GENERAL ASSEMBLY

WHERE TO GO NEXT

John Sandall

Data Science Consultant

@john_sandall

BUSINESS FUNDAMENTALS

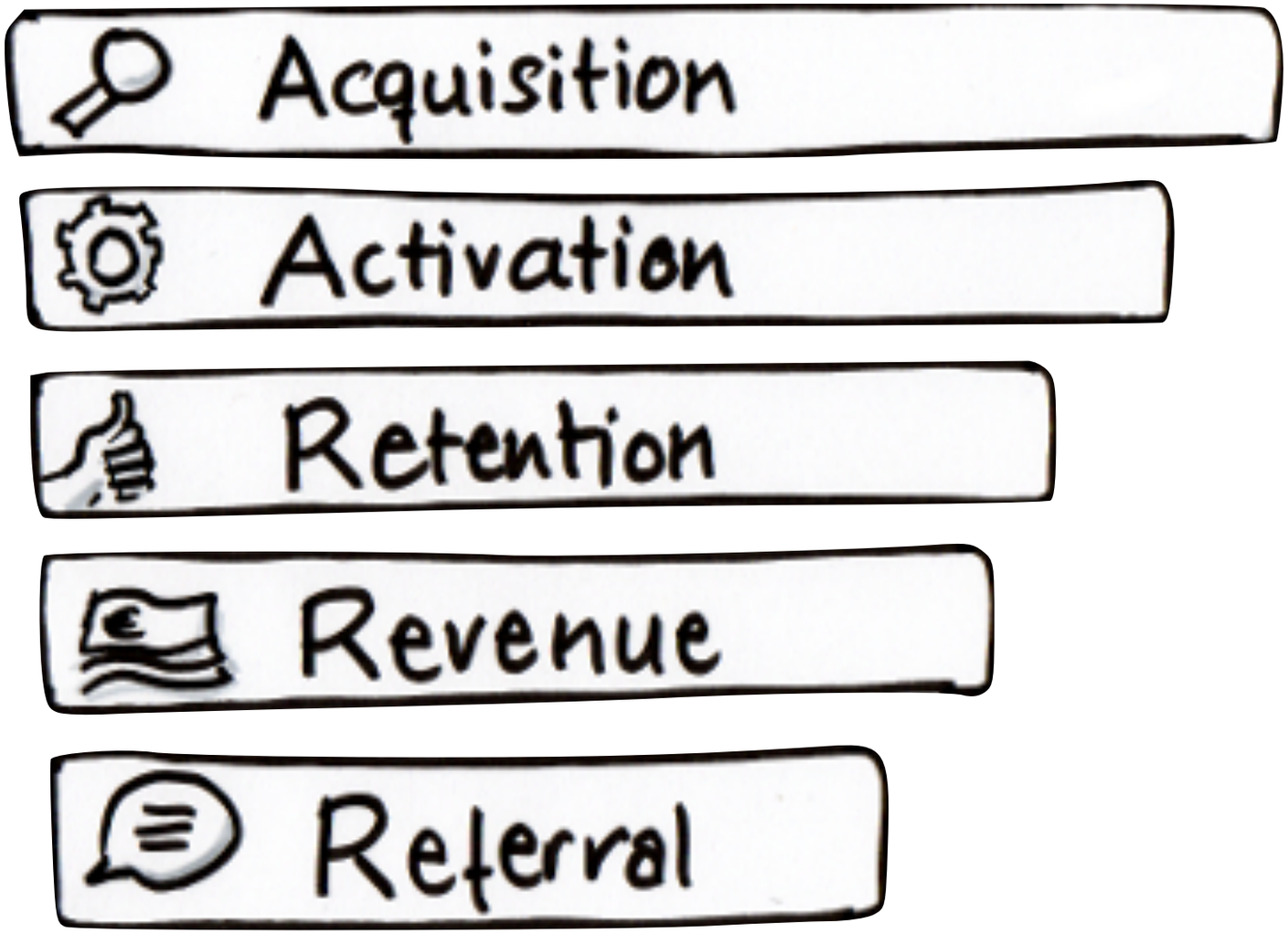
BUSINESS FUNDAMENTALS

- AARRR
- Unit Economics
- Retention/Churn
- Product/Pricing Optimisation
- Operations Research
- Demand Prediction
- Customer Segmentation
- ...

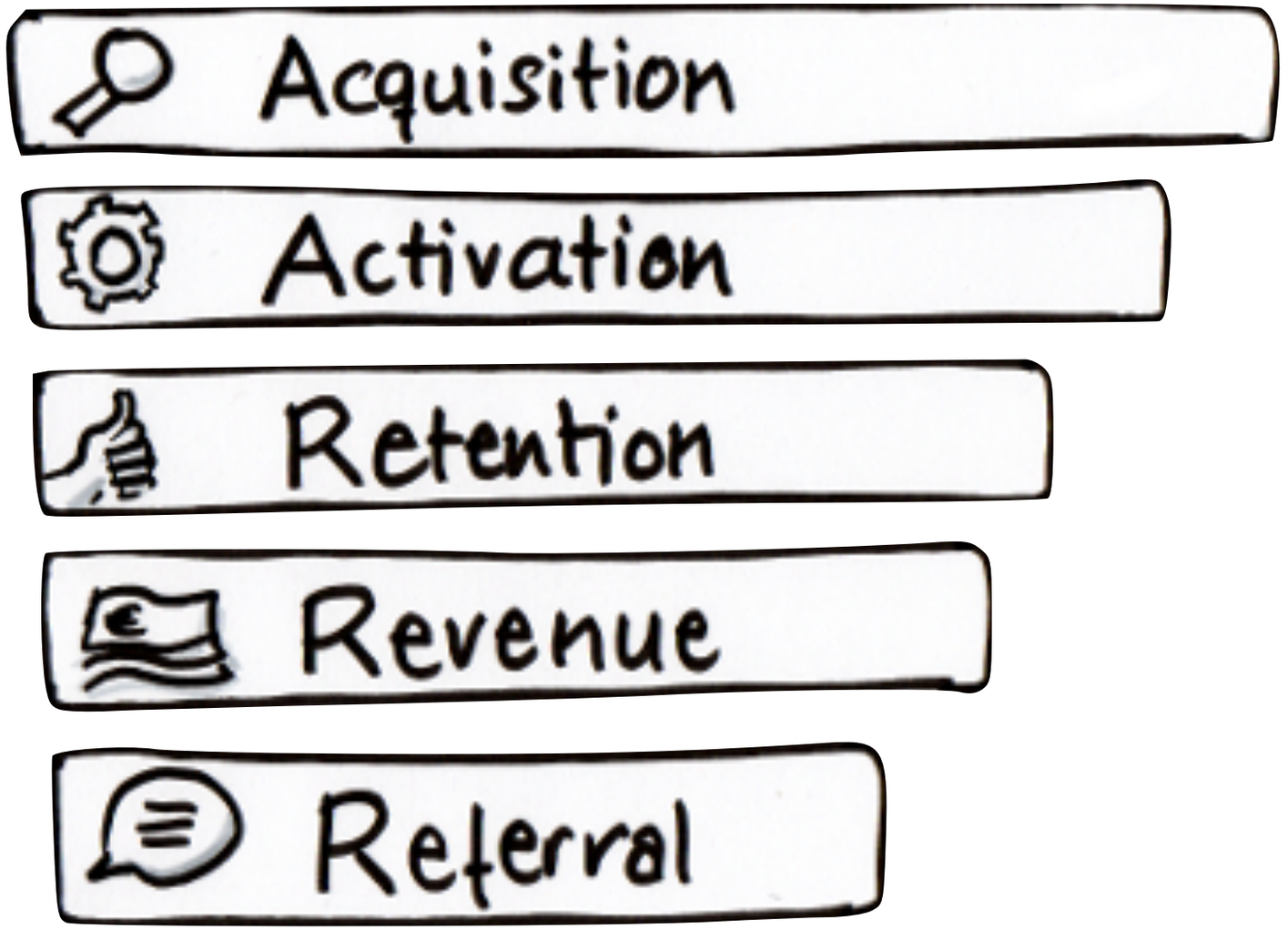
GETTING YOUR METRICS RIGHT

- **Goal:** Use data science to optimise social media strategy.
- **Questions:** What do you mean by "optimise"? What is "success" here?
- In order to optimise anything, we need a metric to optimise.
- My goal here is to provide insight into how a data scientist might approach some common problems in the world of business, marketing and social media.

AARRR ("PIRATE METRICS")



AARRR ("PIRATE METRICS")



Visitor (unregistered)

↓

Registered user

↓

Loyal user

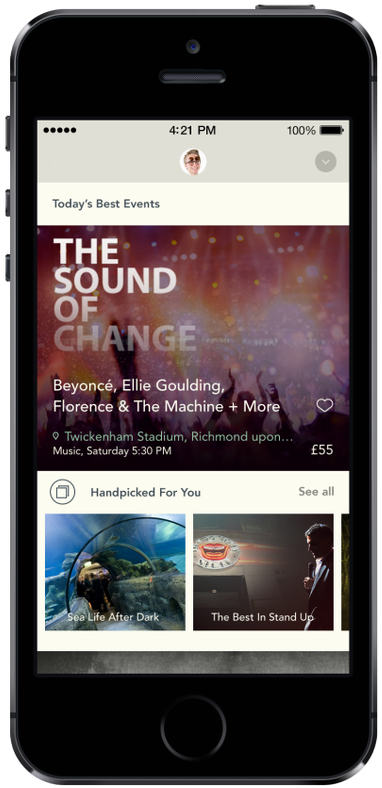
↓

Paying customer

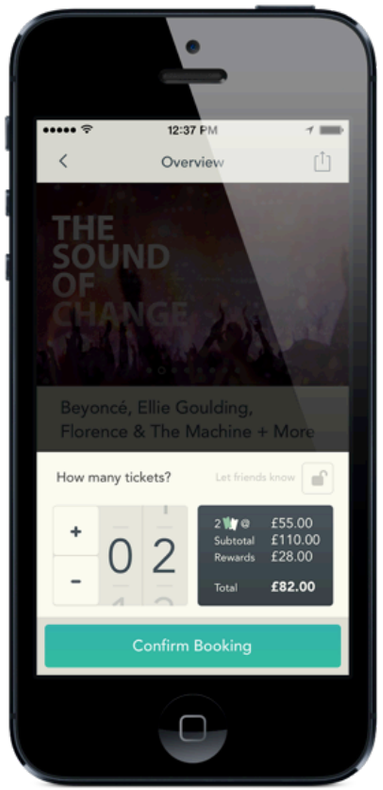
↓

Promoter

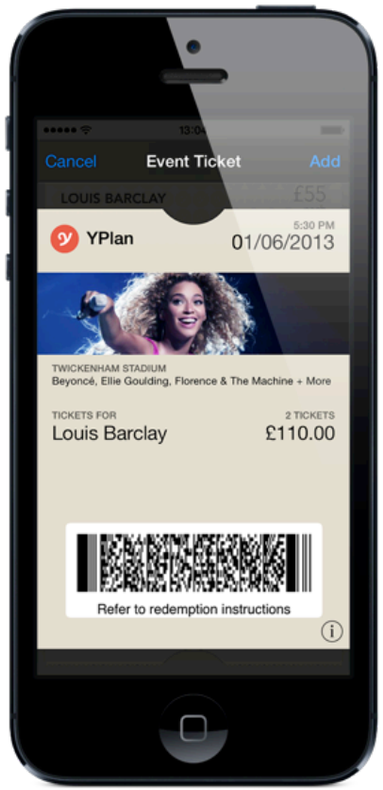
AARRR FOR YPLAN



Spontaneous inspiration



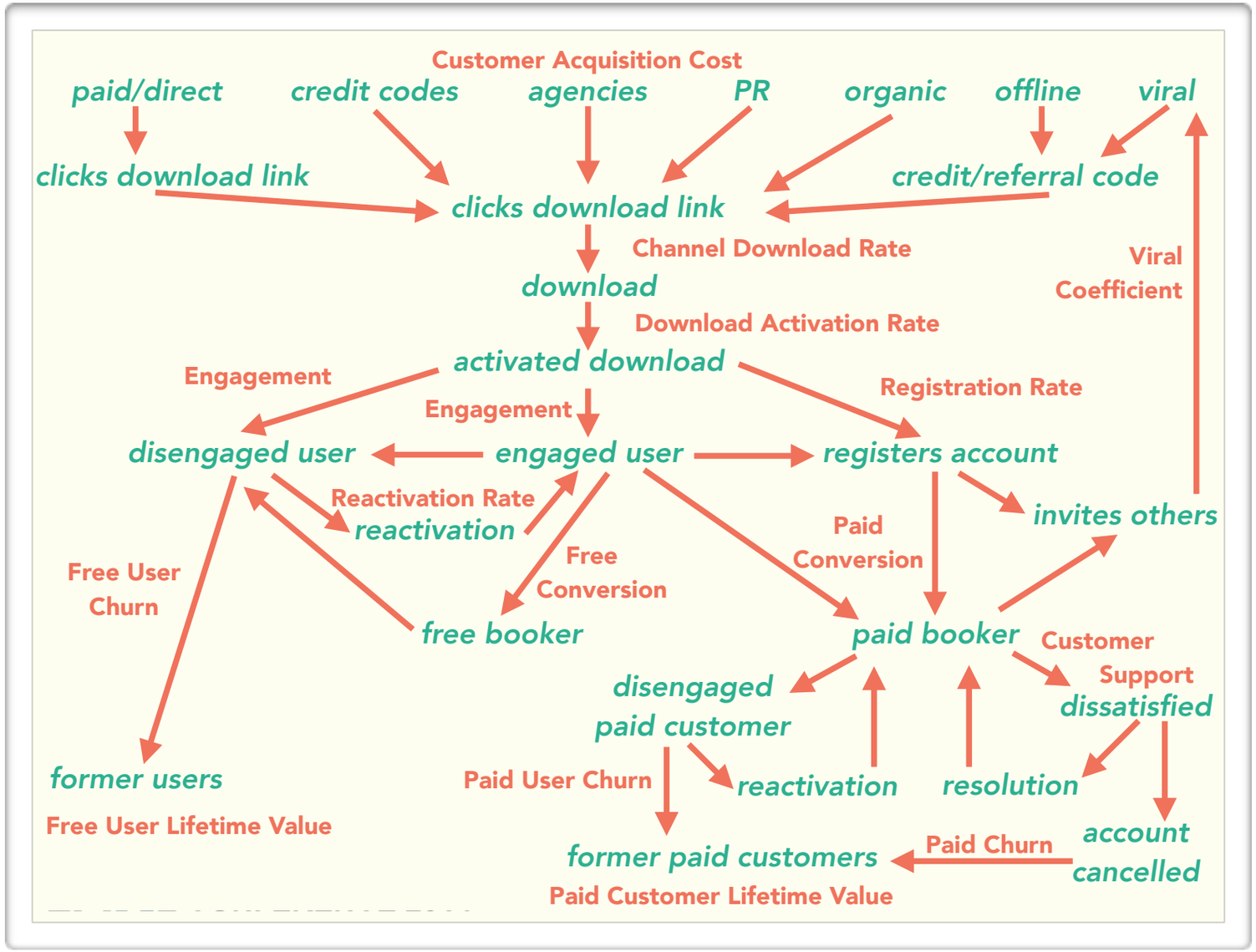
2-tap booking process



Paperless ticketing

- Discovery of new experiences
- Curated events
- Social media integration
- 100% mobile

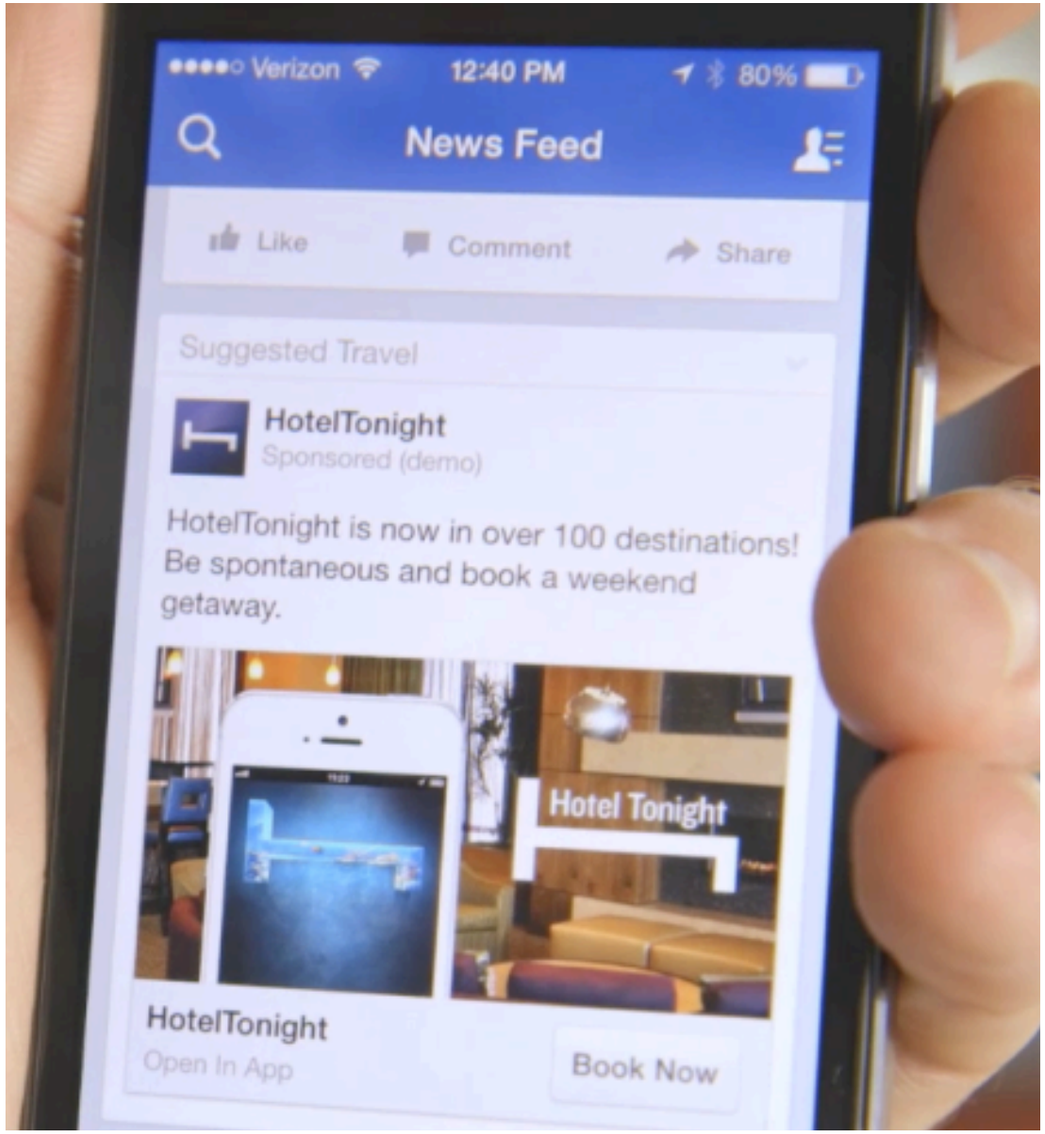
AARRR FOR YPLAN



LESSON: BE VERY CLEAR ABOUT WHAT IT IS YOU'RE TRYING TO IMPACT

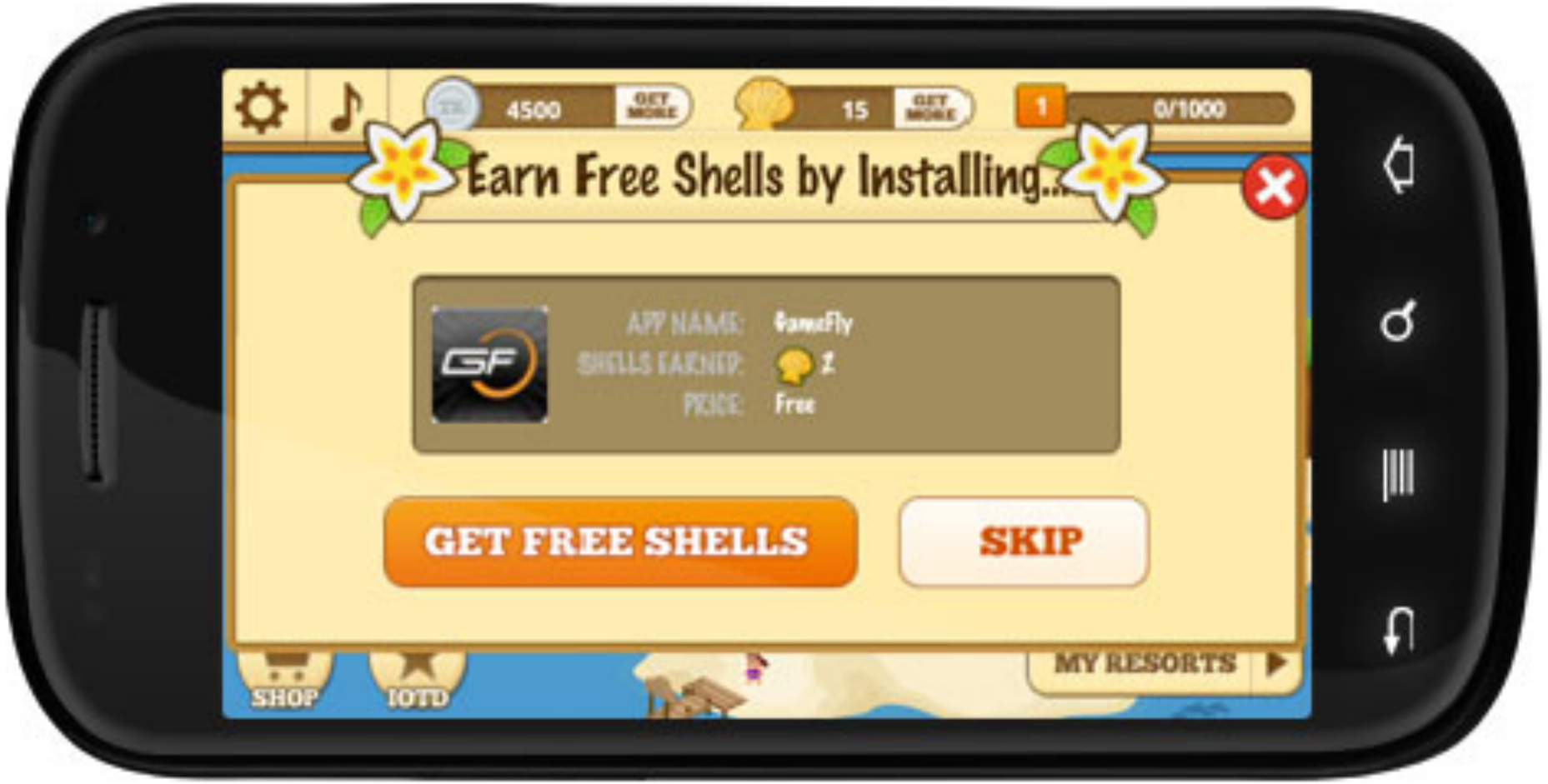
EXAMPLE: ADVERTISING ON GOOGLE/FACEBOOK/AD NETWORKS

- **Goals:** Awareness? Acquisition? Both?
- **Track:** Each activity against these goals. Want to prove causality.
- **Proxy metrics:** Ad impressions are useful…if they follow through!



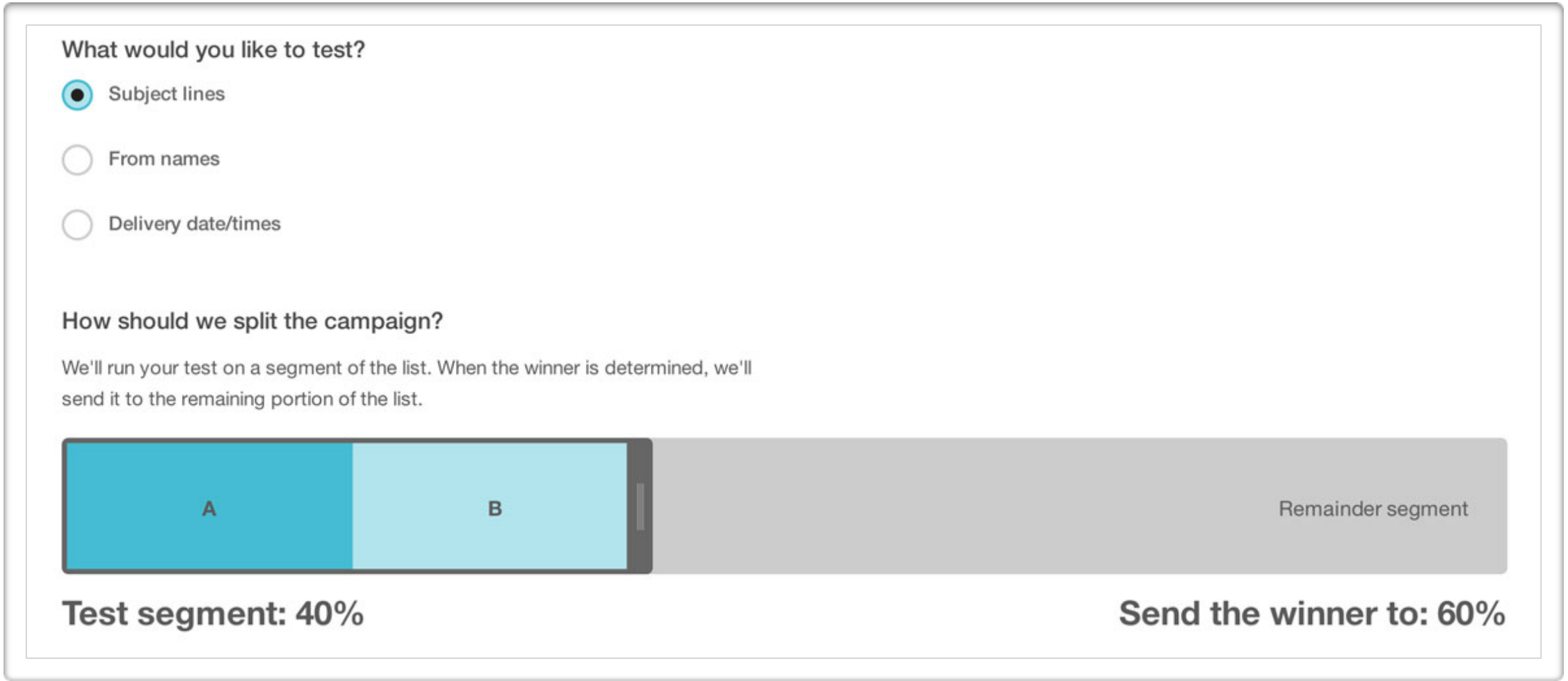
EXAMPLE: ADVERTISING ON GOOGLE/FACEBOOK/AD NETWORKS

- **Proxy metrics going wrong:** Incentivised downloads



EXAMPLE: ADVERTISING ON GOOGLE/FACEBOOK/AD NETWORKS

- **Proxy metrics going wrong:** Optimising subject lines for open rate only



EXAMPLE: SOCIAL MEDIA ACTIVITY

- **Goals:** Difficult to isolate!

- **Guiding question:** "If we stopped doing this, what would happen?"

- **Some example goals:**
  - Engaging new users (social media as an acquisition tool)
  - Engaging existing users (social media as a retention tool)
  - Persuading acquired users to activate into paying customers
  - Building communities shifts people from "like" to "love" to "promoters/defenders"

EXAMPLE: SOCIAL MEDIA ACTIVITY

- **Impacts:** Almost every part of AARRR funnel!

- **Example:** Engaging new users (social media as an acquisition tool)
- **Measure:**
  - # of new users from social media sources
  - % of new users from social media sources
  - segment by social media platform
  - segment by post
  - look at trends over time
  - look at best/worst posts from last month…what can we learn?

EXAMPLE: SOCIAL MEDIA ACTIVITY

- **Impacts:** Almost every part of AARRR funnel!

- **Example:** Measuring impact of social media activities on referrals
- **Problem:** Effectively word of mouth…referred users look like organics.
- **Measure:**
  - Total/proportion of new users from organic sources
  - Experiment: increase/decrease activities, look for correlations
  - Regular surveys of new users: "how did you hear about us?"
  - Directly link customer accounts with social media profiles
    - e.g. Insightly: "*We'll detect virtually every social media profile related to a contact's email address*"
  - Accurately measuring brand equity is notoriously difficult!

CASE STUDY

DRIVING DOWN CAC USING PREDICTIVE MODELS

WHAT IS CAC?

Cost Per Install (CPI)

Signup Conversion Rate %

Booker Conversion Rate %

WHAT IS CAC?

Cost Per Install (CPI)

Signup Conversion Rate %

Booker Conversion Rate %

Customer Acquisition Cost (CAC)

CAC = CPI x Signup Conversion Rate x Signup-to-Booker Conversion Rate
= Cost per Acquired Customer

FIRST CHALLENGE: BRING THE DATA TOGETHER

EXAMPLE: PREDICTING CUSTOMER CONVERSION

- **Metric:** CAC here means "cost to acquire an install that made a paid booking within 30 days of install"

- **Problem:** Have to wait 30 days before you can evaluate if a campaign is working or not

- **Goals:**
  - Determine leading indicators that are predictive for conversion to paid customer.
  - Build a classifier model using these indicators to predict conversion rates.
  - Is it possible to predict within 24h of install which new installs will convert?
  - Putting it all together…we can accurately estimate CAC within 24h of a new campaign going live.
  - This means much faster turnarounds, massively speeds up learning cycle of what works.

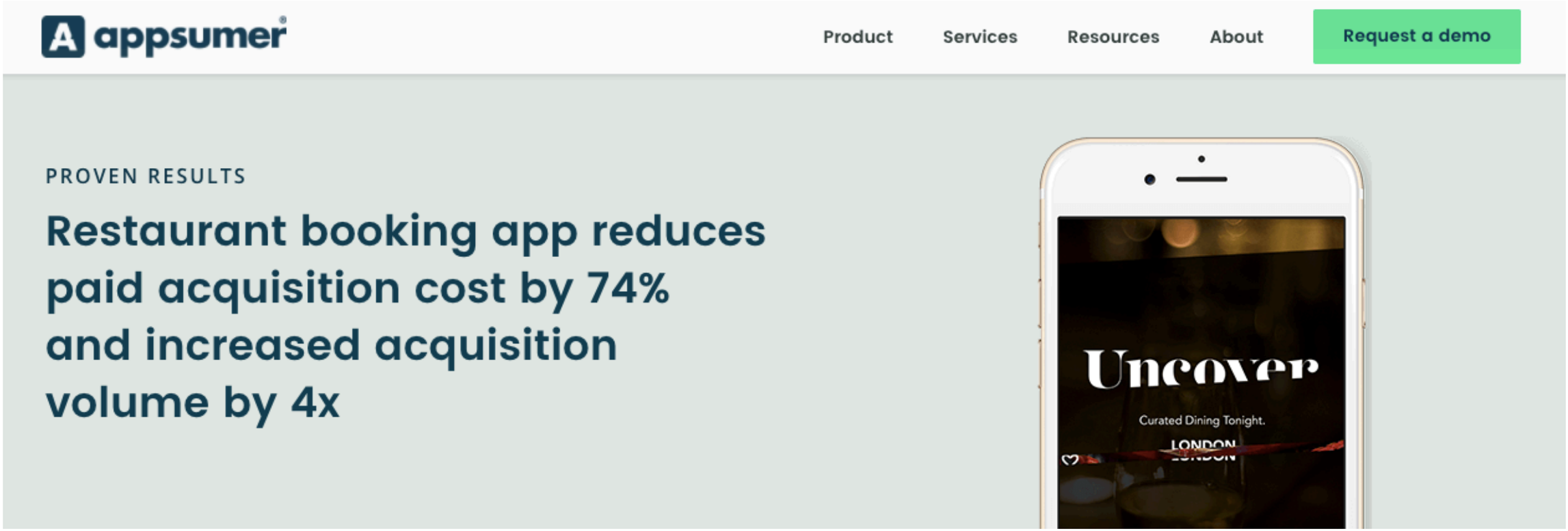
EXAMPLE: PREDICTING CUSTOMER CONVERSION

- **How:**
  - What data is there?
  - Throw this into some machine learning classifier models, e.g. decision tree, random forest
  - Make predictions! Look at them 30 days later. How good were they? Learn, iterate, improve.

user_id	source	device	location	restaurants viewed	skipped tutorial	...	booked (within 30 days)
1	facebook	android	london	5	Y	...	✗
2	twitter	ios	new york	10	N	...	✗
3	organic	ios	bristol	15	N	...	✗
...	...	...	...	...	...	...	✓

EXAMPLE: PREDICTING CUSTOMER CONVERSION

- **Impact…**



EXAMPLE: PREDICTING CUSTOMER CONVERSION

- **Impact...**
  - 74% reduction in CAC within 3 months
  - Precise understanding of behavioural indicators that result in paying customers => product improvements
  - High-probability customers who didn't convert are ideal "low hanging fruit" for personalised targeted offers
  - Data warehouse infrastructure can be reused for other data science projects (e.g. churn prediction, recommendation systems, product optimisation)

CUSTOMER SEGMENTATION

RFM MODELS

Traditional approach to customer segmentation by:

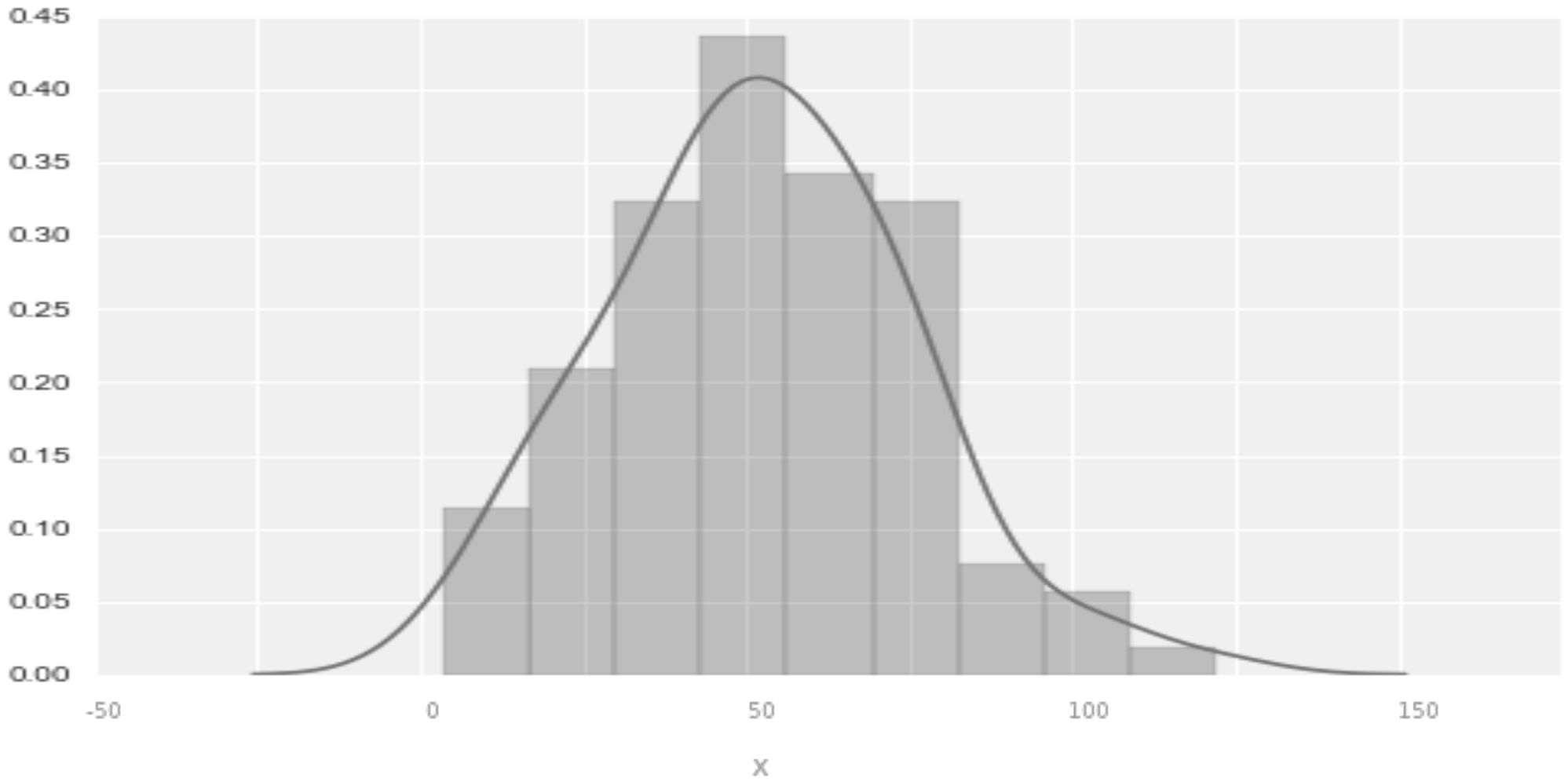
- **Recency:** How recently did we see this customer?
- **Frequency:** How often do we see this customer?
- **Monetisation:** How much does this customer spend per transaction?

MORE ADVANCED STILL: LIFECYCLE MODELS

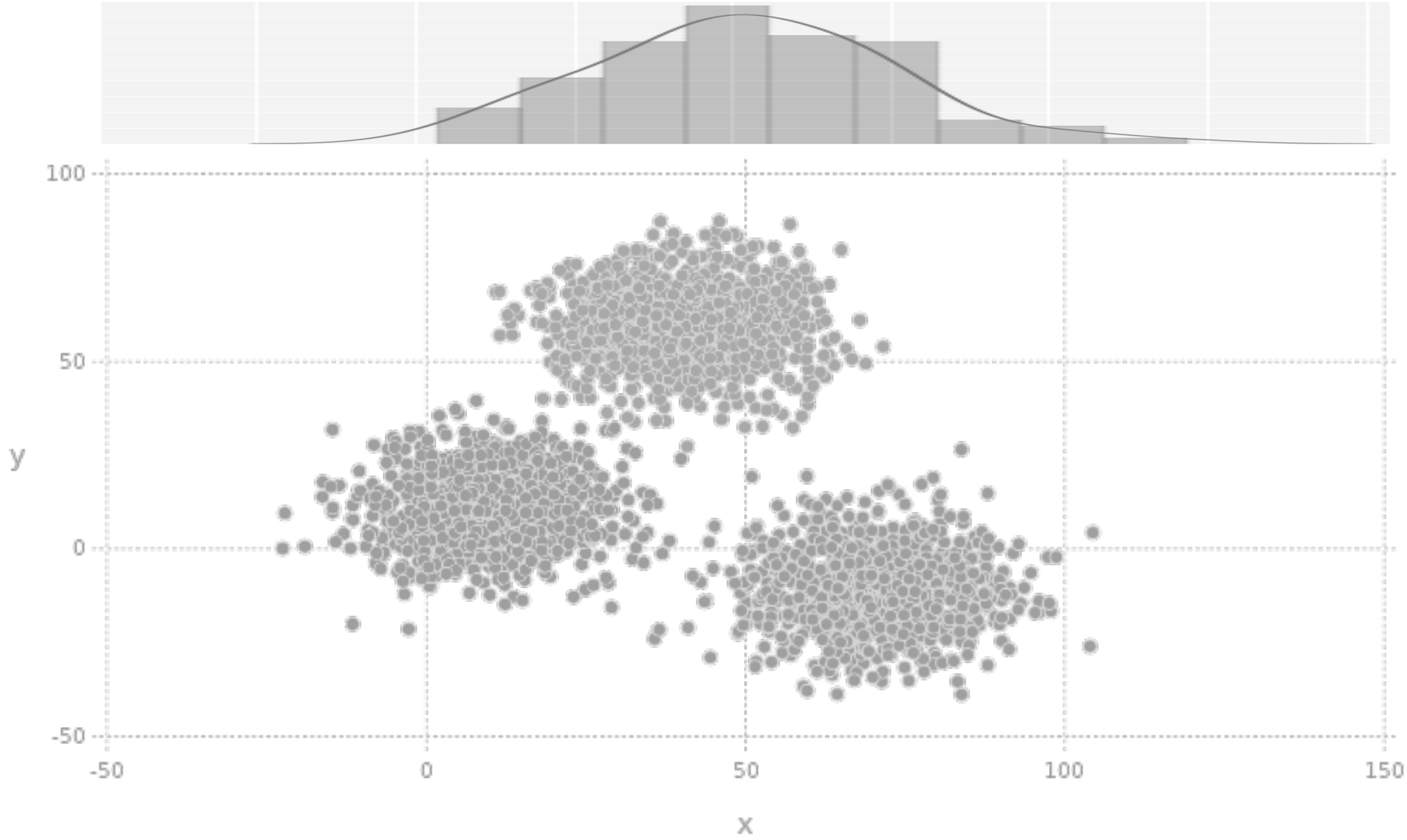
Dimensions to segment on:

- **Conversion Funnel:**
  - Visitor → Registered user → Paying customer → Repeat customer
- **Time since last engagement:**
  - 7 days ("active") → 30 days ("churn risk") → 90 days ("churned")
- **Monetisation:**
  - High vs Medium vs Low value customers
- And more…product, SKU, behavioural groupings

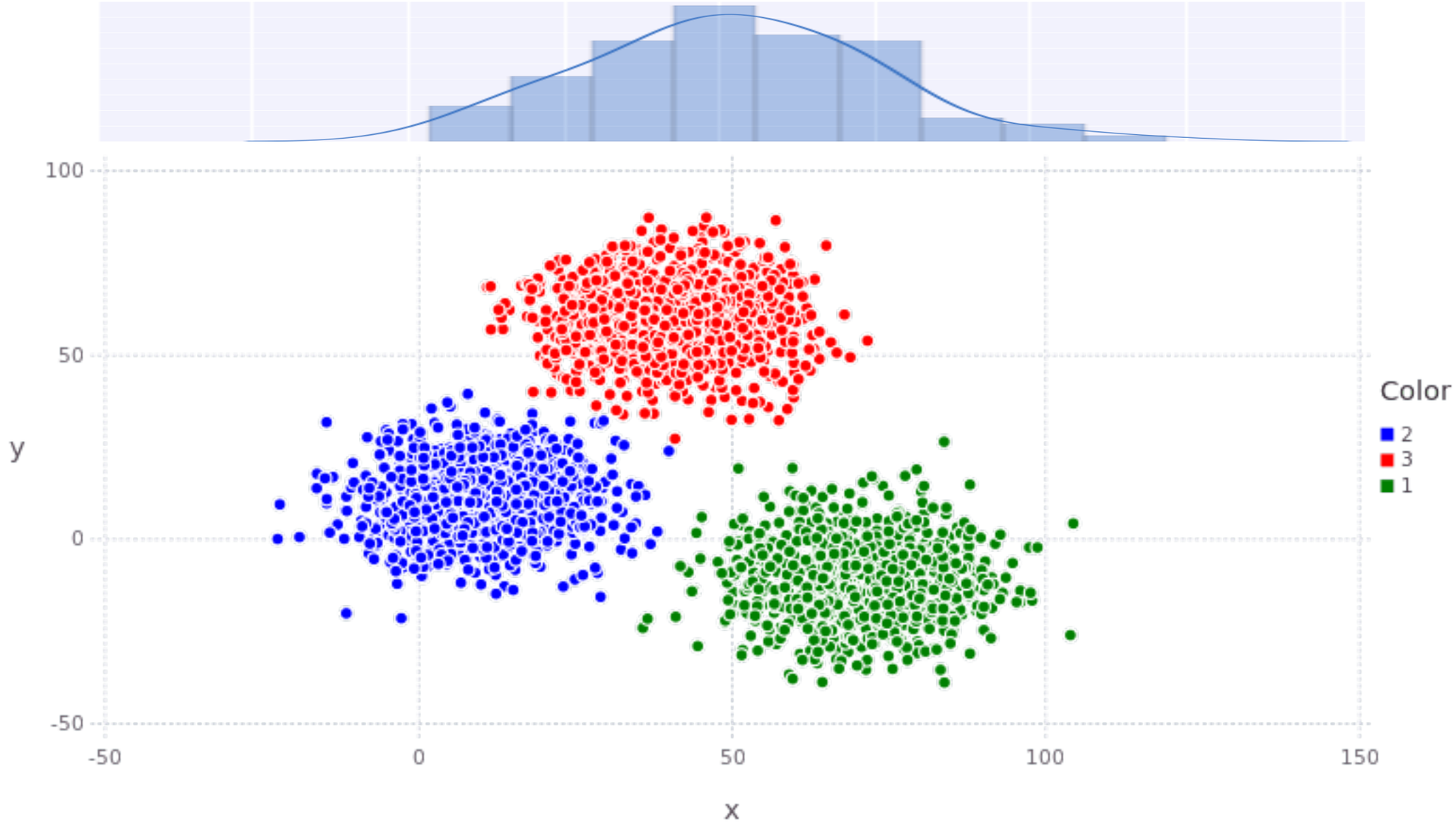
UNSUPERVISED TECHNIQUE: CLUSTERING



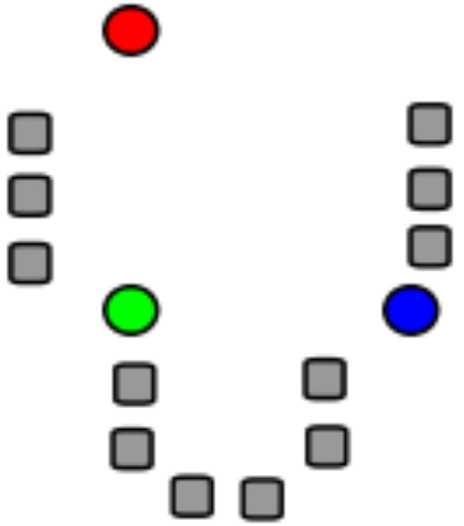
UNSUPERVISED TECHNIQUE: CLUSTERING



UNSUPERVISED TECHNIQUE: CLUSTERING



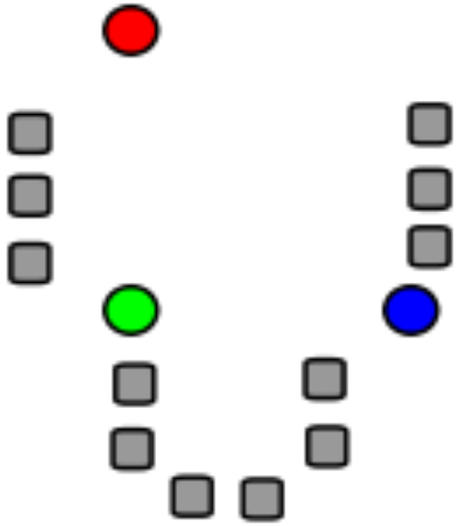
RECAP: K-MEANS CLUSTERING



RECAP: K-MEANS CLUSTERING

Steps

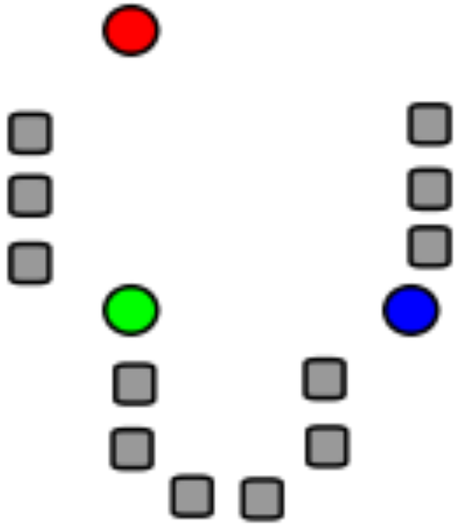
1. Choose k. Here, let's try to cluster into k=3 groups.
2. Choose 3 initial "centroids"...random points.
3. For each grey point:
    - find distance to nearest centroid
    - assign the point to the nearest centroid's "team"
4. Recalculate centroid positions to become the centre of each cluster
5. Repeat steps 3 and 4 until the centroids don't move



RECAP: K-MEANS CLUSTERING

Steps

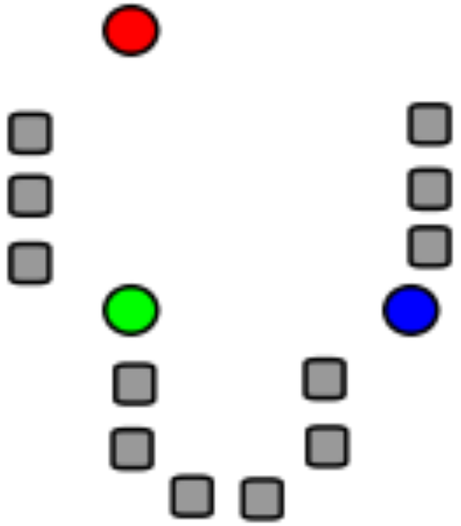
1. Choose k. Here, let's try to cluster into k=3 groups.
2. Choose 3 initial "centroids"...random points.
3. For each grey point:
    - find distance to nearest centroid
    - assign the point to the nearest centroid's "team"
4. Recalculate centroid positions to become the centre of each cluster
5. Repeat steps 3 and 4 until the centroids don't move



RECAP: K-MEANS CLUSTERING

Steps

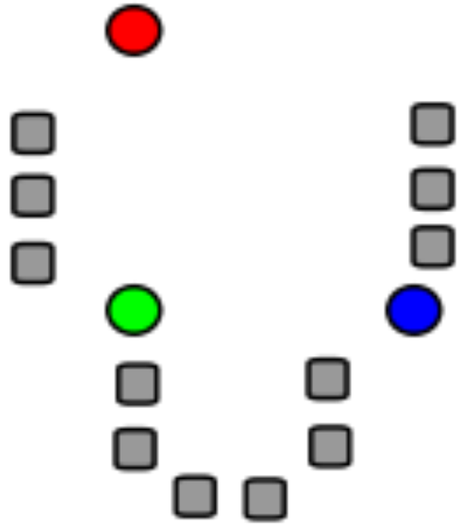
1. Choose k. Here, let's try to cluster into k=3 groups.
2. Choose 3 initial "centroids"...random points.
3. For each grey point:
    - find distance to nearest centroid
    - assign the point to the nearest centroid's "team"
4. Recalculate centroid positions to become the centre of each cluster
5. Repeat steps 3 and 4 until the centroids don't move



RECAP: K-MEANS CLUSTERING

Steps

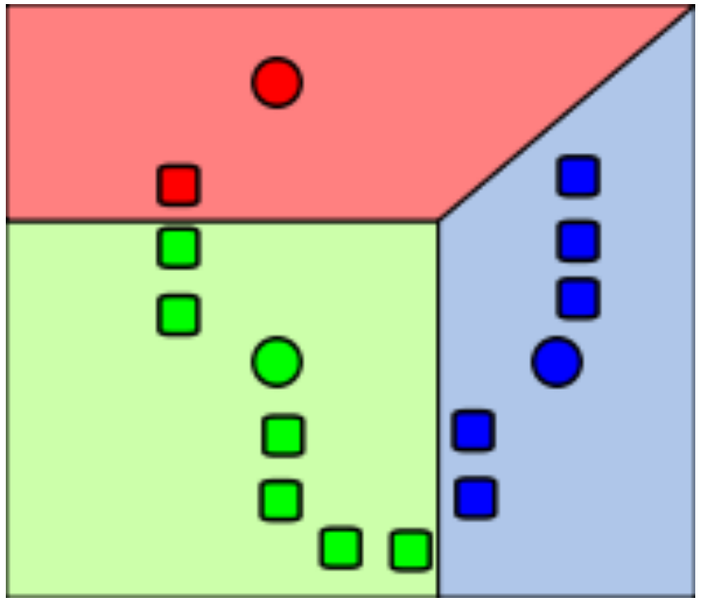
1. Choose k. Here, let's try to cluster into k=3 groups.
2. Choose 3 initial "centroids"...random points.
3. For each grey point:
    - find distance to nearest centroid
    - assign the point to the nearest centroid's "team"
4. Recalculate centroid positions to become the centre of each cluster
5. Repeat steps 3 and 4 until the centroids don't move



RECAP: K-MEANS CLUSTERING

Steps

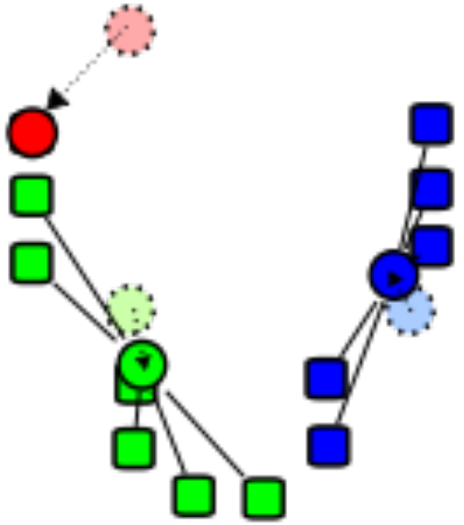
1. Choose k. Here, let's try to cluster into k=3 groups.
2. Choose 3 initial "centroids"...random points.
3. For each grey point:
    - find distance to nearest centroid
    - assign the point to the nearest centroid's "team"
4. Recalculate centroid positions to become the centre of each cluster
5. Repeat steps 3 and 4 until the centroids don't move



RECAP: K-MEANS CLUSTERING

Steps

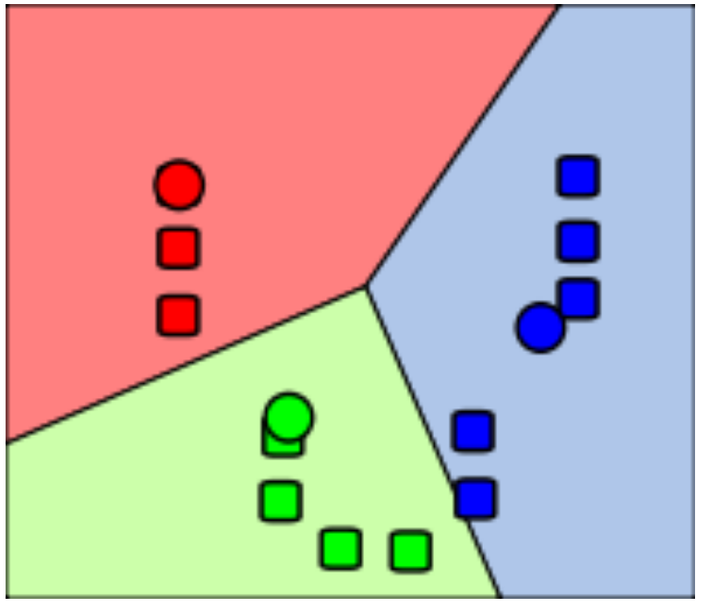
1. Choose k. Here, let's try to cluster into k=3 groups.
2. Choose 3 initial "centroids"...random points.
3. For each grey point:
    - find distance to nearest centroid
    - assign the point to the nearest centroid's "team"
4. Recalculate centroid positions to become the centre of each cluster
5. Repeat steps 3 and 4 until the centroids don't move



RECAP: K-MEANS CLUSTERING

Steps

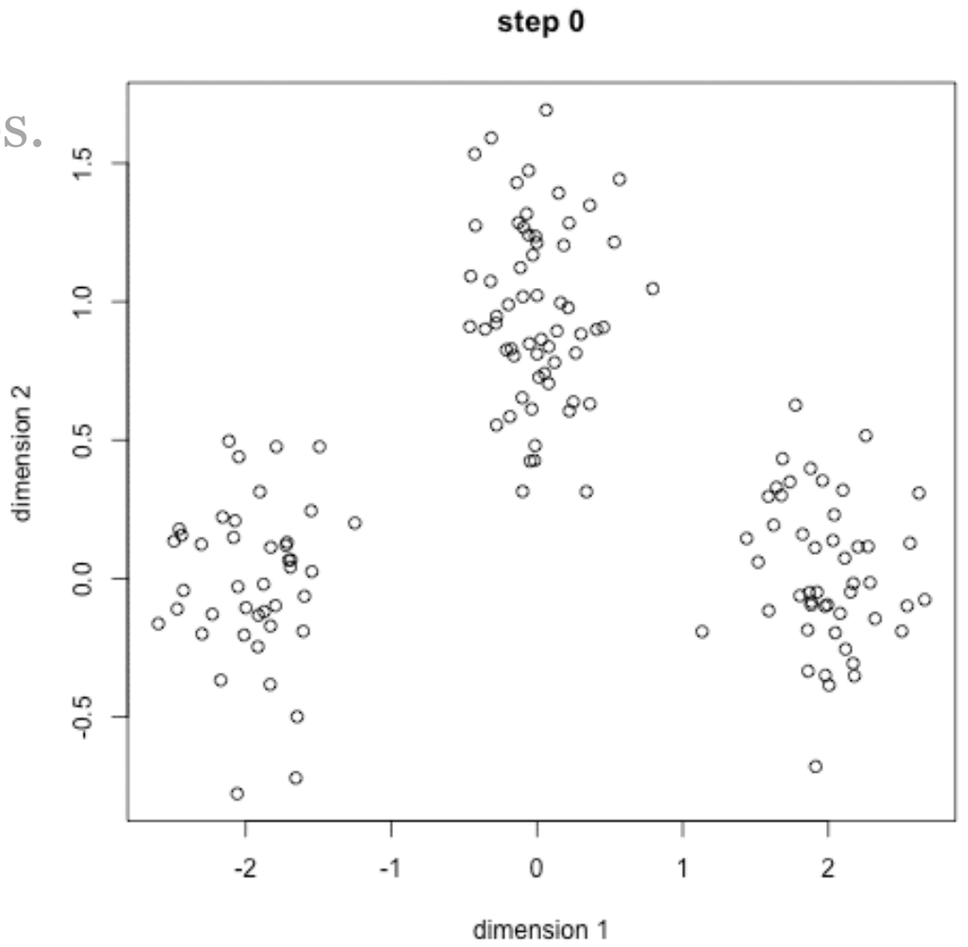
1. Choose k. Here, let's try to cluster into k=3 groups
2. Choose 3 initial "centroids"...random points.
3. For each grey point:
    - find distance to nearest centroid
    - assign the point to the nearest centroid's "team"
4. Recalculate centroid positions to become the centre of each cluster
5. Repeat steps 3 and 4 until the centroids don't move



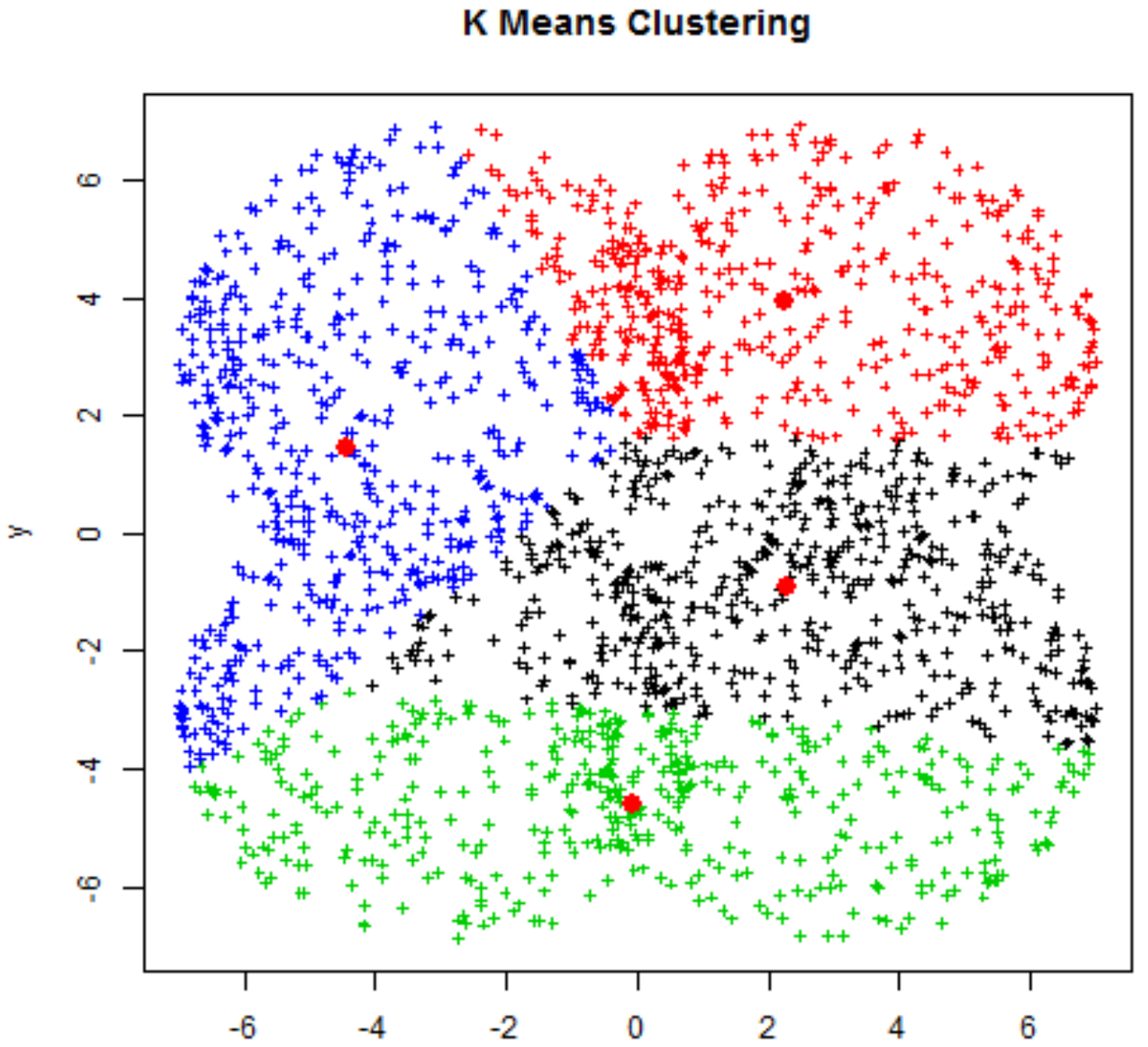
RECAP: K-MEANS CLUSTERING

Steps

1. Choose k. Here, let's try to cluster into k=3 groups.
2. Choose 3 initial "centroids"...random points.
3. For each grey point:
    - find distance to nearest centroid
    - assign the point to the nearest centroid's "team"
4. Recalculate centroid positions to become the centre of each cluster
5. Repeat steps 3 and 4 until the centroids don't move



RECAP: K-MEANS CLUSTERING



RECAP: DISTANCE METRICS

K-Means requires the concept of "distance" between two data points. Let's consider a few different metrics.

- **Euclidean distance...**
  - ...between 1 and 2 is 1
  - ...between point (1,1) and point (2,2) is √2
  - ...this works for any "vector" of numbers

$$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

RECAP: DISTANCE METRICS

K-Means requires the concept of "distance" between two data points. Let's consider a few different metrics.

- **Euclidean distance...**
  - ...between 1 and 2 is 1
  - ...between point (1,1) and point (2,2) is √2
  - ...this works for any "vector" of numbers

$$d(x, y) = \sqrt{\sum(x_i - y_i)^2}$$

- What about the distance between two tweets?

DISTANCE METRICS

K-Means requires the concept of "distance" between two data points. Let's consider a few different metrics.

- **Jaccard distance...**
  - ...between "rock music" and "classical music" is "one word shared out of three words total" or 1/3 = 0.33

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} :$$

  - The same concept can determine distances between words, tweets, books...even the Jaccard distance between actors in two movies

CASE STUDY

BEHAVIOURAL TARGETING

EXAMPLE: BEHAVIOURAL TARGETING

- **Problem:** Can we find smarter customer segmentations than RFM/lifecycle models?

- **Goals:**
  - Build a dataset of indicators on a per customer basis
    - Demographic indicators (e.g. gender, age, location, device type)
    - Acquisition channel indicators (e.g. organic, paid, CAC)
    - Behavioural indicators (e.g. RFM metrics, genre/topics of interest)
    - Social indicators (e.g. number of friends using service)
  - Build a clustering model using this dataset that finds the best "distinct groups"
  - Interpret what each group represents
  - Can we use these groupings for targeted newsletter or social media campaigns, and other retention activities?

EXAMPLE: BEHAVIOURAL TARGETING

- **Results:**
  - Six groupings emerged, upon investigation we had groups such as:
    - Users who open the app almost every day but never book anything
    - Users who only book free events
    - Users who book a "date night" style event once a month
    - Users who book a headline high value item once a year
    - And so on…
  - Subgroups within main groups, e.g.
    - Users who go to free lunchtime concerts
  - The clustering algorithm can automatically assign new users into one of these categories.

- **How to validate if this segmentation is better than RFM/lifecycle models?**

DATA SCIENCE == SCIENCE

DATA + SCIENCE = DATA SCIENCE

- Two options for user segmentation for newsletter campaigns:
  - **Old method:** Most viewed category (e.g. sport, music)
  - **New method:** Algorithmically generated clusters

- **Question:** How to know which is better?

- **Answer:** Run an experiment!

- **Hypothesis:** "New method is significantly better"

- **Question:** What do we mean by "better"?
  - Email campaign opens? CTR? CTR + purchase? Long-term retention metrics?

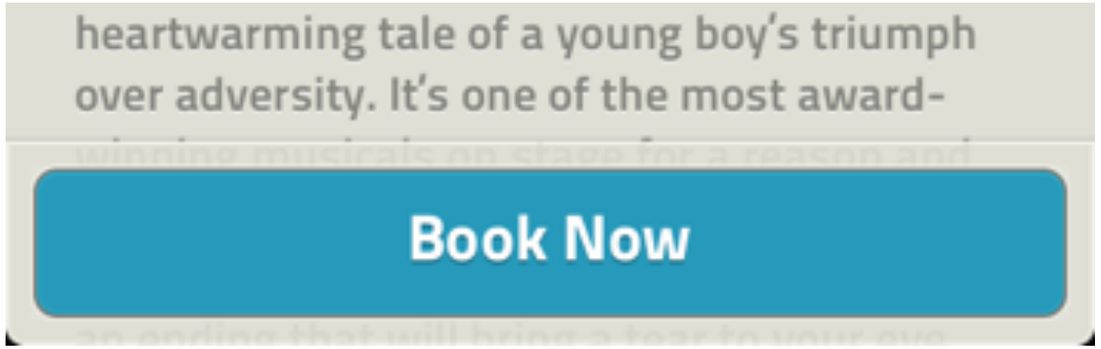
DATA + SCIENCE = DATA SCIENCE

- Let's go with "the campaign that generated more revenue". Difficult to argue with that.

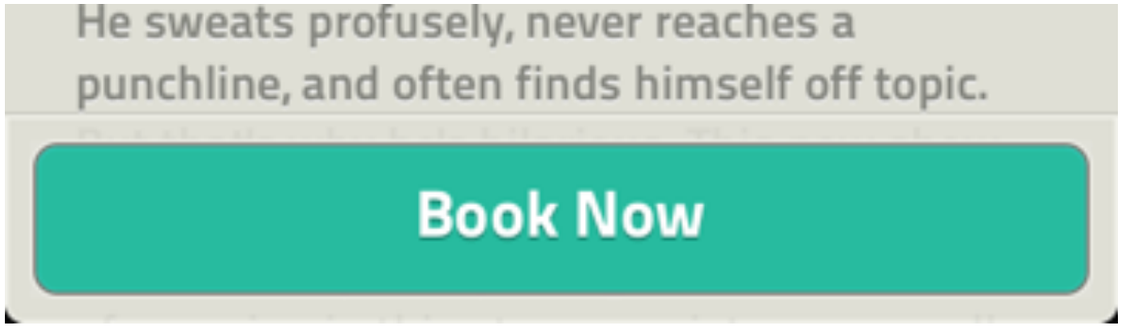
- Run the test:
  - Divide all customers into two groups at random
  - A gets old segmentation
  - B gets new segmentation
  - ...wait...
  - Crunch numbers once sample size is reached

- Welcome to the wonderful world of A/B testing!

A/B TESTING



vs



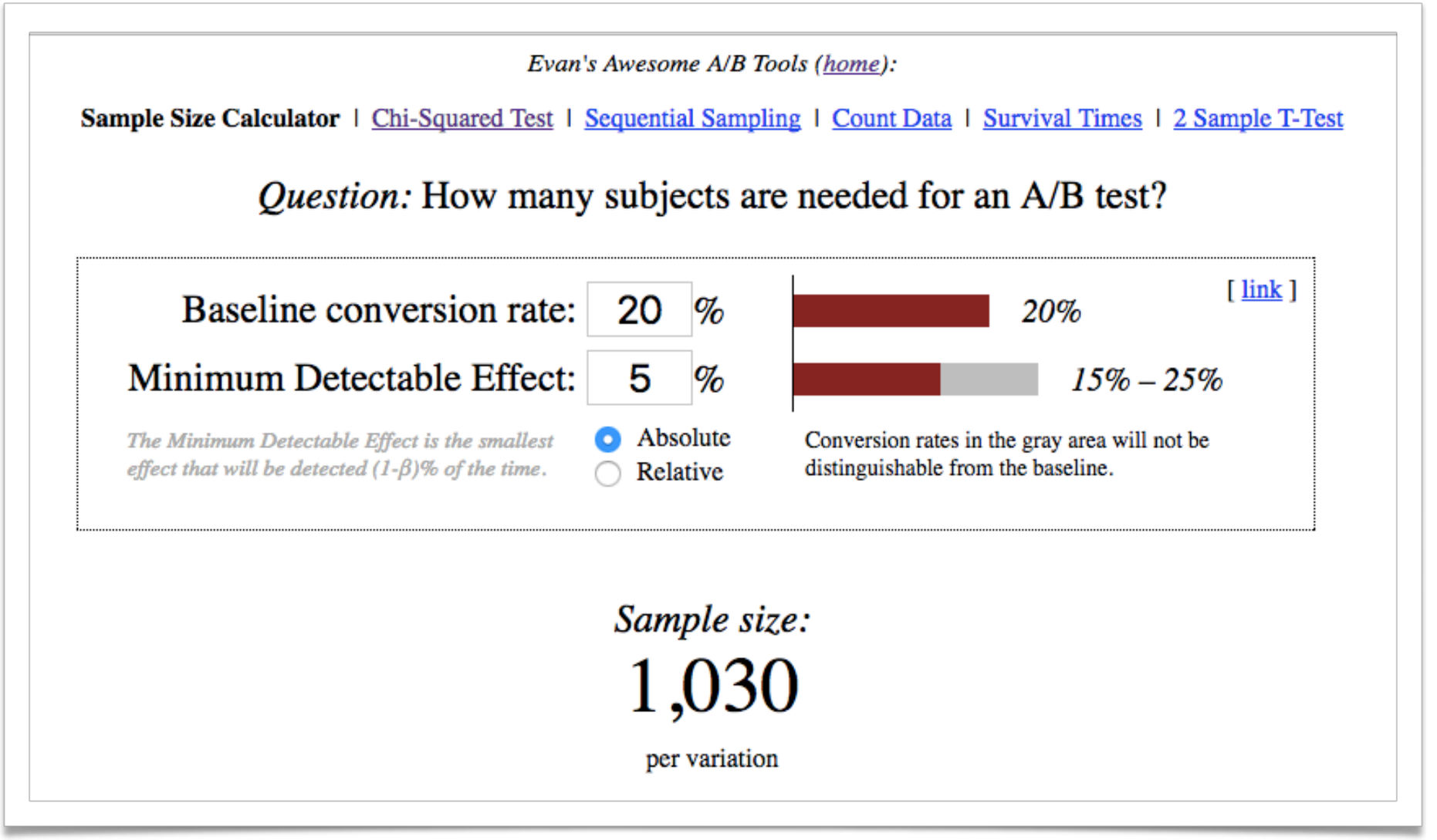
HOW NOT TO RUN AN A/B TEST

Statistical **significance** is for scientists!

Don’t bother calculating a s **sample size**

End the A/B test **early**

TOP TIP: EVANMILLER.ORG



Evan's Awesome A/B Tools (home):

Sample Size Calculator | Chi-Squared Test | Sequential Sampling | Count Data | Survival Times | 2 Sample T-Test

Question: How many subjects are needed for an A/B test?

Baseline conversion rate: 20 %

Minimum Detectable Effect: 5 %

The Minimum Detectable Effect is the smallest effect that will be detected (1-β)% of the time.

(•) Absolute
( ) Relative

[link]

20%

15% – 25%

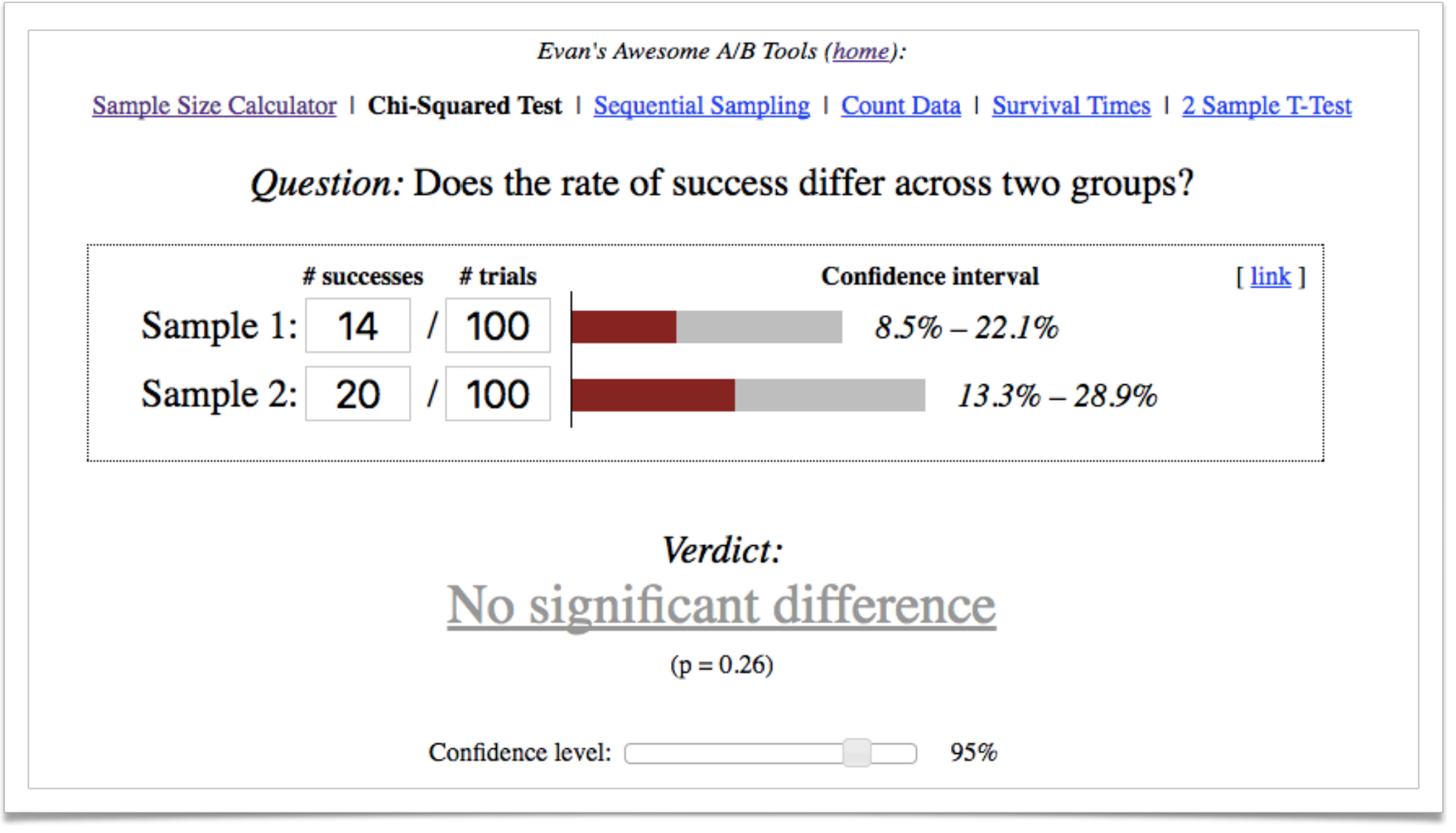
Conversion rates in the gray area will not be distinguishable from the baseline.

Sample size:

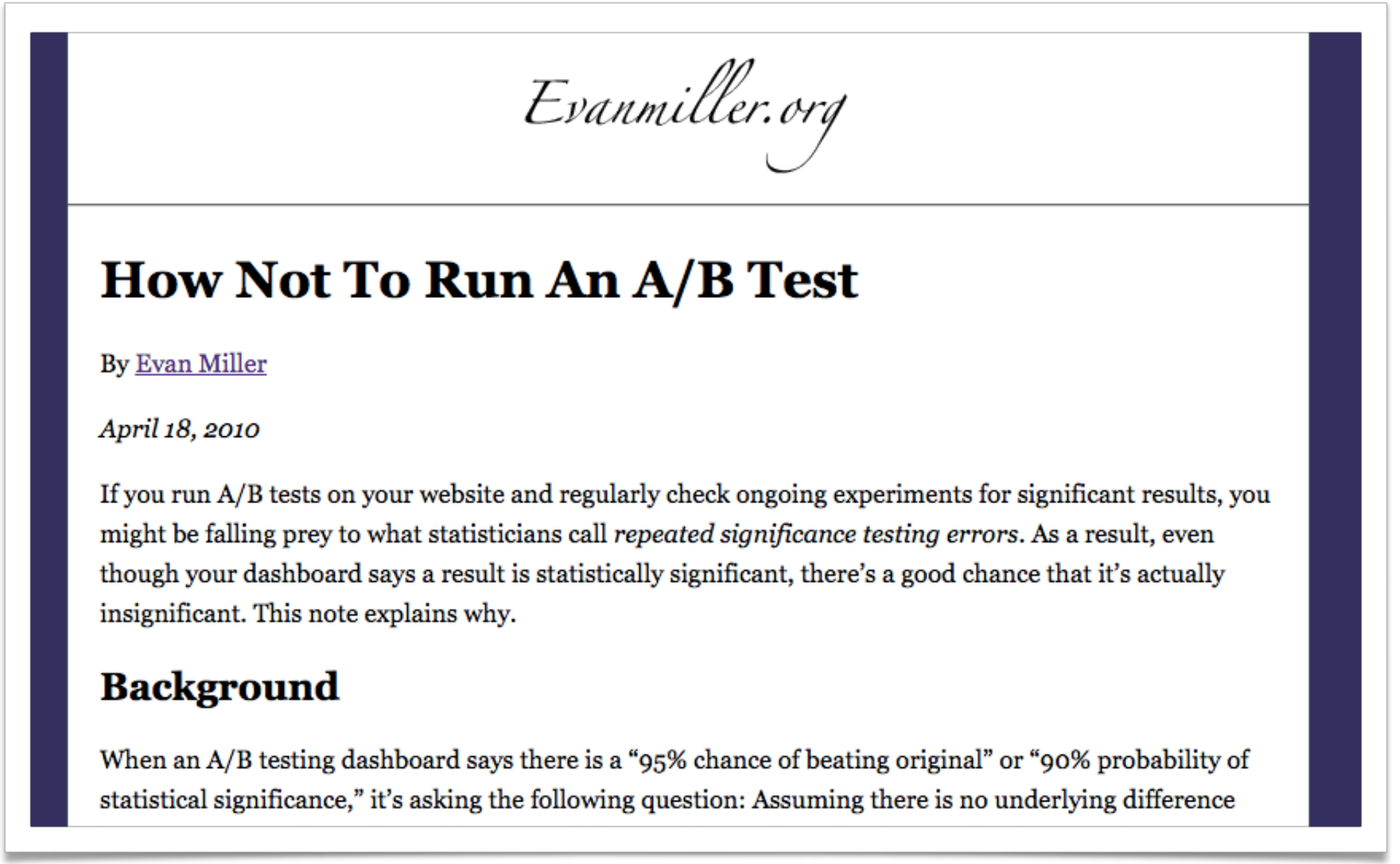
1,030

per variation

TOP TIP: EVANMILLER.ORG



TOP TIP: EVANMILLER.ORG



Evanmiller.org

How Not To Run An A/B Test

By Evan Miller

April 18, 2010

If you run A/B tests on your website and regularly check ongoing experiments for significant results, you might be falling prey to what statisticians call *repeated significance testing errors*. As a result, even though your dashboard says a result is statistically significant, there's a good chance that it's actually insignificant. This note explains why.

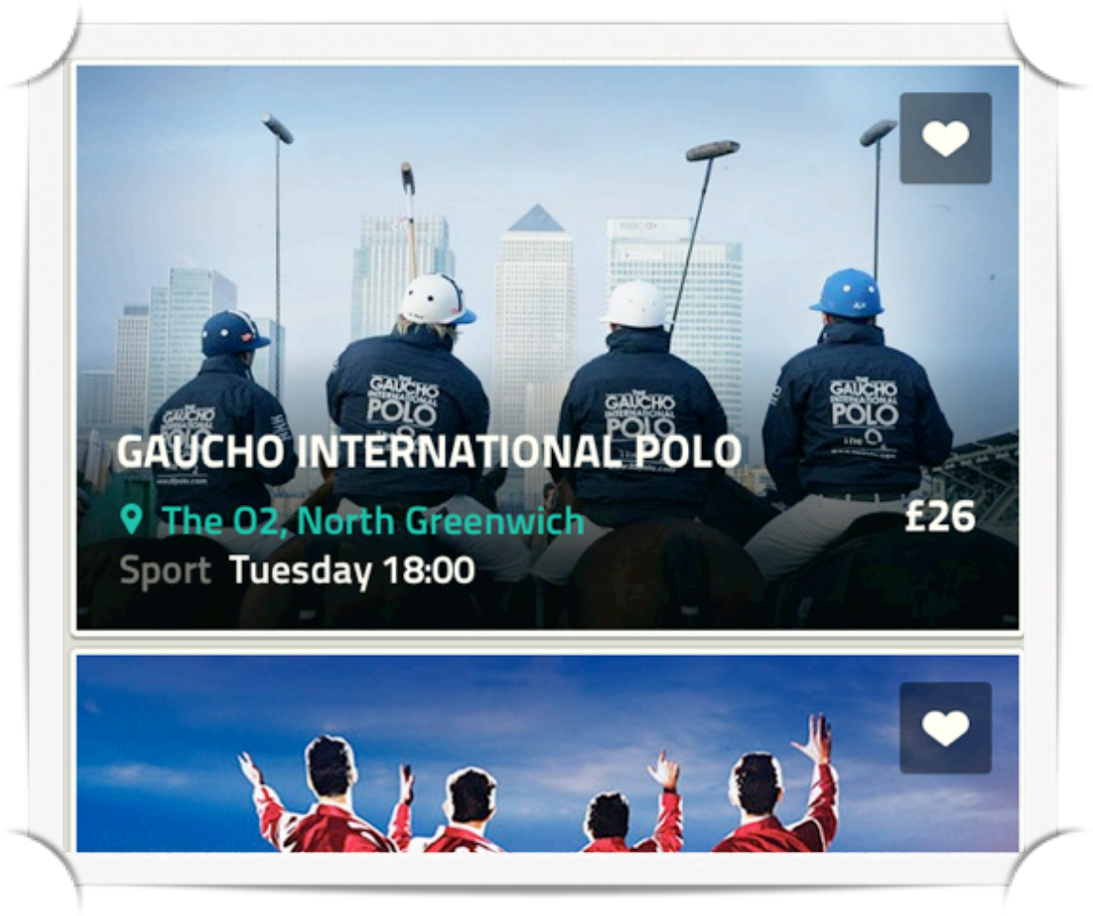
Background

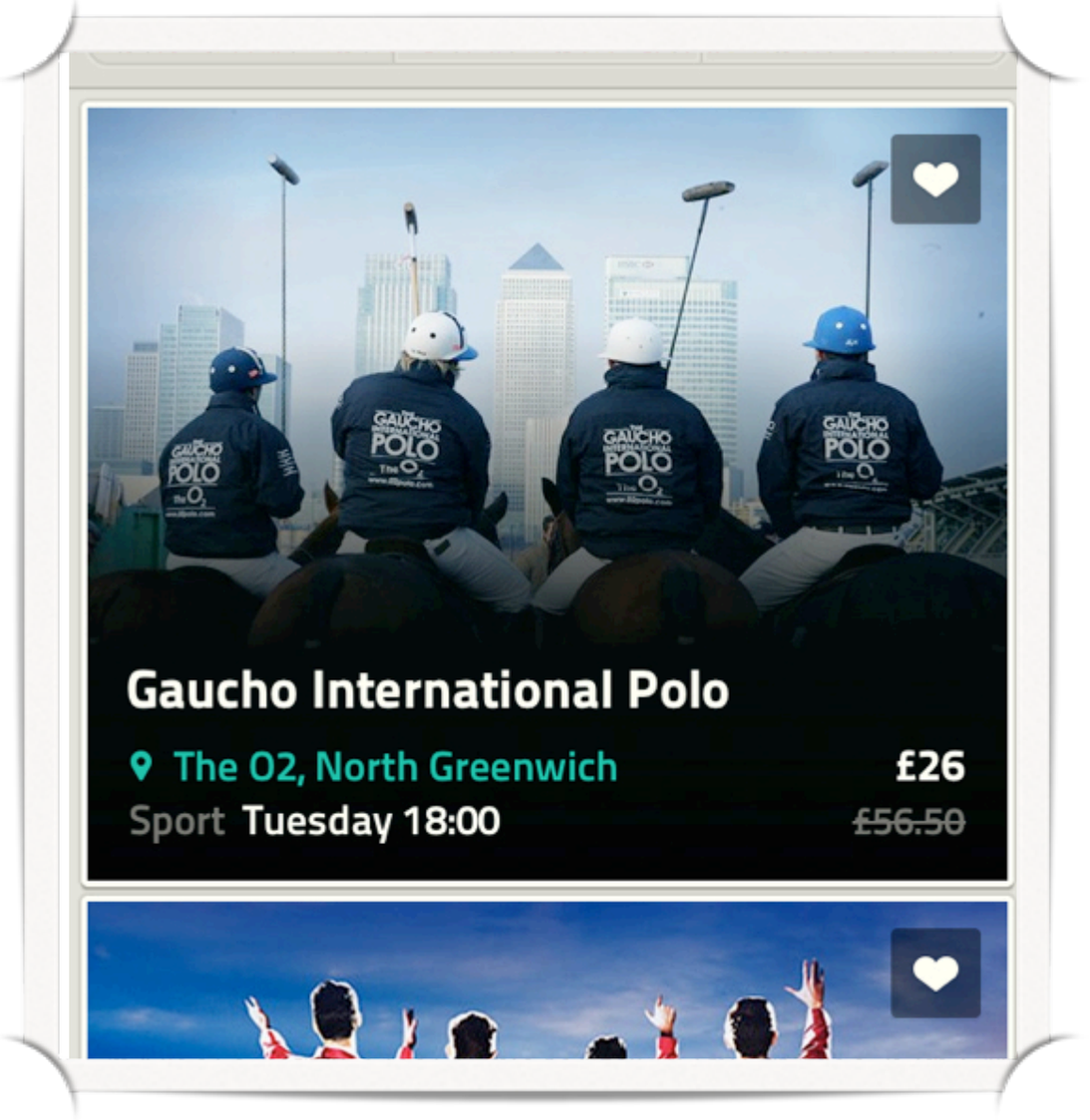
When an A/B testing dashboard says there is a "95% chance of beating original" or "90% probability of statistical significance," it's asking the following question: Assuming there is no underlying difference

A/B TESTING = GROWTH

2048 different versions of the app

A/B TESTING = GROWTH





A/B TESTING = GROWTH

2048 different versions of the app

Fearless experimentation!

A/B TESTING = GROWTH

A/B TESTING = GROWTH

2048 different versions of the app

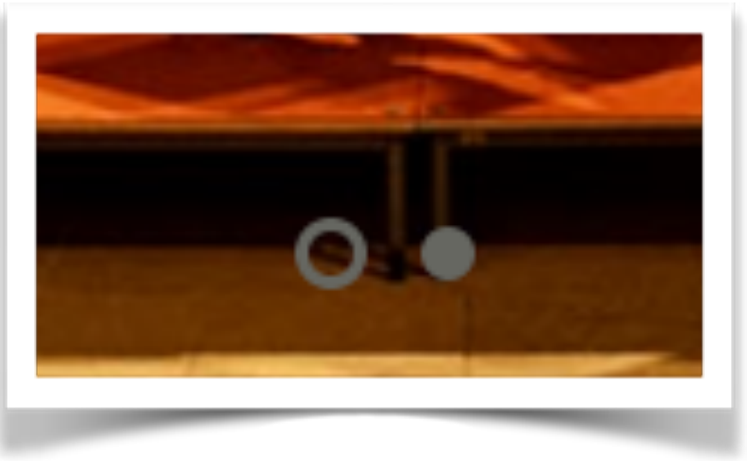
Fearless experimentation!

Not significant doesn’t mean **boring**

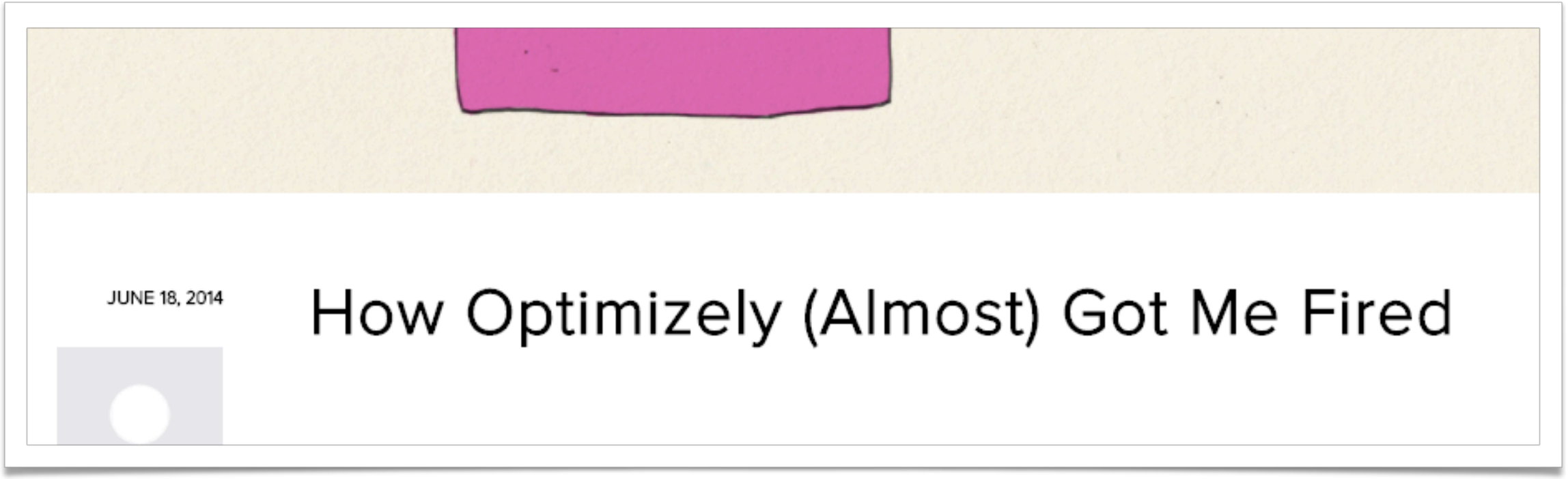
A/B TESTING = GROWTH

4 images

2 images



A WORD OF WARNING



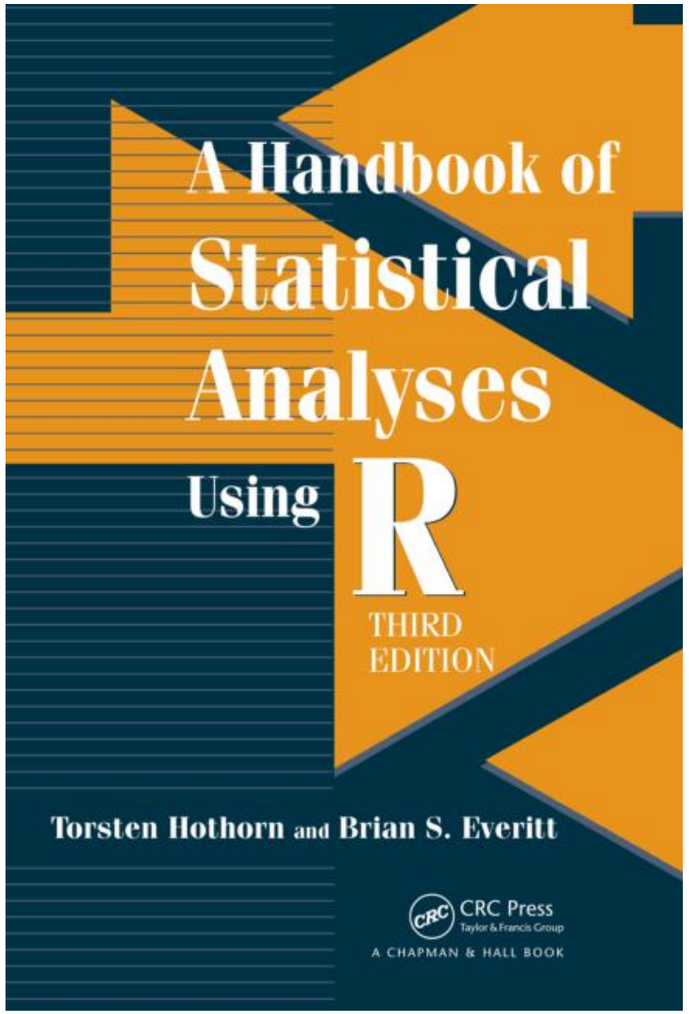
http://blog.sumall.com/journal/optimizely-got-me-fired.html

INFERENCE TESTS

CHI-SQUARED

	Group A	Group B
Viewed signup button	100	100
Clicked signup button	20	25

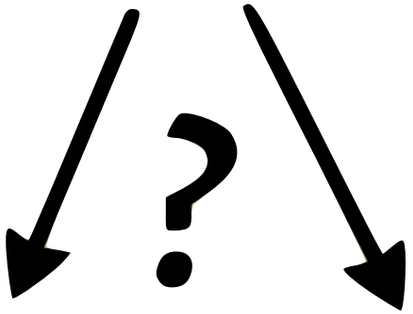
HANDBOOK OF STATISTICAL ANALYSES USING R



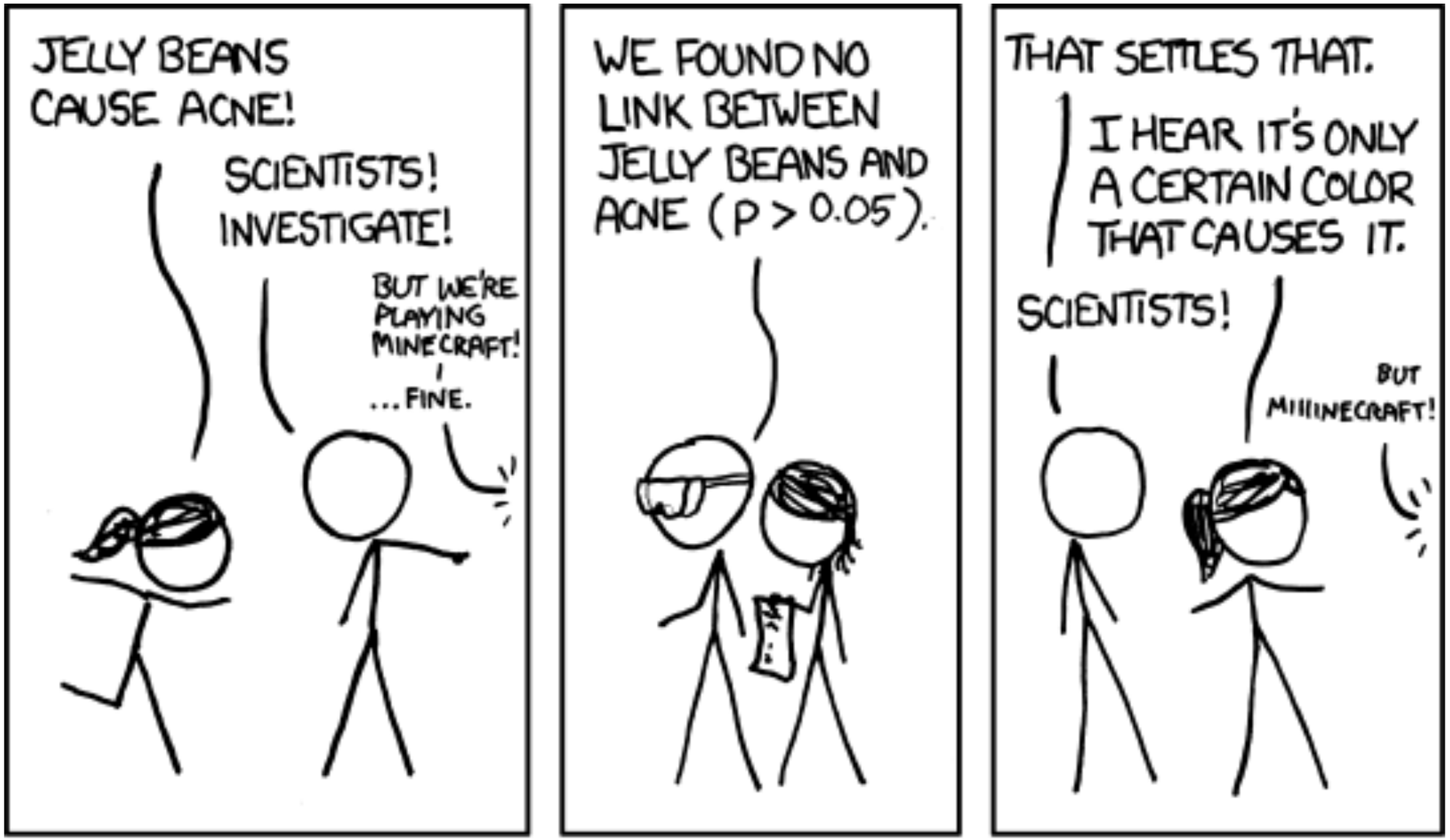
- **2nd edition is freely available online as a PDF**
- **#1 podcast recommendation:**
  - Data Skeptic

P HACKING

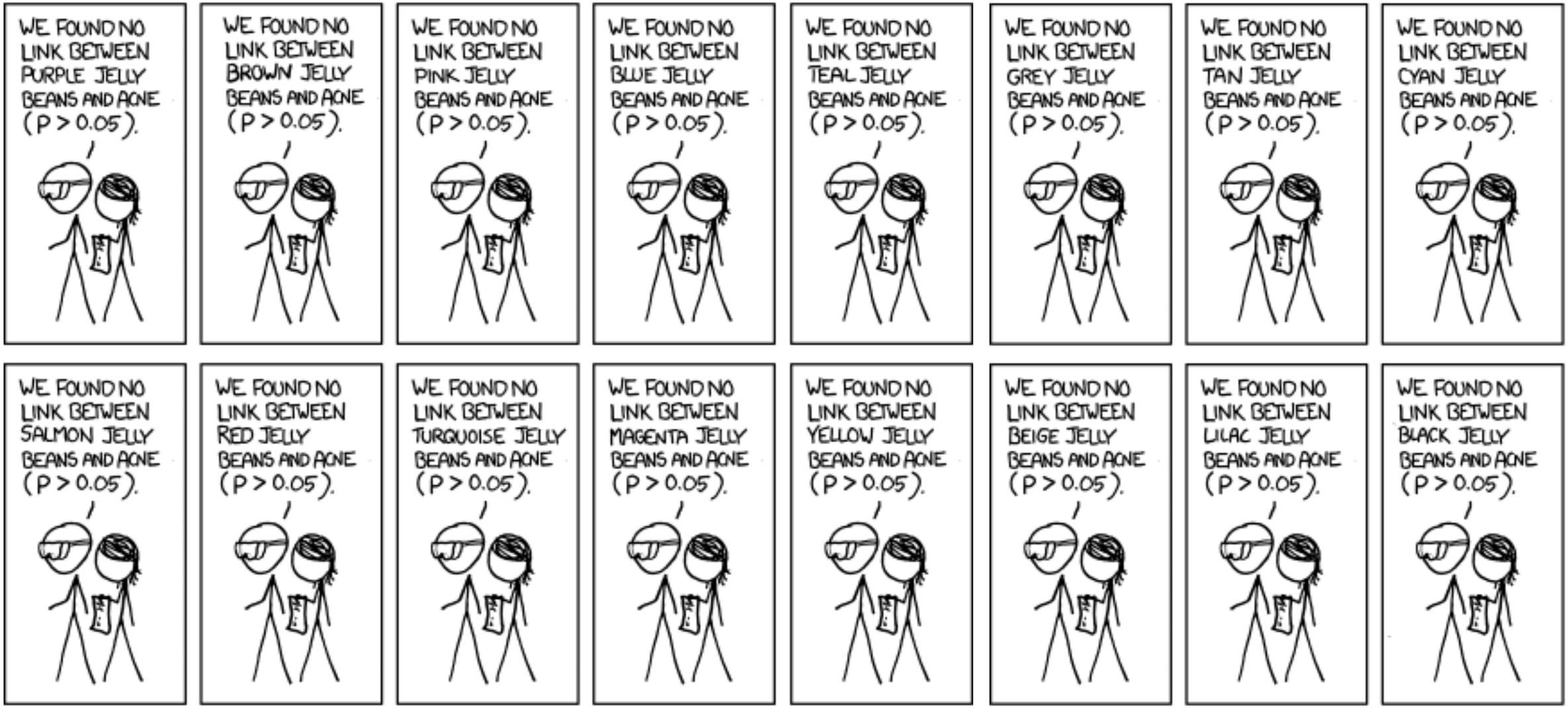
HYPOTHESIS TESTING AND P-VALUES



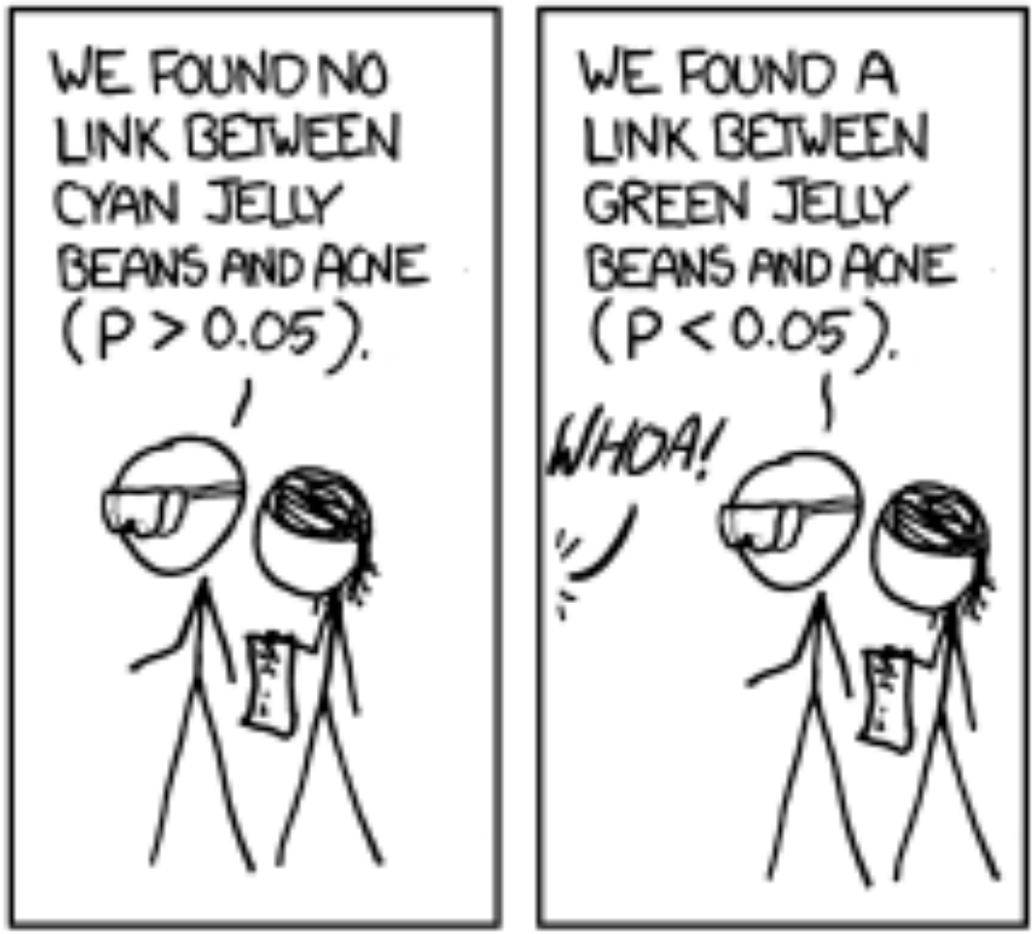
BEWARE P-HACKING

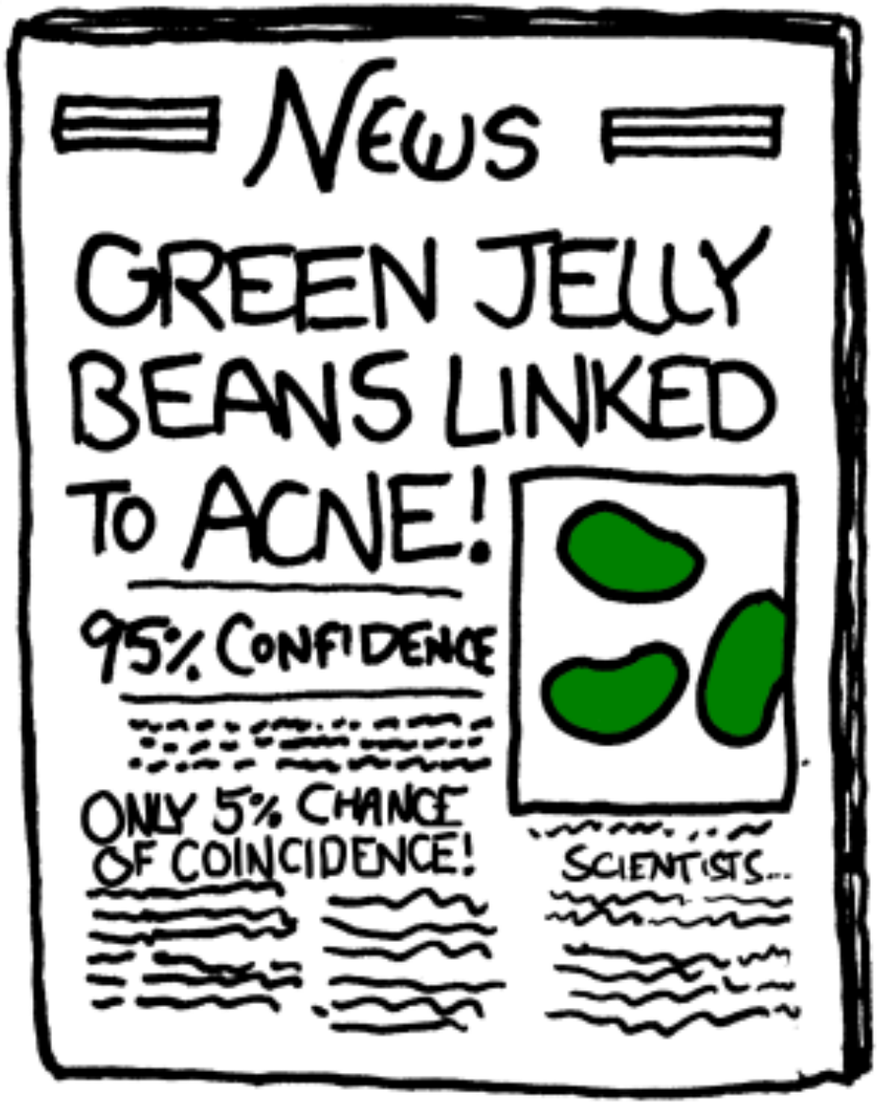


BEWARE P-HACKING

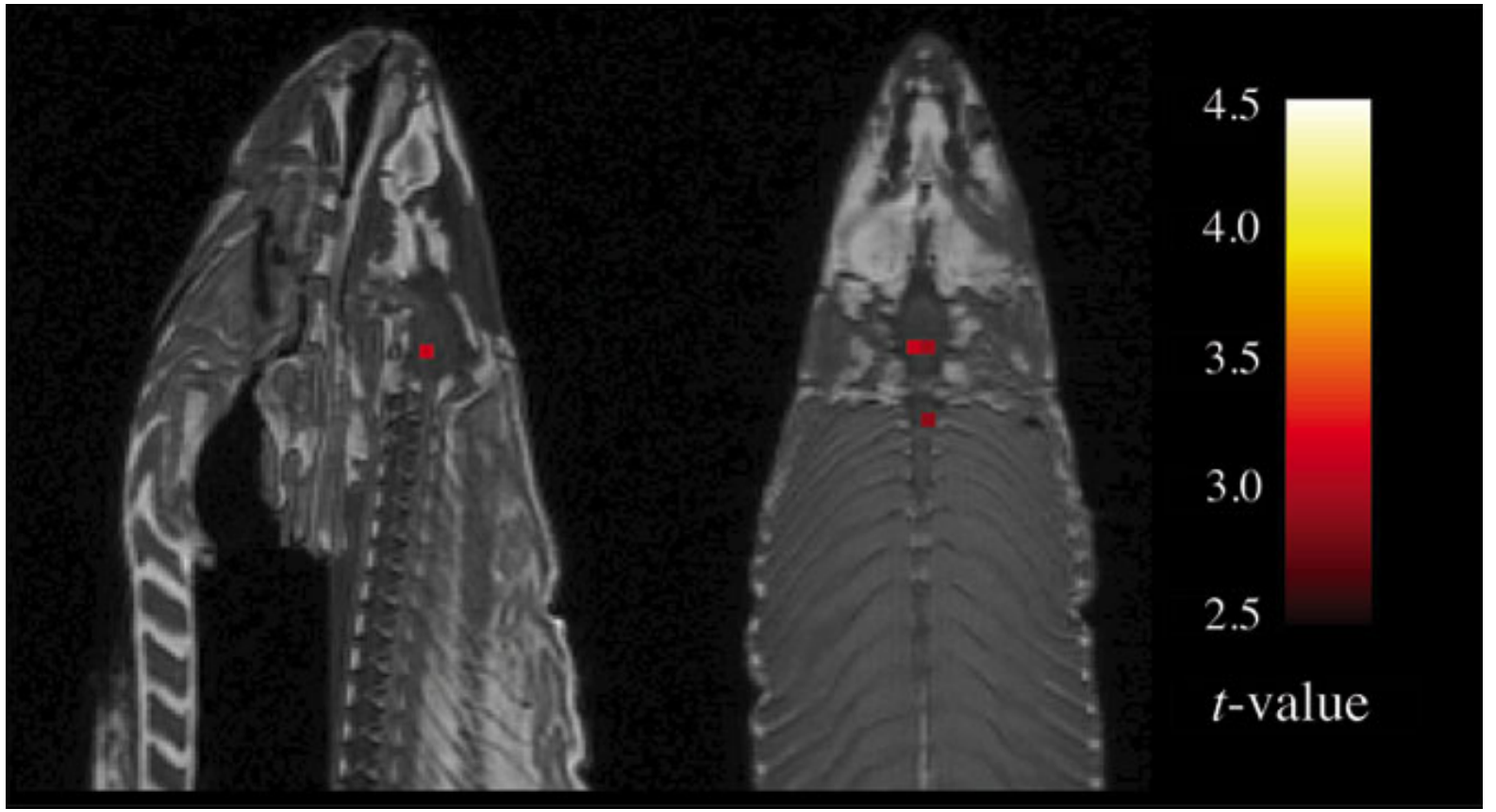


BEWARE P-HACKING





CASE STUDY: THE DEAD SALMON STUDY



*Bennett et al. "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: **An Argument For Proper Multiple Comparisons Correction**". Journal of Serendipitous and Unexpected Results, 2010. IgNobel Award in Neuroscience 2012.*

CASE STUDY: THE DEAD SALMON STUDY

LIVESCIENCE

NEWS TECH HEALTH PLANET EARTH SPACE

Live Science > Strange News

Dead Salmon 'Responds' to Pictures of People

By Robert Roy Britt | September 27, 2009 07:03am ET

A dead salmon has become a scientific celebrity after its brain supposedly lit up when shown pictures of humans during a brain scan.

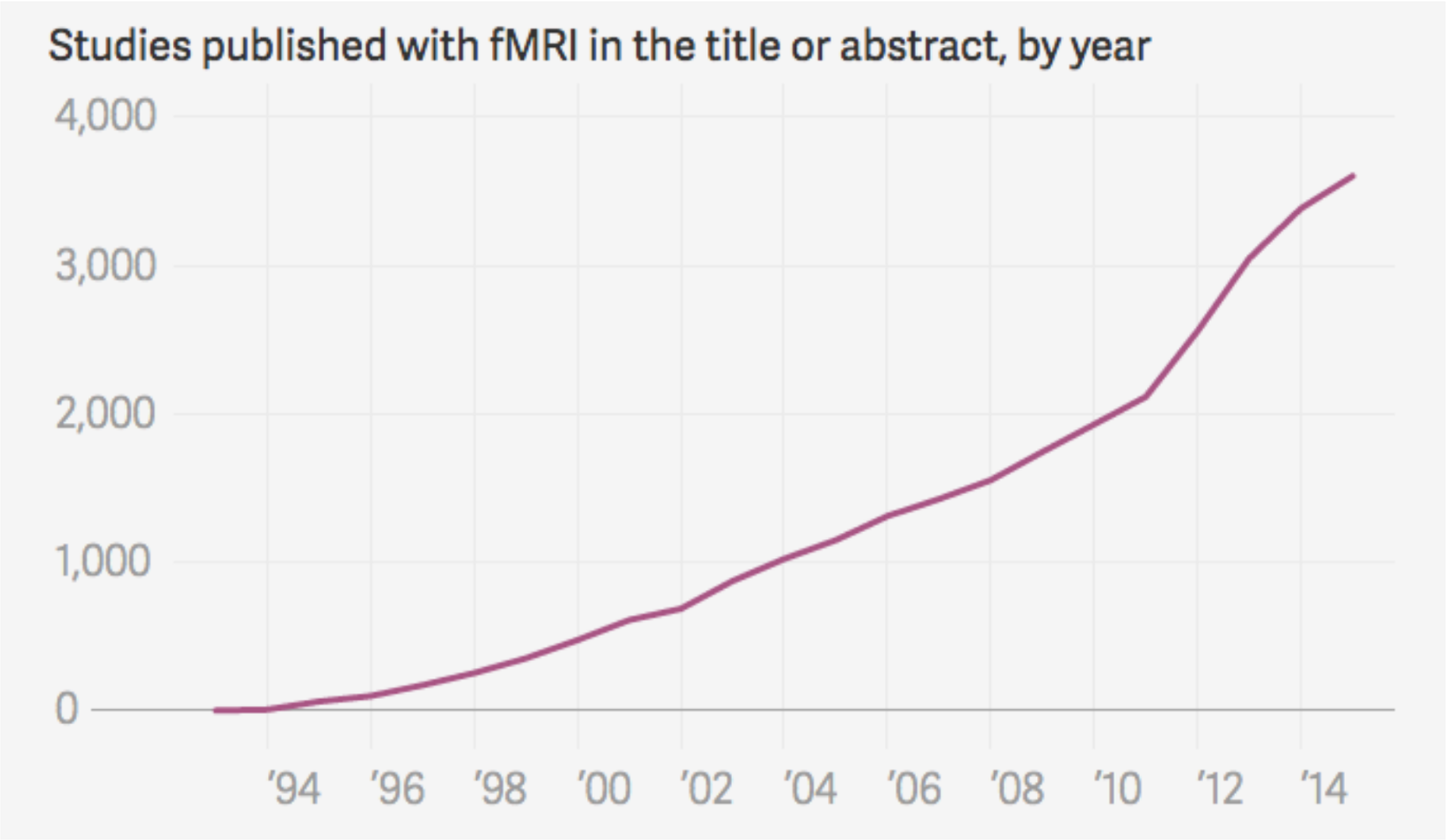
CASE STUDY: THE DEAD SALMON STUDY

WIRED SUBSCRIBE

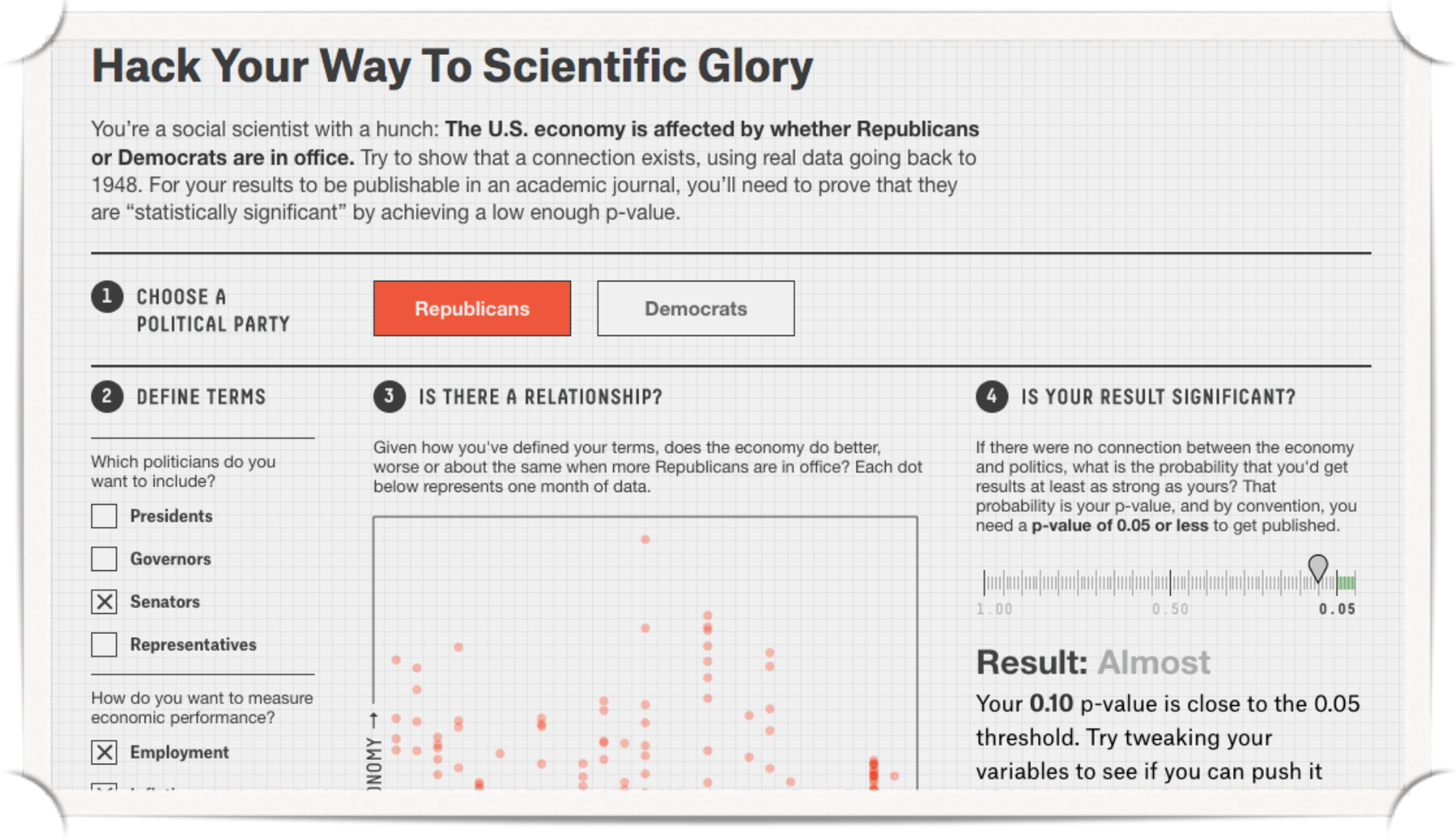
ALEXIS MADRIGAL SCIENCE 09.18.09 5:37 PM

SCANNING DEAD SALMON IN FMRI MACHINE HIGHLIGHTS RISK OF RED HERRINGS

CASE STUDY: THE DEAD SALMON STUDY

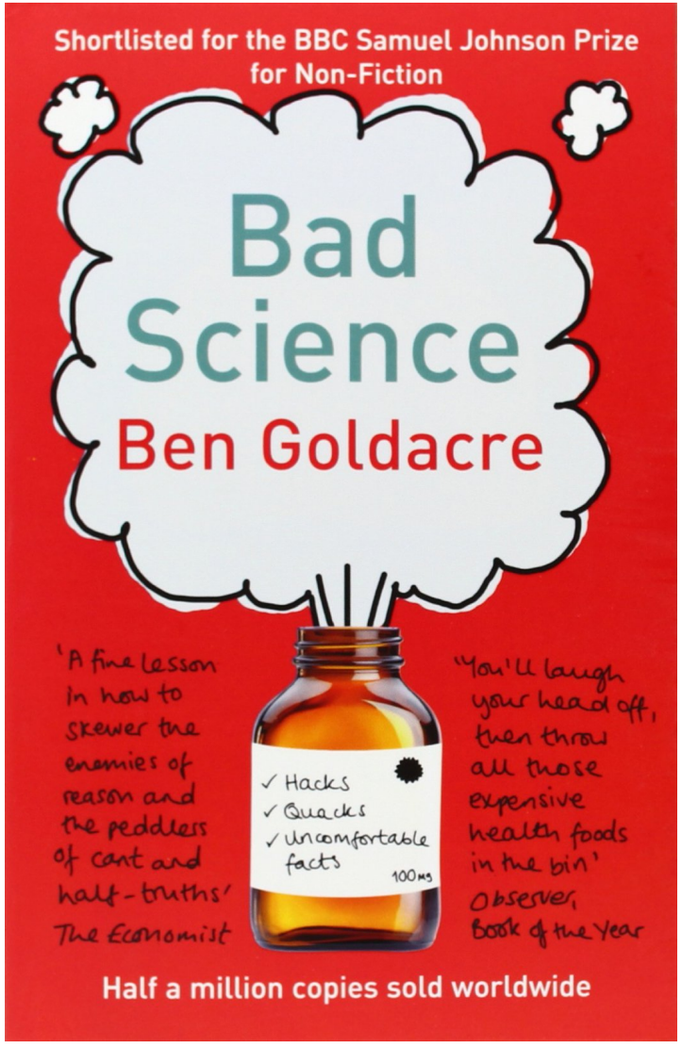


P-HACKING: TRY IT OUT!

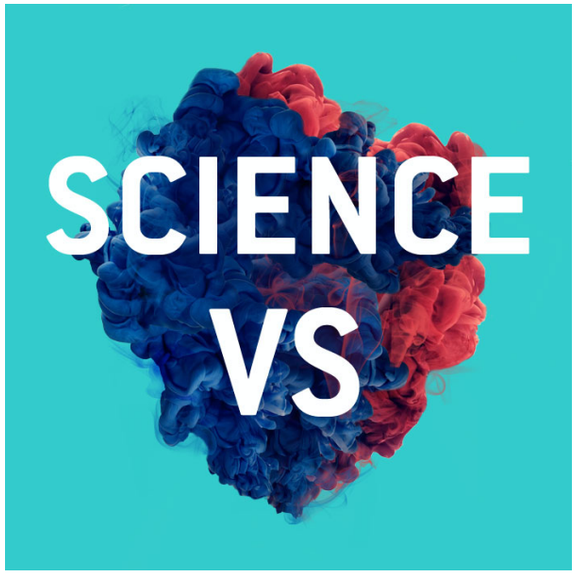


projects.fivethirtyeight.com/p-hacking/

RECOMMENDED READING



Podcasts



CASE STUDY

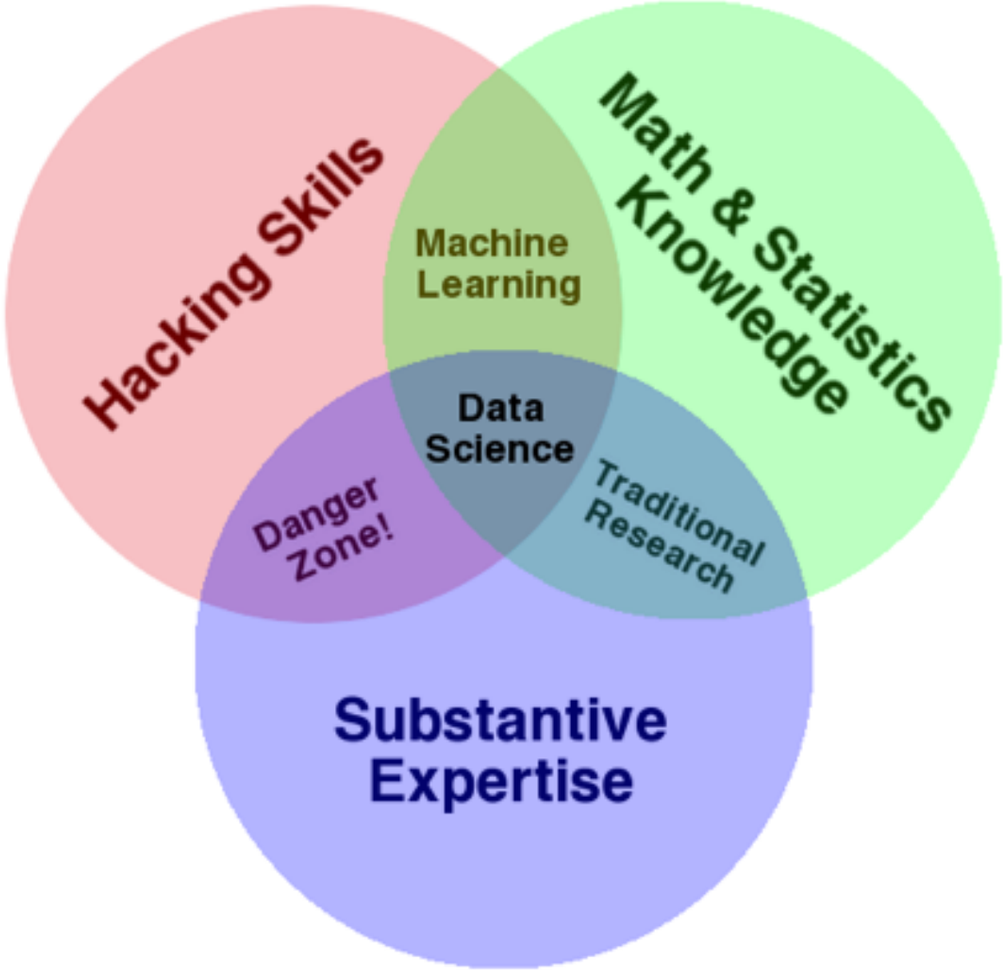
BUILDING YPLAN'S RECOMMENDATION ENGINE

CLOSING

DATA SCIENCE VS DATA ANALYTICS

"Data Analytics"	"Data Science"
Historical reporting.	Predictive forecasting.
Metrics. KPIs. Segmentation.	Statistics. Regression. Machine learning.
Dashboards. BI tools. Pivot tables.	Coding. Flexibility. Automation.
Necessary...keeps the engines running.	Exciting...unexpected insights.
Tools: Excel, SQL, Tableau.	Tools: Python, R, scikit-learn.

DATA ANALYST...OR DATA SCIENTIST?



source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

WHAT IS DATA SCIENCE FOR ME?

"Data Analytics"	"Data Science"
Historical reporting.	Predictive forecasting.
Metrics. KPIs. Segmentation.	Statistics. Regression. Machine learning.
Dashboards. BI tools. Pivot tables.	Coding. Flexibility. Automation.
Necessary...keeps the engines running.	Exciting...unexpected insights.
Tools: Excel, SQL, Tableau.	Tools: Python, R, scikit-learn.

WHAT IS DATA SCIENCE FOR ME?

"Data Analytics"

Historical reporting.
Metrics. KPIs. Segmentation.
Dashboards. BI tools. Pivot tables.
Necessary...keeps the engines running.
Tools: Excel, SQL, Tableau.

"Data Science"

Predictive forecasting.
Statistics. Regression. Machine learning.
Coding. Flexibility. Automation.
Exciting...unexpected insights.
Tools: Python, R, scikit-learn.

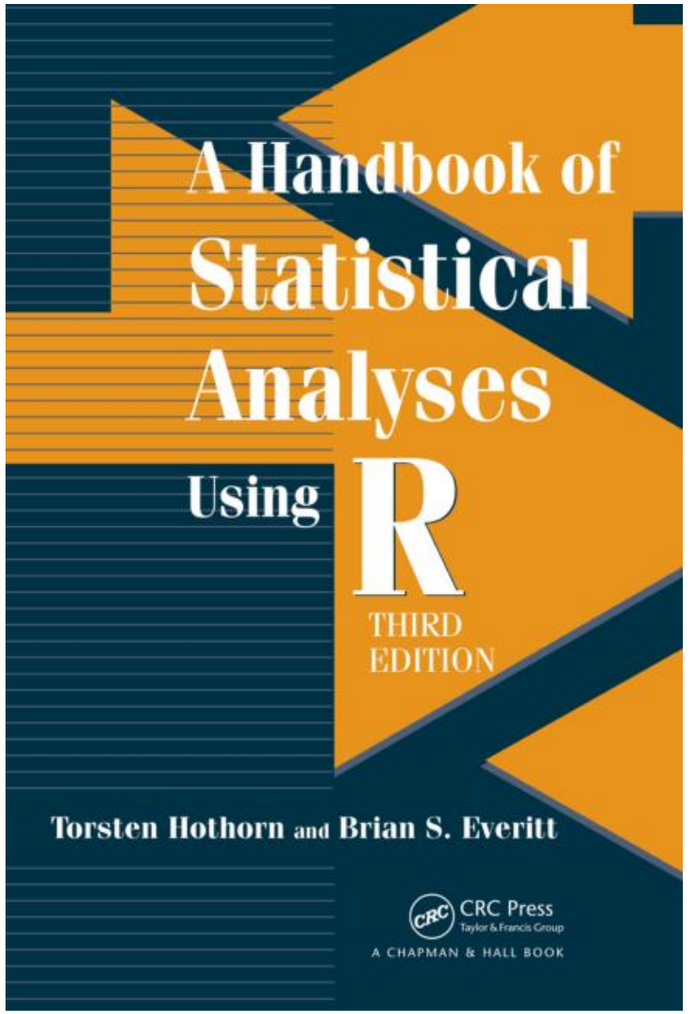
"Data Engineering"

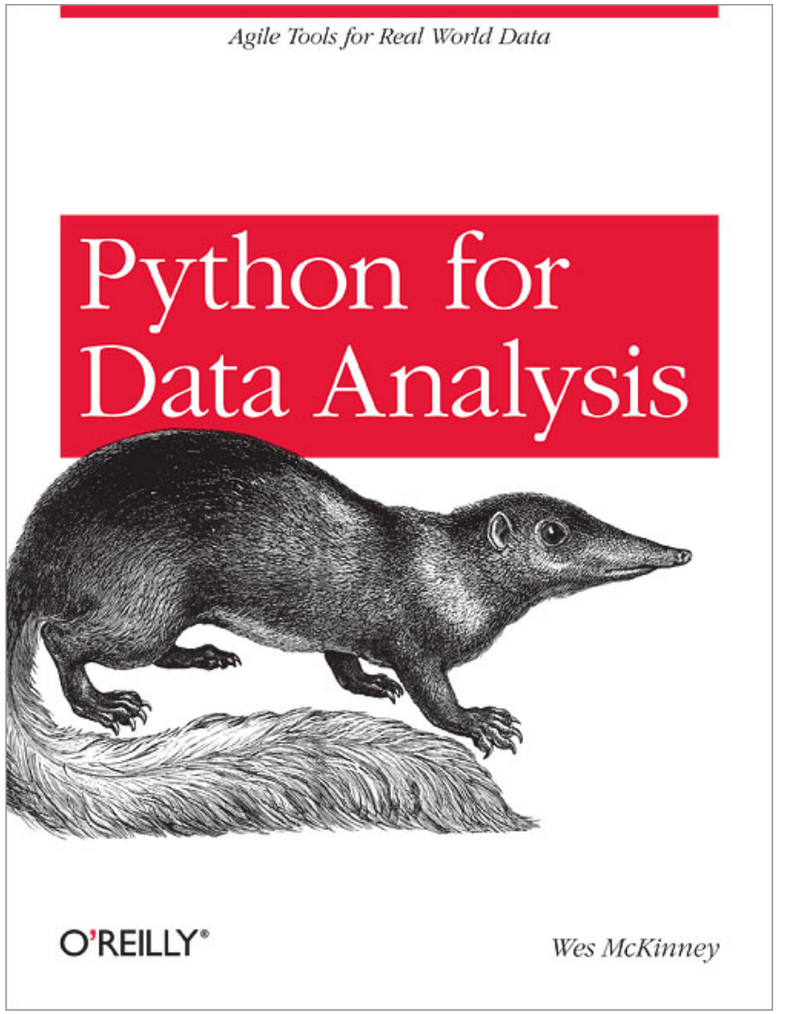
Architecture. Devops. Cloud solutions.
Databases. Data warehouses. Big data.
Integrations (e.g. tracking, channel attribution).
BI tools. Automated reporting. Bespoke solutions.
Version control. Repo management. Code review.

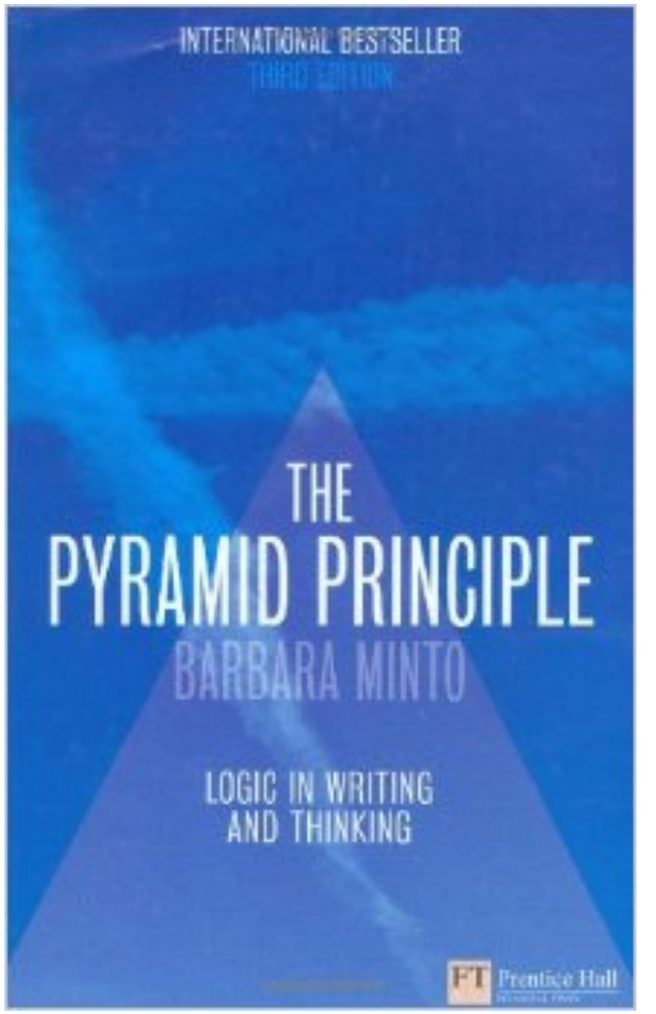
"Strategic Analysis"

Business skills. Startup methodology. Working lean.
Measuring success. KPIs. Data-informed decisions.
Communication. Technical writing. Domain expertise.
Project management. Agile workflows. Problem solving.
Education. Hiring. Mentoring. Advisory.

MY TOP 3 BOOK RECOMMENDATIONS







PODCASTS

- **Data Skeptic** (Kyle Polich, I ❤️ the mini-explainer episodes!)
- **Partially Derivative** (light hearted)
- **Linear Digressions** (Udacity)
- **More or Less** (Tim Harford & BBC Radio 4)
- **O'Reilly Data Show** (Ben Lorica, technical with more focus on data engineering)
- **Planet Money** (NPR, economics/data/finance – A/B testing, multiple comparisons)
- **What's The Point** (FiveThirtyEight, how data is changing our lives)
- **Science Vs** (Gimlet Media, new this summer, controversial issues + rigour)

LONDON MEETUPS

1. PyData London
2. LondonR
3. Data Science Meetup London
4. Big Data London
5. London Machine Learning Meetup
6. Quantified Self
7. Predictive Analytics London Meetup
8. Data Visualization Meetup
9. PyLadies London
10. Women in Data
11. Londata
12. Data Science Journal Club

HACKATHONS AND DATADIVES

- DataKind
- NHS Hack
- Kaggle
- UK Hackathons & James Meetup
- StartupWeekend
- Code for Good

FOUR STEPS TO SUCCESS

1. Learn to code

Python. R. Professional software engineering practices.

2. Get statistical

Significance. Inference. Regression. Machine learning.

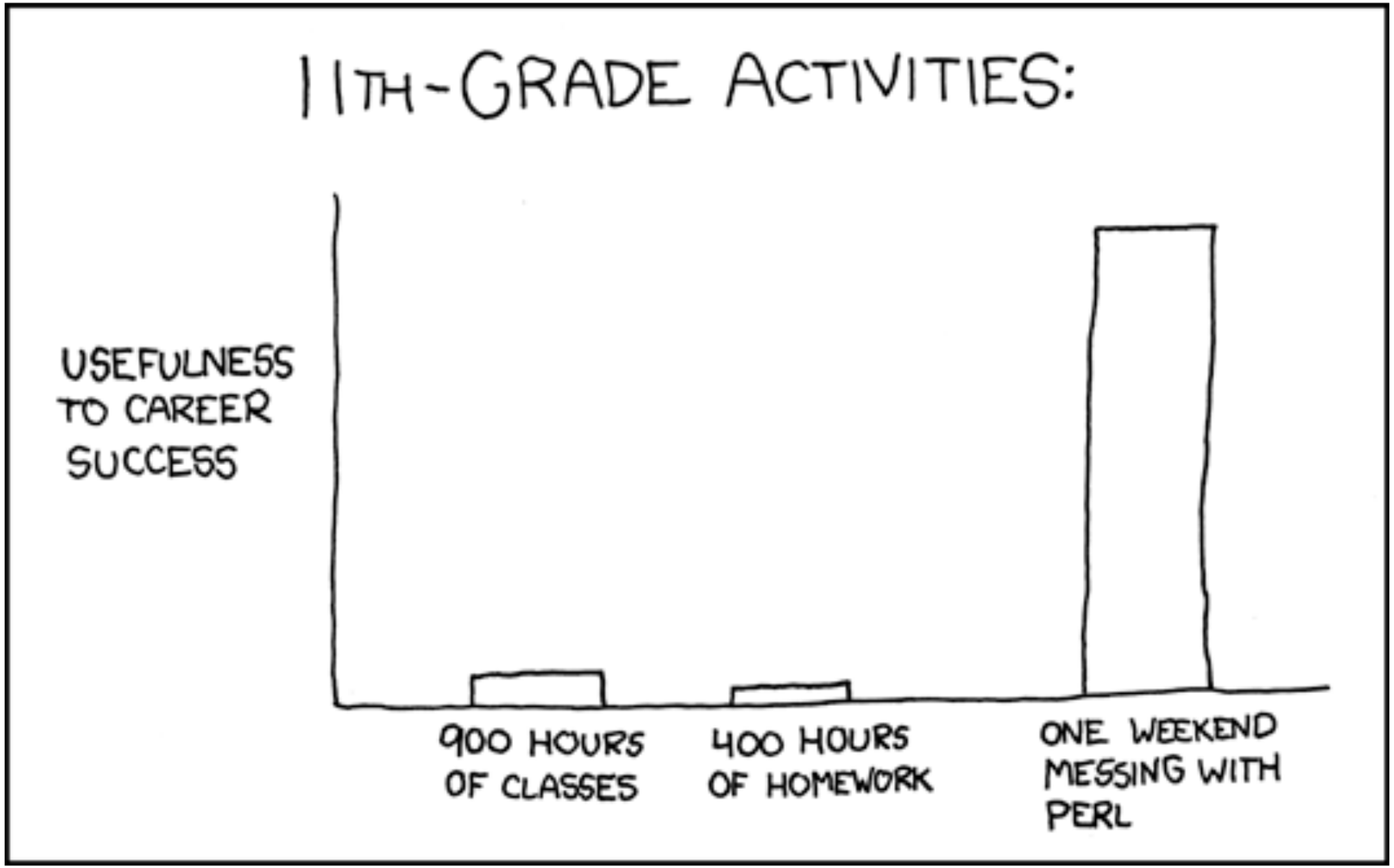
3. Learn lean

Business skills. Startup methodology. Communication.

4. Experience

Side projects. Github. Kaggle. Hackathons. Stand out.

"BECOME A TOP DATA SCIENTIST WITH THIS ONE WEIRD TIP..."



QUESTIONS?