

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

# **Survey Results**

What should we talk about?

Analytics for your app	21
Analytics for marketing	17
Analytics for business	11
Big data	10

**First:** some context on big data and analytics

# “Big Data” and “Analytics” are kind of a thing right now.

The screenshot shows the WSJ homepage with a prominent article titled "Big Data's Big Problem: Little Talent". The article discusses the challenges of finding qualified data scientists. It features a photo of Hilary Mason, chief scientist for Bitly, and a quote from her. The WSJ logo is at the top, along with various navigation links like Home, World, U.S., New York, Business, Tech, Markets, Market Data, and Opinion. The date is Sunday, April 29, 2012, at 9:44 AM EDT.

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

## Big Data's Big Problem: Little Talent

Article   Video   Stock Quotes   Comments (46)

[Email](#) [Print](#) [Save](#) [Facebook](#) [Twitter](#) [Google+](#) [LinkedIn](#) [A](#) [A](#)

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 Bitly, 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.

The screenshot shows the Harvard Business Review website for the October 2012 issue. The article title is "Data Scientist: The Sexiest Job of the 21st Century" by Thomas H. Davenport and D.J. Patil. The page includes a sidebar with related content and a link to buy the PDF.

Harvard Business Review

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) [Email](#) [Print](#) [Save](#) [Facebook](#) [Twitter](#) [Google+](#) [LinkedIn](#)

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"

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives

<http://online.wsj.com/article/SB10001424052702304723304577365700368073674.html>  
<http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/>

## Cool data stories

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

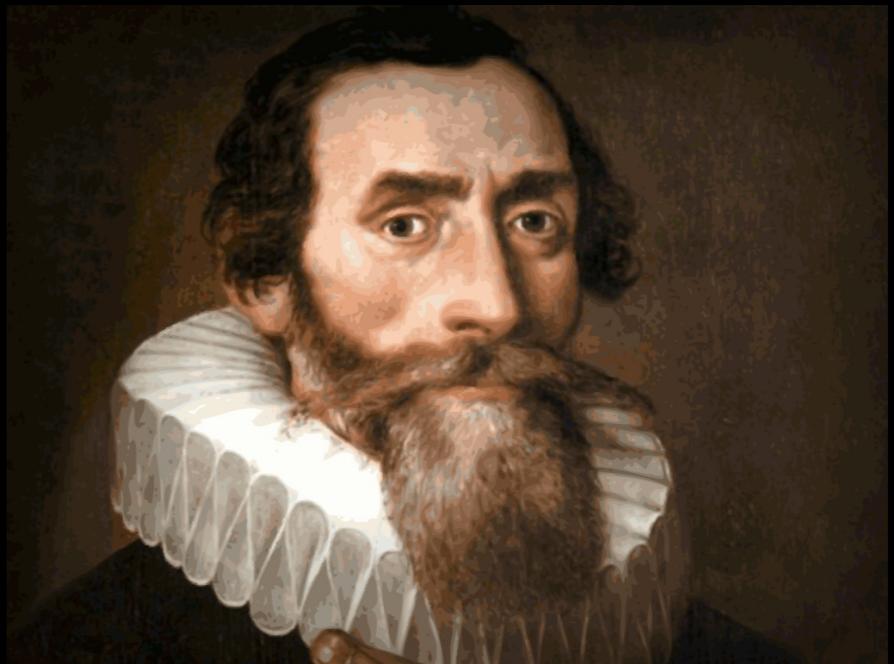
Data mining/analysis - Linked In

Data viz - Infosthetics

Communication - Broad Street Pump



Tycho Brahe



Johannes Kepler



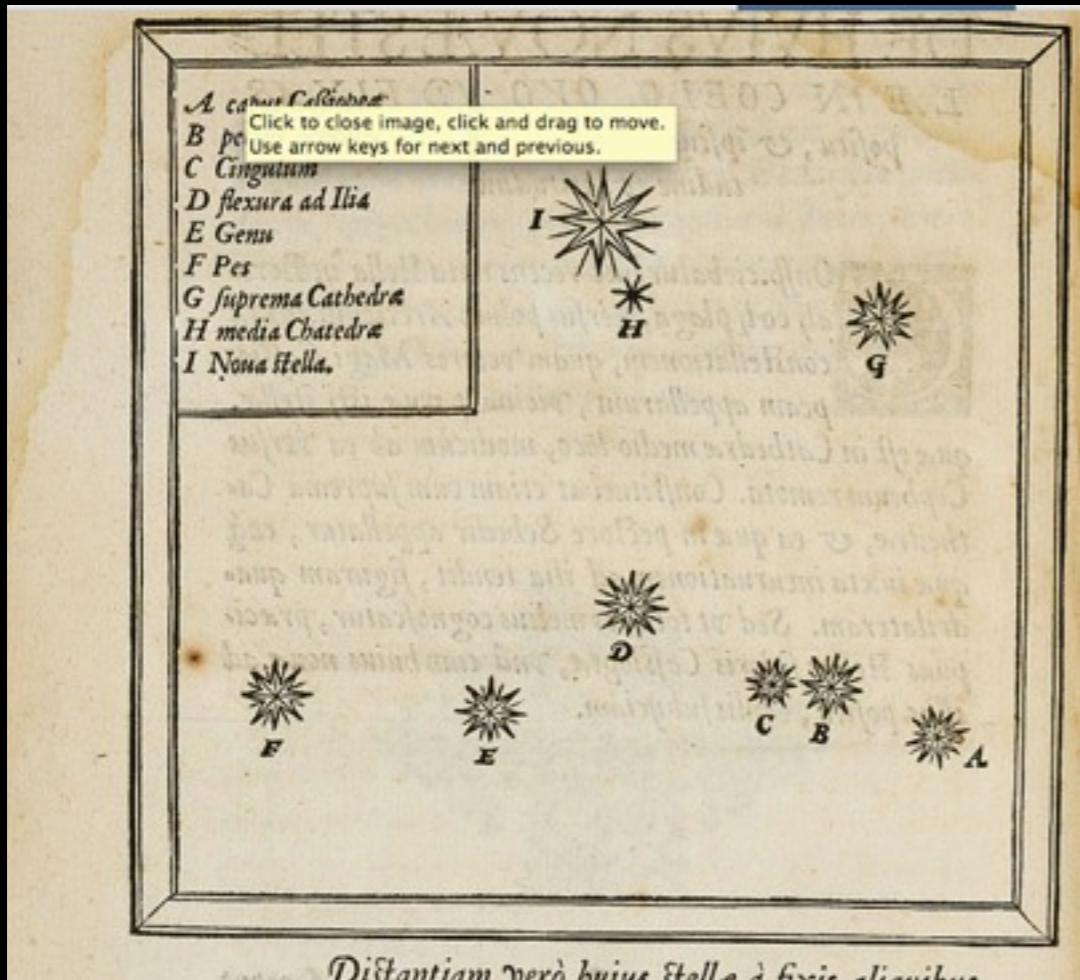
aka science walrus

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

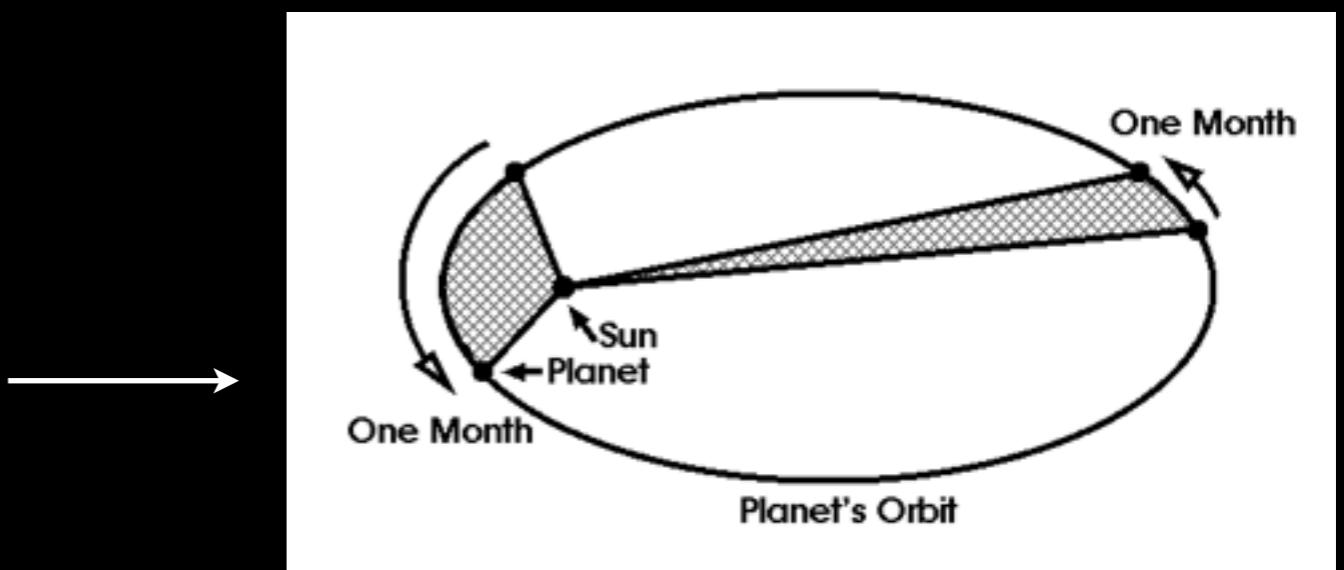
Tycho Brahe – collected an astounding amount of astronomical data. Every night, he would write down the location of every star and every planet in the sky. After 30 years of doing this... he died.

Johannes Kepler – took Brahe's data and single-handedly discovered the laws of planetary motion.

The point: recording stuff is important!



*Distantiam verò huius stellæ à fixis aliquibus  
in hac Cassiopeiae constellatione, exquisito instrumento,  
et omnium minutorum capaci, aliquoties obseruani. In-  
ueni autem eam distare ab ea, quæ est in peclore, Schedir  
appellata B, 7. partibus et 55. minutis: à superiori  
verò*



Tycho Brahe

Johannes Kepler

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

The screenshot shows the LinkedIn homepage. At the top, it displays 'Account Type: Basic | Upgrade' and a notification bar with 23 messages and 1 new message. The main navigation bar includes Home, Profile, Contacts, Groups, Jobs, Inbox, Companies, News, More, People, and a search bar. A red circle highlights the 'PEOPLE YOU MAY KNOW' section on the right side, which lists three recommended connections: Valerie Serrin, Izzy Johnston, and Peggy Hills, each with a 'Connect' button. Below this is an 'ADS BY LINKEDIN MEMBERS' section featuring an ad for building a huge email list.

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

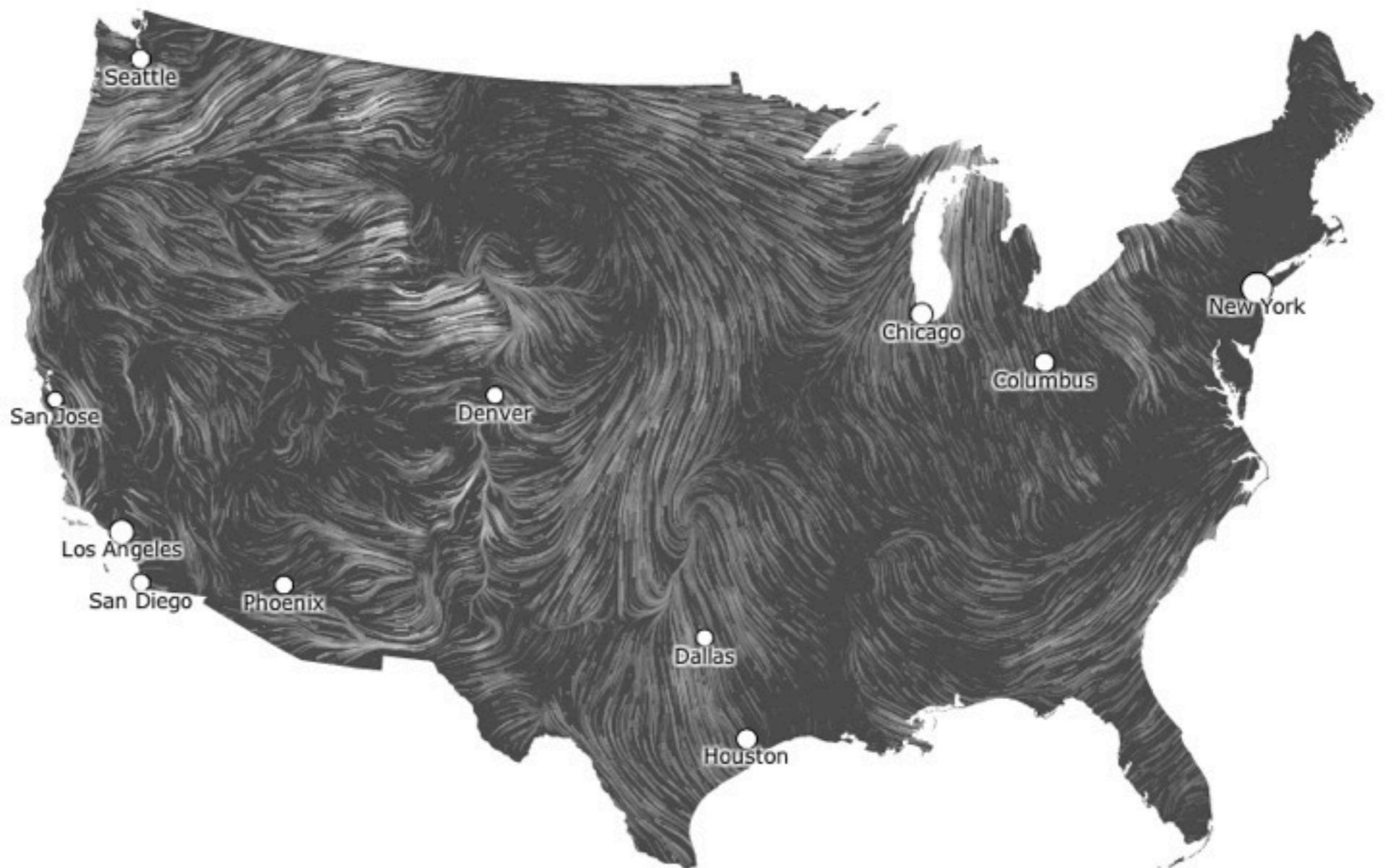
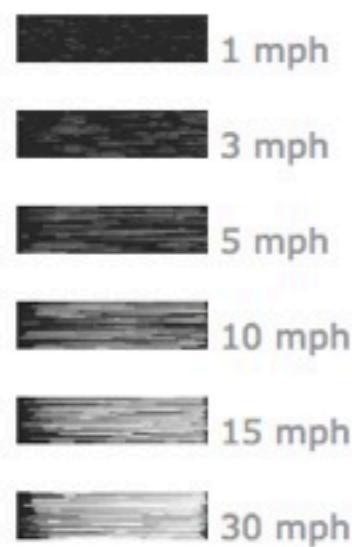
- Goldman, a PhD in physics from Stanford, was intrigued by the linking activity he saw happening on LinkedIn. He began exploring people's connections, forming theories, testing hunches, and finding patterns that allowed him to predict whose networks a given profile would land in.
- He could imagine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

**Dec. 6, 2012**

5:59 pm EST

(time of forecast download)

top speed: **30.2 mph**  
average: **6.2 mph**



<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. Checkout [infosthetics.com](http://infosthetics.com).

<http://hint.fm/wind/>

<http://infosthetics.com/>

# Le Petit Journal

ADMINISTRATION  
61, RUE LAFAYETTE, 61  
Les manuscrits ne sont pas rendus  
On s'abonne sans frais  
Dans tous les bureaux de poste

5 CENT. SUPPLÉMENT ILLUSTRÉ 5 CENT.  
23<sup>e</sup> Année — Numéro 1.150  
DIMANCHE 1<sup>er</sup> DÉCEMBRE 1912

ABONNEMENTS  
PARIS  
SEINE ET SEINE-ET-OISE... 2 fr. 3 fr. 80  
DÉPARTEMENTS..... 2 fr. 4 fr. \*  
ÉTRANGER..... 2.50 6.50 \*



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](http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak)

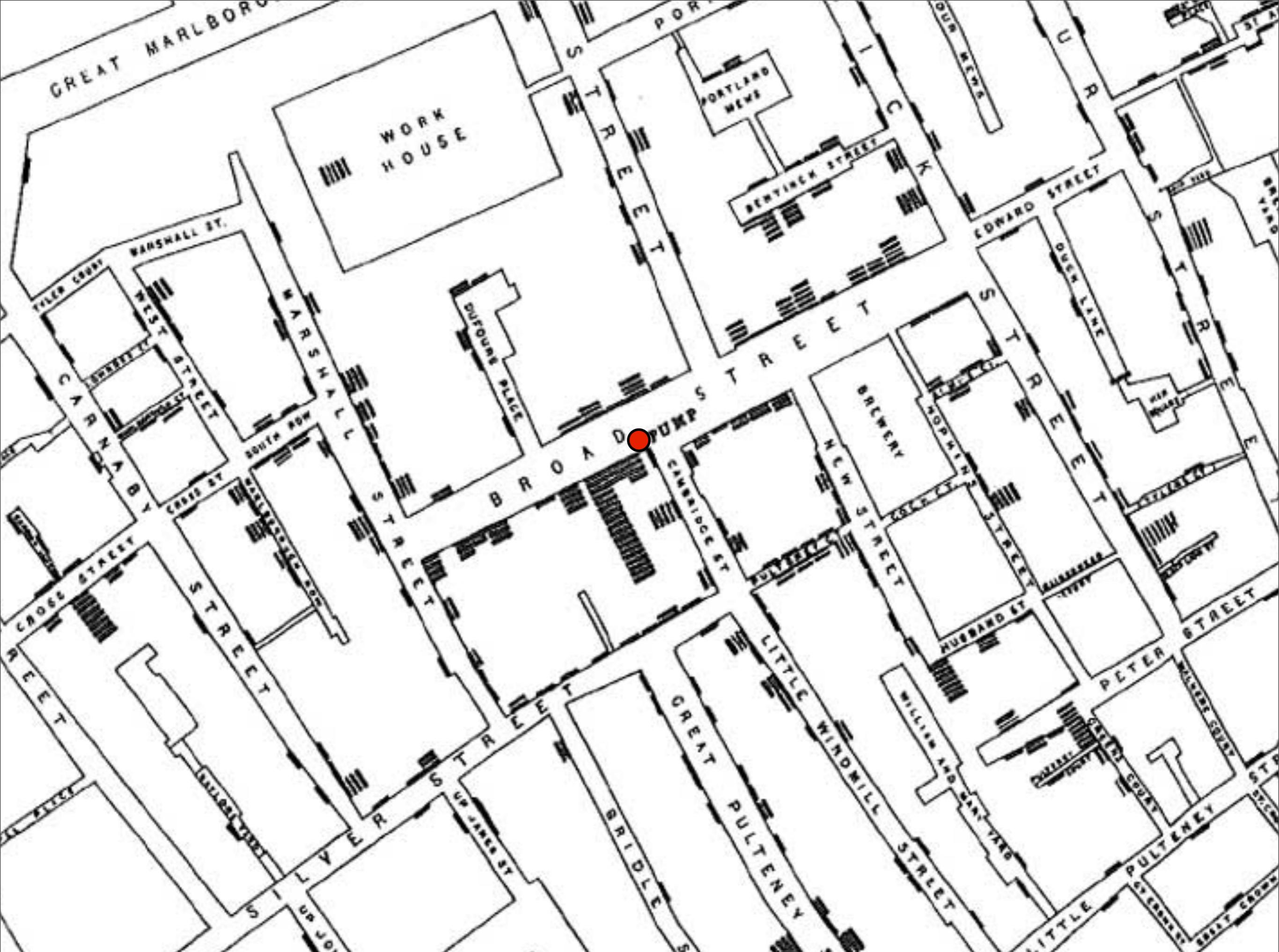


Figured out Cholera  
spreads through water.

No one believed him :(

## John Snow

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



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.

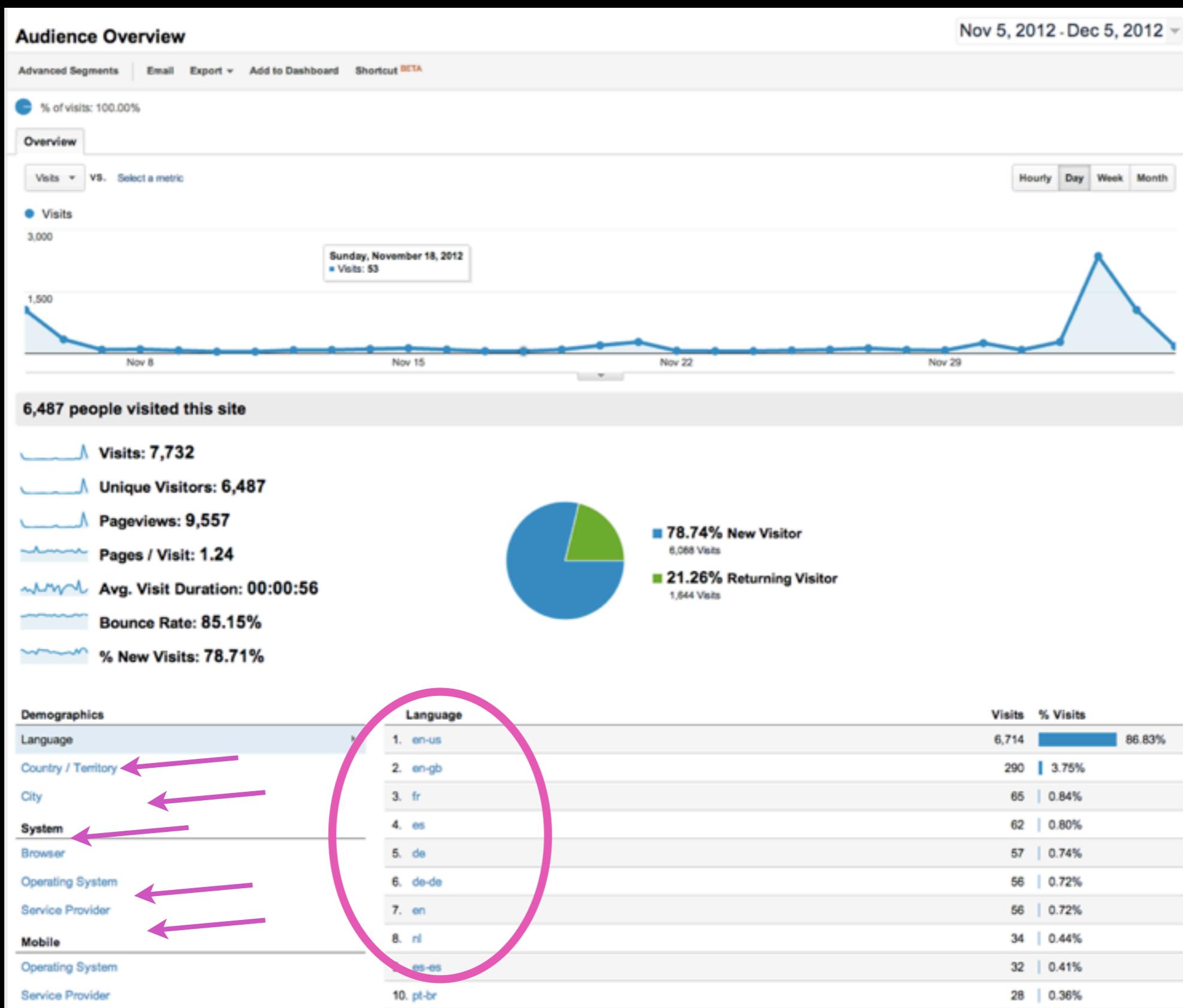
## **Next: analytics in your app.**

---

### **Use analytics to make your app better**

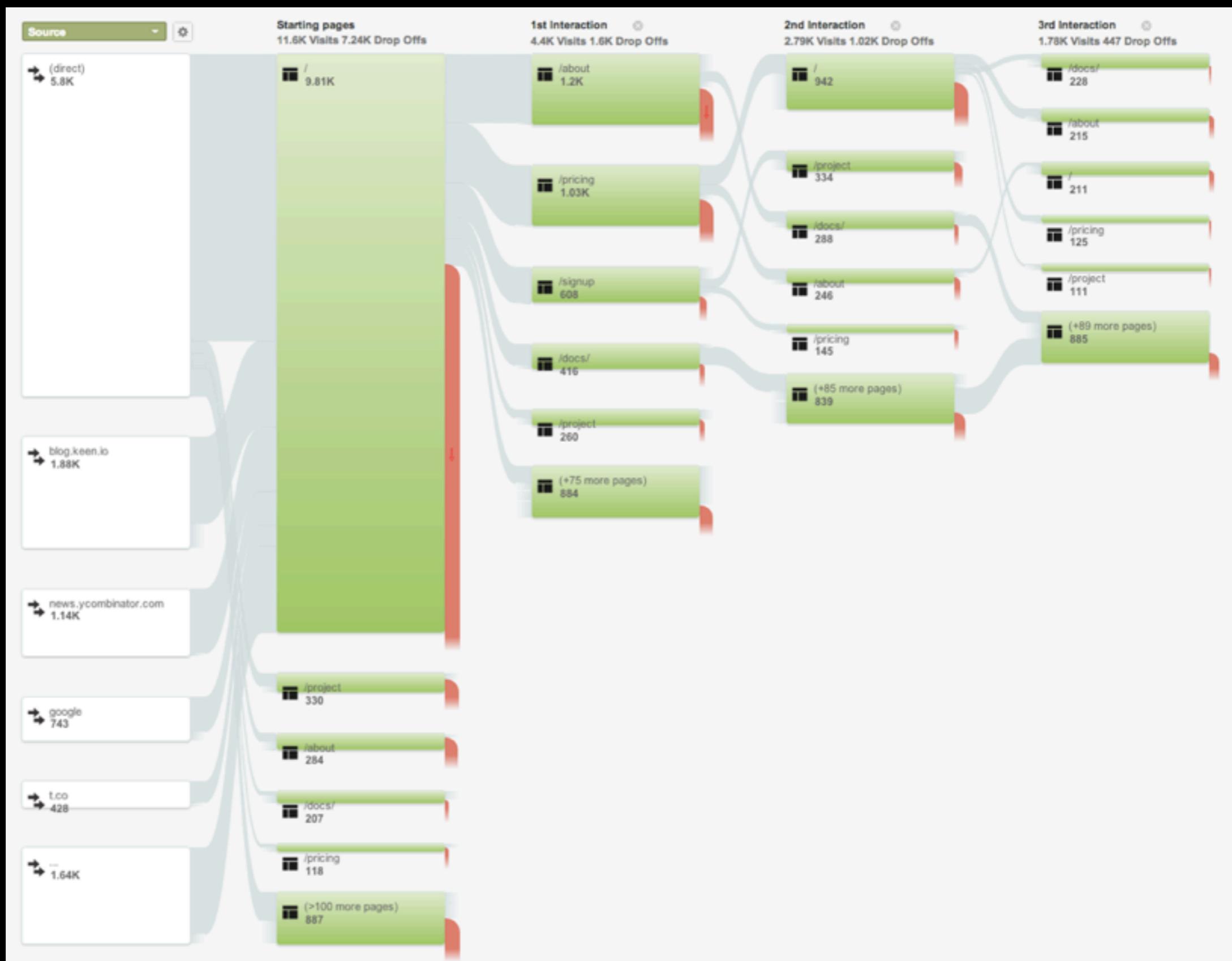
1. Figure out who is using your app
2. Figure out what features are most used
3. Figure out where people are quitting

# Google analytics is freeeeeeee!



# <live demo>

# Google analytics “flows” show user click paths



# Who is using your app?

USER	COL2
thomas@github.com	
micah@uber.com	
shelly@google.com	
sam.livingston@urban.uk	
keith@urbanairship.com	
kelly@myspace.com	
janglingbones@gmail.com	

use a DB to store login data

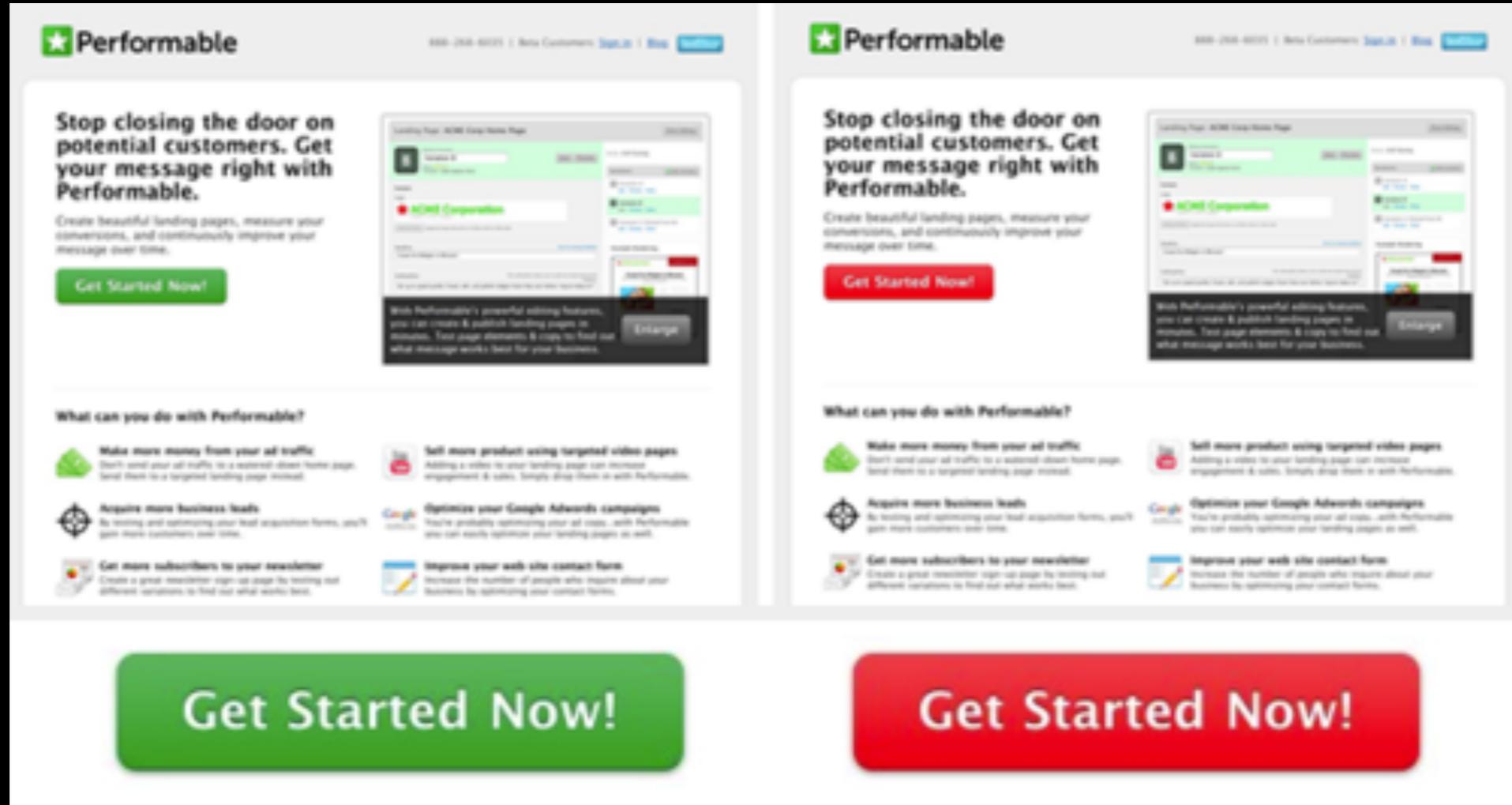
The screenshot shows a window titled "Contacts" with a sidebar menu on the left. The menu includes "Web Visitors", "Buddies", "Accenture (General)", and "Operators". Under "Web Visitors", there is a list of users with their location and ID: USA (San Francisco, CA) #757, USA (San Francisco, CA) #757 / | looking at https://keen..., USA (Orlando, FL) #8484, USA (Orlando, FL) #8484 / | There is no status informati..., kirk (kirk@keen.io), kirk (kirk@keen.io) / | speaking with kirk / | kirk@keen.i..., USA (San Francisco, CA) #6561, USA (San Francisco, CA) #6561 / | looking at https://keen.io/. The main pane displays a detailed status for a user: USA (San Francisco, CA) #6561 (webuser48.128478@im.olark.com). The status information includes: Status: USA (San Francisco, CA) #6561 | looking at https://keen.io/ | already looked at: https://keen.io/ https://keen.io/pricing https://keen.io/ https://keen.io/ https://keen.io/pricing https://keen.io/ https://keen.io/docs/event-data-modeling/event-data-intro/ referred from https://keen.io/docs/getting-started-guide/ located in USA (San Francisco, CA) 6 pages viewed 4 visits using Firefox 16.0 on Mac OSX 173.228.102.82 spent at least 1 hours so far. At the bottom, it says Contacts: USA (San Francisco, CA) #6561.

Olark for real time data

Record logins to your app. Use tools like KissMetrics, Mixpanel, Keen IO.

- who is using your app the most?
- who hasn't logged in for a long time?
- what companies do your users work at?
- track last login for each customer

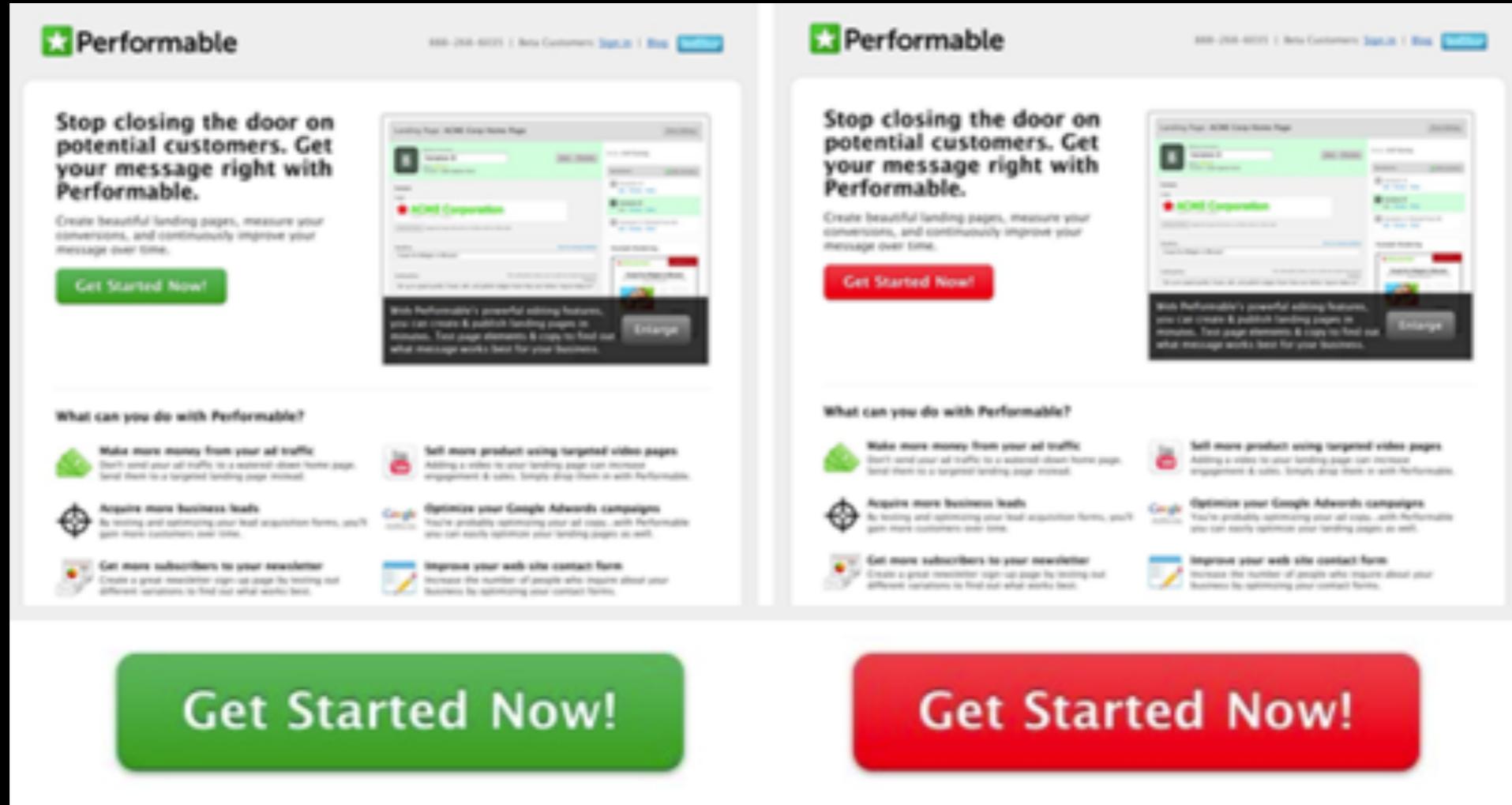
# A/B Testing aka Split Testing



[http://en.wikipedia.org/wiki/A/B\\_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](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

Signup Events

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?

Name an example of something in your app that you would like to A/B test.

# What is Event Data?



**Actions & State + Time**

now we can record more stuff than ever!

we can store information about *every user interaction*.

new analytics tools like KissMetrics, MixPanel, and Keen IO allow you to record event data about your application.

traditional db stores state in tables: users, inventory, accounting data

event data bases store Actions, State, and Time!

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

# Action

```
purchase = {  
    timestamp: 2012-06-06T19:10:39.205000,  
    item: sophisticated orange turtleneck with deer on it,  
    cost: 469.5,  
    payment_method: Bank Simple VISA,  
    customer: {  
        name: Francis Woodbury,  
        age: 28,  
        personality: sullen,  
    },  
    store: {  
        name: Yupster Things,  
        city: San Francisco,  
        state: California,  
        address: 467 West Portal Ave,  
    }  
};
```

# Time

# State

provide rich contextual information with every event.

now you can count the number of purchases by people under the age of 21. In a single query.

you can calculate the average purchase amount for businesses within 10miles of SF.

you can graph the number purchases per house for a given store or state.

## **Exercise I**

Pair with the person next to you.

Spend 5 minutes to come up with an event data model for something you would like to A/B test.

## Funnels

Launched App: 550

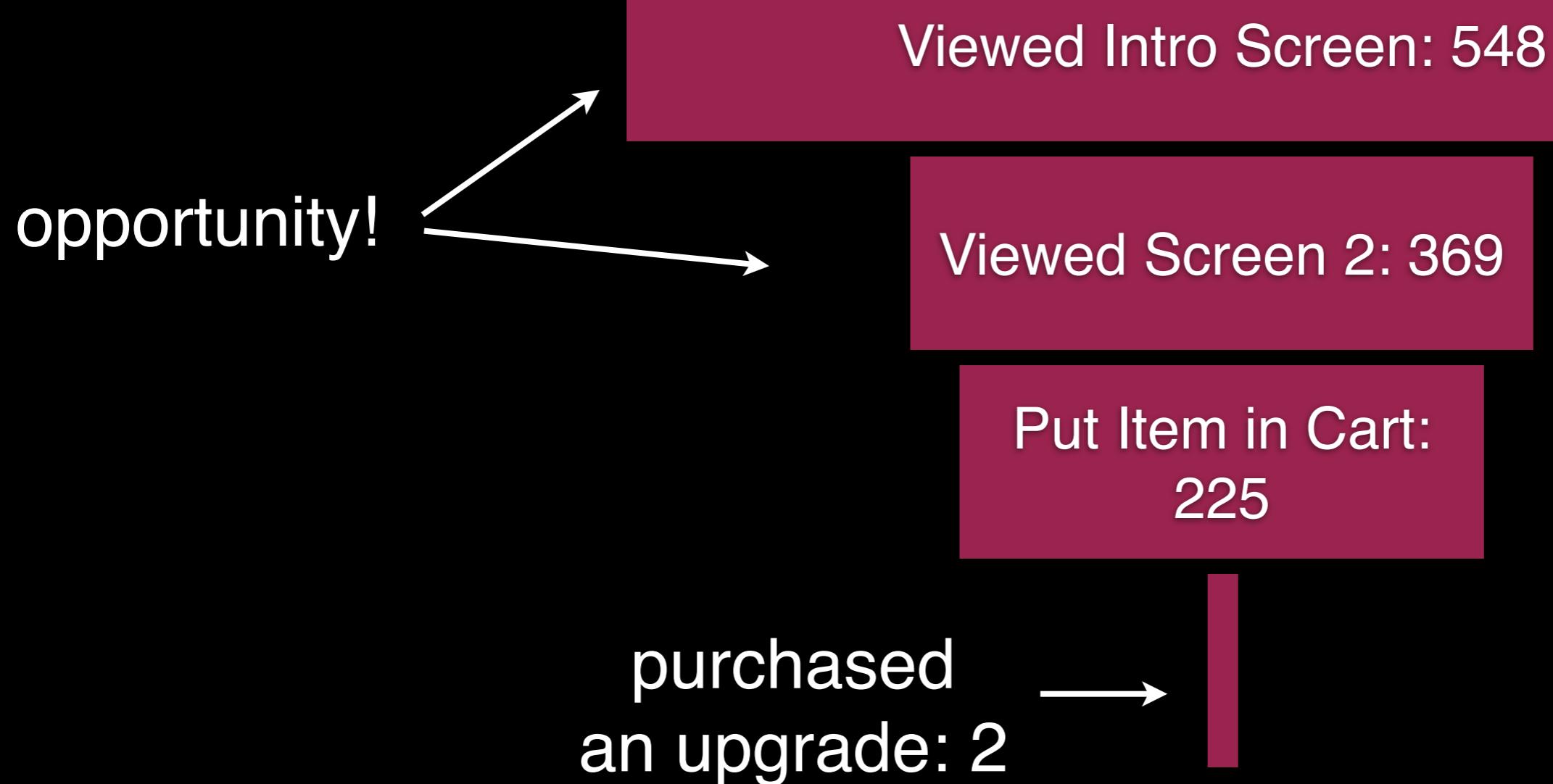
Viewed Intro Screen: 548

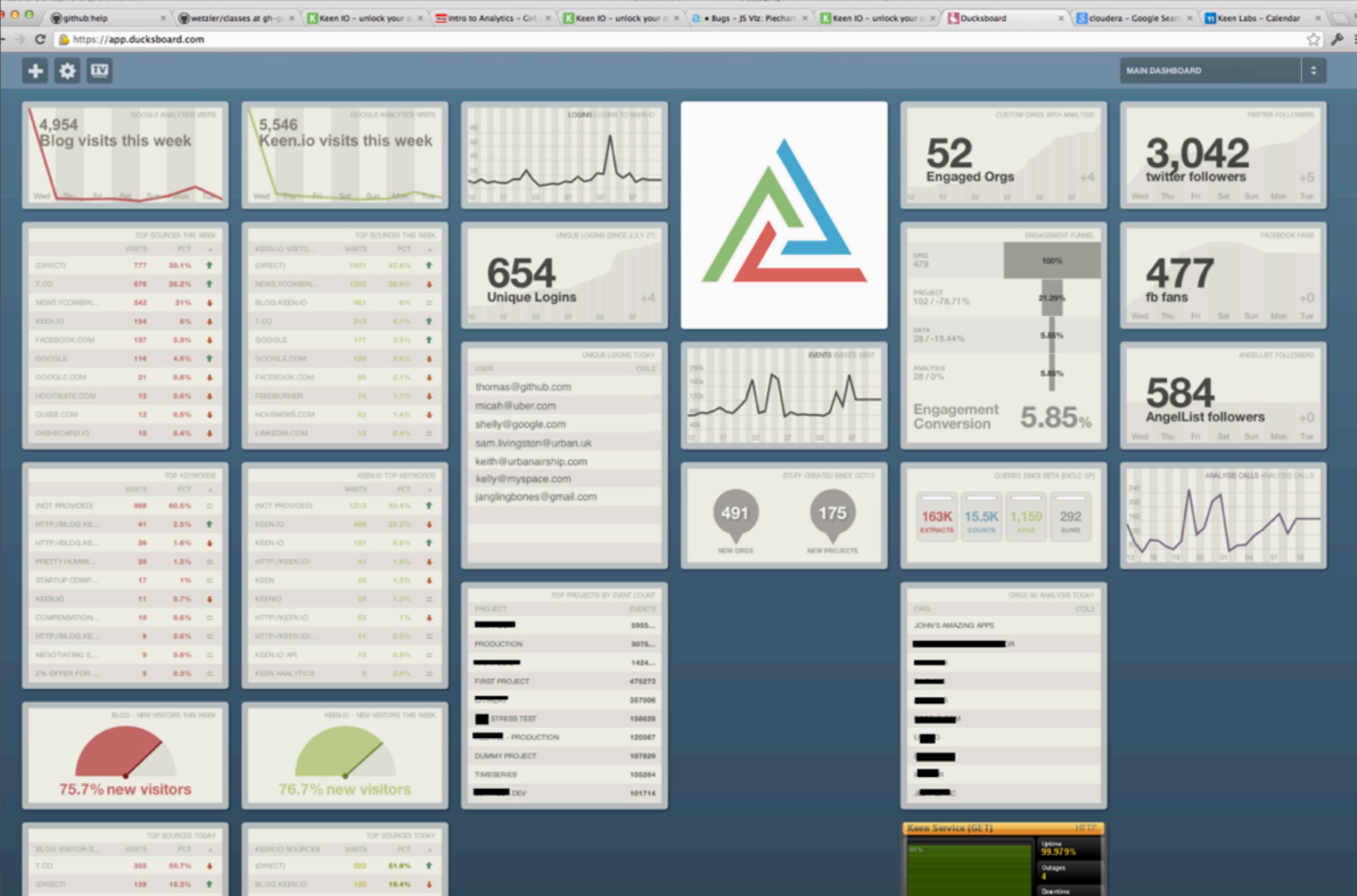
Viewed Screen 2: 369

Put Item in Cart:  
225

purchased  
an upgrade: 2 → |

## Funnels





# New Cool Stuff

What are people highlighting?

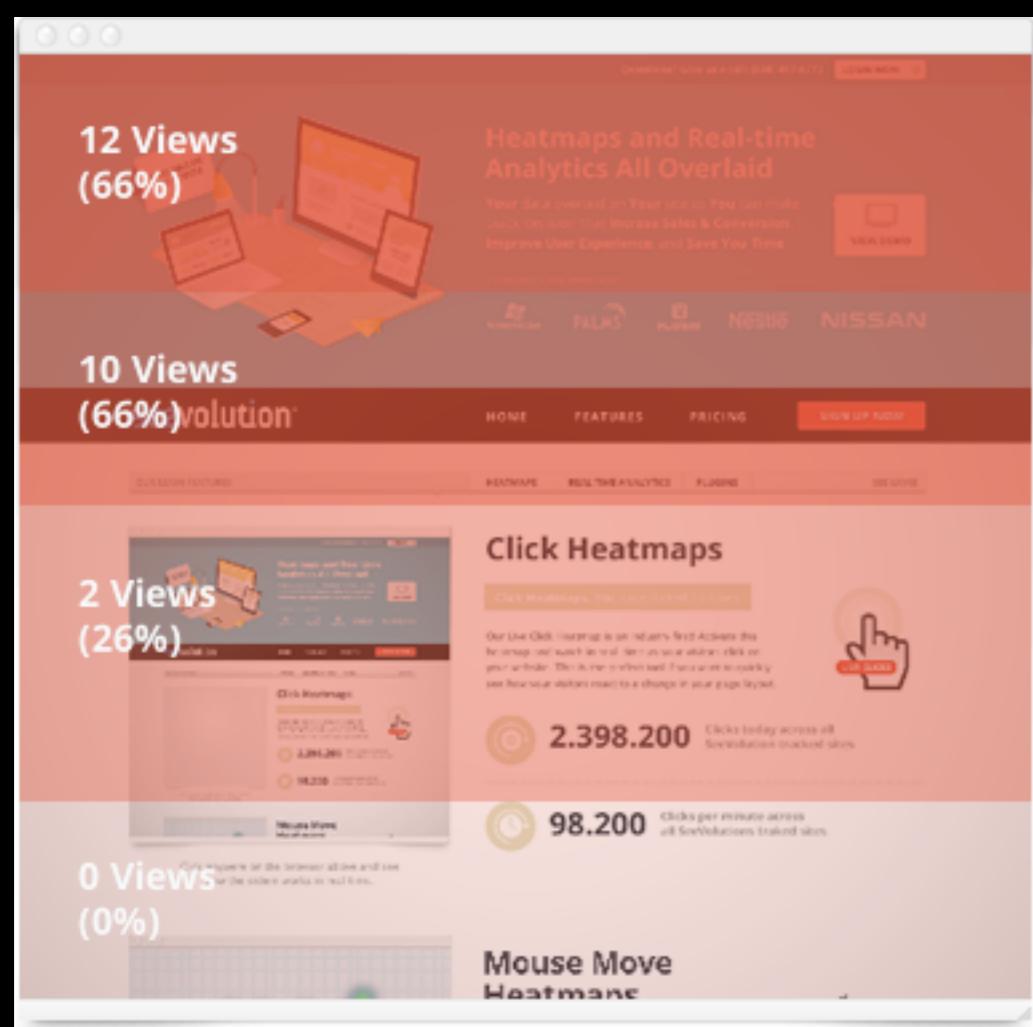
<http://markerly.com/>

Where are people looking?

<http://www.crazyegg.com/>

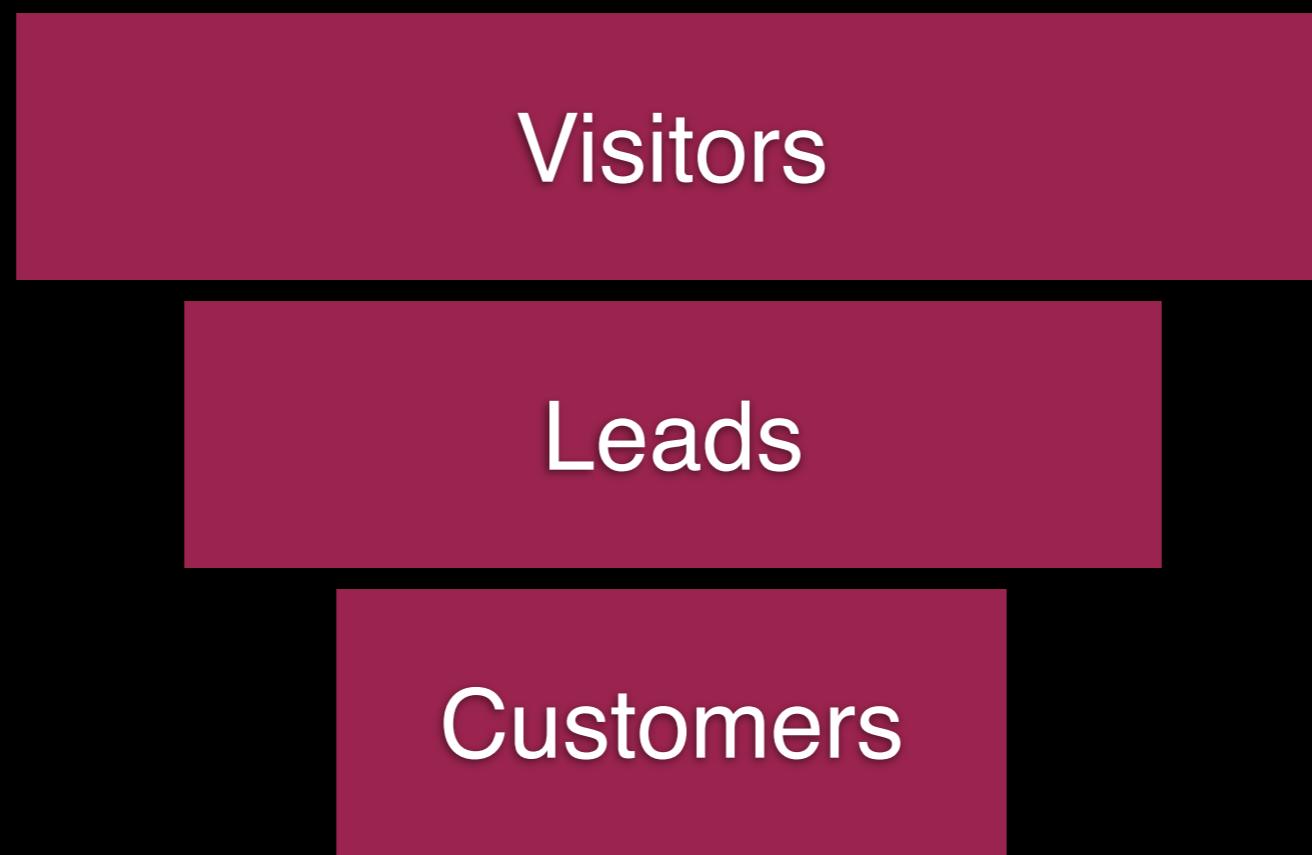
Where are people mousing?

<https://www.seevolution.com/>



**Next:** analytics for marketing

# Conversion Funnel



You can use analytics to improve conversion throughout the funnel.  
So far we've talked about how to make your app better for people who have made it to your app.  
You can also use analytics to increase the traffic to your app or site.

# where are people coming from? use google analytics “sources” feature

Source / Medium	Visits
1. news.ycombinator.com / referral	5,985
2. (direct) / (none)	3,410
3. google / organic	959
4. t.co / referral	854
5. facebook.com / referral	828
6. keen.io / referral	814
7. google.com / referral	424
8. hckrnews.com / referral	170
9. Newsletter / Email	166
10. news.ycombinator.org / referral	132

“dark social” = email and IM sharing  
t.co = twitter

# how successful are your app's emails?



ToutApp Live Feed  
<https://toutapp.com/next#live>

**toutapp™**

your advice for Keen IO

Nov 26 12:44 am Viewed from Alhaur, 51.  
Nov 07 1:27 am Viewed from Sanl, 51.

Show 10 more... ▾

**Tom W.....**  
your advice for Keen IO

Nov 12 9:57 am Viewed from Poway, CA.  
Nov 12 9:57 am Viewed from Poway, CA.

Show 49 more... ▾

**Ted S.....**  
your advice for Keen IO

Nov 08 4:57 pm Viewed your email  
Nov 06 10:35 am Viewed from Sterling, VA.

Show 8 more... ▾

**Tim E.....**  
your advice for Keen IO

Nov 05 1:42 am Viewed your email

**Tom C.....**

watch people respond to your emails in real time!

www.facebook.com/Keen.io?sk=page\_insights

facebook Search for people, places and things Keen IO Home

Keen IO Timeline ▾

Overview Likes Reach Talking About This

All dates and times are in Pacific Time Export Data

Total Likes? Friends of Fans? People Talking About This? Weekly Total Reach? Total Subscribers?

477 +2.36% 232,485 +8.07% 188 +135% 7,975 -2.43% 1 +0%

Posts? People Talking About This? Weekly Total Reach?

Nov 14 Nov 21 Nov 28 Dec 5

Page Posts (Updated 4 minutes ago)

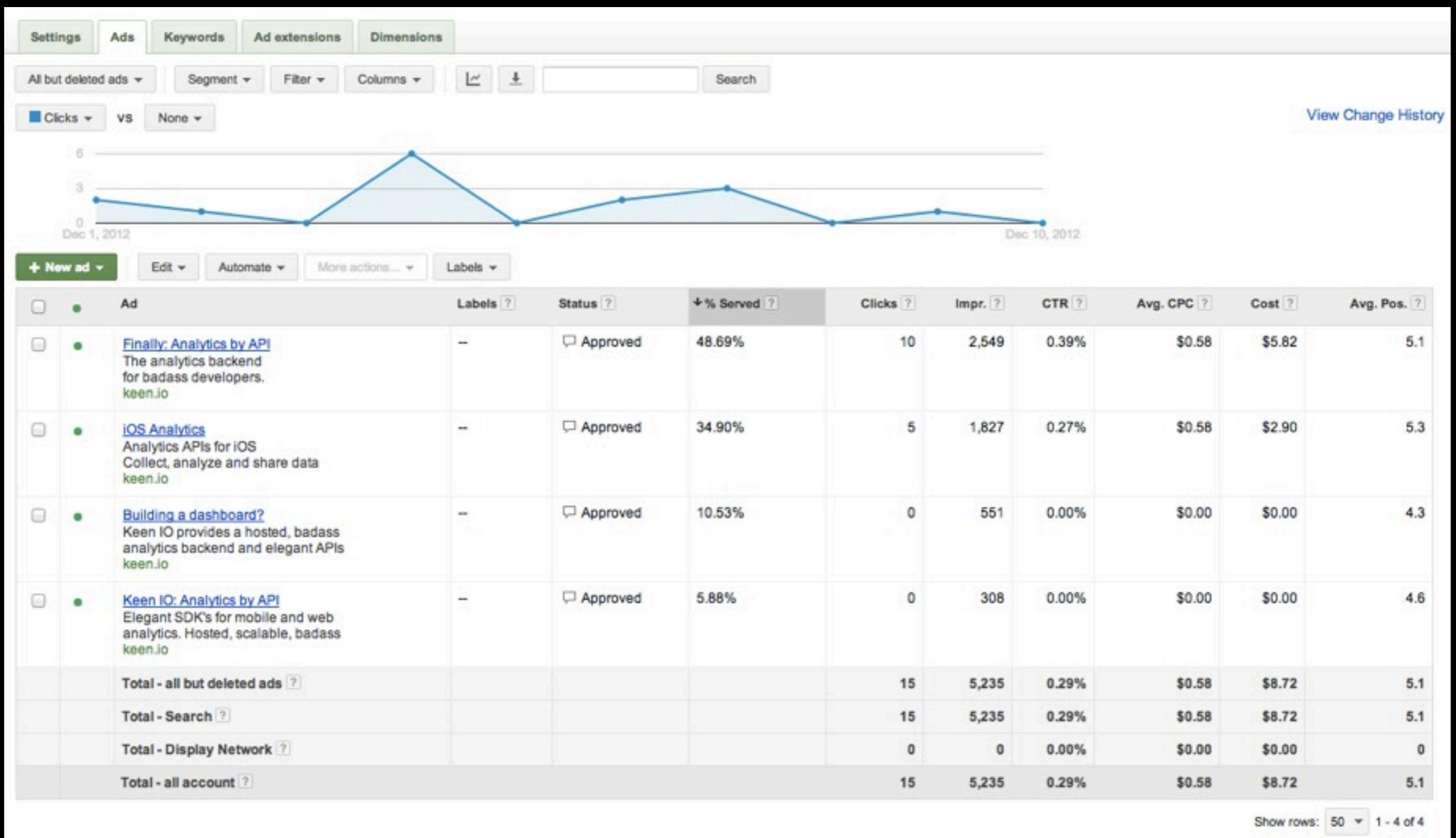
All Post Types

Date	Post	Reach	Engaged Users	Talking About This	Virality
12/10/12	sometimes, you just gotta START so...	70	7	4	5.71%
12/6/12	radiohead singalong-and-work session	178	128	29	16.29%
12/6/12	do you like pie?? just added this feat...	133	17	4	3.01%
12/4/12	Wow! Around midnight, someone po...	241	66	20	8.3%
12/4/12	Oh my! Someone posted www.keen.i...	2,819	112	50	1.77%
12/3/12	How we used Twitter ads to generate...	4,614	157	66	1.43%
12/2/12	developer who came to the hackatho...	--	33	--	--
12/2/12	Keen at work	--	47	--	--
12/2/12	Keen at work	11	72	9	81.82%

Ads Manager See Your Ad Here  
The right amount of ambition, alcohol and lack of plans on a Just Fucking Build It http://blog.keen.io/post/37419311398/just-fucking-build-it Promote This Story

social media sites have AMAZING dashboards for measuring the effectiveness of your efforts.  
See also: Google Ads, Twitter

# Google Adwords



Google adwords is cool because it automatically chooses ads and keywords for you.

# Cohort Analysis

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

Engagement over time													
	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12	
(Joined in) January	100%	20%	19%	13%	13%	10%	12%	11%	7%	7%	7%	7%	?
February	100%	21%	16%	13%	11%	9%	9%	7%	7%	7%	7%	7%	?
March	100%	24%	20%	17%	15%	13%	11%	10%	10%	10%	10%	10%	?
April	100%	31%	27%	24%	19%	15%	12%	12%	12%	12%	12%	12%	?
May	100%	31%	27%	25%	21%	18%	16%	16%	16%	16%	16%	16%	?
June	100%	39%	28%	24%	20%	19%	19%	19%	19%	19%	19%	19%	?
July	100%	40%	33%	27%	23%	23%	23%	23%	23%	23%	23%	23%	?
August	100%	47%	41%	32%	32%	32%	32%	32%	32%	32%	32%	32%	?
September	100%	52%	43%	43%	43%	43%	43%	43%	43%	43%	43%	43%	In this case engagement is improving nicely. Of the January cohort, only 20% were engaged in month 2. Of the October cohort, 53% were engaged in month 2.
October	100%	53%	?	?	?	?	?	?	?	?	?	?	
November	100%	?	?	?	?	?	?	?	?	?	?	?	
December	?	?	?	?	?	?	?	?	?	?	?	?	

<http://redeye.firstround.com/2008/01/after-the-techc.html>

<http://www.cohortanalysis.com/>

<http://jonathonbalogh.com/2012/03/24/introduction-to-cohort-analysis-for-startups/>

<http://52weeksofux.com/post/646711369/cohort-analysis-measuring-engagement-over-time>

**Exercise Time!**

**Conversion Rate** - Percentage of people who made it from one part of the funnel to another part of the funnel

**MAU** - Unique monthly active users. Common in gaming and consumer apps (Facebook has 1B MAU).

**Segmentation** - Dividing things into groups for analysis (e.g. Average purchase amount segmented by age group)

**CPC** - Cost per click. This is how ads are priced on twitter, facebook, and google.