Motivation to experiment

David Sweet

***Experimentation for Engineers*, Manning books**

From A/B testing to Bayesian optimization

Experimentation for Engineers

David Sweet

MANNING

Available for preorder on Amazon

Early access at manning.com

Web tool: cogneato.xyz

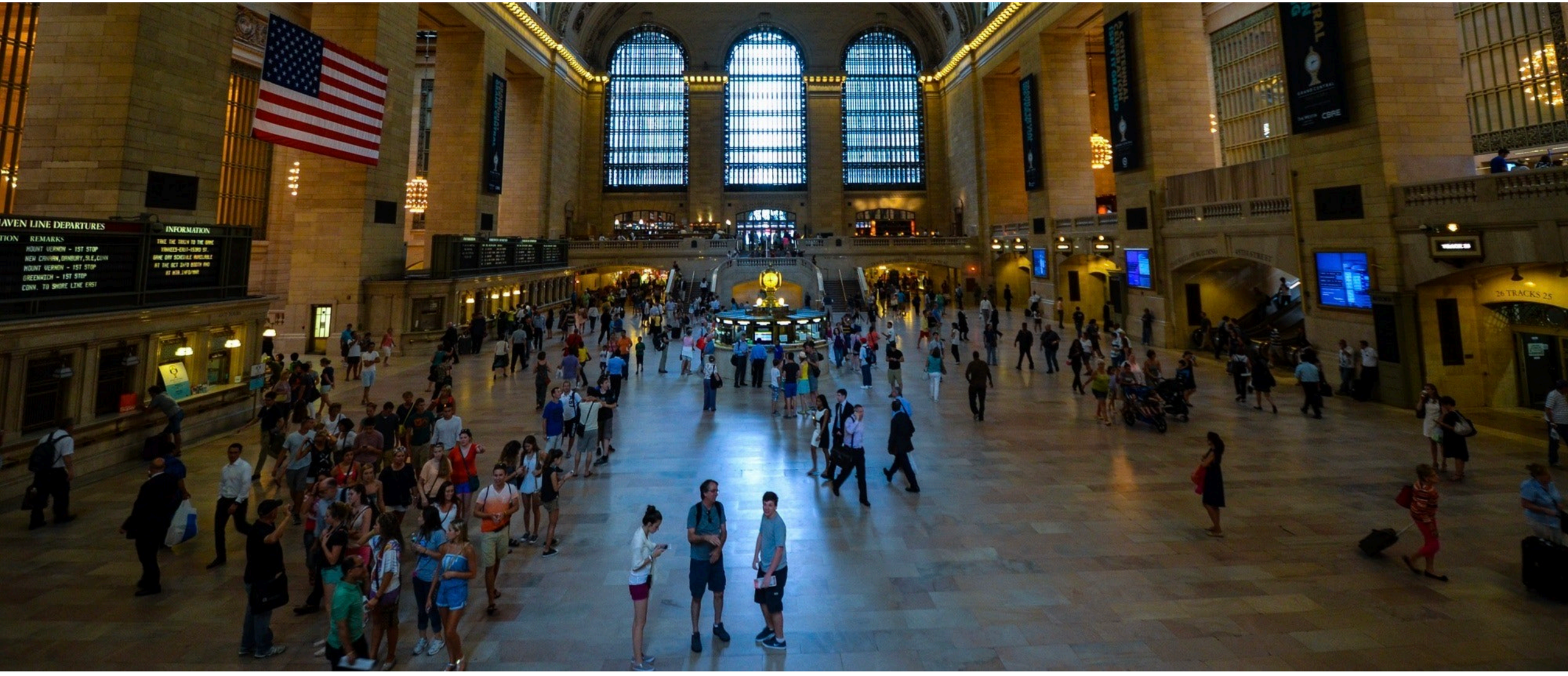
Motivation to experiment

David Sweet

***Experimentation for Engineers*, Manning books**

Observation: Lots

Interaction: Little

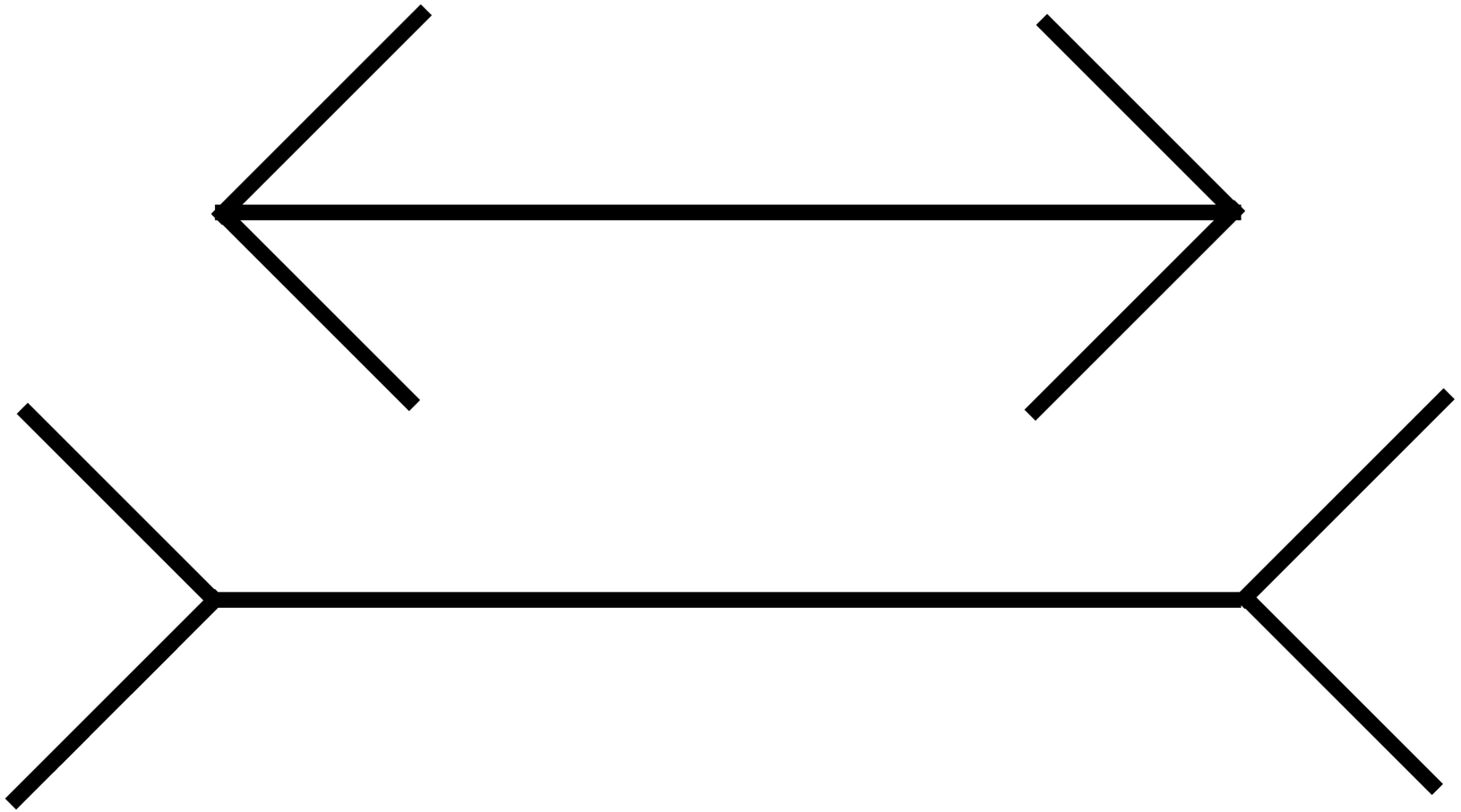


Observation: Lots

Interaction: Little

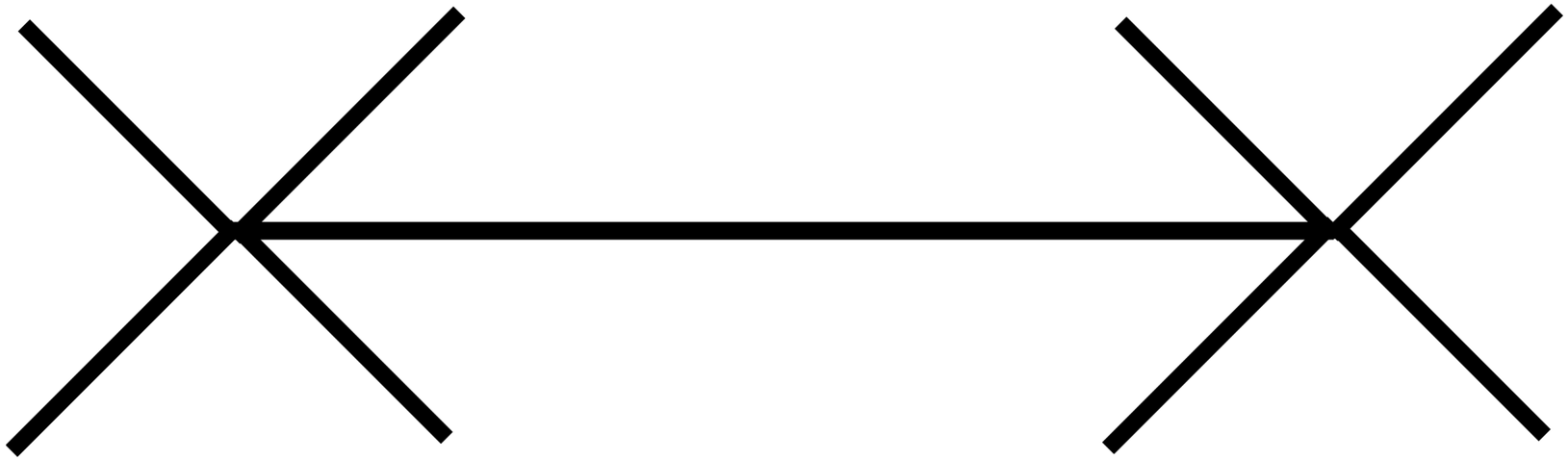
Muller-Lyer illusion: Which is longer?

Observe



Muller-Lyer illusion: Which is longer?

Interact



Observation vs. Experiment

- Observation: passively collect information, form hypotheses
- Experiment: interact with the environment, test hypotheses
- Experiment may contradict observation.
- For example…

Early math ability

- What's Past is Prologue: Relations Between Early Mathematics Knowledge and High School Achievement
  https://pubmed.ncbi.nlm.nih.gov/26806961/

  - **Observation**: "These results demonstrate the importance of pre-kindergarten mathematics knowledge and early math learning for later achievement."

- Risky business: Correlation and causation in longitudinal studies of skill development
  https://pubmed.ncbi.nlm.nih.gov/29345488/

  - **Experiment**: "We first show that experimental manipulation of early math skills generates much smaller effects on later math achievement than the nonexperimental literature has suggested."

Early math ability

- **Plausible claim**: Pre-K math education leads to high school achievement
- **Measured correlation**: It’s right there in the data
- **Morally aligned**: Helping children
- Easy to believe. Easy to sell. Easy to spend scarce time and money on.
- Alas, not real.

Acupuncture

- An Observational Study on Acupuncture for Earthquake-Related Post-Traumatic Stress Disorder
  https://pubmed.ncbi.nlm.nih.gov/31031878/

  - **Observation**: "These results suggest that acupuncture could be a useful tool for reducing pain and psychologic symptoms related to earthquakes, "

- Acupuncture for osteoarthritic pain: an observational study in routine care
  https://academic.oup.com/rheumatology/article/45/2/222/1784739

  - **Observation**: "patients with chronic pain due to osteoarthritis reported clinically relevant improvements after acupuncture treatment.

Acupuncture

- A Randomized Controlled Trial of Acupuncture for Osteoarthritis of the Knee
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3651275/
  - **Experiment**: "[Traditional Chinese Acupuncture] was not superior to sham acupuncture. “
- Randomized, Controlled Trial of Acupuncture for the Treatment of Hot Flashes in Breast Cancer Patients
  https://ascopubs.org/doi/10.1200/JCO.2007.12.0774
  - **Experiment**: ”...compared with sham acupuncture, the reduction by the acupuncture regimen as provided in the current study did not reach statistical significance. “
- Acupuncture for Patients With Migraine: A Randomized Controlled Trial
  https://jamanetwork.com/journals/jama/fullarticle/200822
  - **Experiment**: "Acupuncture was no more effective than sham acupuncture in reducing migraine headaches”

Reiki

- A Large-Scale Effectiveness Trial of Reiki for Physical and Psychological Health
  https://www.liebertpub.com/doi/10.1089/acm.2019.0022

  - **Observation**: “The results from this large-scale multisite effectiveness trial suggest that a single session of Reiki improves multiple variables related to physical and psychological health.”

- Effects of reiki in clinical practice: a systematic review of randomized clinical trials
  https://pubmed.ncbi.nlm.nih.gov/18410352/

  - **Experiment**: "...the evidence is insufficient to suggest that reiki is an effective treatment for any condition.”

Acupuncture / Reiki

- **Plausible claim**: We've been at this for 3000 years.
- **Measured correlation**: You *actually* feel better afterward.
- **Morally aligned**: Helping people who are suffering
- Easy to believe. Easy to sell. Easy to spend scarce time and money on.
- Alas, not real.

Hormone replacement therapy

- Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence (1991; HRT began in 1960's)
  https://pubmed.ncbi.nlm.nih.gov/1826173/

  - **Observation**: "Overall, the bulk of the evidence strongly supports a protective effect of estrogens that is unlikely to be explained by confounding factors. "

- Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial (pub. 2002)
  https://pubmed.ncbi.nlm.nih.gov/12117397/

  - **Experiment**: "The risk-benefit profile found in this trial is not consistent with the requirements for a viable intervention for primary prevention of chronic diseases, and the results indicate that this regimen should not be initiated or continued for primary prevention of CHD." *(actually found increased CHD)*

Alzheimer’s disease

- “Exceptions that prove the rule”–Why have clinical trials failed to show efficacy of risk factor interventions suggested by observational studies of the dementia-Alzheimer's disease syndrome?

  https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1002/alz.12633

  - **Observation**: “Epidemiological studies over three decades have identified associations between the dementia-Alzheimer’s disease syndrome and an array of putative risk factors.”
  - **Experiment**: “Numerous randomized controlled trials (RCTs) have tested the efficacy of interventions suggested by these associations.” … “uniformly disappointing”

Stomach (GI) Ulcers

- GI Society
  https://badgut.org/information-centre/a-z-digestive-topics/nobel-prize-for-h-pylori-discovery/
- **Observation**: Doctors note that more stress is associated with more ulcer pain. Clear-cut. Everybody “knows” this for at least 100 years.
- **Experiment**:
  - 1982 Marshall and Warren propose H. pylori causes ulcers. Nobody listens
  - 1985 Marshall infects himself with H. pylori and gets an ulcer.
  - 2005 Marshall and Warren win Nobel prize for medicine

Stomach (GI) Ulcers

- Effect of triple therapy (antibiotics plus bismuth) on duodenal ulcer healing. A randomized controlled trial
  https://pubmed.ncbi.nlm.nih.gov/1854110/

  - **Experiment**: "Combined therapy with anti-H. pylori agents and ranitidine was superior to ranitidine alone for duodenal ulcer healing. Our results indicate that H. pylori plays a role in duodenal ulcer disease.”

- Cure of duodenal ulcer associated with eradication of Helicobacter pylori
  https://pubmed.ncbi.nlm.nih.gov/1971318/

  - **Experiment**: "17 of the 45 patients who completed the treatment, Helicobacter pylori was eradicated, and there was no ulcer relapse during the first 12 months of follow-up.”

Are we *always* wrong?

Engineering/technology

- Amazon reports < 50% of their A/B tests improve metrics
- Microsoft reports only 1/3
- Netflix reports only 10%
- Failure is probable

A/B test == RCT where no one can die

Your great idea probably won’t work.

Counterfactuals

- Observational data is missing **counterfactuals**
  - “What *would have happened* had we done things differently?”
- Example: barometer
  - Observation: The barometer goes down just before it rains
  - Hypothesis: A decrease on the barometer causes the rain
  - Counterfactual: Barometer reading low, but no rain

Counterfactuals

- Example: Ulcers
  - Patient gets stressed, stomach hurts. Doctor sees an ulcer.
  - Hypothesis: Stress causes ulcers.
- Counterfactuals:
  - Can you be stressed without an ulcer? Yes, but you don’t visit doctor.
  - Can you create an ulcer without stress? <== Marshall, 1985

Causation

- Experiments collect counterfactuals
  - Control: Do it the usual way
  - Treatment: Try it a different way
- Experiments compare control to treatment
  - Control: Give some patients traditional ulcer treatment
  - Treatment: Give other patients an antibiotic, too.
- Experiments establish **causation** <== kill bacteria, no more ulcer

	Low Barometer	High Barometer
No Rain	Experiment: Can you make this happen?	Observed
Yes Rain	Observed	Experiment: Can you make this happen?

Experiment challenges: Statistical

- Variation / uncertainty / noise
  - Metric varies (randomly) from measurement to measurement
  - Solution, **replication**: Take many measurements and average
- Bias
  - Metric different (consistently) for different subsamples (ex., different age groups, different geographies, etc.)
  - Solution, **randomization**: Randomly assign subjects to control or treatment

Experiment challenges: Ethical

- Experimentation used in medicine, social media, food manufacturing, materials science, finance, social science, manufacturing, consumer product design
- Is it too risky to keeping try new things?
- How does this affect the people involved?

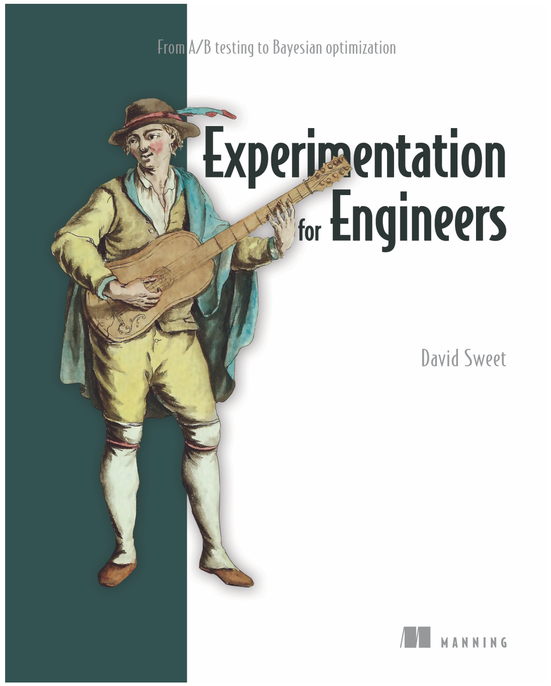
Experiment challenges: Ethical

- Controversy: 2014, Facebook ran “emotion contagion” study on users
  - manipulated the emotional content of users’ feeds: If a user sees more sad posts, does the user create more sad posts? [Yes, BTW.]
  - Experimented on ~600,000 users
  - Could users have been harmed?
  - Would users approve of having their posts used to make friends and family sadder? That’s generally not considered the intent of posting on Facebook.
  - https://www.pnas.org/content/111/24/8788

Experiment challenges: Ethical

- LinkedIn w/Harvard, Stanford, & MIT ran a study (2017-2022) on 20MM users to test whether weak ties provided better job leads than strong ties [Yes, BTW]
- Could some users have missed out on job opportunities because of this?
- Question was considered
  - Not actually experiments, but advanced observational analysis techniques
  - Ok'd by MIT's **Institutional Review Board** beforehand
- https://arstechnica.com/tech-policy/2022/09/experts-debate-the-ethics-of-linkedins-algorithm-experiments-on-20m-users/

Experimental methods



- Ongoing research into experimental methods
- Sequential methods
  - Stop an experiment when you've "seen enough"
- Bayesian optimization
  - Find best design/settings, testing as few versions as possible
- Causal-observational methods
  - Seek evidence for causation in *special* observational data when experimentation is impossible or unethical

Statistical & ethical challenges
Expectation of failure

Why bother?

Experimentation calculus

- You pay an experimentation cost — risk, time, dollars — once
- You reap the benefits many times
- Example:
  - People who volunteer for clinical trials put themselves at risk, but
  - New treatments help save lives for decades (at least)
- Ex, measles: “In the decade before 1963, when a vaccine became available, nearly all children got measles by the time they were 15 years of age.”
  https://www.cdc.gov/measles/about/history.html

Experimentation calculus

- Smallpox
- Diphtheria
- Tetanus
- Pertussis
- Polio
- Measles
- Mumps
- Rubella
- Meningococcal disease
- Chicken pox
- Hepatitis A
- Hepatitis B
- Pneumococcal
- Influenza
- Rotavirus
- COVID-19

Summary: Why do we experiment?

- Experiment to **understand**
  - Observations missing counterfactuals
  - Observations don’t establish causation
- Experiment to **improve**
  - Food, medicine, technology, …
  - Pay experimentation cost once, benefit many times over