Talking Bayes to Business

An A/B testing use case

About me

- Bayesian by belief - Frequentist by practice
- I call myself a “Data Scientist” because I know math, stats & just enough programming to be “dangerous”
- Currently focused on forecasting & causality (for elasticity, optimisation, etc.) and NLP for recommendations & search

Find me on @BigEndianB, Linkedin, github.com/ytoren

Agenda

- Motivation: Is it working?
- Getting the right answers with Bayes: concepts & toolkits
- Beyond A/B testing (with examples)
- Problem Forward vs. Solution Backwards

Meet Nadia

Nadia is a product manager.

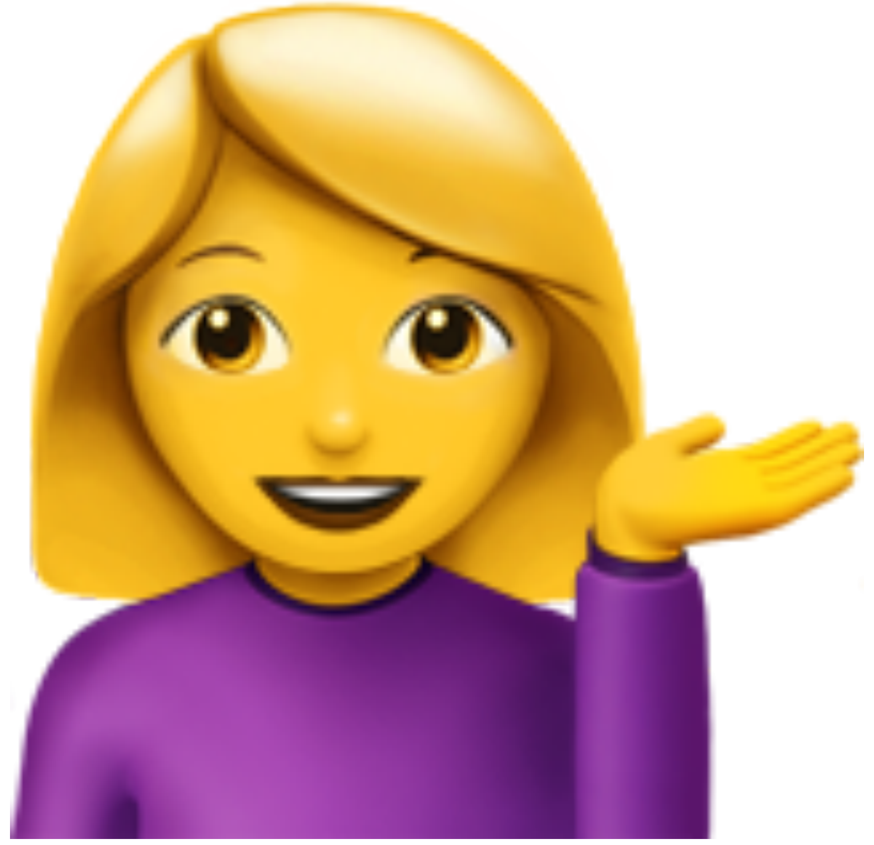
Nadia is smart.

She wants to know if a new feature will be effective.

She talks to you about impact, tracking & KPIs *before* planning the feature.

BE LIKE NADIA

Meet Nadia



Nadia is a product manager.

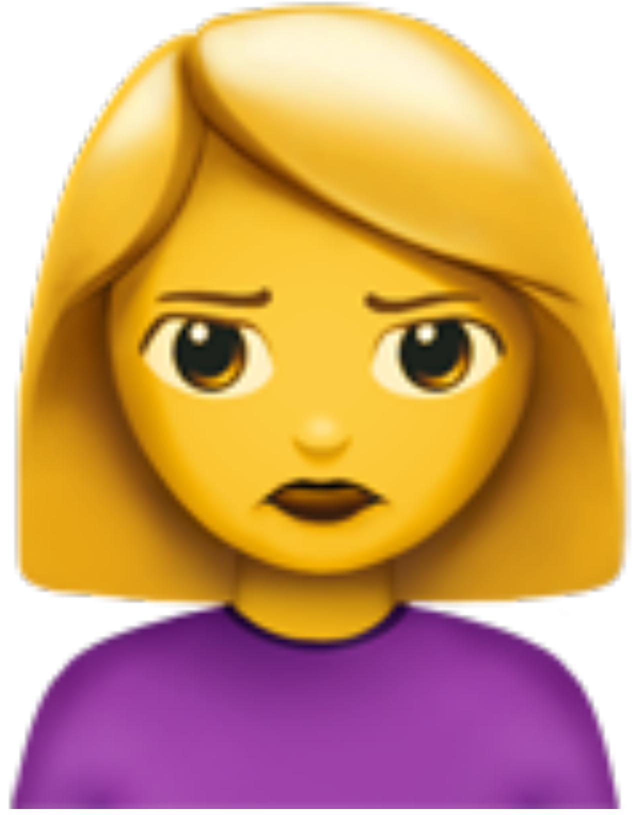
Nadia is ~~smart~~ responsible.

She wants to know if a new feature will be effective.

She talks to you ~~about impact, tracking & KPIs~~ *before* ~~planning~~ releasing the feature.

BE LIKE NADIA, but be better next time

Meet Nadia



Nadia is a product manager.

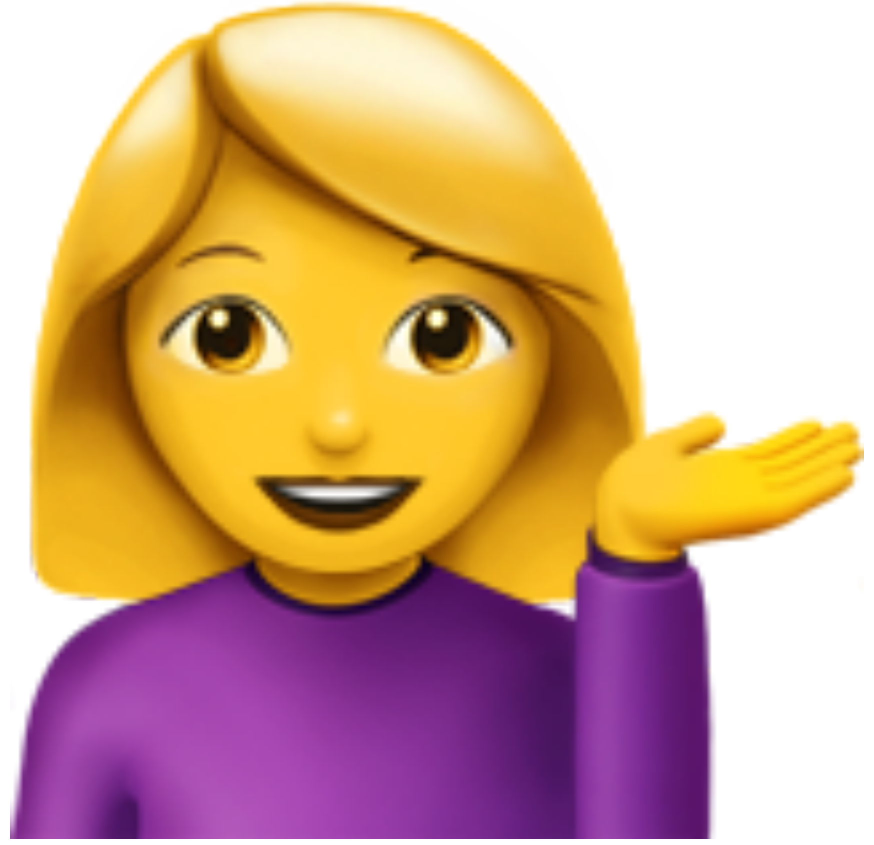
~~Nadia is smart responsible.~~

She wants to know if a new feature will be effective.

She talks to you ~~about impact, tracking & KPIs *before* planning~~ after releasing the feature.

~~BE LIKE NADIA, but be better next time~~

In ~~a perfect~~ the real world

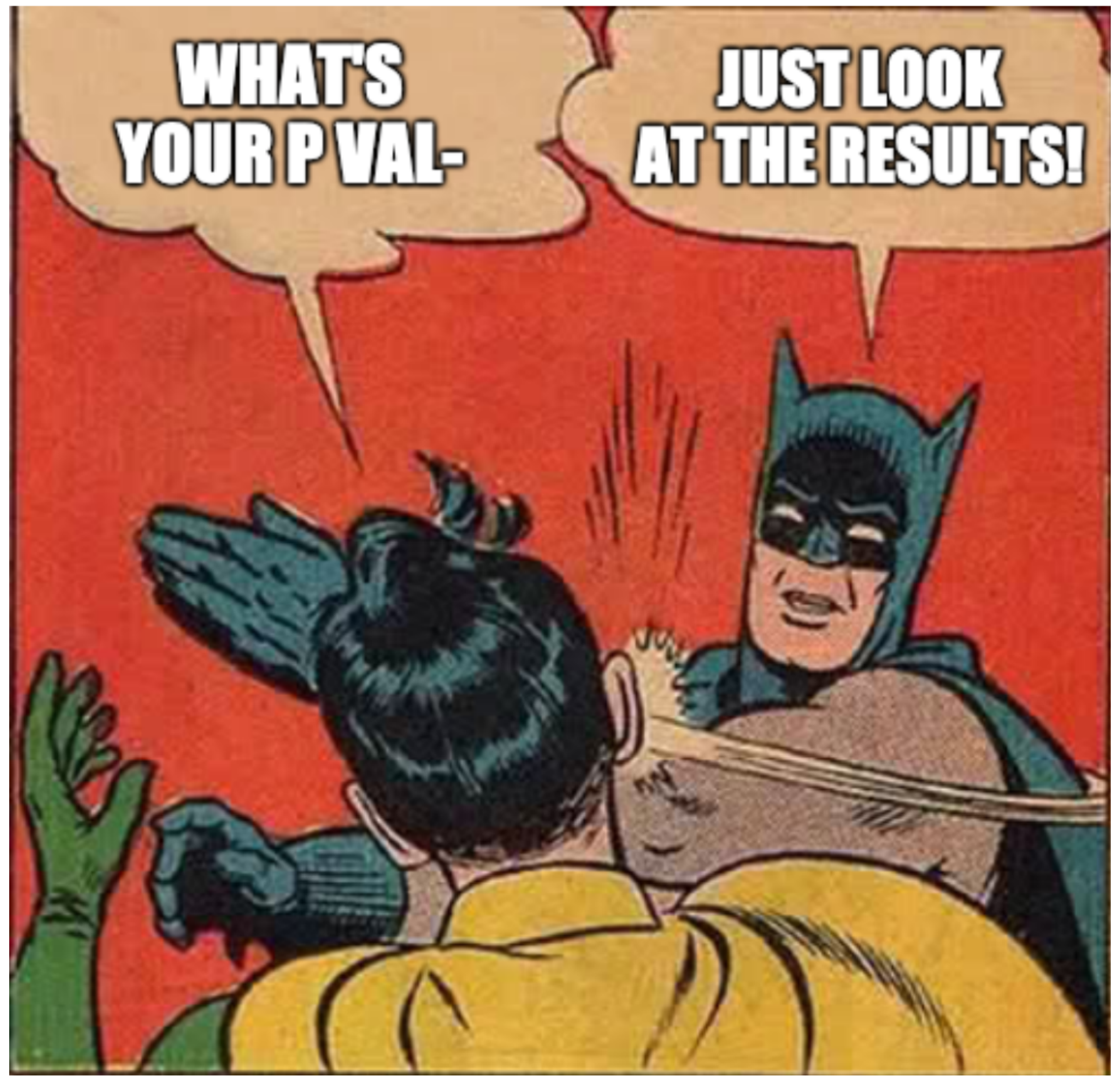


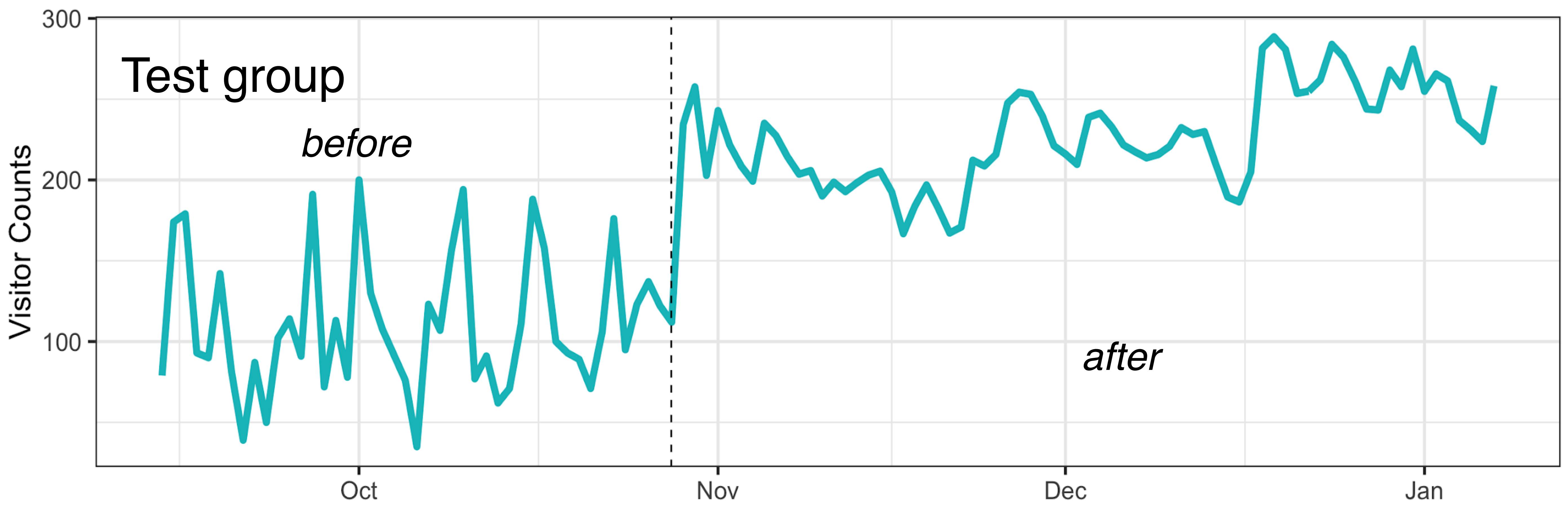
- We have a model of population & causality (e.g. *better feature ➡ more usage)*
- We have well defined KPIs (clicks, sales) and understanding of effect size
- Sufficient volume for significance & power
- Sufficient velocity for timely answer
- Good randomisation & user tracking infra for A/B tests

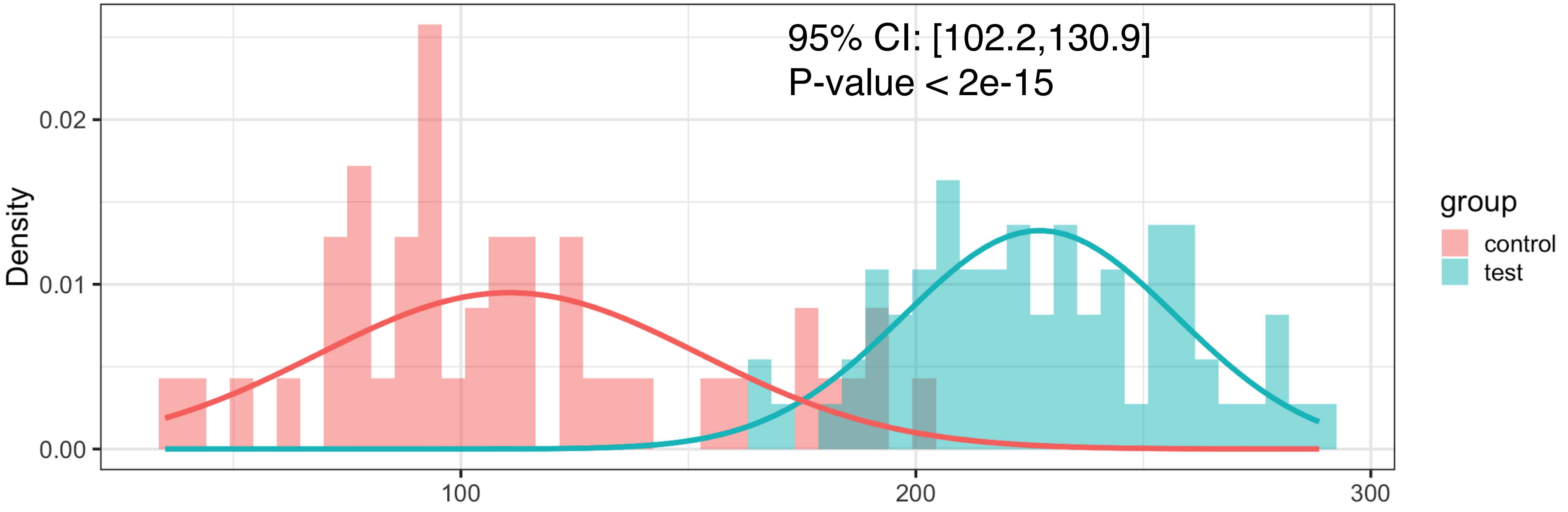
⚠
harder than
you’d think
⚠

Nadia wants to know: Is it working?

Good news! We pass the IOTT (Intra-Ocular Trauma Test)



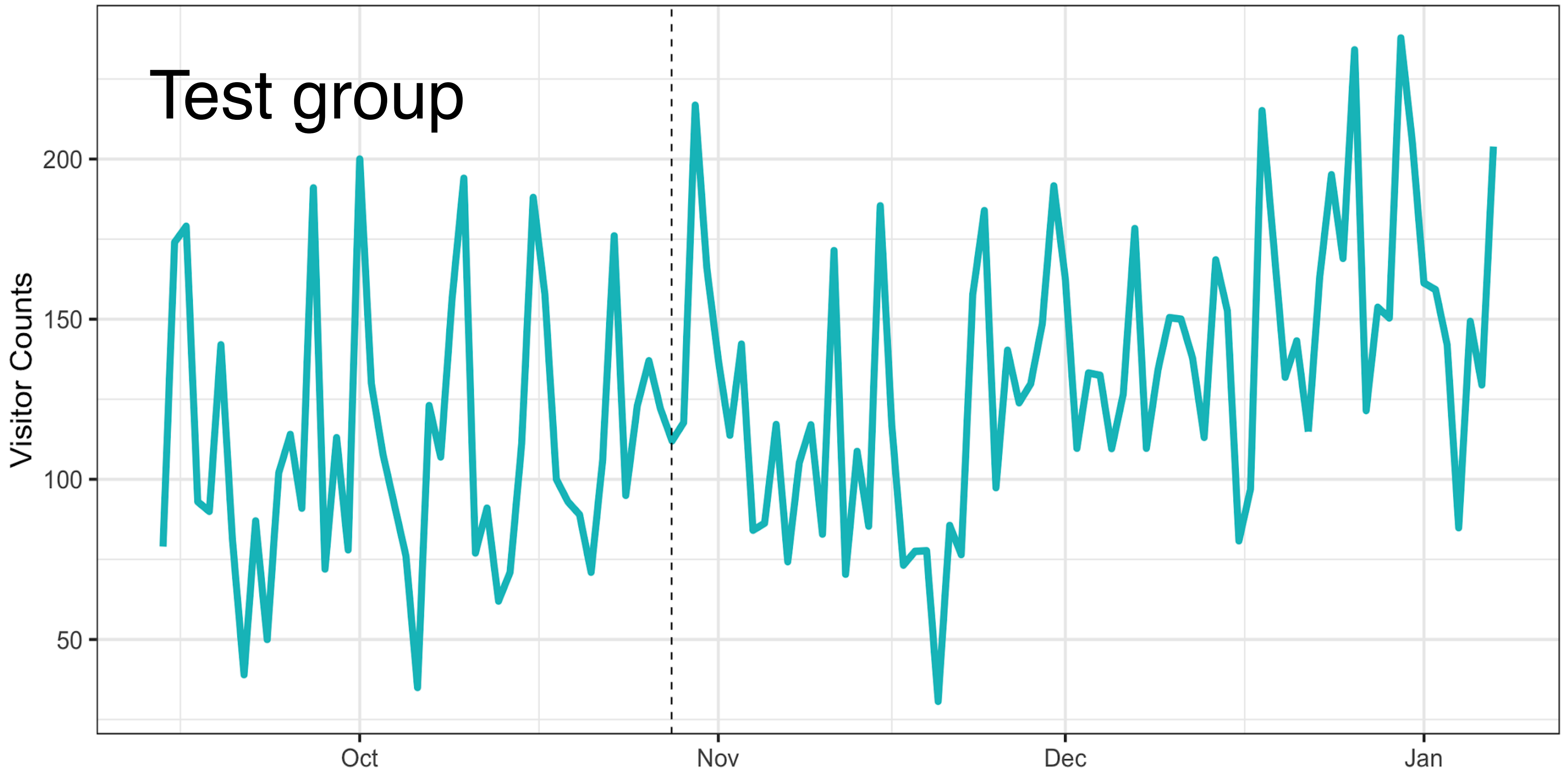




So... Is it working?

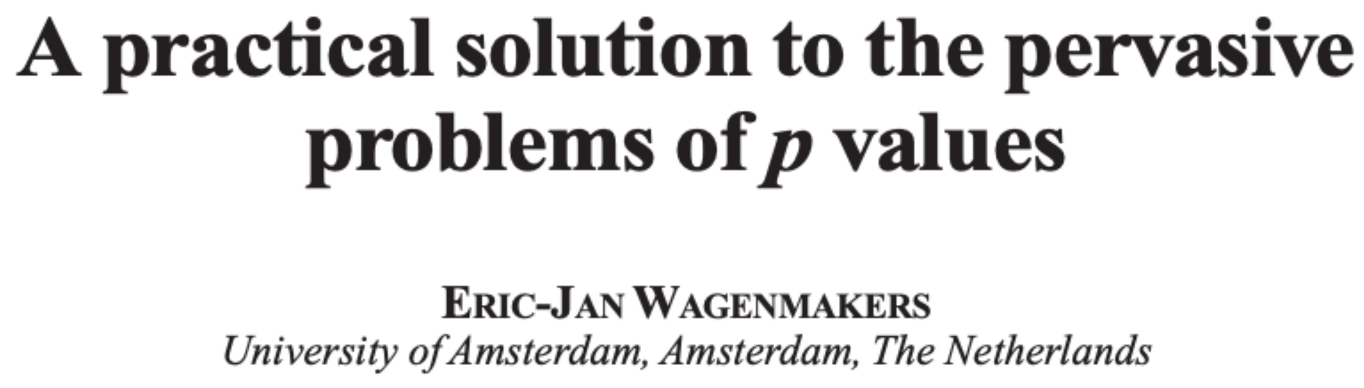
Life is noisy and complicated, so we ran a test:

- Nadia asks: “Can we say the ad campaign worked?”
- You say : “We saw X% increase daily visits, with p < 0.005”
- Nadia hears: “99.5% its working?”



Why Bayes?

- Because you want the right answer: ***Is it working?***
- Because by using p-values you are miss-communicating with your stakeholders (with $p < 0.001$)



A practical solution to the pervasive problems of *p* values

Eric-Jan Wagenmakers
University of Amsterdam, Amsterdam, The Netherlands

- Because it's a good way to think about problems
- Because Bayesian tools support a better processes (and cover more cases)

The answers you want

The answer Nadia wants

Prior

Likelihood (model)

$$P(\text{“it works”}|\text{data}) = \frac{P(\text{“it works”})\ P(\text{data}|\text{“it works”})}{P(\text{data})}$$

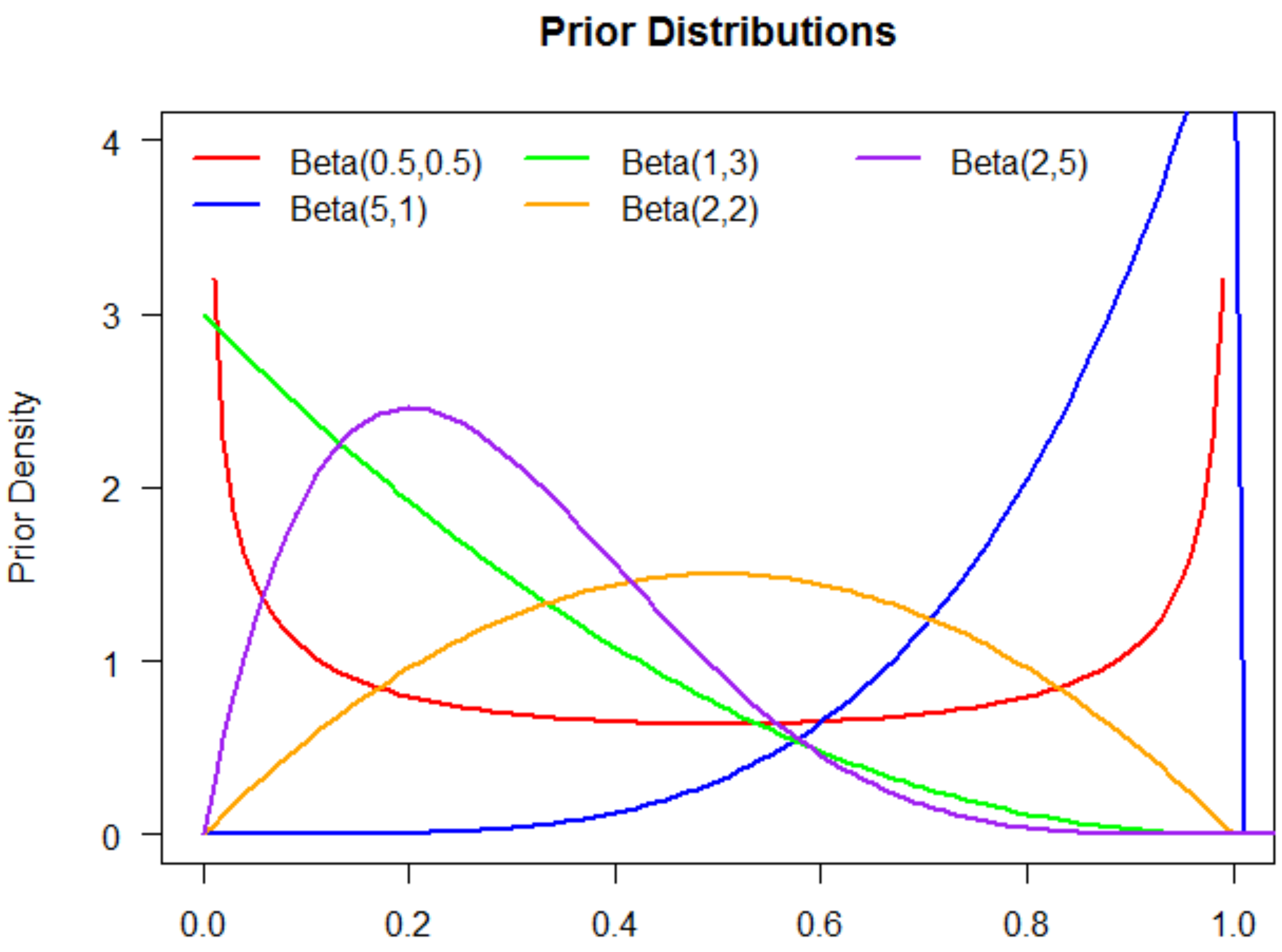
Might be Hard to Compute

$$\text{p-value} = P(\text{data}|\text{”it’s not working”})$$

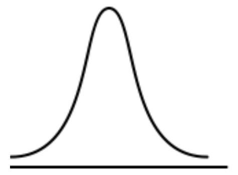
Priors means you have an opinion

"... the probability distribution that would express ***one's beliefs*** (yes, it's subjective 🙀) about this quantity ***before*** some evidence is taken into account."

Adapted from Wikipedia



How do we choose?

- For A/B testing there are some obvious defaults: mean=0, some “natural” limits
- From stakeholders: “if you had to guess”, “from your experience”, surveys, gamification, ...
- If you’re lucky there are industry benchmarks
- Defaults from your tools (when in doubt - )
- Beyond that there are good guidelines

Your new job: Translate business insights into a distribution

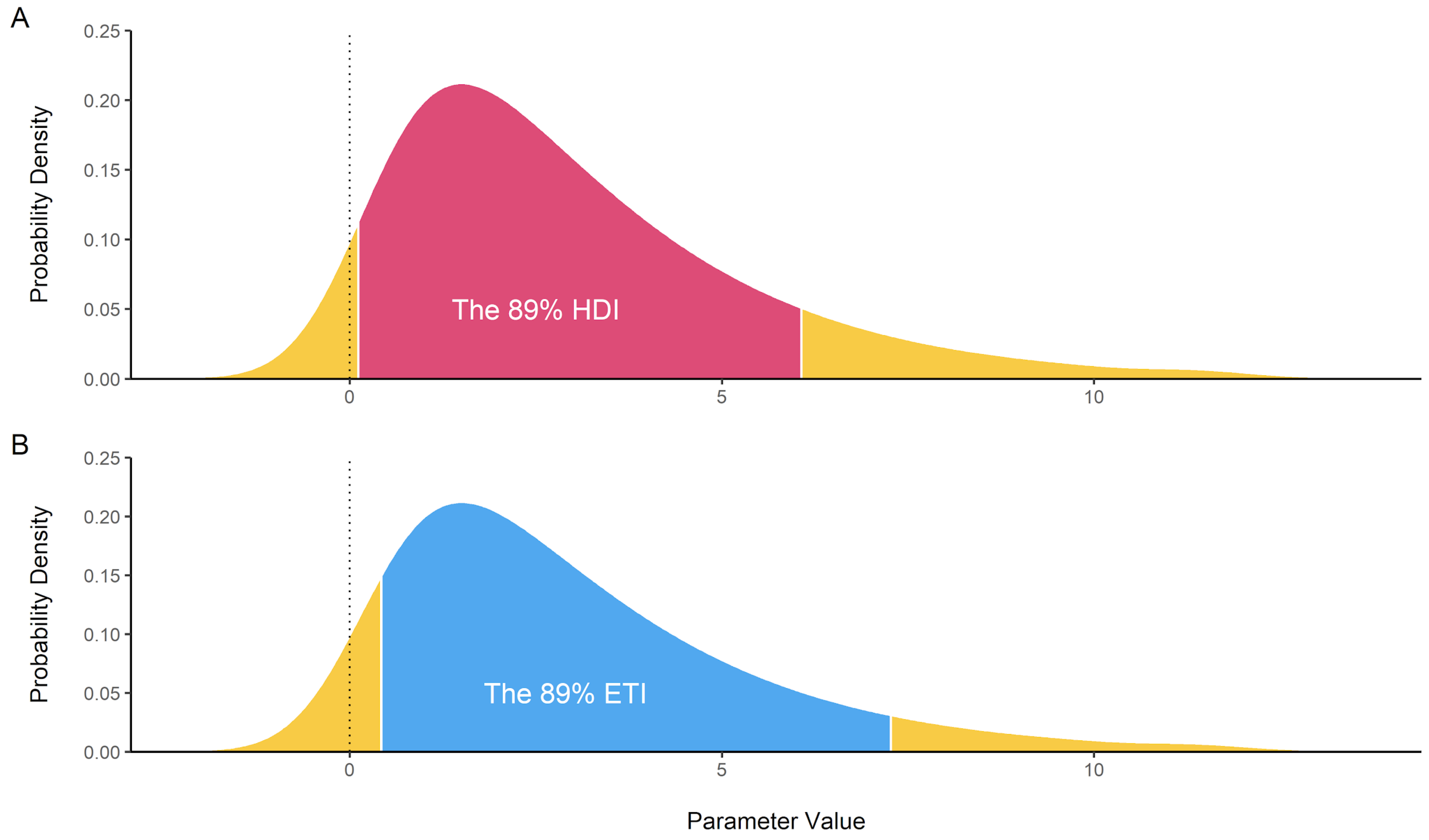
It is working!

Frequentist gives:

Point estimate + CI + p-value (&power) + confusion

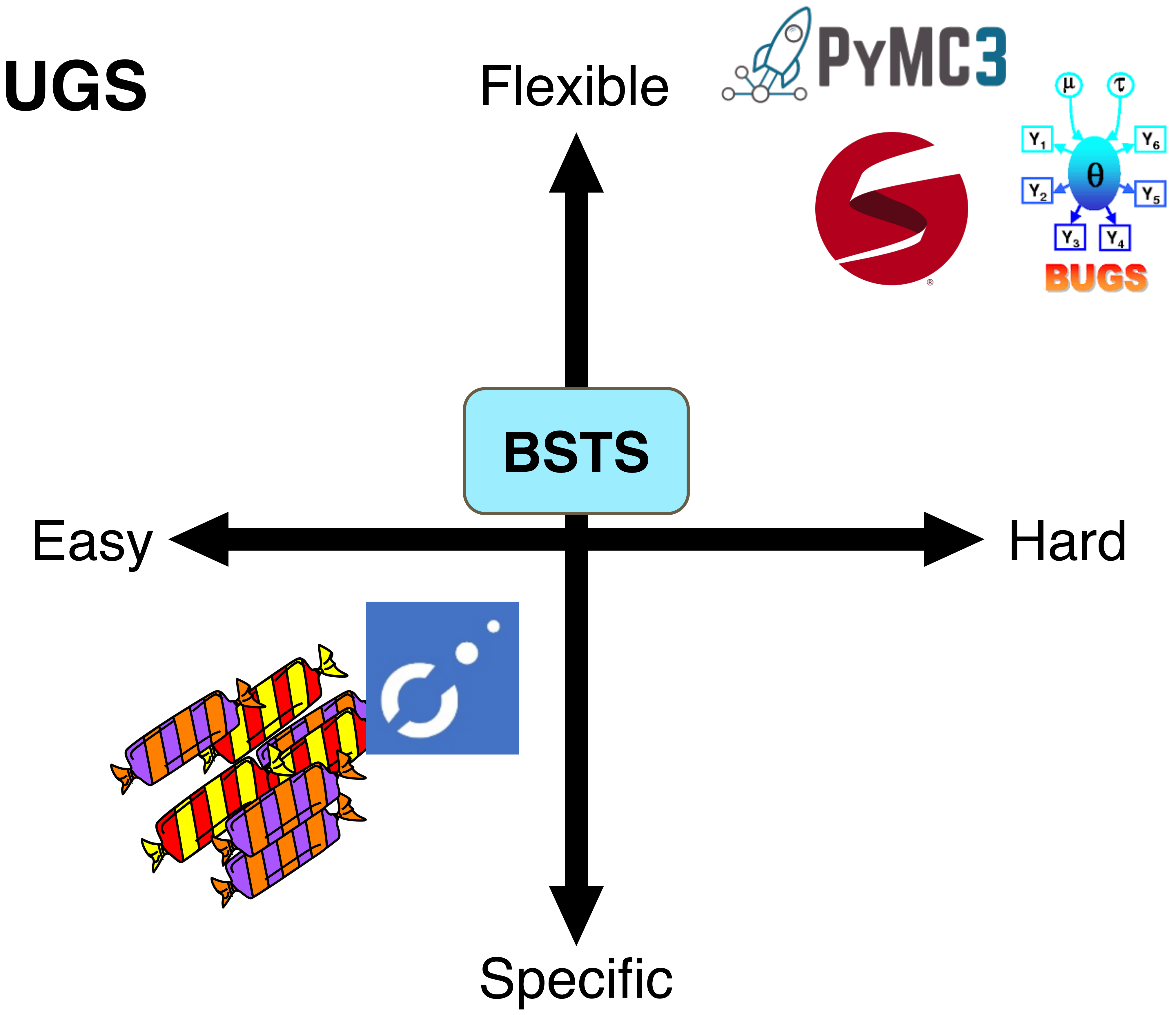
Bayes gives: Posterior distribution, that can answer:

- Where does the difference “live” (HDI/EDI)
- Are doing damage? (Type S)
- Are we off by a magnitude? (Type M)
- Are below an arbitrary minimal threshold?
- How crazy do you have to be to think there was no difference? (Bayes factors)



Some Toolkits

- **Low level frameworks: Stan/pyMC3/BUGS/JUGS**
  - Fully flexible & powerful
  - New syntax
  - Cross platform
- **Mid level frameworks: BSTS**
  - Topical (solve a specific problem)
  - Flexibility ⇔ structure trade-off
- **Wrappers** 🍬🍬🍬
  - Stan/R ecosystem: Prophet, BRMS, stanARM, ...
  - BSTS: CausalImpact
  - R packages: BEST / BayestestR / …



Flexible

BSTS

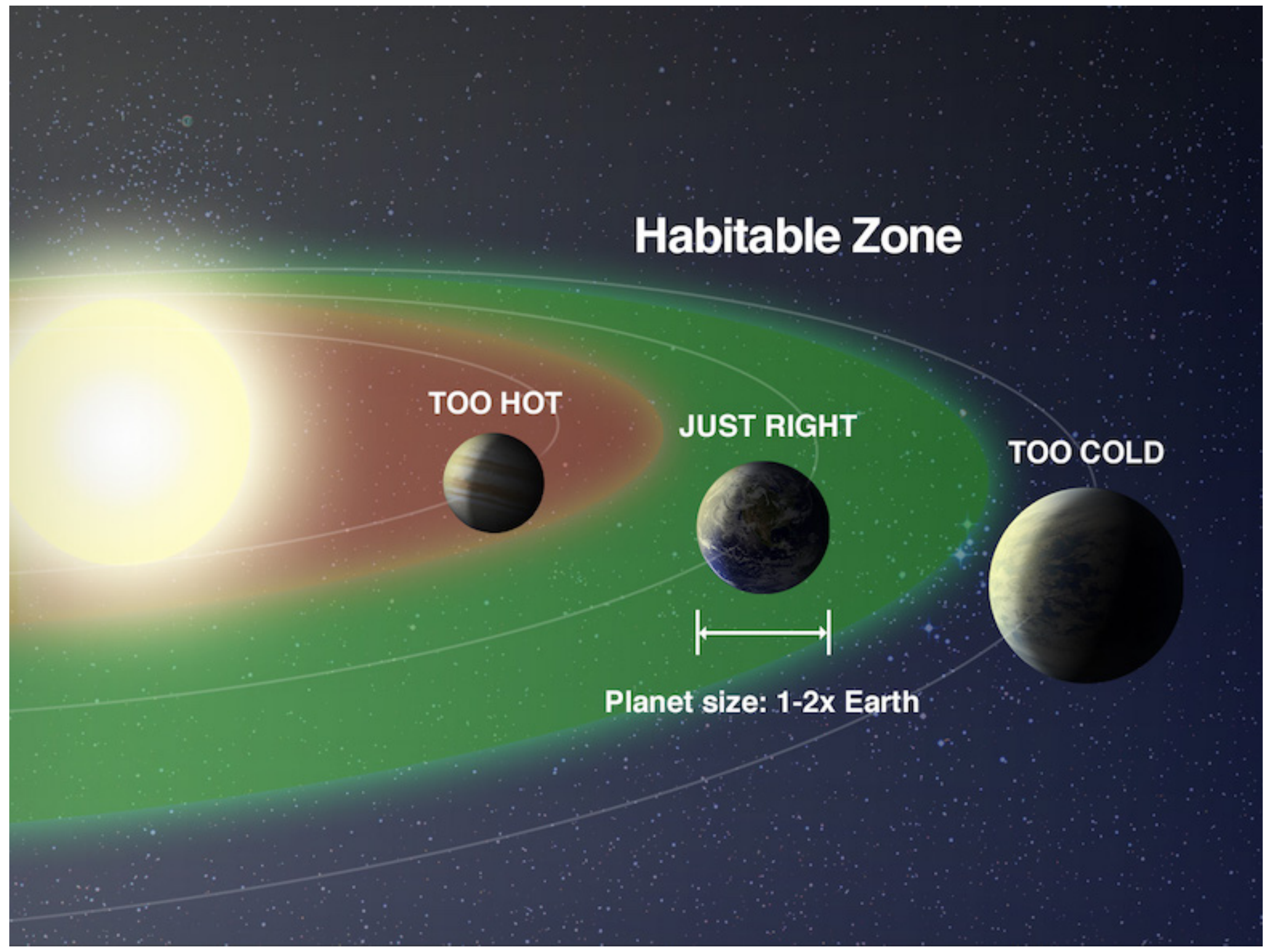
Easy

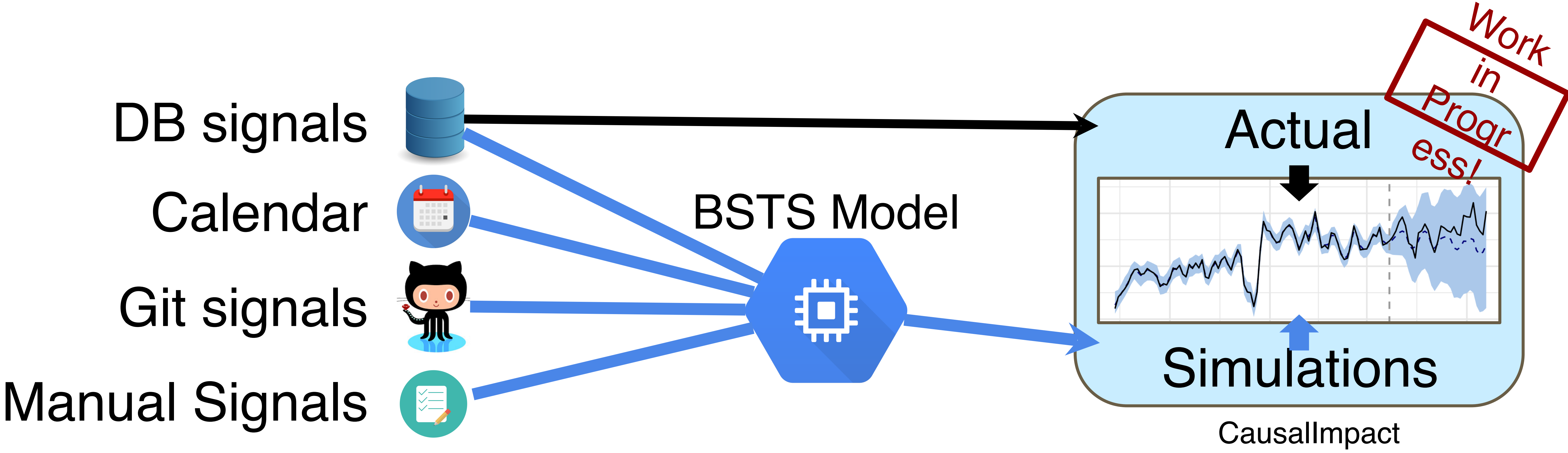
Hard

Specific

A/B testing is the answer to everything, except...

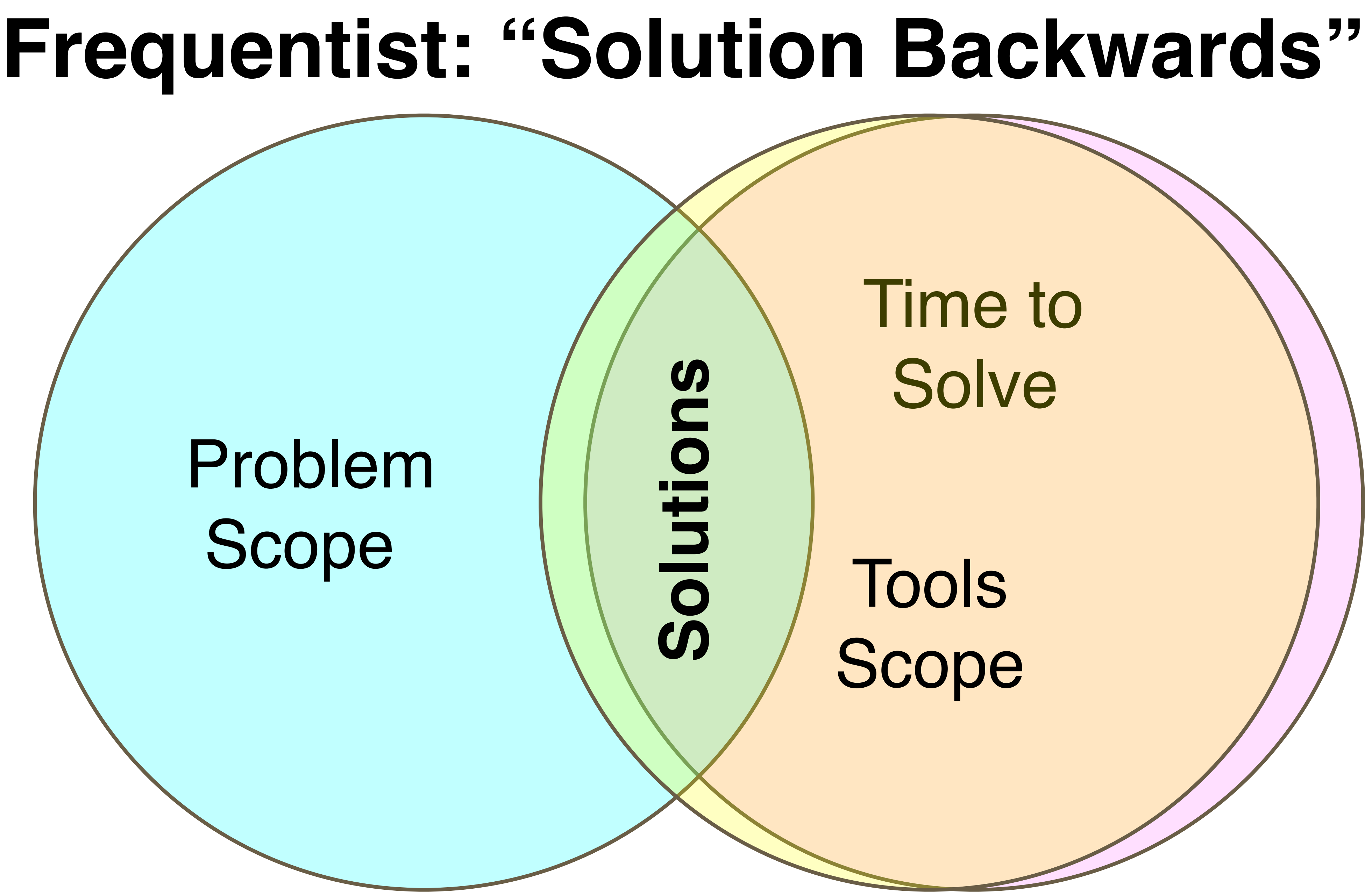
- When you are out of the “Goldilocks Zone”
  - Too fast / slow (time matters)
  - Too broad / specific (pooling)
- When you just can’t test:
  - Public campaigns
  - Tracking gaps
  - Legal issues

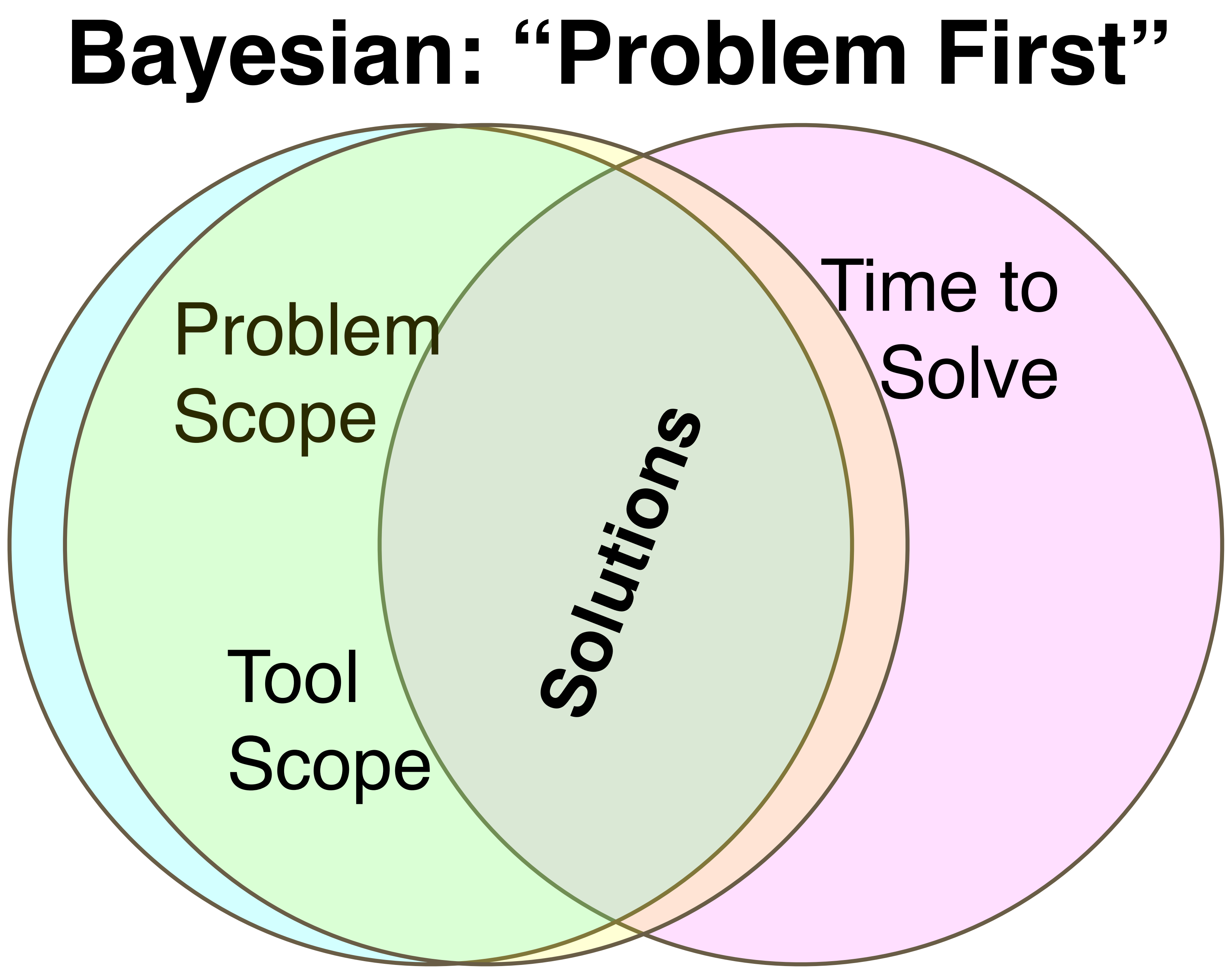




More at: https://github.com/ytoren/presentation-bsts

Thinking & Framing





- Frequentist tools: phrase the problem to fit the tools
- Bayesian tools: find a model that fits the problem (but in a finite time…)

Summary

- P-value is a good answer, just to the wrong question
  (“are we surprised?”)
- Bayesian models can give you *the answers you need*,
  as long as you have an opinion and you are willing to change it
  (both are not so easy)
- Bayesian tools allow you to ask *good questions*
- But - with great power comes great responsibility 🕸
  so use powerful tools with care!

Questions?

Thank you!

We’re Hiring!

Find me on @BigEndianB, Linkedin, github.com/ytoren