Welcome to

SMWi

BRISTOL

AN INDEPENDENT SOCIAL MEDIA WEEK

SMWi BRISTOL

AN INDEPENDENT SOCIAL MEDIA WEEK

Thank you to all of our partners and sponsors

GENERAL ASSEMBLY

HOW DATA SCIENCE COULD TRANSFORM YOUR SOCIAL MEDIA STRATEGY

John Sandall

Data Science Consultant

Wednesday 16th November 2016

@john_sandall

WELCOME!

I. WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

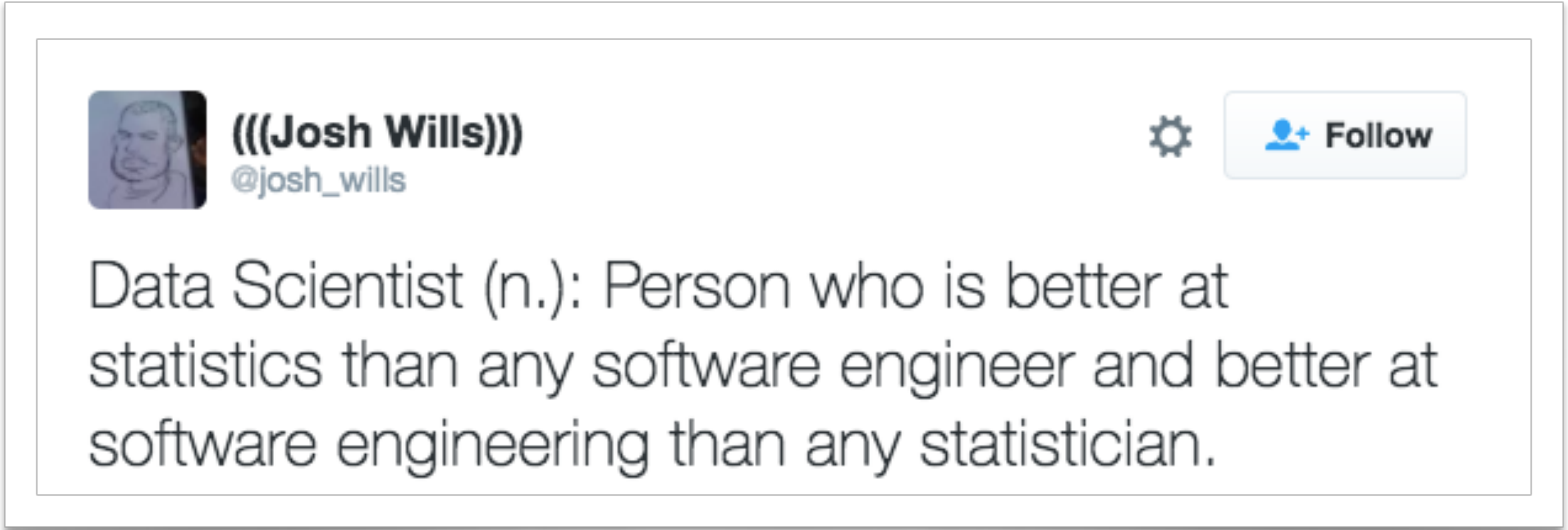
WHAT IS DATA SCIENCE?

Big Data Borat
@BigDataBorat

Follow

Data Science is statistics on a Mac.

WHAT IS DATA SCIENCE?

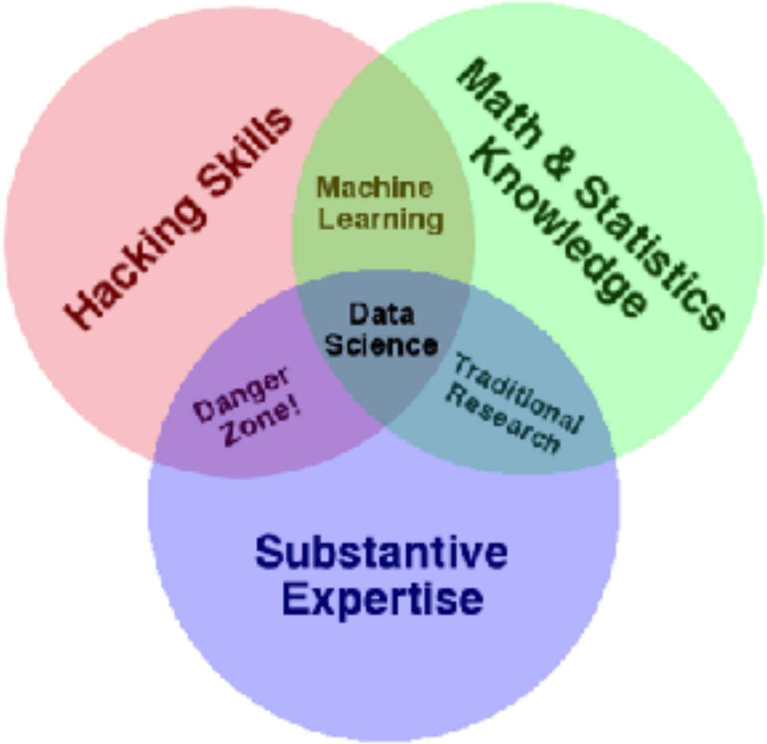


(((Josh Wills)))
@josh_wills

Follow

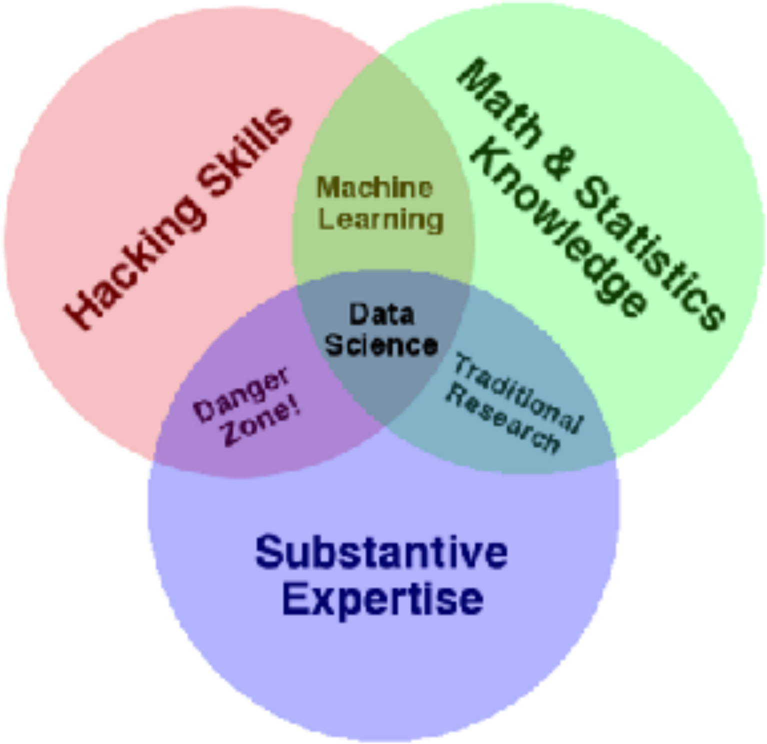
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

THE QUALITIES OF A DATA SCIENTIST



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

THE QUALITIES OF A DATA SCIENTIST



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

ONE MORE THING!

Communication skills

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.

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.

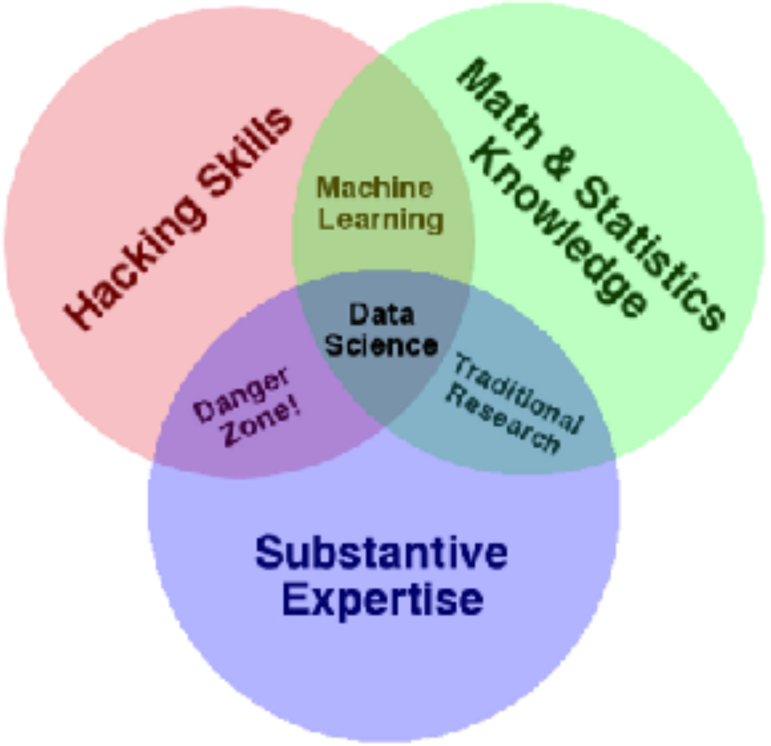
Measuring success. KPIs. Data-informed decisions.

Communication. Technical writing. Domain expertise.

Project management. Agile workflows. Problem solving.

Education. Hiring. Mentoring. Advisory.

DATA ANALYST...OR DATA SCIENTIST?



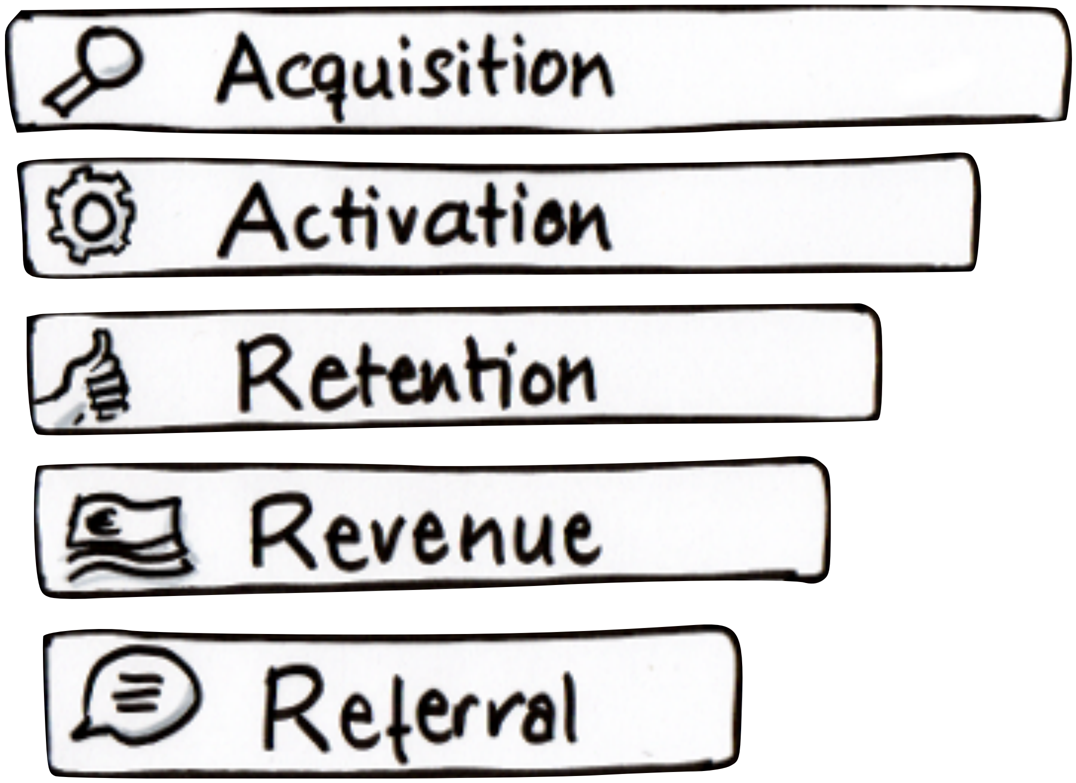
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

II. GET YOUR METRICS RIGHT

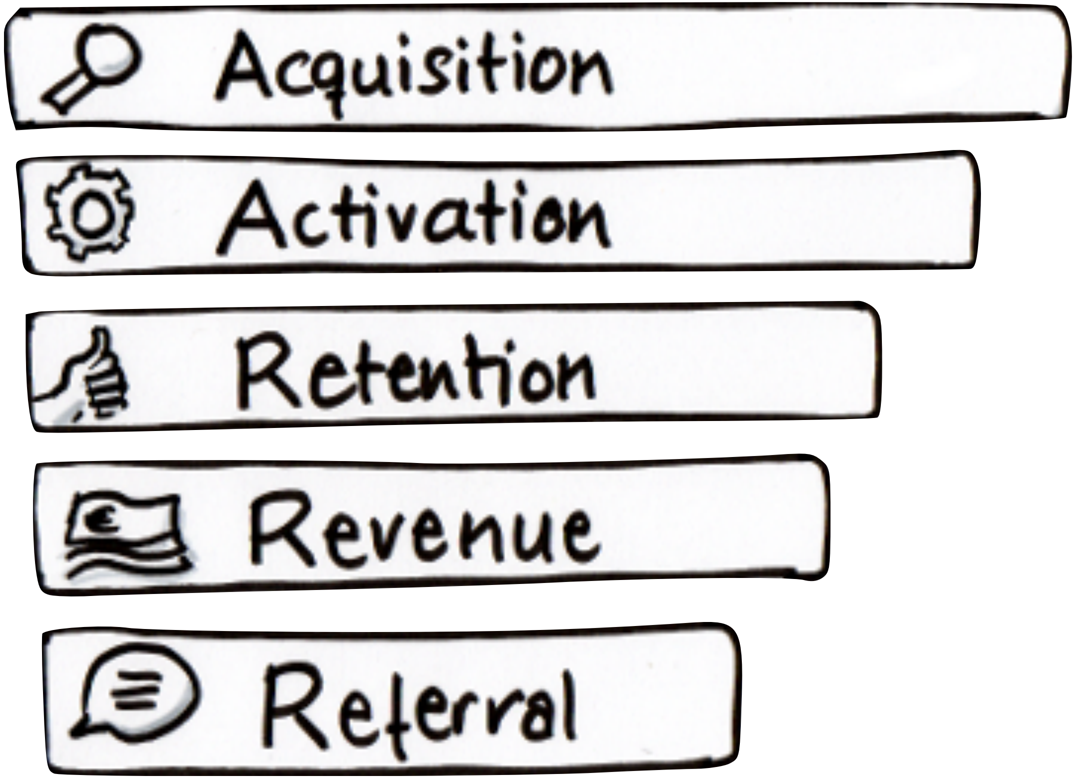
GETTING YOUR METRICS RIGHT

- **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

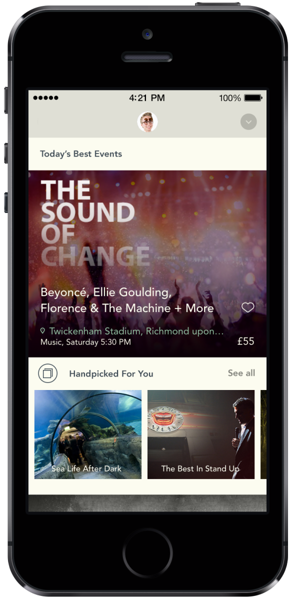
↓

Paying customer

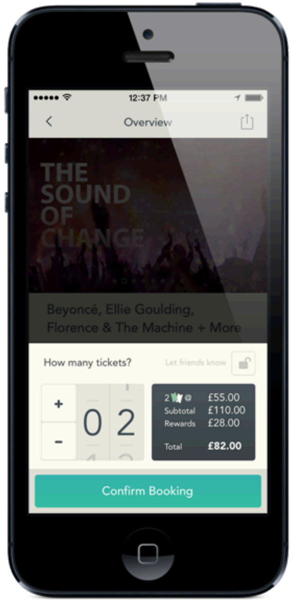
↓

Promoter

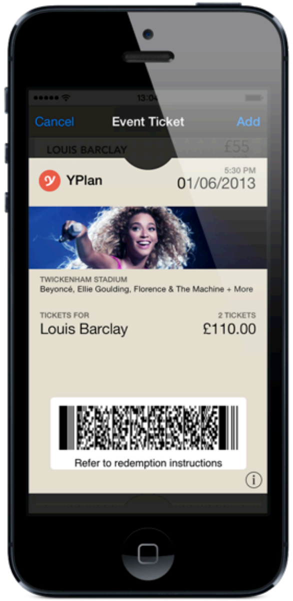
AARRR FOR YPLAN



Spontaneous inspiration



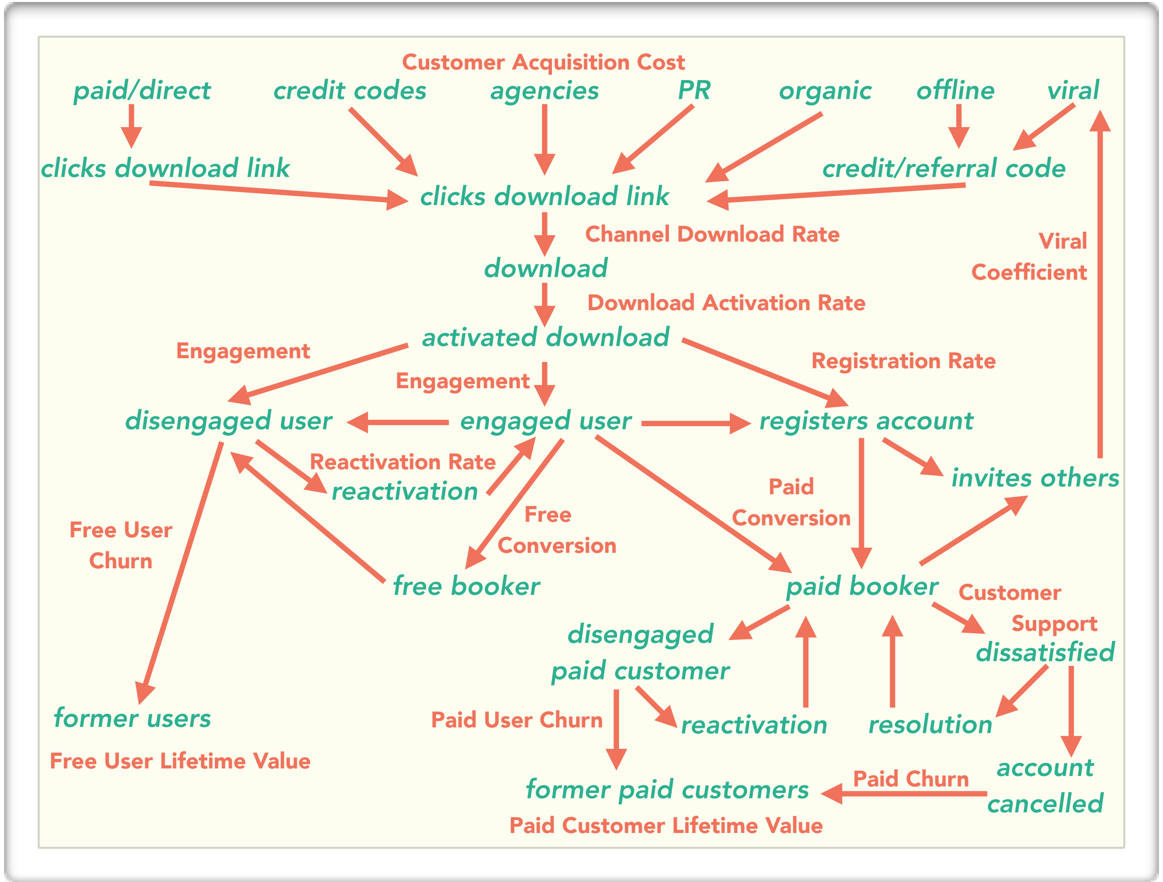
2-tap booking process



Paperless ticketing

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

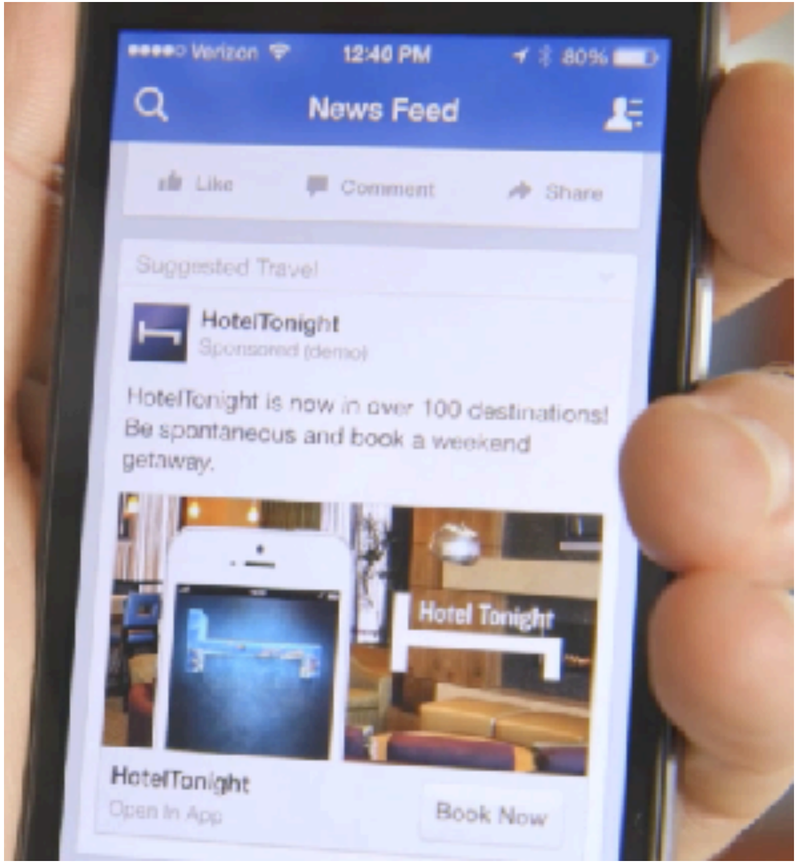
AARRR FOR YPLAN



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

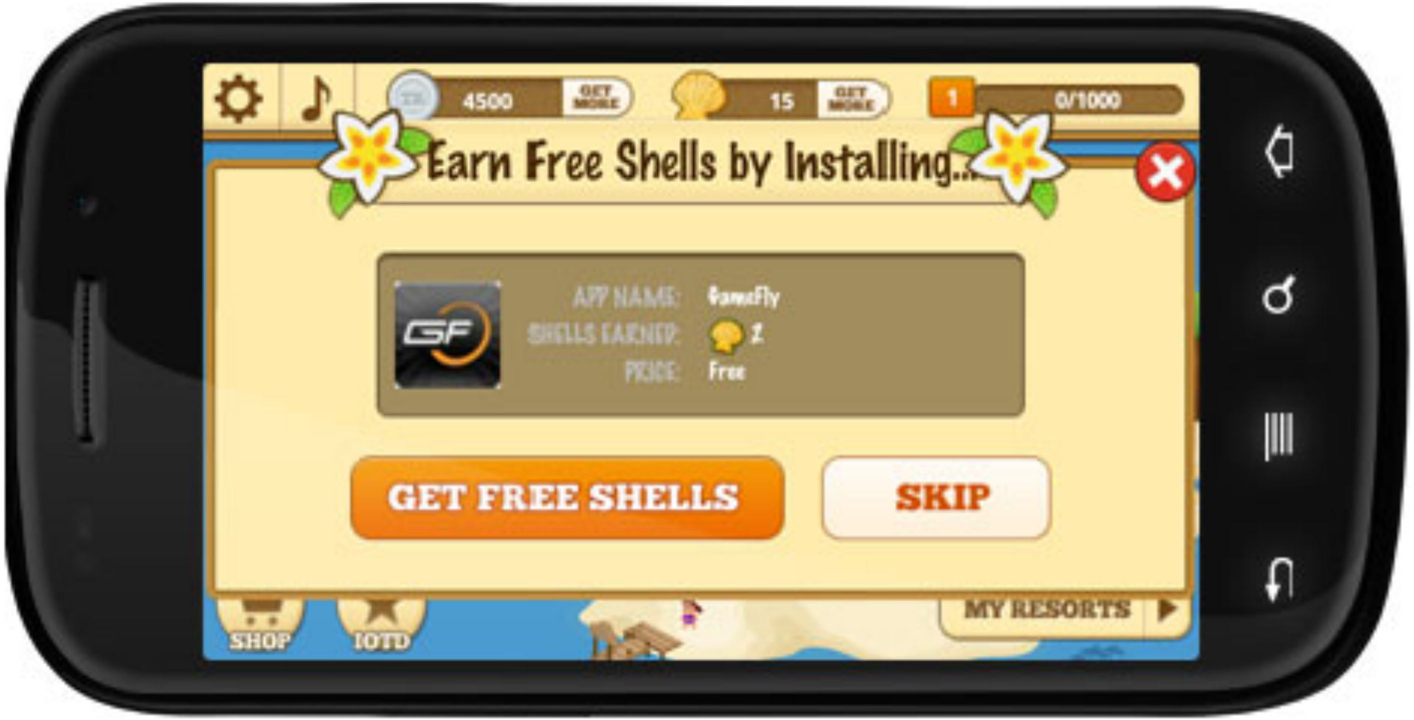
EXAMPLE: ADVERTISING ON GOOGLE/FACEBOOK/AD NETWORKS

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



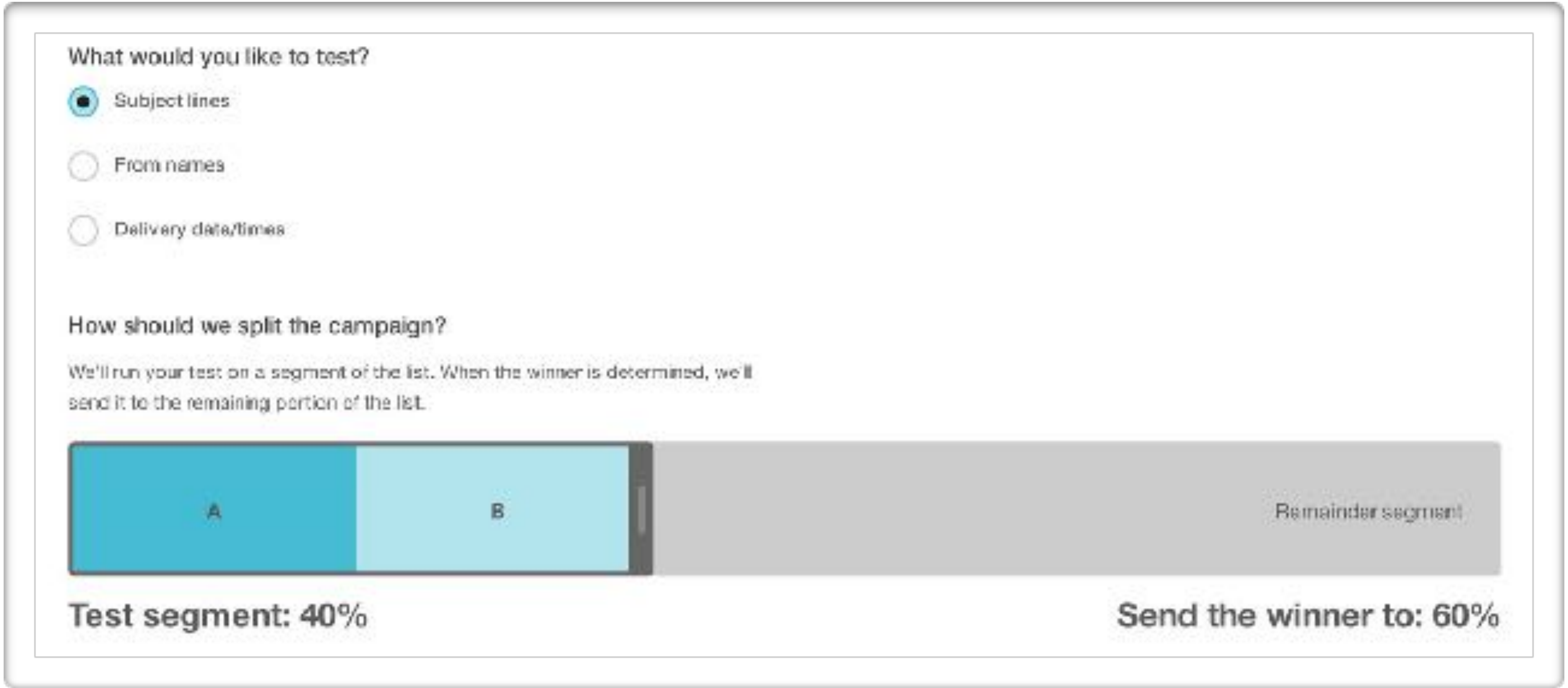
EXAMPLE: ADVERTISING ON GOOGLE/FACEBOOK/AD NETWORKS

- **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!

III. PREDICTIVE MODELLING

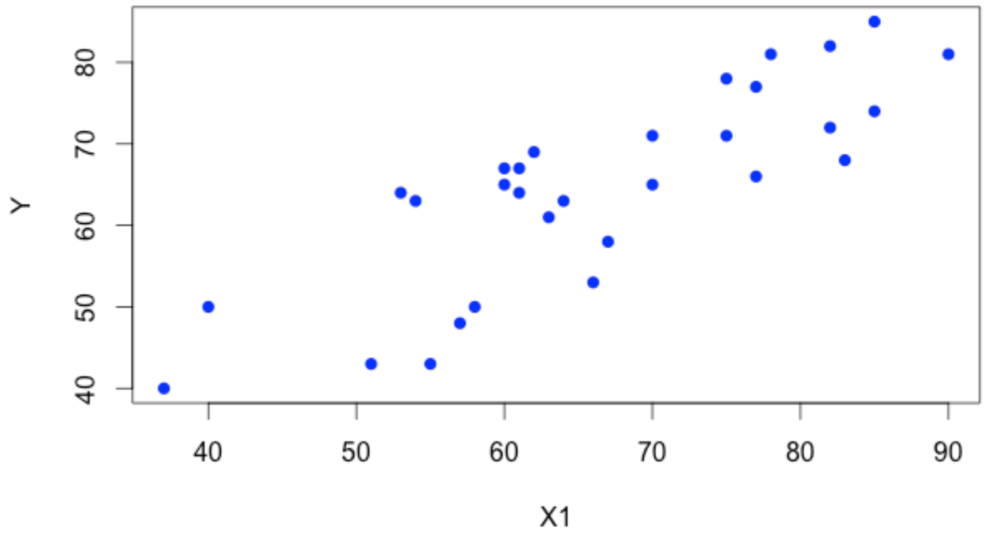
SO FAR...

- Advanced tracking, data management, reporting, segmentation, ...
- Setting up a data warehouse that can easily answer such questions isn't trivial, but once in place there's so much more that can be done!
- Welcome to the wonderful world of machine learning!
- **What is machine learning?**

supervised

Y	X1
43	51
63	64
71	70
61	63
81	78
43	55

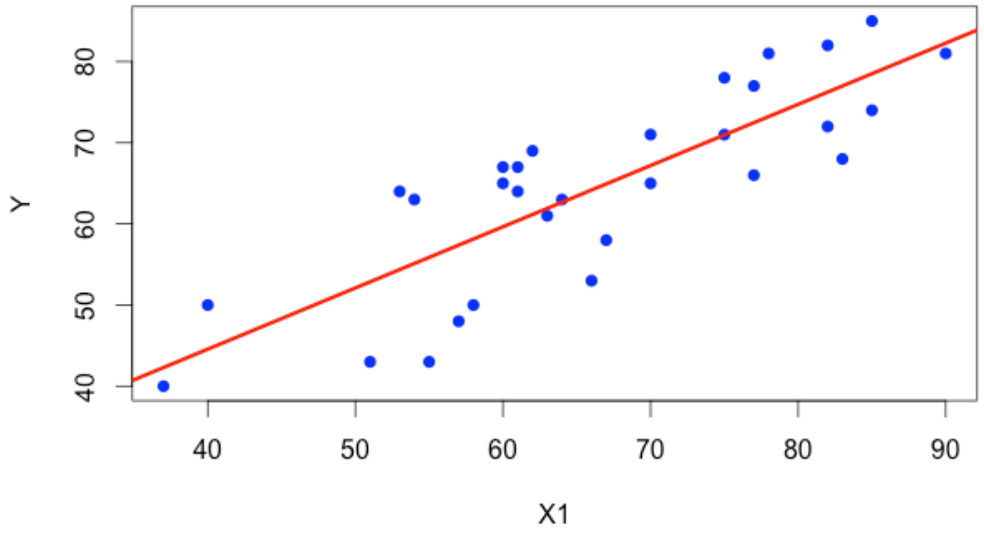
making predictions



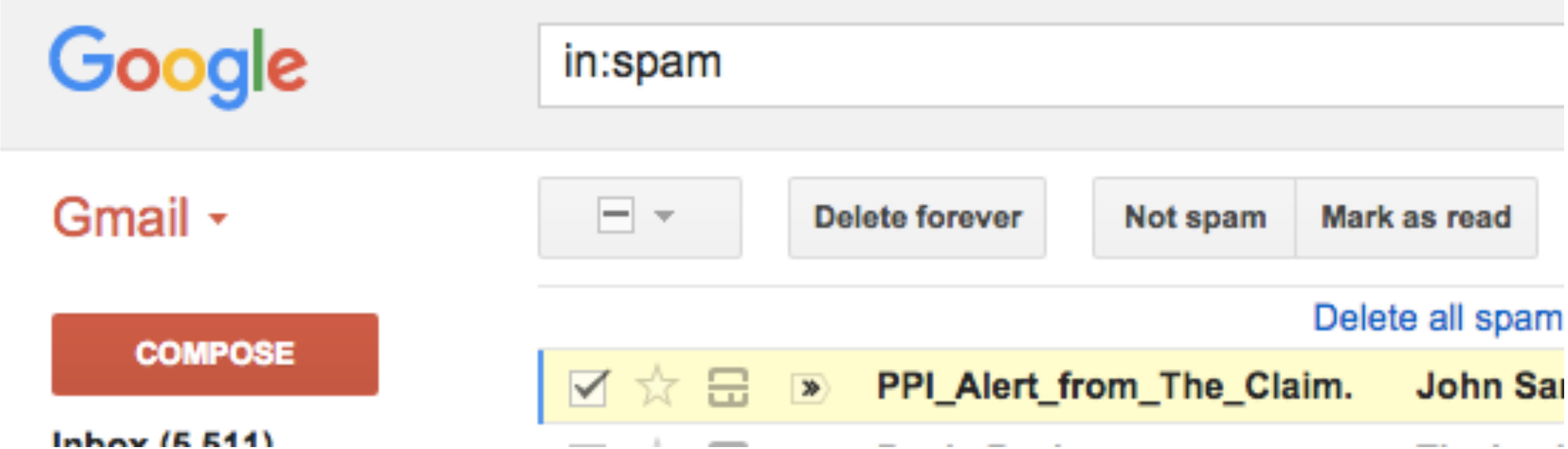
supervised

Y	X1
43	51
63	64
71	70
61	63
81	78
43	55

making predictions



CLASSIFICATION EXAMPLE: SPAM FILTER



$$$ Act now! As seen on Satisfaction guaranteed

100% free All natural Bargain !!!

THEORY ALERT

DECISION TREES

AN ILLUSTRATIVE EXAMPLE

EXAMPLE: PREDICTING MOVIE PREFERENCES

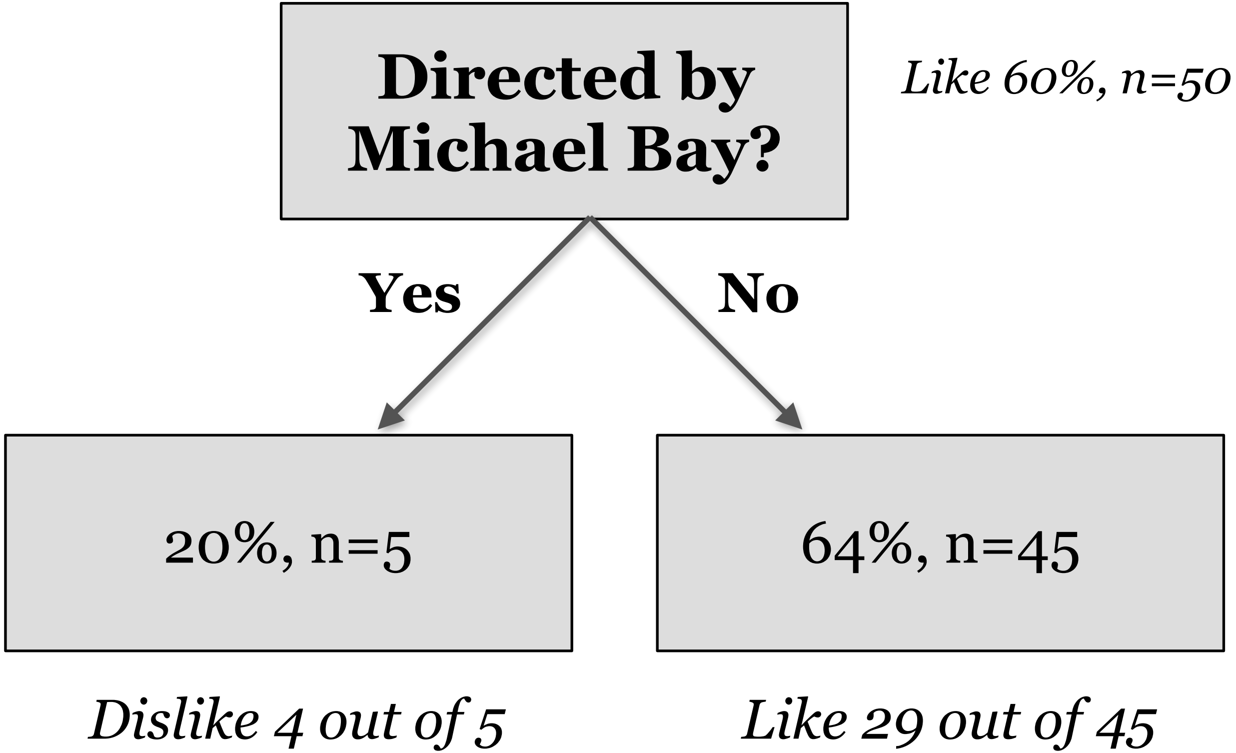
Film	Released	Director	Starring	IMDB rating	Genre	Like/ Dislike?
Frozen	2013	Chris Buck, Jennifer Lee	Kristen Bell, Idina Menzel, Jonathan Groff	7.6	Animation, Adventure, Comedy	✔
Moneyball	2011	Bennett Miller	Brad Pitt, Robin Wright, Jonah Hill	7.6	Biography, Drama, Sport	✔
Transformers	2007	Michael Bay	Shia LaBeouf, Megan Fox, Josh Duhamel	7.1	Action, Adventure, Sci-Fi	❌

EXAMPLE: PREDICTING MOVIE PREFERENCES

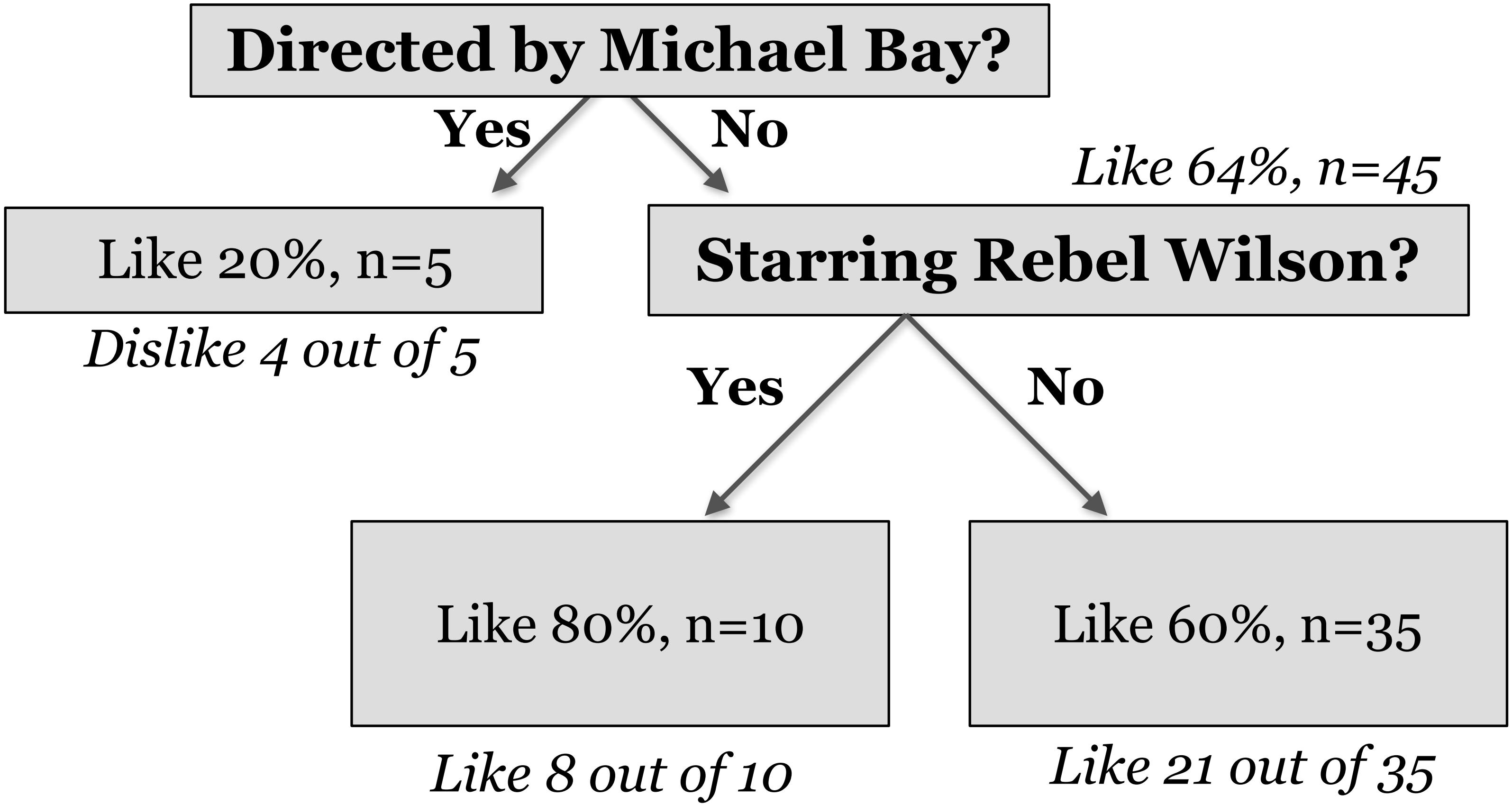
Like 60%, n=50

I like 30 out of 50 films I've watched on Netflix ("60% like")

EXAMPLE: PREDICTING MOVIE PREFERENCES



EXAMPLE: PREDICTING MOVIE PREFERENCES



EXAMPLE: PREDICTING MOVIE PREFERENCES

- **Some questions don't help, e.g.**
    - "Is the film length (in minutes) even or odd?"
    - This won't help separate into distinct groups that I either love or hate
- **Questions could be numeric cutoffs, e.g.**
    - "Was the film produced before 1970?"
    - "Is the film between 2 and 3 hours in length?"
- **Continue the tree until we can accurately predict for new movies**

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)

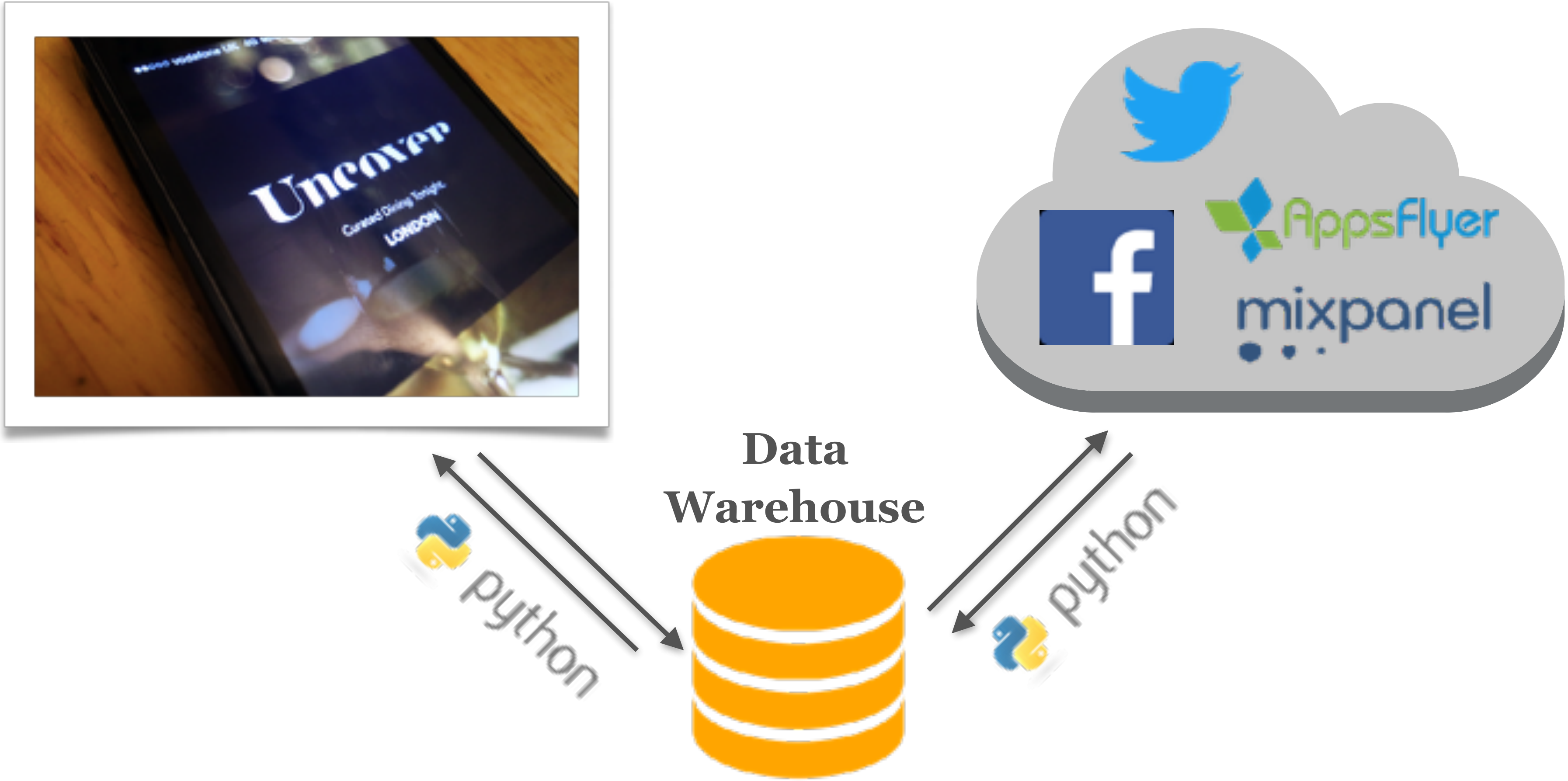
Signup Conversion Rate %

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…**
  - 30% 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)

IV. EXTRACTING HIDDEN PATTERNS

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?

			M			
	R	F	1	2	3	4
ACTIVE	1	1				
		2				
		3				
		4				
AT RISK	2	1				
		2				
		3				
		4				
CHURNED	3	1				
		2				
		3				
		4				

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

THEORY ALERT

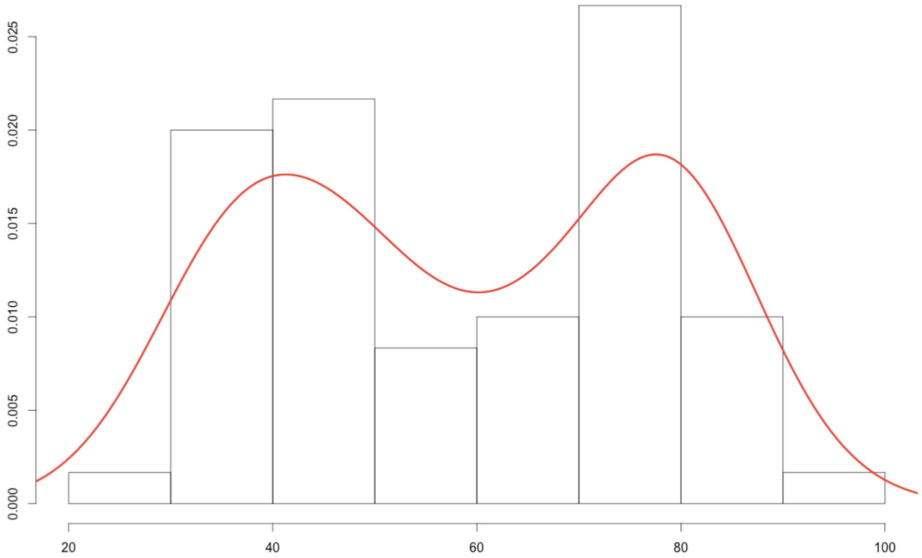
CLUSTERING

UNSUPERVISED MACHINE LEARNING

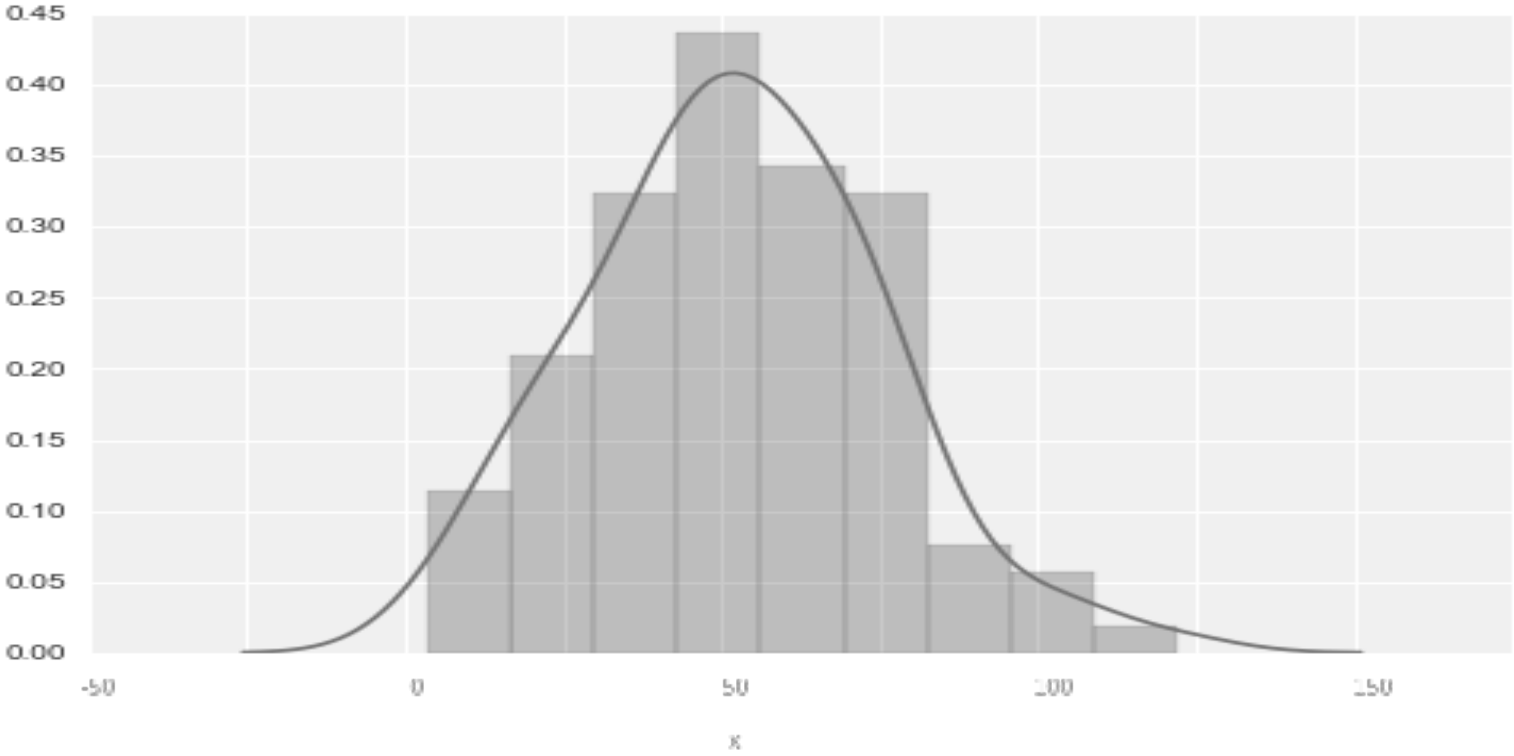
unsupervised

92
73
86
84
83
49
68
66
83
80
67
74
63

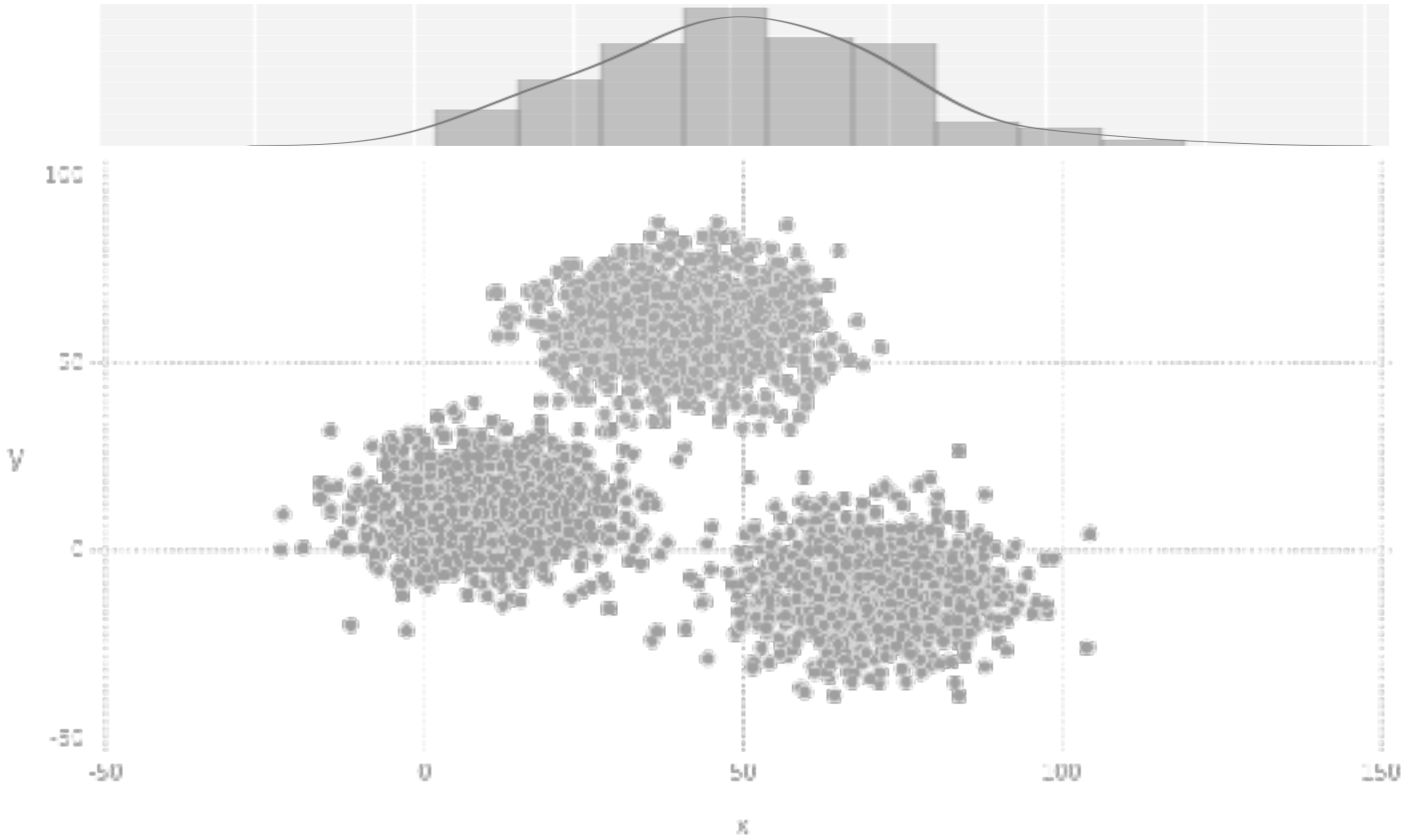
extracting structure



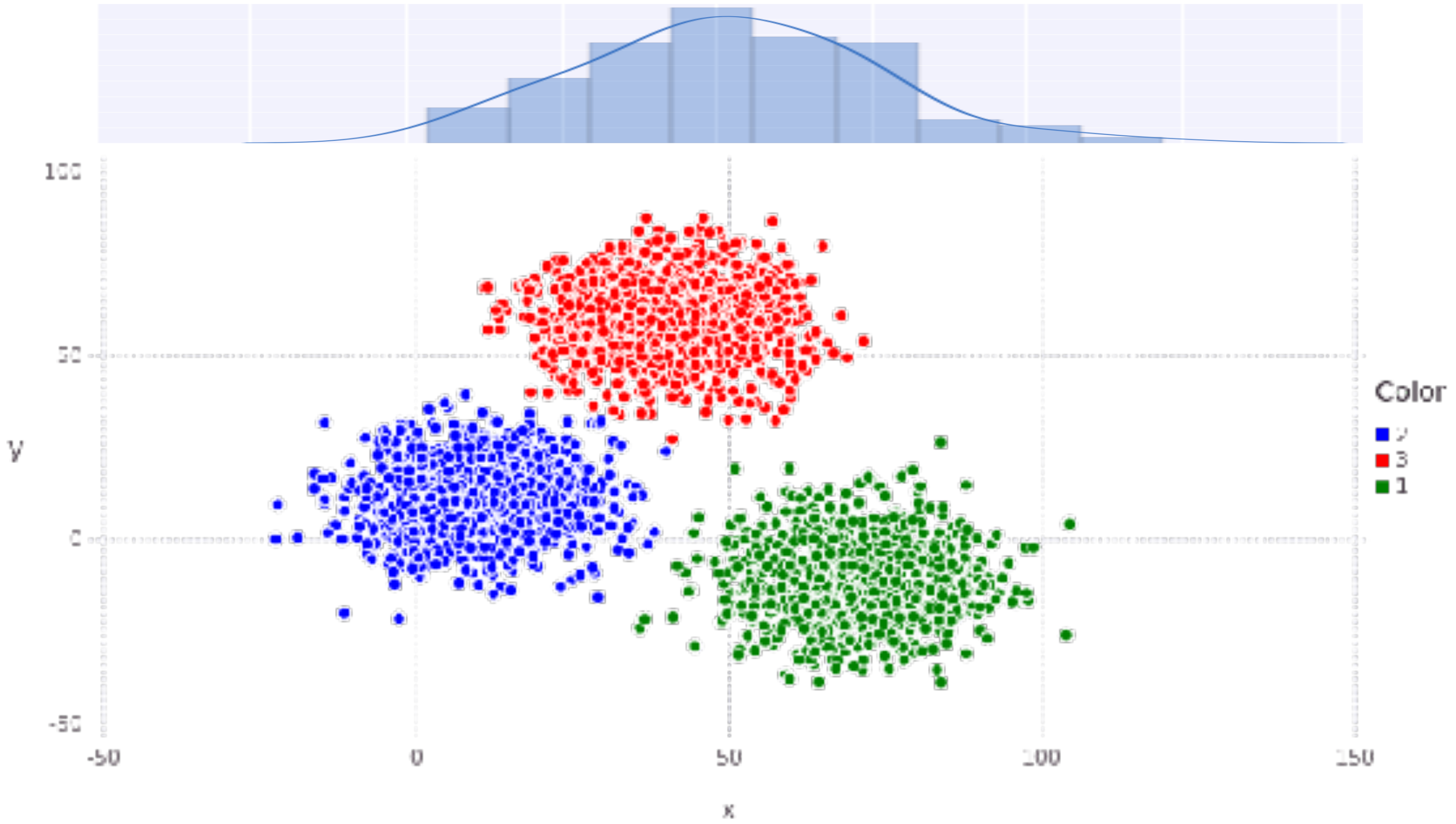
UNSUPERVISED TECHNIQUE: CLUSTERING



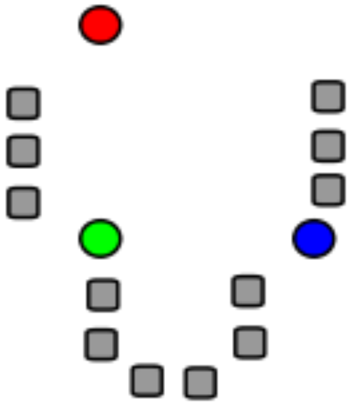
UNSUPERVISED TECHNIQUE: CLUSTERING



UNSUPERVISED TECHNIQUE: CLUSTERING



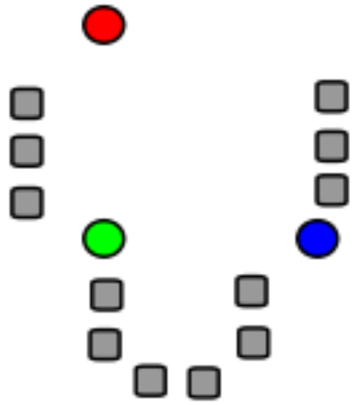
K-MEANS CLUSTERING



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



K-MEANS CLUSTERING

Steps

1. Choose k. Here, let's try to cluster into k=3 groups.

2. Choose 3 initial "centroids"...random points.

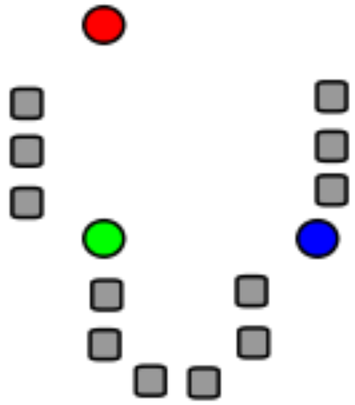
3. For each grey point:

    - 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

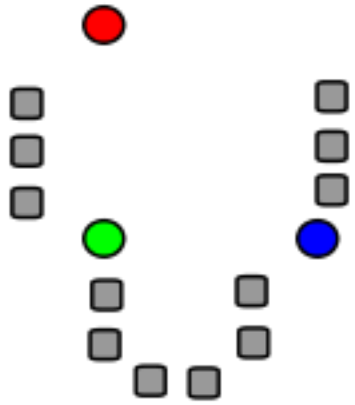
5. Repeat steps 3 and 4 until the centroids don't move



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



K-MEANS CLUSTERING

Steps

1. Choose k. Here, let's try to cluster into k=3 groups.

2. Choose 3 initial "centroids"...random points.

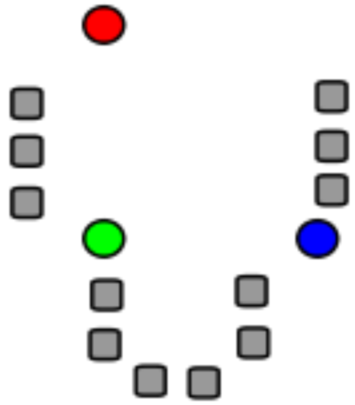
3. For each grey point:

   - 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

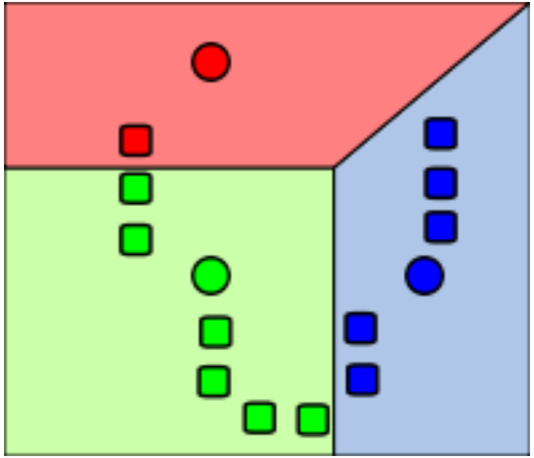
5. Repeat steps 3 and 4 until the centroids don't move



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



K-MEANS CLUSTERING

Steps

1. Choose k. Here, let's try to cluster into k=3 groups.

2. Choose 3 initial "centroids"...random points.

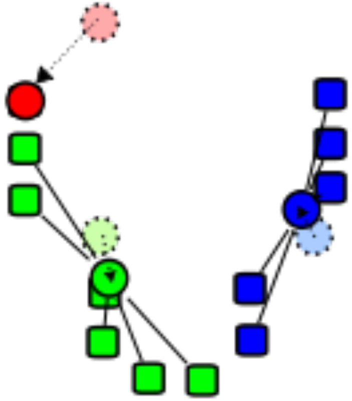
3. For each grey point:

   - 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

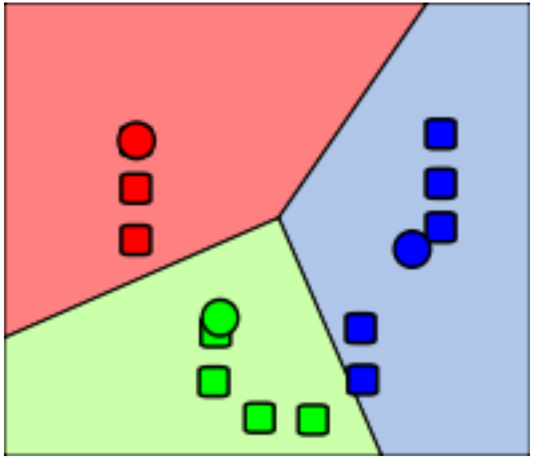
5. Repeat steps 3 and 4 until the centroids don't move



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 centr
of each cluster
5. Repeat steps 3 and 4 until the centroids don't move



K-MEANS CLUSTERING

Steps

1. Choose k. Here, let's try to cluster into k=3 groups.

2. Choose 3 initial "centroids"...random points.

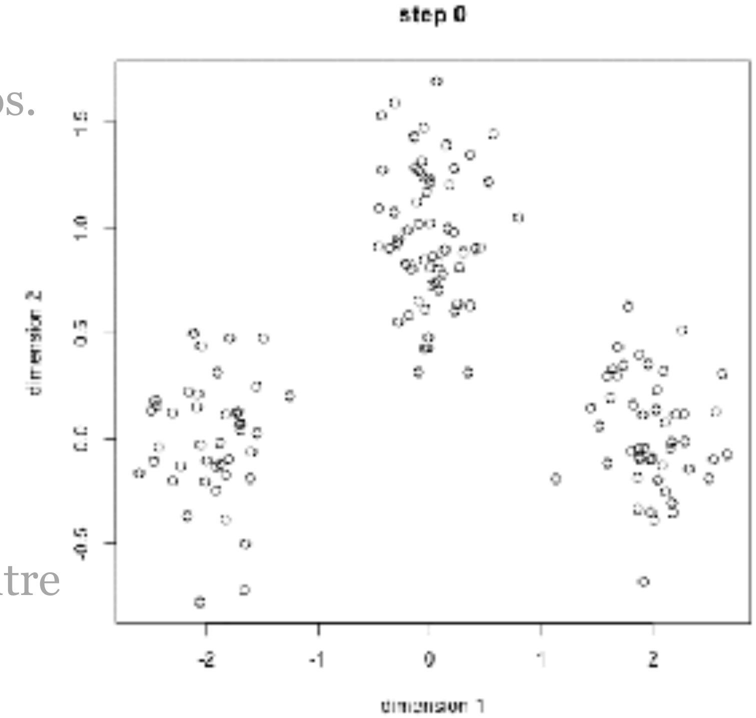
3. For each grey point:

    - find distance to nearest centroid

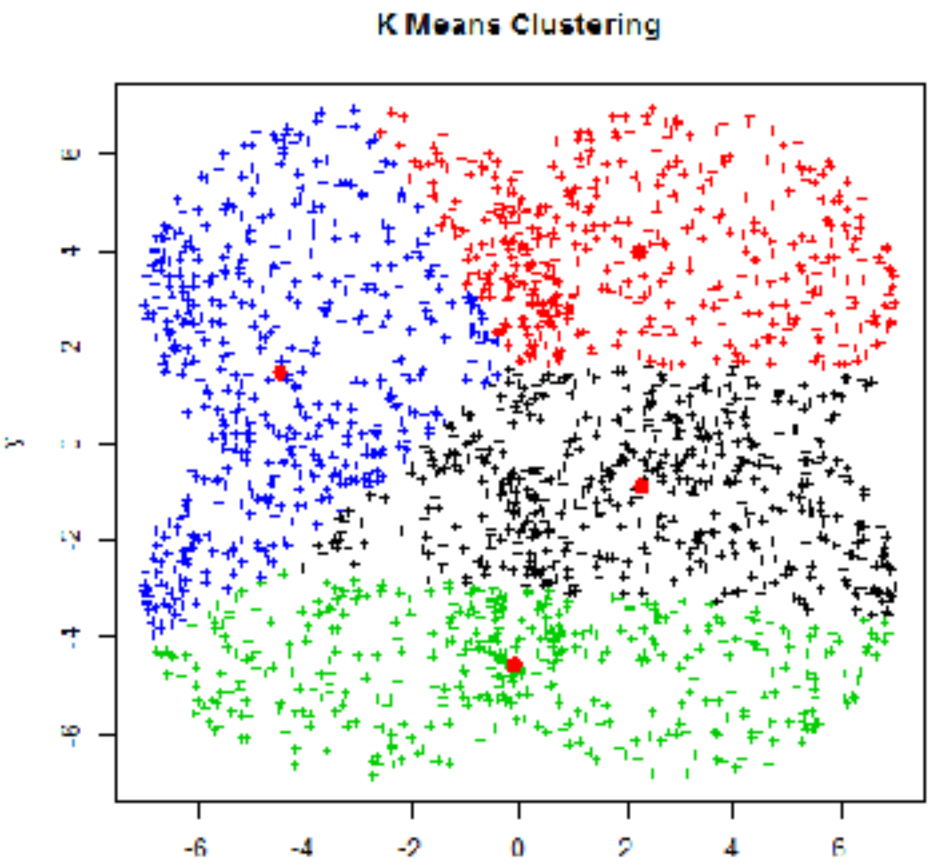
    - assign the point to the nearest centroid's "team"

4. Recalculate centroid positions to become the centre of each cluster

5. Repeat steps 3 and 4 until the centroids don't move



K-MEANS CLUSTERING



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

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

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

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

- 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?**

V. DATA SCIENCE IS…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?

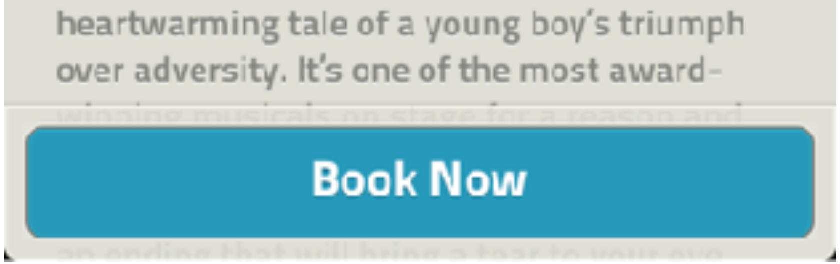
DATA + SCIENCE = DATA SCIENCE

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

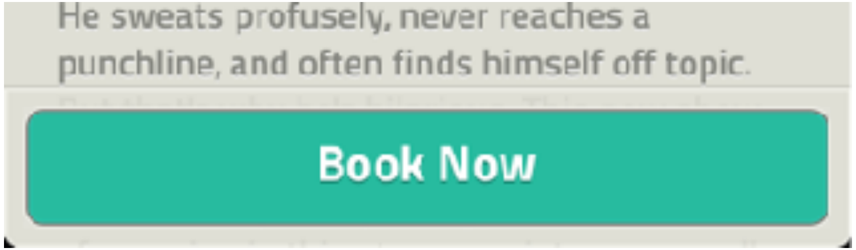
- Run the test:
    - Divide all customers into two groups at random
    - A gets old segmentation
    - B gets new segmentation
    - ...wait...
    - Crunch numbers once sample size is reached

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

A/B TESTING



vs



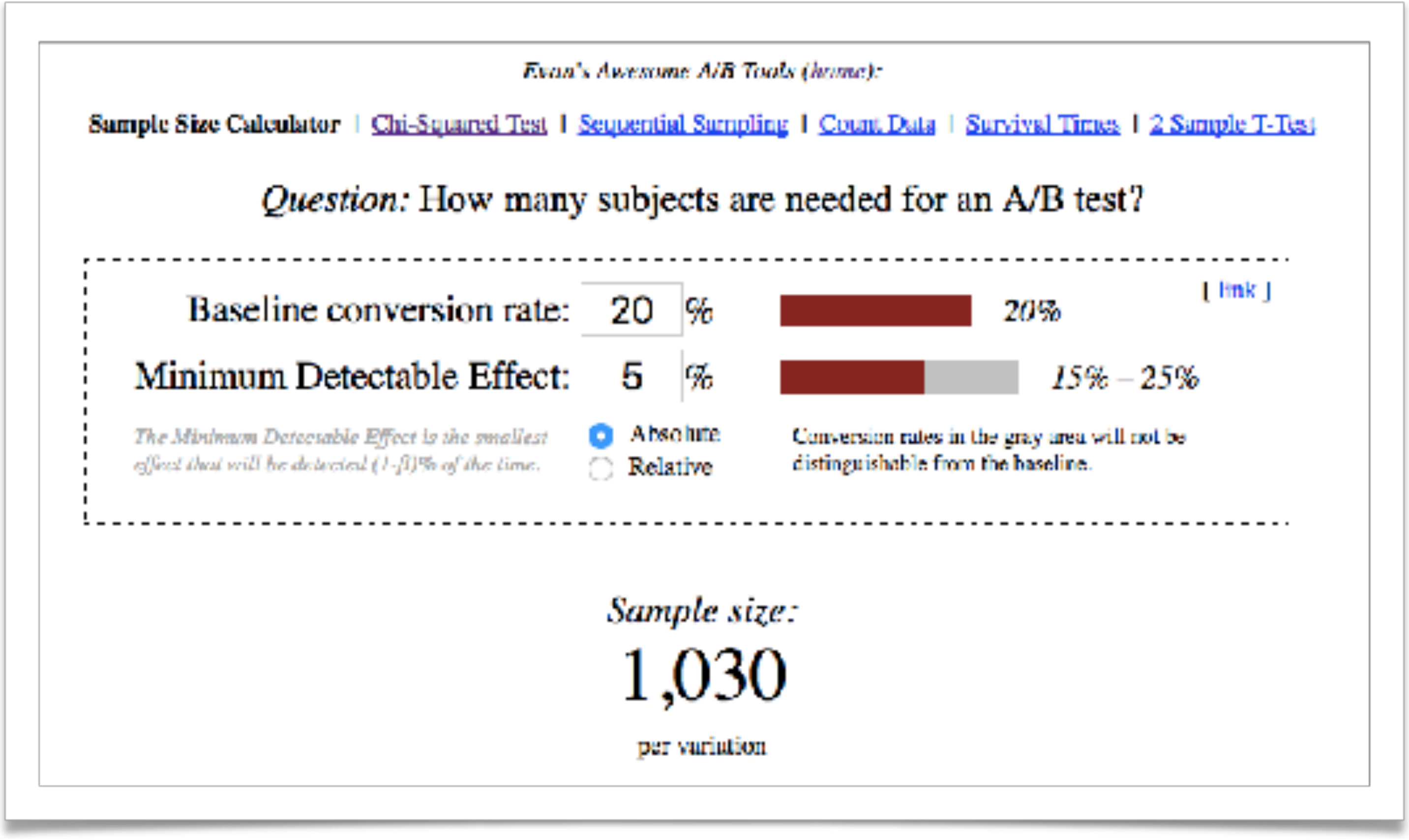
HOW NOT TO RUN AN A/B TEST

Statistical **significance** is for scientists!

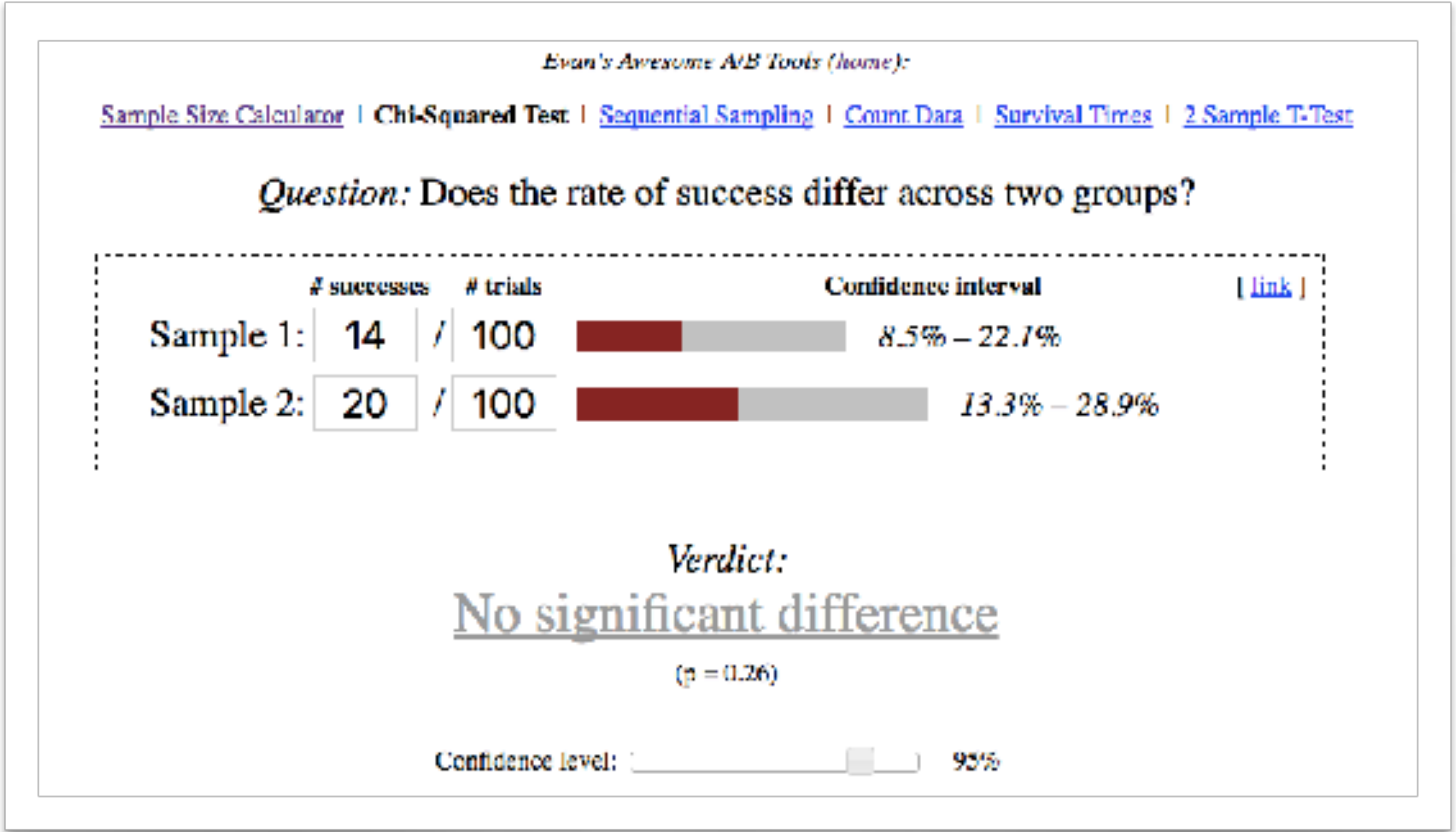
Don't bother calculating a **sample size**

End the A/B test **early**

TOP TIP: EVANMILLER.ORG



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

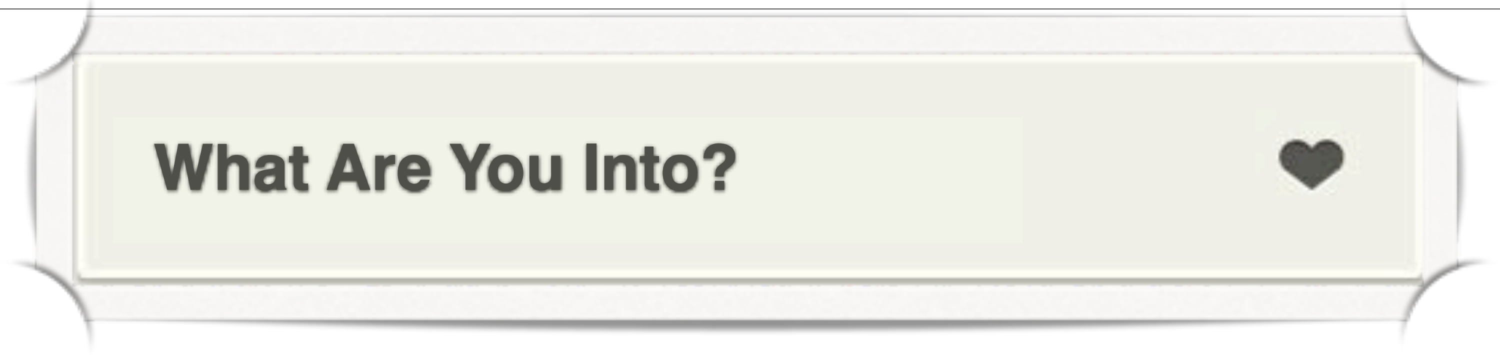
A/B TESTING = GROWTH

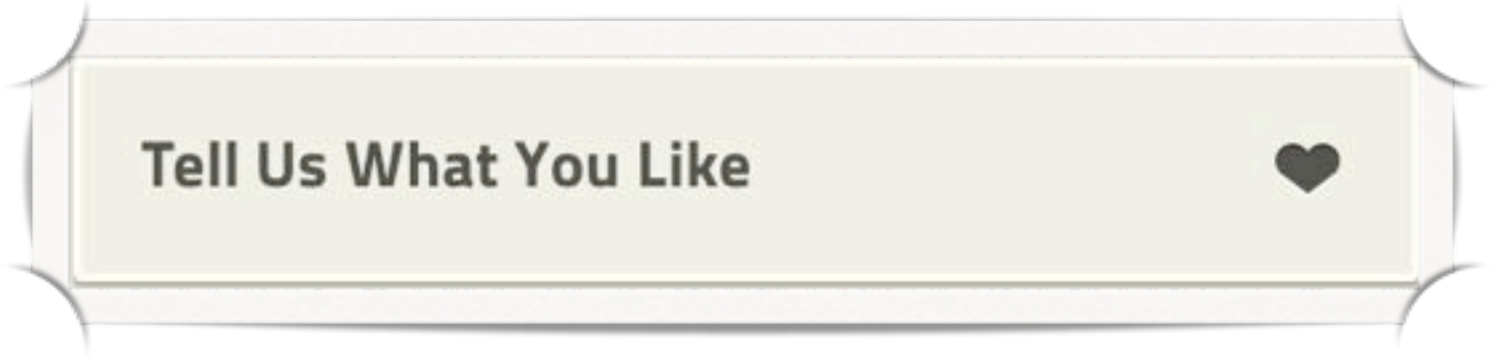
A/B TESTING = GROWTH

2048 different versions of the app

Fearless experimentation!

A/B TESTING = GROWTH





A/B TESTING = GROWTH

2048 different versions of the app

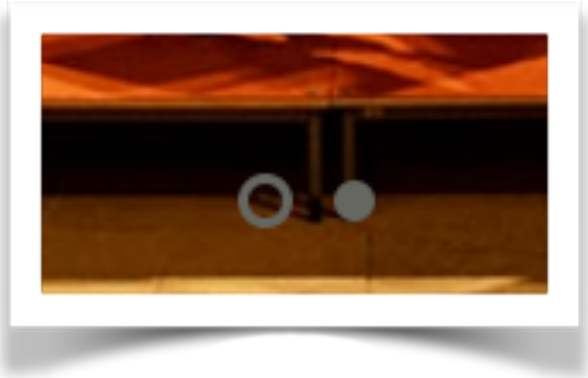
Fearless experimentation!

Not significant doesn't mean **boring**

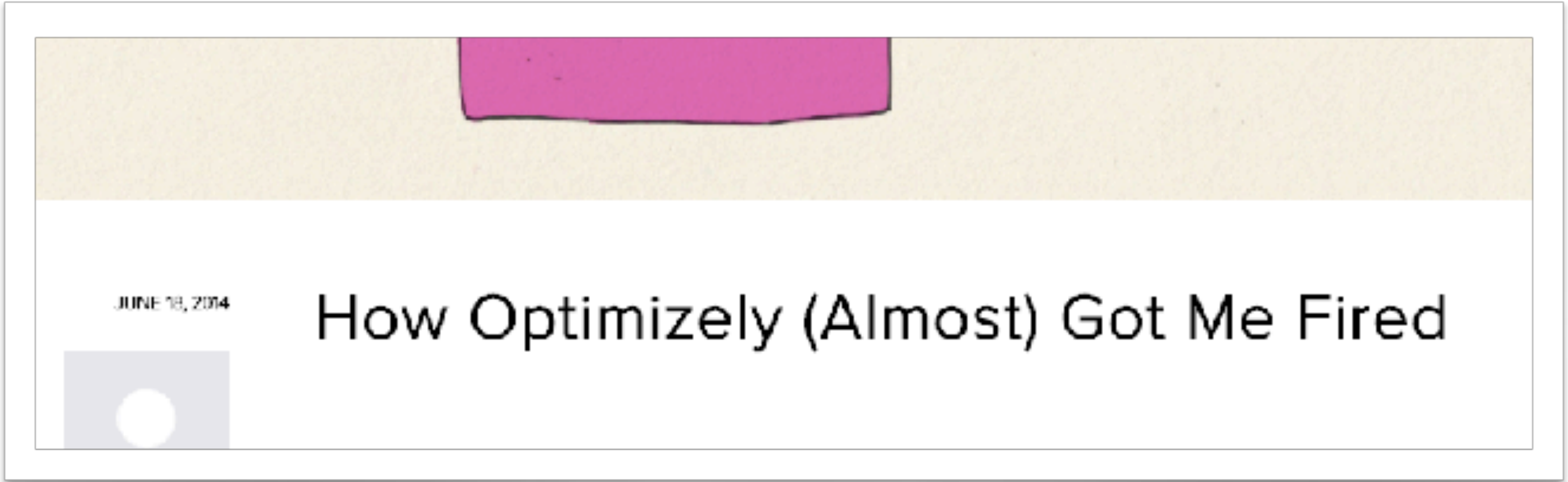
A/B TESTING = GROWTH

4 images

2 images



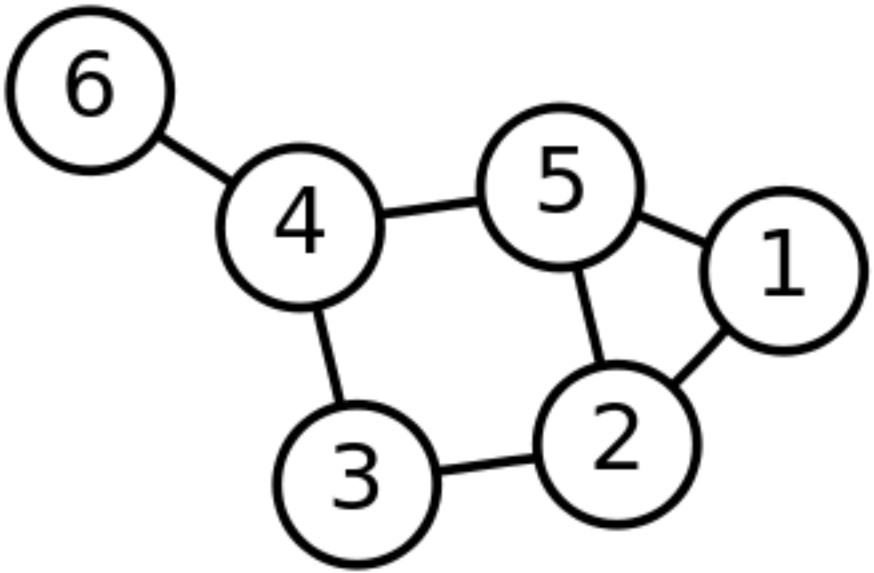
A WORD OF WARNING



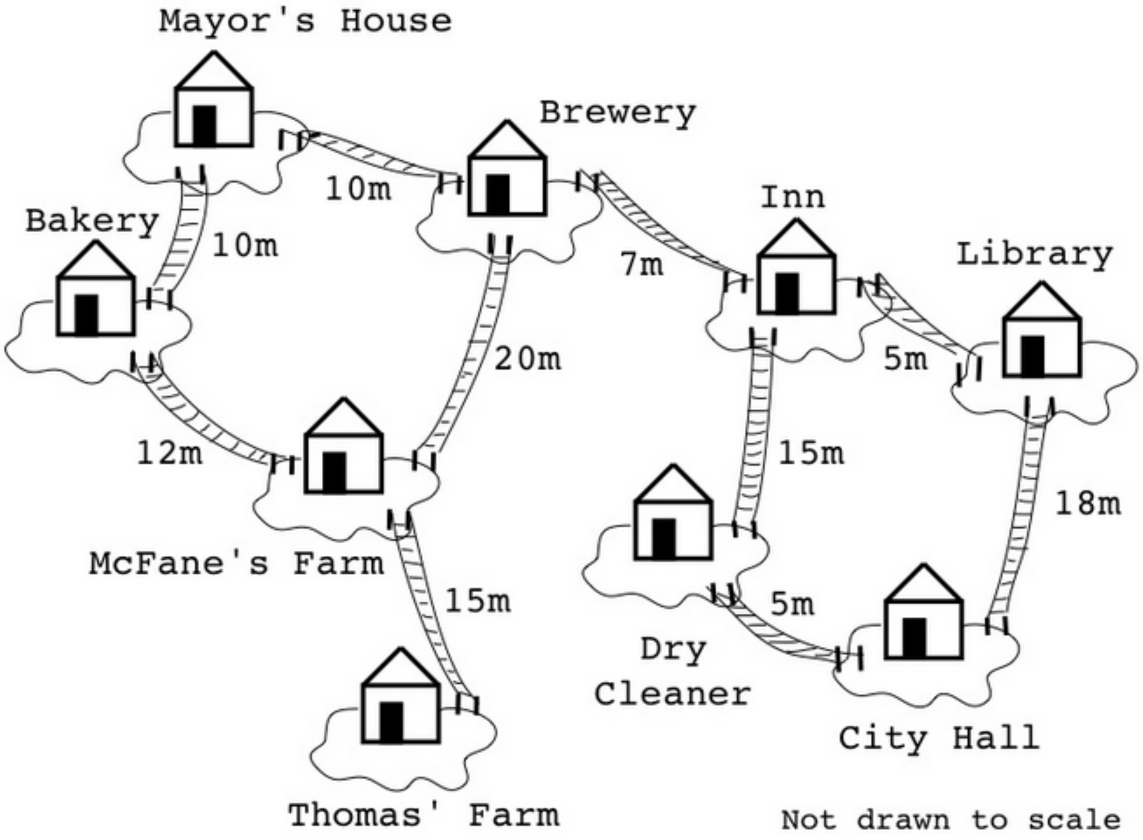
http://blog.sumall.com/journal/optimizely-got-me-fired.html

VI. NETWORKS

Nodes represent actors in the graph, and edges represent the relationships between actors.



NETWORK THEORY



NOTE

A *weighted graph* contains edges associated with real-valued numbers, eg to measure distance or importance.

NETWORK THEORY

SOCIAL MEDIA NETWORKS

THEORY ALERT

NETWORK THEORY

NETWORK DYNAMICS

Suppose we're interested in the idea of how information (or behavior) spreads through a network:

- *How do members of a social network influence each other to adopt a new technology/product/behavior?*
- *How did information about the bin Laden raid spread over Twitter?*
- *What's the best way to use a social network to market your product?*

NETWORK DYNAMICS

There are two primary methods of influence in social networks:

informational effects – people observe the decisions of their network neighbors & gain indirect information that lead them to try the innovation themselves

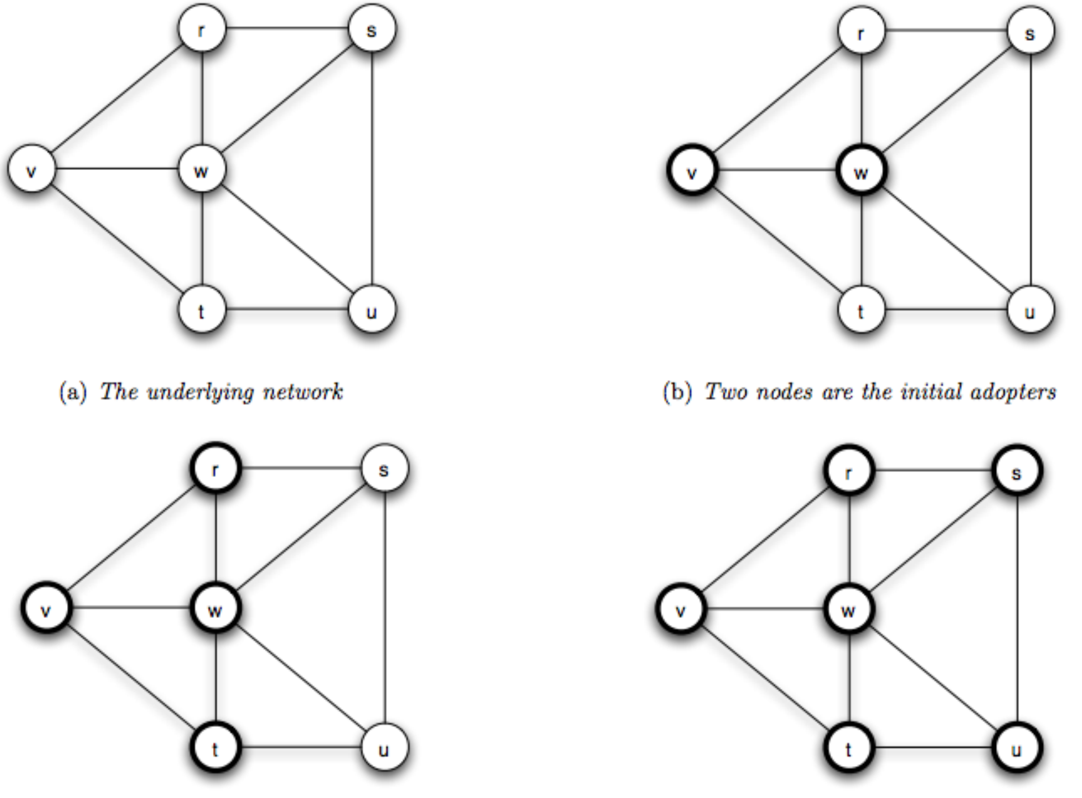
direct benefit effects – people may have incentives to use the same products/technology/etc as their network neighbors

NETWORK DYNAMICS

Studies of informational effects have shown that while initial lack of information makes innovations risky to adopt, adopters ultimately benefit.

Furthermore, early adopters share certain common traits (eg higher socio-economic status, wider travel experience), and they influence their neighbors by providing indirect information about the innovation.

NETWORK DYNAMICS



(a) *The underlying network*

(b) *Two nodes are the initial adopters*

(c) *After one step, two more nodes have adopted*

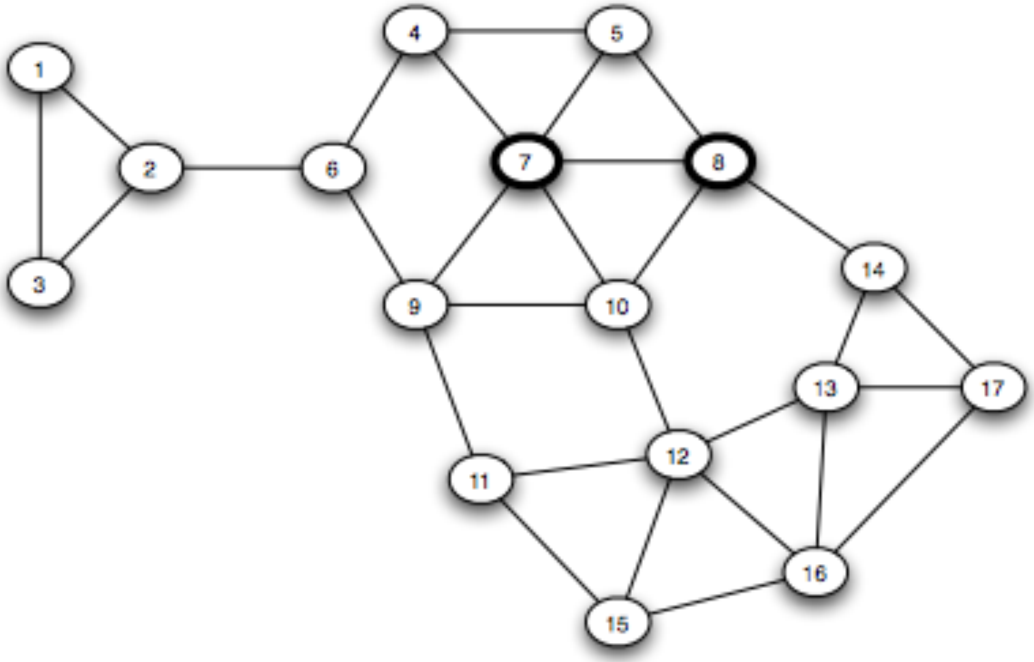
(d) *After a second step, everyone has adopted*

NOTE

Since all nodes have adopted, this is called a *complete cascade* (at threshold q).

DIFFUSION EFFECTS

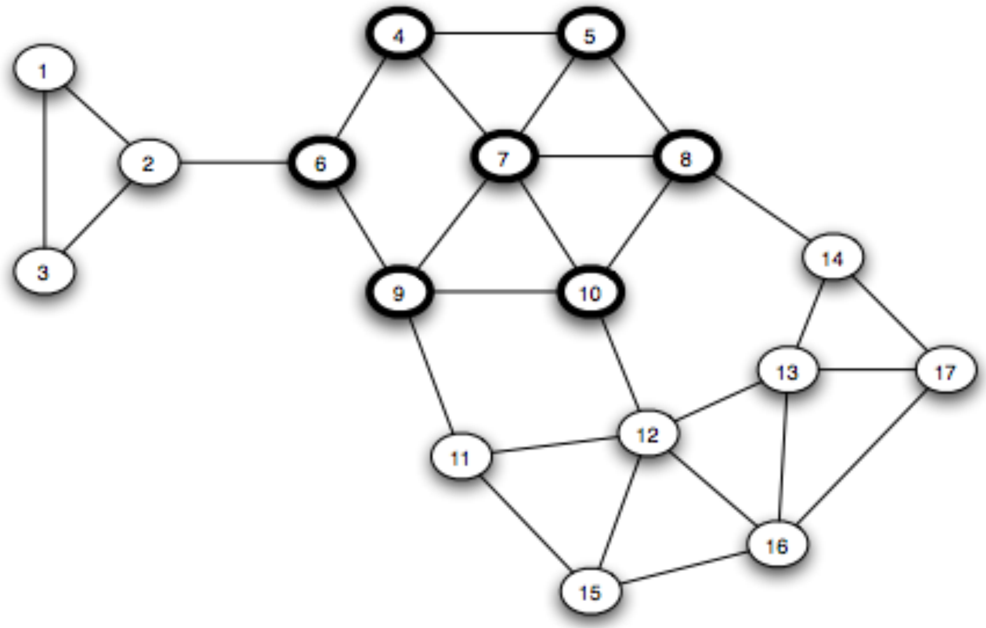
Consider the same diffusion now on another graph.



(a) *Two nodes are the initial adopters*

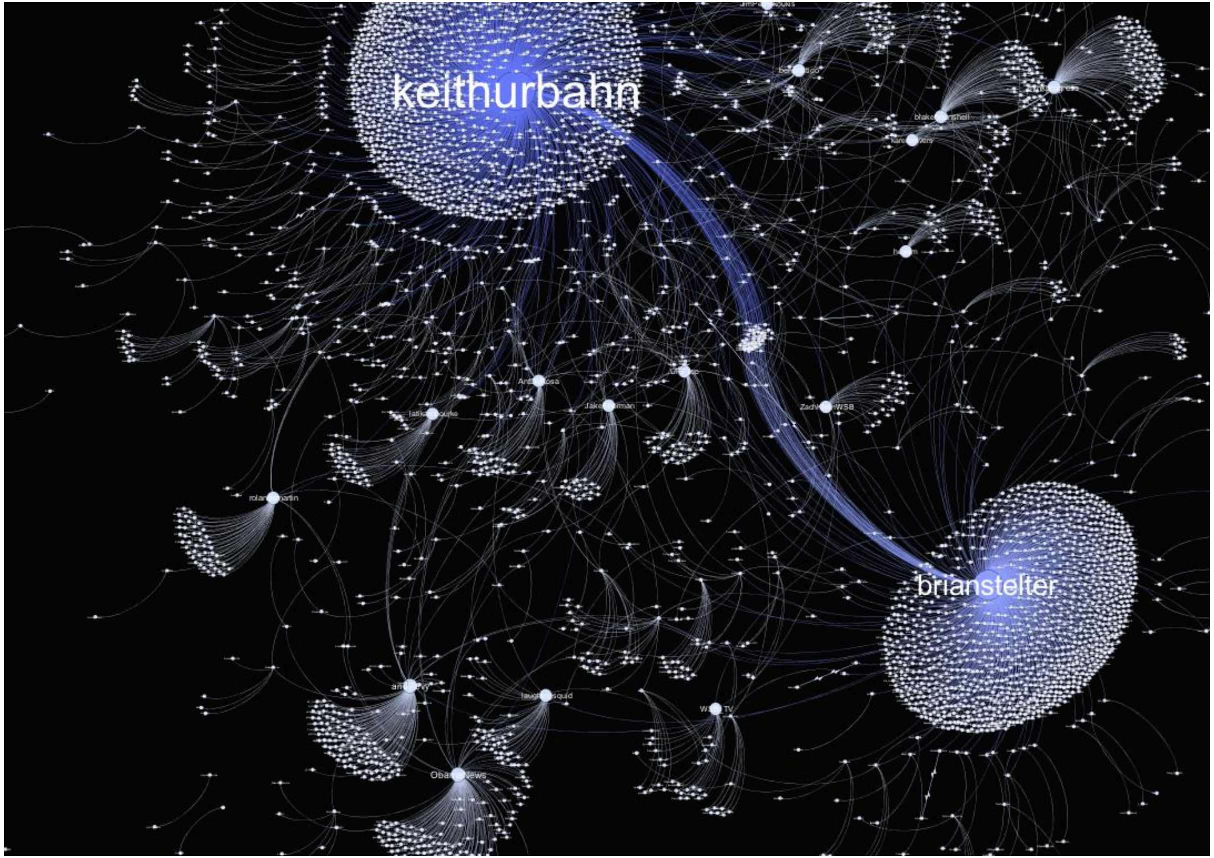
DIFFUSION EFFECTS

Since not all nodes adopt, this is called a partial cascade.



(b) *The process ends after three steps*

EXAMPLE: DIFFUSION EFFECTS IN TWITTER



IDENTIFYING INFLUENCERS

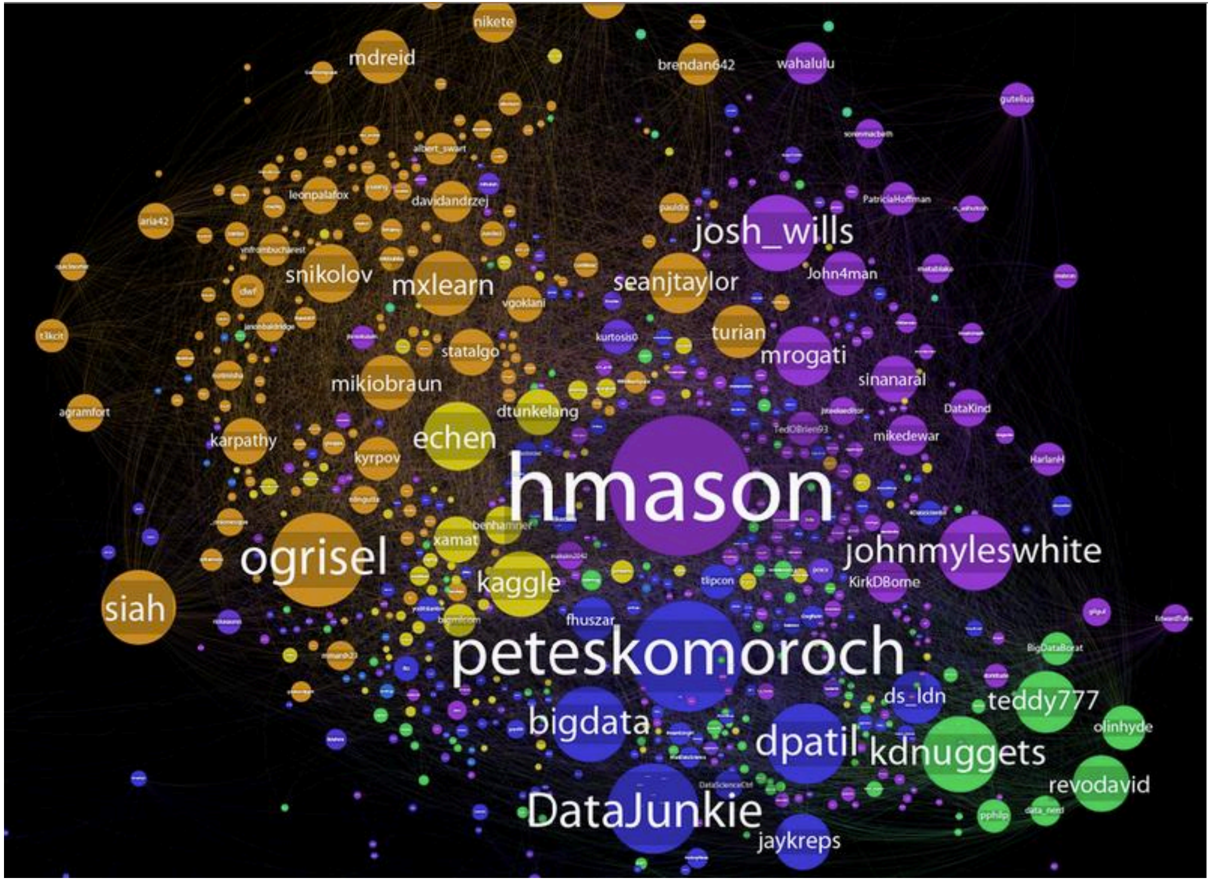
Here's an interesting question: how can you identify which (non-adopting) nodes are most important to allowing the cascade to continue?

IDENTIFYING INFLUENCERS

Here's an interesting question: how can you identify which (non-adopting) nodes are most important to allowing the cascade to continue?

Answering this question effectively is the idea behind viral marketing.

IDENTIFYING INFLUENCERS

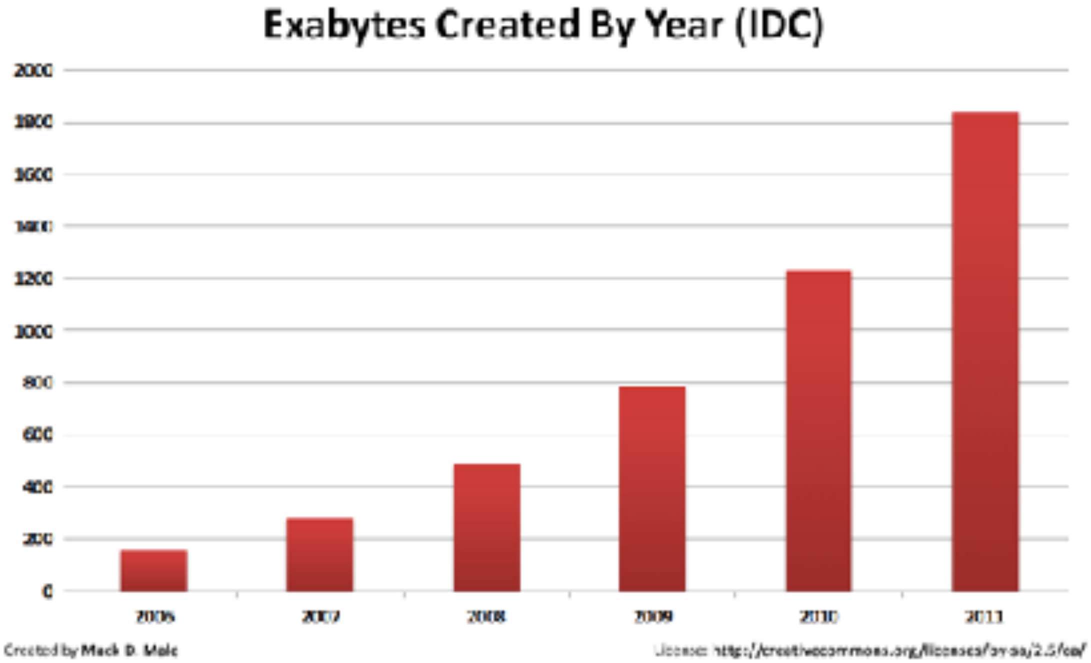


VII. FINAL THOUGHTS

REMEMBER THIS?

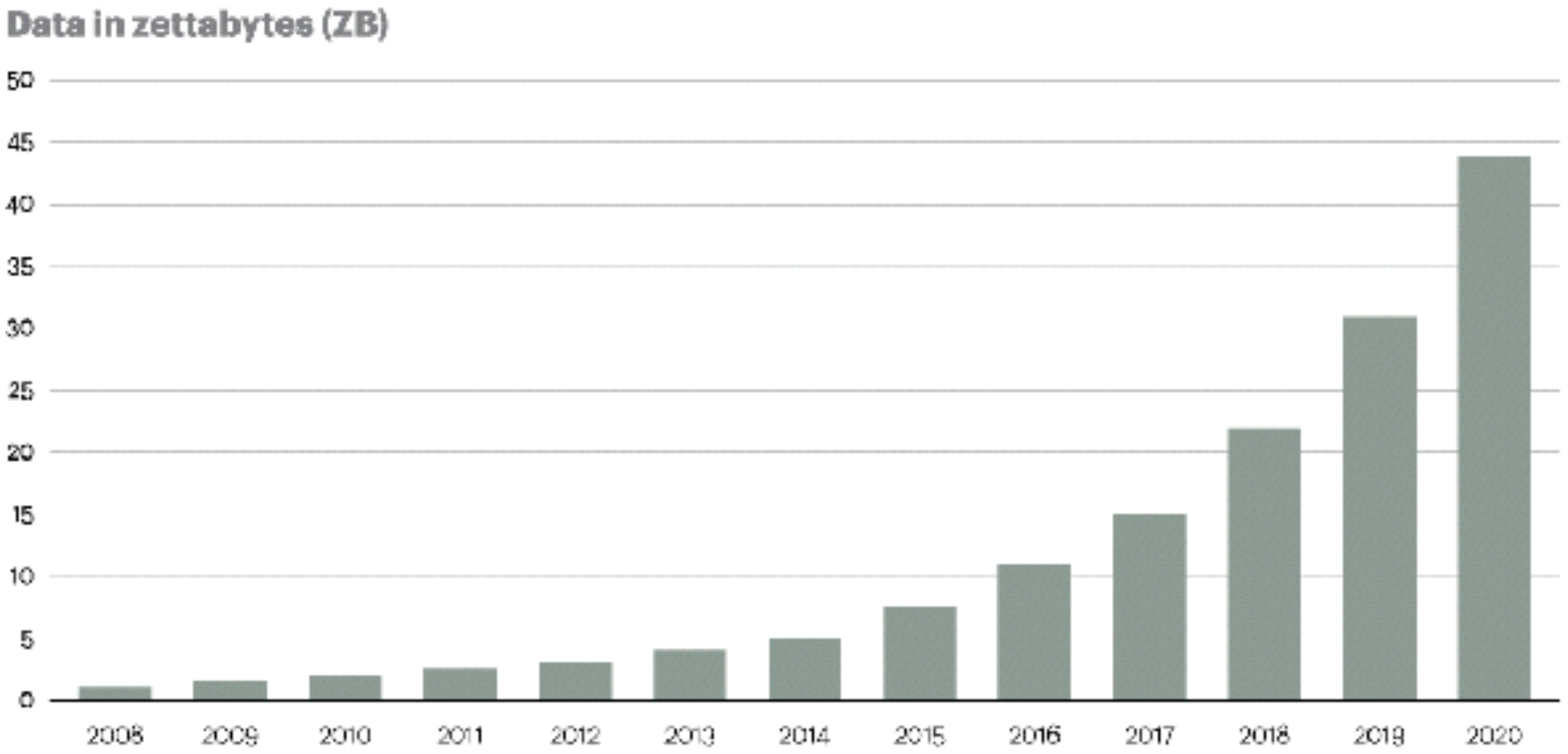
In 2011:

"Every two days we create more information than we did up until 2003 (around two exabytes)."

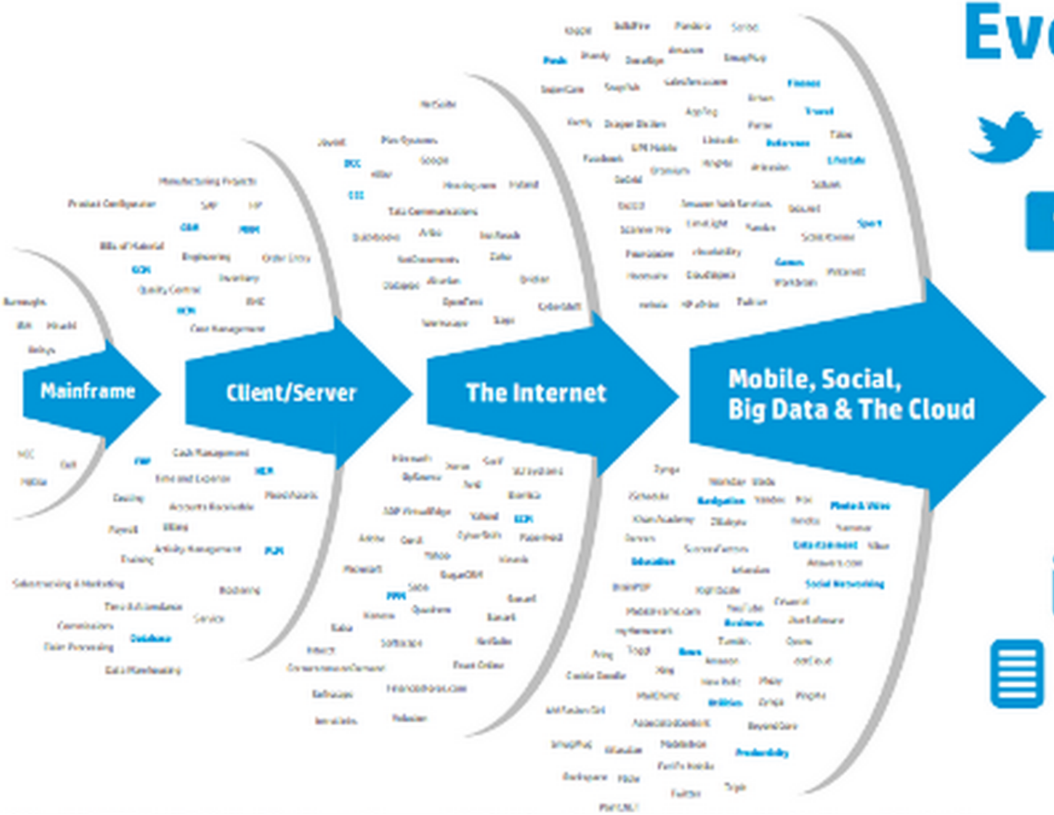


1 exabyte (EB) = 1000 petabytes (PB) = 1 billion gigabytes (GB)

ANNUAL DATA CREATION WILL HIT NEARLY 45ZB BY 2020



1 zettabyte (ZB) = 1000 exabytes (EB) = 1 trillion gigabytes (GB)



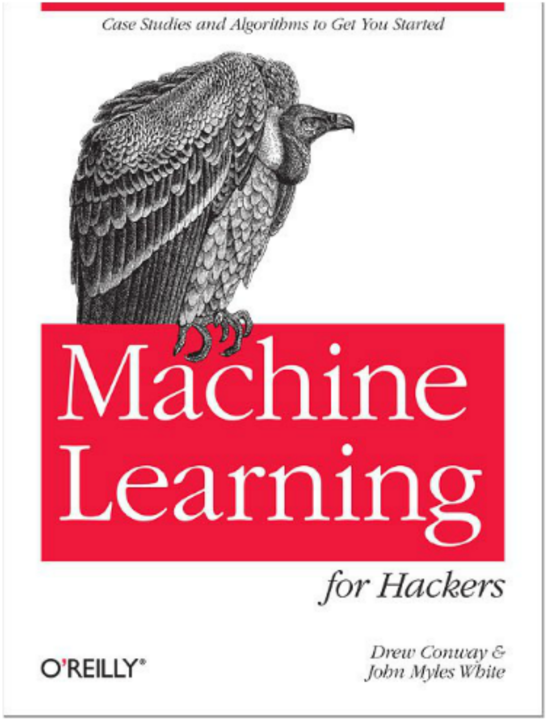
Every 60 seconds

- 98,000+ tweets
- 695,000 status updates
- 11million instant messages
- 698,445 Google searches
- 168 million+ emails sent
- 1,820TB of data created
- 217 new mobile web users

http://www.datasciencecentral.com/profiles/blogs/basic-understanding-of-big-data-what-is-this-and-how-it-is-going

LEARN MORE!

- Data Skeptic
- Partially Derivative
- Linear Digressions
- More or Less
- O'Reilly Data Show



“BECOME A DATA SCIENTIST WITH THESE 4 WEIRD TIPS”

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.

Q&A