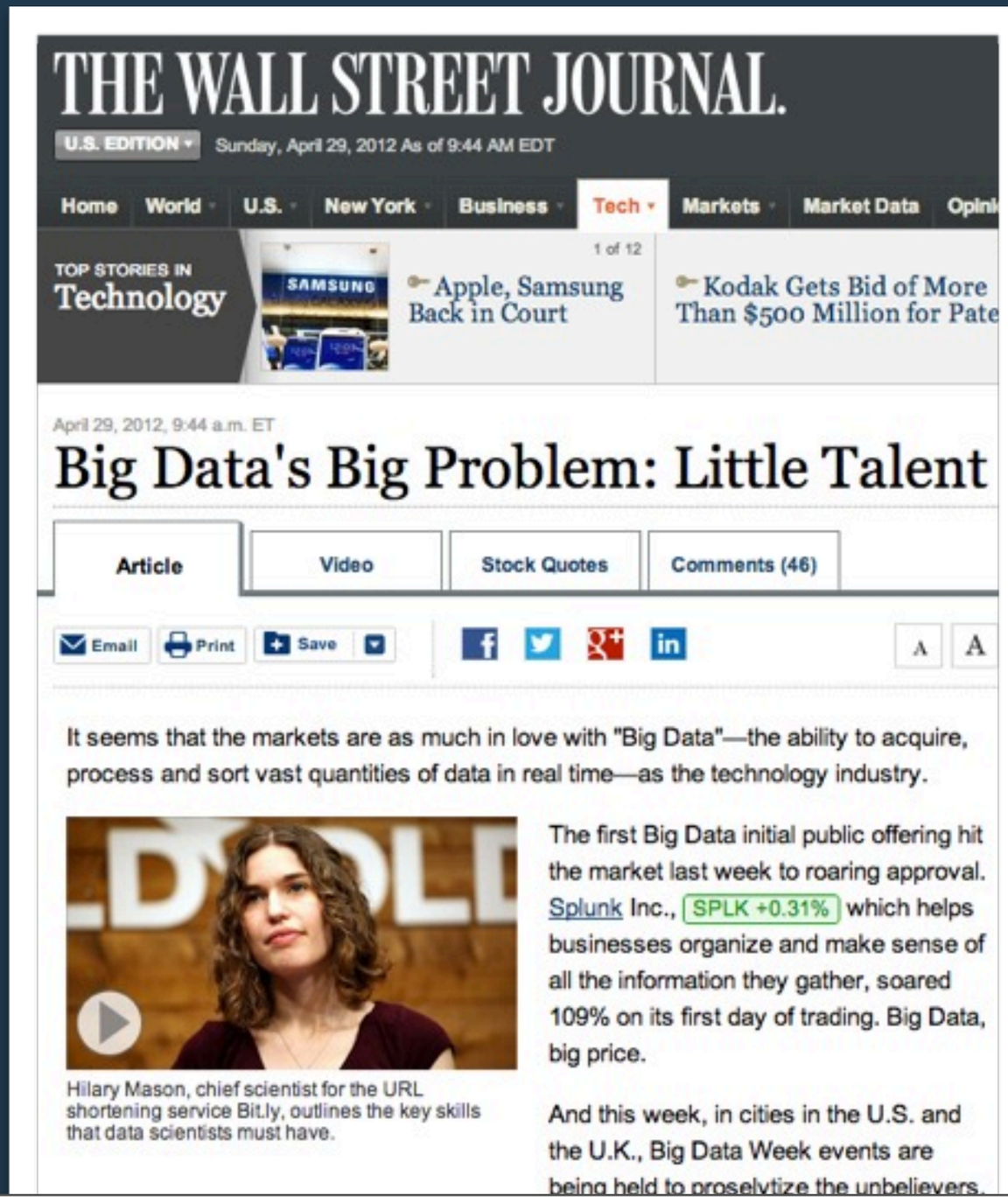
Analytics is the discovery and communication of meaningful patterns in data.

AGENDA

- Some context on big data & analytics
- What is the goal of your app?
- Event data
- Common analytics methods
- Analyze some data

Every company is becoming a **software company**.
Every **software company** is becoming a **data company**.

Big Data and **Analytics** are kind of a thing right now.



THE WALL STREET JOURNAL.

U.S. EDITION · Sunday, April 29, 2012 As of 9:44 AM EDT

Home · World · U.S. · New York · Business · Tech · Markets · Market Data · Opini

TOP STORIES IN Technology

1 of 12

Apple, Samsung Back in Court

Kodak Gets Bid of More Than $500 Million for Pate

April 29, 2012, 9:44 a.m. ET

Big Data's Big Problem: Little Talent

Article · Video · Stock Quotes · Comments (46)

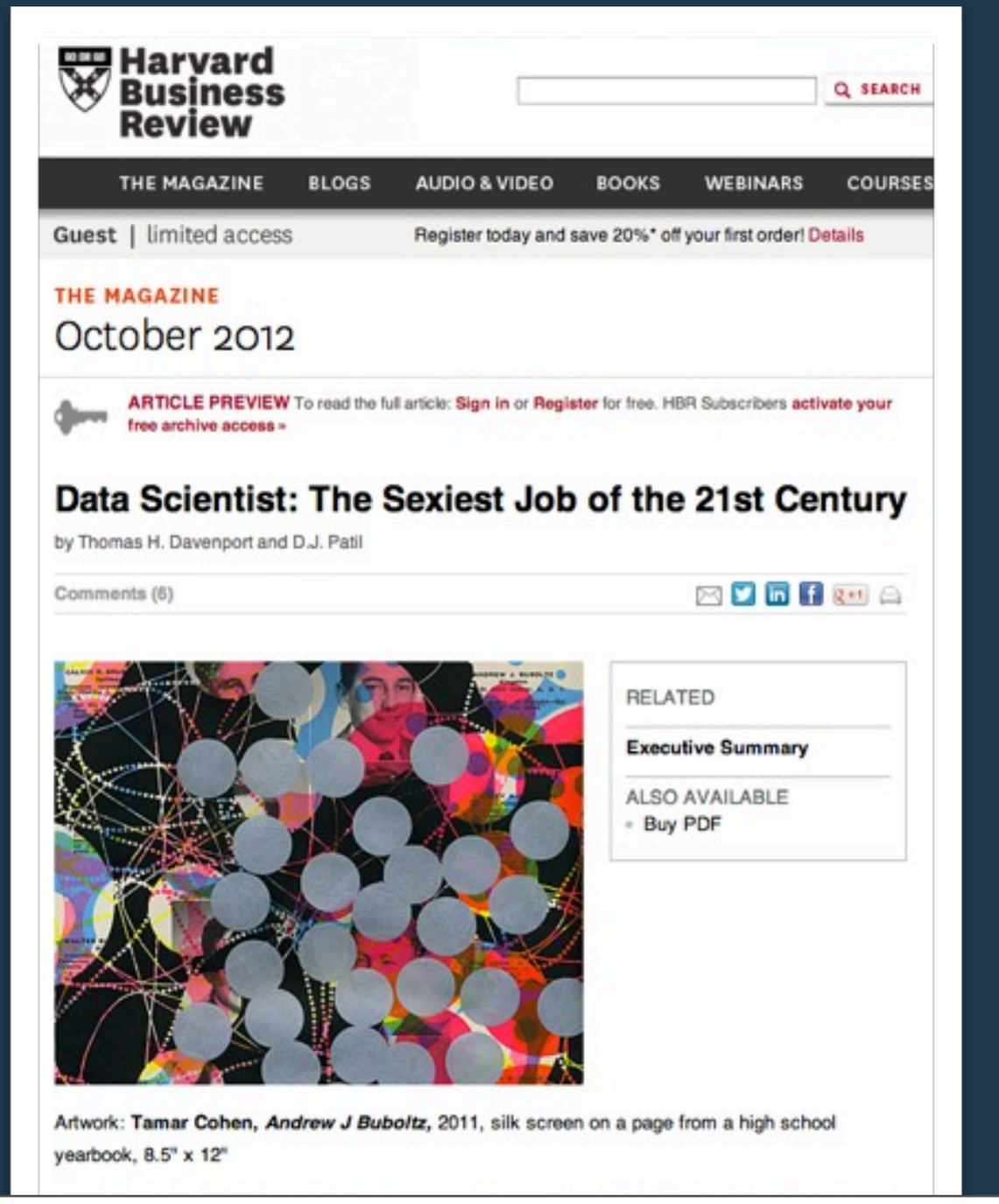
Email · Print · Save

It seems that the markets are as much in love with "Big Data"—the ability to acquire, process and sort vast quantities of data in real time—as the technology industry.

Hilary Mason, chief scientist for the URL shortening service Bit.ly, outlines the key skills that data scientists must have.

The first Big Data initial public offering hit the market last week to roaring approval. Splunk Inc., SPLK +0.31% which helps businesses organize and make sense of all the information they gather, soared 109% on its first day of trading. Big Data, big price.

And this week, in cities in the U.S. and the U.K., Big Data Week events are being held to proselytize the unbelievers.



Harvard Business Review

SEARCH

THE MAGAZINE · BLOGS · AUDIO & VIDEO · BOOKS · WEBINARS · COURSES

Guest | limited access

Register today and save 20%* off your first order! Details

THE MAGAZINE

October 2012

ARTICLE PREVIEW To read the full article: Sign in or Register for free. HBR Subscribers activate your free archive access »

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (6)

RELATED

Executive Summary

ALSO AVAILABLE

- Buy PDF

Artwork: **Tamar Cohen,** ***Andrew J Buboltz,*** 2011, silk screen on a page from a high school yearbook, 8.5" x 12"

Wednesday, July 17, 13

http://online.wsj.com/article/SB10001424052702304723304577365700368073674.html

http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

COOL DATA STORIES

Wednesday, July 17, 13

There are entire professions and really amazing work happening in different areas of analytics.

I have some stories to illustrate work in each of these realms.

- Data modeling - Brahe & Kepler
- Data mining/analysis - Linked In
- Data viz - Infosthetics
- Communication - Broad Street Pump

Tycho Brahe

Johannes Kepler

Wednesday, July 17, 13

Story 1: Data Modeling - what data should we record and how should we record it?

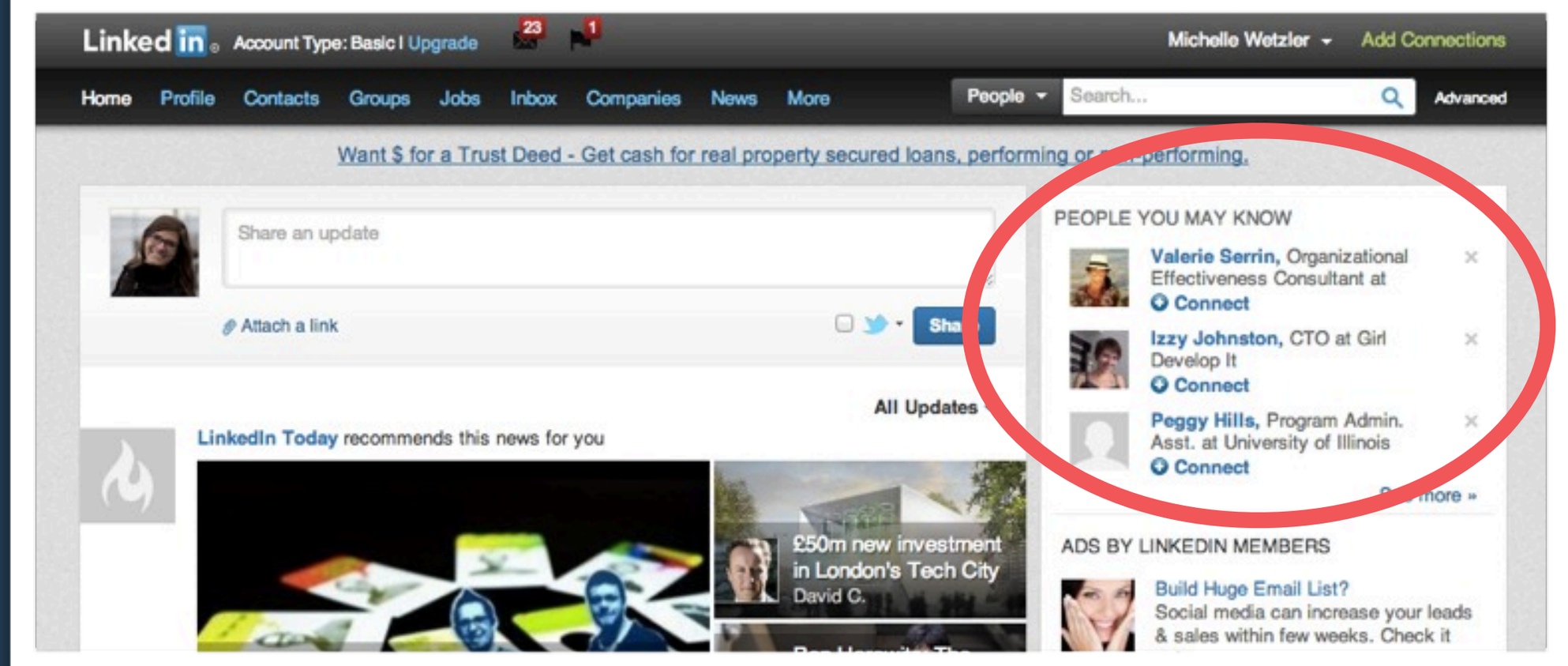
Tycho Brahe - collected an astounding amount of astronomical data. Every night, he would write down the location of every star and every planet in the sky. After 30 years of doing this... he died.
Johannes Kepler - took Brahe's data and single-handedly discovered the laws of planetary motion.

The point: recording stuff is important!

Wednesday, July 17, 13

Second point: sometimes you don’t know how the data you are collecting will be used.

Other cool data collection: quantified self movement, 23andme, sensor data on vehicles,



Wednesday, July 17, 13

Story 2: Data Analysis - what can we learn from data? what tools and algorithms can we use to unlock new meaning?

- Goldman, a PhD in physics from Stanford, was intrigued by the linking activity he saw happening on linkedIn. He began exploring people's connections, forming theories, testing hunches, and finding patterns that allowed him to predict whose networks a given profile would land in.

- He could imagine that new features capitalizing on the heuristics he was developing might provide value to users. But LinkedIn's engineering team, caught up in the challenges of scaling up the site, seemed uninterested. Some colleagues were openly dismissive of Goldman's ideas. Why would users need LinkedIn to figure out their networks for them? The site already had an address book importer that could pull in all a member's connections.

- LinkedIn's cofounder and CEO at the time had faith in the power of analytics because of his experiences at PayPal, and he had granted Goldman a high degree of autonomy. For one thing, he had given Goldman a way to circumvent the traditional product release cycle by publishing small modules in the form of ads on the site's most popular pages.

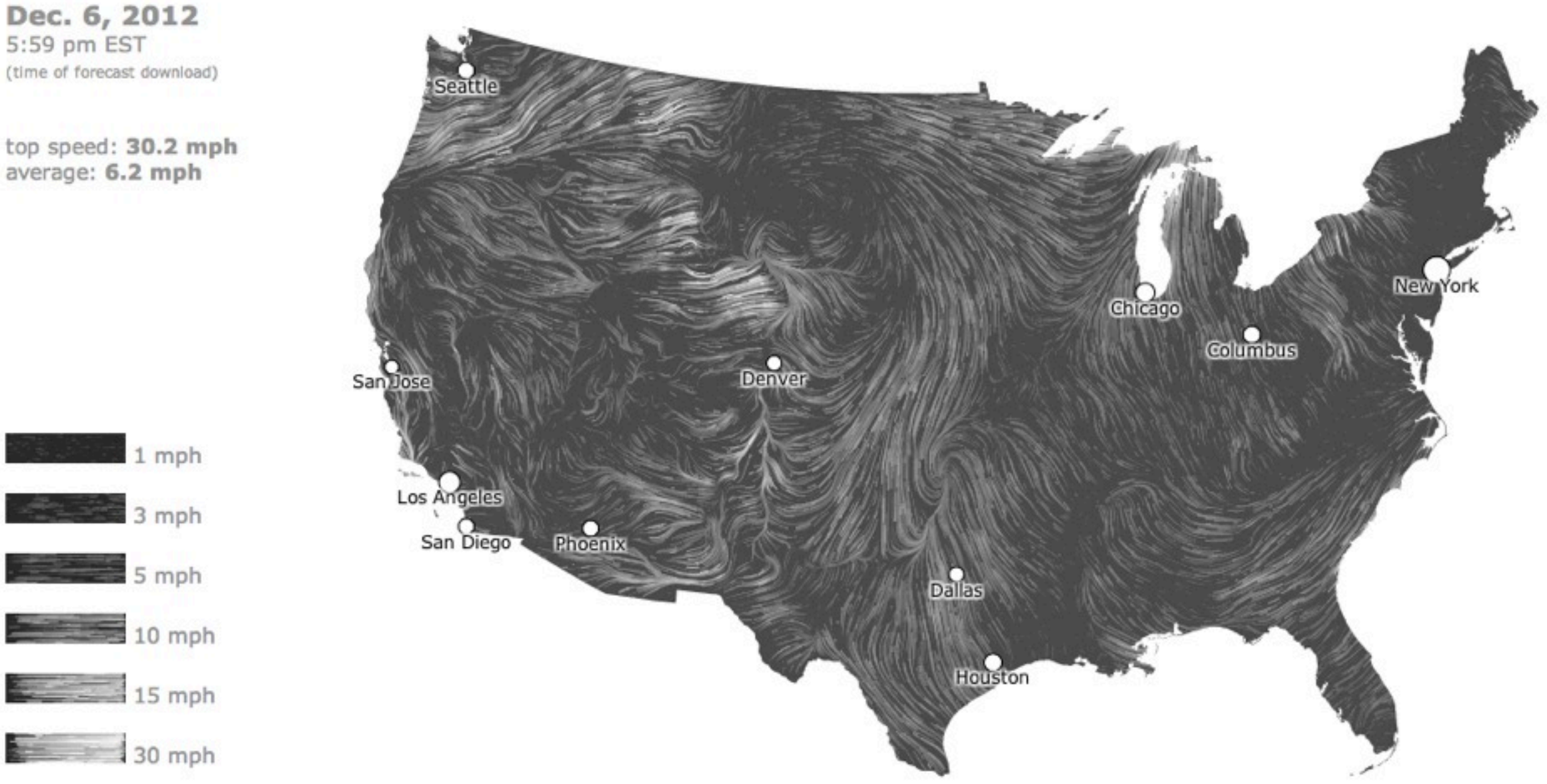
- Goldman started to test what would happen if you presented users with names of people they hadn't yet connected with but seemed likely to know—for example, people who had shared their tenures at schools and workplaces. He did this by ginning up a custom ad that displayed the three best new matches for each user based on the background entered in his or her LinkedIn profile. Within days it was obvious that something remarkable was taking place.

- The click-through rate on those ads was the highest ever seen. Goldman continued to refine how the suggestions were generated, incorporating networking ideas such as "triangle closing"—the notion that if you know Larry and Sue, there's a good chance that Larry and Sue know each other. Goldman and his team also got the action required to respond to a suggestion down to one click.

- It didn't take long for LinkedIn's top managers to recognize a good idea and make it a standard feature. That's when things really took off. "People You May Know" ads achieved a click-through rate 30% higher than the rate obtained by other prompts to visit more pages on the site. They generated millions of new page views. Thanks to this one feature, LinkedIn's growth trajectory shifted significantly upward.

The point: Analytics can be used to unlock value. It can also be used to prove value. Companies that recognize this and devote some effort to it can potentially reap huge benefits.

source: http://online.wsj.com/article/SB10001424052702304723304577365700368073674.html



http://hint.fm/wind

http://infosthetics.com

<click link for animation>

Story 3: Data Visualization - how can we use visuals to discover trends?
Sometimes visuals can tell us things we can't see from raw data.
There are some amazing projects out there (see infosthetics)

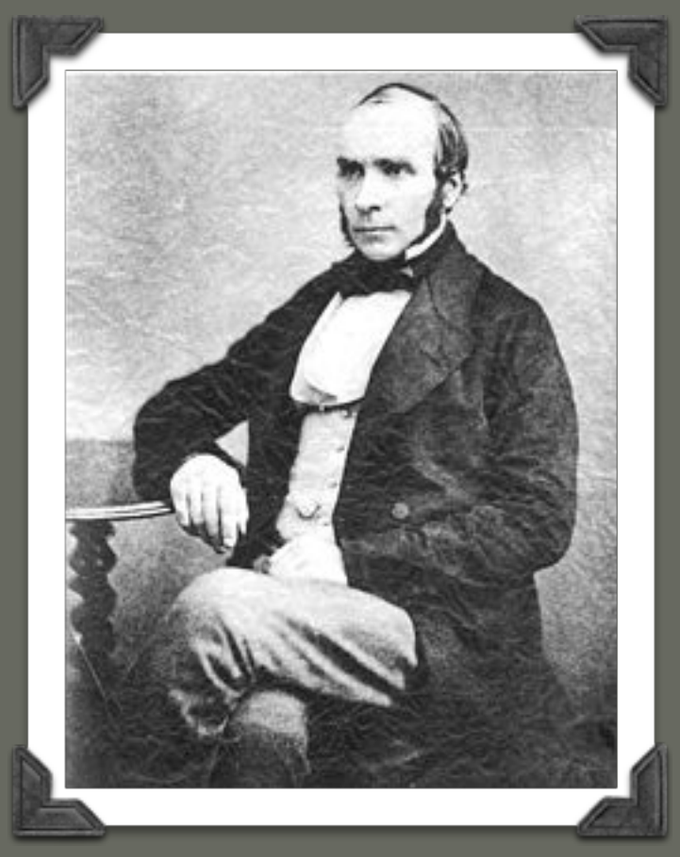
http://hint.fm/wind/
http://infosthetics.com/

LE CHOLÉRA

Story 4: Data Communication - How do we share what we have learned with others? What does the data tell us (and what does it leave out)? One critical skill required of a data analyst is how to communicate what the results mean and what we should do next. The following is the story of how data sampling and visualization were used to communicate something very important.

Cholera hit london in 1854. Waste management systems were really bad and consisted of vats in people's basements.
In 3 days, 127 people near Broad Street Died.
The mortality rate was 12.8 percent in some parts of the city.
By the end of the outbreak, 616 people had died.

http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

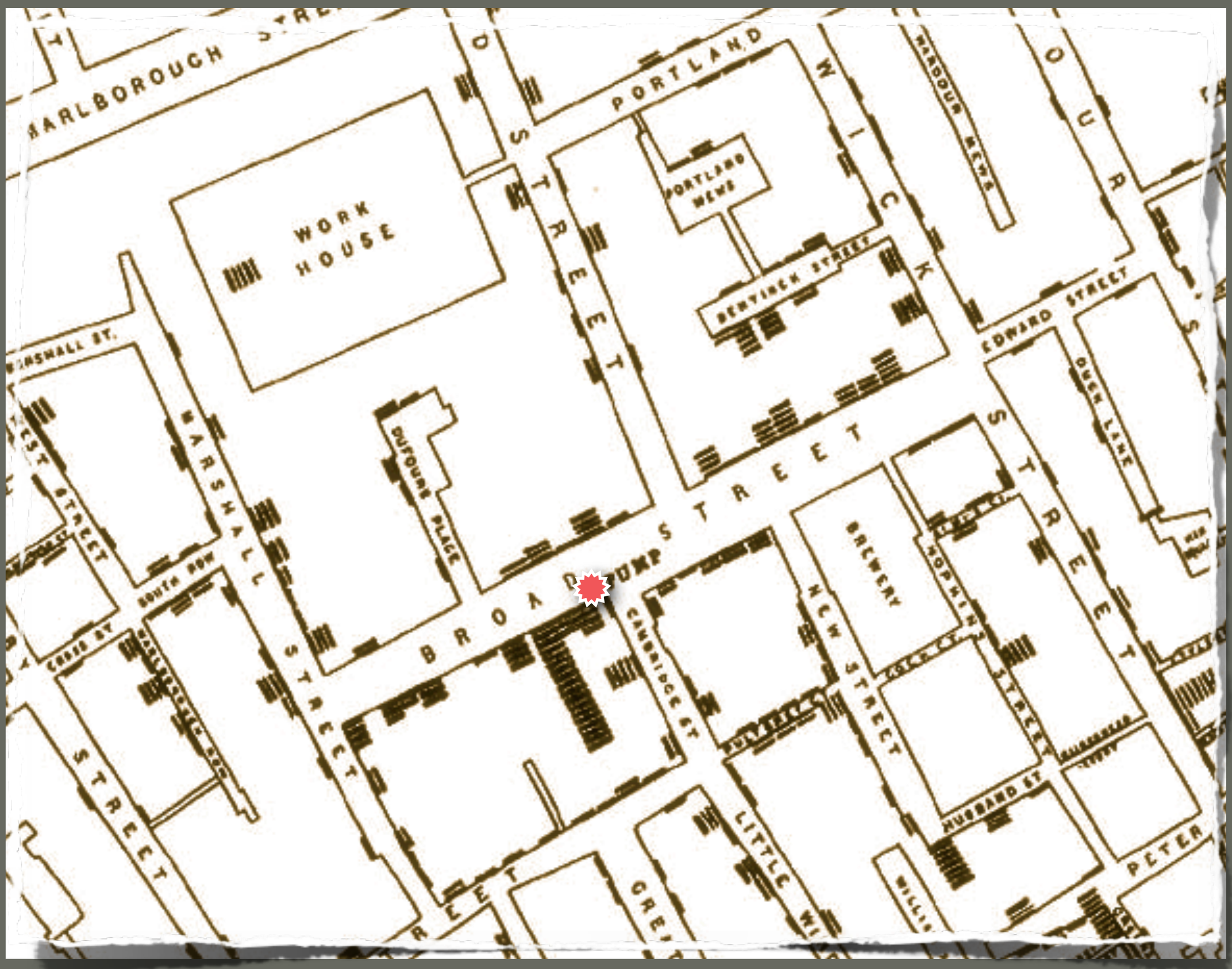


John Snow figured out Cholera spreads through water.

No one believed him :(

Wednesday, July 17, 13

John Snow hypothesized that Cholera was being spread through the water supply, but no one believed him.



Wednesday, July 17, 13

Until he created this awesome image.
He counted deaths in each building and used this data visualization to prove his theory.
Little black bars are deaths.
The broad street pump was shut down and deaths stopped.

APPLYING ANALYTICS TO YOUR BUSINESS

Wednesday, July 17, 13

Use analytics to measure progress toward a goal.

Use analytics to test new hypotheses.

Use analytics to explore.

Avoid using analytics just to be using analytics.

EXERCISE!

1 MINUTE

WHAT IS YOUR GOAL?

Examples:

Vine: reach 1M user-generated videos.

Spotify: increase conversions to paying subscriptions.

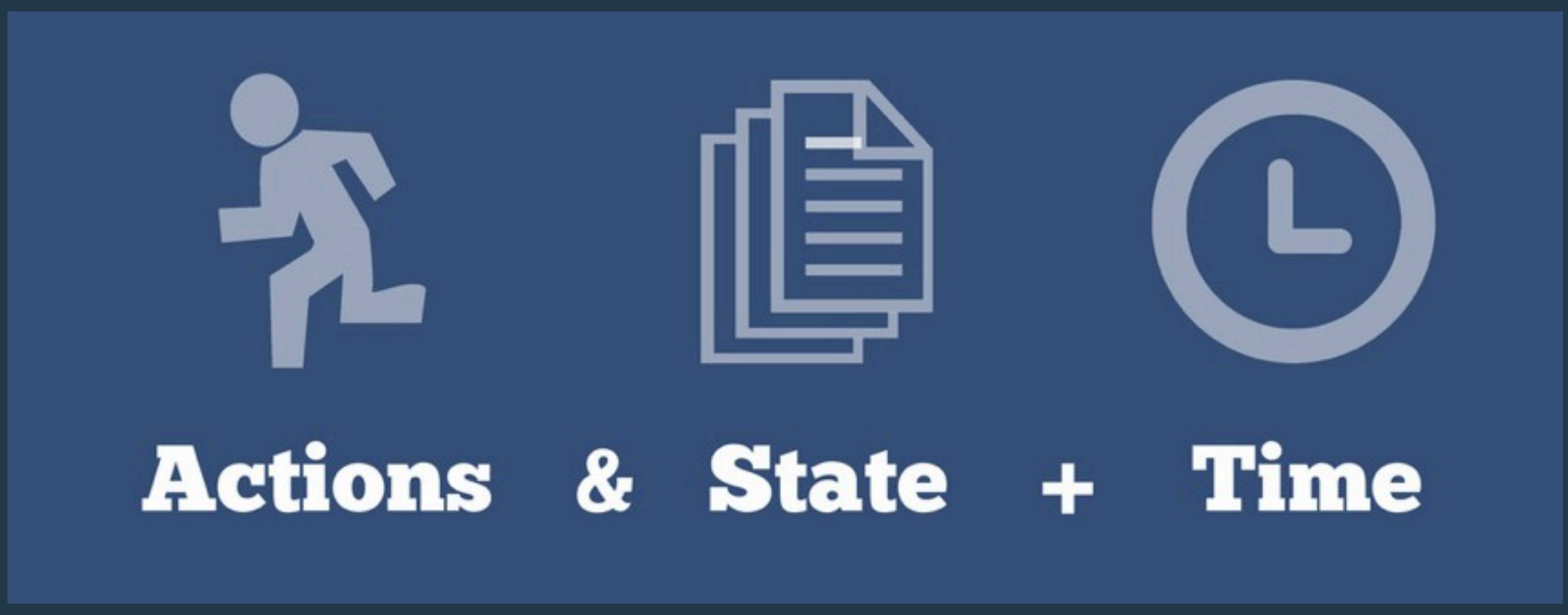
Wednesday, July 17, 13

Think of a goal for the product you are working on right now. Be prepared to share it with the class.
Example goal for Vine: reach 1M user-generated videos.
Example goal for Spotify: increase subscription revenue
Example goal of an enterprise contacts app: make it easier for colleagues to contact each other.

A COMMON GOAL: ENGAGEMENT

- Account creations
- Deploys
- Purchases
- App Launches
- Views
- Posts
- Shares/Tweets/Likes

INTRODUCING EVENT DATA



Wednesday, July 17, 13

All of the actions on the previous page would be considered event type data.
Event data isn't new. You've probably used it before in the form of logs or massive, frequently archived tables.
Event data is different than state data, because events happen A LOT, and they don't easily fit into your schema'd database.
Events are actions that happen all the time, like an app launch, share, post, view, or delete.

Certain types of databases and a lot of new, modern analytics tools allow you do to powerful analysis with event data.

Ref: https://speakerdeck.com/benbjohnson/behavioral-databases

UID	twitter handle	age	Account ID
773345	@hipsterhacker	29	443556
773346	@TNG_S8	27	432354
773347	@modernseinfeld	28	336658
773348	@michellewetzler	28	2115789

It's easier to describe event data if we start out by comparing it to entity data.
Entity data is the type of data we're most familiar with.

Entity data captures the current state of things in our database. We have a user table, a products table, a photos table, etc.

Most databases have historically been designed for entity-type data.

```
{
  "event": "death",
  "timestamp": "2013-05-23T1:50:00-0600",
  "cause": "creeper explosion",
  "enemy": {
    "type": "creeper",
    "power": .887,
    "distance_from_player": 3.43,
    "age": .6677,
  },
  "player": {
    "UID": "99234890823",
    "experience": 8873729,
    "age": 338,
    "inventory": [“diamond sword”, “torches”]
  }
}
```


Here’s another example.
Notice a few things
- nested
- denormalized
- rich
- dynamic -- see enemy example --- in other words the schema is really flexible

Here’s another example.
Notice a few things
- nested
- denormalized
- rich
- dynamic -- see enemy example --- in other words the schema is really flexible

Ok. So you get what entity data is.
But so far I've only shown you boring examples!! What I want you to take away from this talk is that there is tons of interesting applications and opportunities for event data.

Ok. So you get what entity data is.

But so far I've only shown you boring examples!! What I want you to take away from this talk is that there is tons of interesting applications and opportunities for event data.

entity data	event data
strict schema	flexible schema

Ok. So you get what entity data is.
But so far I've only shown you boring examples!! What I want you to take away from this talk is that there is tons of interesting applications and opportunities for event data.

entity data	event data
strict schema	flexible schema
normalized	denormalized

Ok. So you get what entity data is.
But so far I've only shown you boring examples!! What I want you to take away from this talk is that there is tons of interesting applications and opportunities for event data.

entity data	event data
strict schema	flexible schema
normalized	denormalized
shorter	wider

Ok. So you get what entity data is.
But so far I've only shown you boring examples!! What I want you to take away from this talk is that there is tons of interesting applications and opportunities for event data.

entity data	event data
strict schema	flexible schema
normalized	denormalized
shorter	wider
describes nouns	describes verbs

Ok. So you get what entity data is.
But so far I've only shown you boring examples!! What I want you to take away from this talk is that there is tons of interesting applications and opportunities for event data.

entity data	event data
strict schema	flexible schema
normalized	denormalized
shorter	wider
describes nouns	describes verbs
describes now	describes trends over time

Ok. So you get what entity data is.
But so far I've only shown you boring examples!! What I want you to take away from this talk is that there is tons of interesting applications and opportunities for event data.

entity data	event data
strict schema	flexible schema
normalized	denormalized
shorter	wider
describes nouns	describes verbs
describes now	describes trends over time
updates	appends

Ok. So you get what entity data is.
But so far I've only shown you boring examples!! What I want you to take away from this talk is that there is tons of interesting applications and opportunities for event data.

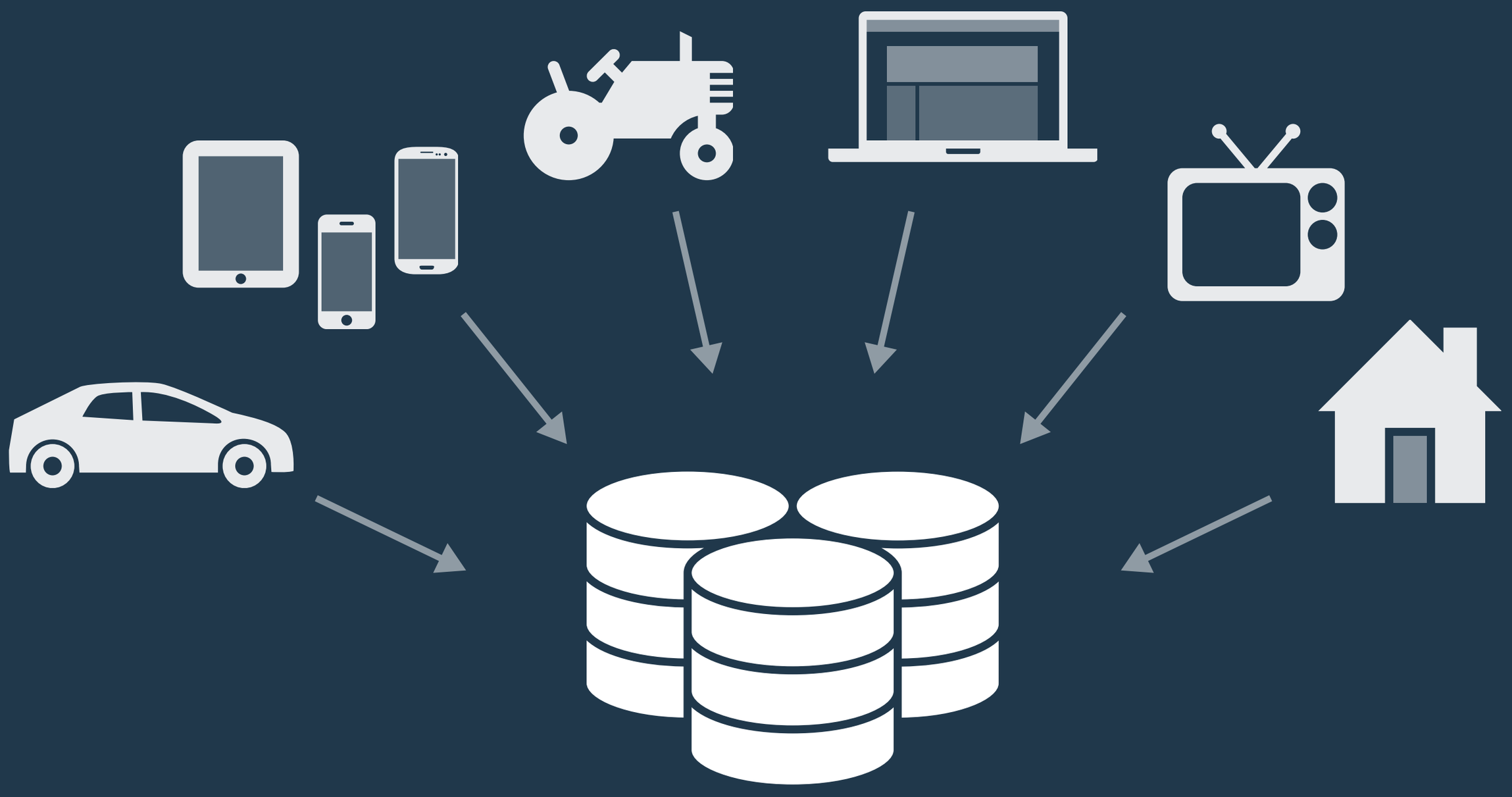
entity data	event data
strict schema	flexible schema
normalized	denormalized
shorter	wider
describes nouns	describes verbs
describes now	describes trends over time
updates	appends
big data	big big big data

Ok. So you get what entity data is.
But so far I've only shown you boring examples!! What I want you to take away from this talk is that there is tons of interesting applications and opportunities for event data.

In fact events are happening everywhere... mobile example, web example, fitbit example. Now we can collect event data from all sorts of apps and devices. Billions of events happening all the time and we finally have the technology to start capturing them.

We're really just now at a point where storage is cheap enough that we can record pretty much everything.

What is cool about this?



In fact events are happening everywhere... mobile example, web example, fitbit example. Now we can collect event data from all sorts of apps and devices. Billions of events happening all the time and we finally have the technology to start capturing them.

We’re really just now at a point where storage is cheap enough that we can record pretty much everything.

What is cool about this?

MORE EXAMPLES

VERBS: Signup, Login, Upgrade, Submit, Scroll, Send, Share, Search, Check-In, Vote, Update, Purchase, Level Up, Fail, Favorite, Vote, Crash, Rate, Start, Modify, Check, View, Capture

NOUNS: User, Company, Organization, Team, Platform, Device, App, Level, Garden, Favorites, Interests, Inventory, Cart, Video, Location, Item, Record, Product, Account, Form, Picture, Story

ELEVATOR PITCH TECHNIQUE

- Describe your app to a stranger and listen to the words you use.
- Verbs are the actions you should record.
- Nouns are the important contextual information you should include in your data model.
- Most apps can be very robustly described by 5-10 key events and 5-10 key nouns.

Sometimes it seems like EVERYTHING could be recorded. How do you know what to record?

EXERCISE!

2 MINUTES

Recall your goal from the previous exercise.

What actions & properties should you record to measure your progress toward your goal?

Example goal for Vine: reach 1M user-generated videos. Track the number of videos being generated.
Example goal for Spotify: increase subscription revenue. Track the revenue from subscriptions.
Example goal of an enterprise contacts app: make it easier for colleagues to contact each other. Track app launches, contact searches, calls.

COMMON ANALYTICS TECHNIQUES

Wednesday, July 17, 13

95% of analytics involves
what mathematical operation?

How many users signed up today? Downloaded our app?
How many people played our game this week?
Is the count increasing or decreasing over time?

95% of analytics involves
what mathematical operation?

COUNTING!

How many users signed up today? Downloaded our app?
How many people played our game this week?
Is the count increasing or decreasing over time?

MORE BASICS

- Sum
- Average
- Min
- Max
- Division

EXAMPLE

What was the average revenue per active user last month?

EXAMPLE

What was the average revenue per active user last month?

1. **Count** the number of unique users who performed some action in June (2300)

2. **Sum** all of the purchases from last month ($5564)

3. Divide 2 by 1 ($2.40)

ADVANCED

- Statistical Analysis
- Correlation Analysis
- Predictive Analysis

Fancy Terms for Counting Stuff

Wednesday, July 17, 13

SEGMENTATION

- Sorting data into buckets. Commonly used to sort users or products into categories.
- Examples: Gender, Age, Location, Department, Referrer, Version

	Country / Territory	Visits	% Visits
1.	United States	33,399	59.17%
2.	Canada	3,578	6.34%
3.	United Kingdom	3,244	5.75%
4.	Australia	1,588	2.81%
5.	India	1,384	2.45%
6.	Germany	1,183	2.10%
7.	(not set)	1,153	2.04%
8.	France	838	1.48%
9.	Netherlands	697	1.23%

Segmentation example: Say you are graphing revenue over time. Segment by product category to see which categories have the biggest sales.
Filtering example: Use filtering to exclude events with certain properties. E.g. only look at events above a certain price, or users in a certain category.

FILTERING

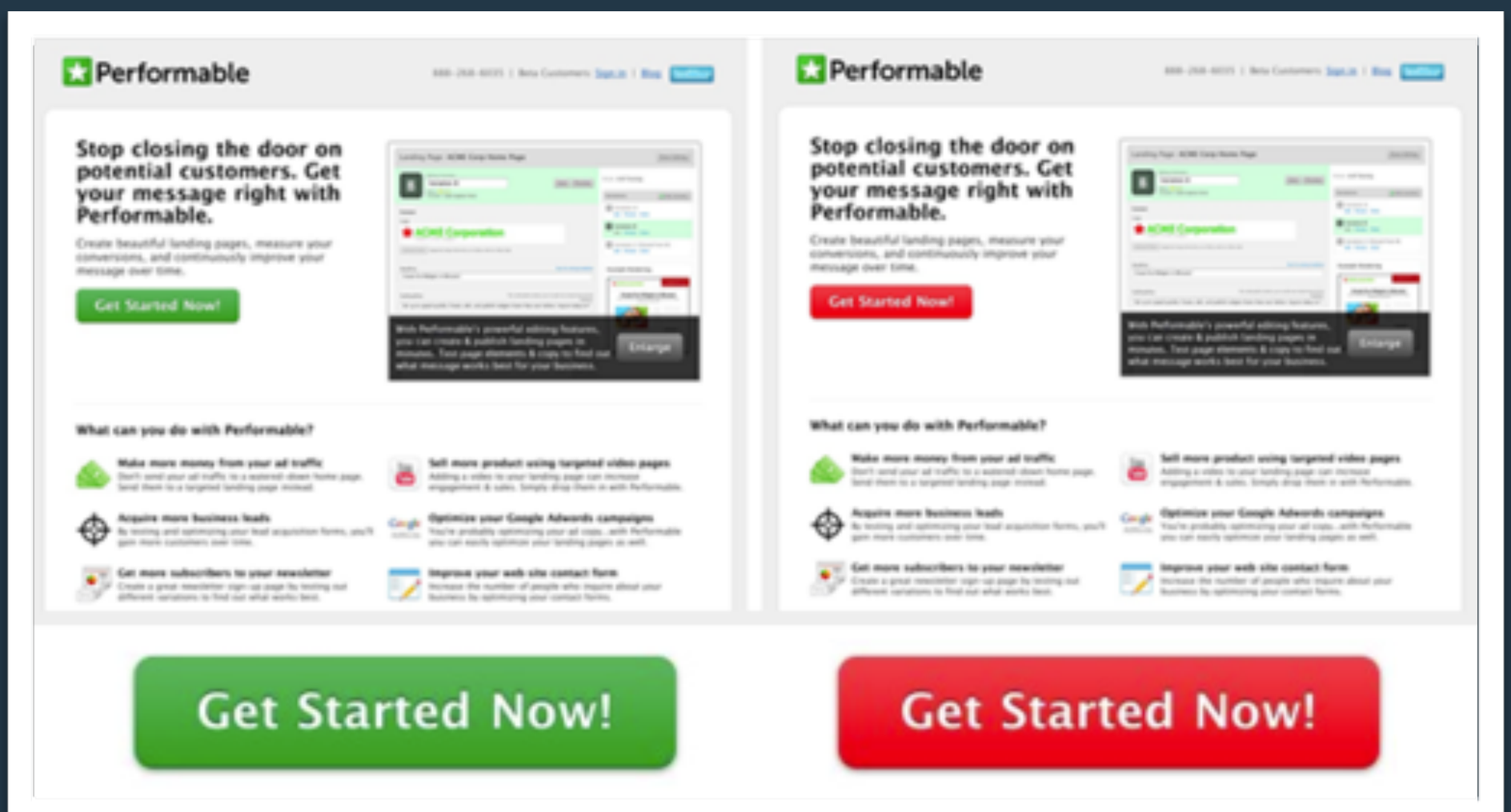
- Translation: only count stuff that meets these criteria.
- Use any of your event properties to do filtering.
- Example: Count the number of purchases events where item.category = “add-ons” and item.price > $100.

EXERCISE!

Recall the event you modeled in the previous exercise.

Think of at least one property you would like to use for filtering or segmentation.

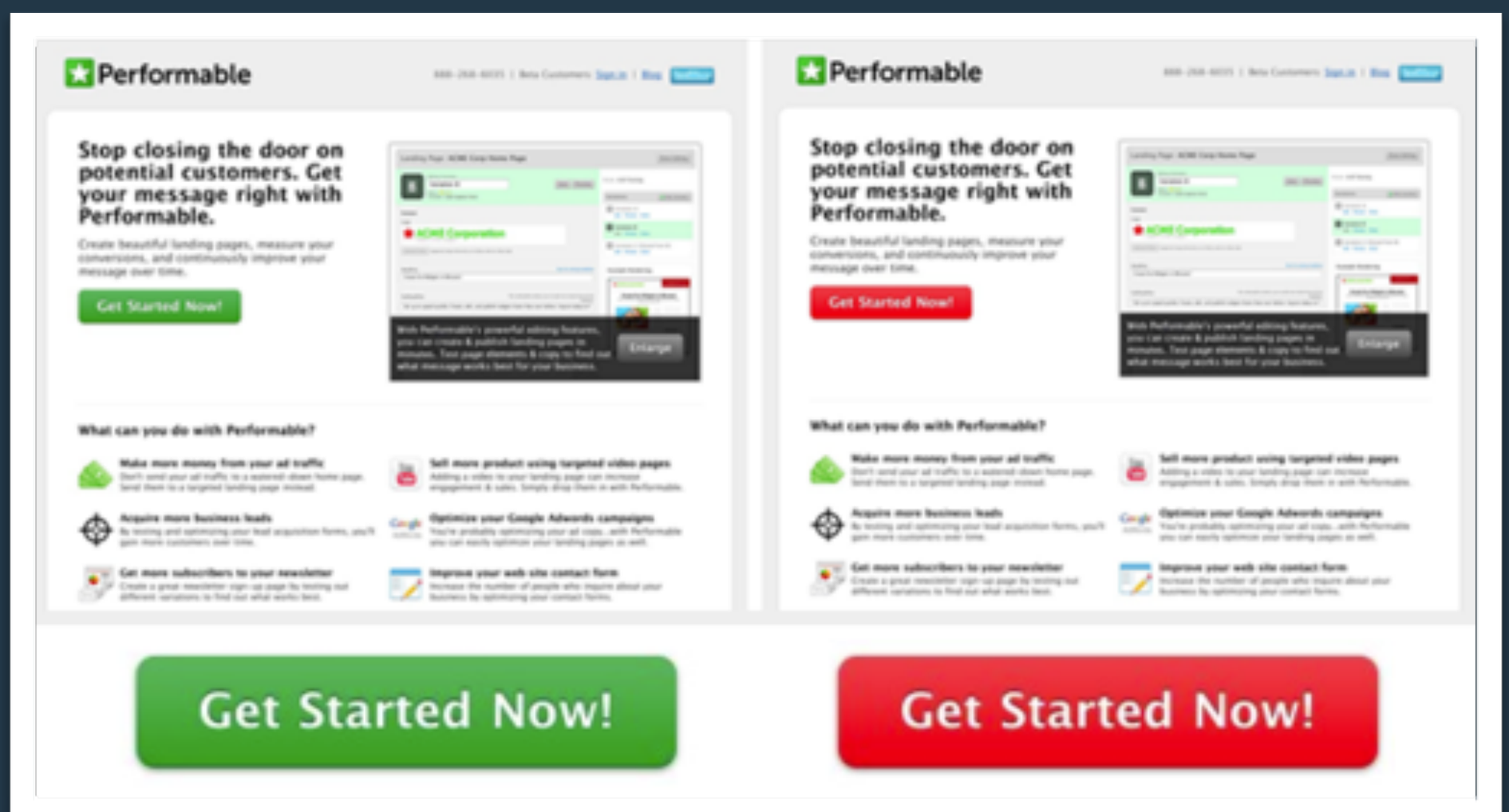
A/B TESTING AKA SPLIT TESTING



http://en.wikipedia.org/wiki/A/B_testing

- can be applied to even the smallest things (e.g. button color)
- used extensively in advertising & marketing
- used to get “the last 20%” in app optimization
- tells you which option is better, but won’t tell you if both of them suck!
- some tools now automate this and will automatically serve the most popular version of site

A/B TESTING AKA SPLIT TESTING



21% more people clicked on the red button than on the green button!

http://en.wikipedia.org/wiki/A/B_testing

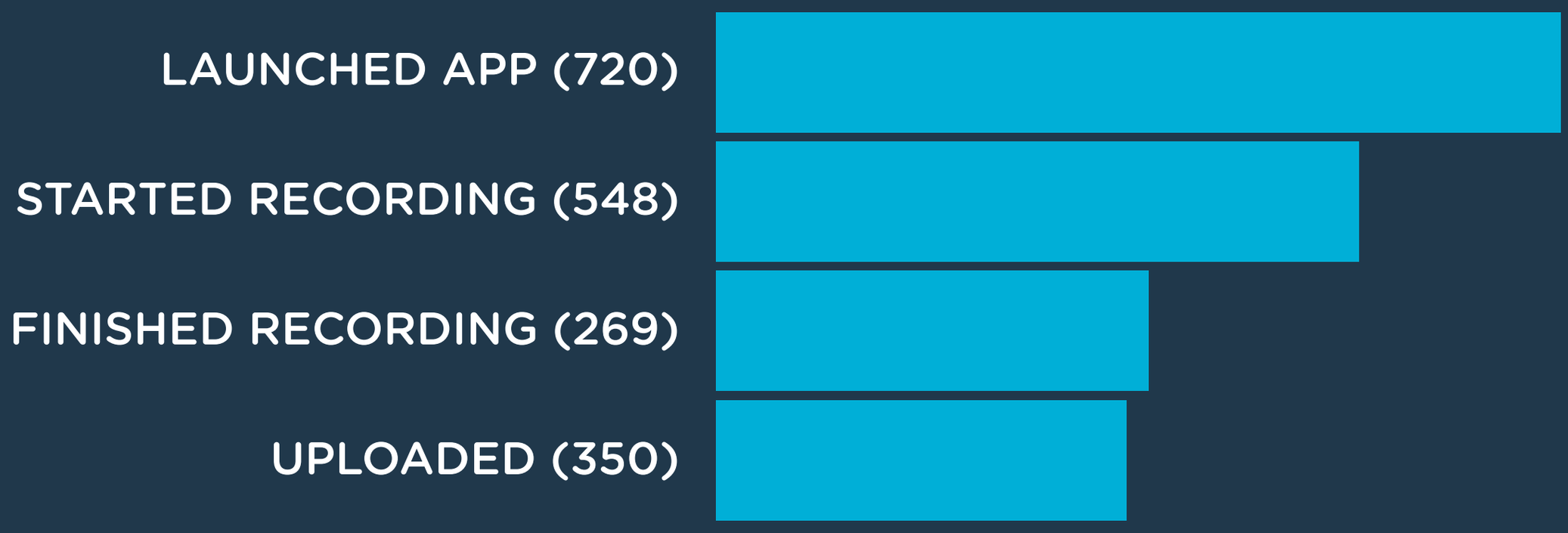
- can be applied to even the smallest things (e.g. button color)
- used extensively in advertising & marketing
- used to get “the last 20%” in app optimization
- tells you which option is better, but won’t tell you if both of them suck!
- some tools now automate this and will automatically serve the most popular version of site

EXAMPLE OF SPLIT TESTING DATA

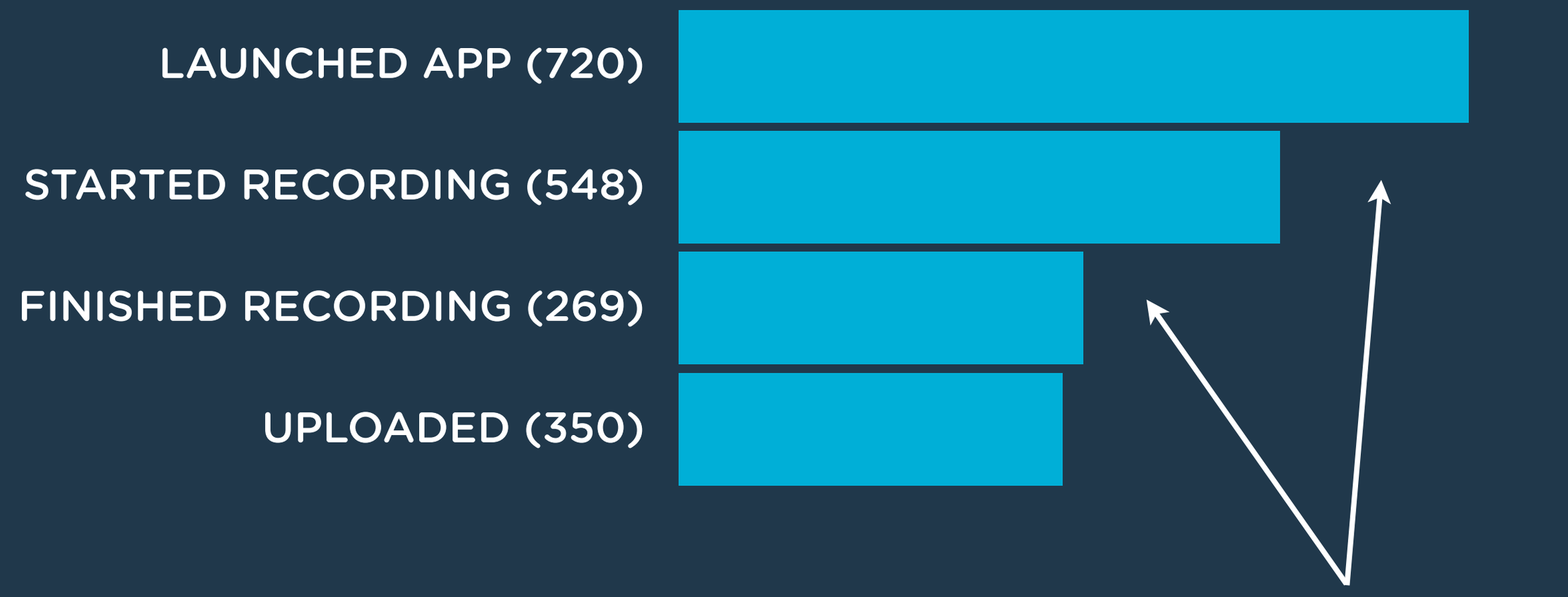
user.id	user.name.first	user.name.last	form.version	form.fields
223655	zach	morris	A	[first name, middle name, last name, organization, gender, age, email, password
223656	kelly	kapowski	B	[email, password]
223657	screech	powers	A	[first name, middle name, last name, organization, gender, age, email, password
223658	lisa	turtle	A	[first name, middle name, last name, organization, gender, age, email, password
223659	ac	slater	B	[email, password]
223660	jessie	spano	B	[email, password]
223661	mr.	belding	B	[email, password]
223662	mrs.	culpepper	B	[email, password]
223663	stacey	carosi	A	[first name, middle name, last name, organization, gender, age, email, password
223664	allison	fox	B	[email, password]
223665	tori	scott	B	[email, password]
223666	mr.	dewey	B	[email, password]
223667	ollie	creeky	B	[email, password]
223668	violet	bickerstaff	B	[email, password]
223669	rhonda	robistelli	B	[email, password]

Which version of the form was more effective?

FUNNELS

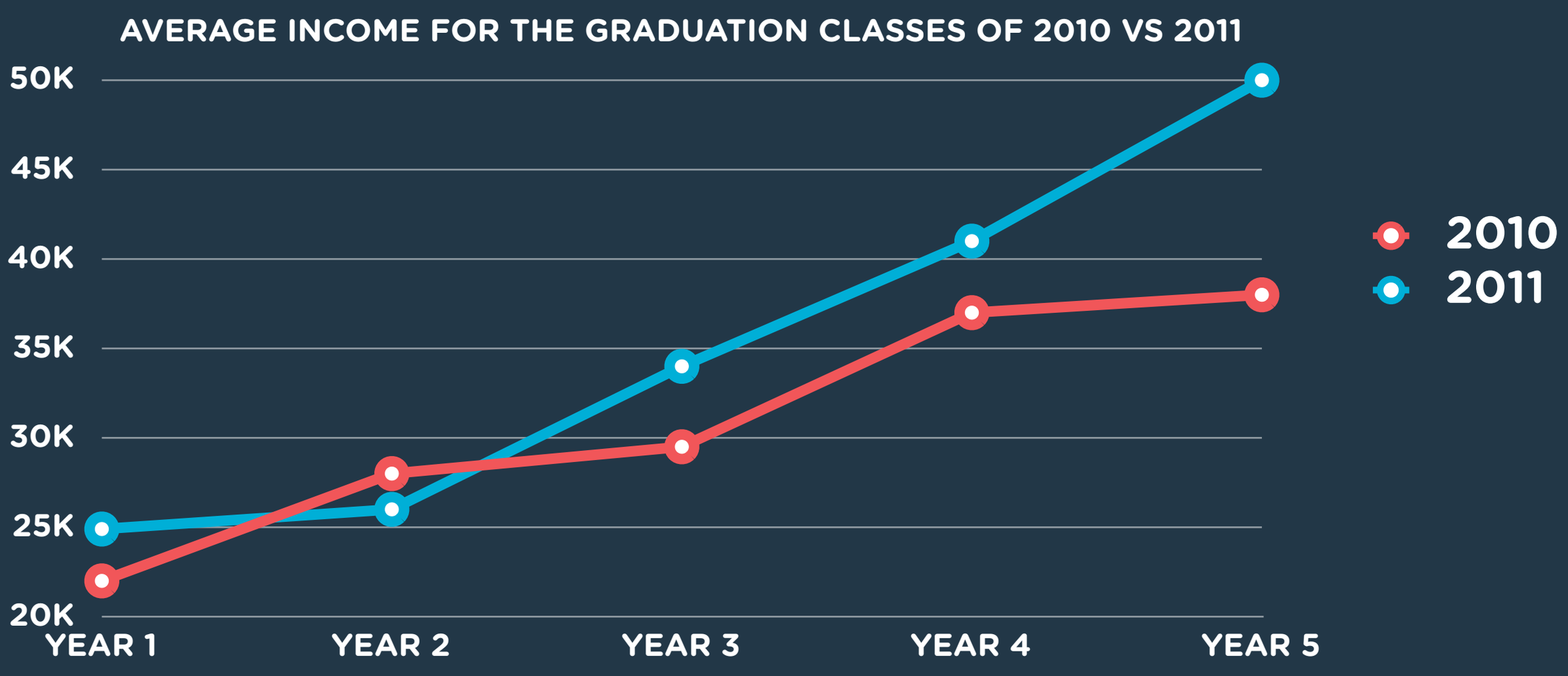

LAUNCHED APP (720)
STARTED RECORDING (548)
FINISHED RECORDING (269)
UPLOADED (350)

FUNNELS


LAUNCHED APP (720)
STARTED RECORDING (548)
FINISHED RECORDING (269)
UPLOADED (350)
opportunity!

COHORT ANALYSIS

A cohort is a group of people who share a common characteristic over a certain period of time.



RETENTION

How many customers remain customers?

How many users came back a second time?

Do my customers value my product?

Measure retention by counting how many users did an action X days after their first usage.

RETENTION ANALYSIS BY COHORT

Date	People	Weeks Later 1	2	3	4	5	6	7	8	9	10	11	12
Nov 19	5,768	38.62%	46.21%	43.17%	40.09%	37.94%	35.59%	32.16%	27.82%	23.43%	18.55%	13.73%	8.64%
Nov 26	6,984	47.01%	44.19%	41.76%	38.82%	32.09%	29.73%	33.72%	27.94%	25.55%	20.71%	16.73%	7.82%
Dec 3	6,172	44.36%	41.66%	36.52%	32.2		26.03%	24.44%	21.48%	18.63%	12.57%	6.56%	1.72%
Dec 10	5,930	42.32%	38.98%	31.44%				19.74%	15.43%	11.68%	9.81%	2.36%	
Dec 17	4,784	36.74%	31.21%	26.7				19.39%	12.44%	10.06%	3.42%		
Dec 24	5,836	33.10%	29.67%	25.				10.08%	5.43%	0.97%			
Dec 31	4,236	31.55%	26.81%	22.63				7.83%	1.32%				

41.76%
36.52%
31.44%

36.52% of people who signed up
came back and uploaded a file 4 months later.

Wednesday, July 17, 13

CHURN

How many users are we losing?

Your app has 500 customers at the beginning of the month and only 450 customers at the end of the month. What is your churn?

Churn = (500-450)/500 = 50/500 = 10%

http://www.evergage.com/blog/how-calculate-churn

http://www.provisibly.com/2013/05/why-is-measuring-customer-churn-so-important/

CHURN

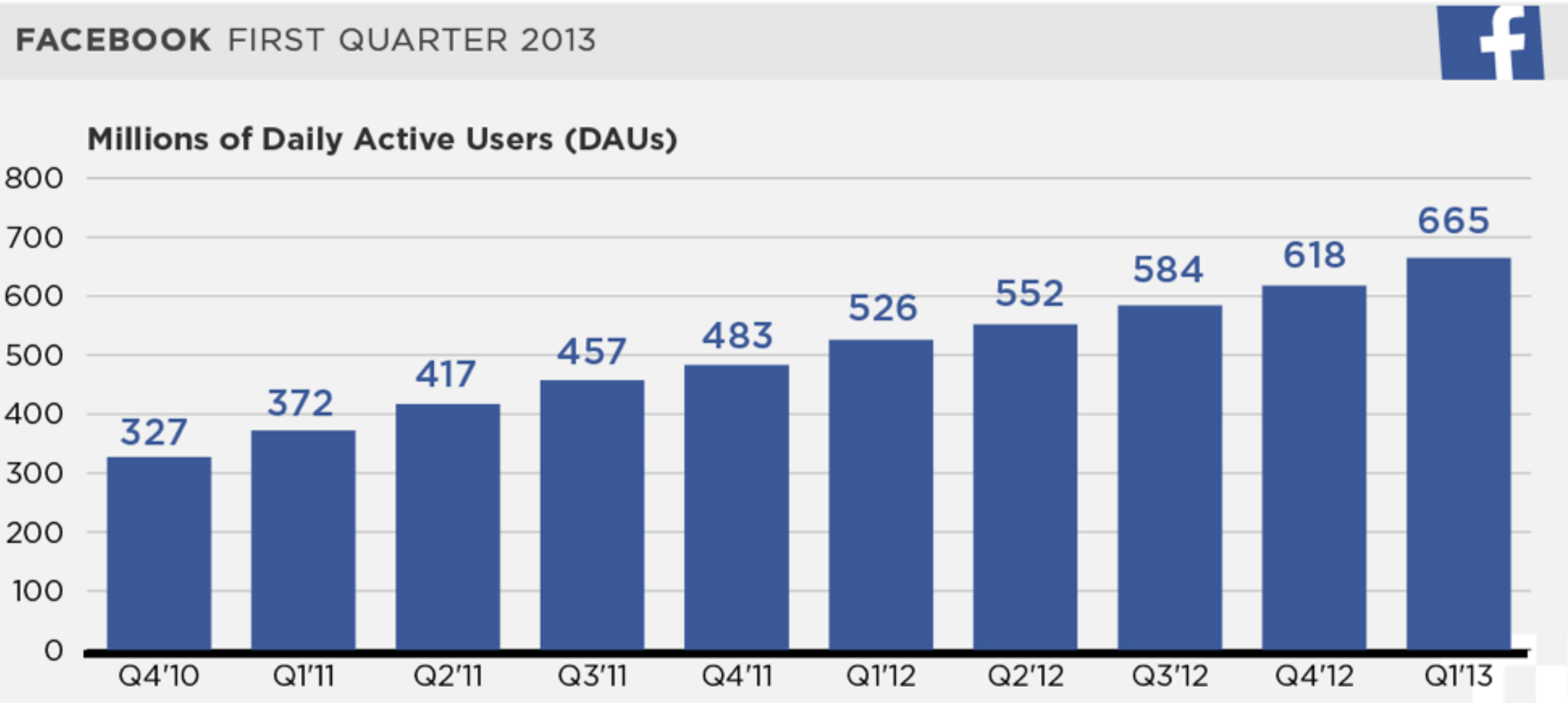
Churn impacts growth & profits significantly.

Wednesday, July 17, 13

http://www.forentrepreneurs.com/why-churn-is-critical-in-saas/

DAU/MAU

FACEBOOK FIRST QUARTER 2013



CUSTOMER LIFETIME VALUE

(CLV, CLTV, LCV, LTV)

How much is a customer worth?

CLV = Monthly Revenue x Margin x Number of Months

CLV = $100/mo x 25% x 10 months = $250

Wednesday, July 17, 13

CLV = (Avg Monthly Revenue per Customer * Gross Margin per Customer) ÷ Monthly Churn Rate

http://en.wikipedia.org/wiki/Customer_lifetime_value

Customer Acquisition Cost

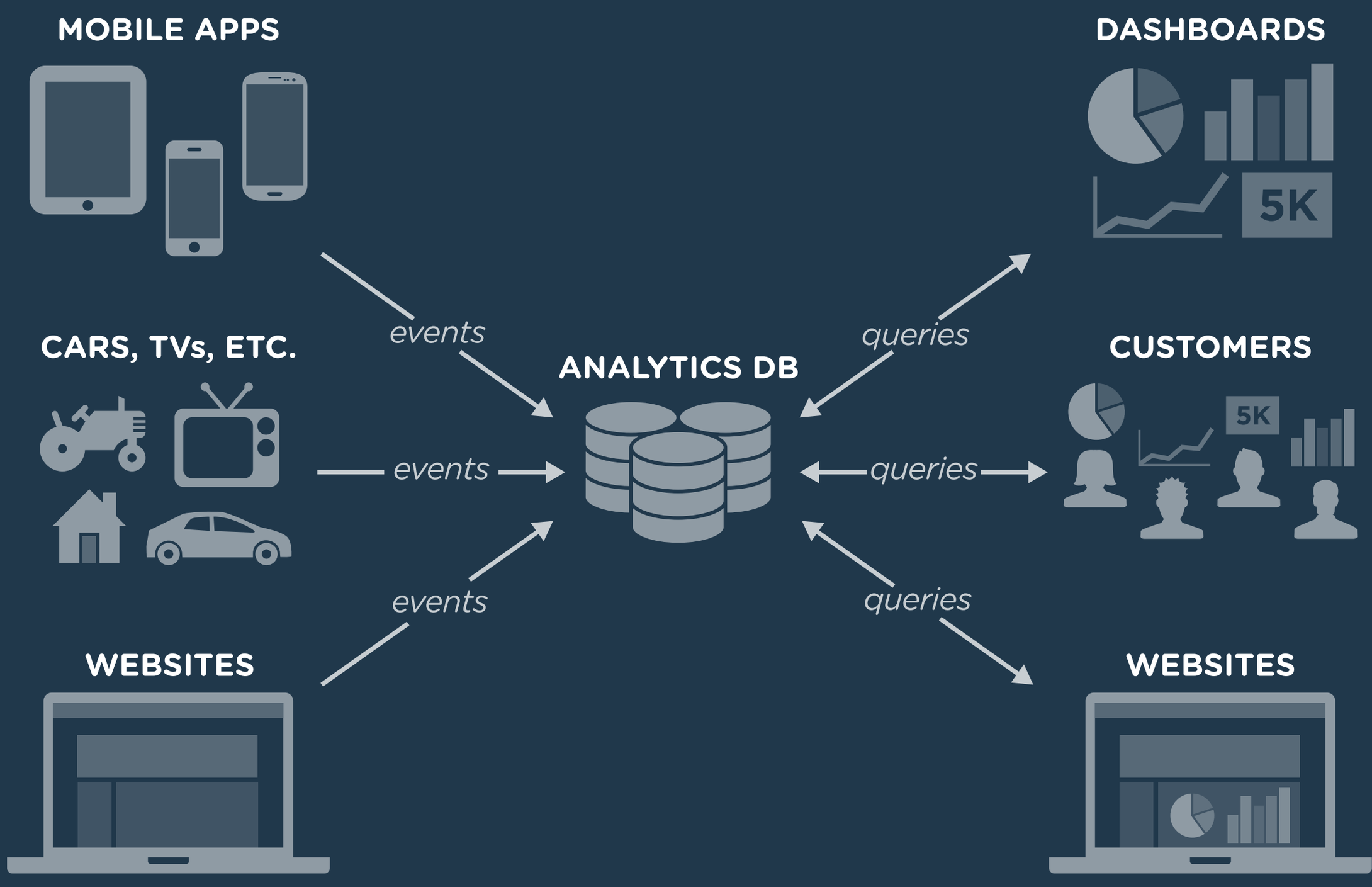
How much did it cost to get that user?

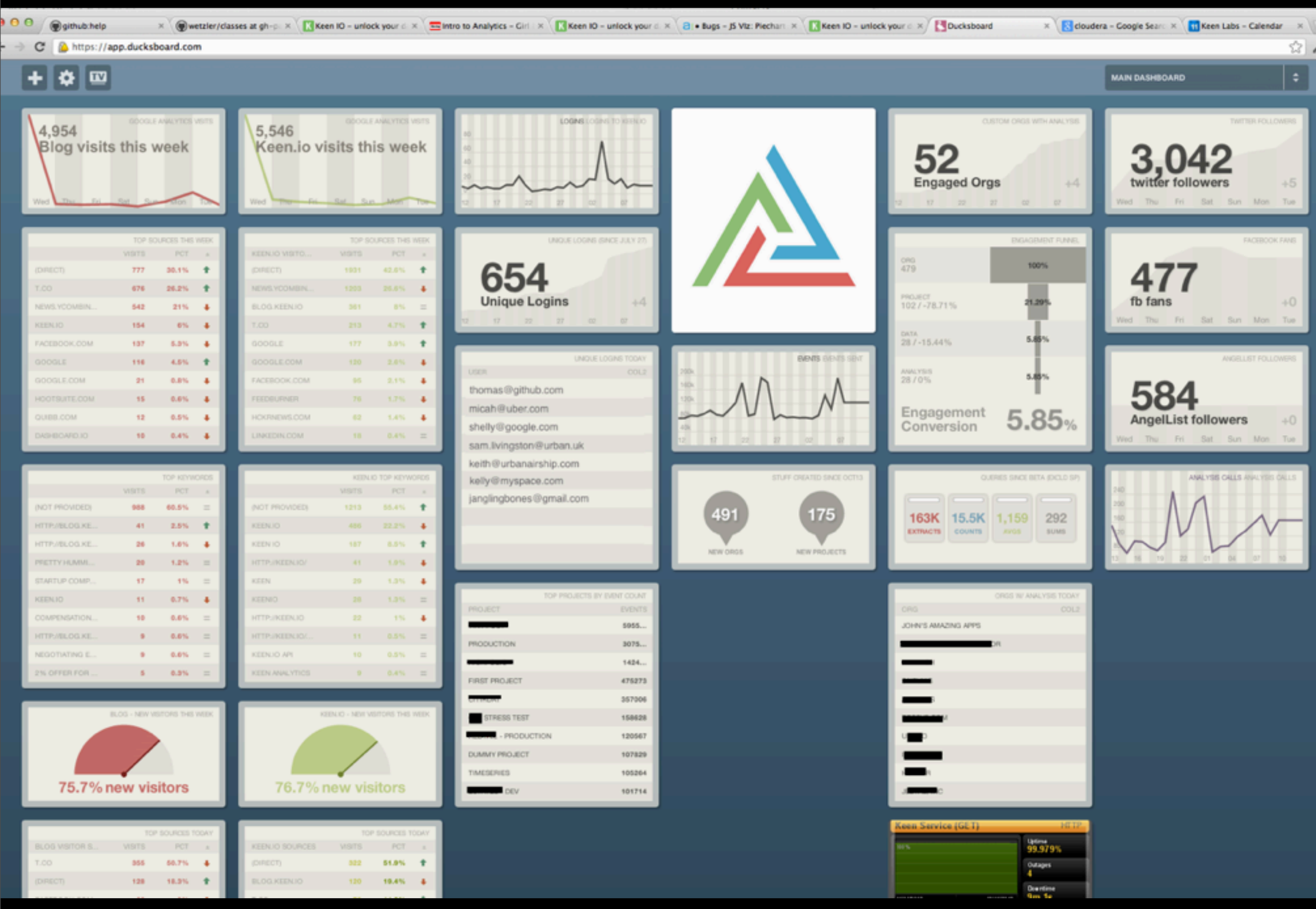
CAC = $ spent / number of users

Include amount invested in marketing, advertising, and sales.

Tools

Wednesday, July 17, 13

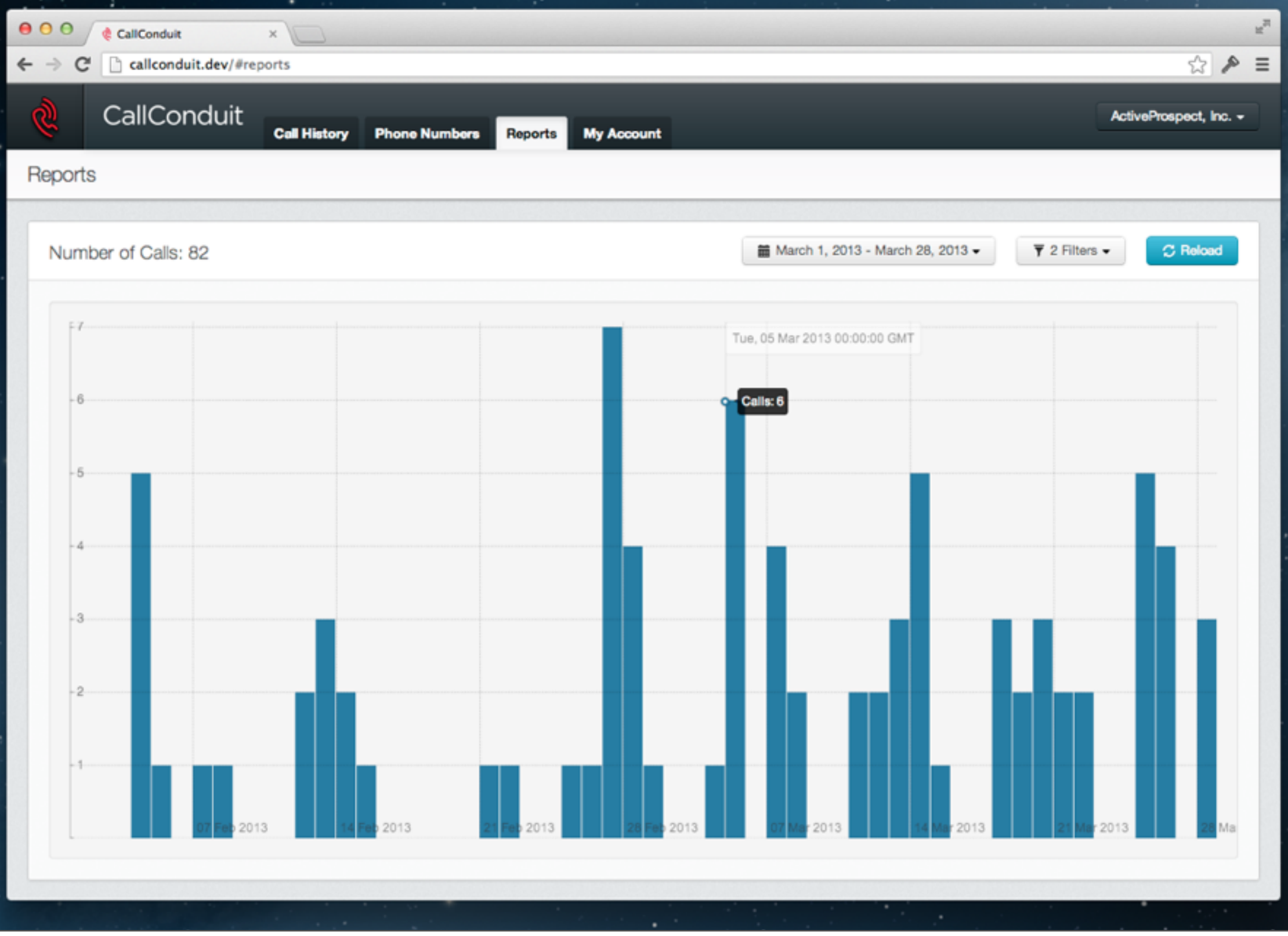

MOBILE APPS
DASHBOARDS
5K
CARS, TVs, ETC.
events
ANALYTICS DB
queries
CUSTOMERS
5K
events
queries
events
queries
WEBSITES
WEBSITES



Wednesday, July 17, 13

Use a dashboard to share analytics and motivate your team

A completely custom app (secretly powered by Keen IO)



FINAL EXERCISE!

1. Find exercise instructions at www.dataclassroom.com

2. The rest of the class time is devoted to this exercise & your questions. Feel free to leave at any time!

In the Peeps app, segment the number of records contacted by department to find out which departments are contacted most.
Segment the searches by searcher's department to see which department has the most active users.

In the peeps app segment the number of "contacts" by type (e.g. text, mail, phone call)