Recommendation Systems

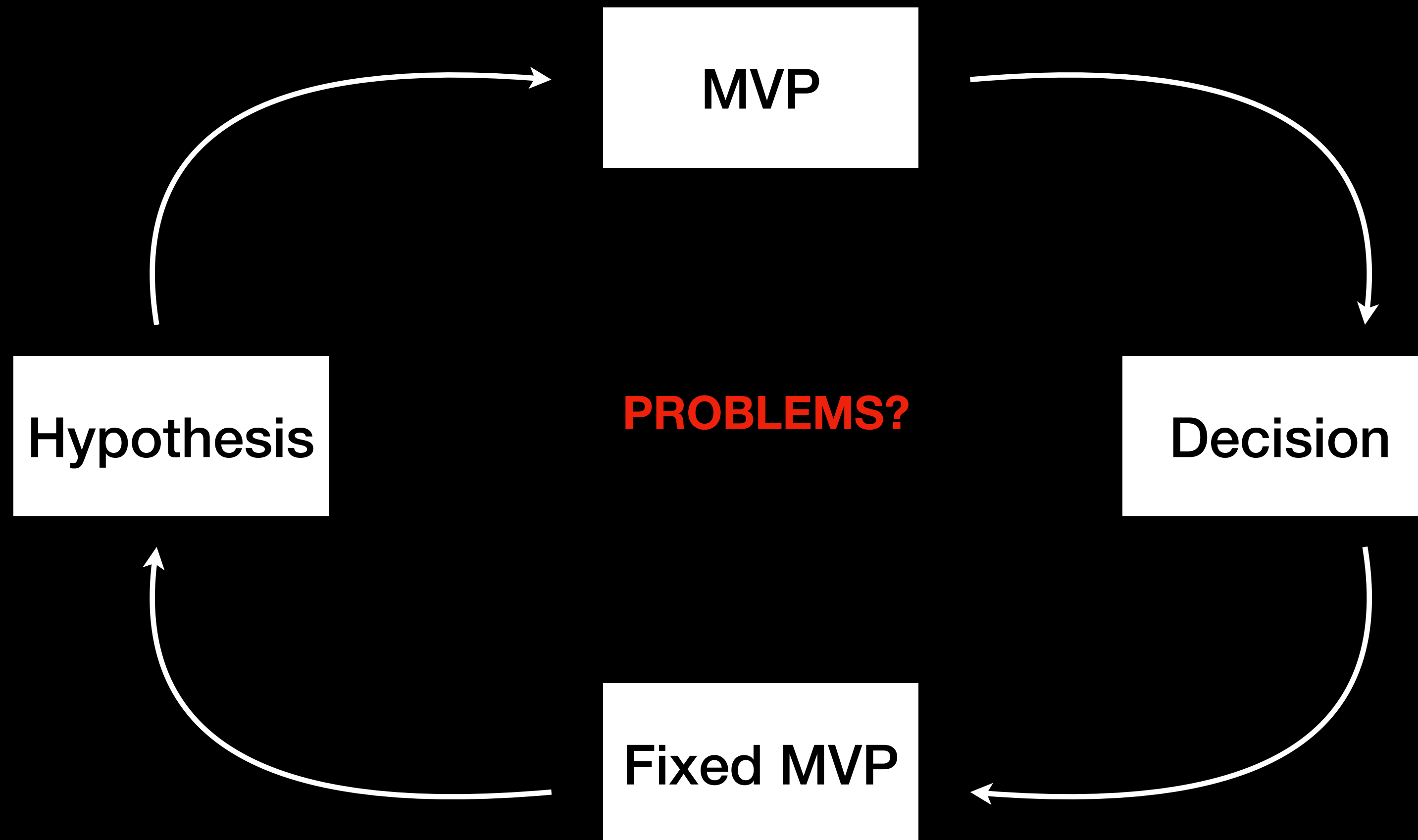## BI, metrics, AB-tests, bandits

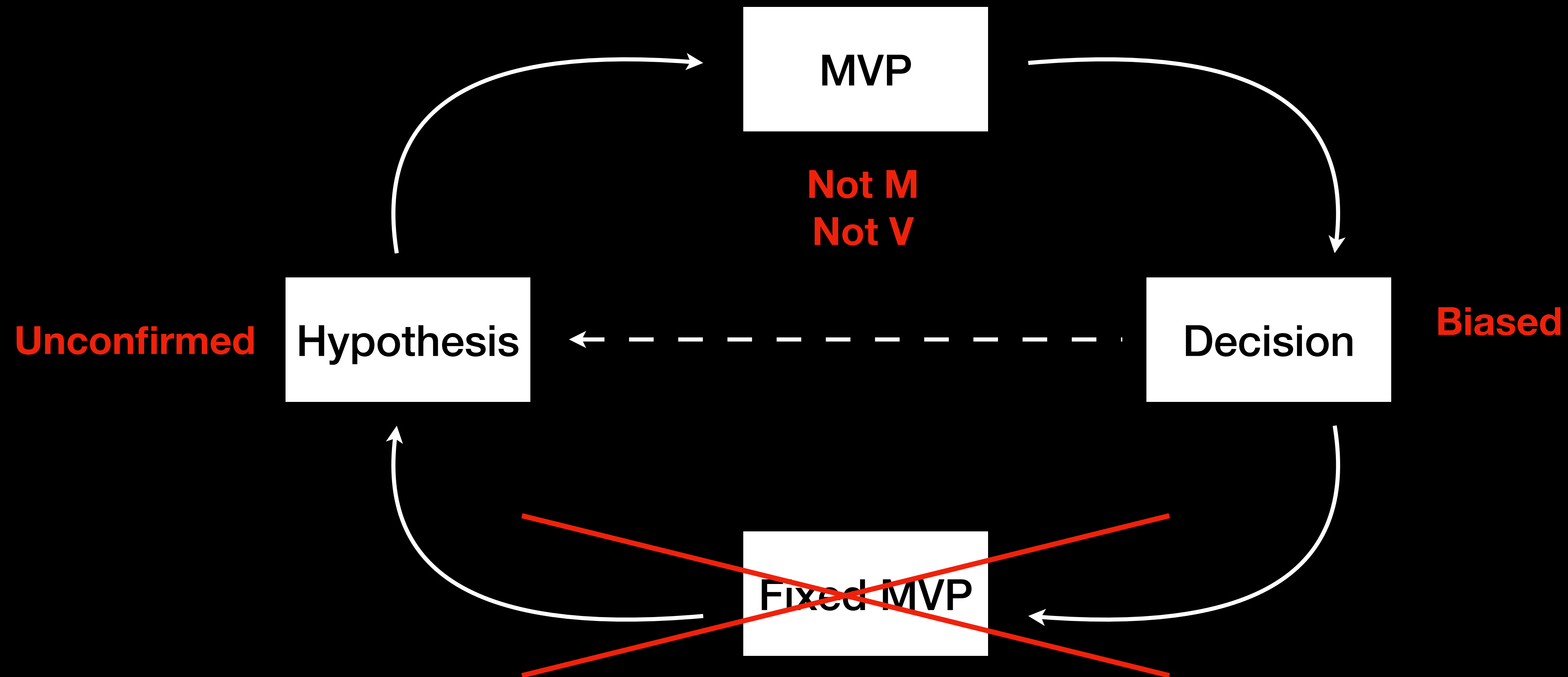Eugeny Malyutin / Sergey Dudorov

To discuss:

- A good startup is a **data-driven** «startup»

- Data-driven «startup» - is not **isn't always good** startup

- Data **fatigue**

- **No** direction

«Startup» progress:

MVP → Decision → Fixed MVP → Hypothesis → MVP

PROBLEMS?

«Startup» progress:

MVP

Not M
Not V

Unconfirmed Hypothesis ← — — — — — Decision **Biased**

Fixed MVP

Metrics canvas discussion:

- Why should you care about it?

    - Most time you are not just an «ML-engineer», but self-sufficient product maker

    - Your product = your bonuses

- Canvas = schema = table

- Answers question:

    - What's going on? Why is the audience not growing?

    - Where to go next?

    - We got AB result sheet with 100+ metrics, what's next?

MATH & STATISTICS
- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE
- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS
- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau
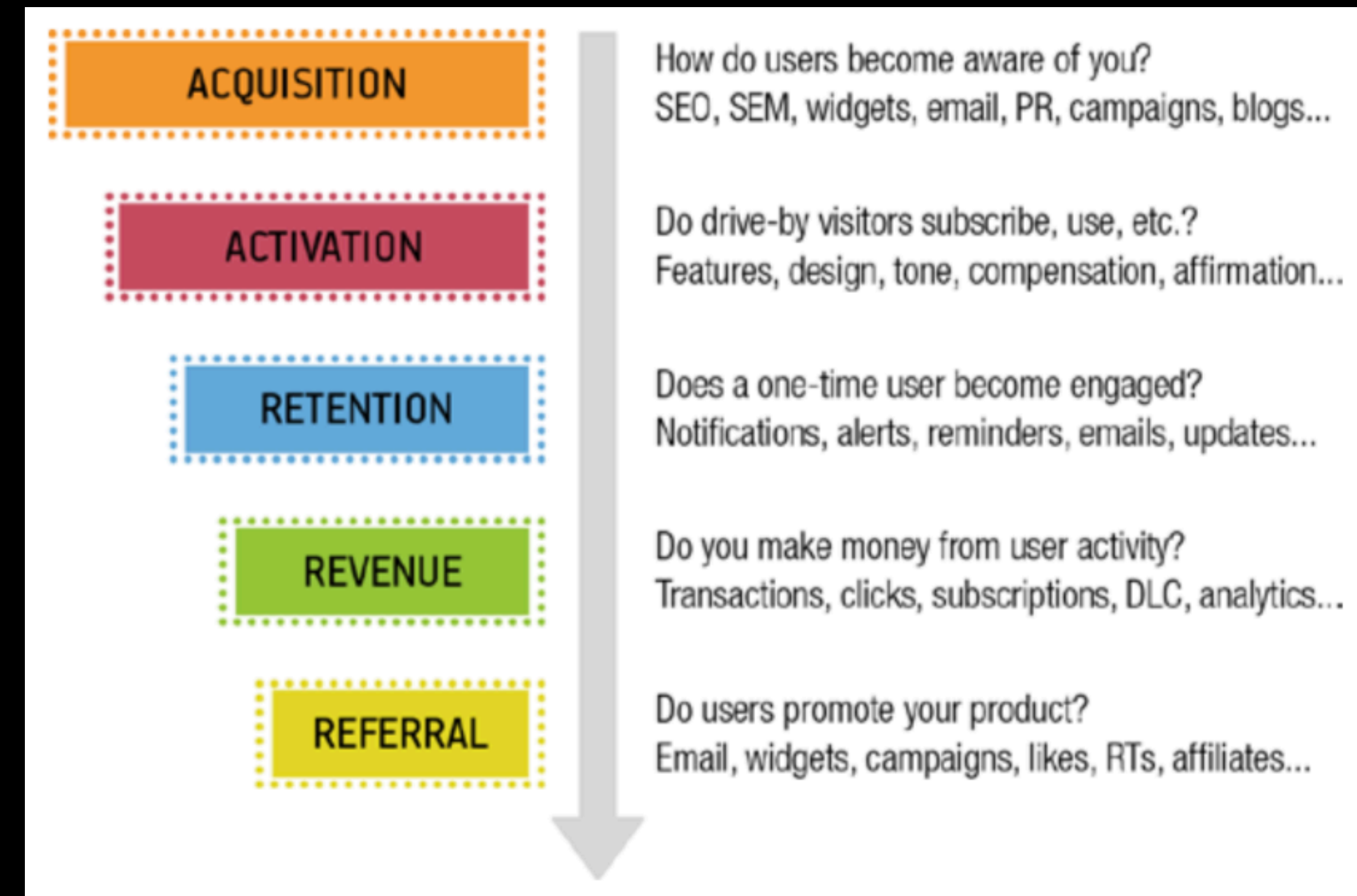
«AARRR» framework:

- **A**cquisition

- **A**ctivation

- **R**etention

- **R**evenue

- **R**eferral

Good at question: where to go next?
Whats to look in AB-sheet?

Bad at question: How concretely move further?
A bit b2c-centered.

AARRR

- **What:** Acquisition

- **About-what:** How do users become aware of you?

- **Instruments:** SEO / SEM / widgets / email / PR / campaigns / blogs

- **Utility:** Generate attention and traffic (either organic or not)

- **Metrics:** traffic / mentions / cost-per-click / search results / cost of acquisition / open rate

AA**R**RR

- **What:** Activation

- **About-what:** Are new users subscribing?

- **Instruments:** Features / design / tone / confirmations

- **Utility:** Turn attracted users into engaged

- **Metrics:** Engagement / registrations / onboarding finished / used service at least once

AA**R**RR

- **What:** Retention

- **About-what:** Are users returns?

- **Instruments:** notifications / alerts / reminders / updates / emails / game mechanics

- **Utility:** Encourages users to come back

- **Metrics:** engagement / time since last visit / DAU/MAU / «retention rate»: 1d, 7d, 30d …

AAR**R**R

- **What:** Revenue

- **About-what:** Are we earn moneys?

- **Instruments:** transactions / dlc / pricing / subscriptions …

- **Utility:** Money-money-money

- **Metrics:** ARPU (avg revenue per user) / conversion rate / shopping cart / click-through-revenue

AARR**R**

- **What:** Refferal

- **About-what:** Are our users promote our product?

- **Instruments:** promo / referral system / affiliates …

- **Utility:** Virality

- **Metrics:** invites / viral coefficient / viral cycle lifetime

What to track?!

OMTM:

One metric that matters:

• It answers the most important question you have

• It focuses the entire company

• It inspire a culture of experimentation

OMTM:

One metric that matters:

- Easy

- Every-day measured

- Easy to move

- Easy to compare

- Fundamental

OMTM:

One metric that matters:

- Easy

- Every-day measured

- Easy to move

- Easy to compare

- Fundamental

PROBLEMS?

OMTM:

One metric that matters:

- Easy

- Every-day measured

- Easy to move

- Easy to compare

- Fundamental

PROBLEMS?

YEP, it does not exists!

OMTM:

One metric that matters:

- Easy

- Every-day measured

- Easy to move

- Easy to compare

- Fundamental

Don't worry, we will fix it further!

«Startup» progress:
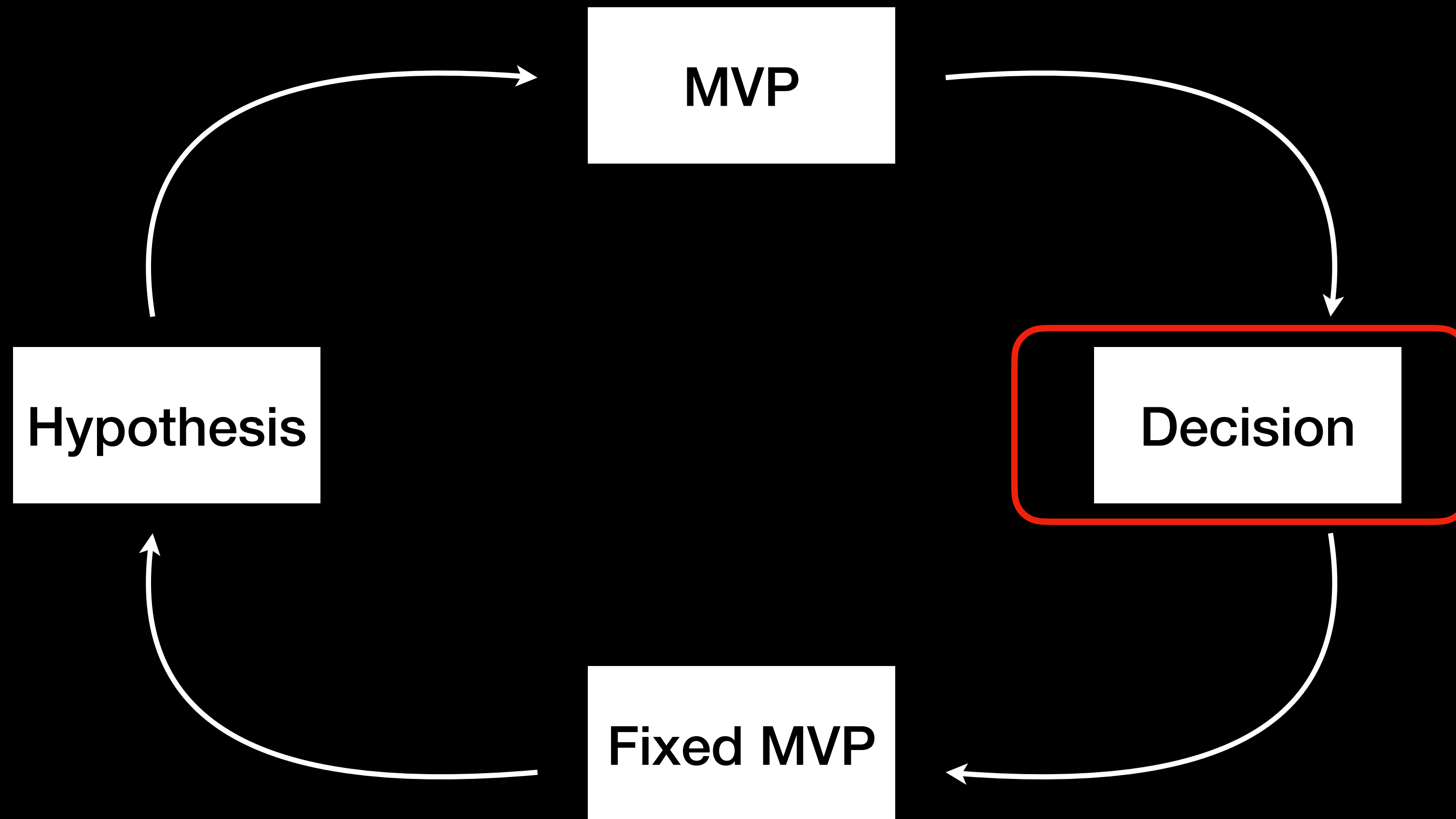
A/B tests:

- Two options of banner. Which one is better?

- Split user into two groups,
  show default banner for control group,
  new one — for test group.

- Wait a bit

- Measure performance.

- Compare
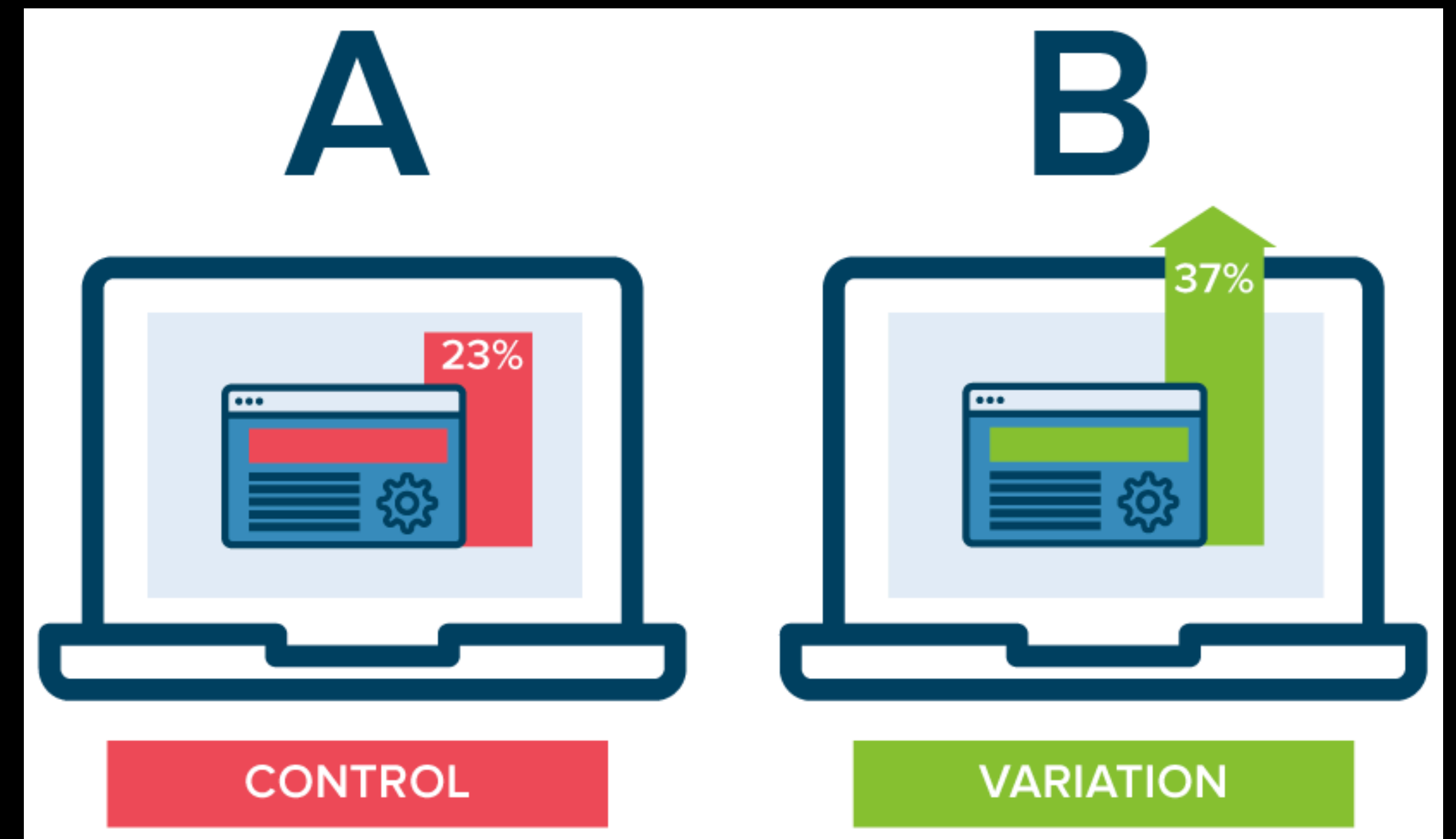
- Choose best

A/B tests:

- Two options of banner. Which one is **better?**

- **Split user** into two groups,
  show default banner for control group,
  new one — for test group.

- Wait **a bit**

- Measure **performance**.

- **Compare**

- Choose **best**
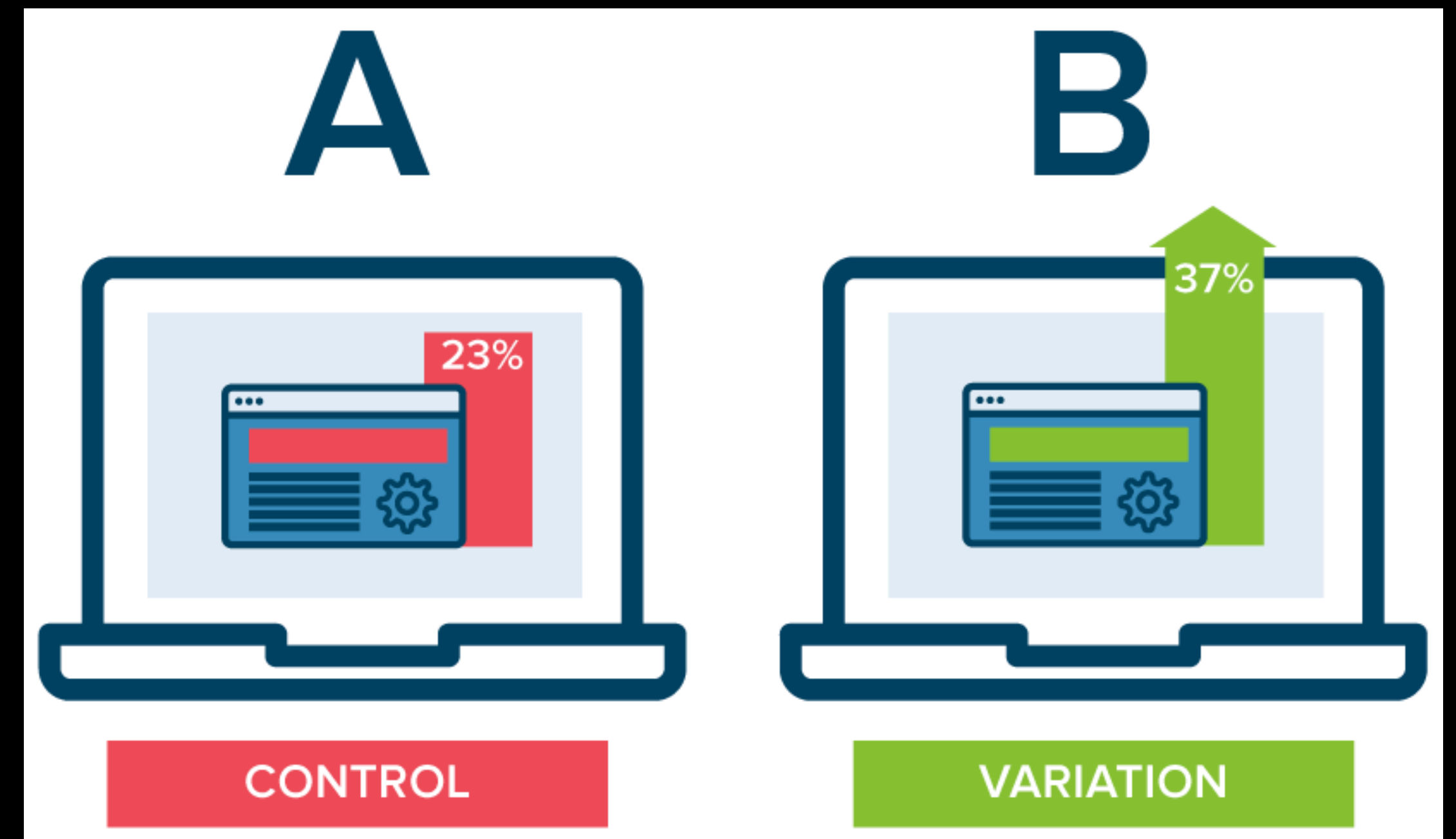
How to compare?

- $k \sim Bernoulli(\theta)$, k — binary click or not
  θ — Bern. parameter, probability of a click, mean…
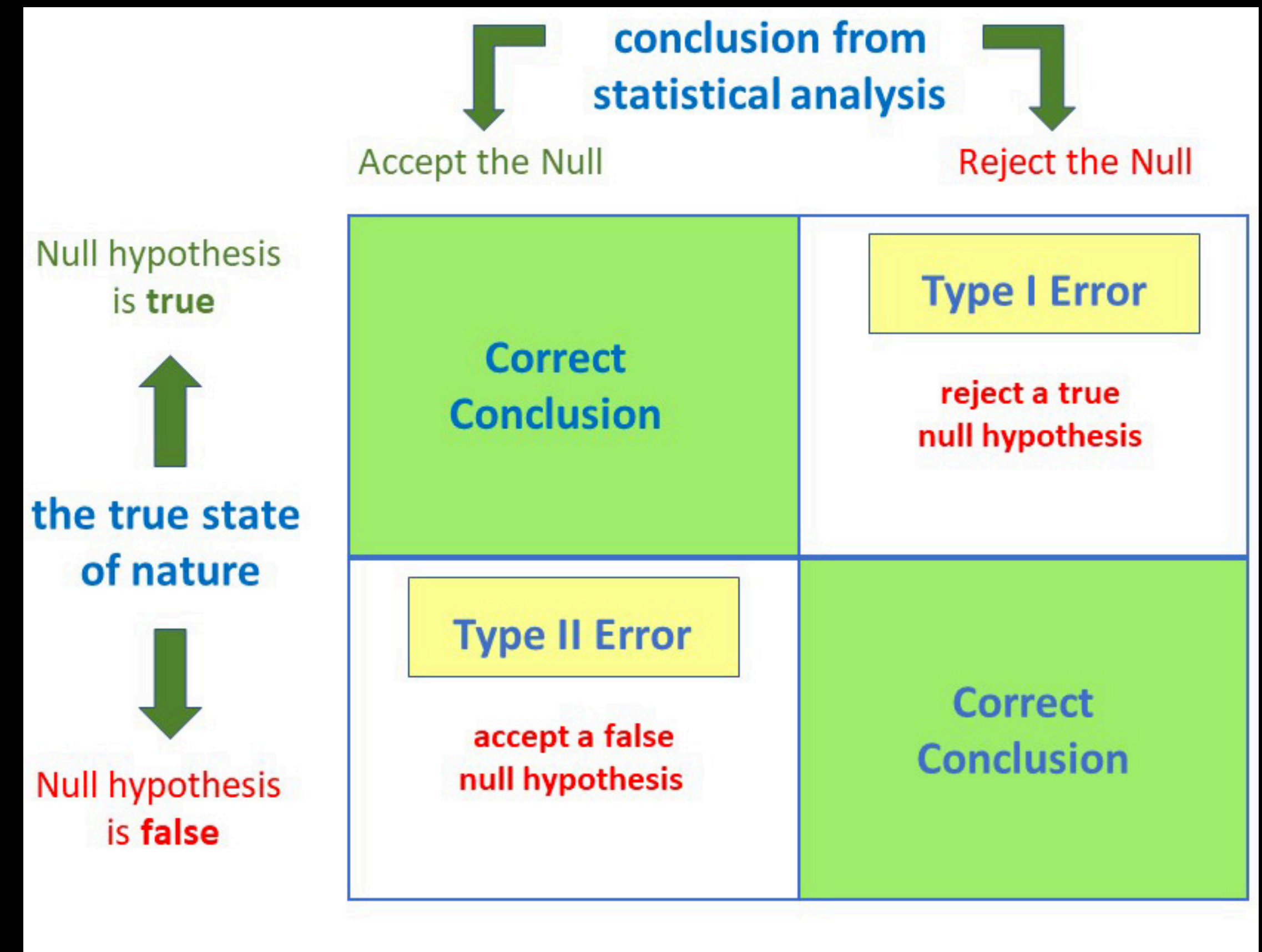
  $p(k) = \theta^k (1-\theta)^{1-k}$, $\sigma^2 = \theta(1-\theta)$ — dispersion

- $H_0 : \theta_c = \theta_t$
  $H_1 : \theta_c < \theta_t$
  $\theta_c \sim clicks/views$

- $\bar{\theta} = N(\mu, \dfrac{\sigma^2}{n})$ —> **T**-test for comparing
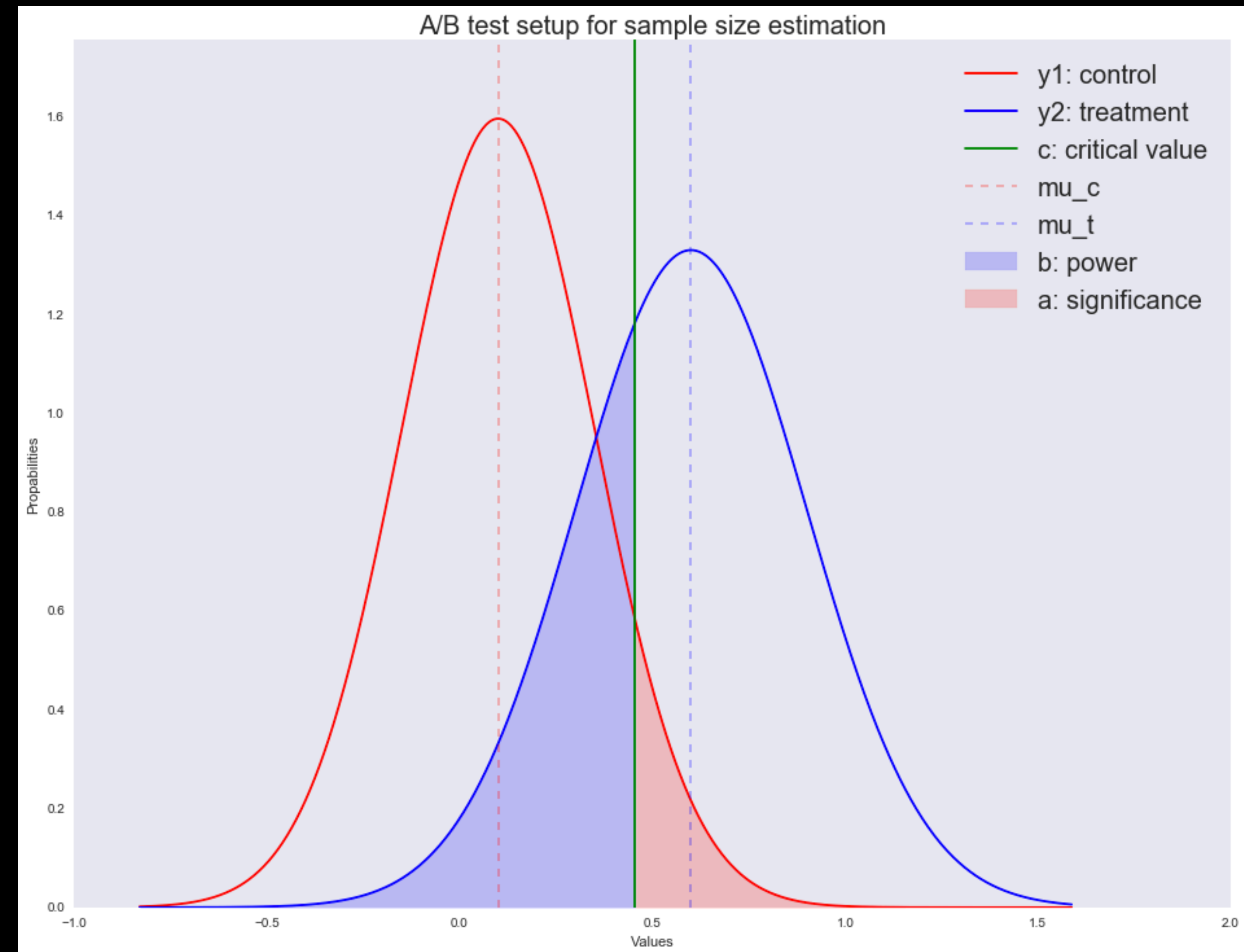
How to **compare?**

- $\alpha = P(H_1 | H_0)$ (0.05)
  $\beta = P(H_0 | H_1)$ (0.2),
  $(1 - \beta) -$ statistical power)

- Stat test concept:
  $T(X) = t$
  $P(T(x) > t) = p_{value}$
  $p_{value} < \alpha -$ H_0 rejected

How long to hold?

- $c = \mu + t\dfrac{\sigma}{\sqrt{n}}$, t − quantile N(0,1)

- $c = \theta_c + t_\alpha\sqrt{\dfrac{\theta_c(1 - \theta_c)}{n}}$

  $c = \theta_t + t_\beta\sqrt{\dfrac{\theta_t(1 - \theta_t)}{n}}$

- ^^ with known θ_c, α, β and expected θ_t solvable for n

A/B test setup for sample size estimation

Dead salmon case

- Happens on multiple tests

- Take dead salmon, put them into MRI

- Conduct low-quality research

- Salmon is alive

- Holm–Bonferroni method $\alpha := \alpha/n$

$$P\left(\text{хотя бы один результат значимый}\right) = 1 - P\left(\text{все результаты незначимы}\right)$$
$$= 1 - (1 - 0.05)^5$$
$$= 1 - 0.95^5$$
$$\approx 0.2262$$
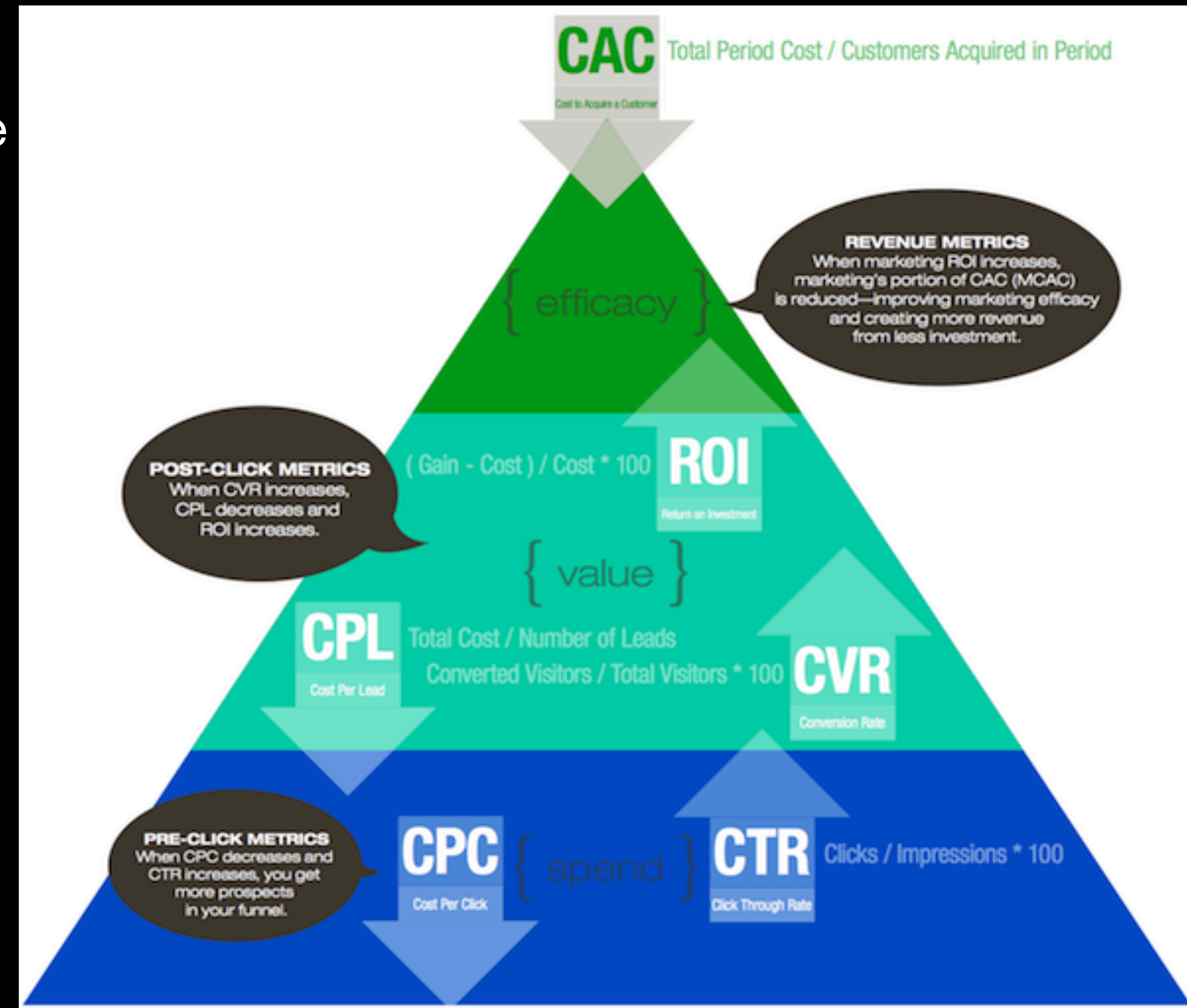
AB tests IRL:

- Tons of metrics:

  - No one knows what amount of change to care

  - How to make a decision?

  - How long should we hold for 100 metrics on radar

  - Have you heard about novelty effect?

Действие	Метрика	Эффект (%)	Увереность в отличии (%)	Увереность в новизне (%)
	users	-62.16%	99.00%	97.38%
	counts	-57.08%	99.00%	97.38%
	users	36.40%	99.00%	97.38%
	counts	35.66%	99.00%	97.38%
	counts	28.75%	99.00%	97.38%
	counts	28.53%	99.00%	97.38%
	users	24.04%	99.00%	97.38%
	avgs	14.60%	99.00%	97.38%
	avgs	8.10%	99.00%	97.38%
	users	7.48%	99.00%	97.38%
e_click:Content	avgs	-6.60%	99.00%	97.38%
e_click:Content	users	5.83%	99.00%	96.10%

AB tests IRL:

- Tons of metrics:

  - No one knows what amount of change to care

  - How to make a decision?

  - How long should we hold for 100 metrics on radar

  - Have you heard about novelty effect?

Good AB tests IRL:

- Automatised

- Based on metrics. Plug-in plug-out

- Based on separation unit. Flexible separation

- A/A test auto and stable

- Easy for everyone

- Based on world-wide experience:
  https://exp-platform.com/

AB tests nature:

- Origins from medicine and agriculture:

  - High cost of one **experiment**

  - High cost of an single **error**

- Classical AB/tests allows you to estimate budget beforehand, set different α, β and estimate result
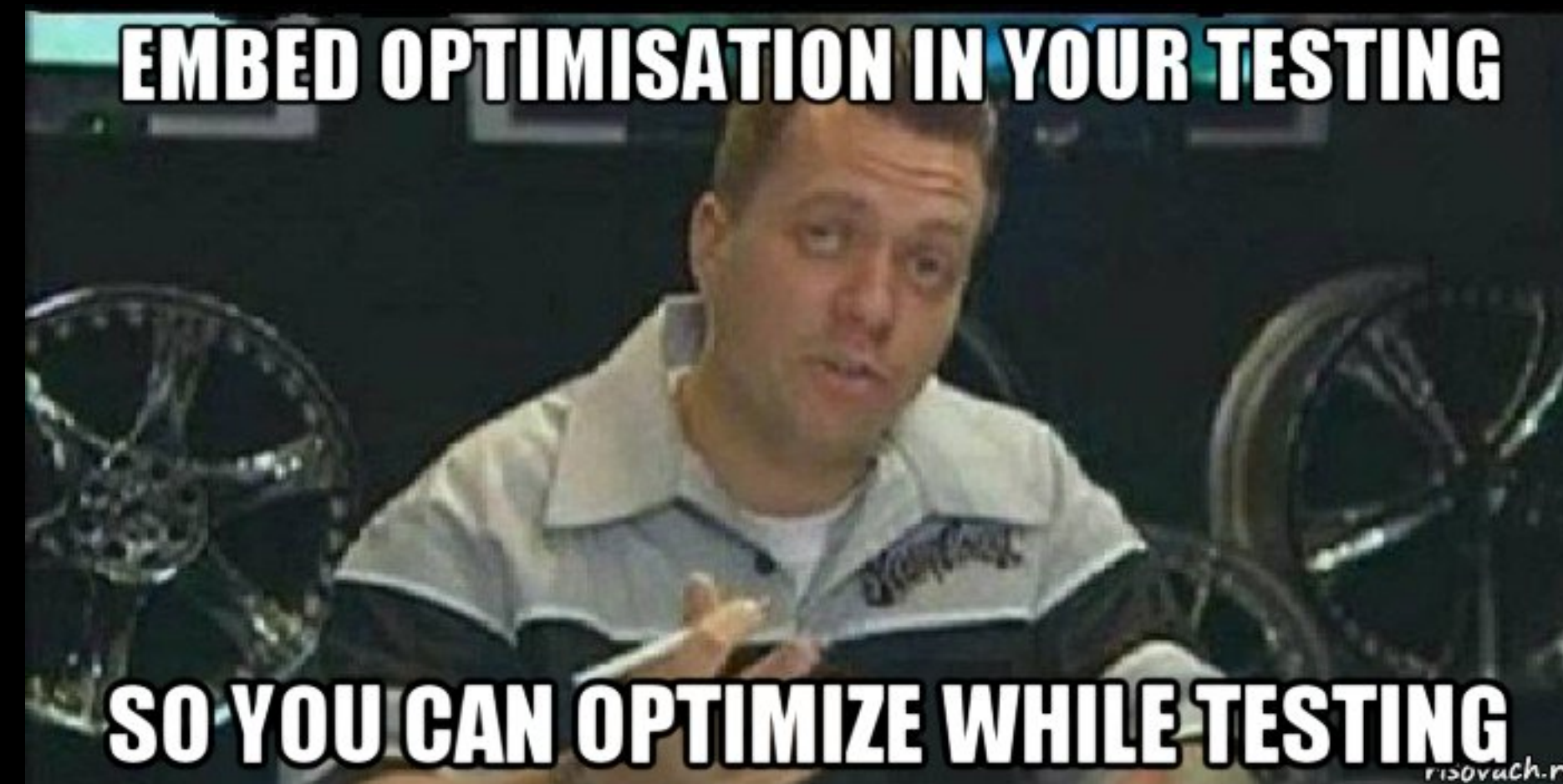
Bayesian multi-armed bandits:

Unformally:

- You ended up in casino with fixed amount money and time.

- You want to find the best one and maximise your gain.

- Let's look at Thompson sampling schema.

- Yahoo use it to optimise banners. Microsoft too. Netflix and artwork optimisation.

Bayesian multi-armed bandits:

Formally:

- Let us observe the sequence $y_t = (y_1, y_2 \dots y_t)$ at time t

- Let us denote action a_t taken at moment t

- $y_t \sim f_{a_t}(y \,|\, \vec{\theta})$, where $\vec{\theta}$ some unknown parameters vector

- We don't know the actual distribution and/or $\vec{\theta}$ thats why we can't optimise mean directly

Bayesian multi-armed bandits:

- $f_a(y \mid \theta_a) = \theta_a^y (1 - \theta_a)^{1-y}$ and expected reward $\mu_a = \theta_a$. So let we **assume** that $\theta \sim Beta(\alpha, \beta)$

- Beta is conjugate prior to Bernoulli

$$
\begin{aligned}
p(\theta \mid y) \;&\propto\; p(\theta) \cdot p(y \mid \theta) \\
&\propto\; \tfrac{1}{\mathrm{B}(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \theta^y (1-\theta)^{1-y} \\
&\propto\; \theta^{\alpha-1+y}(1-\theta)^{\beta-1+1-y}
\end{aligned}
$$

$$
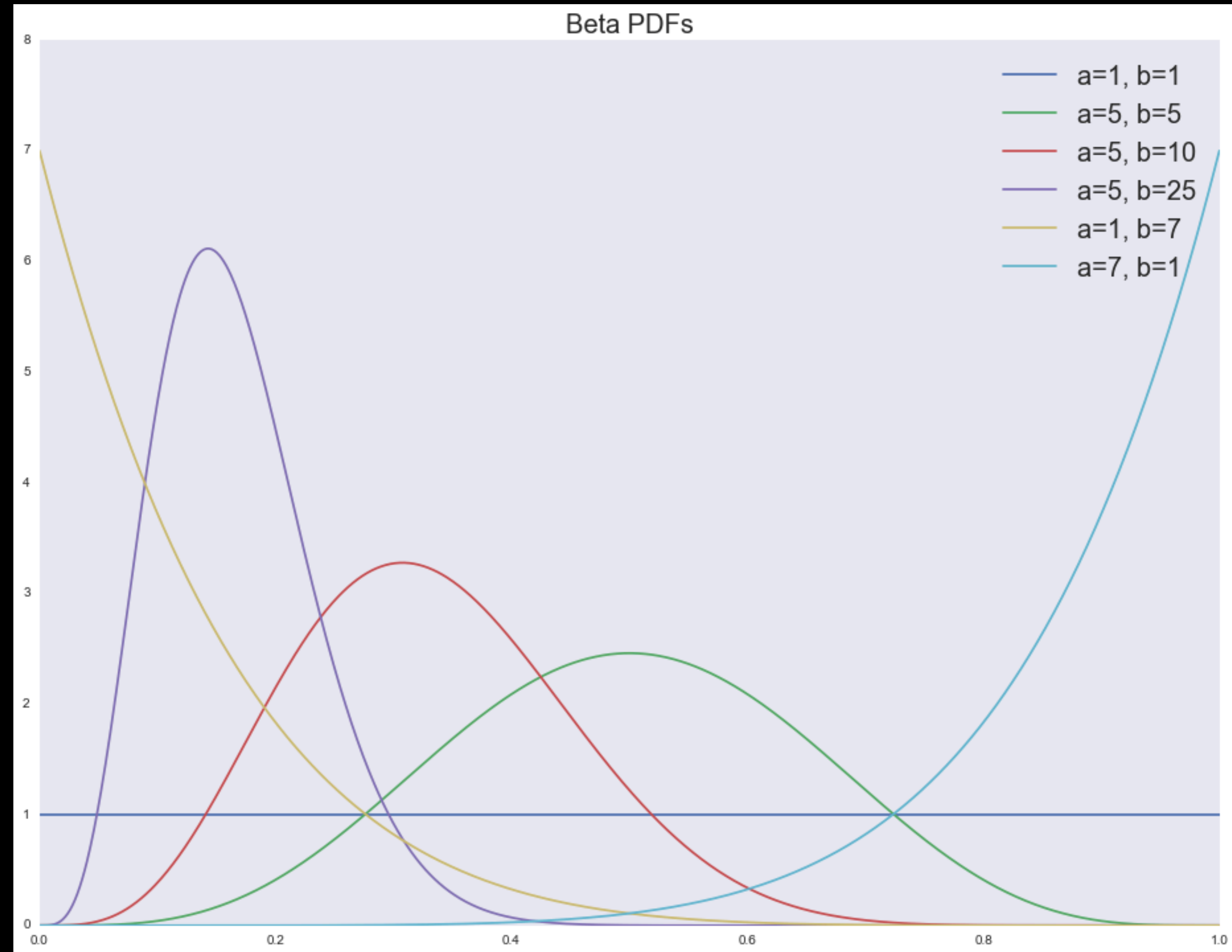\mathrm{Beta}(\alpha + y, \beta + 1 - y) \;=\; \mathrm{Beta}(\theta \mid \alpha, \beta) \cdot \mathrm{Bernoulli}(y \mid \theta)
$$

$$
p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{\displaystyle\int_{\text{range } \theta} p(x|\theta)\, p(\theta)\, d\theta}.
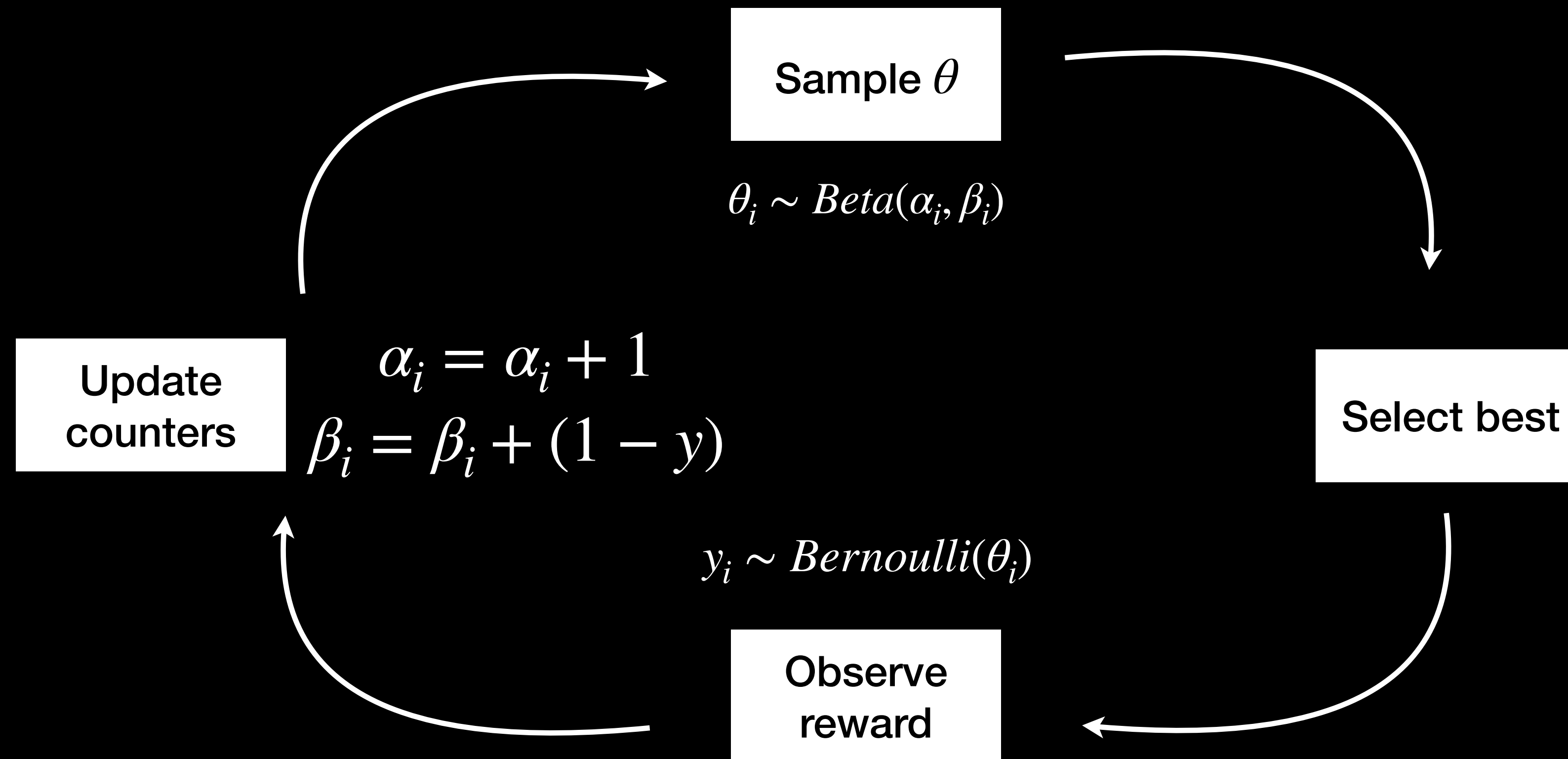$$

- Beta becomes uniform when $\alpha = \beta = 1$

- α as num of successes, β as num of failures

Beta-distribution:

$$f(\theta, \alpha, \beta) = \frac{1}{\mathbf{B}(\alpha, \beta)} \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$$
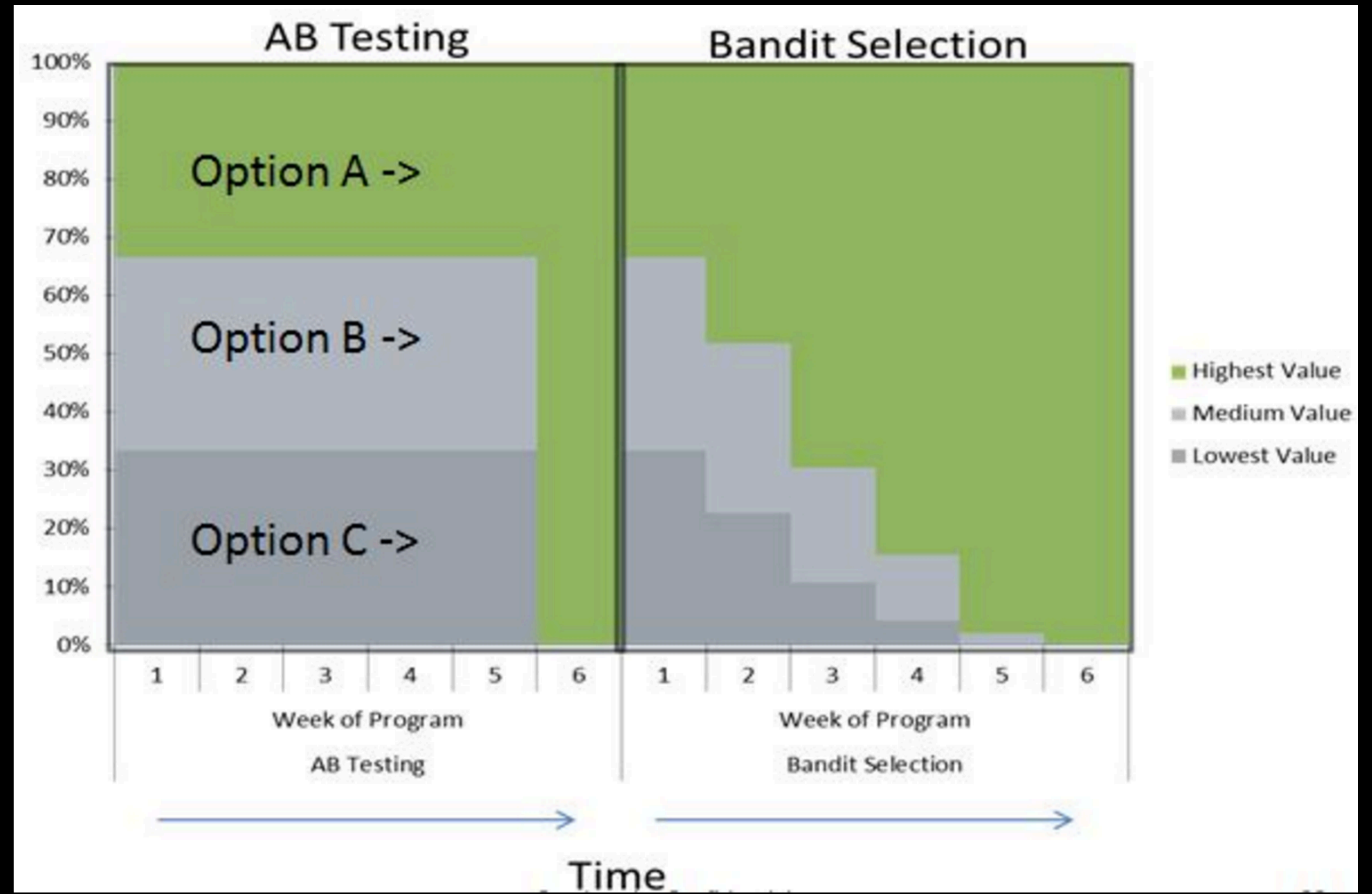

Beta PDFs

- a=1, b=1
- a=5, b=5
- a=5, b=10
- a=5, b=25
- a=1, b=7
- a=7, b=1

Beta-distribution sampling

Sample θ

$\theta_i \sim Beta(\alpha_i, \beta_i)$

Update counters

$\alpha_i = \alpha_i + 1$
$\beta_i = \beta_i + (1 - y)$

Select best

$y_i \sim Bernoulli(\theta_i)$

Observe reward

Beta-distribution sampling

- Traffic proportion ~ probability of the best bandit $\theta*$

- Have all needed counters to perform AB-test calculation on runtime

- Maximises CTR directly

Bandits as recommender:

- Allows to add new options during experiments

- Allows to take account of a-priori by setting non-default α, β

- Greediness problem -> $$ewma_i = (1-\alpha) \bullet ewma_{i-1} + \alpha \bullet freq_i$$

- Relatively easy to code. Especially if you have no historic data and need to implement something on launch.

Sources

- Metrics: Lean Analytics: How to build startup faster

- Bandits https://habr.com/ru/company/ods/blog/325416/