Studying Online Behavior at Scale

COMM 4940

Kennedy Hall 213

Notes: bit.ly/36RTkdJ

J. Nathan Matias

@natematias

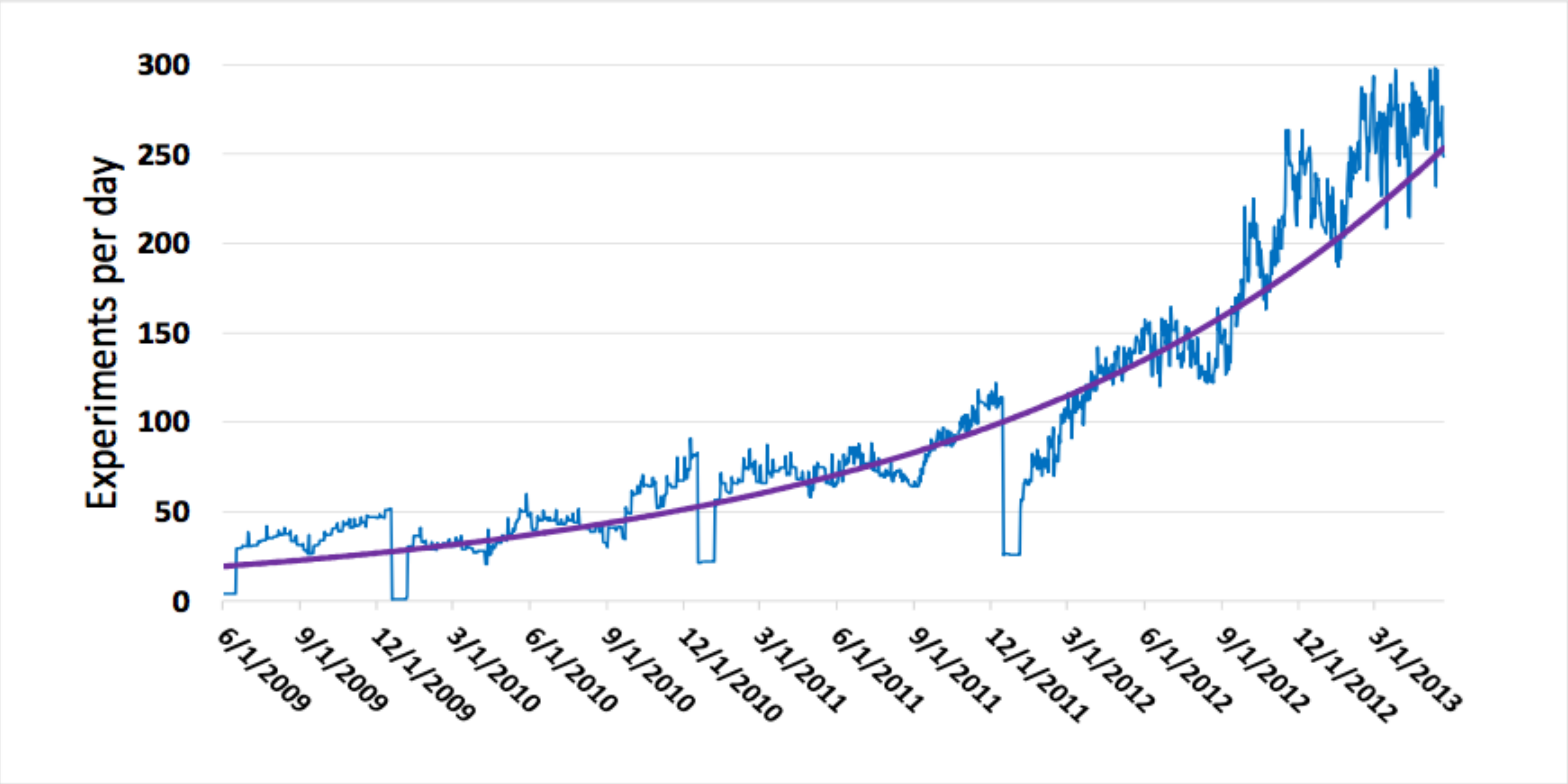
citizensandtech.org

natematias.com

CORNELL COMM #ONECOMMunity

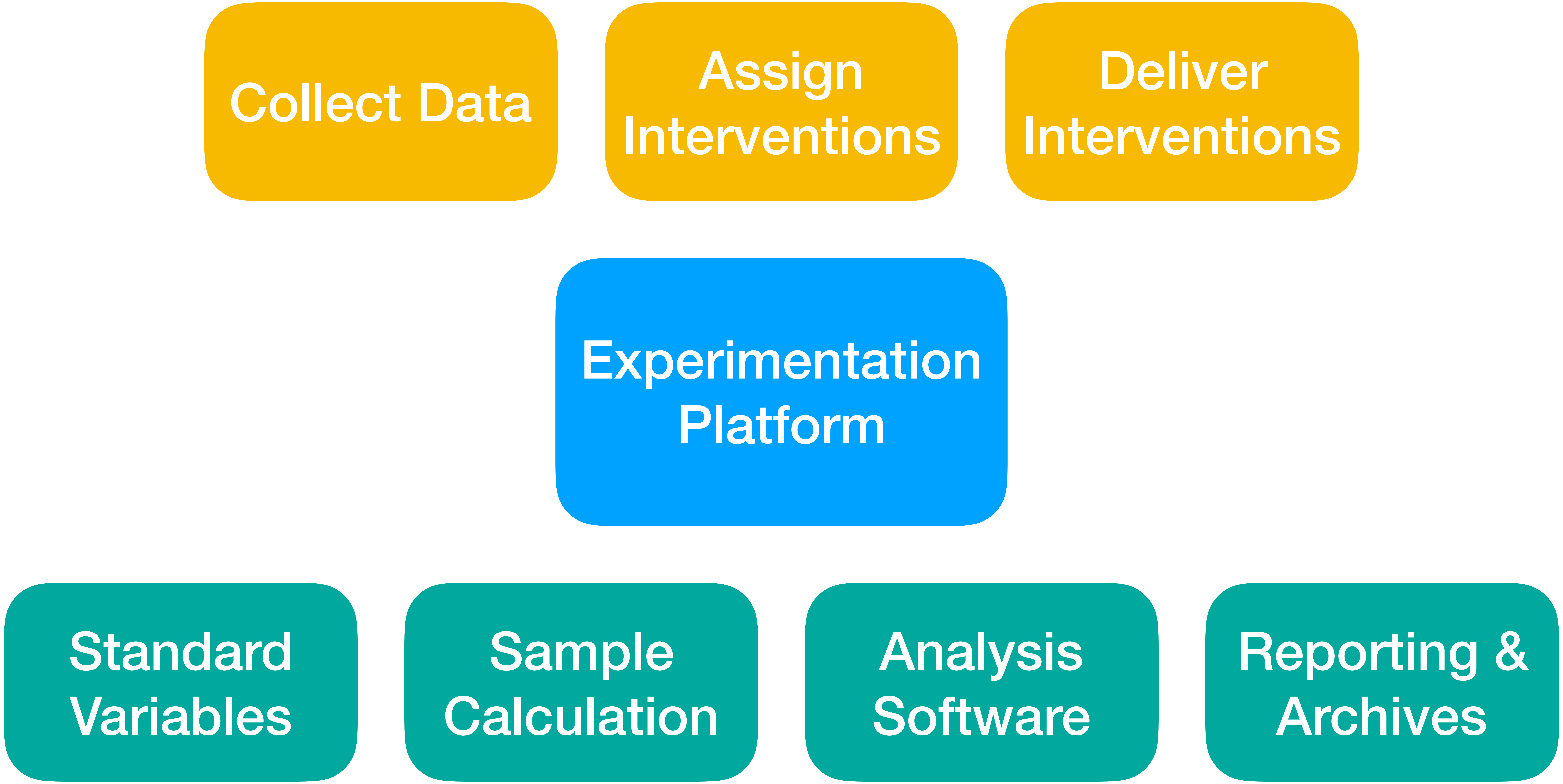
Cornell CALS

- Causal Knowledge
- **Studying Online Behavior at Scale**
- Audits & Account-ability
- Listening Manipulation Or Both?
- Will it Work More Than Once?
- Clickbait & Viral Content
- Why Scientists do Field Research
- Do Firms Learn from A/B Tests?
- Experiments in Democracy
- Interpreting, Using, Misusing Results
- Debriefing, Harm & Consent
- Answering Scientific Questions



Experiments Per Day on bing.com

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013, August). **Online controlled experiments at large scale**. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1168-1176). ACM.



Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013, August). **Online controlled experiments at large scale**. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1168-1176). ACM.

	Category/ Phase	Crawl	Walk	Run	Fly
Technical					
Technical Evolution	Technical focus of product dev. Activities	(1) Logging of signals (2) Work on data quality issues (3) Manual analysis of experiments Transitioning from the debugging logs to a format that can be used for data-driven development.	(1) Setting-up a reliable pipeline (2) Creation of simple metrics Combining signals with analysis units. Four types of metrics are created: debug metrics (largest group), success metrics, guardrail metrics and data quality metrics.	(1) Learning experiments (2) Comprehensive metrics Creation of comprehensive set of metrics using the knowledge from the learning experiments.	(1) Standardized process for metric design and evaluation, and OEC improvement
	Experimentation platform complexity	**No experimentation platform** An initial experiment can be coded manually (ad-hoc).	**Platform is required** 3rd party platform can be used or internally developed. The following two features are required: • Power Analysis • Pre-Experiment A/A testing	**New platform features** The experimentation platform should be extended with the following features: • Alerting • Control of carry-over effect • Experiment iteration support	**Advanced platform features** The following features are needed: • Interaction control and detection • Near real-time detection and automatic shutdown of harmful experiments • Institutional memory
	Experimentation pervasiveness	**Generating management support** Experimenting with e.g. design options for which it's not a priori clear which one is better. To generate management support to move to the next stage.	**Experiment on individual feature level** Broadening the types of experiments run on a limited set of features (design to performance, from performance to infrastructure experiments)	**Expanding to (1) more features and (2) other products** Experiment on most new features and most products.	**Experiment with every minor change to portfolio** Experiment with any change on all products in the portfolio. Even to e.g. small bug fixes on feature level.
Operational					
Organizational Evolution	Engineering team self-sufficiency	**Limited understanding** External Data Scientist knowledge is needed in order to set-up, execute and analyse a controlled experiment.	**Creation and set-up of experiments** Creating the experiment (instrumentation, A/A testing, assigning traffic) is managed by the local Experiment Owners. Data scientists responsible for the platform supervise Experiment Owners and correct errors.	**Creation and execution of experiments** Includes monitoring for bad experiments, making ramp-up and shut-down decisions, designing and deploying experiment-specific metrics.	**Creation, execution and analyses of experiments** Scorecards showing the experiment results are intuitive for interpretation and conclusion making.
	Experimentation team organization	**Standalone** Fully centralized data science team. In product teams, however, no or very little data science skills. The standalone team needs to train the local product teams on experimentation. We introduce the role of Experiment Owner (EO).	**Embedded** Data science team that implemented the platform supports different product teams and their Experiment Owners. Product teams do not have their own data scientists that would analyse experiments independently.	**Partnership** Product teams hire their own data scientists that create a strong unity with business. Learning between the teams is limited to their communication.	**Partnership** Small data science teams in each of the product teams. Learnings from experiments are shared automatically across organization via the institutional memory features.
Business					
Business Evolution	Overall Evaluation Criteria (OEC)	OEC is **defined** for the first set of experiments with a few key signals that will help ground expectations and evaluation of the experiment results.	OEC **evolves** from a few key signals to a structured set of metrics consisting of Success, Guardrail and Data Quality metrics. Debug metrics are not a part of OEC.	OEC is **tailored** with the findings from the learning experiments. Single metric as a weighted combination of others is desired.	OEC is **stable**, only periodic changes allowed (e.g. 1 per year). It is also used for setting the performance goals for teams within the organization.

Fabijan, A., Dmitriev, P., Olsson, H. H., & Bosch, J. (2017, May). **The evolution of continuous experimentation in software product development: from data to a data-driven organization at scale.** In 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE). IEEE.

“ **Trapped administrators** have so committed themselves in advance to the efficacy of reform that they cannot afford honest evaluation.

Experimental administrators have justified the reform on the basis of the importance of the problem, not the certainty of their answer.

Campbell, D. T. (1969). **Reforms as experiments.** American psychologist, 24(4), 409.

Goals for Today

- Reminders
- Upworthy / Columbia Journalism Review
- Kohavi / Online controlled experiments at large scale
- Answer questions about the assignment
- If you joined in the last week, I’l have a Q&A afterward

Upworthy Reading

Online Controlled Experiments at Large Scale

Tuesday's Assignment

For this assignment, create a report for Upworthy that describes what you learned and proposes which headline to use. You should also explain the benefits of causal inference, and argue why field experiments could help the foundation test its headlines and beyond.

Your essay should include:

- a paragraph **describing the experiment design**, including the intervention being tested, the outcome measures being used, and how many participants were included.
- a paragraph **summarizing the findings**. It should summarize the outcome variable, the means for each condition, and include a statement of the effect size.
- a paragraph that **suggests a course of action**, contextualizing the findings in a way that the organization would normally think about, such as the payoff per thousand people who see the headline. Think about whether the result could inform future headline writing. Make sure to reflect on the limitations of the sample, which is drawn from the Upworthy's homepage.
- include a **table of results** and an **illustration of the average treatment effect**. You could (a) show the effect with error bars or (b) show fitted(predicted) values for each condition, with error bars for the treatment (color). If you show fitted values, document details of any covariates(predictors) used to generate the fitted values (such as weekend).
- a paragraph that builds on this finding in the attempt to **convince Upworthy to do more testing** with headlines and in the organization.