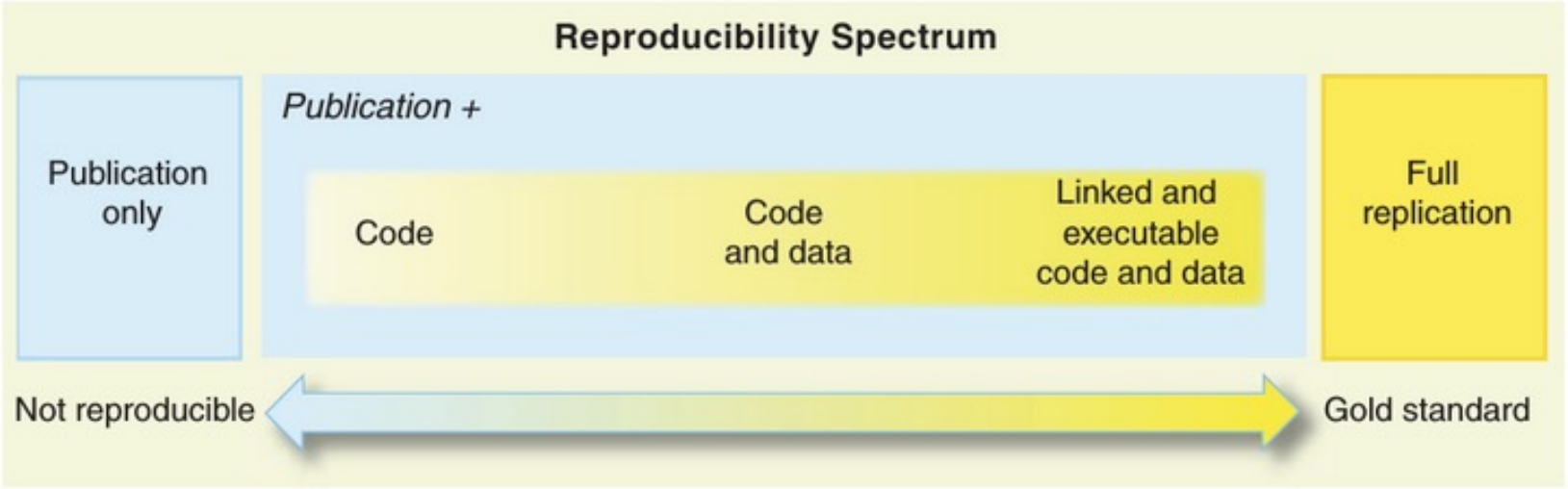
The database file was then sorted by quadkey and converted to a .csv format. SAS was able to do this within an hour without crashing.

Processing 2.0.1 for 64-bit Windows was used to create the map tiles. The Java code reads each point from the .csv file and plots a dot on a 512x512 .png map tile using the quadkey reference and x-y coordinates. The racial categorical variable is used to color-code each plotted dot. This process used the default JAVA2D renderer, but other platforms may work better using P2D. Map tiles were created for Google Maps' zoom levels 4 through 13 to make the final map. A non-color-coded map was also produced to help add more contrast for lightly populated areas. In total, the color-coded and non-color-coded maps contain 1.2 million .png files totaling about 7 GB. Producing all of the map tiles in Processing took about 16 hours for the two maps.

The Google Maps API is used to display the map tiles. Map tiles with zero population are never created using the above method. Therefore, an index was used to tell the map application whether a tile exists in order to prevent 404 errors.

The entire code is up on GitHub and was adapted from code developed by Brandon Martin-Anderson and Peter Richardson in order to account for the racial coding and errors in reading the shapefiles.

Reproducible workflows: Dynamic documents



Peng (2011) doi:10.1126/science.1213847

This course:

Real data, real tools

- Things will break
- Things will change

This course:

Some things will be in keeping with good pedagogy:

- We'll start with graphs first
- Active learning

Some things aren't:

- We'll jump forward and back at times
- Starting at the deep end

How we'll do it:

- Different modules in which we transition from learning and coding as a class to independently
- Frequent practice via four problem sets
- An independent project in which you go from raw data to a reproducible final report
- A free online text to help you along! https://r4ds.had.co.nz/

Who we are:

What best describes your familiarity with R? (I love teaching R to new-comers, and I do not expect you to have any familiarity, but it helps me to know where you're starting from).

12 responses

A smattering of what we think about: ecology, conservation/restoration, freshwater ecosystems, snow pack science, sustainable agriculture, environmental justice, animal behavior, wildlife, TEK

Imposter Syndrome

Language: Why R?

- Scripted languages allow reproducibility
- Not about technical attributes
- Pick your language based on what people in your area speak/code
- Rstudio, Rmarkdown and other tools for communication and sharing
- A great open-source community (on that note - thanks to Carl Boettiger and Jenny Bryan for many of todays slides)

Exploratory data analysis exercise

Goals for this exercise:

- Highlight the value of plotting our data before we analyze it
- Get to know one another
- Viscerally appreciate how much easier R will make our lives!