BELIV Provocations Fireside Chat

Tamara Munzner

Department of Computer Science

University of British Columbia

BELIV 2020 Provocations Fireside Chat
October 25 2020, Utah/virtual

http://www.cs.ubc.ca/~tmm/talks.html#beliv20fireside **@tamaramunzner**

Provocations

- agree ✓
  - applied vis research (design studies) are n=1 case studies

- disagree ✗
  - all implications of that framing
    - *case studies are near-useless "anecdata"*
  - many other things
    - methodological & rhetorical

What Do We Actually Learn from Evaluations in the "Heroic Era" of Visualization?
Michael Correll
BELIV 2020 Position Paper
https://arxiv.org/abs/2008.11250

Metaphors matter

- viz researcher = biologist
  - in design studies, field biologist
- collaborators = specific group of animals
  - mob of meerkats
- domain = species
- task abstraction = behavior
- analysis process = context

http://nomadnaturalist.com/archives/761

https://www.mentalfloss.com/article/62843/10-fascinating-furry-facts-about-meerkats

Metaphors matter

- does case study merit a paper?
  - should biologist publish **every** time they observe animal behavior?
    - yes!
    - iff they learn something new to biology - they usually do
- do we only need one case study per domain?
  - should biologist publish **only** if they identify a new species?
    - no!
    - existence proof of species is cool but rare
    - document and analyze existence of meerkat behaviors in contexts
      - how do these meerkats act in the summer in the desert in the presence of coyote predators?

viz DS researcher = field biologist

collaborators = group of animals

domain = species

task abstraction = behavior

analysis process = context

Metaphors matter

- are abstractions for tasks & data useful?
  - *MC: no! "avoid the idiosyncratic and often impenetrable "task analyses" that generate the n = 1 paper experimental conditions for our work"*
  - TM: yes! exactly need to transfer between contexts
    - avoiding them would eviscerate a DS paper
      - miss the whole point if you skip abstractions!
  - what could we learn from n=1, single mob of meerkats?
    - what are their behaviors and how does context affect them?
      - do meerkats act differently in deserts than fields? in summer than winter? from badgers or shrews?
    - develop theories that might transfer beyond specific setting
      - what matters: seasonality? terrain? body size?

viz DS researcher = field biologist

collaborators = group of animals

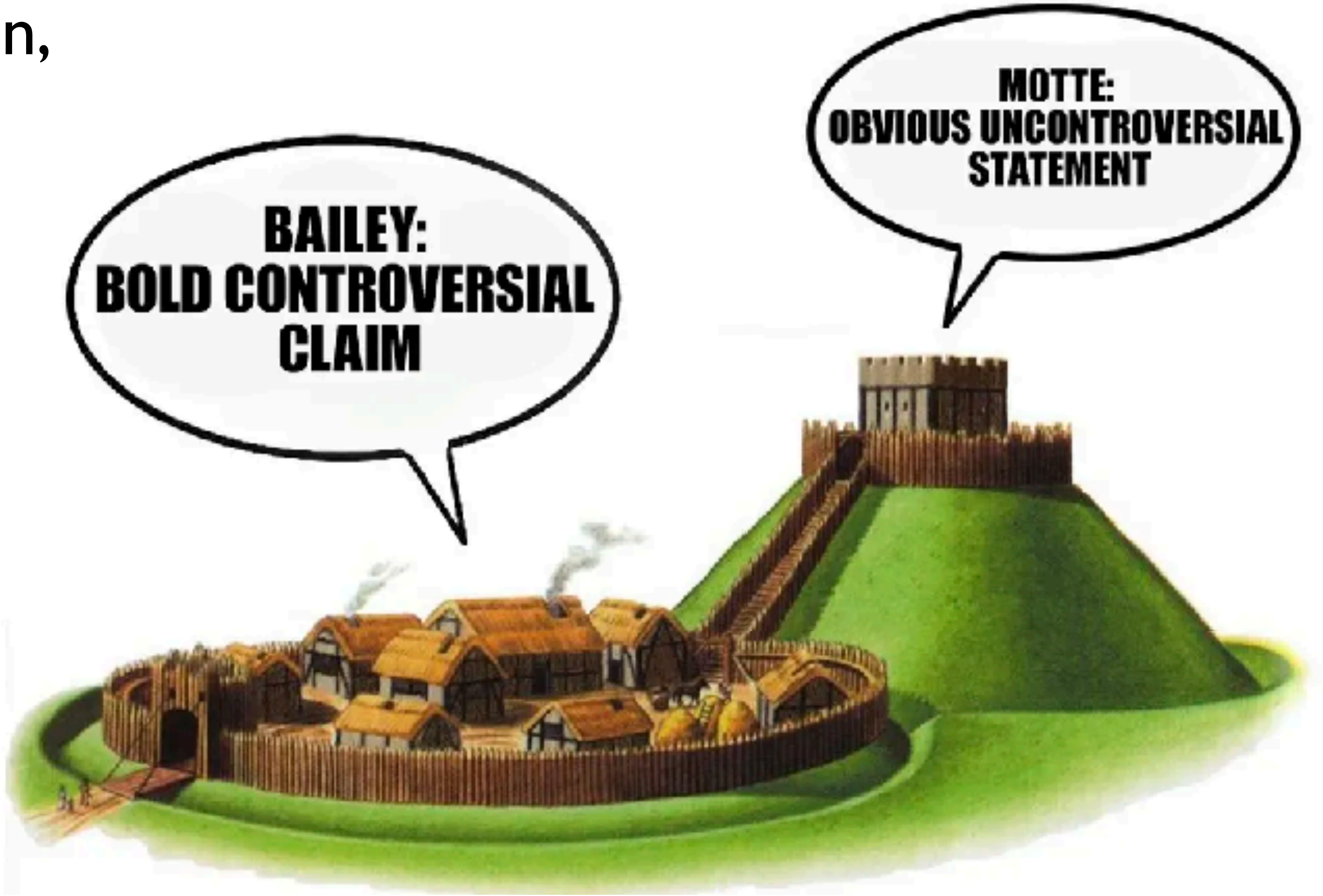
domain = species

task abstraction = behavior

analysis process = context

Motte-and-bailey fallacy (aka bait-and-switch shenanigans)

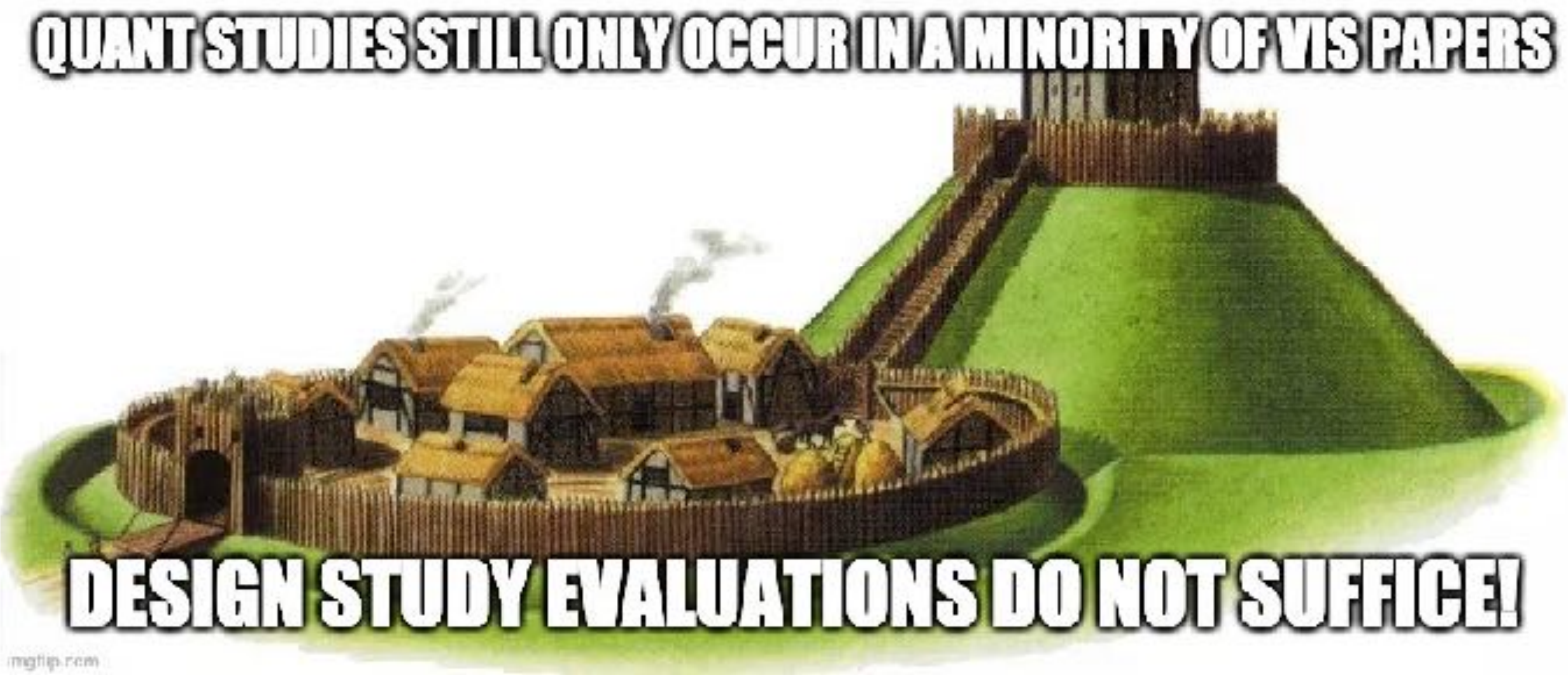
- conflating two positions with similar properties
  - one modest and easy to defend (the "motte")
  - one more controversial (the "bailey")
  - arguer first states controversial position,
    but when challenged states
    they're advancing modest position



https://bigthink.com/culture-religion/motte-bailey-meme

Motte-and-bailey fallacy (aka bait-and-switch shenanigans)

- qual vs quant methods
  - bailey: (earlier) claim that design study evaluations do not suffice
  - motte: **quantitative** studies only occur in minority of **all** paper types
  - reality: not relevant, since almost all **design study** eval with **qualitative** studies


QUANT STUDIES STILL ONLY OCCUR IN A MINORITY OF VIS PAPERS
DESIGN STUDY EVALUATIONS DO NOT SUFFICE!
imgflip.com

https://imgflip.com/memegenerator/172979893/motte-and-bailey

Qualitative research methods misconstrued

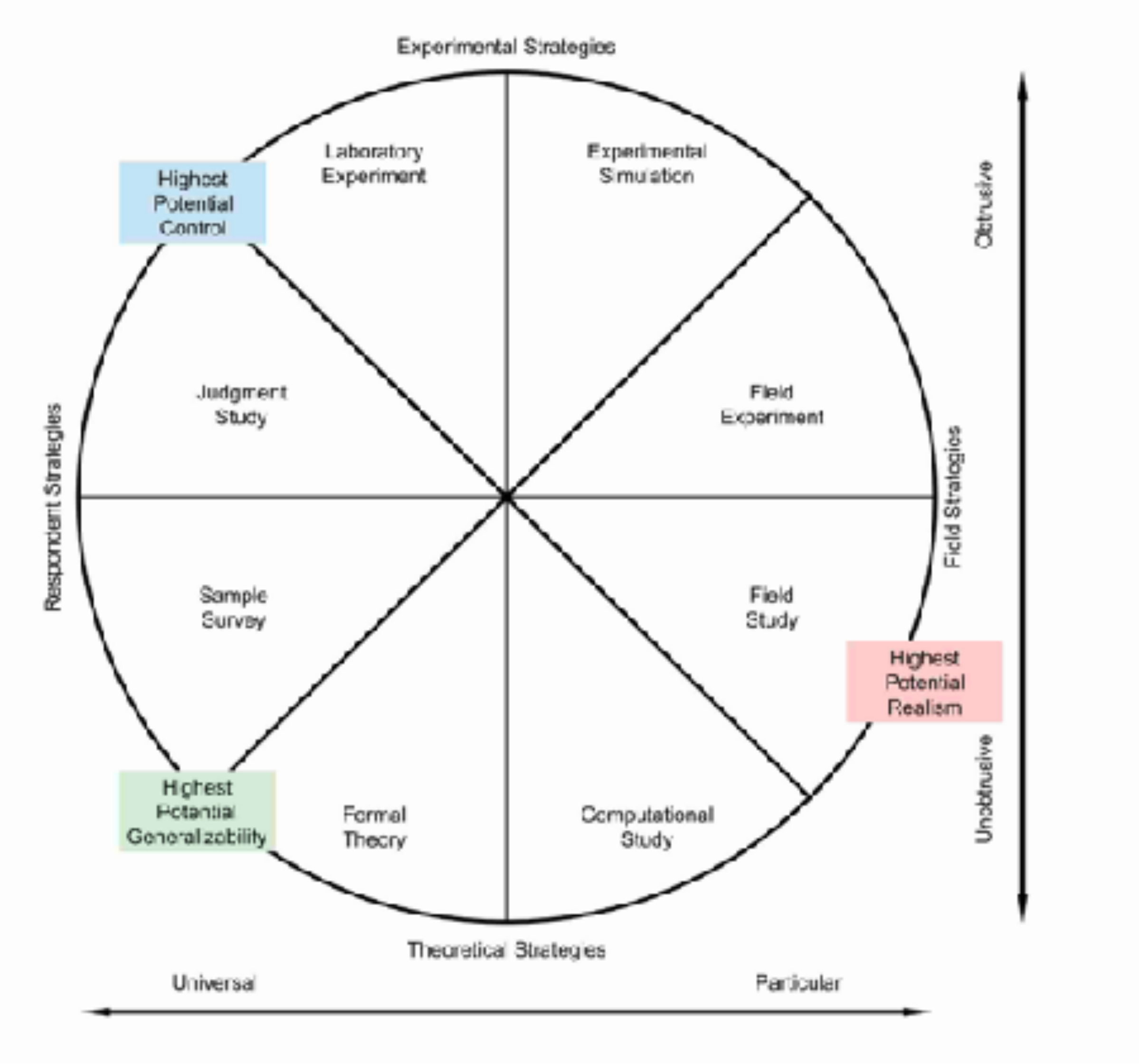
https://www.flickr.com/photos/yorkntu/48279104311/

- *MC: existence proofs are small contributions*
  - no!
  - existence proofs can require dramatic shifts our theories
  - biologist: wow, I just saw this meerkat do a backflip!
    - now can disprove previous theory that it's anatomically impossible
- *MC multiverse thought experiment setup*
  - *they cure cancer, they thank you in Nobel Prize speech, then you get study email*
    - *your favorite eval method: "quant, qual, insight-based, whatever floats your boat"*
  - no! setup is **not** agnostic to eval method
    - no surprises in email if qual field study w/ longterm deployment after iterative refinement
  - no! they wouldn't have thanked you in prize speech if your system was crap
    - rules out half the scenarios
    - when deploy in field, they can vote with their feet (in contrast to quant lab studies)

Methods matter: qualitative, quantitative, mixed methods

- no single method answers all questions
  - science is all about choosing the right method!

Methodology matters: Doing research in the behavioral and social sciences.
Joseph E McGrath.
In Readings in Human–Computer Interaction. Elsevier, 152–169, 1995.



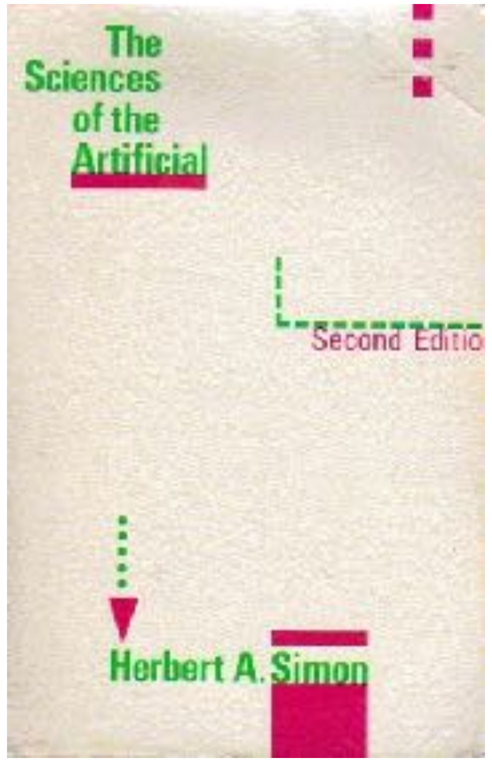
- plug for BELIV 2018 paper
  - detailed discussion of qual, quant, & mixed methods & their use in visualization

How to Evaluate an Evaluation Study? Comparing and Contrasting Practices in Vis with Those of Other Disciplines.
Anamaria Crisan and Madison Elliott.
BELIV 2018
https://amcrisan.github.io/assets/files/papers/beliv-2018.pdf

Discussion Slides

Is system building a "heroic" (aka excessive) measure?

- MC: extreme measure, we should do less of it
  - need more theory before we do more practical work
- TM: fundamental way to engage & learn in design/engineering
  - DSM: operating in huge tradeoff spaces so cannot just optimize, must satisfice
    - need to build and iteratively refine to get it right, theory alone isn't enough
    - design as crucial driver to **develop** theory!
      - continues to be most important opportunity for applied vis research
  - Herb Simon, Sciences of the Artificial, 1969
    - coiner of *satisficing*, only Turing-Nobel laureate
    - engineering as instance of design
      - *"how to make artifacts that have desired properties & how to design"*
        Ch 5, The Science of Design: Creating the Artificial
    - key difference from natural sciences: must build before can observe



The Sciences of the Artificial

Second Editio

Herbert A. Simon

Backup Slides

Strawman arguments (aka nobody said that!)

- "*design study... evaluated by n=500 Mechanical Turk workers...*"
  - no
  - almost nobody does that. quant MTurk studies are mismatch for DS
    - they mostly do qual evaluation. if it's quant, it's of domain experts not MTurk randoms

- "*emphasis on individual herculean actions by individual actors...* "
  "*... assumption that other labs would not have produced the same positive results*"
  - no
  - that's not heroism - it's the polar opposite, realistic humility!
    - noting that another researcher wouldn't recreate the identical system is basic tenet of qualitative research

Misapprehensions (aka we said the opposite of that!)

- "*did they really need a new system?... wrong questions for the heroic age*"
  - Huh?! These are **precisely** the questions we ask!
    DSM Pitfall #6: no need for visualization
    DSM Pitfall #9: no need for change: existing tools are good enough

- "*standard design study procedure doesn't necessarily advance field*"
- "*lacking... empirical and rhetorical tools to supplant the old theory with the new*"
  - Huh?! DSM Pitfall #27: **don't fail to advance theory**, **must improve guidelines**
    - confirm, refine, reject, propose theory as a fundamental expectation for publication!
    - what distinguishes practice from research
- "*need... greater willingness to detect (and report on) our design failures*"
  - Huh?! documenting iterative refinement **does** report on failures along the way

Misapprehensions (aka we said the opposite of that!), cont.

- MC: can we learn from "we built it and they liked it"?
- TM: misconstrues DS
  - it's not "did they like it?"
  - it's "did it help them?"

Other thoughts

- we each argue extreme case
  - MC argues about worst possible & TM argues about best possible
    - what about common case in the middle, some flaws and some strengths?
    - methods vs their execution - any method can be carried out poorly
- do we actually do too little comparison?
  - MC: yes, need to compare to Excel 'placebo'
  - TM: no, previous workflow (plus variations during iteration) covers a lot of ground
    - Excel may well be something they're already using
- expense of bespoke solution
  - yes, very high cost.
    - worth it if improve theory in addition to building practical tool?
- where's the bar for publication?
  - does get higher as years go by. will it ever get so high can't publish?
    - I don't know, but not for a while at least

Dubious thought experiments, prolog

- MC argues against three tacit premises
  - kind of work we do suggests kind of evaluations to perform and metrics to use
    - yup! that's not tacit at all, cornerstone of my Nested Model
  - evaluations can succeed or fail in illustrating utility
    - yup!
  - success or failure of evaluation is informative for the field
    - yup!
- MC claim: evaluations may be uninformative even if designed appropriately
  - no. thought experiments do not hold up.
  - snark about magical thinking and Tarot cards isn't enough to make the case

Dubious thought experiments, I

- Unique
  - *MC claim: problem so idiosyncratic nobody else can benefit from your solution*
  - TM counter: I don't believe there's any such thing
    - always can abstract up from domain specifics! design studies without abstractions get rejected
- Obvious:
  - *MC claim: obvious how to go from textbook guidelines to a system*
  - TM counter: no, no, no. it's a huge tradeoff space!
    - I should know, I wrote textbook & I teach out of it & do in-class exercises
    - let me tell you, students sure aren't channelling me (if only!...). many variants proposed.
- Worse Than Baseline:
  - *MC claim: almost never test against baselines like Excel ("placebos")*
  - TM counter: yes we do! many design studies compare against previous workflows
    - claims of success based on massive speedups (hours vs days). Excel is workhorse not placebo

Dubious thought experiments, 2

- Detestable
  - *MC claim: they perform better but they absolutely hate it*
  - TM counter: in real world, they just wouldn't use it. deploy requirement is high bar!
    - DSM PF-25: lack of case study
      - usage by developers much weaker validation than usage from domain experts.
- Serendipitous
  - *MC claim: one anecdote of successful use shows nothing, maybe just got lucky. insight found by chance, if sliders set differently wouldn't have seen it*
  - TM counter: case studies report on weeks or months of use, not single thing
    - mostly about systematic speedup of workflow, not *just* single glorious insight
  - *MC claim: system worked for designed tasks, but they didn't do those*
  - TM counter: iterative refinement to understand tasks is cornerstone of DS
  - **TM anti-counter: nevertheless, this critique has some merit**

Dubious thought experiments, 3

- Super Serendipitous
  - *MC claim: system so wrong and buggy they figured it out just to disprove you*
  - TM counter: <eyeroll>

Qualitative research methods misconstrued, cont

- other quant/qual swapperoos
  - "*we're just showing that our design seems to do what we claimed it does, which may not require any sort of quantitative evaluation at all*"
  - qualitative evaluation is exactly required to show that claims are correct.
    - of course doesn't require quant evaluation, that's why we don't do it!