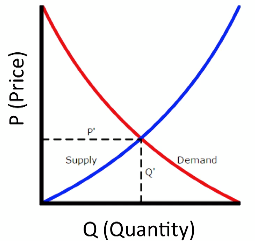
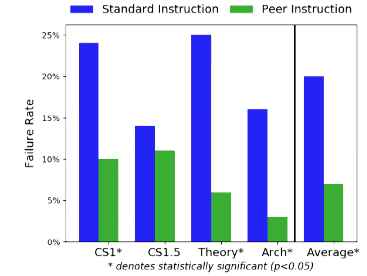
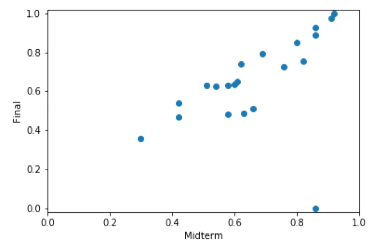
* Quote from Google's chief economist back in 2008:
* "The ability to **take data**, to be able to **understand it,** to process it, to **extract value** from it, to **visualize it**, to **communicate it**, that's going to be a hugely important skill in the next decade, b/c now we really do have essentially free + ubiquitous data.
* This quote because it points out the need to data scientists + statisticians well before big data + data science became buzzwords
* Statistics is one of the core areas of data science, + using statistics is 1 way to try + understand data.
* The other part of understanding data is visualization.
* Processing data + extracting value from it/understanding it often requires visualizing it to gain that kind of insight
* When we want to communicate results, you'll often want to use visualizations to do that too.
* So visualization is central to the skills that he describes.
* There are 2 good definitions of defining visualization.
* "The use of CPU-supported, interactive, visual representations of abstract data to amplify cognition."
* "The representation + presentation of data to facilitate understanding."
* Both discuss the importance of representing + presenting data, each of which is central to data visualization.
* Key part of data visualization = **Improving how we understand + think about the data.**
* If we think about it, a pretty large chunk of our brain is dedicated to vision + we've got some really powerful processing associated w/ that vision in our heads.
* In contrast, we're not very good at deciphering raw data.
* W/ 100 raw XY coordinates, it could take hours working by hand to see what they mean.
* But w/ a plot w/ 1000 points on a trend line, one can often tell you in seconds what the data is telling us.
* The human visual cortex is really talented.
* Another key point is that visualization of data means *nothing* w/out the **context** of the data.
* When we didn't know what X + Y were, we didn't shouldn’t really care about the data, b/c context is what makes us value a visualization, + gives us the ability to dig deeper into the data, looking to explain outliers, + also to be able to come to conclusions about the data itself.
* There are 2 key ways to categorize data visualization.
* Whether it’s conceptual or data-driven.
* Visualization of *concepts* is important, particularly when we aim to explain how things work conceptually.
* Ex: Economists seeking to visualize the notion of the classic supply + demand curve w/out using real data + then back it up w/ data supporting the concept from, say, Uber surge pricing.
* B/c data-driven is where we spend much of our time as data scientists, let's look at a second categorization in that context.
* Declarative or Exploratory
* **Declarative** = a point when we've analyzed the data + we have data-supported conclusion we wish to articulate to our audience.
* At the point of presenting, we want our visualization to convey this conclusion to the observer in the most straightforward possible manner.
* **Exploratory** = We spend much of our time exploring data + visualization plays a key role in it
* Visualizations often encourage + enable us to look deeper into the data.
* Ex: Classic supply-demand curve from economics explains the relationship between supply (quantity of goods available) + demand (amount someone is willing to pay for the item)



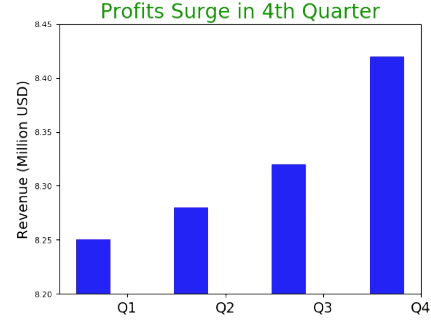
* When quantity is low, demand is high. As quantity goes up, demand will decline.
* This is conceptual, + these lines are hypothetical.
* Ex: Peer instruction pedagogy resulted in a significant decrease in failure rates for students, relative to standard instruction in CS classes at UC San Diego.

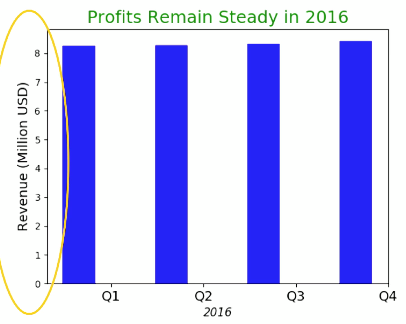
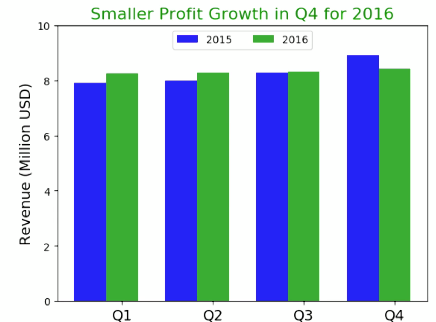


* Goal in presenting these results was to convey to the reader the impact peer instruction had on failure rates in classes.
* This has failure rates from a # of classes at UC San Diego, where instructors either taught using the active learning pedagogy of peer instruction, or using standard lecture-style pedagogy.
* Ex: Correlation between midterm and final exam scores is likely explorative.



* Results would likely cause me to explore the data more.
* W/ explorative data visualizations, don’t spend as much time polishing the appearance, so long as one can interpret it, that's fine.
* Likewise, we often want to be able to quickly plug in different parts of a data set into the figure to explore different relationships (HOMER scores vs final exam, rather than midterms).
* Exploration is really at the heart of the data science process,
* When we're finding outliers or trends, we're often using visualization tools.
* And those visualizations lead us to dig deeper into the data.
* And as we dig deeper, we do the same thing.
* Ex: Looking at data distributions using histograms, exploring relationships between variables, or seeking other trends.
* Ultimately, we find ourselves zooming in and out of various parts of the data set as we try to gain a better understanding.
* This process of zooming in + out of the data is almost always accompanied by + facilitated by data visualization.
* Data Visualizations needs to be (Andy Kirk):
* **Trustworthy 🡪** data presented is honestly portrayed
* *Take trust seriously, it’s hard to earn + easy to lose*
* Ex: find 1 seemingly intentionally misleading claim/figure, can now doubt everything an author has said
* *Honesty is not limited to visualization stages 🡪 should be present everywhere in the data science process*
* Need evidence for relationships being displayed



* Notice the y-axis only jumps about 0.5% between quarters 🡪 zoomed-in
* Not really a surge 🡪 dishonest/misleading (y-axis font seems small, maybe on purpose
* 2% is not a surge, would expect a Q4 surge in profits due to holiday season
* More honest graph may have the prior year’s profits to compare against (YOY)
* 
* Looks like steady revenue in 2016, so we can rename the plot
* Should plot prior year data if we have it
* 
* Do indeed see a surge in Q4 2016 profits (12% boost between Q1 and Q4), but not in 2015 (only 2% growth)
* Want to dive deeper into the data to figure out why we didn’t see the Q4 revenue boost again in 2016
* Want to do this before presenting results so we have an answer as to why this happened
* **Accessible**
* Focus on and know the audience and their ability to use a visualization you make (how they’ll use it
* Understand what they understand and know how they’d interpret results
* A graph could be useless or great, depending on the audience, or depending on the intent of what it’s trying to display
* Understand the purpose of a visualization
* Are we presenting it or exploring it
* Helps us craft if in an appropriate manner
* Also take into account the expected amount of time for an audience to read + understand the results from a visualization
* **Elegant**
* More time is put into visualization elegance when *presenting* results
* Can be nice in exploration, but not necessary (NYT article vs. 1st exploratory analysis)
* Make sure additional features to a graph only add value and do not detract from it or distract an audience
* Overall, style and beauty is subjective
* So, *focus on what is relevant + remove anything not adding to the figure*
* Make the design invisible so that a user can take as much from the figure as possible w/out being distracted (NOT the same as minimalism)
* *Be stylish if possible*
* *Think about decorations (but make sure they’re not contrary to honesty)*
* Ex: shaping a gauge graph about blood donations in the shape of flowing blood for a Halloween fundraiser ad vs. in a scientific paper
* May bring extra attention to the graph, *depending on the audience*