



# DESIGNING WITH DATA

Seth Familian  
Founder + Principal, Familian&1



## SETH FAMILIAN, FOUNDER + PRINCIPAL, FAMILIAN&1

### BUSINESS STRATEGY



BERTELSMANN  
media worldwide



### PRODUCT MANAGEMENT



### WEB PRESENCE

bergamot:station

BLUE FIELD  
STRATEGIES



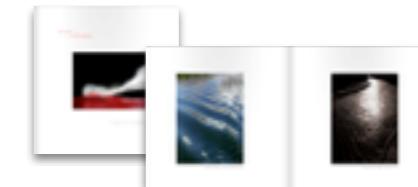
### GROWTH HACKING



### TEACHING + EDUCATION



### PROCRAFTINATION



GENERAL ASSEMBLY

# WORKING WITH BIG DATA

Seth Familian  
Founder + Principal, Familian&I

GENERAL ASSEMBLY

# VISUAL DESIGN WITH DATA

Seth Familian  
Founder + Principal, Familian&I

GENERAL ASSEMBLY

# DIGITAL TOOLS FOR BUSINESS

Seth Familian  
Founder + Principal, Familian&I

FOLLOW ALONG!

<http://bit.ly/ga-dt4b>

- What makes good (data) design?
- Creating effective charts
- Data viz tools + techniques

# WHAT'S GOOD (DATA) DESIGN?

# WHAT'S GOOD (DATA) DESIGN?

CONSISTENT.

6

	A	B	C	D
1	Date	VSS	SVS	Total
2	Jan-09	\$39,123	\$27,310	\$66,433
3	Feb-09	\$39,044	\$27,041	\$66,085
4	Mar-09	\$39,064	\$26,453	\$65,517
5	Apr-09	\$38,276	\$25,979	\$64,255
6	May-09	\$37,440	\$25,660	\$63,100
7	Jun-09	\$36,321	\$25,019	\$61,340
8	Jul-09	\$37,352	\$24,156	\$61,508
9	Aug-09	\$36,577	\$23,625	\$60,202
10	Sep-09	\$37,262	\$22,558	\$59,820
11	Oct-09	\$37,362	\$21,844	\$59,206
12	Nov-09	\$37,684	\$23,603	\$61,287
13	Dec-09	\$38,861	\$22,949	\$61,810

	A	B	C	D
1	Date	VSS	SVS	Total
2	Jan-09	\$39,123	\$27,310	\$66,433
3	Feb-09	\$39,044	\$27,041	\$66,085
4	Mar-09	\$39,064	\$26,453	\$65,517
5	Apr-09	\$38,276	\$25,979	\$64,255
6	May-09	\$37,440	\$25,660	\$63,100
7	Jun-09	\$36,321	\$25,019	\$61,340
8	Jul-09	\$37,352	\$24,156	\$61,508
9	Aug-09	\$36,577	\$23,625	\$60,202
10	Sep-09	\$37,262	\$22,558	\$59,820
11	Oct-09	\$37,362	\$21,844	\$59,206
12	Nov-09	\$37,684	\$23,603	\$61,287
13	Dec-09	\$38,861	\$22,949	\$61,810

# WHAT'S GOOD (DATA) DESIGN?

7

## UNCLUTTERED.

	A	B	C	D
1	Date	VSS	SVS	Total
2	Jan-09	\$39,123	\$27,310	\$66,433
3	Feb-09	\$39,044	\$27,041	\$66,085
4	Mar-09	\$39,064	\$26,453	\$65,517
5	Apr-09	\$38,276	\$25,979	
6	May-09	\$37,440	\$25,660	
7	Jun-09	\$36,321	\$25,019	
8	Jul-09	\$37,352	\$24,156	
9	Aug-09	\$36,577	\$23,625	
10	Sep-09	\$37,262	\$22,558	
11	Oct-09	\$37,362	\$21,844	
12	Nov-09	\$37,684	\$23,603	
13	Dec-09	\$38,861	\$22,949	

	A	B	C	D
1	Date	VSS	SVS	Total
2	Jan-09	\$39K	\$27K	\$66K
3	Feb-09	\$39K	\$27K	\$66K
4	Mar-09	\$39K	\$26K	\$66K
5	Apr-09	\$38K	\$26K	\$64K
6	May-09	\$37K	\$26K	\$63K
7	Jun-09	\$36K	\$25K	\$61K
8	Jul-09	\$37K	\$24K	\$62K
9	Aug-09	\$37K	\$24K	\$60K
10	Sep-09	\$37K	\$23K	\$60K
11	Oct-09	\$37K	\$22K	\$59K
12	Nov-09	\$38K	\$24K	\$61K
13	Dec-09	\$39K	\$23K	\$62K

# WHAT'S GOOD (DATA) DESIGN?

8

## ROLLED-UP.

	A	B	C	D	E	F	G	H	I
1	Monthly Trends				Quarterly Trends				
2	Date	VSS	SVS	Total	Quarter	VSS	SVS	Total	
3	Jan-09	\$39K	\$27K	\$66K	Q1-09	\$117K	\$81K	\$198K	
4	Feb-09	\$39K	\$27K	\$66K	Q2-09	\$112K	\$77K	\$189K	
5	Mar-09	\$39K	\$26K	\$66K	Q3-09	\$111K	\$70K	\$182K	
6	Apr-09	\$38K	\$26K	\$64K	Q4-09	\$114K	\$68K	\$182K	
7	May-09	\$37K	\$26K	\$63K	Annual Totals				
8	Jun-09	\$36K	\$25K	\$61K	Year	VSS	SVS	Total	
9	Jul-09	\$37K	\$24K	\$62K	2009	\$454K	\$296K	\$751K	
10	Aug-09	\$37K	\$24K	\$60K					
11	Sep-09	\$37K	\$23K	\$60K					
12	Oct-09	\$37K	\$22K	\$59K					
13	Nov-09	\$38K	\$24K	\$61K					
14	Dec-09	\$39K	\$23K	\$62K					

# WHAT'S GOOD (DATA) DESIGN?

9

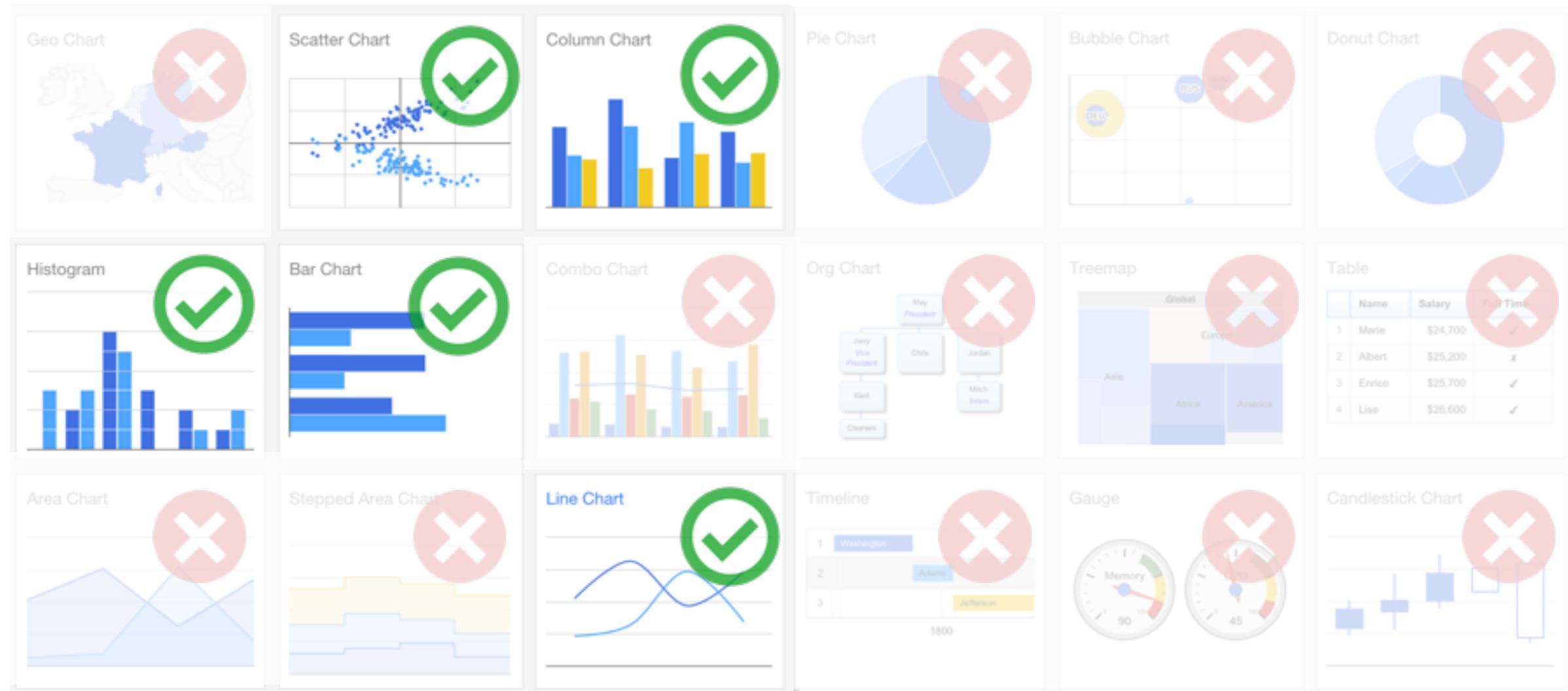
## MUTED.

	A	B	C	D	E	F	G	H	I
1	Monthly Trends				Quarterly Trends				
2	Date	VSS	SVS	Total	Quarter	VSS	SVS	Total	
3	Jan-09	\$39K	\$27K	\$66K	Q1-09	\$117K	\$81K	\$198K	
4	Feb-09	\$39K	\$27K	\$66K	Q2-09	\$112K	\$77K	\$189K	
5	Mar-09	\$39K	\$26K	\$66K	Q3-09	\$111K	\$70K	\$182K	
6	Apr-09	\$38K	\$26K	\$64K	Q4-09	\$114K	\$68K	\$182K	
7	May-09	\$37K	\$26K	\$63K	Annual Totals + Trends				
8	Jun-09	\$36K	\$25K	\$61K	Year	VSS	SVS	Total	
9	Jul-09	\$37K	\$24K	\$62K	2009	\$454K	\$296K	\$751K	
10	Aug-09	\$37K	\$24K	\$60K					
11	Sep-09	\$37K	\$23K	\$60K					
12	Oct-09	\$37K	\$22K	\$59K					
13	Nov-09	\$38K	\$24K	\$61K					
14	Dec-09	\$39K	\$23K	\$62K					

# CREATING EFFECTIVE CHARTS

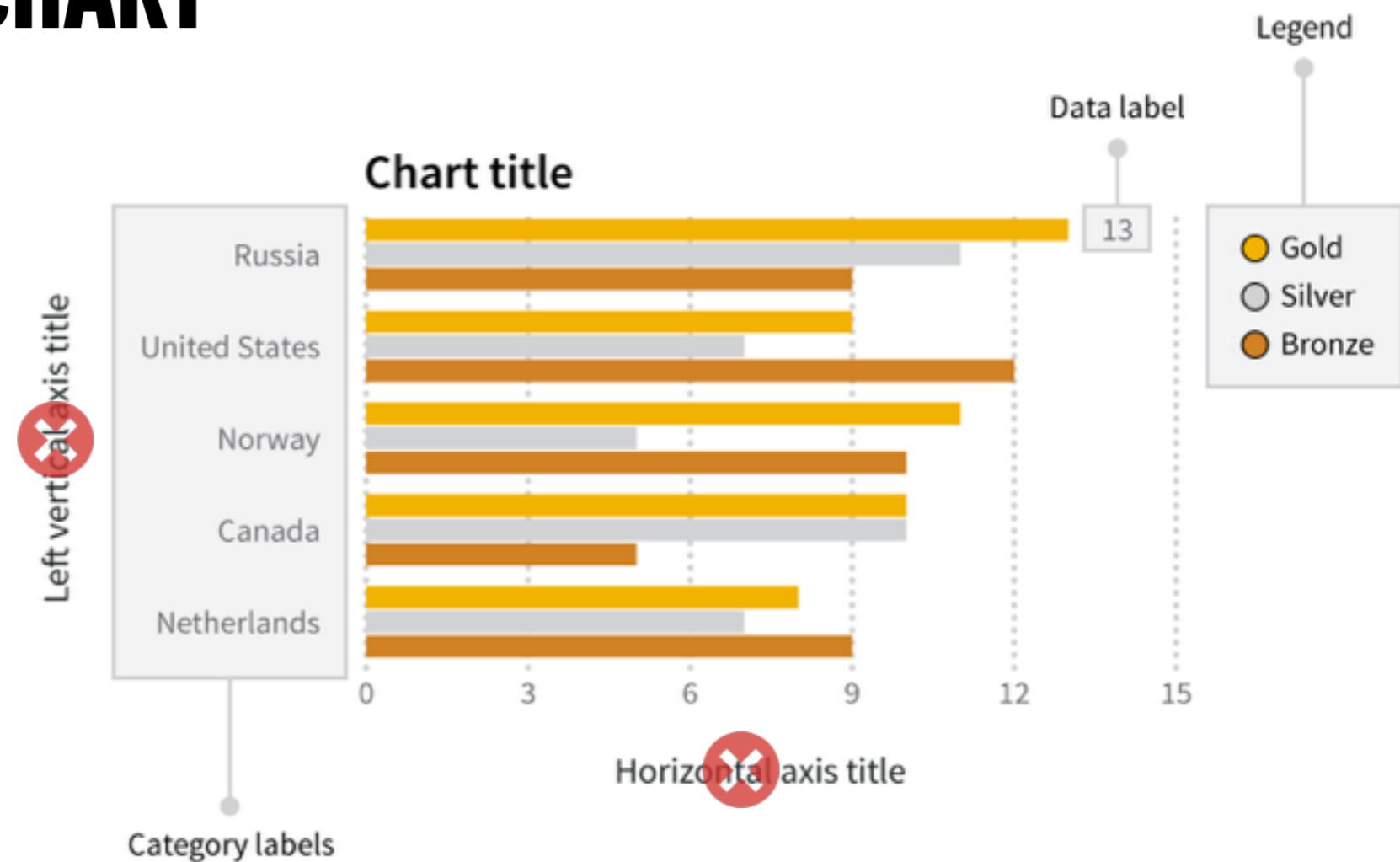
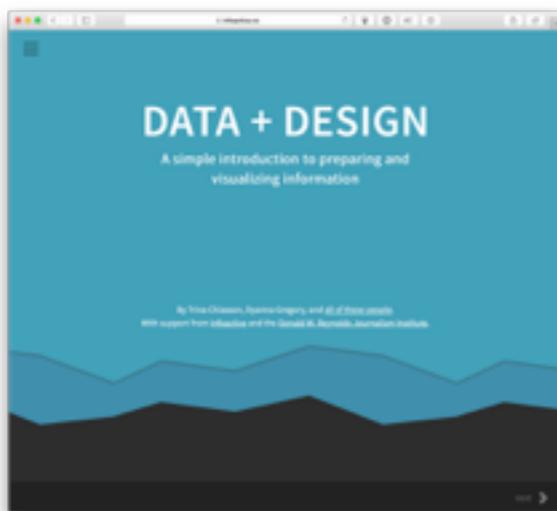
## CHART TYPES

<https://developers.google.com/chart/interactive/docs/gallery>



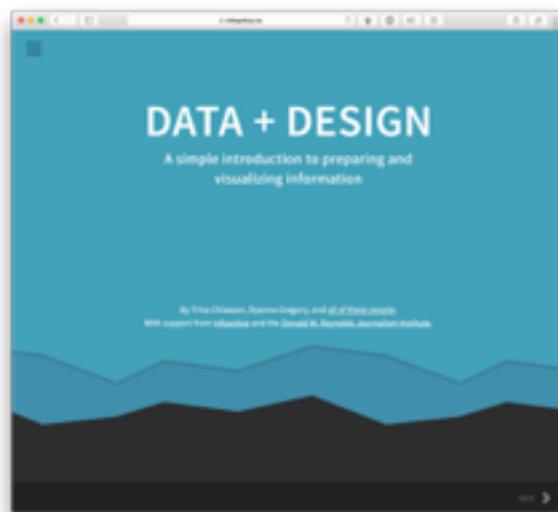
## ANATOMY OF A CHART

FROM



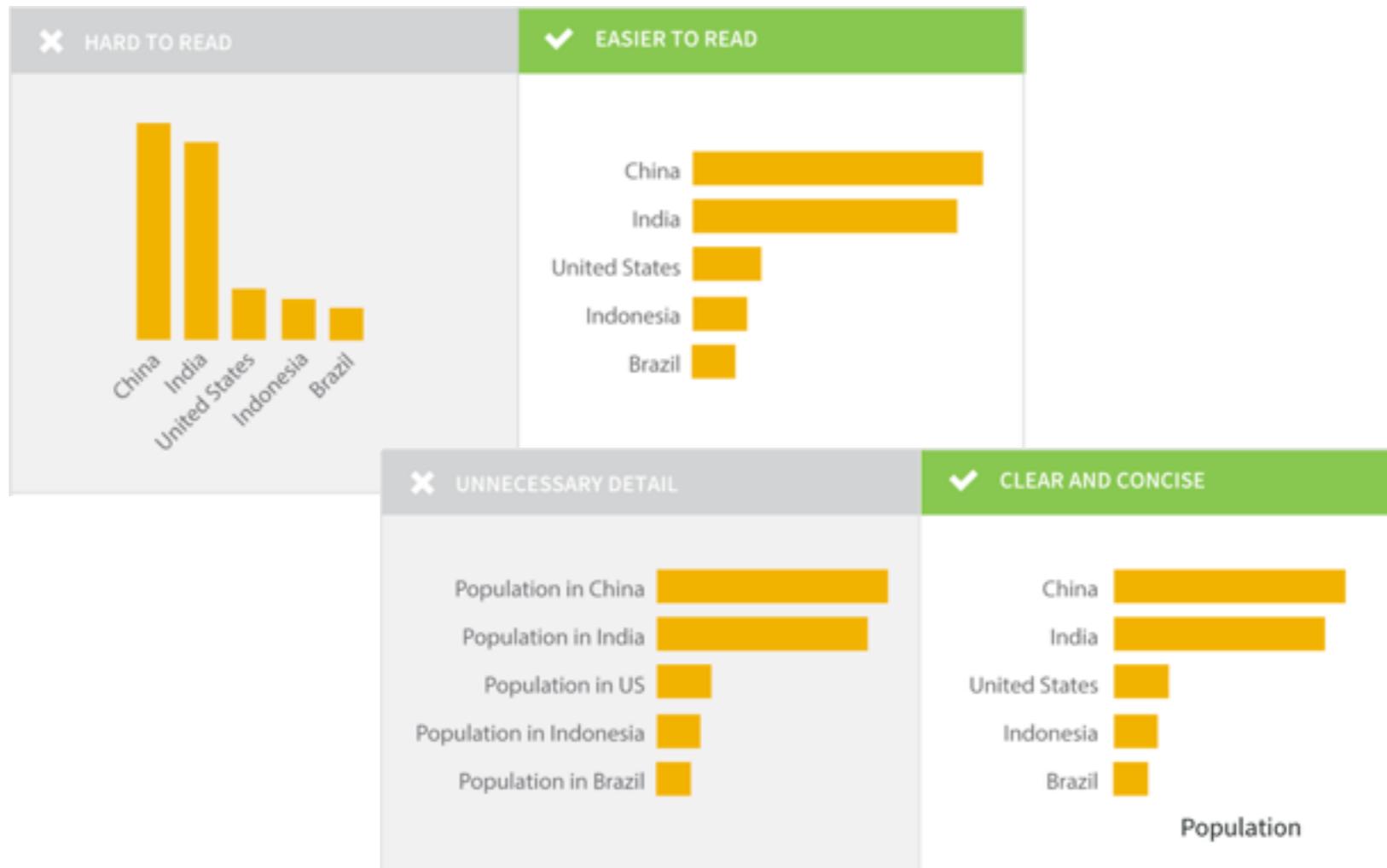
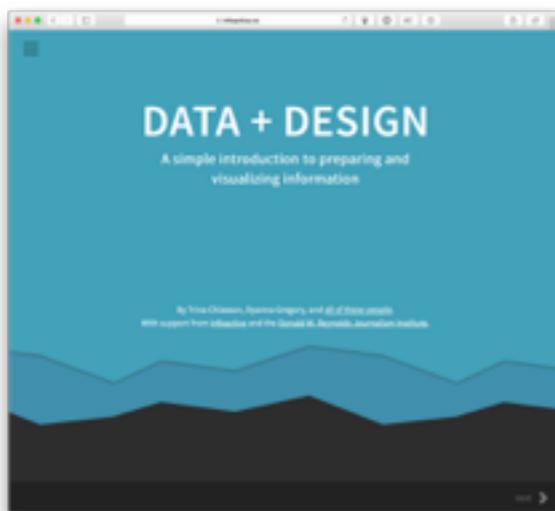
## AXIS TITLES

FROM



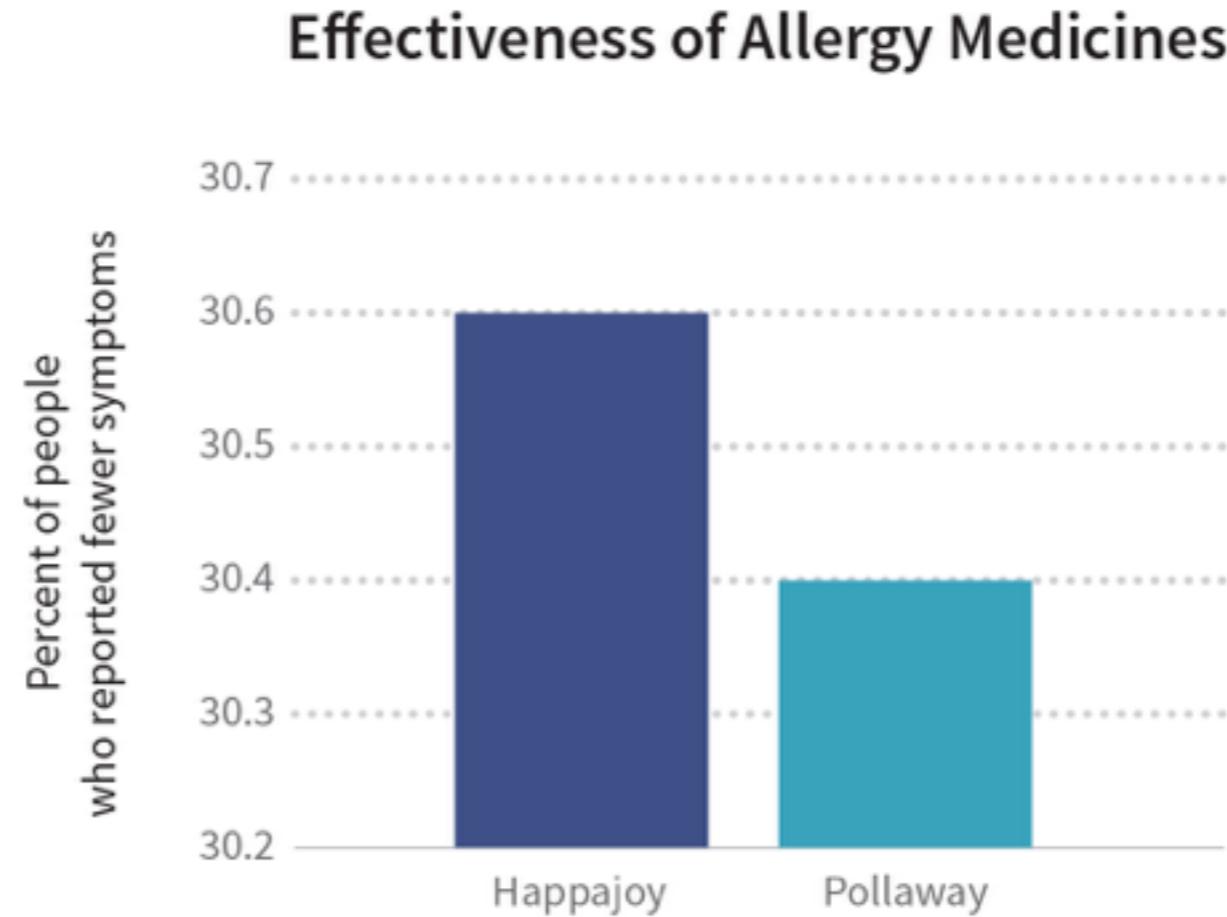
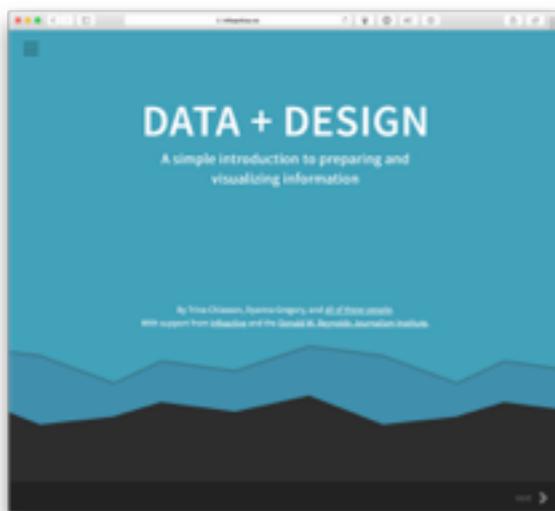
## AXIS TITLES

FROM



## DON'T TRUNCATE AXES

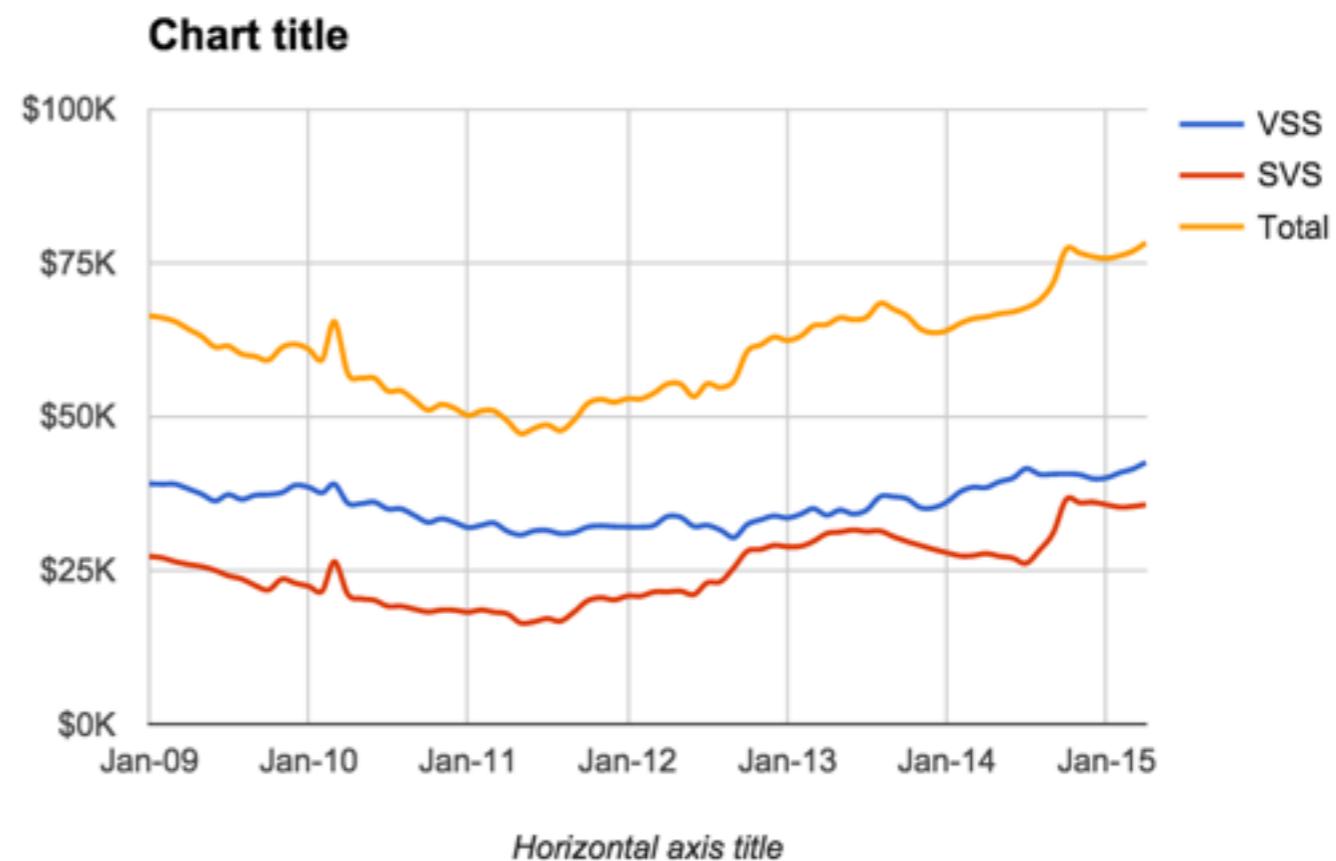
FROM



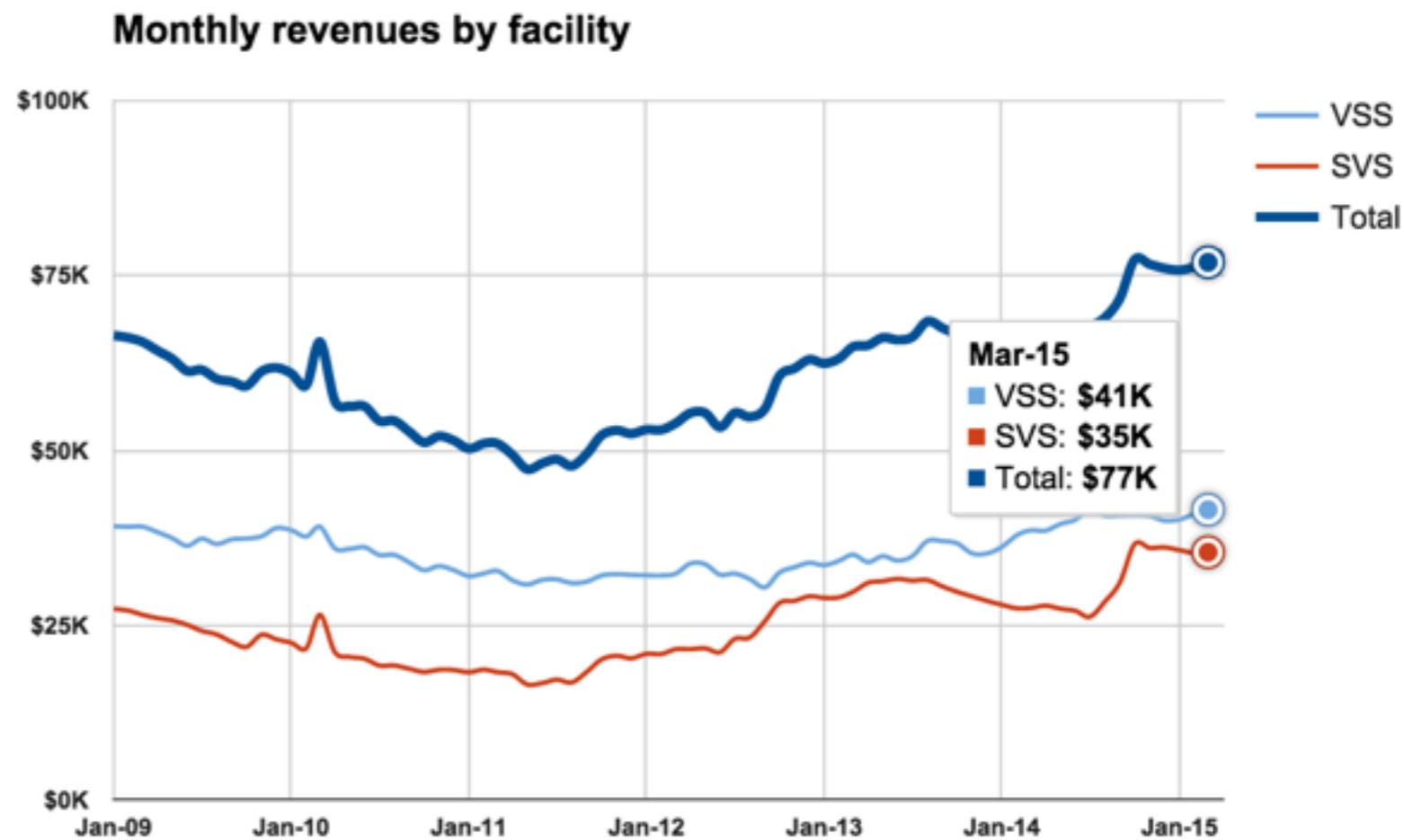
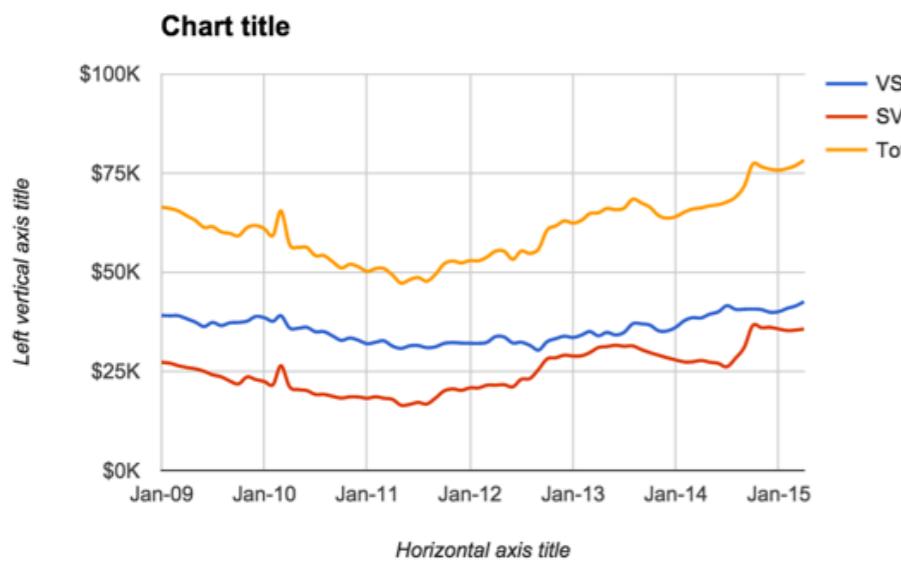
## BACK TO OUR DATA...

Monthly Trends			Quarterly Trends				
Date	VSS	SVS	Total	Quarter	VSS	SVS	Total
Jan-09	\$39K	\$27K	\$66K	Q1-09	\$117K	\$81K	\$198K
Feb-09	\$39K	\$27K	\$66K	Q2-09	\$112K	\$77K	\$189K
Mar-09	\$39K	\$26K	\$66K	Q3-09	\$111K	\$70K	\$182K
Apr-09	\$38K	\$26K	\$64K	Q4-09	\$114K	\$68K	\$182K
May-09	\$37K	\$26K	\$63K	Q1-10	\$115K	\$71K	\$186K
Jun-09	\$36K	\$25K	\$61K	Q2-10	\$108K	\$62K	\$170K
Jul-09	\$37K	\$24K	\$62K	Q3-10	\$104K	\$57K	\$161K
Aug-09	\$37K	\$24K	\$60K	Q4-10	\$99K	\$55K	\$155K
Sep-09	\$37K	\$23K	\$60K	Q1-11	\$97K	\$55K	\$152K
Oct-09	\$37K	\$22K	\$59K	Q2-11	\$94K	\$51K	\$145K
Nov-09	\$38K	\$24K	\$61K	Q3-11	\$94K	\$52K	\$146K
Dec-09	\$39K	\$23K	\$62K	Q4-11	\$98K	\$61K	\$157K
Jan-10	\$39K	\$22K	\$61K	Q1-12	\$97K	\$63K	\$160K
Feb-10	\$38K	\$22K	\$59K	Q2-12	\$100K	\$64K	\$164K
Mar-10	\$39K	\$26K	\$66K	Q3-12	\$94K	\$72K	\$166K
Apr-10	\$36K	\$21K	\$57K	Q4-12	\$100K	\$66K	\$165K
May-10	\$36K	\$20K	\$56K	Q1-13	\$103K	\$68K	\$190K
Jun-10	\$36K	\$20K	\$56K	Q2-13	\$103K	\$64K	\$197K
Jul-10	\$35K	\$19K	\$54K	Q3-13	\$109K	\$63K	\$202K
Aug-10	\$35K	\$19K	\$54K	Q4-13	\$107K	\$67K	\$194K
Sep-10	\$34K	\$19K	\$53K	Q1-14	\$113K	\$63K	\$195K
Oct-10	\$33K	\$18K	\$51K	Q2-14	\$118K	\$62K	\$200K
Nov-10	\$33K	\$19K	\$52K	Q3-14	\$123K	\$66K	\$209K
Dec-10	\$33K	\$19K	\$51K	Q4-14	\$121K	\$68K	\$230K
Jan-11	\$32K	\$18K	\$50K	Q1-15	\$122K	\$66K	\$229K
Feb-11	\$32K	\$19K	\$51K	Q2-15	\$43K	\$36K	\$78K

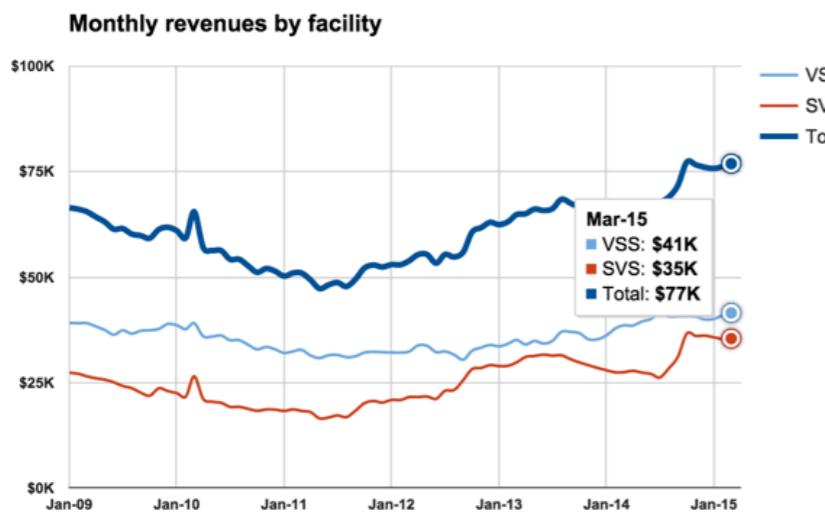
*Left vertical axis title*



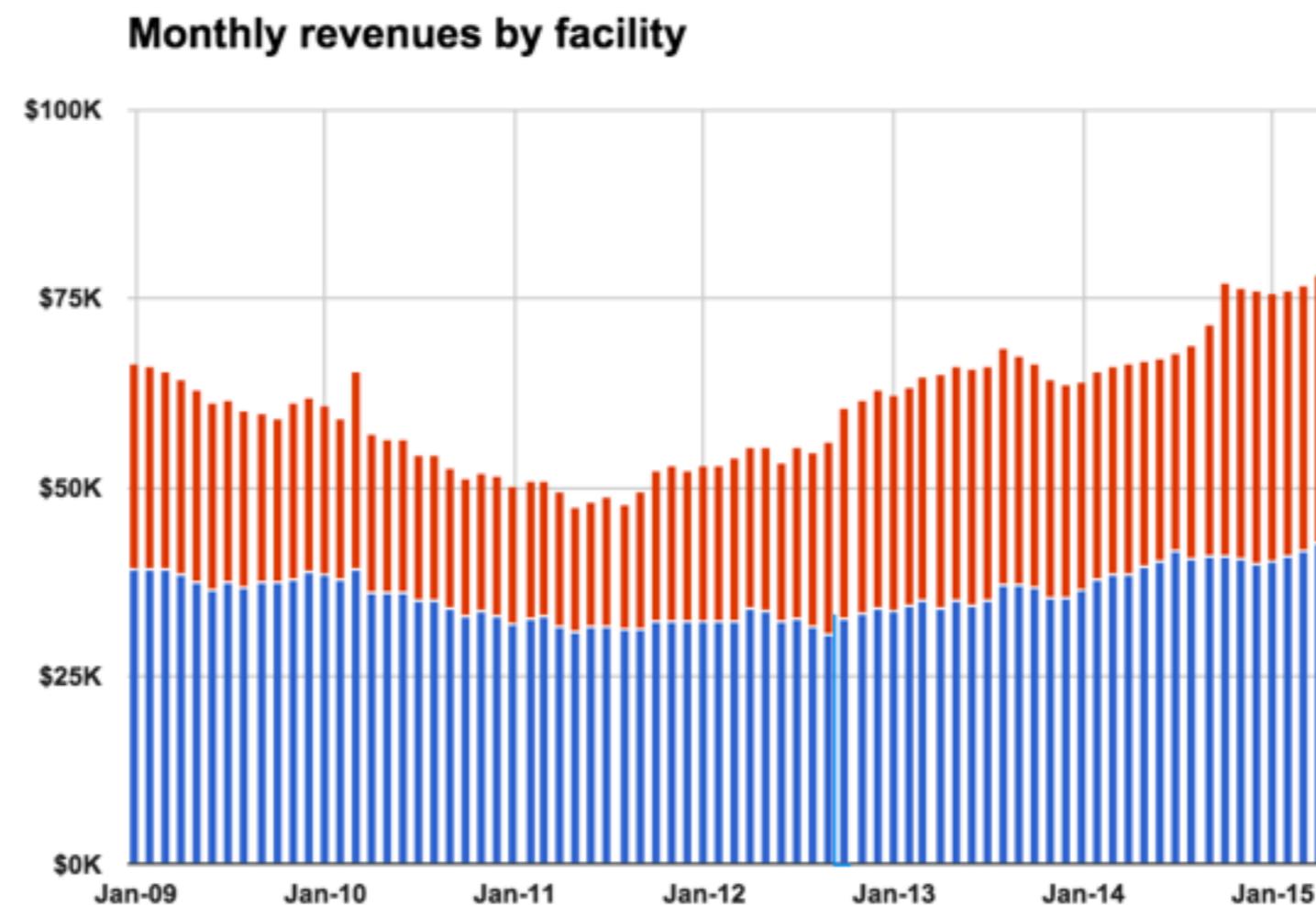
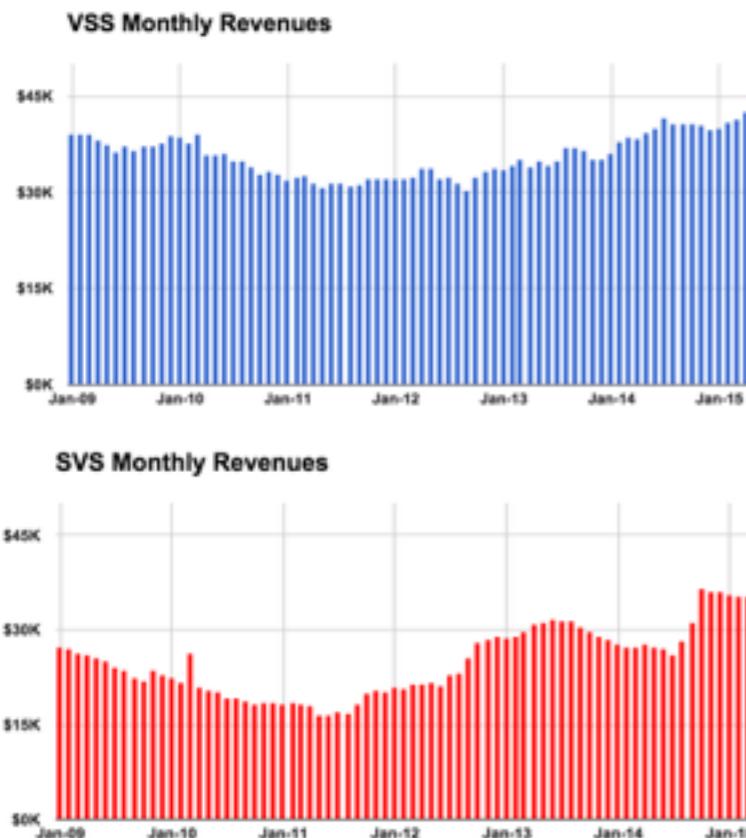
# APPLYING DESIGN PRINCIPLES



# ALTERNATIVE TIME-SERIES



## STACKED BARS



# DATA VIZ TOOLS + TECHNIQUES

## NESTED CHARTS + SMALL MULTIPLES



## SPARKLINES

Tasks completed by team members

(last 26 weeks, YoY change shown in %s)

Team Member	Total Tasks Completed	w1	w2	w3	w25	w26
Julie		13	15	19	11	19
John		11	18	11	14	16
Jabba the hut		15	14	14	19	12
Johnson		18	17	14	12	19
Jeremy		14	20	10	12	20
Josh		15	12	19	11	10



### Types of Sparklines

Regular Sparkline



Min and Max Points



First and Last Points



Markers



Column Chart



Win Loss Chart



Sum of Net Sales		Month	Jan 07	Feb 07	Mar 07	Apr 07	May
Salesman	Region						
Joseph			4655	3928	4462	4171	64
			2680	4604	4727	5668	5
			5423	5566	3503	4008	561
Lawrence			3840	3925	5928	5132	396
			4627	4219	5205	5309	770
			4896	5240	3516	6609	472
Maria			6580	2984	5375	5078	391
			3467	4710	4575	3661	523
			5152	6215	2783	6549	5025
Matt			4204	5886	3238	5634	477
			4512	5330	4052	3061	34
			6850	5277	4257	4901	64
Grand Total			56886	57884	51621	59781	624



### Sample Usage

```
SPARKLINE(A1:F1)
```

```
SPARKLINE(A2:E2, {"charttype", "bar"; "max", 40})
```

```
SPARKLINE(A2:E2,A4:B5)
```

<https://support.google.com/docs/answer/3093289?hl=en>

# SPARKLINES

=SPARKLINE(B7:B36)

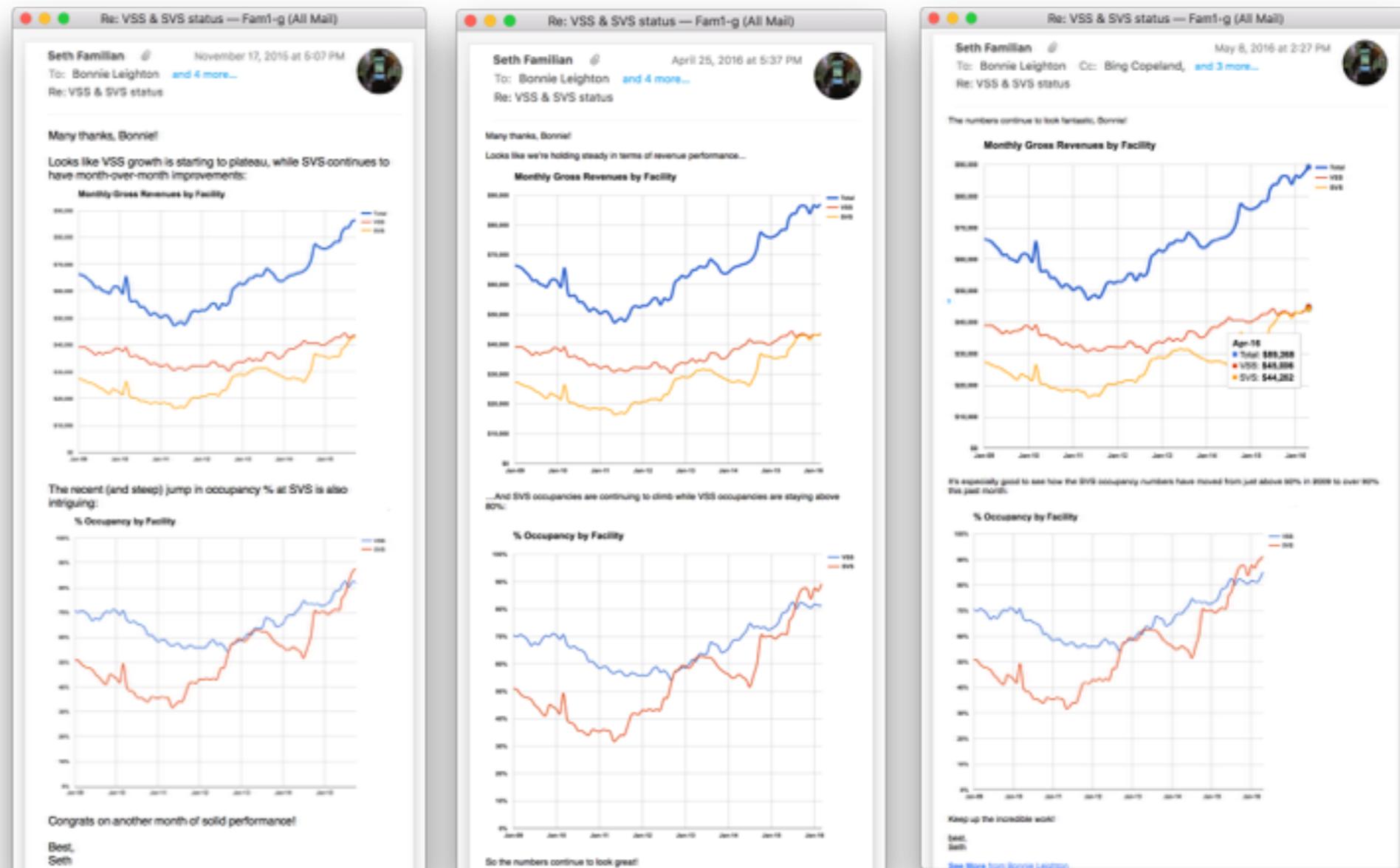
A	B	C	D	E	F	G	H	I	J	K	L	M
				SIGNUPS			VIRAL CONVERSIONS			COMPANIES		
Past 30 Days	5,702	5,402	41	199	60	9,109	5,443	60%	1,747			
	Total	Glip Viral	RC Viral	Glip Org.	RC Org.	Invites	Signups	Conv.	New Co's.			
04-30-2016 0:00:00	268	41	0	53	174	37	41	111%	16			
05-01-2016 0:00:00	329	79	0	71	179	59	79	134%	27			
05-02-2016 0:00:00	1,916	294	1	265	1,356	276	295	107%	83			
05-03-2016 0:00:00	1,892	340	1	350	1,201	299	341	114%	91			
05-04-2016 0:00:00	1,860	325	1	320	1,214	298	326	109%	109			
05-05-2016 0:00:00	1,701	285	2	255	1,159	264	287	109%	73			
05-06-2016 0:00:00	1,532	225	0	206	1,101	212	225	106%	43			
05-07-2016 0:00:00	283	74	0	61	148	38	74	195%	16			
05-08-2016 0:00:00	314	78	0	97	139	58	78	134%	26			
05-09-2016 0:00:00	1,937	309	2	284	1,342	384	311	102%	77			
05-10-2016 0:00:00	1,760	278	0	266	1,216	264	278	105%	71			
05-11-2016 0:00:00	1,861	265	1	255	1,340	279	266	95%	72			
05-12-2016 0:00:00	1,828	265	1	211	1,351	278	266	96%	89			
05-13-2016 0:00:00	227	209	1	10	7	372	210	56%	68			
05-14-2016 0:00:00	61	58	1	2	0	78	59	76%	23			
05-15-2016 0:00:00	84	78	0	4	2	93	78	84%	23			
05-16-2016 0:00:00	287	268	4	8	7	471	272	58%	68			
05-17-2016 0:00:00	283	270	1	9	3	492	271	55%	87			

# DATA VIZ TOOLS + TECHNIQUES

SHIFT + CONTROL + COMMAND + 4 → SELECT / SPACEBAR-CCLICK → PASTE

24

## NARRATIVE!



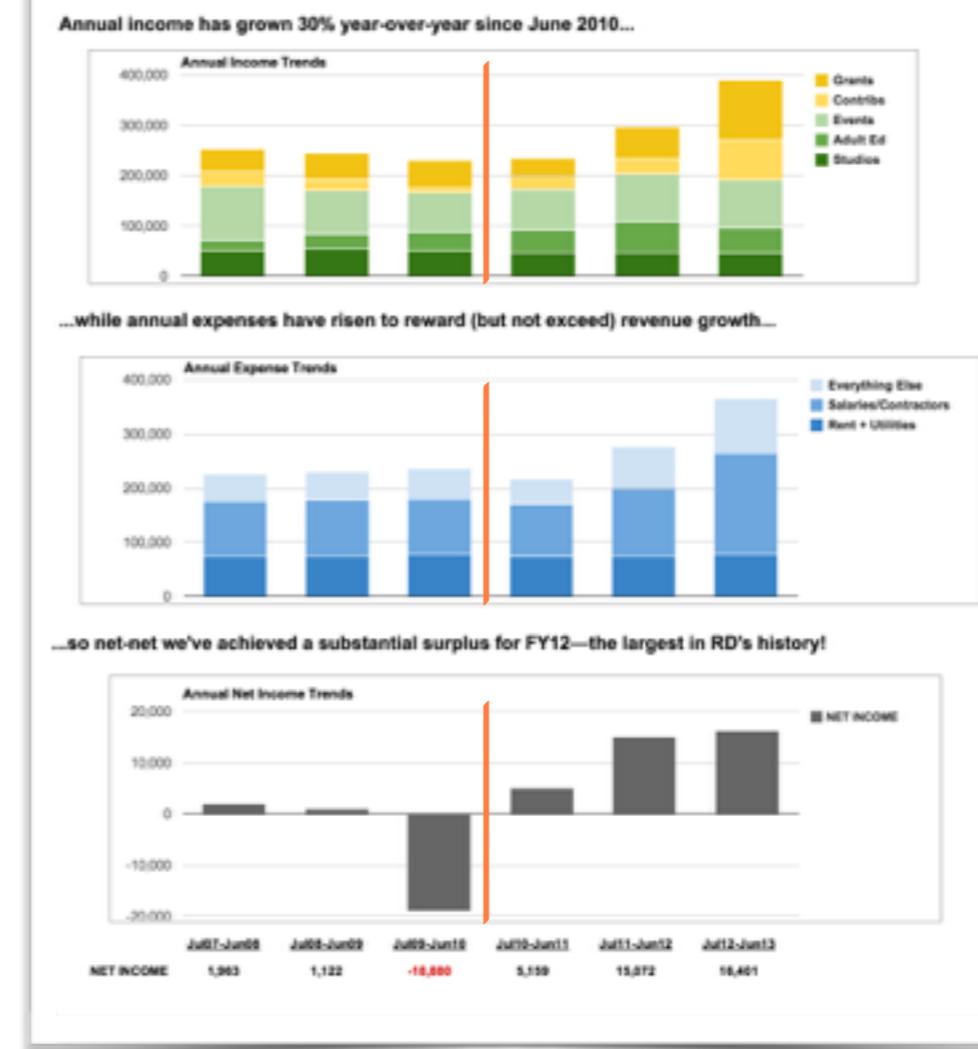
# DATA VIZ TOOLS + TECHNIQUES

SHIFT + CONTROL + COMMAND + 4 → SELECT / SPACEBAR-CCLICK → PASTE

25

## NARRATIVE!

Budget/Actual Combined	Annual Aggregates						
	FY12	FY13	FY14	FY15	FY16	FY17	FY18
<b>Income</b>							
4000 — Event Revenue	23,250	19,870	16,910	17,470	20,020	21,900	26,000
4540 — Misc Event Revenue	5,914	4,100	7,284	8,000	9,996	8,725	7,900
2600 — Education	26,046	27,474	27,876	33,073	34,944	36,129	36,000
3500 — All Art Sales	57,080	40,852	33,444	33,957	35,850	40,305	42,000
4300 — Status of Inventory	0	23	230	208	177	986	1,100
4200 — Rent + Program Use	50,770	54,977	49,129	48,407	47,879	44,955	47,950
4010 — Use of Space Income	3,875	7,000	1,980	4,359	7,000	1,840	8,000
7010 — Interest Income	500	200	11	27	170	410	400
4501 — Public Support (GRANT)	14,180	14,820	12,774	10,020	12,798	13,258	12,000
4010 — Fundraising Events	7,850	7,850	7,644	8,399	9,170	11,930	10,000
4300 — Contributions Income	29,080	23,810	10,063	27,013	30,150	78,910	80,000
4110 — Grants	46,400	52,800	53,800	34,900	63,200	118,300	100,000
4800 — Street Artists Income	2,204	0	0	0	0	0	0
4100 — Remunerated Expenses	0	0	0	0	0	0	0
<b>TOTAL INCOME</b>	<b>354,820</b>	<b>249,860</b>	<b>229,330</b>	<b>233,422</b>	<b>291,058</b>	<b>386,228</b>	<b>378,300</b>
<b>Cost of Goods Sold</b>							
3002 — Cost of Goods Sold-NonArt	300	710	0	0	0	0	0
5000 — Cost of Goods Sold	27,520	13,840	12,523	12,070	7,372	8,470	8,000
<b>TOTAL COGS</b>	<b>27,820</b>	<b>13,840</b>	<b>12,523</b>	<b>12,070</b>	<b>7,372</b>	<b>8,470</b>	<b>8,000</b>
<b>GROSS PROFIT</b>	<b>227,200</b>	<b>231,985</b>	<b>217,814</b>	<b>221,352</b>	<b>290,687</b>	<b>381,458</b>	<b>379,300</b>
<b>Expense</b>							
4270 — Salary & Fringe	95,292	85,315	58,737	79,426	94,742	108,791	120,000
4271 — Professional Fees	10,200	17,416	18,160	20,798	20,894	21,875	20,000
8200 — Artist Fees	9,100	9,742	16,118	15,500	24,710	33,571	30,000
2300 — Education Expenses	2,147	3,887	3,098	3,379	3,882	7,884	7,000
4260 — Event & Exhibition Expenses	8,520	7,867	7,955	5,711	3,454	11,215	9,500
4160 — Marketing (D)	1,660	1,660	1,664	1,454	2,176	3,001	4,100
6280 — Printing and Reproduction	8,640	10,104	9,699	4,953	12,545	15,300	12,000
6281 — Postage and Shipping	3,800	4,800	3,632	1,881	1,800	1,481	1,000
6290 — Rent	67,200	67,200	64,200	69,800	67,200	67,200	67,000
6300 — Utilities	7,077	5,762	7,080	7,180	8,800	7,036	7,000
6340 — Telephone	1,487	916	932	1,280	1,118	794	800
4260 — Insurance & Indemnity	800	747	889	870	2,361	2,884	2,000
6100 — Insurance	5,523	5,419	4,627	5,579	4,954	4,650	8,000
4230 — Facilities	2,139	1,163	329	898	2,098	3,781	2,000
6710 — Supplies	1,590	2,467	2,311	1,836	2,944	6,229	8,000
4400 — Capital Purchases	8,467	8,471	1,438	218	2,391	3,379	1,000
6121 — Bank Fees & Commissions	1,300	2,330	4,482	3,813	4,382	4,258	4,000
6262 — Catering & Hospitality	100	1,084	968	701	1,342	1,817	1,000
6365 — Travel & Ent.	12	1,084	105	369	660	1,744	1,000
6163 — Professional Development	0	0	80	800	412	2,348	2,000
6164 — Conferences & Meetings	0	0	94	110	30	271	200
6165 — dues and Subscriptions	112	247	248	90	260	210	200
6220 — Licenses and Permits (S)	220	58	47	23	56	235	100
2300 — Expenses	2,190	280	0	0	0	0	0
<b>TOTAL EXPENSES</b>	<b>326,240</b>	<b>250,473</b>	<b>256,984</b>	<b>216,414</b>	<b>364,267</b>	<b>378,304</b>	<b>378,300</b>
Balances as a % of Total Expenses	64.81%	44.40%	44.57%	44.51%	45.21%	51.39%	55.11%



## 2x2s



### GARTNER'S MAGIC QUADRANT

Figure 1. Magic Quadrant for Business Intelligence and Analytics Platforms



## 2x2s



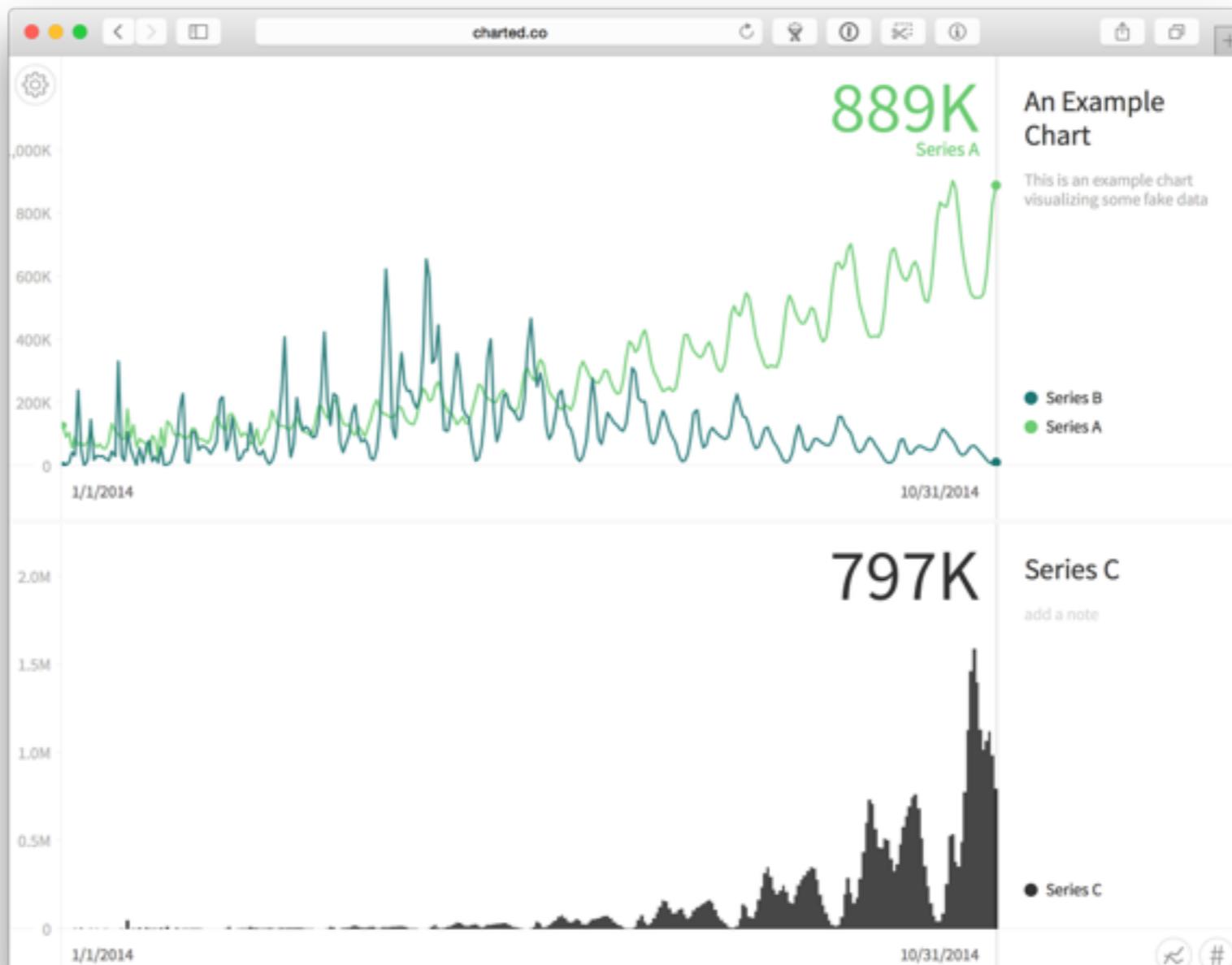
*privately  
visualize  
your love life*



## CHARTED

FOR SUPER SIMPLE CHARTS

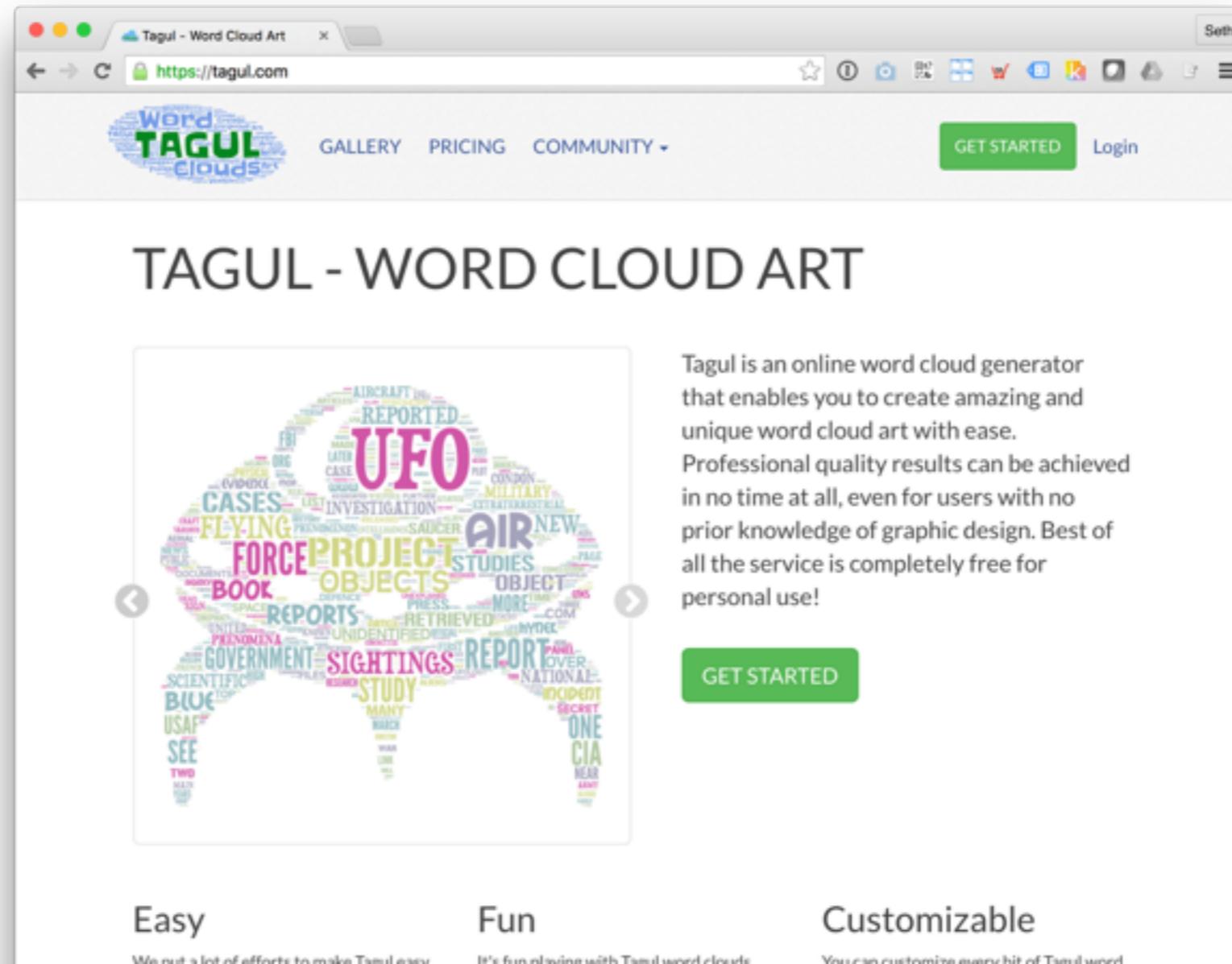
[CHARTED.CO](#)



## TAGUL

FOR GORGEOUS WORD CLOUDS

[TAGUL.COM](https://tagul.com)



The screenshot shows the Tagul website's homepage. At the top, there's a navigation bar with links for GALLERY, PRICING, and COMMUNITY, along with a GET STARTED button and a Login link. The main title "TAGUL - WORD CLOUD ART" is centered above a large, colorful word cloud. This word cloud is shaped like a map of the United States and contains numerous terms related to UFOs, such as "UFO", "PROJECT", "REPORTS", "STUDIES", "GOVERNMENT", "SCIENTIFIC", and "CIA". Below the word cloud, there are three descriptive cards: "Easy", "Fun", and "Customizable". Each card has a small explanatory text and a "GET STARTED" button.

**TAGUL - WORD CLOUD ART**

Tagul is an online word cloud generator that enables you to create amazing and unique word cloud art with ease. Professional quality results can be achieved in no time at all, even for users with no prior knowledge of graphic design. Best of all the service is completely free for personal use!

GET STARTED

Easy

Fun

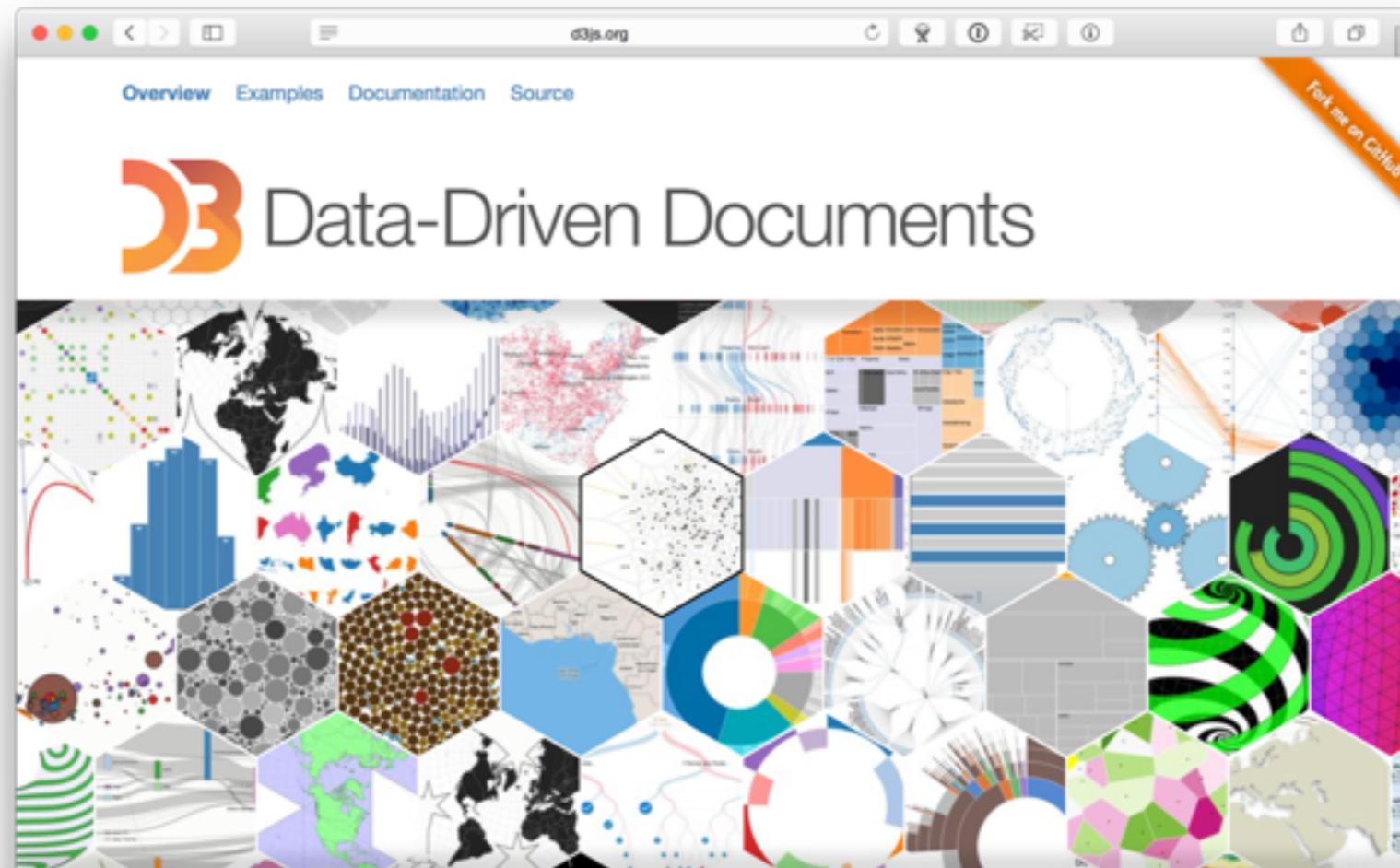
Customizable

# ADVANCED DATA VIZ TOOLS

## D3.JS

FOR AMAZING(LY COMPLEX) DATA VIZ

[D3JS.ORG](https://d3js.org)



D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

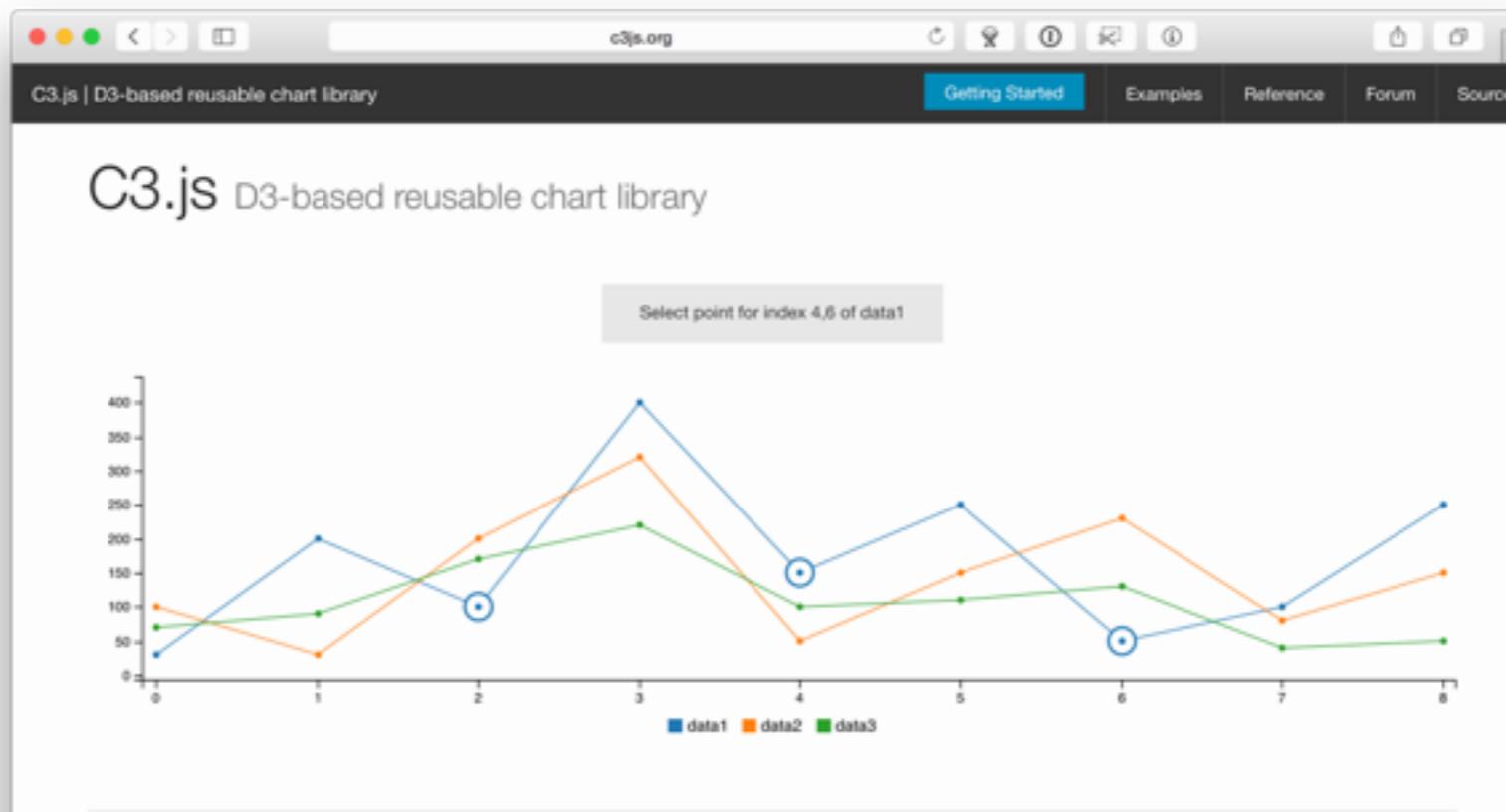
[See more examples.](#)

Download the latest version (3.5.5) here:

## C3.JS

D3 META-LIBRARY

[C3JS.ORG](http://c3js.org)



### Why C3?

#### Comfortable

C3 makes it easy to generate D3-based charts by wrapping the code required to construct the entire chart. We don't need to write D3 code any more.

#### Customizable

C3 gives some classes to each element when generating, so you can define a custom style by the class and it's possible to extend the structure directly by D3.

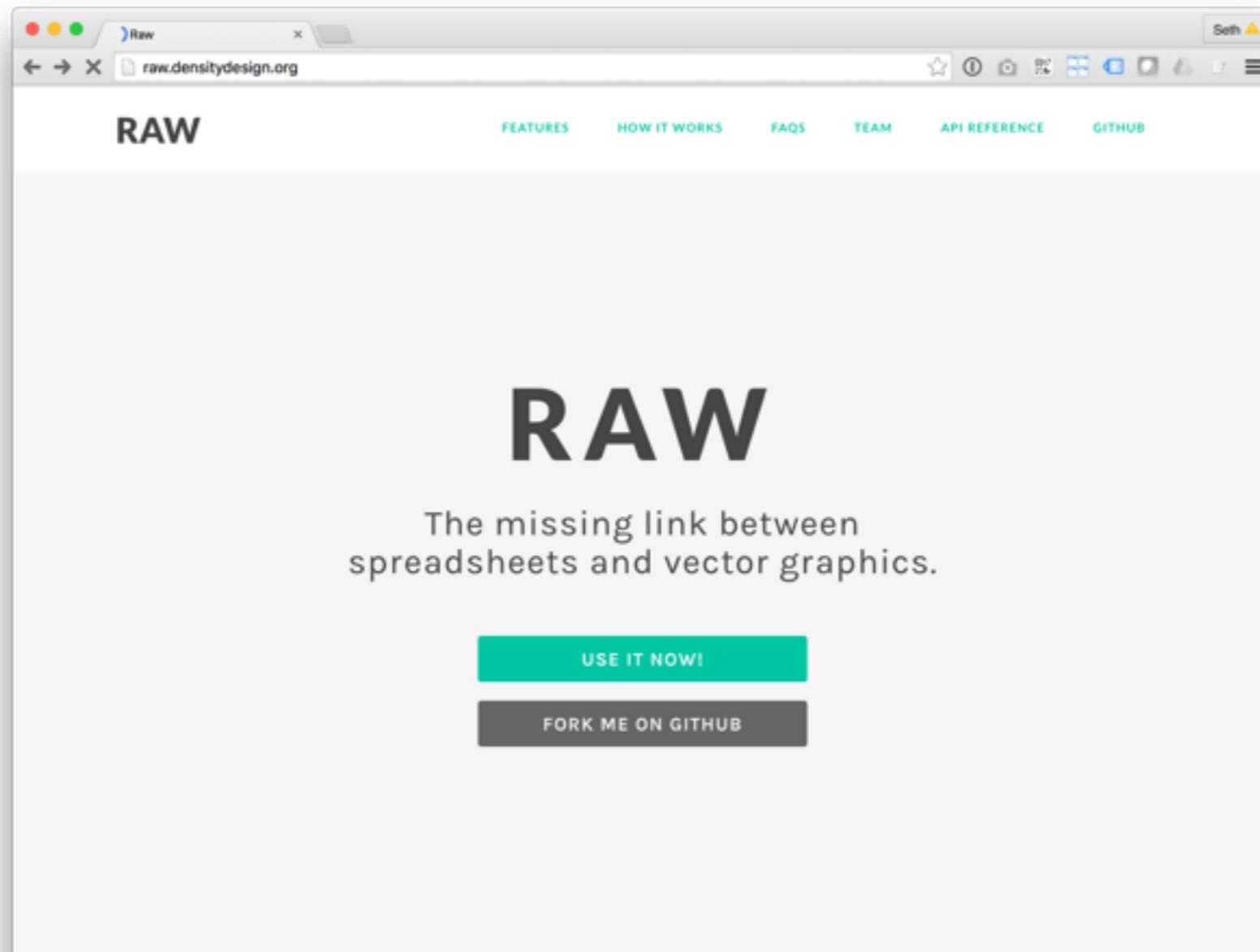
#### Controllable

C3 provides a variety of APIs and callbacks to access the state of the chart. By using them, you can update the chart even if after it's rendered.

# RAW

D3 META-LIBRARY

[RAW.DENSITYDESIGN.ORG](http://raw.densitydesign.org)



## QUID

FOR UNSTRUCTURED ANALYSIS

[QUID.COM](https://quid.com)

The screenshot shows a web browser window for the Quid website. The title bar reads "Quid" and the URL is "https://quid.com/product#/knowledge-download". The page features a navigation bar with links for "PRODUCT", "OUR STORY", "CAREERS", "LOG IN", and "TRY QUID". A large section on the left is titled "Competitive Intelligence" and contains a bulleted list:

- A look at what your competitors are doing, saying, investing in and acquiring. Follow company footprints over time and gain insights on how to get ahead of them.

To the right is a network visualization on a tablet screen. The network consists of several nodes (circles) connected by lines (edges). The nodes are colored according to their primary mention: Google (yellow), Facebook (green), IBM (cyan), Yahoo (magenta), Microsoft (blue), Hewlett-Packard (light blue), NTT (orange), and Qualcomm (light green). A legend titled "Coloring by Company" provides the percentage distribution of mentions for each company. The network visualization is overlaid on a dark background with a sidebar containing options like "Network", "Nodes", "Labels", "Links", "Color by", and "Company (Primary Mention)".

Company	Percentage
Google	13%
Facebook	10%
IBM	9%
Yahoo	6.5%
Microsoft	5.4%
Hewlett-Packard	4.9%
NTT	4.7%
Qualcomm	4.7%

# SPLUNK

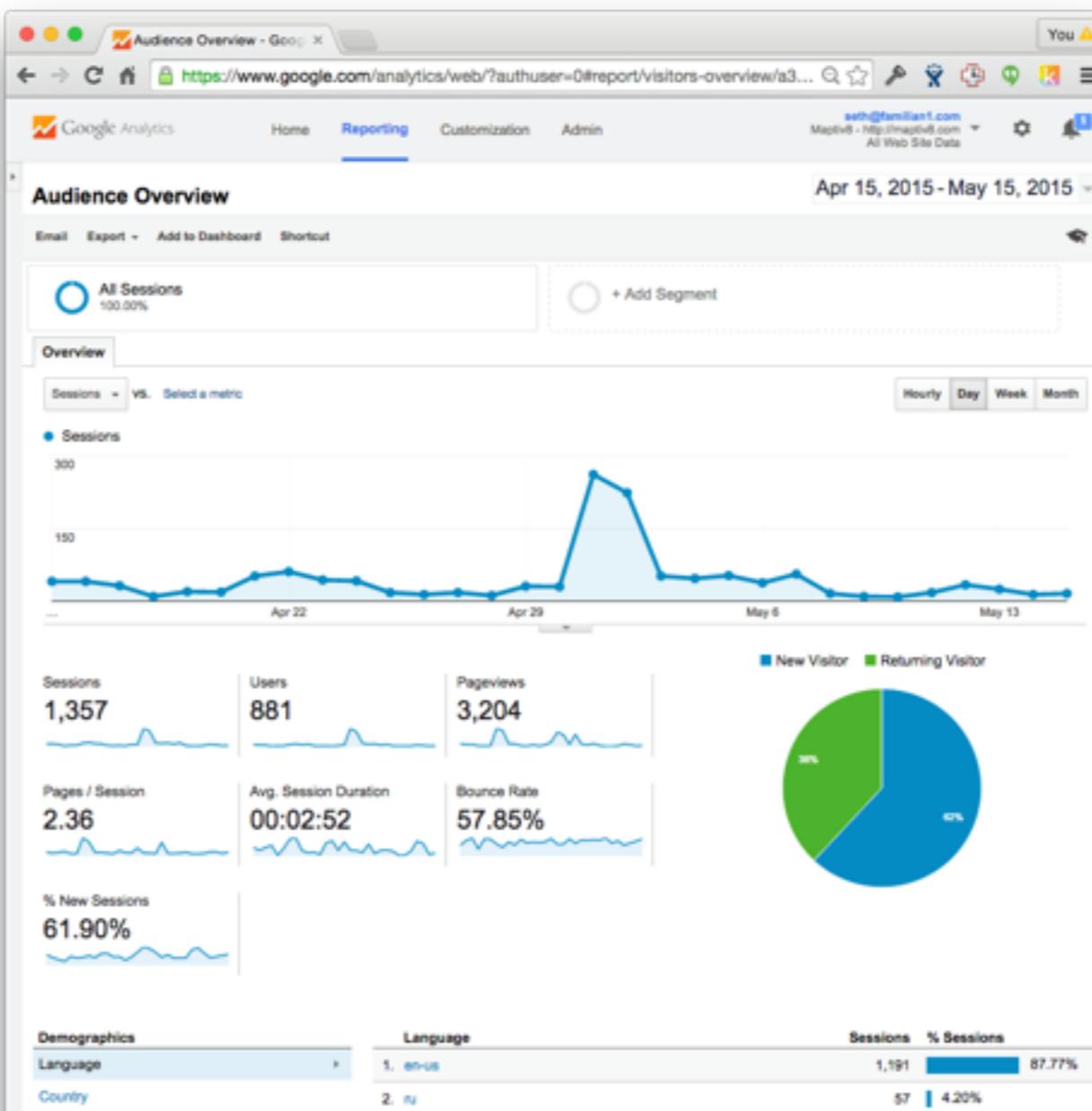
FOR ANY MACHINE DATA

[SPLUNK.COM](https://splunk.com)

The screenshot shows the Splunk website's main navigation bar with tabs: Collect and Index Data, Search and Investigate, Correlate and Analyze, Visualize and Report (which is highlighted in green), Monitor and Alert, and Access from Anywhere. Below the navigation, a heading says "Visualize and Report". A descriptive text block follows, stating: "Visualize trends and characteristics in custom dashboards and reports suited to any business, operational or security need. Analyze further with chart overlay and pan and zoom controls. Predictive visualizations let you forecast highs and lows, plan system resources and anticipate workloads. You can also personalize dashboards and reports for anyone, share them as PDFs, or embed them into other applications." To the right of this text is a large image of a tablet displaying a complex dashboard titled "Activations Over Time". The dashboard includes a bar chart for "Sales" and a line graph for "Prediction", both over time. Below this are two smaller charts: "Revenue by Plan, Price, and Location" (a bubble chart) and "Sales Map" (a world map with colored dots representing sales data). A vertical blue sidebar on the right contains the text "Ask an Expert".

# GOOGLE ANALYTICS

FOR WEBSITE TRAFFIC

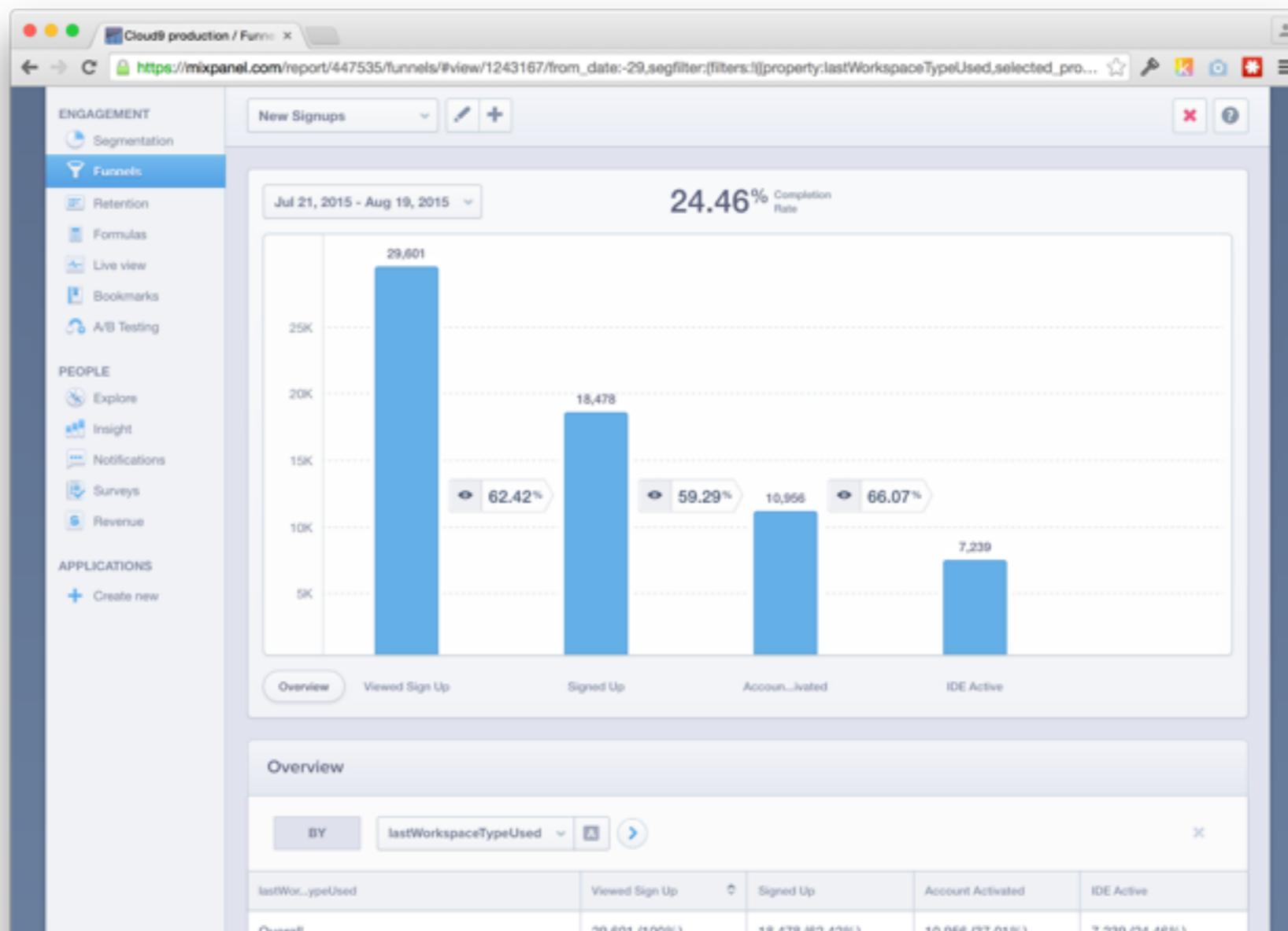


[ANALYTICS.GOOGLE.COM](https://analytics.google.com)

# MIXPANEL

FOR USER-EVENT DATA

MIXPANEL.COM



# GECKOBOARD

FOR AGGREGATING DATA SOURCES

GECKOBOARD.COM

The screenshot shows the Geckoboard website at https://www.geckoboard.com/product/. The page features a dark header with the Geckoboard logo and navigation links for PRODUCT, BENEFITS, INTEGRATIONS, PRICING, LOGIN, and TRY GECKOBOARD. Below the header, a large heading reads "THE DATA DASHBOARD YOU'VE BEEN LOOKING FOR". Two devices are displayed below the heading: a laptop and a smartphone, both showing a complex data dashboard. The dashboard includes various metrics such as "1,799", "130", "1,012", "3:15PM", "18", "2,983", "2,078", and several line graphs and bar charts. A sidebar on the left lists dates from October 29th to December 26th. At the bottom of the page are two buttons: "VIEW EXAMPLE DASHBOARDS" and "TRY GECKOBOARD".

# SEGMENT

[SEGMENT.COM](https://segment.com)

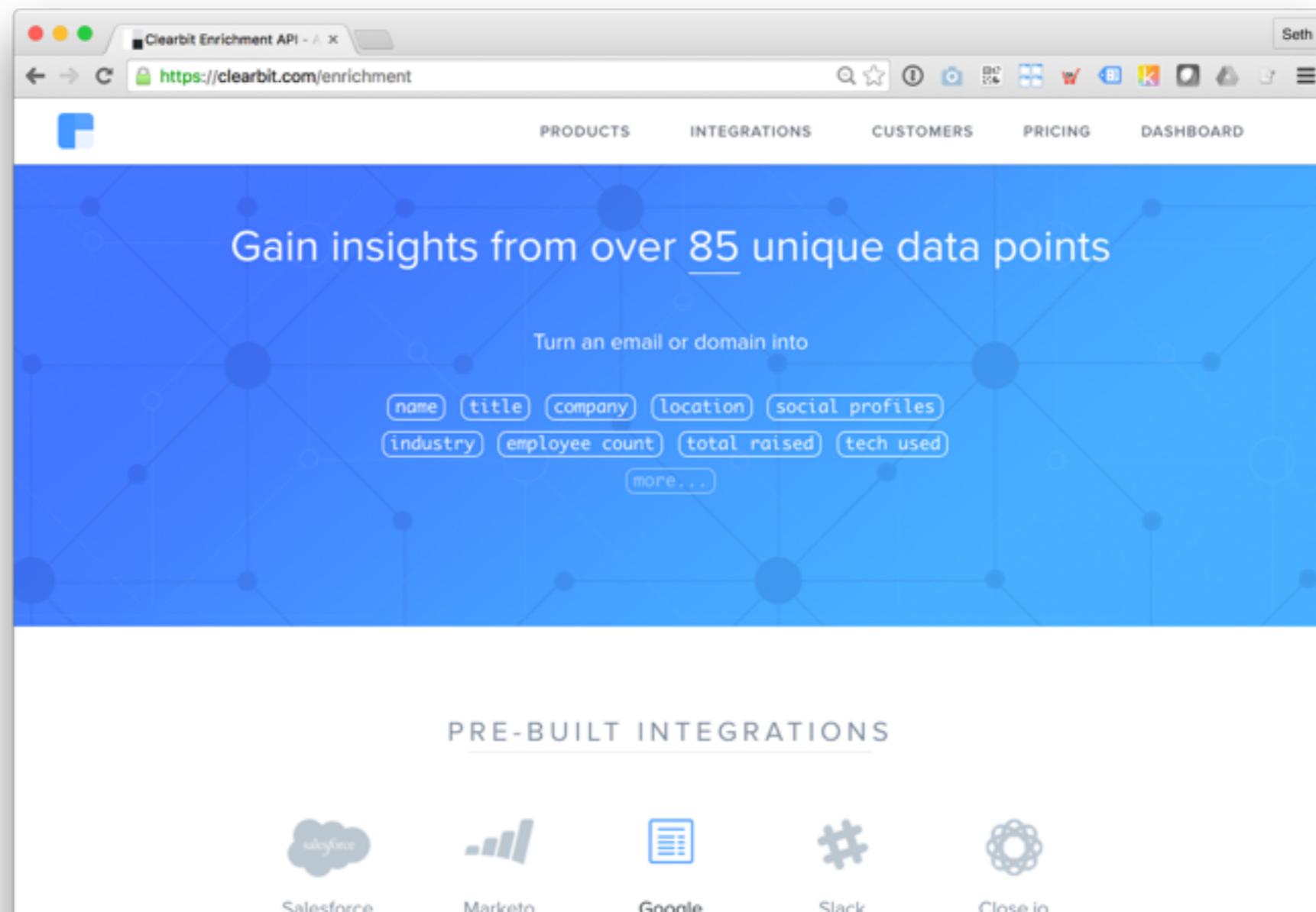
The screenshot shows a web browser window for Segment.com. The title bar reads "Analytics API and customer data integration". The address bar shows "https://segment.com". The main content area features a headline "Integrations for every team." followed by a subtext "Access the data and tools you need without being blocked by technical implementations or complexities." A green "View Integrations →" button is present. Below this, there's a search bar and a table titled "INTEGRATIONS". The table has three columns: "SCHEMA", "DEBUGGER", and "DEVELOPER". It lists various data sources with their corresponding logos and toggle switches. Most switches are in the 'on' position, indicated by green icons.

SCHEMA	DEBUGGER	DEVELOPER
	Google Analytics <input checked="" type="checkbox"/>	Marketo <input checked="" type="checkbox"/>
	Adobe Analytics <input checked="" type="checkbox"/>	HEAP <input checked="" type="checkbox"/>
	Customer.io <input checked="" type="checkbox"/>	facebook Conversions <input checked="" type="checkbox"/>
	INTERCOM <input checked="" type="checkbox"/>	kissmetrics <input type="checkbox"/>
		mixpanel <input type="checkbox"/>
		FLURRY <input checked="" type="checkbox"/>
		CRITTERCISM <input checked="" type="checkbox"/>

# CLEARBIT

FOR DATA ENRICHMENT

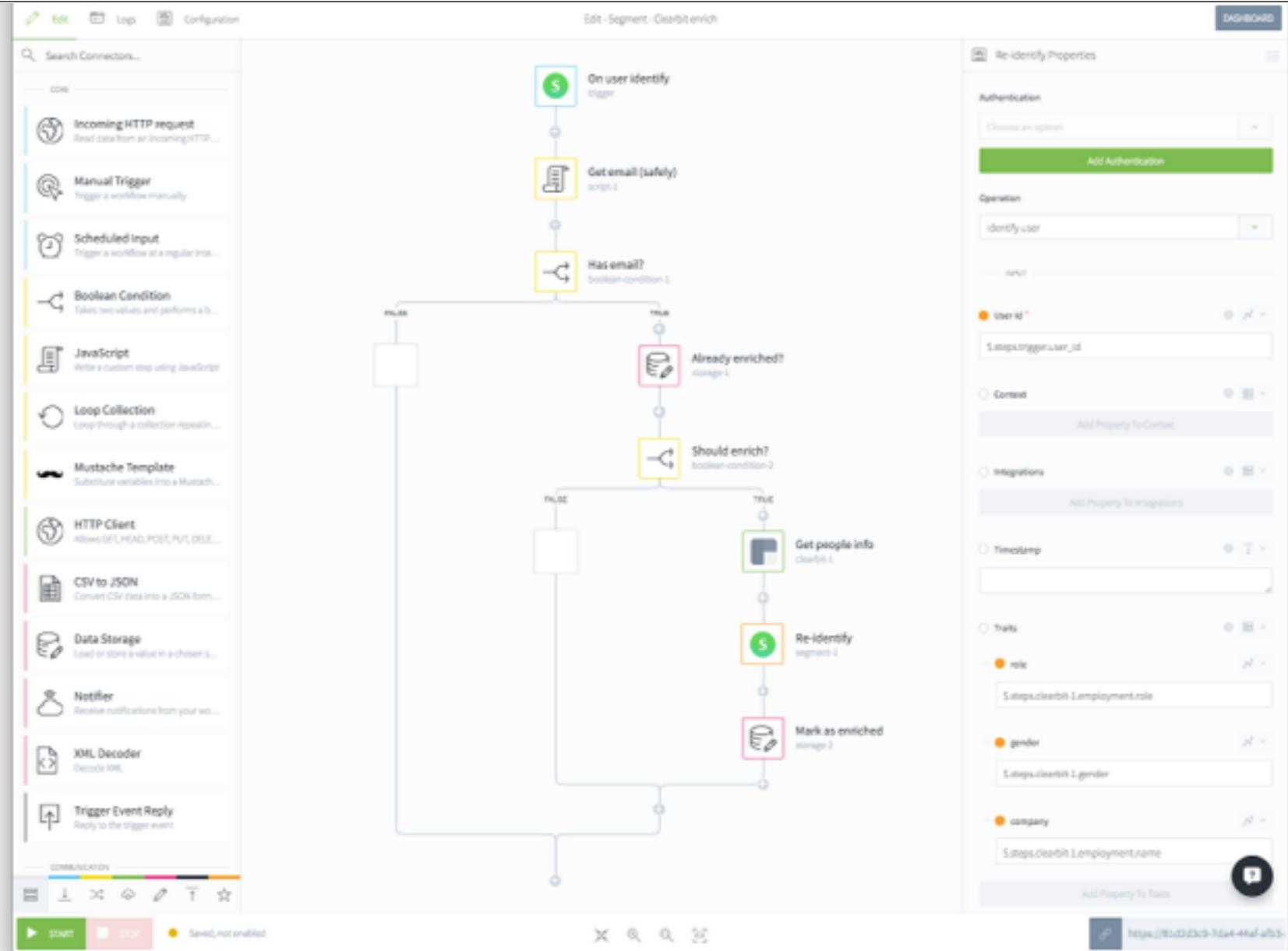
[CLEARBIT.COM](https://clearbit.com)



# USEFUL TOOLS

## TRAY.IO

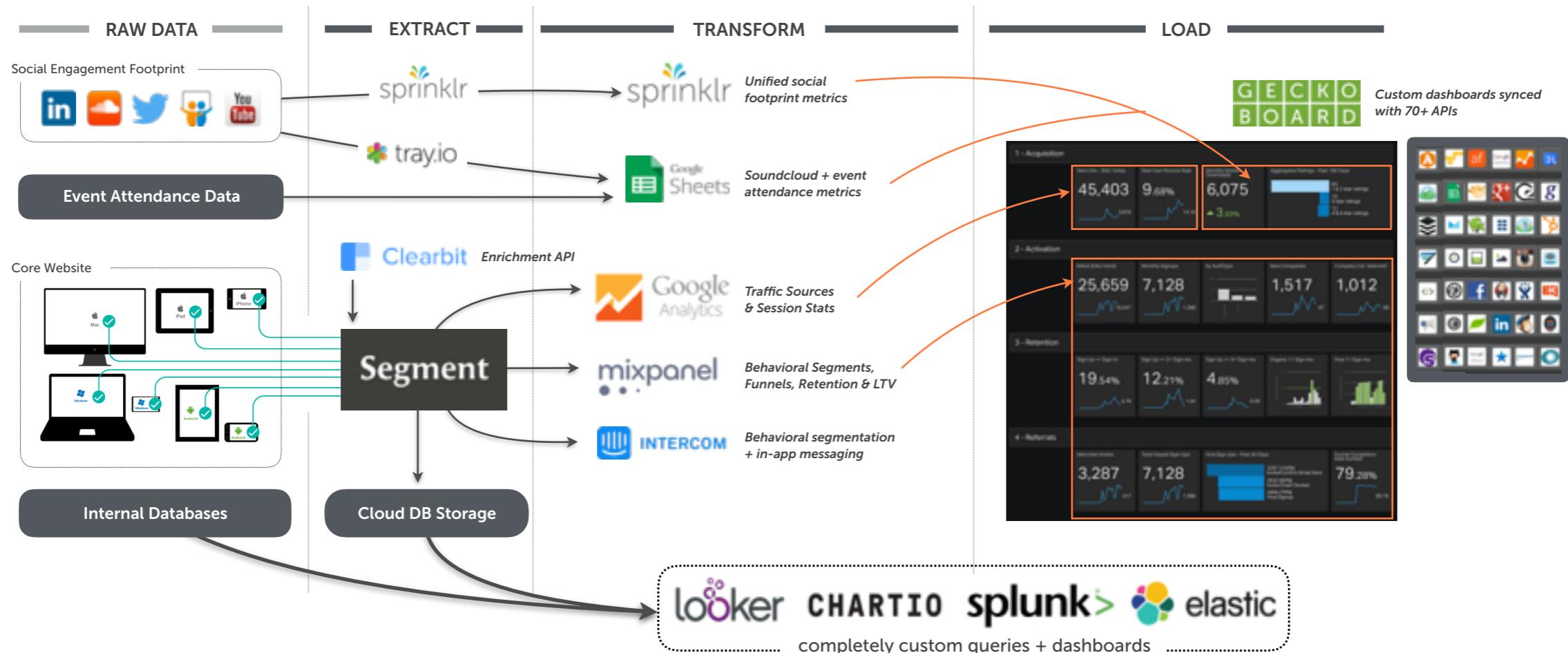
COMPLETELY CUSTOM FLOWS



[TRAY.IO](#)

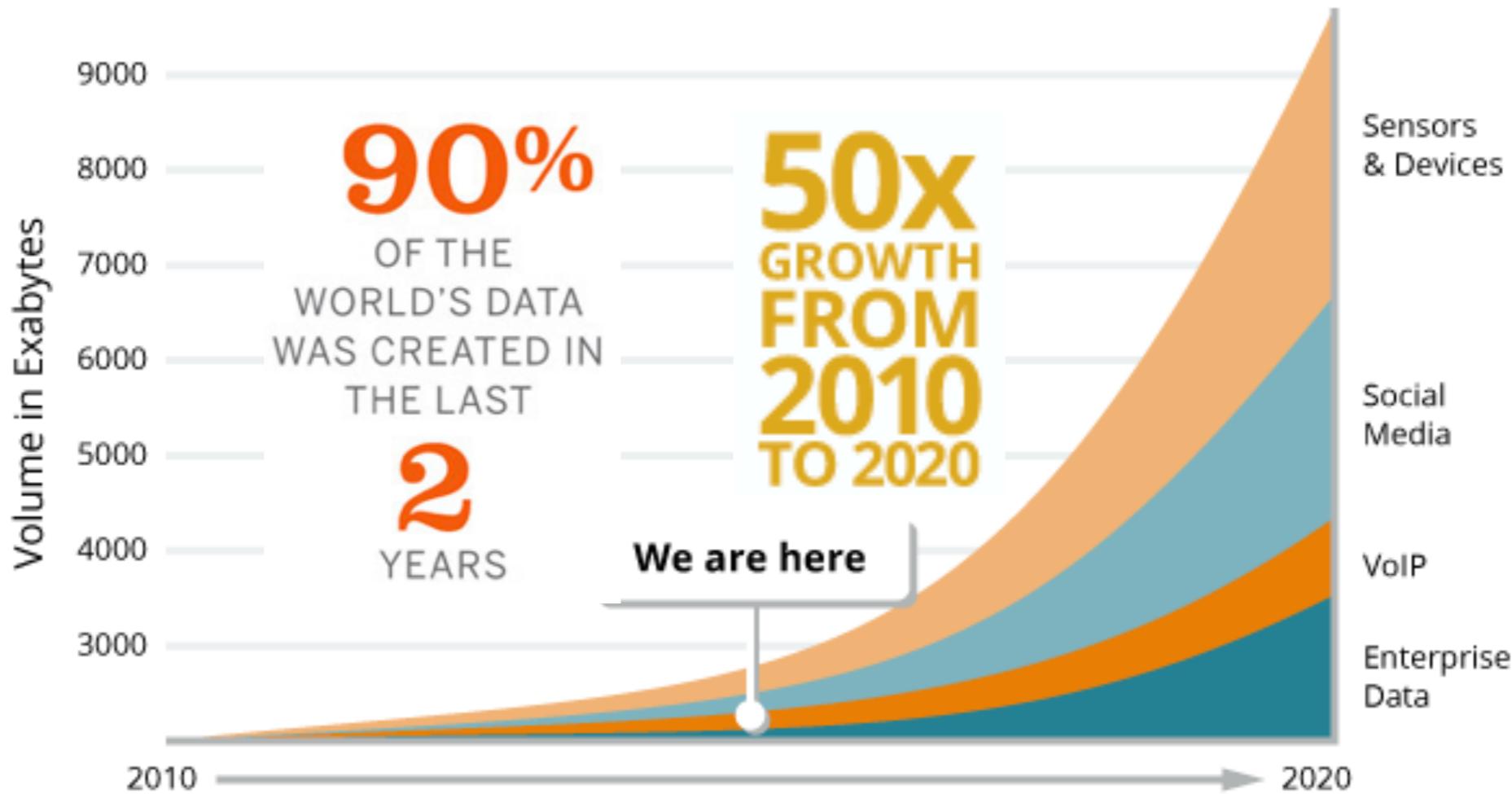
## CONNECTING THE DOTS

WITH THE BIG DATA SAAS VALUE CHAIN!

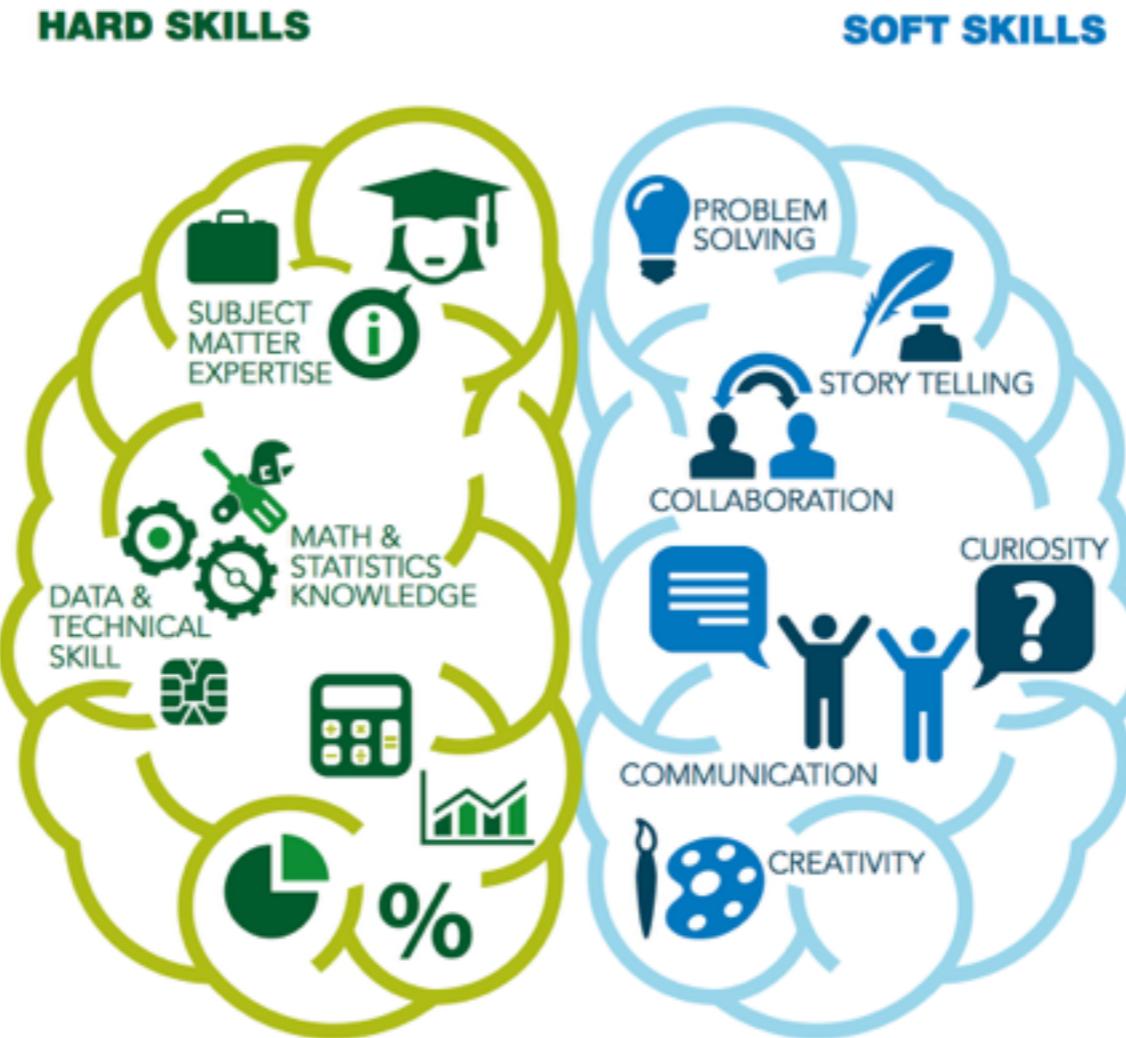


# FINAL THOUGHTS

# DATA OBESITY!



# A NEW TYPE OF KNOWLEDGE WORKER



## AN INCREDIBLY VALUABLE SKILL

**"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding."**

— HAL VARIAN,  
chief economist at Google

Source: "For Today's Graduate, Just One Word: Statistics," New York Times, August 5, 2009

### TREND AMONG MAJOR COMPANIES

**90%**

Source: "Billions and billions: big data becomes a big deal." Deloitte, Technology, Media & Telecommunications Predictions 2012.



Source: "Counting on Analytical Talent," Accenture, March 2010

of **Fortune 500** companies are predicted to have at least some big data initiatives under way.

### GROWING JOB MARKET



**\$97,000-\$108,000**

The **mean salaries** for positions in the data analytics field.

Source: Burning Glass International report of job postings for bachelor's and graduate degree holders in the data analytics field during 2012



**140,000-190,000**

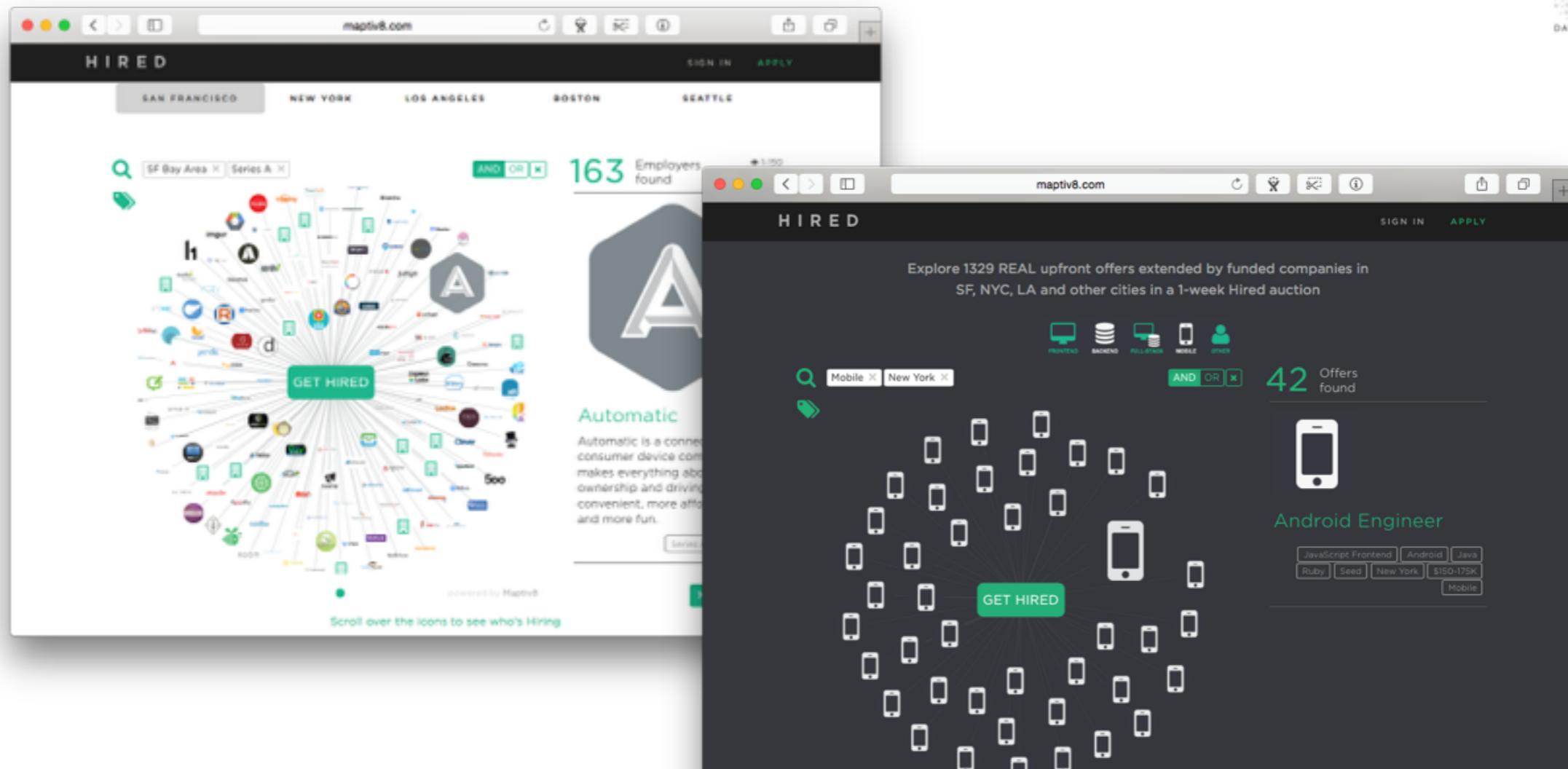
The predicted **shortage of talent** in the next five years with deep analytical skills to take advantage of big data.

Source: "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, May 2011

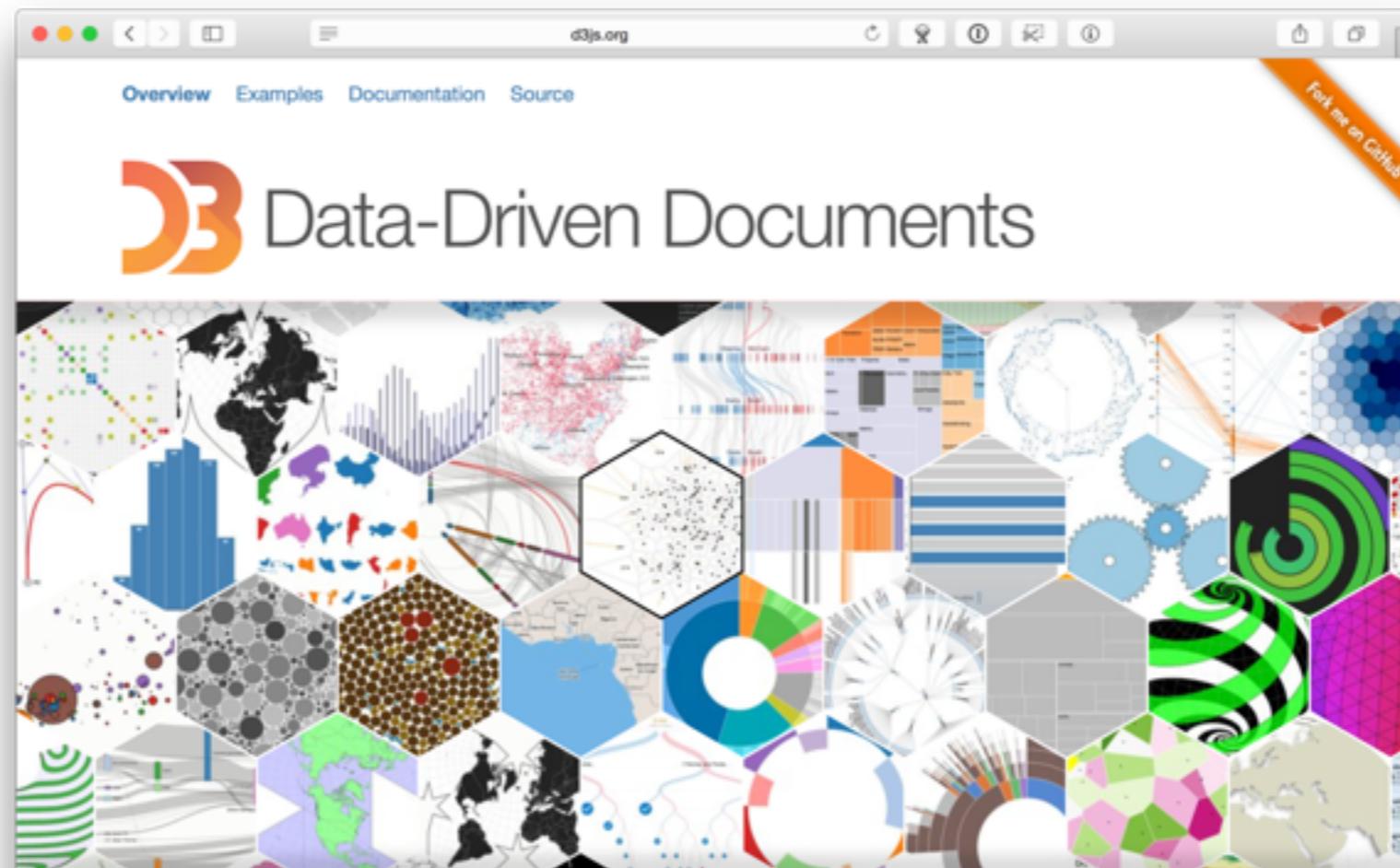
# DATA AS INTERFACE

for HIRED

using Maptiv8



# START HERE

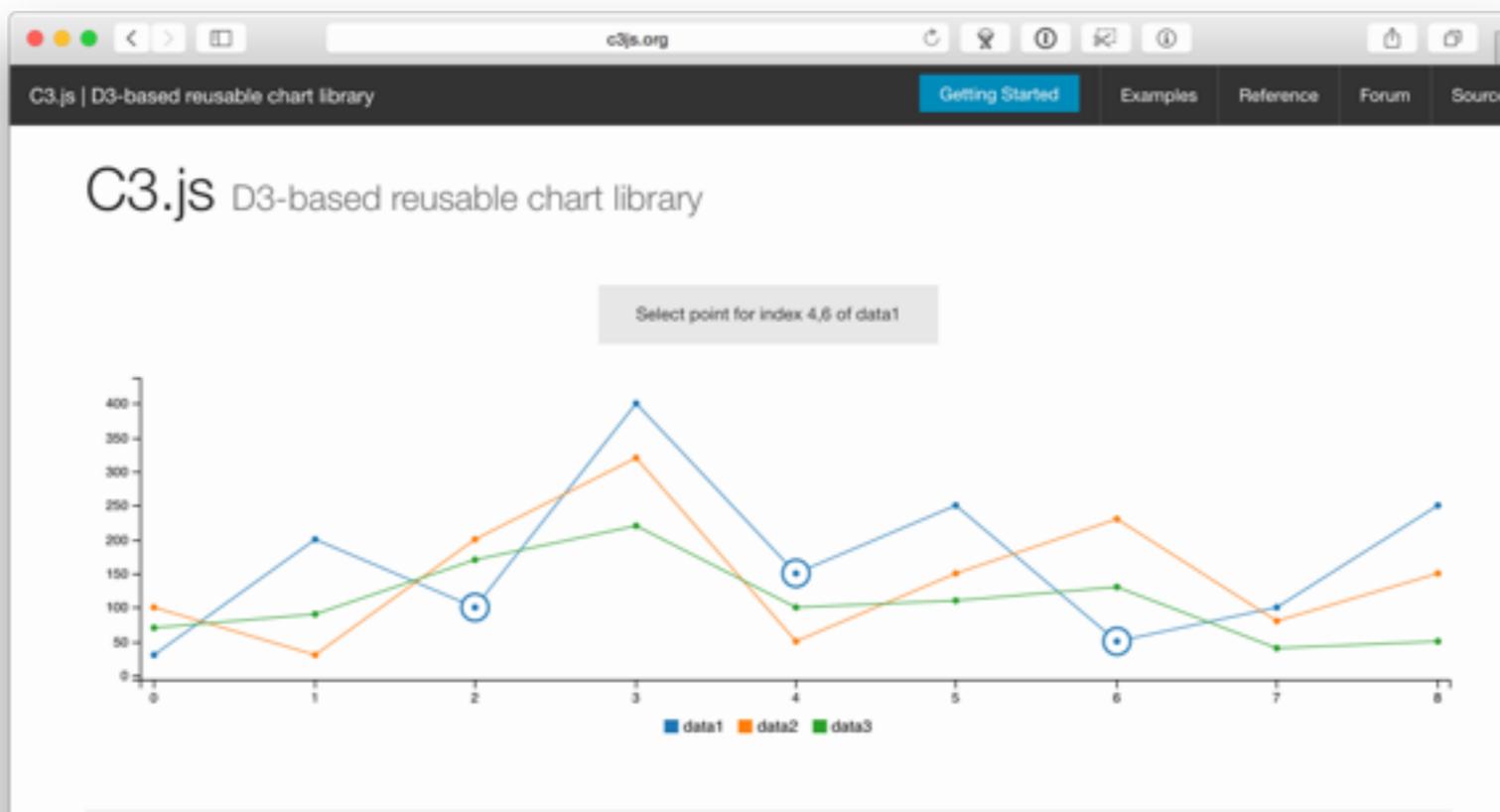


D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

[See more examples.](#)

Download the latest version (3.5.5) here:

# OR HERE



## Why C3?

### Comfortable

C3 makes it easy to generate D3-based charts by wrapping the code required to construct the entire chart. We don't need to write D3 code any more.

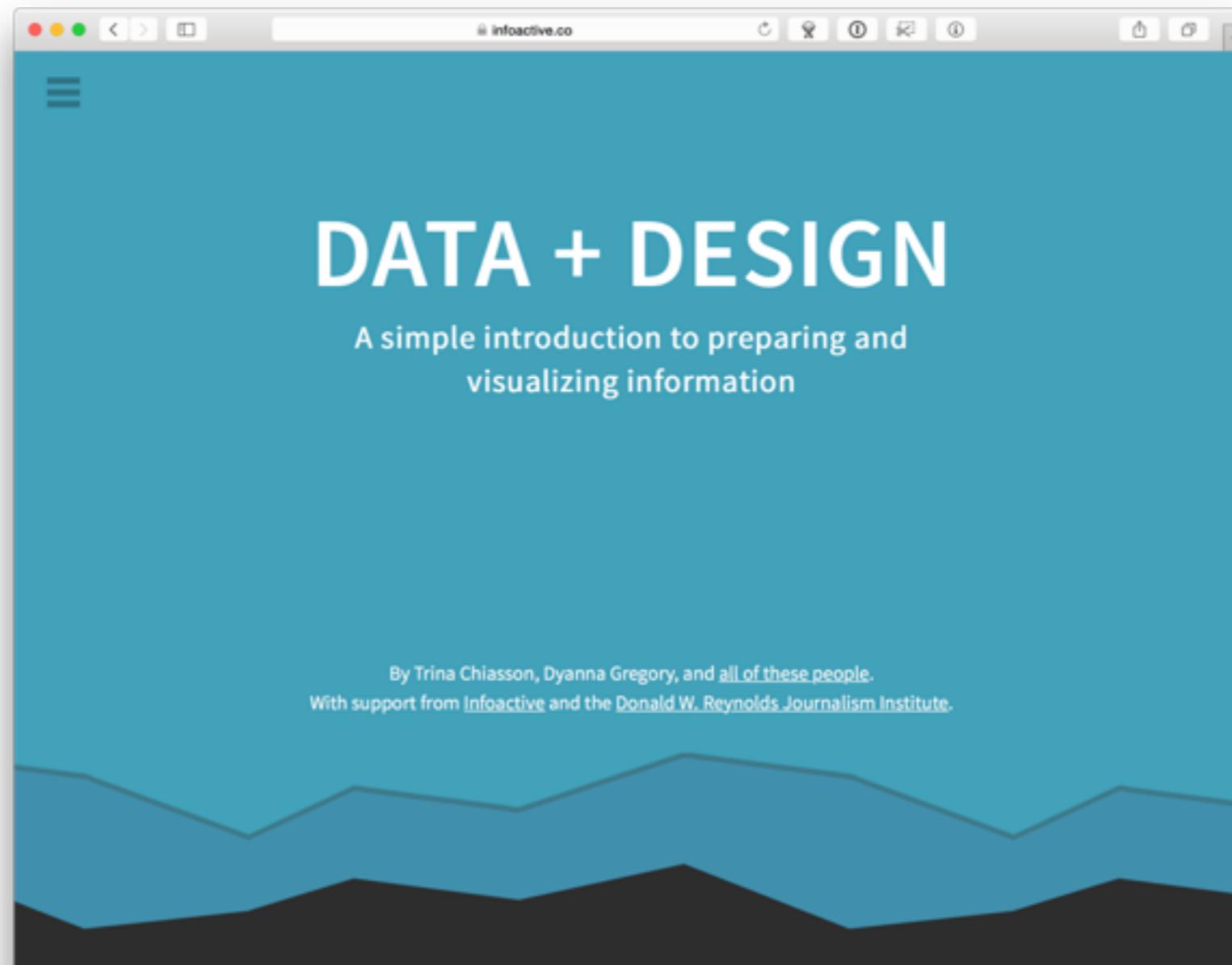
### Customizable

C3 gives some classes to each element when generating, so you can define a custom style by the class and it's possible to extend the structure directly by D3.

### Controllable

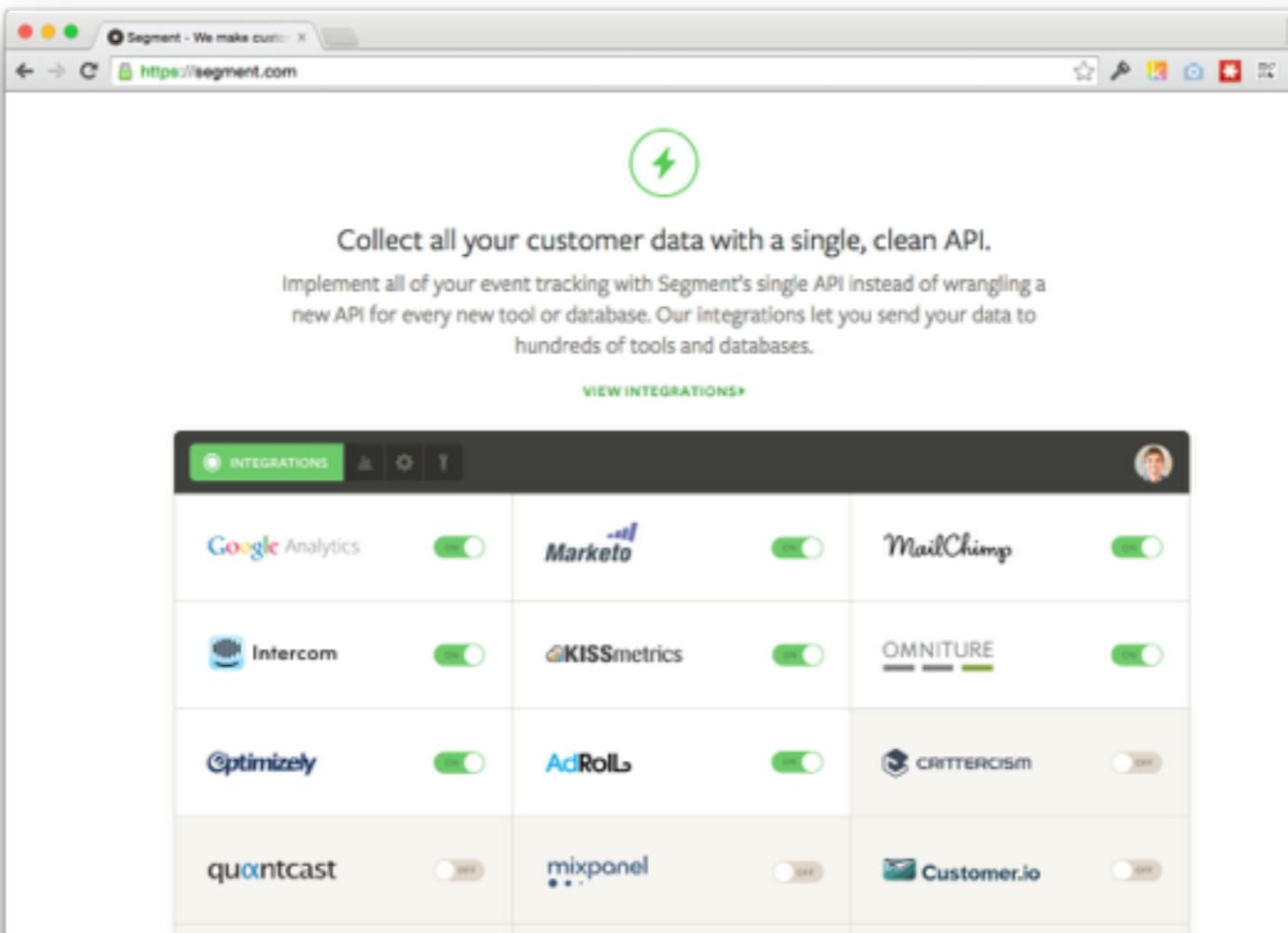
C3 provides a variety of APIs and callbacks to access the state of the chart. By using them, you can update the chart even if after it's rendered.

[INFOACTIVE.CO/DATA-DESIGN](http://infoactive.co/data-design)



# OR HERE

[SEGMENT.COM](https://segment.com)



# OR HERE

[SEGMENT.COM/INTEGRATIONS](https://segment.com/integrations)

The screenshot shows a web browser window with the URL <https://segment.com/integrations>. The page has a navigation bar with tabs: ADVERTISING (highlighted in green), ANALYTICS, MARKETING, SALES, SUPPORT, DEVELOPER, and USER TESTING. The main content area is titled "Advertising" and contains the following text:

"Which ad is driving the most sign ups?"

Advertising tools allow you to add event tracking and ad retargeting to better understand your site visitors, where they came from, and if your campaigns are working.

To the right of this text is a grid of 18 integration logos arranged in four rows of four. The logos are:

- Row 1: AdLearn Open Platform, adometry, AdRoll
- Row 2: Google AdWords, appnexus, AppsFlyer
- Row 3: Attribution, awe.sm, bing ads
- Row 4: blueshift, comSCORE, CONVERTRO
- Row 5: DataXU, facebook Conversions, facebook Audiences
- Row 6: FLURRY, Google Tag Manager, improvely
- Row 7: InsideVault, INTERSTATE, Kenshoo

# OR HERE

FAMILIAN1.COM

The screenshot shows a web browser window displaying the [familian1.com](http://familian1.com) website. The page title is "Tools + Technologies". The header includes the familiant logo and the tagline "insights + interfaces for digital business". Below the header, there's a search/filter icon and a tag icon. A large circular network graph in the center represents various tools, with labels like "mixpanel", "D3", "quicksand", "in", "G", "P", "f", "a:", "K", "plus", "minus", "asterisk", "checkmark", "person", "cloud", "database", "chart", "map", and "S" scattered around. To the right of the graph, a large "D3" logo is shown with the text "43 Tools in total" above it and "Open source data visualization platform" below it. At the bottom, there are "Analytics + Viz" and "Free" buttons, along with navigation links for "Map", "List", and a gear icon. The footer is powered by "Haptiv®".



---

**BUT MOST IMPORTANTLY...**

---

**HAVE FUN!**

**THANK YOU!**

---

**56**

**KEEP IN TOUCH.**

---

**SETH@FAMILIAN1.COM • @SFAM • @MAPTIV8**

---

# Advanced Metrics and Communicating Results

*Ivan Corneillet*

*Data Scientist*

# Learning Objectives

After this lesson, you should be able to:

- › Evaluate a model using advanced metrics such as confusion matrix and ROC/AUC curves
- › Explain the trade-offs between false positives vs. false negatives
- › Describe the difference between visualization for presentations vs. exploratory data analysis
- › Identify the components of a concise and convincing report and how they relate to specific audiences/stakeholders



# Announcements and Exit Tickets



# Q & A



# Review

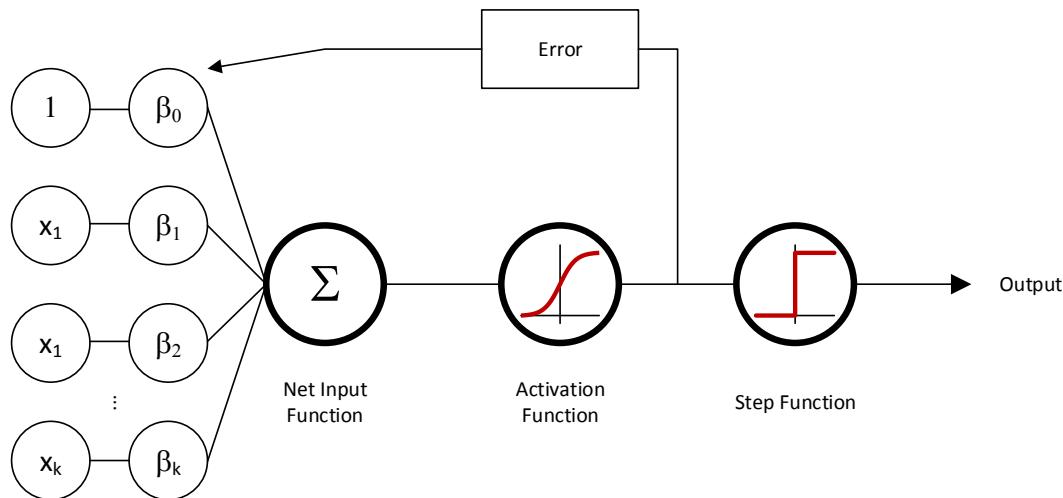


# Review

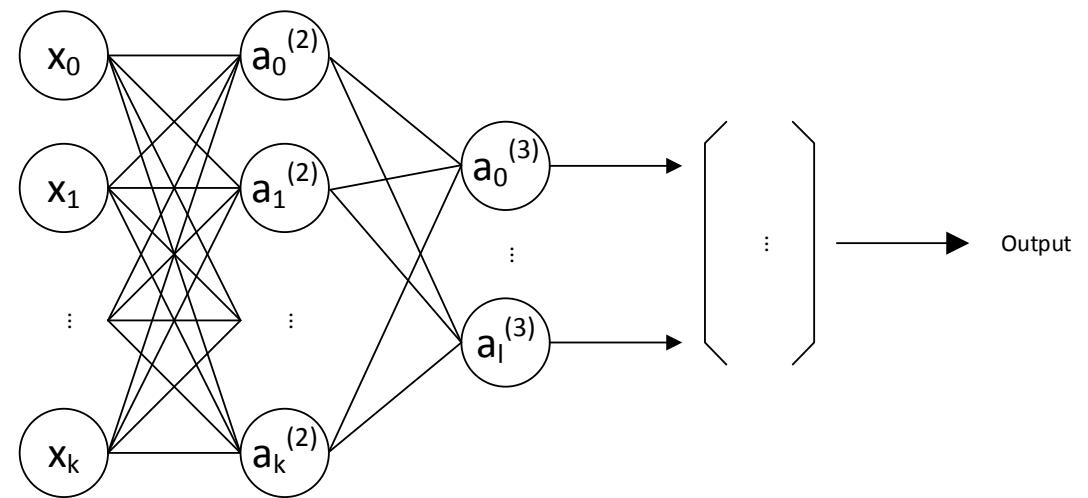
*Logistic Regression*

Going further | Neural networks are related to logistic regression; you can think of logistic regression as a one-layer neural network

**One-layer neural network**



**Multi-layer neural network**





# Today

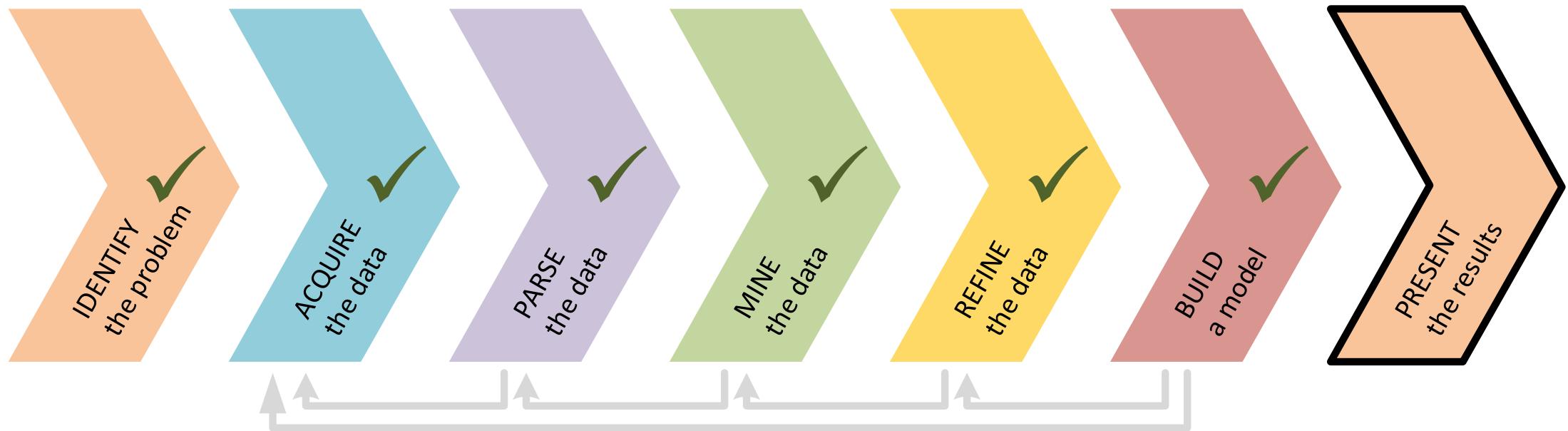
# Outline

- Unit Project 4 (due today)
- Announcements and Exit Tickets
- Review
- Advanced metrics
  - Confusion Matrix
  - True Positive, False Positive Rates, ROC, and AUC
  - Codealong for ROC/AUC
- Communicating Results
  - Showing our Work
  - Codealong to pretty up graphs
- Review

# Today, we are wrapping Unit 2 – Foundation of Modeling

Research Design and Data Analysis	Research Design	Data Visualization in <i>pandas</i>	Statistics	Exploratory Data Analysis in <i>pandas</i>
Foundations of Modeling	Linear Regression (sessions 6, 7, and 11)	Classification Models (KNN, Logistic Regression) (sessions 8, 9, and 11)	Evaluating Model Fit (sessions 5, 6, and 7)	Presenting Insights from Data Models (session 10)
Data Science in the Real World	Decision Trees and Random Forests	Time Series Data	Natural Language Processing	Databases

... as well as the first full pass of the Data Science Workflow





# Pre-Work

# Pre-Work

Before this lesson, you should already be able to:

- Create and interpret results from a binary classification problem
- Know what a decision line is in logistic regression



# ⑥ Build a Model

*Advanced Metrics*

# Advanced Metrics

- Accuracy is only one of several metrics used when solving for a classification problem
  - E.g., if we know a prediction is 75% accurate, accuracy doesn't provide any insight into why the 25% was wrong. Was it wrong *equally* across all class labels? Did it just guess one class label for all predictions and 25% of the data was just the other label?
- It's important to look at other metrics to fully understand the problem

## ‣ Accuracy

- How many observations that we predicted were correct? This is a value we'd want to increase (like  $R^2$ )

## ‣ Misclassification rate

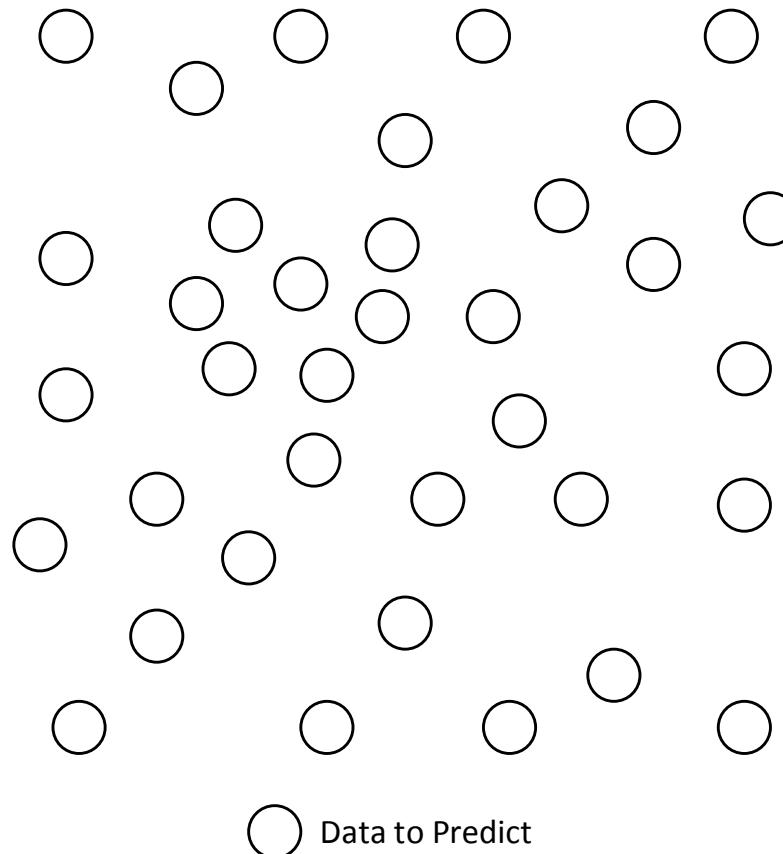
- Directly opposite of accuracy
- Of all the observations we predicted, how many were incorrect? This is a value we'd want to decrease (like the mean squared error)



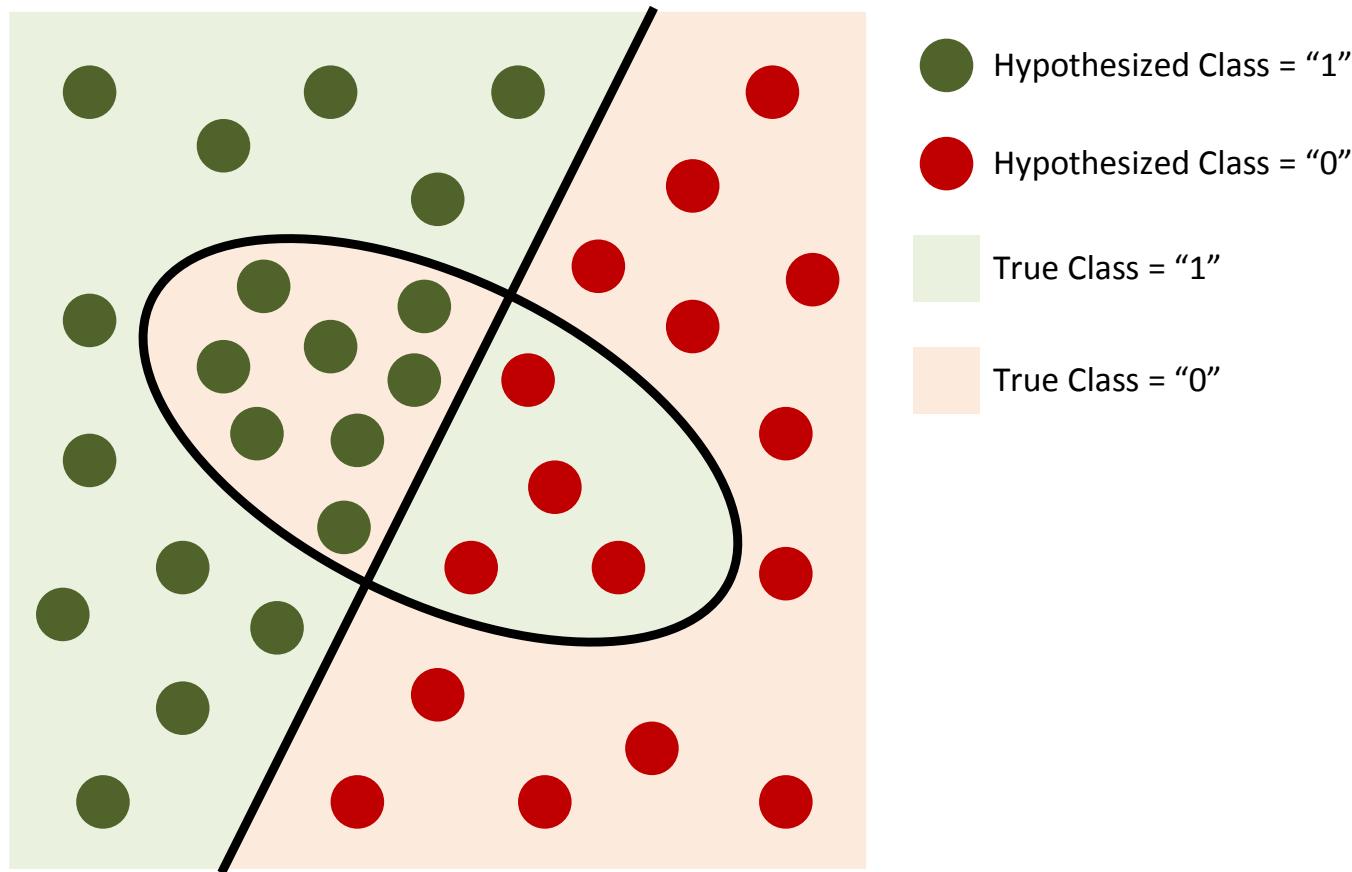
# ⑥ Build a Model

*Confusion Matrix*

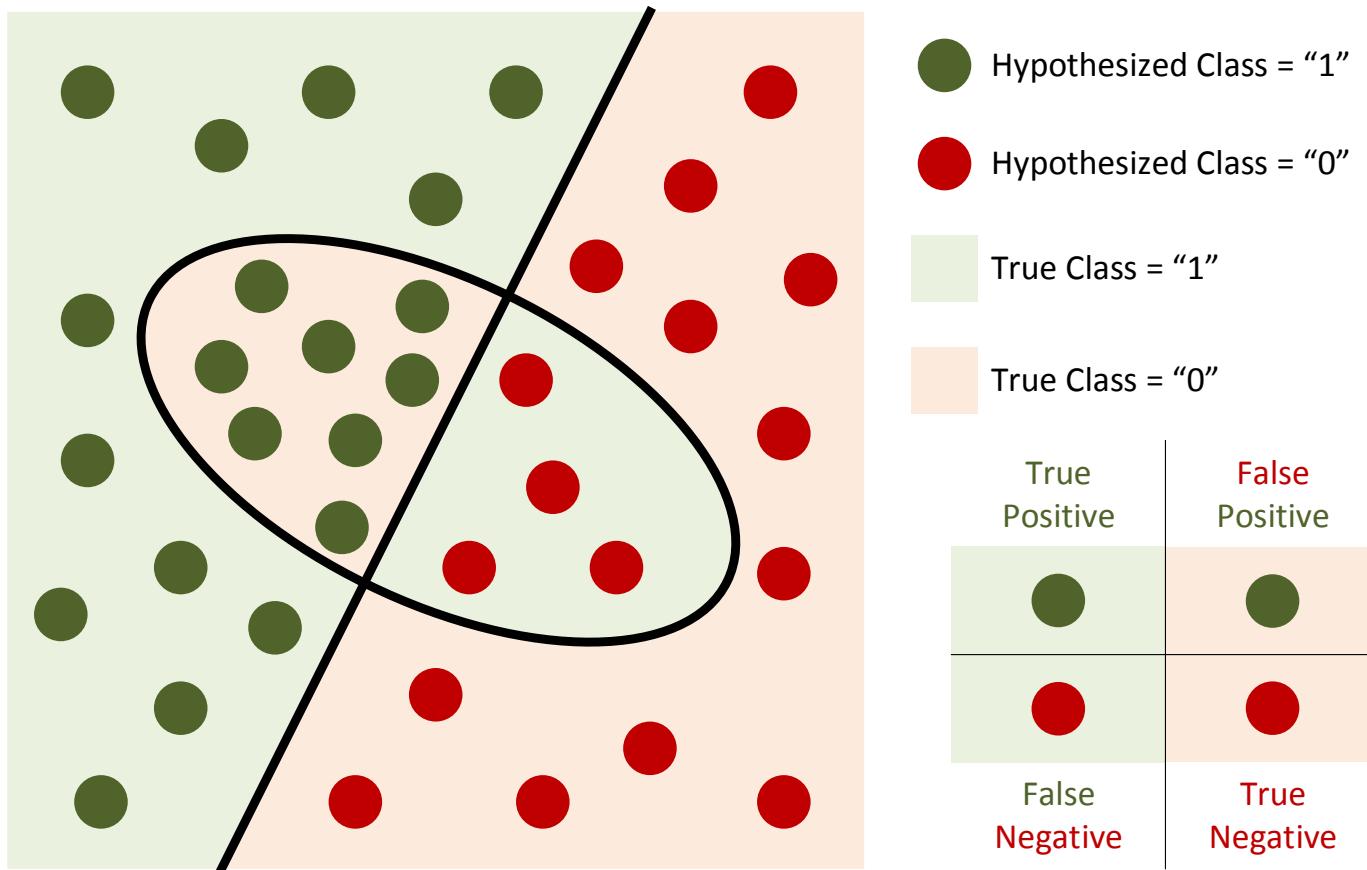
Stepping back | Let's say we want to classify this data:



# Putting hypothesized and true classes together, we get 4 possibilities



We can rearrange these 4 possibilities into a 2x2 table



# Confusion Matrix (a.k.a., Contingency Table or Error Matrix)

		True Class	
		1	0
		True Positives ( $TP$ )	False Positives ( $FP$ ) <i>(type I error)</i>
Hypothesized Class	1	●	●
	0	■	■
Total Columns		$P = TP + FN$	$N = FP + TN$

- A confusion matrix is a specific table layout that allows visualization of the performance of a supervised learning algorithm
- Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class
- The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e., commonly mislabeling one as another)



## ⑥ Build a Model

*Activity | Interpreting the confusion matrix*

# Activity | Interpreting the confusion matrix

EXERCISE

## DIRECTIONS (20 minutes)

1. Use the variables defined in the confusion matrix ( $TP$ ,  $FN$ ,  $FP$ ,  $TN$ ,  $P$ , and  $N$ ) to calculate the answers to the following questions:
  - a. Overall, how often is the classifier correct?
  - b. When the classifier predicts yes, how often is it correct?
  - c. How often does the yes condition actually occur in our sample?
  - d. When it's actually yes, how often does the classifier predict yes?
  - e. When it's actually no, how often does the classifier predict yes?
  - f. When it's actually no, how often does it predict no?
  - g. Overall, how often is the classifier wrong?

# Activity | Interpreting the confusion matrix (cont.)

EXERCISE

## DIRECTIONS (cont.)

2. Given a medical exam that tests for cancer ( $1 = \text{Cancer}$ ,  $0 = \text{Cancer free}$ ), use the variables defined in the confusion matrix ( $TP$ ,  $FN$ ,  $FP$ ,  $TN$ ,  $P$ , and  $N$ ) to calculate the answers to the following questions:
  - a. How often is it correct when it identify patients with cancer?
  - b. How often does it correctly identify patients without cancer?
  - c. How often does it trigger a “false alarm” by saying a patient has cancer when they actually don’t?
  - d. How often does it correctly identify patients with cancer?
3. When finished, share your answers with your table

## DELIVERABLE

Answers to the above questions

# Activity | Interpreting the confusion matrix (cont.)

		True Class		Question: Overall, how often is the classifier correct?  Answer: $\frac{TP+TN}{P+N}$	<i>When the classifier predicts yes, how often is it correct?</i>		
Hypothesized Class	1	● <b>True Positives</b> (TP)	● <b>False Positives</b> (FP) <i>(type I error)</i>				
	0	■ <b>False Negatives</b> (FN) <i>(type II error)</i>	■ <b>True Negatives</b> (TN)				
	Total Columns	$P = TP + FN$	$N = FP + TN$				
<i>How often does the yes condition actually occur in our sample?</i>		<i>When it's actually yes, how often does the classifier predict yes?</i>					
<i>When it's actually no, how often does the classifier predict yes?</i>		<i>When it's actually no, how often does it predict no?</i>					
<i>Overall, how often is the classifier wrong?</i>							

# Activity | Interpreting the confusion matrix (cont.)

		True Class			
		Has Cancer	Doesn't have cancer		
Hypothesized Class	Predict Cancer ●	● <b>True Positives (TP)</b>	● <b>False Positives (FP)</b> (type I error)	<i>How often is it correct when it identify patients with cancer?</i>	<i>How often does it correctly identify patients without cancer?</i>
	Predict No Cancer ■	■ <b>False Negatives (FN)</b> (type II error)	■ <b>True Negatives (TN)</b>	<i>How often does it trigger a “false alarm” by saying a patient has cancer when they actually don’t?</i>	<i>How often does it correctly identify patients with cancer?</i>
	Total Columns	$P = TP + FN$		<i>How many patients have cancer?</i>	

# Activity | Interpreting the confusion matrix (cont.)

$\frac{TP+TN}{P+N}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FP}$
$\frac{FN+TP}{P+N}$	$\frac{P}{P+N}$	$\frac{TP+FN}{TP+FP}$	$\frac{TP}{TP+FN}$
$\frac{TP+FN}{P+N}$	$\frac{TP}{P+N}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$
$\frac{FP}{TN+FP}$	$\frac{TP}{FP+TN}$	$\frac{TN}{TN+FP}$	$\frac{TN}{TN+FP}$
$\frac{FP}{N}$	$\frac{TP}{FP+TN}$	$\frac{TN}{FP+TN}$	$\frac{TN}{N}$
$\frac{FP+FN}{P+N}$	$\frac{FP+FN}{P+N}$	$\frac{FP+FN}{P+N}$	$\frac{FP+FN}{P+N}$

$\frac{TP}{P+N}$	$\frac{TP}{FP+TP}$	$\frac{TP}{P+N}$	$\frac{TN/P+N}{P+N}$	$\frac{TN}{N}$
$\frac{FP}{N}$	$\frac{FP}{FP+TP}$	$\frac{FP}{N}$	$\frac{FP}{P+N}$	$\frac{FP}{P+N}$
$\frac{FP}{N}$	$\frac{FP}{FP+TP}$	$\frac{FP}{N}$	$\frac{FP}{P+N}$	$\frac{FP}{N}$
$\frac{TP+FN}{P+N}$	$\frac{TP+FN}{P+N}$	$\frac{TP+FN}{P+N}$	$\frac{TP+FN}{P+N}$	$\frac{TP}{P}$
$\frac{TP+FN}{P+N}$	$\frac{TP+FN}{P+N}$	$\frac{TP+FN}{P+N}$	$\frac{TP+FN}{P+N}$	$\frac{TP}{P}$

# Activity | Interpreting the confusion matrix (cont.)

		True Class		Question: Overall, how often is the classifier correct?	When the classifier predicts yes, how often is it correct?
		1	0	Answer: $\frac{TP+TN}{P+N}$ (accuracy)	Answer: $\frac{TP}{TP+FP}$ (precision)
Hypothesized Class	1	True Positives (TP) 	False Positives (FP)  (type I error)	How often does the yes condition actually occur in our sample?	When it's actually yes, how often does the classifier predict yes?
	0	False Negatives (FN)  (type II error)	True Negatives (TN) 	Answer: $\frac{P}{P+N}$ (prevalence)	Answer: $\frac{TP}{P}$ (TPR, sensitivity, recall)
Total Columns		$P = TP + FN$	$N = FP + TN$	When it's actually no, how often does the classifier predict yes?	When it's actually no, how often does it predict no?
				Answer: $\frac{FP}{N}$ (FPR, fall-out)	Answer: $\frac{TN}{N}$ (specificity)
				Overall, how often is the classifier wrong?	
				Answer: $\frac{FP+FN}{P+N}$ (misclassification rate)	

# Activity | Interpreting the confusion matrix (cont.)

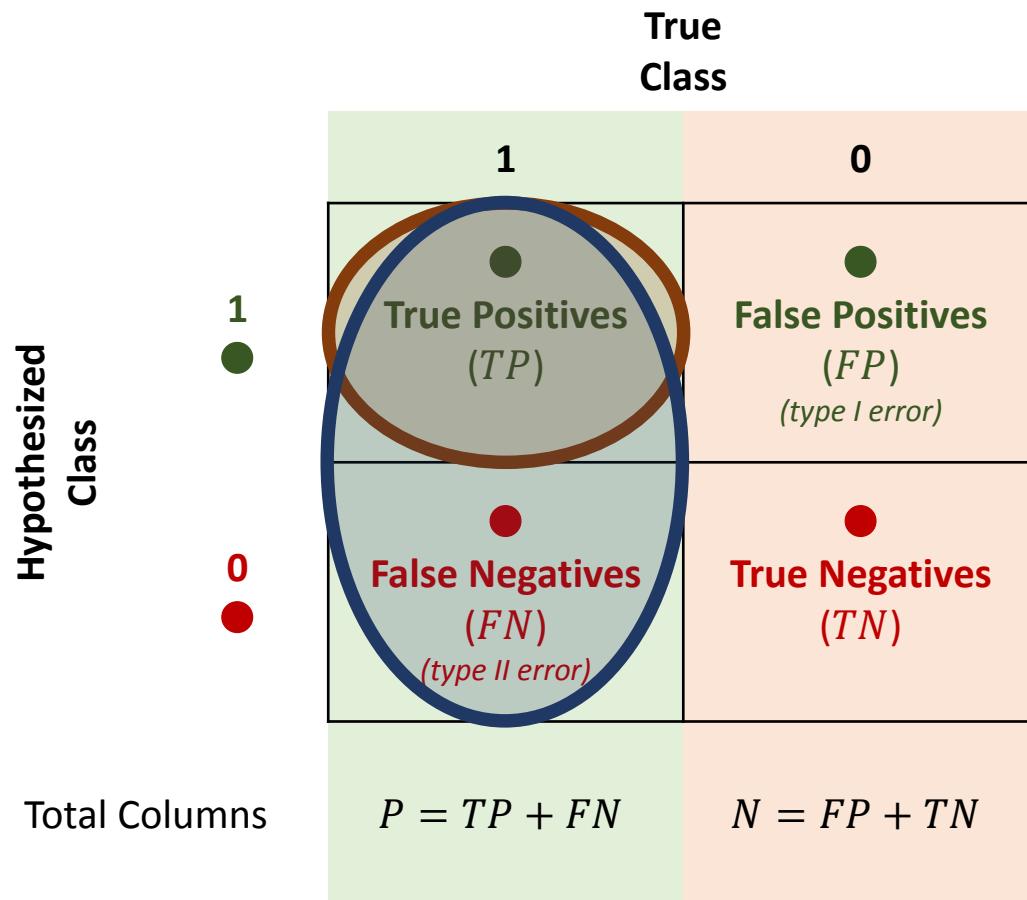
		True Class			
		Has Cancer	Doesn't have cancer		
Hypothesized Class	Predict Cancer ●	● <b>True Positives (TP)</b>	● <b>False Positives (FP)</b> (type I error)	How often is it correct when it identify patients with cancer?	How often does it correctly identify patients without cancer?
	Predict No Cancer ■	■ <b>False Negatives (FN)</b> (type II error)	■ <b>True Negatives (TN)</b>	How often does it trigger a "false alarm" by saying a patient has cancer when they actually don't?	How often does it correctly identify patients with cancer?
	Total Columns	$P = TP + FN$	$N = FP + TN$	Answer: $\frac{TP}{TP+FP}$ (precision)	Answer: $\frac{TN}{N}$ (specificity)
				How many patients have cancer?	
				Answer: $\frac{P}{P+N}$ (prevalence)	



## ⑥ Build a Model

*True and False Positive Rates, ROC, and AUC*

$$\text{True Positive Rate, } TPR = \frac{TP}{P}$$



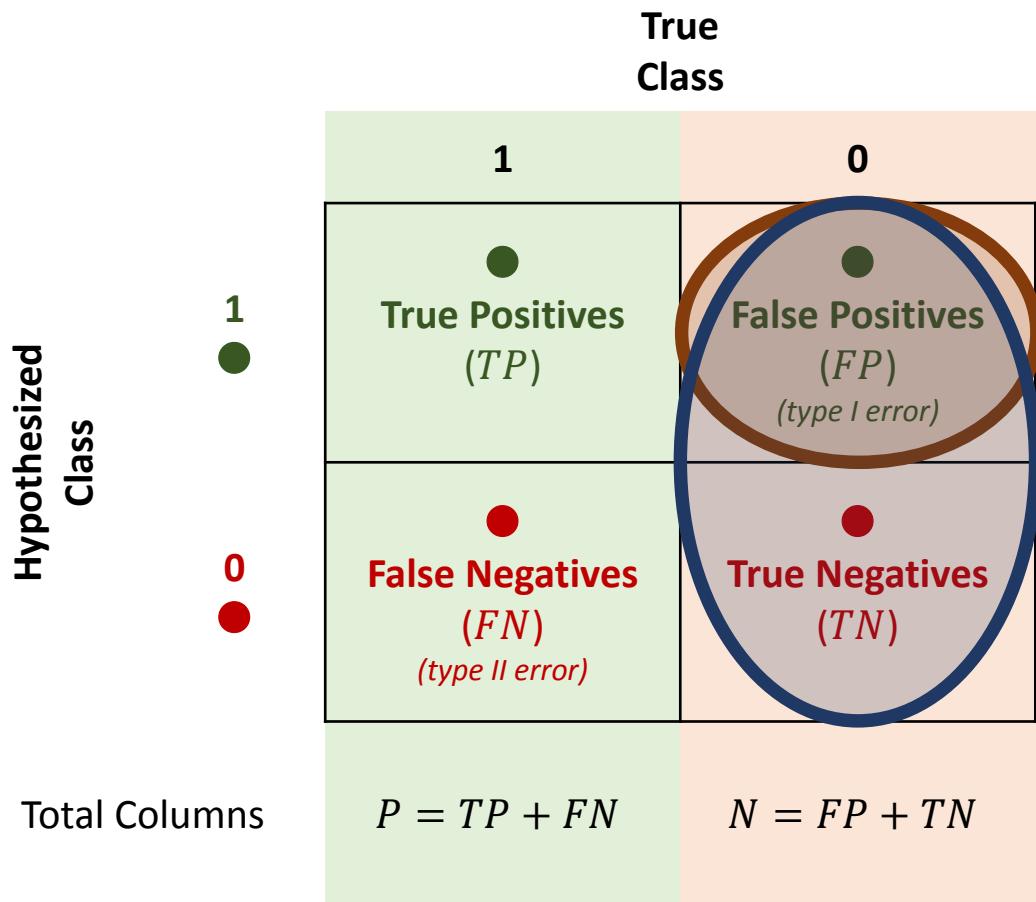
- When it's actually yes, how often does the classifier predict yes?

- A.k.a., “Sensitivity”

- E.g., given a medical exam that tests for cancer, how often does it correctly identify patients with cancer?

- Likewise, this can be inverted: how often does a test *correctly* identify patients without cancer

$$\text{False Positive Rate, } FPR = \frac{FP}{N}$$



- When it's actually no, how often does the classifier predict yes?
- A.k.a., “Fall-out”

- E.g., given a medical exam that tests for cancer, how often does it trigger a “false alarm” by saying a patient has cancer when they actually don’t?
- Likewise, this can be also inverted: how often does a test *incorrectly* identify patients as being cancer-free when they might actually have cancer!

# True positive and false positive rates

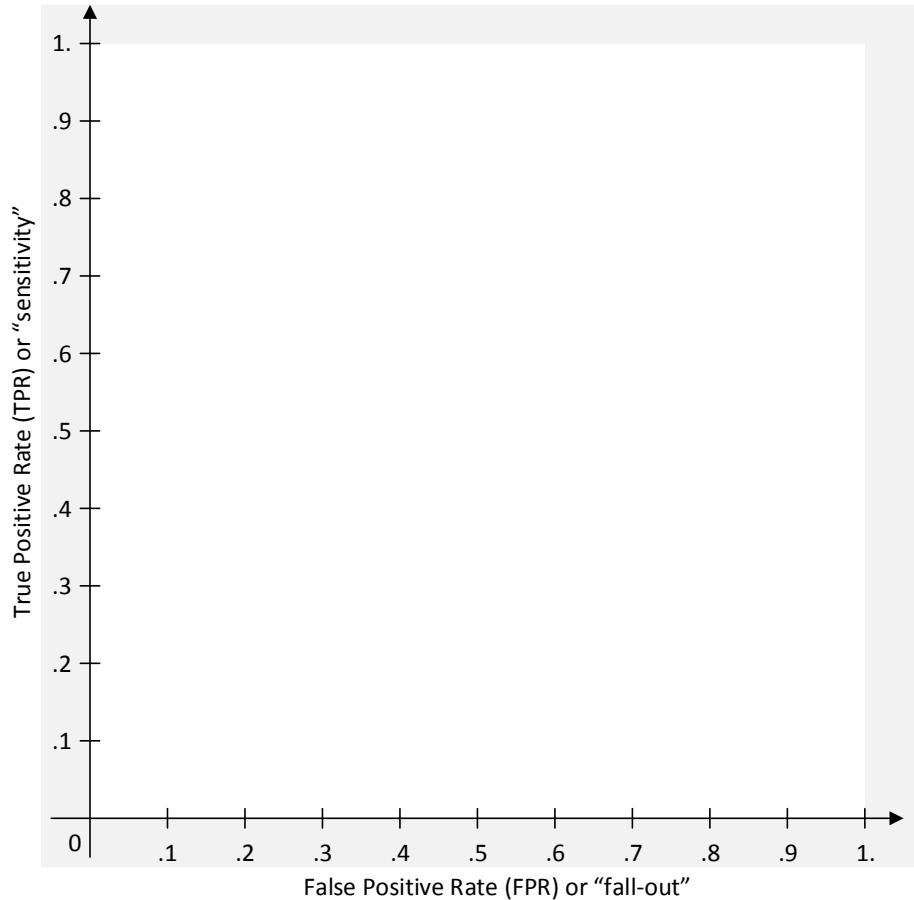
- We can split up the accuracy of each label by using true positive and false positive rates. Using them, we can get a much clearer picture of where predictions begin to fall apart
- A good classifier would have a true positive rate approaching 1, and a false positive rate approaching 0. In a binary problem (say, predicting if someone smokes or not), it would accurately predict all of the smokers as smokers, and not accidentally predict any of the non-smokers as smokers

# ⑥ Build a Model

*ROC (receiver operating characteristic or relative operating characteristic) curve*

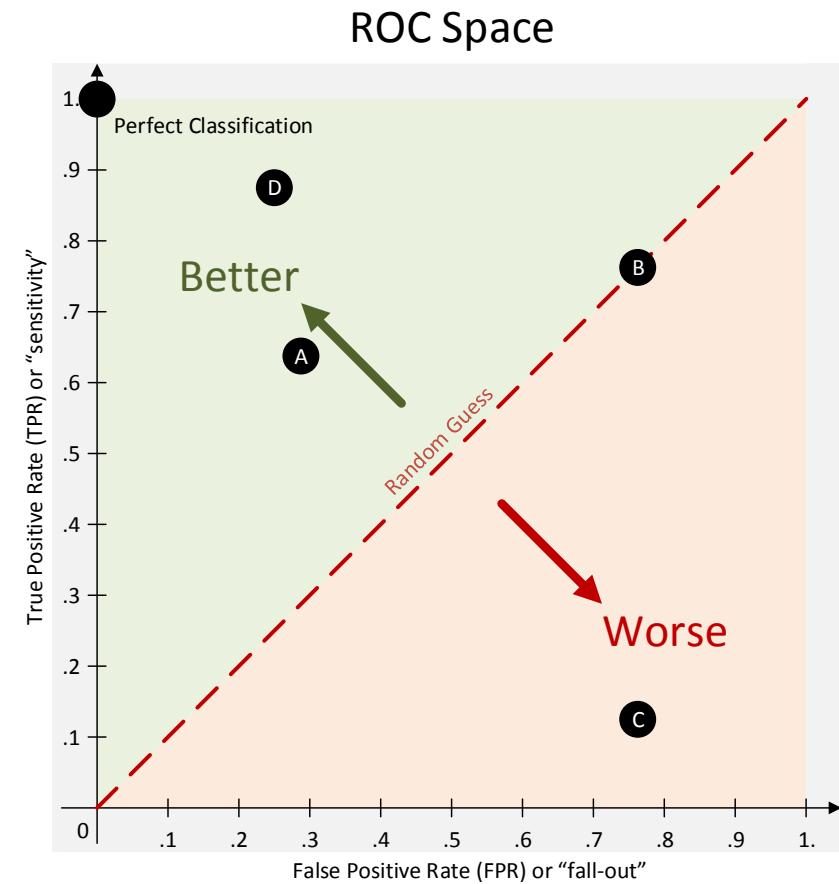
# ROC (receiver operating characteristic) curve (a.k.a., relative operating characteristic curve)

- An ROC curve plots the true positive rate (TPR) (or “sensitivity”) against the false positive rate (FPR) (or “fall-out”) at various threshold settings to illustrate the performance of a binary classifier system. The ROC curve is thus the sensitivity as a function of fall-out



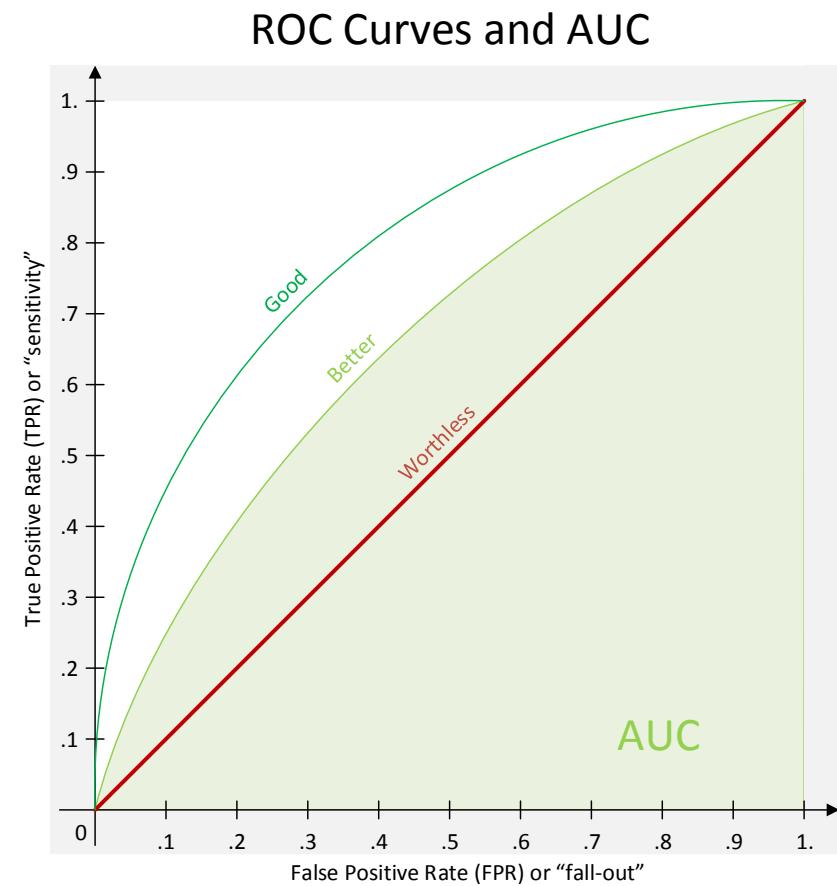
# The ROC space demonstrates several things:

- It shows the tradeoff between sensitivity and fall-out (any increase in sensitivity will be accompanied by an increase in fallout)
  - The closer the **points** are in the left-hand border and then the top border of the ROC space, the more accurate the classifier is
  - The closer the **points** come to the 45-degree diagonal of the ROC space, the less accurate the classifier is

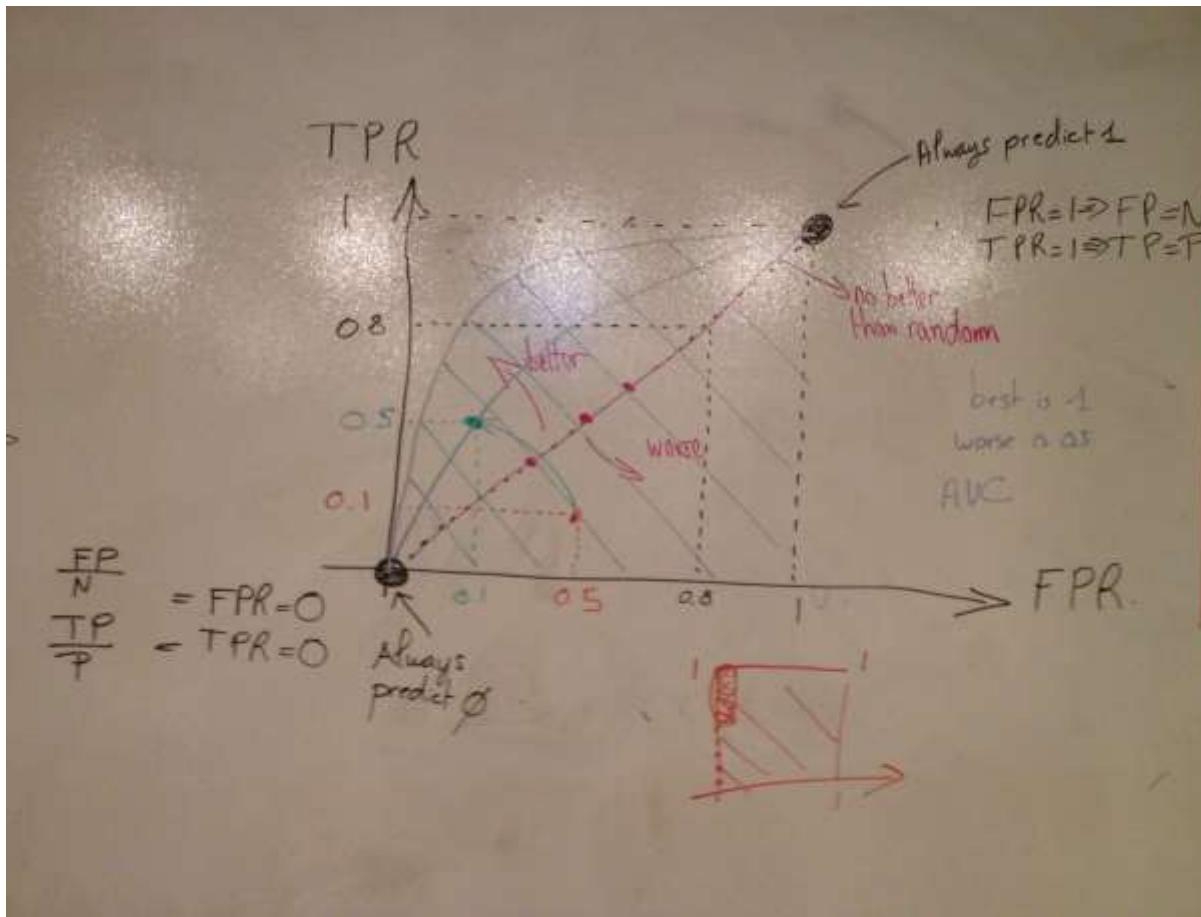


# The ROC curves demonstrate several things:

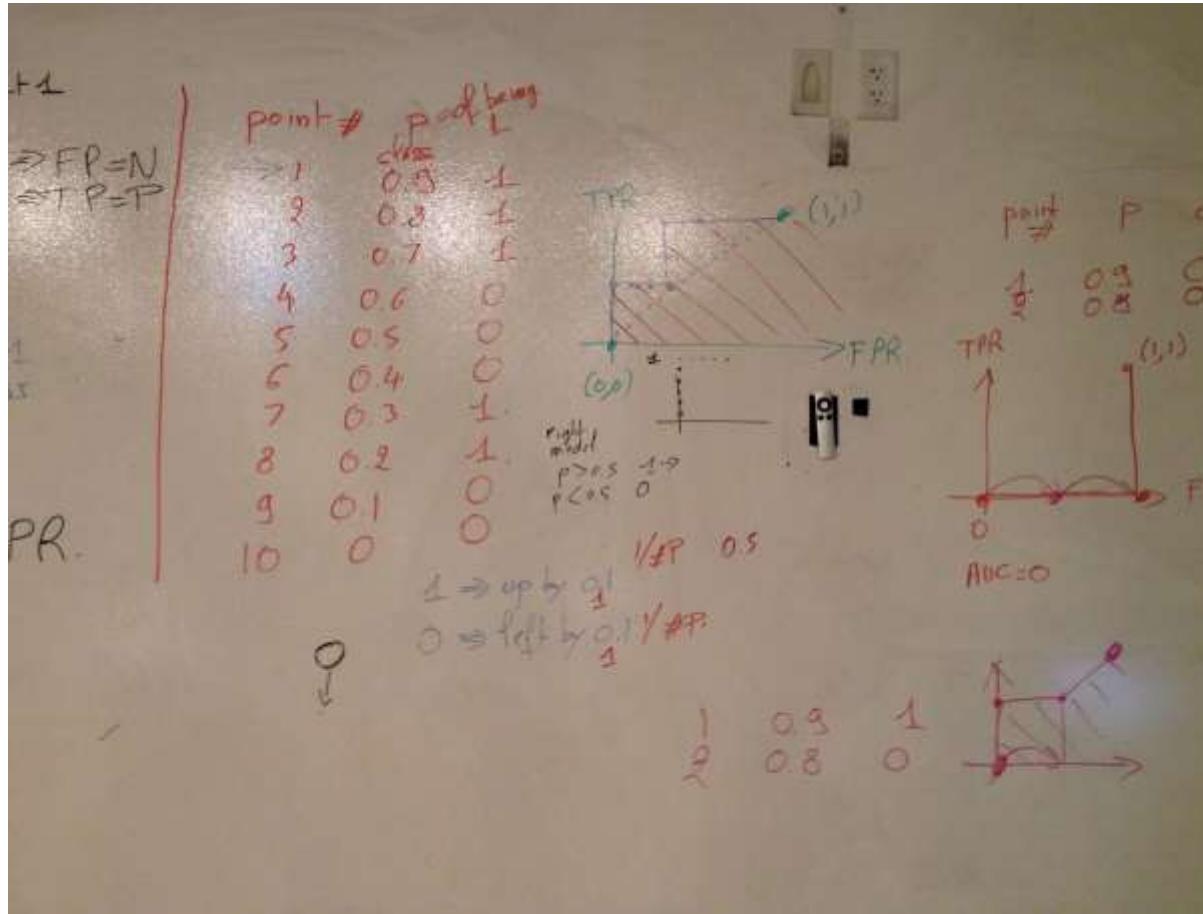
- The area under the curve (AUC) is a measure of classifier accuracy
  - The closer the **curve** follows the left-hand border and then the top border of the ROC space, the more accurate the classifier is
  - The closer the **curve** comes to the 45-degree diagonal of the ROC space, the less accurate the classifier is



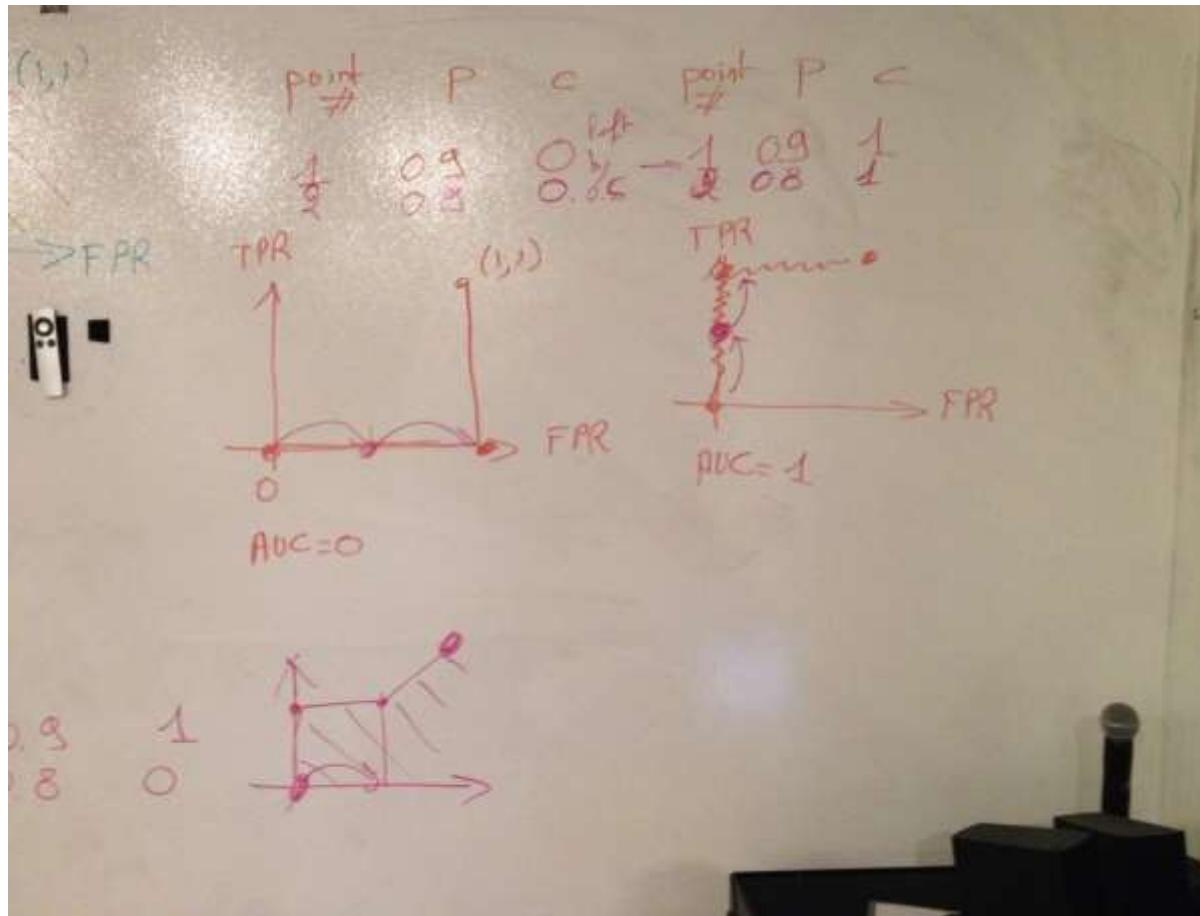
# ROC and AUC



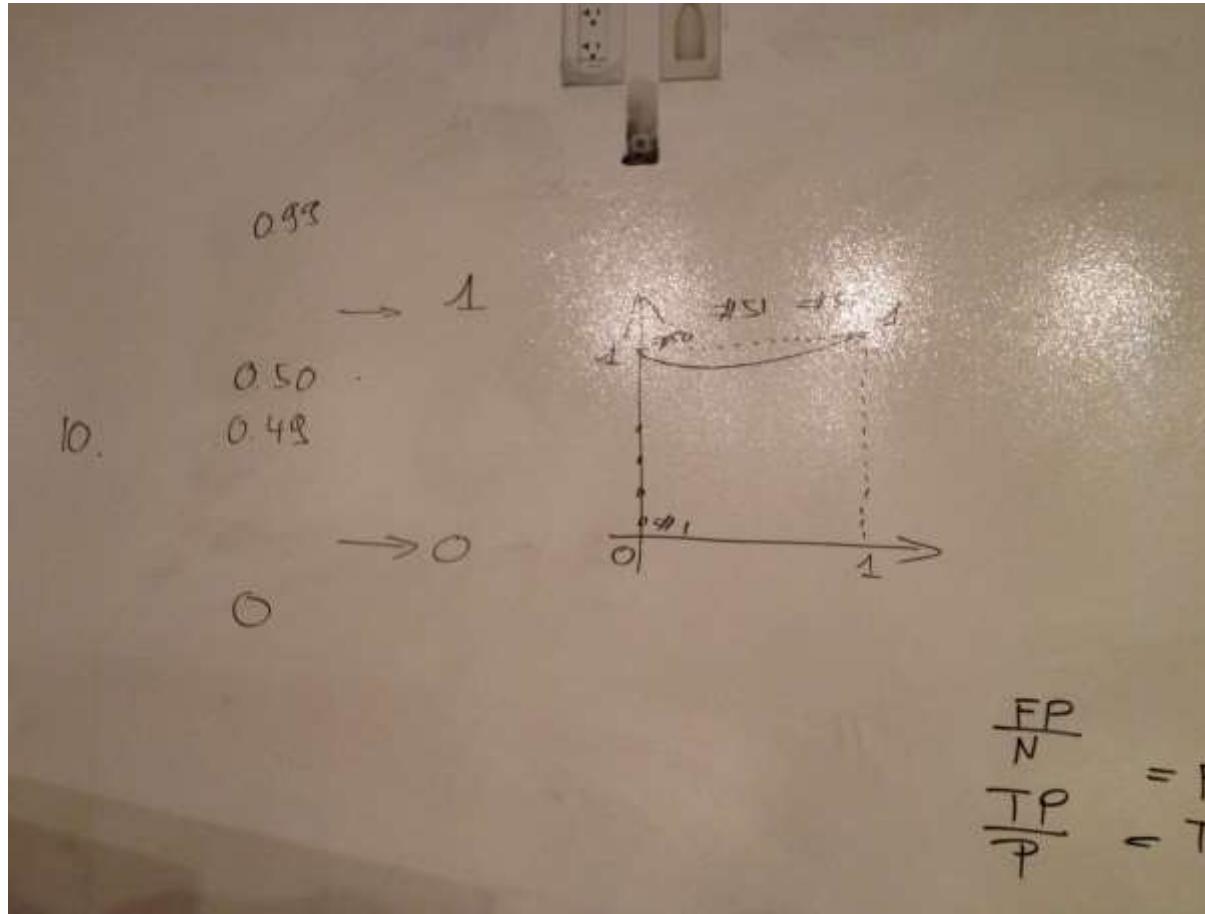
# ROC and AUC (cont.)



# ROC and AUC (cont.)



# ROC and AUC (cont.)





# ⑥ Build a Model

*Codealong – Part A*

*ROC/AUC*



7

# Present the Results

*Communicating Results*

# We built a model! Now what?

- We've built our model, but there is still a gap between our iPython notebook with its plots and figures and a slideshow needed to present our results
- Classes so far have focused on two core concepts:
  - Developing consistent practices
  - Interpreting metrics to evaluate and improve model performance
- But what does that mean to your audience?

# We built a model! Now what? (cont.)

- › Imagine how a non-technical audience might respond to the following statements:
  - › “The predictive model I built has an accuracy of 80%”
  - › “Logistic regression was optimized with L2 regularization”
  - › “Gender was more important than age in the predictive model because it has a ‘larger coefficient’”
  - › “Here’s the AUC chart that shows how well the model did”

# We built a model! Now what? (cont.)

- Who is your audience? Are they technical? What are their concerns?
  - In a business setting, you may be the only person who can interpret what you've built
- Some people may be familiar with basic visualization, but you will likely have to do a lot of “hand holding”
- You need to be able to efficiently explain your results in a way that makes sense to all stakeholders (technical or not)

# We built a model! Now what? (cont.)

- Today, we'll focus on communicating results for “simpler” problems, but this applies to any type of model you may work with



# 7 Present the Results

*Showing our Work*

# Showing our Work

- We've spent a lot of time exploring our data and building a reasonable model that performs well
- However, if we look at our visuals, they are most likely:
  - Statistically heavy: most people don't understand histograms
  - Overly complicated: scatter matrices produce too much information
  - Poorly labeled: code doesn't require adding labels, so you may not have added them

To convey important information to your audience, make sure your charts are simplified, easily interpretable, and clearly labeled

### Simplified

- At most, you'll want to include figures that either explain a variable on its own or explain that variable's relationship against a target
- If your model used a data transformation (like natural log), just visualize the original data
- Remove unnecessary complexity

### Easily interpretable

- Any stakeholder looking at a figure should be seeing the exact same thing you're seeing
  - A good test for this is to share the visual with others less familiar with the data and see if they come to the same conclusion
  - How long did it take them?

### Clearly labeled

- Take the time to clearly label your axis, title your plot, and double check your scales – especially if the figures should be comparable
- If you're showing two graphs side by side, they should follow the same Y axis

When building visuals for another audience, ask yourself who, what, and how

### **Who**

- Who is my target audience for the visual?

### **What**

- What do they already know about this project?
- What do they need to know?

### **How**

- How does my project affect this audience? How might they interpret (or misinterpret) the data?

# Visualizing Models over Variables

- One effective way to explain your model over particular variables is to plot the predicted values against the most explanatory variables
- E.g., in logistic regression, plotting the probability of a class against a variable can help explain the range of effect of the model

# Visualizing Performance Against Baseline

- Another approach of visualization is the effect of your model against a baseline, or – even better – against previous models
- Plots like this will also be useful when talking to your peers – other data scientists or analysts who are familiar with your project and interested in the progress you've made

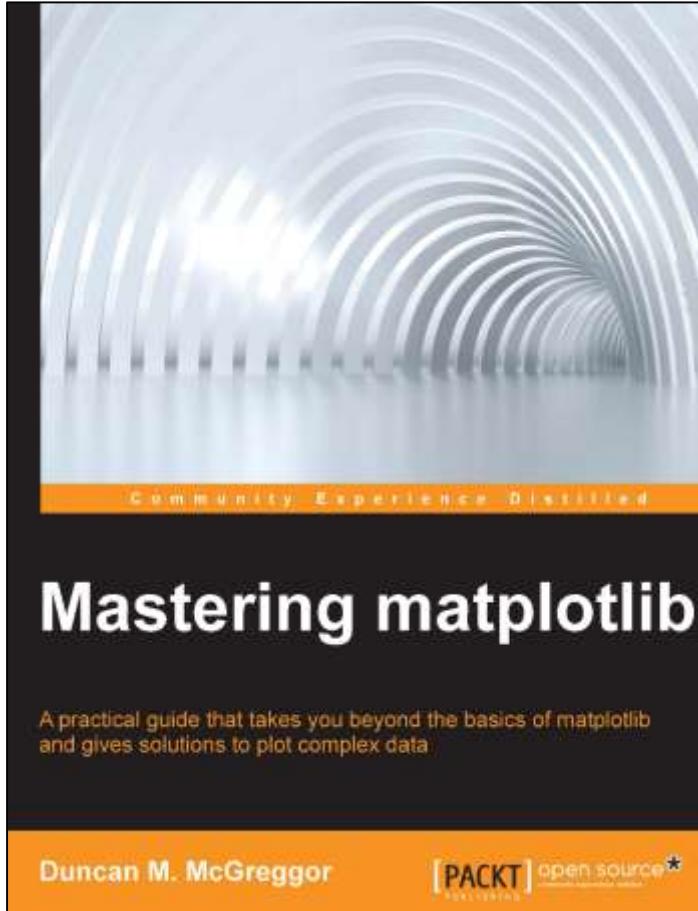


# 7 Present the Results

*Codealong – Part B*

*Prettying up Graphs*

A good resource to learn more about *matplotlib*  
(optional; not required for the course)





# Review

# Review

- What do precision and recall mean? How are they similar and different to True Positive Rate and False Positive Rate?
- What are at least two very important details to consider when creating visuals for a project's stakeholders?
- Why would an AUC plot work well for a data science audience but not for a business audience? What would be a more effective visualization for that group?

# Review (cont.)

You should now be able to:

- Evaluate a model using advanced metrics such as confusion matrix and ROC/AUC curves
- Explain the trade-offs between false positives vs. false negatives
- Describe the difference between visualization for presentations vs. exploratory data analysis
- Identify the components of a concise, convincing report and how they relate to specific audiences/stakeholders

# Next Class

*Flexible Class Session #2 / Modeling*

# Learning Objectives

After this next lesson, you should be able to:

- Review Steps **⑤** Refine the Data and **⑥** Build a Model and more specifically
  - Linear Modeling (OLS)
  - Classification Modeling (KNN and Logistic Regressions)
- Have fun doing Data Science!



# Exit Ticket

*Don't forget to fill out your exit ticket [here](#)*

Slides © 2016 Ivan Corneillet Where Applicable  
Do Not Reproduce Without Permission