Sara Altman
STATS 337

Executive summary

Data analysis is a process carried out by minds and for minds, but I've found that the intersection between cognitive science and data analysis is usually not explicitly addressed. I read papers related to understanding data analysis and visualization as a process done by human minds. This involved researching computational thinking, cognitive and theoretical models of data science, perception of data visualization, and cognitive biases on tasks related to analysis.

Several of the papers I read explicitly discuss data analysis as a cognitive process (Grolemund & Wickham, 2014; Jolaoso, Burtner, & Endert, 2015; Wild & Phannkuch, 1999). Once you view data analysis as a cognitive process, you realize that this means we need to be aware of how cognitive biases, flaws, and features can influence the data analysis process. Grolemund and Wickham (2014) give an interpretation of data science as a sensemaking process. Sensemaking is a process for maintaining mental models of phenomena, called *schemas*. When people encounter new information, that information can either fit or conflict with their existing schemas. If the information conflicts with an existing schema, people either update their schema to reflect this new information or reject the information.

If Grolemund and Wickham are correct to characterize data analysis as a sensemaking activity, then data analysis is subject to various flaws in the sensemaking process, and, more generally, to various cognitive features. Some of these have the potential to infer with the ultimate goal of data analysis. To Grolemund and Wickham, this goal is the same as that of sensemaking: "to create reliable ideas of reality from observed data" (Grolemund & Wickham, 2014). For example, people are biased towards maintaining their schemas, even if they encounter information that contradicts those schemas. They are more likely to reject information than alter their schemas.

Perceptual biases and features are also important once you consider data analysis as an activity done by minds. Cleveland and McGill (1985) discuss the relationship between perception and data visualization. There are many ways to encode data in a visualization: position, length, angle, color hue, etc.. People can more accurately judge position, however, than color hue. Cleveland and McGill argue we should create graphics with these differences in mind, so that people understand the data underlying the visualization as accurately as possible. This implies, however, that people may misinterpret data if the wrong encodings are used, or may be biased towards some conclusion. Wu, Xu, Chang, & Wu (2017) develop a framework for studying this kind of bias. They argue for a Bayesian model of data visualization cognition, in which people have some prior probability on an event that gets updated according to Bayes' rule after seeing information encoded in a visualization. Assuming that model, the authors argue you can study visualization bias by setting individuals' priors, showing them visualizations that should update those priors in a certain way, and then measuring their posteriors. The amount by which the posteriors deviate from the Bayesian posterior is the bias. The authors conducted a simple experiment to demonstrate how this might be carried out. Although the process seems like it might not work very well in complicated scenarios, research about visualization bias is important if analyzers are going to draw actionable conclusions from visualizations.

Another area of interaction between minds and data analysis pertains to reproducibility. One way to interpret the reproducibility crisis, that I've seen often, is as a moral failing. The idea is that scientists, interested only in advancing their own careers, are purposefully running test after test looking for a low p-value, excluding data that doesn't fit their hypothesis, and obfuscating their analyses. However, another interpretation is that various cognitive features are at odds with reproducible science. For example, Gelman & Loken (2014) argue that problematic multiple comparisons are not necessarily carried out consciously and with the goal of finding a low p-value. Instead, they argue, it is common for researchers to make decisions that are dependent upon the data that they obtain, but those decisions feel to them like the correct decisions to make at the time. Researchers are biased to make decisions that will produce statistically significant results, but will justify those decisions as the correct ones to make from a scientific standpoint (Simmons, Nelson, & Simonsohn, 2011). This suggests that the problem is not bad actors, but researchers misattributing the reasons behind their decisions.

However, there are also positives to having a human mind involved in the data analysis process.
People can quickly detect patterns in visualizations and pick out anomalies in data. They can also easily consider how higher-level knowledge or outside information relates to the data, a skill that is currently difficult to automate (Samulowitz, Sabharwal, & Reddy, 2014). It is important to note, however, that humans are not good at data analysis by accident. The various activities conducted in data analysis (visualization, transformation, modeling, etc.) were created for humans to use, as we typically cannot gain insights from looking at raw data (Grolemund & Wickham, 2014).

Finally, once you view data analysis as a process conducted by individual minds, it also opens up questions about individual difference. Different analysts likely have different features that influence how they conduct data analysis, such as biases, starting beliefs, and dispositions (Wild & Phannkuch, 1999). Buja et al. (2009) develop one way for studying these differences. They construct a method called "The Rorschach," in which analysts are shown various plots. The underlying data for some of these plots contains a real pattern, and for some contains no pattern. For each plot, each analysts has to decide if there is some structure to the data. This test could allow researchers to understand individual differences in pattern attribution to data, as well as for individual analysts to have a better understanding of their own analysis behavior.

Recommended papers

Gratzl, S., Lex, A., Gehlenborg, N., Cosgrove, N., & Streit, M. (2016). From Visual Exploration to Storytelling and Back Again. In *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization* (pp. 491–500). Goslar Germany, Germany: Eurographics Association. Retrieved from http://data.caleydo.org/papers/2016_eurovis_clue.pdf

This paper discusses how to document and communicate data exploration so that others can follow what you've done and understand how you arrived at your conclusions. The authors argue that documenting a data exploration is important so that others can reproduce what you've done or check that your insights are justified.

The authors created an interactive system called CLUE which allows you to create and view "Vistories." Vistories are interactive visual summaries of your exploration that others can step through. You can see a demo here.

This paper did a couple things I haven't seen elsewhere: 1) emphasize that the insights from exploratory analysis should be reproducible, 2) take a "storytelling" approach to insights form exploratory analysis, 3) create a tool for viewing and creating summaries of data explorations.

I think the CLUE tool, or something similar, could be really useful for teaching exploratory data analysis, or just for presenting findings, because it allows users to easily see an example of an exploration and play with the data themselves. It would be particularly useful for people who don't yet know how to manipulate data through code themselves.

Grolemund, G., & Wickham, H. (2014). A Cognitive Interpretation of Data Analysis. *International Journal of Statistics*, *82*(2), 184–204. Retrieved from http://vita.had.co.nz/papers/sensemaking.pdf

The authors explicitly look at data analysis from a cognitive perspective, and discuss how data analysis is a sensemaking process. This perspective is important because human minds are the ones conducting and consuming data analyses, and therefore various features of human minds are relevant. We transform data, create graphs, etc. because our minds typically cannot draw out insights just by looking at raw data. Additionally, we need to be cognizant of the various biases present in human minds because those can influence which insights we get from our data. The authors write, "...a cognitive view suggests that cognitive phenomena may adversely affect data analysis---often in unnoticed ways."

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., & Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4361–4383. Retrieved from http://rsta.royalsocietypublishing.org/content/roypta/367/1906/4361.full.pdf

This paper develops two methods for using human minds like statistical tests, which is not something I have seen before. I think this is a useful paper to read if you have ever, as I have, worried that the conclusions that you draw from visualizations are not rigorous enough, or will always be wrong because you are an unreliable human.

The authors discuss two methods. The first is called the “Rorschach,” and involves showing an analyst plots with some real pattern amongst many plots with no pattern. The analyst, blind to what the real plot looks like, has to determine if each plot has a pattern or not. This serves as a way to understand that analyst’s tendency to find pattern where there is none and which features can influence that tendency. This idea---that different analysts might systematically differences in how they interpret plots---seems obvious in retrospect but is not something I had previously thought about.

The second method is called “The Lineup.” The analyst has to pick out the real plot from a lineup that includes some number of fake plots. If she can pick out the true plot, this provides evidence that there is some actual pattern in the data. If there are 20 plots, then her false positive rate is 5%, and thus you can assign a p-value of .05 to her discovery if she picks out the correct plot. In addition to this being an interesting way to incorporate visualizations into statistical testing, I think it’s also a good way to explain p-values to people who find them confusing.

Bibliography

1. Barba, L. (2016). Computational thinking: I do not think it means what you think it means. Retrieved from http://lorenabarba.com/blog/computational-thinking-i-do-not-think-it-means-what-you-think-it-means/

- Says it's "what we can *do* while interacting with computers, as extensions of our mind, to create and discover"
- Jeannette Wing article popularized the idea
    - Characterized computational thinking as "an attitude and skill set" that everyone can learn and use
    - Emphasis on using concepts of computer science to solve problems: abstraction, decomposition, recurions, etc.
        - "Thinking like a computer science"
- Original idea from seymour papert
    - Learning is enhanced when learn is engaged in "constructing a meaningful product"
    - Saw programming as communicating with the computer
        - "Learning to communicate with a computer may change the way other learning takes place"
    - Ideas
        - Power principle
            - Using comes before understanding
                - Natural mode is to use first and then build towards understanding
        - Project before problem
        - Media defines content
            - Can do different things with different media
        - Object before operation
            - Giving math idea a "thing-like representation" to help think about them (?)
        - Computer allow us to reverse the order of learning things
            - Without computers, needed to teach statics before dynamics
                - But with computers can expand what you can teach
- Argues that computational thinking is not thinking like a computer scientist

2. Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*(3), 199–231. https://doi.org/10.2307/2676681

- Discusses two different cultures with different views about statistical modeling
    - Data modeling culture: assume data are generated from a given stochastic data model, try to figure out that model, draw insights about data from that model

    - Algorithmic culture: Treat the data mechanism as unknown and focus on its inputs and outputs
        - Neural nets., decision trees
        - Focus on the properties of the algorithms
- Breiman argues for the algorithmic approach, but says that the statistical community has traditionally focused on the first, leading to bad consequences
    - These consequences include “irrelevant theory”, unjustified conclusions, statisticians working on problems that aren’t interesting
        - You can find models that fit data and draw conclusions about the model's mechanism, but if the model doesn't actually emulate nature well the conclusions might be wrong
- Discusses problems in checking data model fit
    - Thinks of goodness-of-fit tests as largely arbitrary, lacking power, and too subjective
- The approach of tackling problems by looking for a model, "imposes an a priori straight jacket that restricts the ability of statisticians to deal with a wide range of statistical problems"

3. Bryan, J., & Wickham, H. (2017). Data science: A three ring circus or a big tent? *Journal of Computational and Graphical Statistics*, *26*(4), 784–785. Retrieved from https://arxiv.org/pdf/1712.07349.pdf%0A

- Statistics departments are likely not at the center of data science learning *because* they tend not to focus on topics like “data preparation, presentation, and prediction”
    - Undervalue work that doesn’t involve mathematical statistics and new models
        - Creates lack of education around other skills involved in working with data

4. Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., & Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4361–4383. https://doi.org/10.1098/rsta.2009.0120

- Develop two ways to use visualizations like test statistics; human cognition as a statistical test
- Idea 1: The Rorschach
    - Show the analyst a plot with a pattern among a bunch of plots with no pattern; blind the analyst to what the real plot looks like
    - Use this to assess the individual analyst’s tendency to find structure in plots with no structure, as well as what features affect her tendency to see structure
        - These will be different for different analysts
- Idea 2: The Lineup
    - Have analysts pick out the plot of real data from among a bunch of null plots

        - If there are 20 plots, then the false positive rate is .05 and so, if the analyst is correct, can assign p = .05 to the discovery
    - Idea blend EDA and model diagnostics with confirmatory statistics
        - Each side is missing what the other has
            - With EDA and MD, can’t confirm discoveries or be sure you aren’t over-interpreting
            - With confirmatory statistics, you might miss the obvious

5. **Cleveland, W., & McGill, R. (1985). Graphical Perception and Graphical Methods for analyzing scientific data. *Science*, *229*, 828–833. Retrieved from https://pdfs.semanticscholar.org/565d/843c2c0e60915709268ac4224894469d82d5.pdf**
    - *Graphical perception* --- visual decoding of a graphic by a human visual system
        - Argues that understanding of graphical perception has traditionally been missing from graphics research/use
    - Identifies elementary *graphical-perception tasks* (e.g., decoding position, length, angle, area, or hue)
        - Figure out how good humans are at these different tasks and order them accordingly
            - For example, humans are better at judging position along a common scale than judging position along non-aligned scales
    - Use the ordering to figure out how to construct graphics
        - Encode data in a way that humans will perceive accurately
            - For example, use position before hue
        - This is a guide for more effective graphical perception
        - We will need for experimental perception research

6. **Dahly, D. (2017). The perils of exploratory data analysis. Retrieved May 30, 2018, from https://darrendahly.github.io/post/2017-02-03-exploratory/**
    - Asks the question, “Is it ok to conduct exploratory data analysis in order to generate hypotheses?”
        - He says no
            - Unscientific to look for areas where your “‘hypothesis works’” (quoting Wansink)
            - “It’s exactly the logic that practioners of so called alternative medicine use, when they point out that the treatment didn’t work, except in the people that it worked in.”

7. **Gelman, A. (2010). Exploratory and confirmatory data analysis. Retrieved from http://andrewgelman.com/2010/02/16/exploratory_and/**
    - Responds to Seth Roberts’s post about exploratory and confirmatory analysis
    - He agrees with Seth that exploratory and confirmatory analysis should go together
    - Disagrees that useful statistical research is low status research

    - Argues that everyone wants to be useful, but that it is hard to know what work is useful
- Disagrees that anyone can graph and transform their data, and instead says that graphics are harder than running a regression, etc.
    - One reason for this is that there’s no “default” graph to make for a given situation
- Also thinks things are changing and exploratory analysis is becoming recognized as more important

8. Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, *13*(4), 755–779. https://doi.org/10.1198/106186004X11435

- Exploratory methods are more effective when using with models; model-based inference is more effective if checked with plots
- Idea: formulate all plots as model checks
    - "new models and new graphical methods go hand-in-hand"
- A definition of model checking: comparison of data to replicated data under the model
    - Should include both exploratory plots and confirmatory calculations
- Theories of graphics
    - General guidelines (e.g., Tufte’s ink-minimizing rule)
    - Psychological findings
    - Theoretical arguments
    - Gelman’s idea: focus on the graph as a comparison/way to check models

9. Gelman, A. (2016). What has happened down here is the winds have changed. Retrieved from http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/

- Gelman is responded to an article written by Susan Fiske about the replication crisis in psychology
    - Fiske criticized the use of social media to give negative comments on published research
- Gelman thinks Fiske is following the “research incumbency rule,” which says that once an article is published in a respectable place it should be taken as truth, which doesn’t really have a place in psychology anymore given the replication crisis
- Discusses the history of thoughts on replication, from Paul Meehl’s warning in the 60s to today
    - Went from people not believing Meehl that there would be/was a problem to today when we are not surprised when things in social psychology don’t replicate

10. Gelman, A., & Loken, E. (2014). The garden of forking paths: Why multiple comparisons

can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Psychological Bulletin*, *140*(5), 1272–1280. https://doi.org/dx.doi.org/10.1037/a0037714

- “Researcher degrees of freedom” are not necessarily the result of conscious p-hacking, where a researcher does a bunch of analyses just trying to get a significant result
    - Instead, what can happen is that researchers conduct analyses that are dependent upon the data they got
        - If the data had looked different, they would have carried out different analyses
- Possible solutions
    - Preregistration
        - Problem: in some fields, you will already be familiar with the data or it will be impossible to form good hypotheses without first looking at the data
    - Pre-publication replication
        - Only possible in some fields
    - Analysis of all data instead of focusing on a single comparison
    - Keeping a firmer distinction between exploratory and confirmatory analysis

11. *Gra*tzl, S., Lex, A., Gehlenborg, N., Cosgrove, N., & Streit, M. (2016). From Visual Exploration to Storytelling and Back Again. In *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization* (pp. 491–500). Goslar Germany, Germany: Eurographics Association. Retrieved from http://data.caleydo.org/papers/2016_eurovis_clue.pdf

- Visual exploration needs to be documented to justify insights and make them reproducible
    - But this is hard to to
        - Videos and images often used but these aren’t great solutions
    - Present the CLUE model for this
        - Integrates exploration and presentation of discoveries from exploration
        - Provenance graph with all actions performed during exploration
- Three modes
    - Exploration
        - Explore the data yourself
    - Authoring
        - Create a story that others can step through
    - Presentation
        - The presentation is kind of like a slideshow you can step through to see their exploration

12. *Gro*lemund, G., & Wickham, H. (2014). A Cognitive Interpretation of Data Analysis. *International Journal of Statistics*, *82*(2), 184–204. Retrieved from http://vita.had.co.nz/papers/sensemaking.pdf

- Construct a theory of data analysis as a sensemaking process
- A theory of data analysis can be useful for several reasons
    - Might lead to better analysis techniques and practice
    - Improve data analysis education
    - Give conceptual answers to various questions: what is data analysis? What goals does it accomplish and how? When and why does it fail?
- Why use theories from cognitive science to understand data analysis?
    - Data analyses are constructed and understood by minds, and are therefore constrained by mental capabilities (to learn, understand, etc.)
- Data analysis as sensemaking
    - Schemas are mental models with information that people maintain through sensemaking. When you encounter new information, you can either realize it fits with your current schema or realize it is incompatible with your current schema and then reject/alter that schema.
- Difference between EDA and CDA under sensemaking theory
    - EDA is used to generate hypotheses from data; CDA is used to test hypotheses
- Bias in sensemaking and data analysis
    - If data analysis is a sensemaking activity, it will take on the flaws of sensemaking
        - There's a bias towards already accepted schemas
        - Sensemaking and DA cannot prove their conclusions
- "a cognitive view suggests that cognitive phenomena may adversely affect data analysis---often in unnoticed ways"
- Have to augment the sensemaking theory to accommodate the fact that data analysis works with measured data
    - Can think about what predictions different schemas will make about the data and then investigate those
    - Also need to think about how to compare patterns, because data will contain variation

13. *Ha*ig, B. D. (2015). Commentary: Exploratory data analysis. *Frontiers in Psychology*, *6*(August), 1–2. https://doi.org/10.3389/fpsyg.2015.01247

- Responds to Behrens et al. paper that criticizes EDA as an abductive process
- Haig says that they mischaracterize abductive processes and what they are actually describing is "inference to the best explanation"
- Characterizes EDA as an "empirical, descriptive, pattern detection process."
    - It does not in of itself create explanatory theories---that comes later.
        - EDA can help you detect new empirical phenomena, and then you make claims about those phenomena and then you try to understand those phenomena by constructing different explanatory theories
            - This is when abduction happens---not at the EDA stage

14. *J*olaoso, S., Burtner, R., & Endert, A. (2015). Toward a Deeper Understanding of Data

Analysis, Sensemaking, and Signature Discovery. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, & M. Winckler (Eds.), *Human-Computer Interaction -- INTERACT 2015* (pp. 463–478). Cham: Springer International Publishing. Retrieved from https://hal.inria.fr/hal-01599858/document

- 2 models of data analysis process
  - Sensemaking
    - Focuses on cognitive processes
  - Signature discovery process
    - More computational
- Interviewed and observed data analysts from different domains to get a sense of how well these models fit the behavior of data analysts
  - Found that neither fully captured the processes of the analysts
    - People didn’t follow the linear process outlined in the signature discovery model
      - Also didn’t do certain steps like inventory of observables or exploring the data
  - Results did vary by domain
- Design implications
  - Tracking provenance is important
    - Participants remarked that recalling process is hard
  - Importance of hypotheses
    - Often not formal

15. Lundberg, C. G. (2000). Made sense and remembered sense: Sensemaking through abduction. *Journal of Economic Psychology*, *21*(6), 691–709. Retrieved from https://pdfs.semanticscholar.org/cce4/b4fa69ed4c7cf997f1fbf38542c247bb19ea.pdf?_ga=2.80098007.1354196306.1528325345-1426359698.1528325345

- Discusses sensemaking as an abductive process
- Sensemaking vs. nonmonotonic reasoning
  - Nonmonotonic reasoning = drawing conclusions which can be invalidated by new information vs monotonic reasoning, which assumes that once a conclusion is reached it can’t be altered with more information or reasoning

16. McKenna, S., Lex, A., & Meyer, M. (2017). Worksheets for Guiding Novices through the Visualization Design Process. Retrieved from http://arxiv.org/abs/1709.05723

- Worksheet for four different design activities involved in vis: *understand, ideate, make, deploy*
- Understand sheet
  - Asks about challenges and users, questions, design requirements
- Ideate sheet
  - Select a design requirement, sketch some ideas

- Make
    - Create a prototype
- Deploy
    - Finalize audience
    - Fix and refine

17. **McInerny, G. J., Chen, M., Freeman, R., Gavaghan, D., Meyer, M., Rowland, F., … Hortal, J. (2014). Information visualisation for science and policy: engaging users and avoiding bias. *Trends in Ecology & Evolution*, *29*(3), 148–157. https://doi.org/https://doi.org/10.1016/j.tree.2014.01.003**
    - Call for more visualization in science
    - Think that ignoring visualization could lead to missed discoveries

18. **Peng, R. D. (2017). Comment on "50 Years of Data Science." *Journal of Computational and Graphical Statistics*, *26*(4), 767. Retrieved from https://www.tandfonline.com/doi/pdf/10.1080/10618600.2017.1384734?needAccess=true**

- Data cleaning and visualization are hard to teach because they are difficult to generalize

19. **Roberts, S. (2010). Exploratory Versus Confirmatory Data Analysis? Retrieved from http://archives.sethroberts.net/blog/2010/02/15/exploratory-versus-confirmatory-data-analysis/**

- He takes issue with Tukey's distinction between exploratory data analysis and confirmatory data analysis, as he views them more collaboratively.
- Draws a distinction between "low-status" and "high-status" work instead, where "low status" work refers to data transformations and graphs, and "high-status" work is more theoretical and abstract. He argues that statistics professors pursued high-status work to impress their peers, and that explains why exploratory data analysis has largely been neglected.
- Also thinks that the "low-status" work is actually more valuable than the "high-status" work, but that's also kind of the point---it's more prestigious to do less useful things.
- Comments: I think his useful/not useful distinction ignores other factors at play. It's seems that there's a clearer path forward if your job is to improve theoretical aspects of statistics versus aspects that rely on many different moving parts, like graphing (have to consider software, theory, cognition, etc.). Also, I think there's a bias to consider heavily mathematical research more prestigious and more difficult than less mathematically inclined research, or to consider work not fully grounded in math to not be truly scientific. I don't think you can explain it all in terms of usefulness/not usefulness.

20. **Samulowitz, H., Sabharwal, A., & Reddy, C. (2014). Cognitive automation of data science. *ICML AutoML Workshop*. Retrieved from http://www.cs.toronto.edu/~horst/cogrobo/papers/CADS.pdf**

- AutoML methods don't entirely capture how humans conduct machine learning
    - Humans can use higher-level knowledge to guide what they do; current autoML techniques can't and are entirely dependent on the data
- Define cognitive automation as having the following three properties
    - Integrates knowledge from data sources, past experience, and current state
    - Interacts with user and reasons based on those interactions
    - Can generate novel hypotheses and techniques and then test those hypotheses

21. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

- Even though there's the endorsement of a maximum false-positive rate of 5%, current practice of science actually makes false positives more likely
- Argue that "researcher degrees of freedom"---essentially decisions that researchers make about how to collect and analyze data---are to blame
    - Researchers are biased to make the decisions that will produce statistically significant results, and will justify those decisions as the correct thing to do
- Provide suggestions for both researchers and reviewers
    - Researchers: decide the rule for stopping data collection before starting collection; list all variables collected; report all conditions
- Things that they don't think will help: Bayesian statistics, corrected alpha values, conceptual replications (because you don't have to follow the original decisions exactly)

22. Unwin, A. (2001). Patterns of Data Analysis? *J. Kor. Stat. Soc.*, *30*(2), 219–230. Retrieved from https://pdfs.semanticscholar.org/89bb/084ac04d15cb2d45594096d23a30ce4fcd8f.pdf

- Says there's no real theory of data analysis
- Create a theory from patterns for data analysis

23. Why you should master R (even if it might eventually become obsolete). (2016). Retrieved from http://www.sharpsightlabs.com/blog/master-r-obsolete/

- By mastering one language, you will learn higher-level concepts that are generalizable to other languages
    - Language agnostic skills
        - E.g., if you know how to use visualization to gain insights in R, can then do in any other language

24. Wickham, H., & Grolemund, G. (2017). Exploratory Data Analysis. In *R for Data Science*. O'Reilly Media. Retrieved from http://r4ds.had.co.nz/exploratory-data-analysis.html

- Describe the process of EDA as:
    - Generate questions
    - Look for answers to those questions by visualizing, transforming, modeling
    - Then refine questions + make new ones based on what you find
- “EDA is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you.”
- Need to do EDA for every analysis to at least investigate data quality
- Goal is to start to understand data
- Things to do
    - Understand variation
    - Visualize distributions of your variables
    - Figure out what values are typical and unusual for different variables
    - Investigate and deal with missing values and unusual values
    - 2d visualizations
    - Understand patterns in data with models

25. Wild, C. J., & Phannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, *67*(3), 223–248. Retrieved from https://iase-web.org/documents/intstatreview/99.Wild.Pfannkuch.pdf

- Discuss a 4 dimensional framework that characterize the thinking they observed in students and statisticians
- Dimension 1: Investigative cycle
    - When you plan what you are going to, what questions you are going to ask
- Dimension 2: Types of thinking
    - Various types of thinking involved in statistical thinking
        - Recognize why you need data and why personal experiences/anecdotes are not enough
                - Transnumeration: "numeracy transformations made to facilitate understanding"
                - Variation---a lot of uncertainty comes from omnipresent variation
                - Set of models---"all thinking uses models"
        - Dimension 3: Interrogative cycle
            - Always active throughout the process
            - Steps
                - Generate possibilities
                - Seek information to investigative those possibilities
                - Interpret the results of the seeking
                - Criticize the results
                    - Do the conclusions/information make sense? Look for internal reference points
                - Judge---this is the “decision endpoint”

        - Dimension 4: Dispositions
            - People have different dispositions that can influence how they approach statistical thinking
            - Curiosity, awareness, engagement with the data/domain, logicalness, skepticism
        - Variation as the cause of a need for statistical thinking and source of uncertainty

26. Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1881), 3717–3725. https://doi.org/10.1098/rsta.2008.0118

- Abstraction is the foundation of computational thinking
    - Thinking about what details can be ignored and which are crucial
    - Because computational abstractions often rely on hardware, you also need to worry about implementation, edge cases, and failure cases
- Computational thinking is now everywhere (in all sciences, starting in humanities as well) and therefore everyone should be comfortable with it
    - This implies that we should start teaching/encouraging computational thinking in schools from a young age
        - We should think about what parts of computational thinking are innate (or present from a very young age), what progression of skills to teach, how to integrate it with learning about various tools
- Comments: many parts of this paper are vague, and I still don't have a good idea of what teaching computational thinking to young children would look like. I'm also not really clear on what Wing means when she says "computational thinking", besides abstraction. She says that computational thinking builds on abstraction, but doesn't really give specifics.

27. Wu, Y., Xu, L., Chang, R., & Wu, E. (2017). Towards a Bayesian Model of Data Visualization Cognition. *VIS 2017: Dealing with Cognitive Biases in Visualisations*. Retrieved from http://decisive-workshop.dbvis.de/wp-content/uploads/2017/09/0110-paper.pdf

- Idea: study how cognition of data visualization by comparing people's probabilities of some event before and after seeing a visualization to Bayesian priors and posteriors
    - Bias will be difference between stated posterior and the Bayesian posterior
    - Have to assume that Bayesian models of cognition are appropriate
- Dealt with challenge of accurately eliciting priors by explicitly giving participants the priors
- Relationships to other theories
    - Prospect theory might come into play when sure losses or sure gains are present

- When participants don't trust the data, they might discount the observations
- People might have inherent priors beyond the explicitly stated probabilities