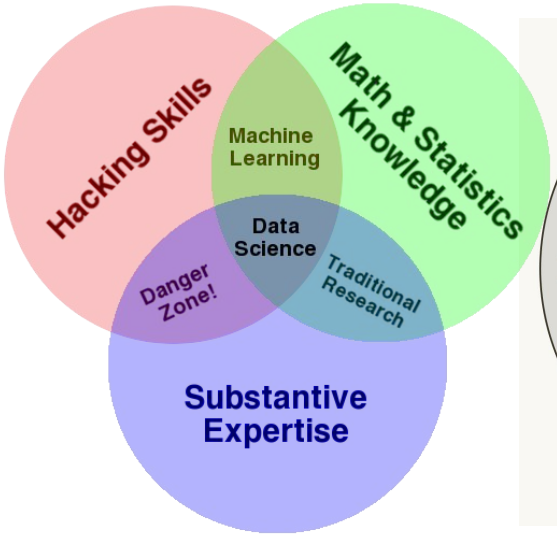
Is reasoning about variability part of data science?

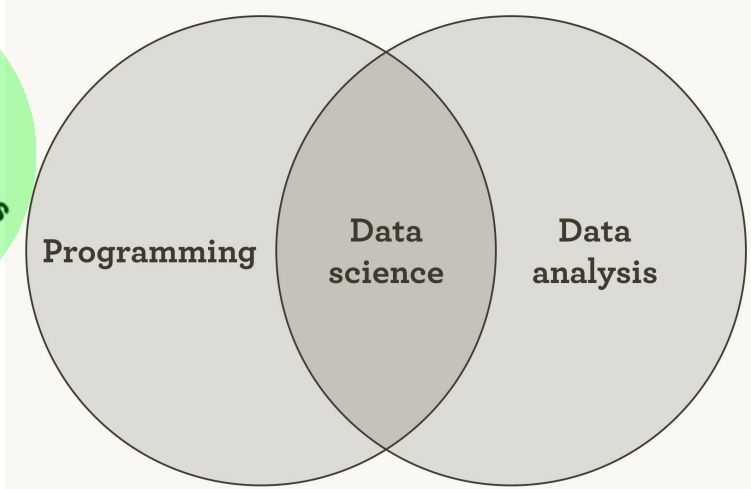
Amelia McNamara @AmeliaMN
Smith College Program in Statistical and Data Sciences

flickr: elora.daphne

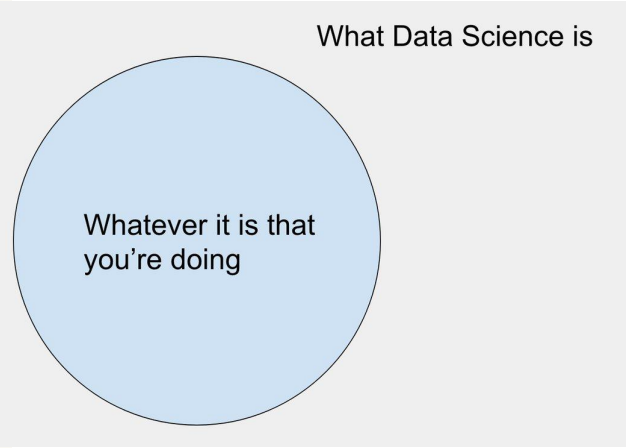
Data science



Drew Conway, 2013



Hadley Wickham, 2017



Sean Kross, 2017

The three V’s of big data

- Volume
- Velocity
- Variety
- (Veracity)

Edd Wilder-James. “What is big data? An introduction to the big data landscape.” O’Reilly, 2012. https://www.oreilly.com/ideas/what-is-big-data

Crawford/boyd critical questions for big data

1. Big Data changes the definition of knowledge
2. Claims to objectivity and accuracy are misleading
3. Bigger data are not always better data
4. Taken out of context, Big Data loses its meaning
5. Just because it is accessible does not make it ethical
6. Limited access to Big Data creates new digital divides

danah boyd and Kate Crawford. “Critical questions for big data.” Information, Communication & Society. 2012. http://bit.ly/CriticalQuestionsForBigData

Leo Breiman. “Statistical Modeling: The Two Cultures.” Statistical Science, 2001

Understanding variability is a key goal of statistics

"Students should recognize and be able to explain the central role of variability in the field of statistics."

Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016. http://bit.ly/GAISE2016

“Konold and colleagues (Konold & Higgins, 2002; Konold et al., 2003) argue that children see data in several simpler ways before ever noticing aggregate and emergent features of data sets. Their fourfold schema includes the following different ways of viewing data, which we consider useful for examining the thinking of adults as well as children:

1. Data as a pointer to the data collection event but without a focus on actual data values—in this view, data remind children of their experiences, “We looked at plants. It was fun.”
2. Data as a focus on the identity of individual cases—these can be personally identifiable, “That’s my plant! It’s 18 cm tall,” extreme values, “The tallest plant was 37 cm,” or interesting in some other way.
3. Data as a classifier which focuses on frequencies of particular attribute values, or “slices,” without an overall view—“There were more plants that were 15 to 20 cm than 10 to 15 cm.”
4. Data as an aggregate, focusing on overall and emergent characteristics of the data set as a whole, for example, seeing it as describing variability around a center, or “noise” around an underlying “signal” (Konold & Pollatsek, 2002)—“These plants typically grow to between 15 and 20 cm.”

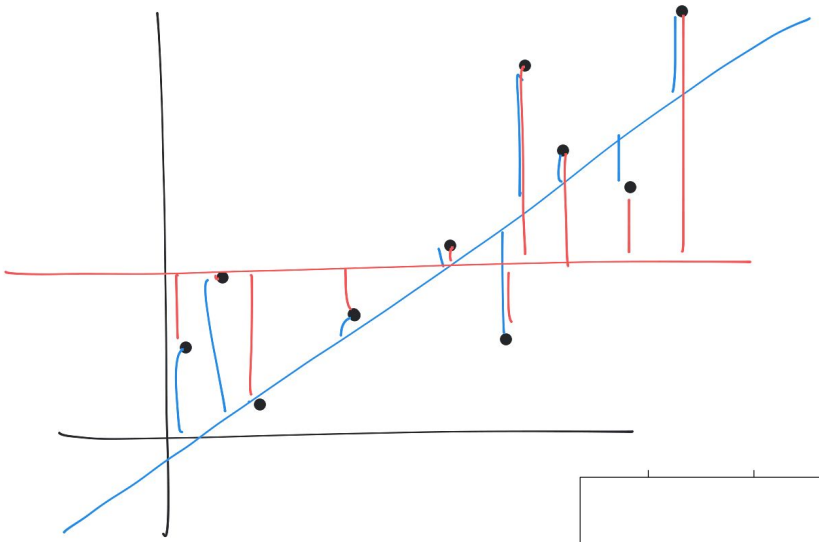
James Hammerman and Andee Rubin. “Strategies for Managing Statistical Complexity Using New Software Tools.” Statistics Education Research Journal, 3(2), 2004.
https://iase-web.org/documents/SERJ/SERJ3(2)_Hammerman_Rubin.pdf

But, is reasoning about variability part of data science?

Types of variability

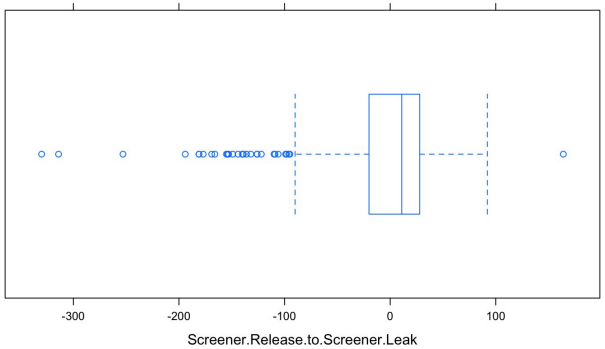
- Variability within an attribute
  - Within individual, and between individuals
- Variability that isn’t-- errors in data that make it look variable
- Sample to sample variability
- Measurement error and process error
- Modeling error
  - Uncertainty of model parameters, and errors in modeling predictions
- Variability due to parameter choices

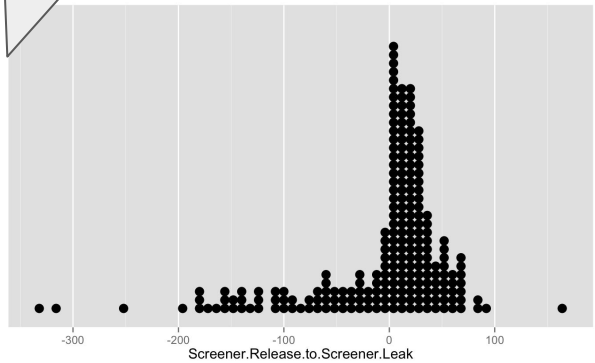
Variability within an attribute



"Data as an aggregate, focusing on overall and emergent characteristics of the data set as a whole, for example, seeing it as describing variability around a center, or "noise" around an underlying "signal""

- Hammerman and Rubin, referencing Konold





Sample to sample variability

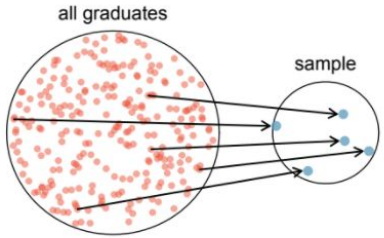
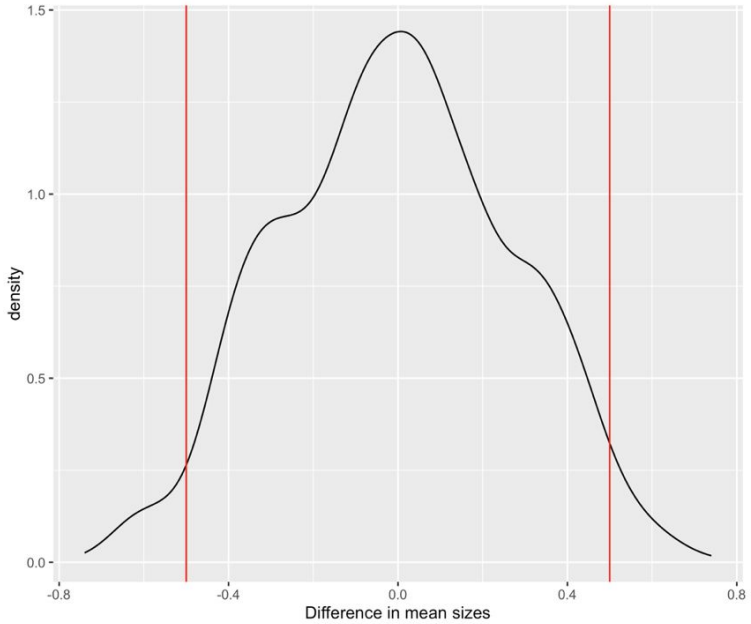


Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

David Diez, Christopher Barr, and Mine Cetinkaya-Rundel. "Introductory Statistics with Randomization and Simulation." OpenIntro, 2014

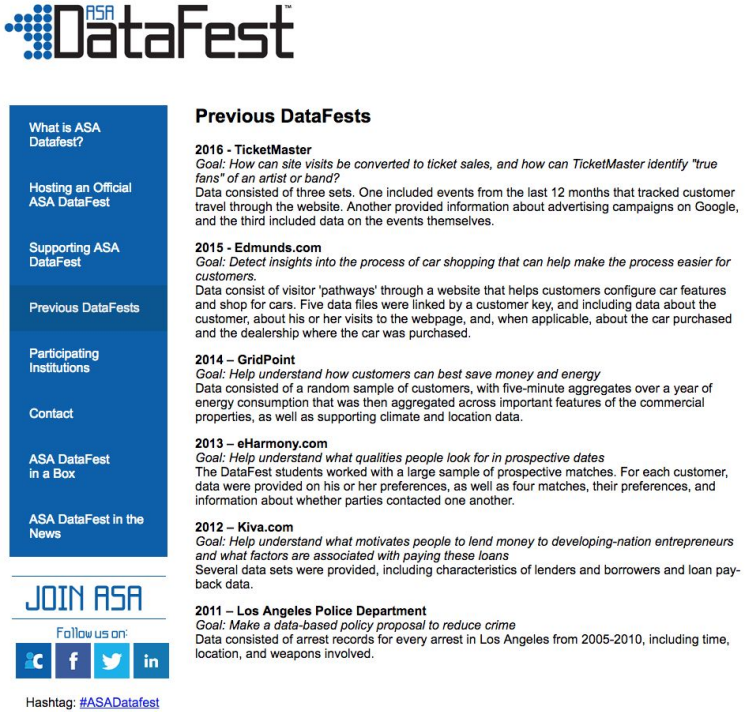


Amelia McNamara. "Do you know Nothing when you see it?" OpenVisConf, 2016. https://www.youtube.com/watch?v=hps9r7JZQP8

Sample to sample variability

Does Big Data have this?

"Claims to objectivity and accuracy are misleading" - boyd and Crawford

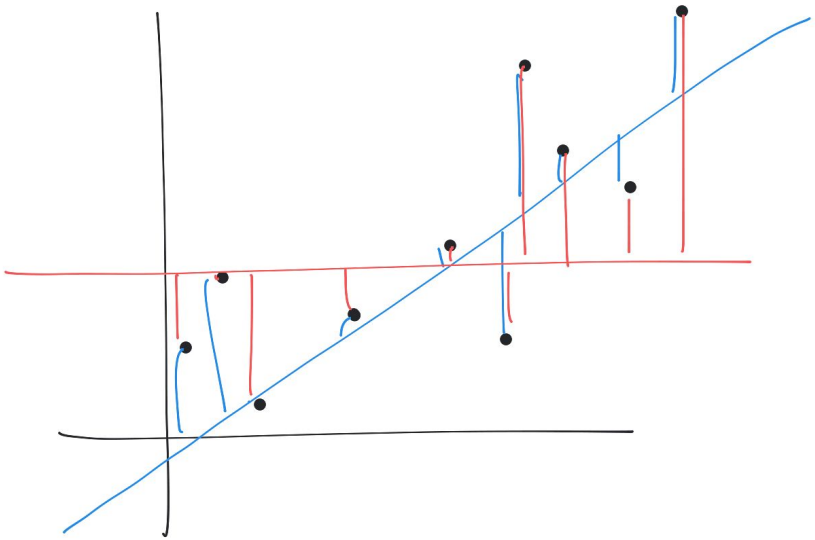


Measurement error and process variation

Is this part of data science?

(Bill or Cliff?)

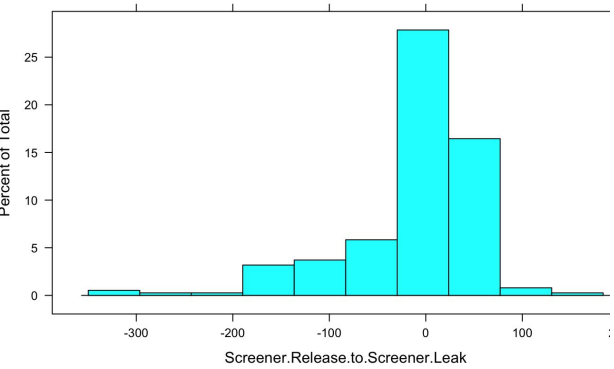
Modeling error

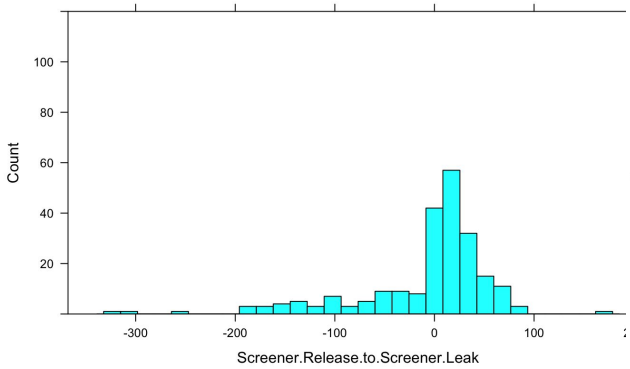


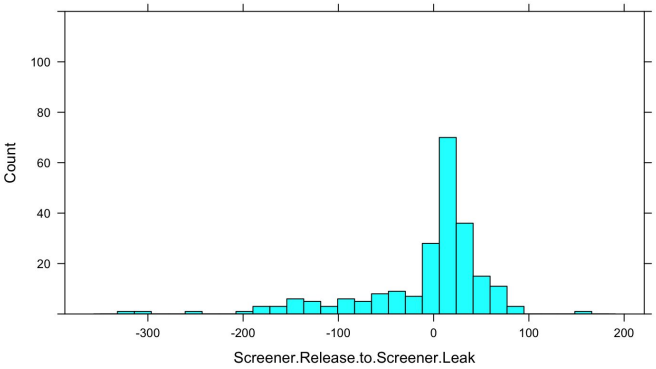
R^2 characterizes the amount of variability in the response that can be explained by the model

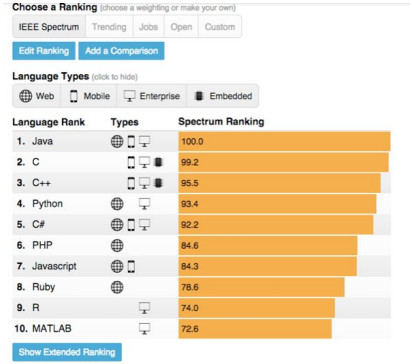
Sport modeling example from last Concord Consortium webinar. Were the predictions correct? How good was your model? How do you quantify this?

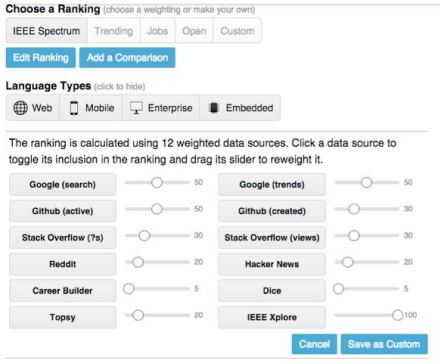
Variability due to parameter choices











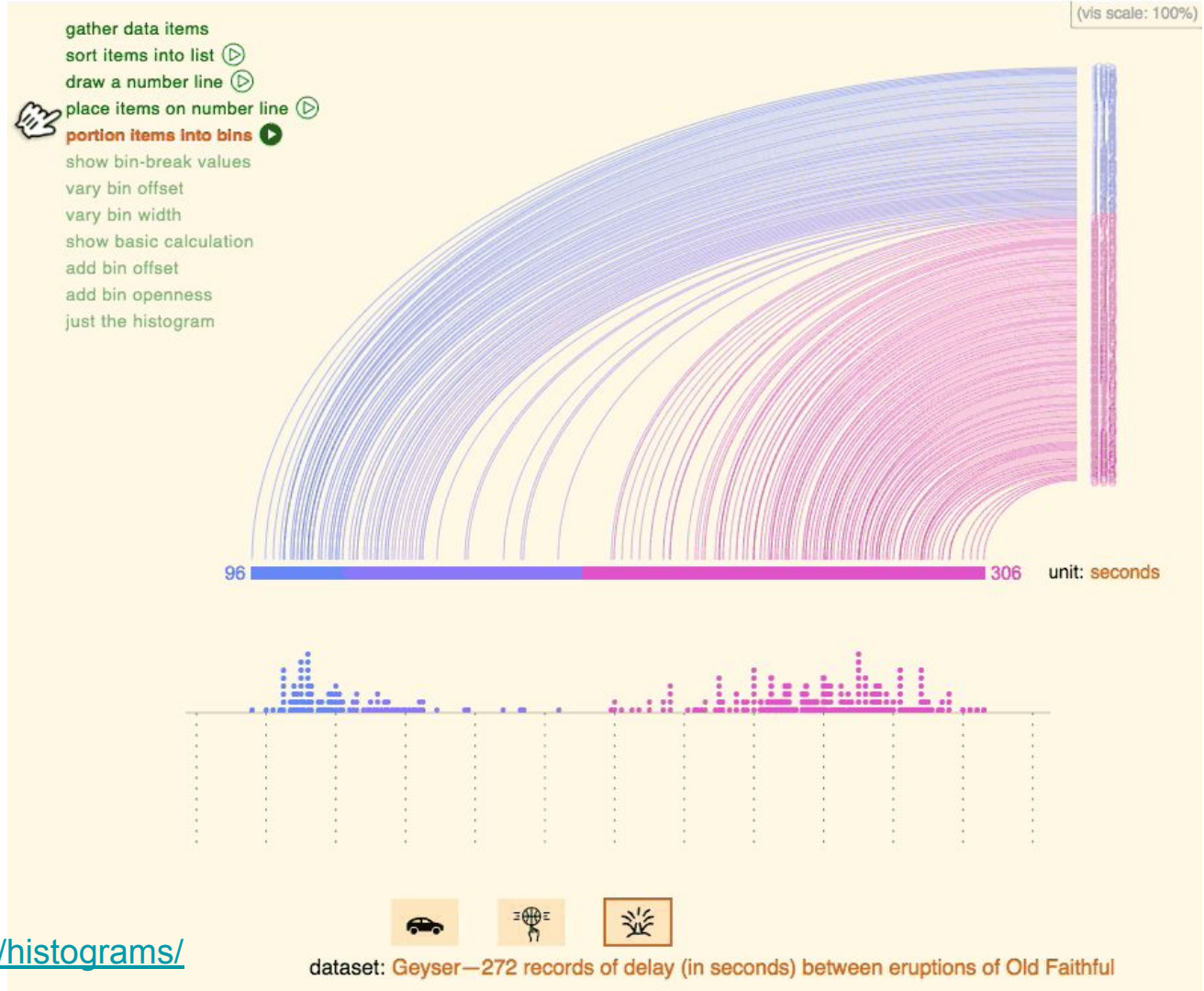
Nick Diakopoulos and Stephen Cass. “Interactive: The Top Programming Languages 2017.” http://bit.ly/IEEE_languagerank

R/ggplot2

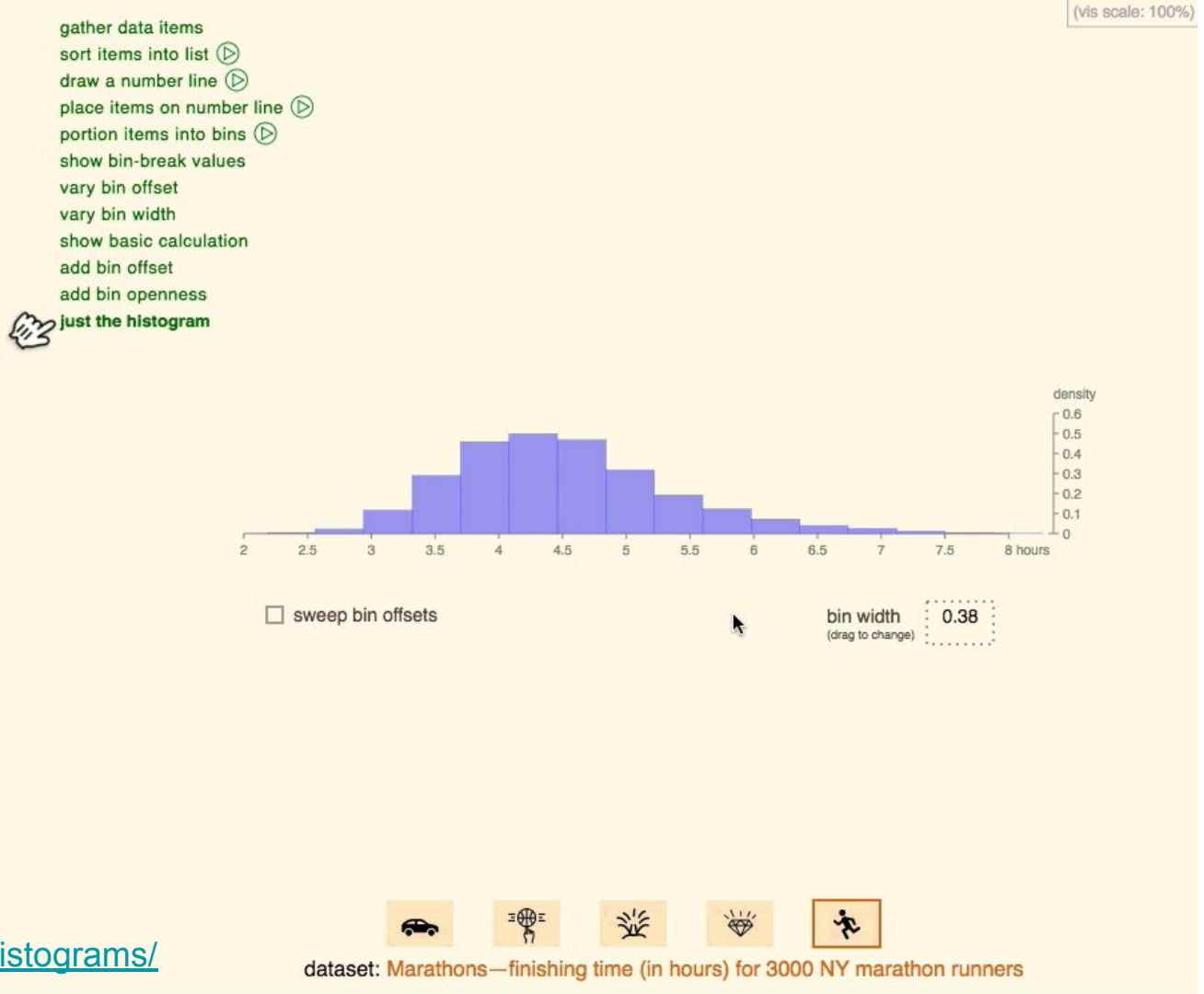
R/manipulate

Tableau

Fathom



http://tinlizzie.org/histograms/



http://tinlizzie.org/histograms/

Other forms of variability?

Thank you

Amelia McNamara @AmeliaMN
Smith College Program in Statistical and Data Sciences

flickr: elora.daphne