GENERAL ASSEMBLY

DESIGNING WITH DATA

Seth Familian
Founder + Principal, Familian&1

SETH FAMILIAN, FOUNDER + PRINCIPAL, FAMILIAN&1

BUSINESS STRATEGY

Apple Inc.

BERTELSMANN media worldwide

get satisfaction

PRODUCT MANAGEMENT

JPG

Everywhere

GradMap

Alumni engagement, transformed.

Maptiv8

DATA MADE BEAUTIFUL

WEB PRESENCE

bergamot:station

BLUE FIELD STRATEGIES

Living Room

GROWTH HACKING

if(we)

Glip

9

CISCO

TEACHING + EDUCATION

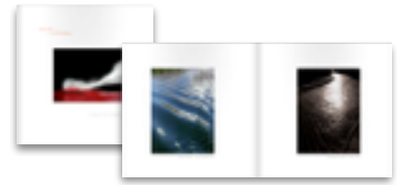
PRESIDIO GRADUATE SCHOOL

IMPACT HUB

GA GENERAL ASSEMBLY

galvanize

PROCRAFTINATION



GA WORKSHOPS I TEACH

AGENDA

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

WHAT'S GOOD (DATA) DESIGN?

CONSISTENT.

	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	$	
6	May-09	$37,440	$	
7	Jun-09	$36,321	$	
8	Jul-09	$37,352	$	
9	Aug-09	$36,577	$	
10	Sep-09	$37,262	$	
11	Oct-09	$37,362	$	
12	Nov-09	$37,684	$	
13	Dec-09	$38,861	$	

	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

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

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					
8	Jun-09	$36K	$25K	$61K			Annual Totals		
9	Jul-09	$37K	$24K	$62K		Year	VSS	SVS	Total
10	Aug-09	$37K	$24K	$60K		2009	$454K	$296K	$751K
11	Sep-09	$37K	$23K	$60K					
12	Oct-09	$37K	$22K	$59K					
13	Nov-09	$38K	$24K	$61K					
14	Dec-09	$39K	$23K	$62K					

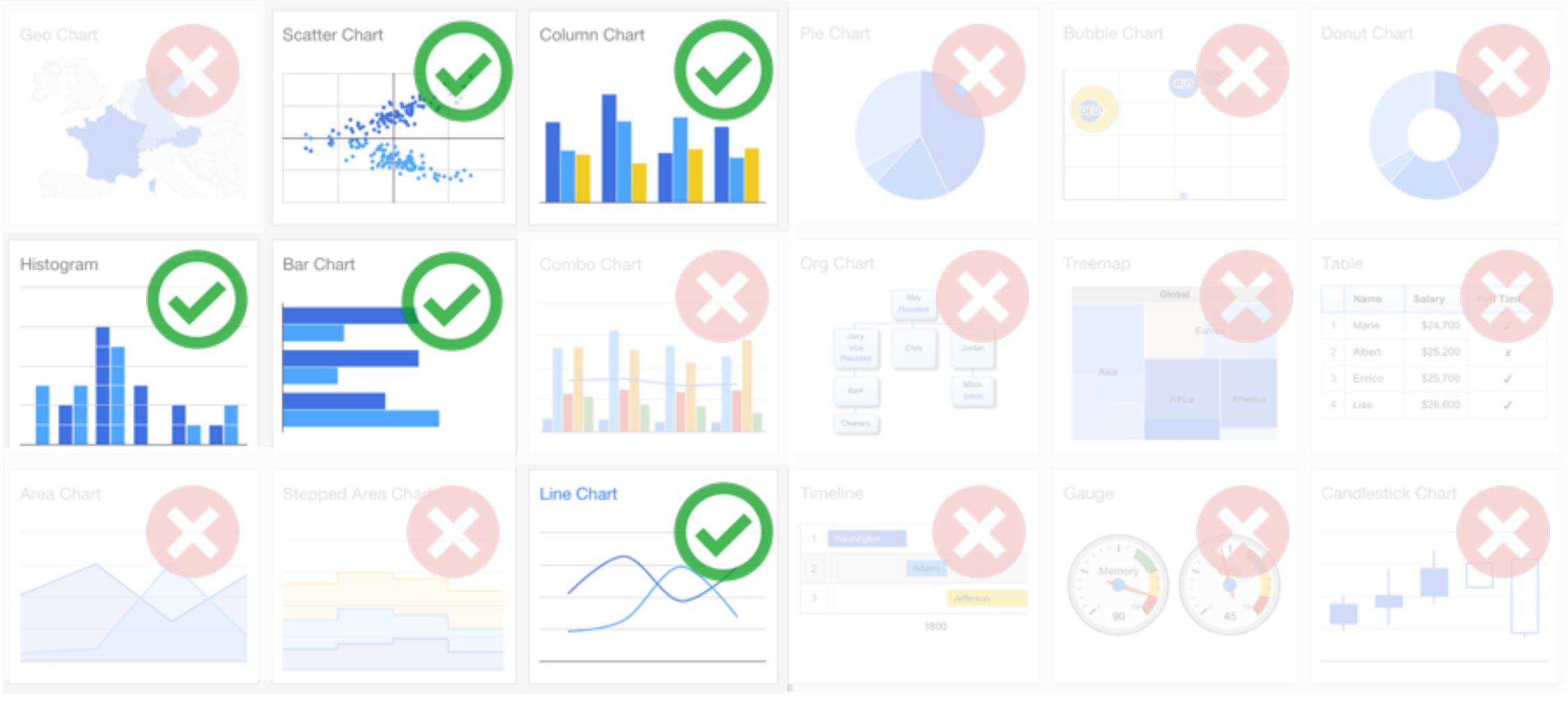
MUTED.

	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					
8	Jun-09	$36K	$25K	$61K		Annual Totals + Trends			
9	Jul-09	$37K	$24K	$62K		Year	VSS	SVS	Total
10	Aug-09	$37K	$24K	$60K		2009	$454K	$296K	$751K
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

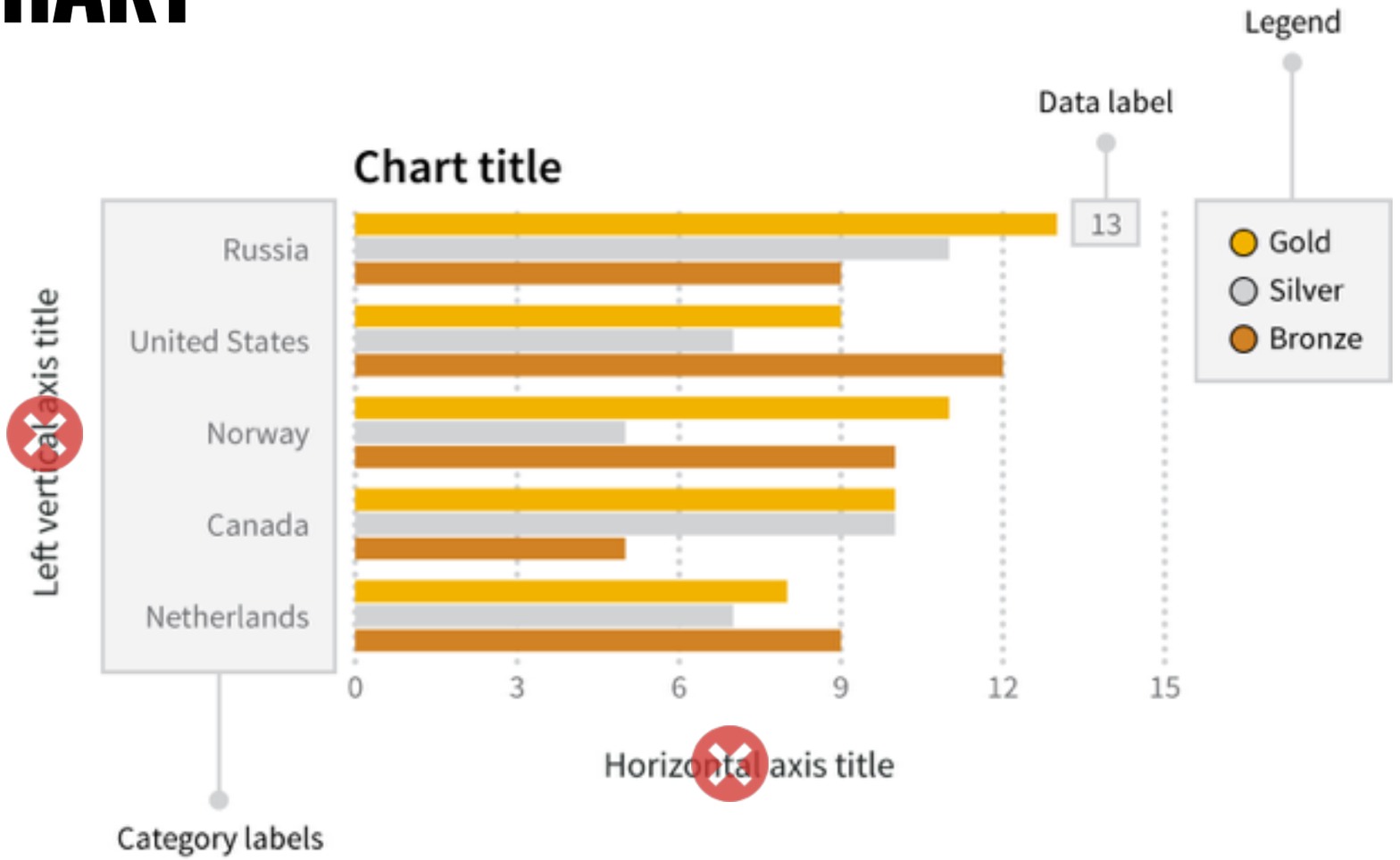
CHART TYPES

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



ANATOMY OF A CHART

FROM



https://infoactive.co/data-design/ch14.html

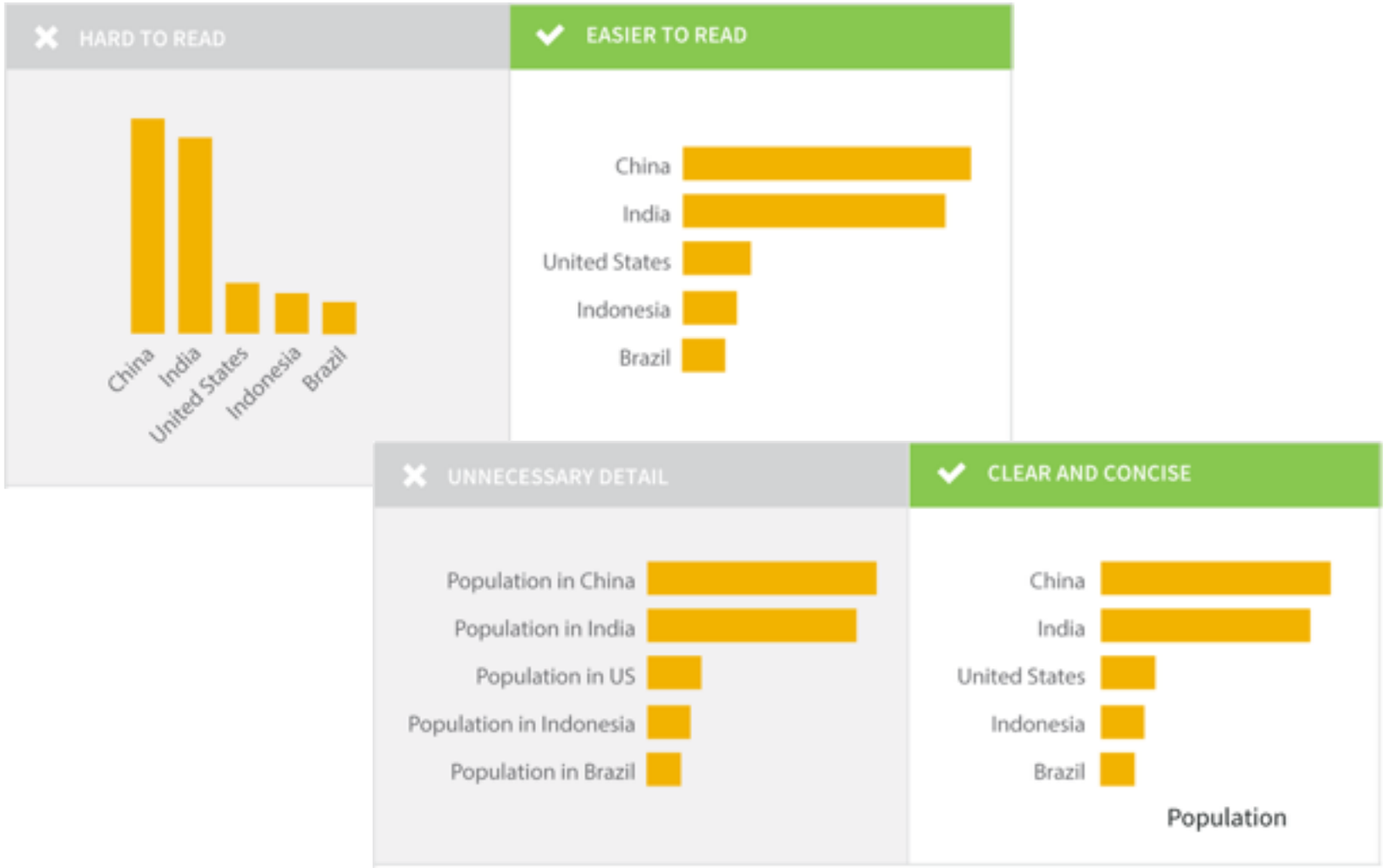
AXIS TITLES

FROM

https://infoactive.co/data-design/ch14.html

AXIS TITLES

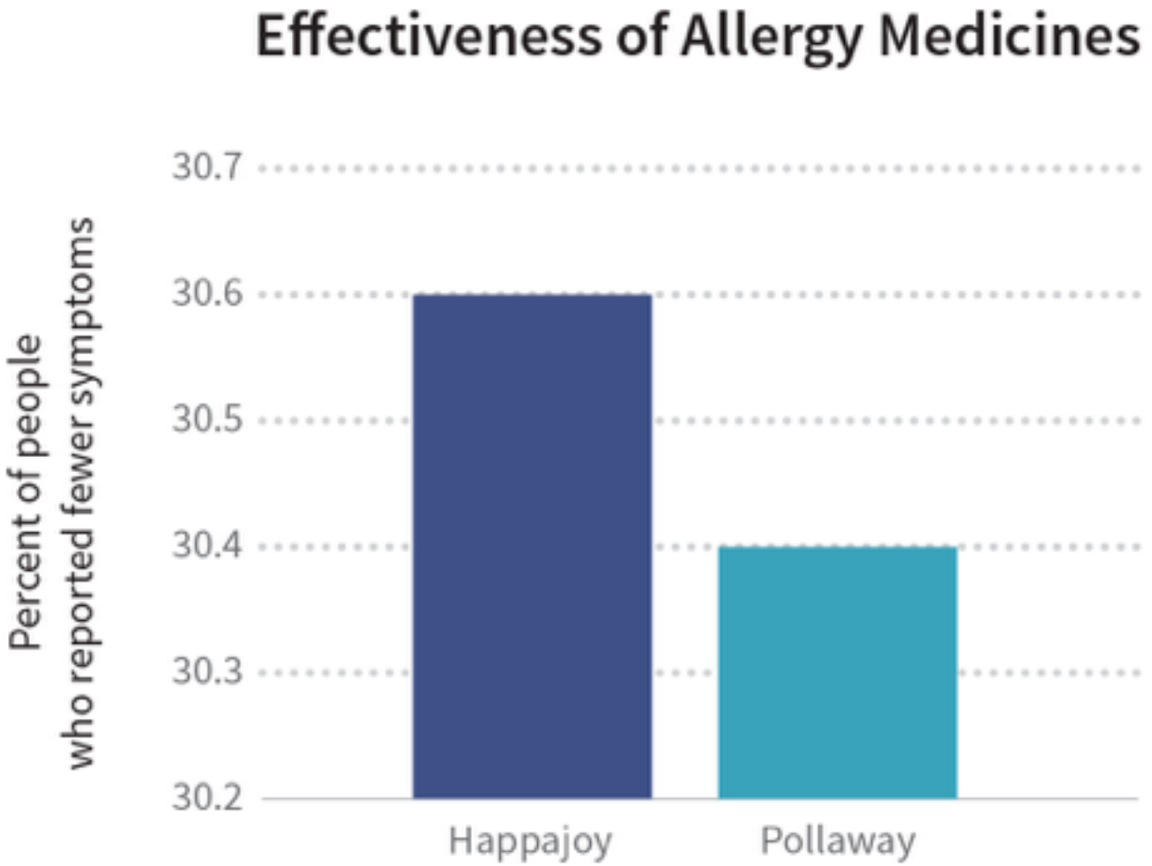
FROM



https://infoactive.co/data-design/ch14.html

DON'T TRUNCATE AXES

FROM

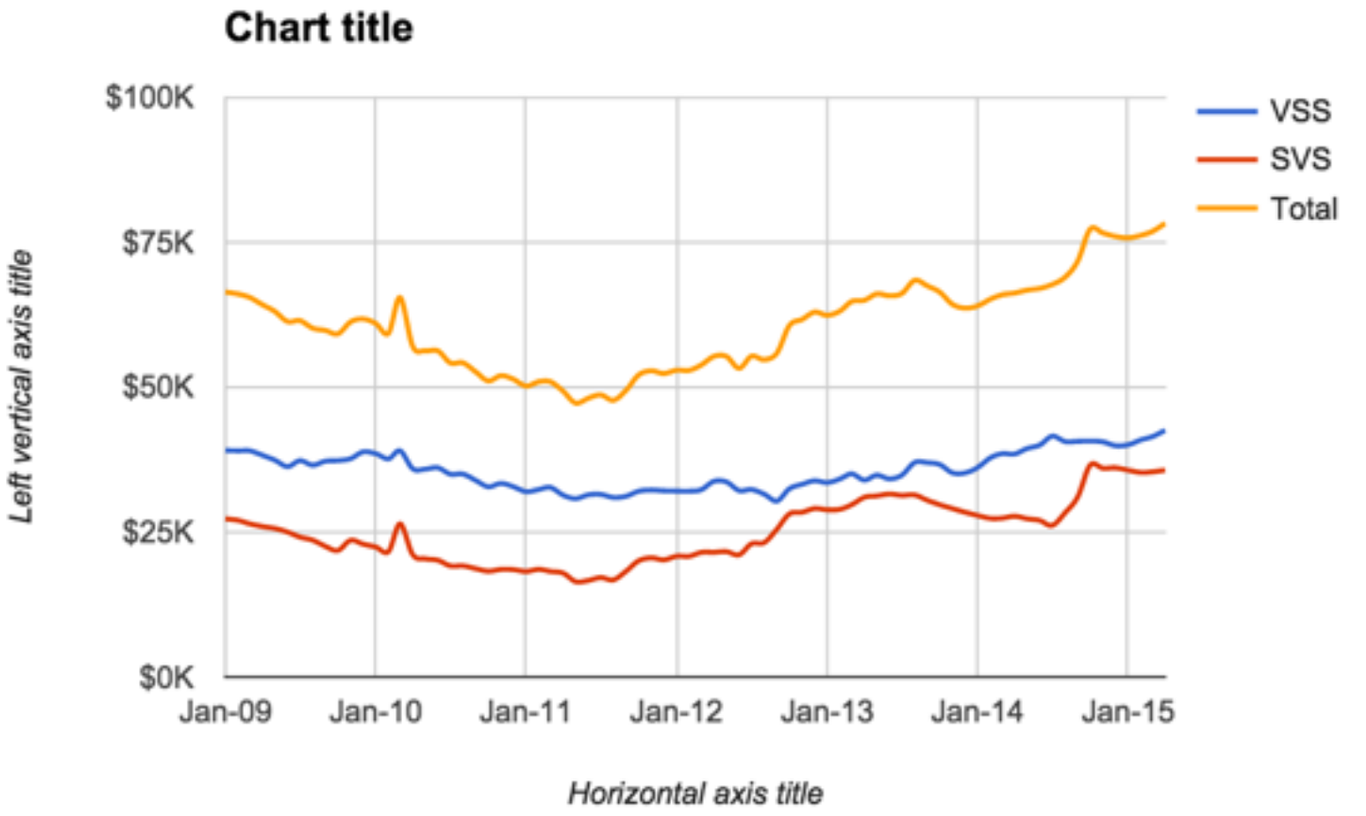


https://infoactive.co/data-design/ch18.html

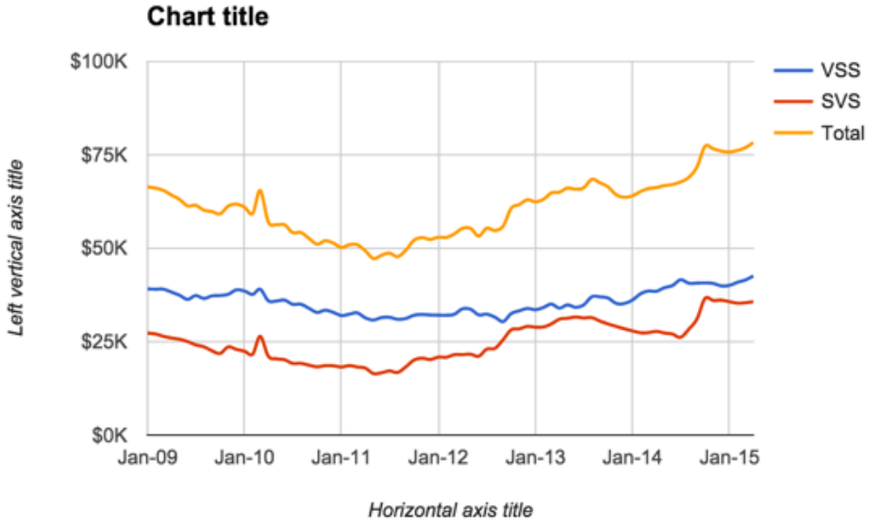
BACK TO OUR DATA…

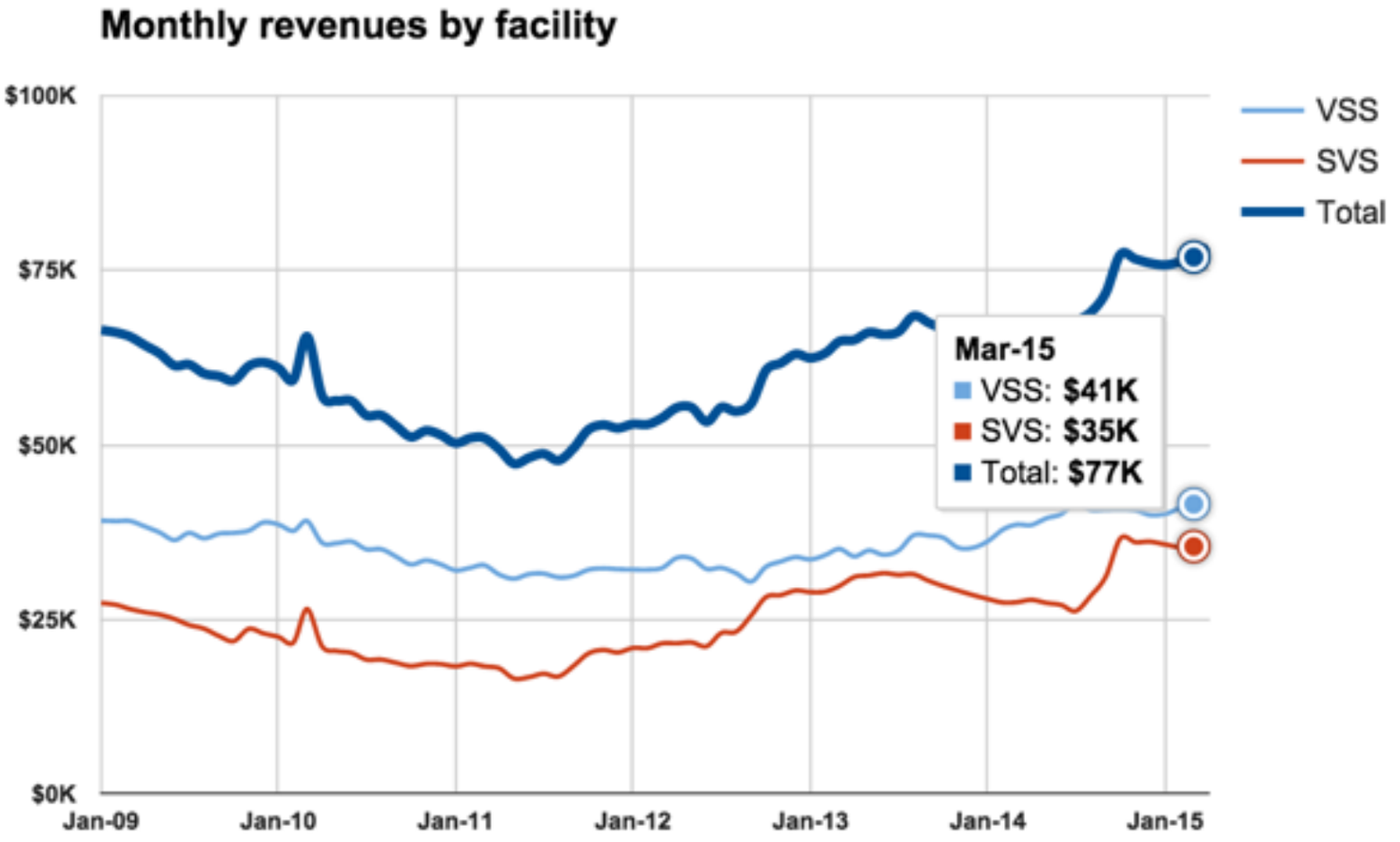
Monthly Trends			
Date	VSS	SVS	Total
Jan-09	$39K	$27K	$66K
Feb-09	$39K	$27K	$66K
Mar-09	$39K	$26K	$66K
Apr-09	$38K	$26K	$64K
May-09	$37K	$26K	$63K
Jun-09	$36K	$25K	$61K
Jul-09	$37K	$24K	$62K
Aug-09	$37K	$24K	$60K
Sep-09	$37K	$23K	$60K
Oct-09	$37K	$22K	$59K
Nov-09	$38K	$24K	$61K
Dec-09	$39K	$23K	$62K
Jan-10	$39K	$22K	$61K
Feb-10	$38K	$22K	$59K
Mar-10	$39K	$26K	$66K
Apr-10	$36K	$21K	$57K
May-10	$36K	$20K	$56K
Jun-10	$36K	$20K	$56K
Jul-10	$35K	$19K	$54K
Aug-10	$35K	$19K	$54K
Sep-10	$34K	$19K	$53K
Oct-10	$33K	$18K	$51K
Nov-10	$33K	$19K	$52K
Dec-10	$33K	$19K	$51K
Jan-11	$32K	$18K	$50K
Feb-11	$32K	$19K	$51K

Quarterly Trends			
Quarter	VSS	SVS	Total
Q1-09	$117K	$81K	$198K
Q2-09	$112K	$77K	$189K
Q3-09	$111K	$70K	$182K
Q4-09	$114K	$68K	$182K
Q1-10	$115K	$71K	$186K
Q2-10	$108K	$62K	$170K
Q3-10	$104K	$57K	$161K
Q4-10	$99K	$55K	$155K
Q1-11	$97K	$55K	$152K
Q2-11	$94K	$51K	$145K
Q3-11	$94K	$52K	$146K
Q4-11	$96K	$61K	$157K
Q1-12	$97K	$63K	$160K
Q2-12	$100K	$64K	$164K
Q3-12	$94K	$72K	$166K
Q4-12	$100K	$86K	$185K
Q1-13	$103K	$88K	$190K
Q2-13	$103K	$94K	$197K
Q3-13	$109K	$93K	$202K
Q4-13	$107K	$87K	$194K
Q1-14	$113K	$83K	$195K
Q2-14	$118K	$82K	$200K
Q3-14	$123K	$86K	$209K
Q4-14	$121K	$109K	$230K
Q1-15	$122K	$106K	$229K
Q2-15	$43K	$36K	$78K

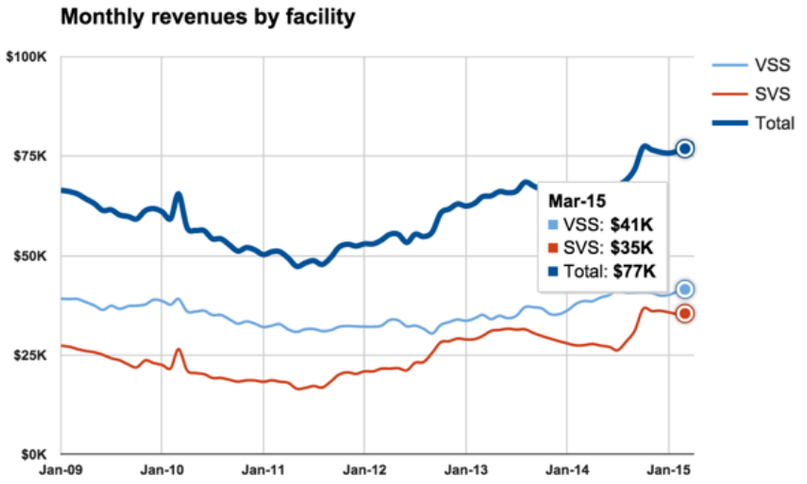


APPLYING DESIGN PRINCIPLES

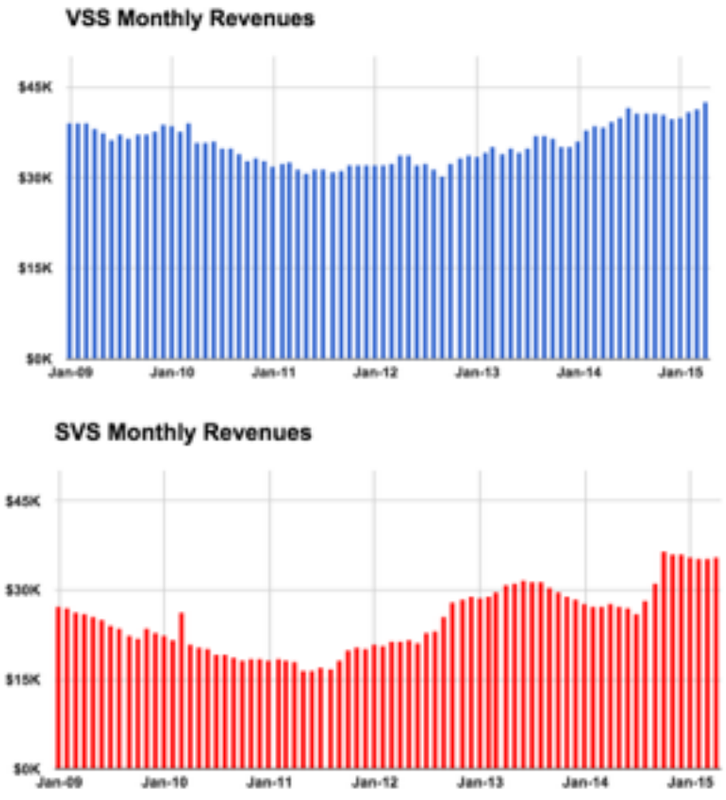


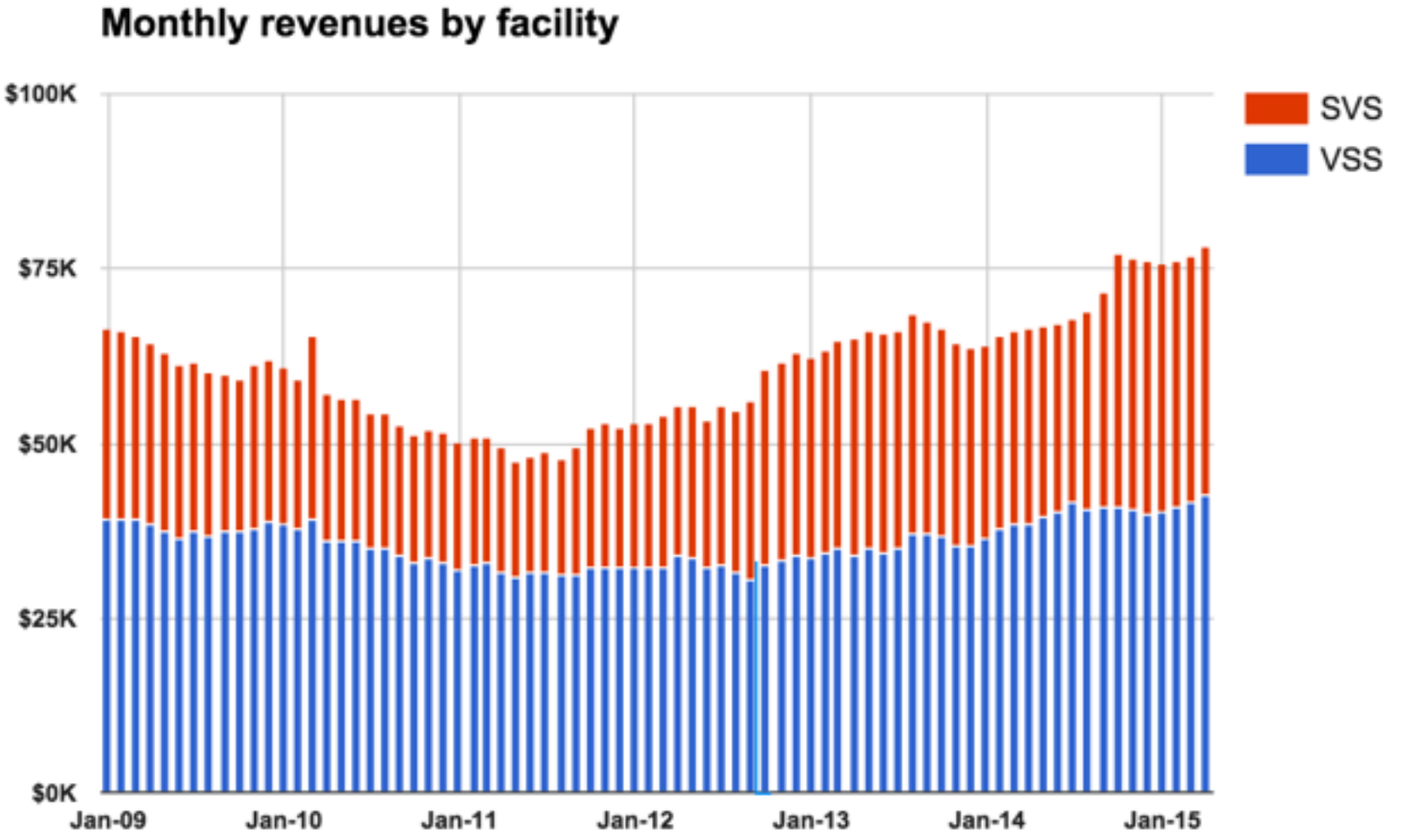


ALTERNATIVE TIME-SERIES



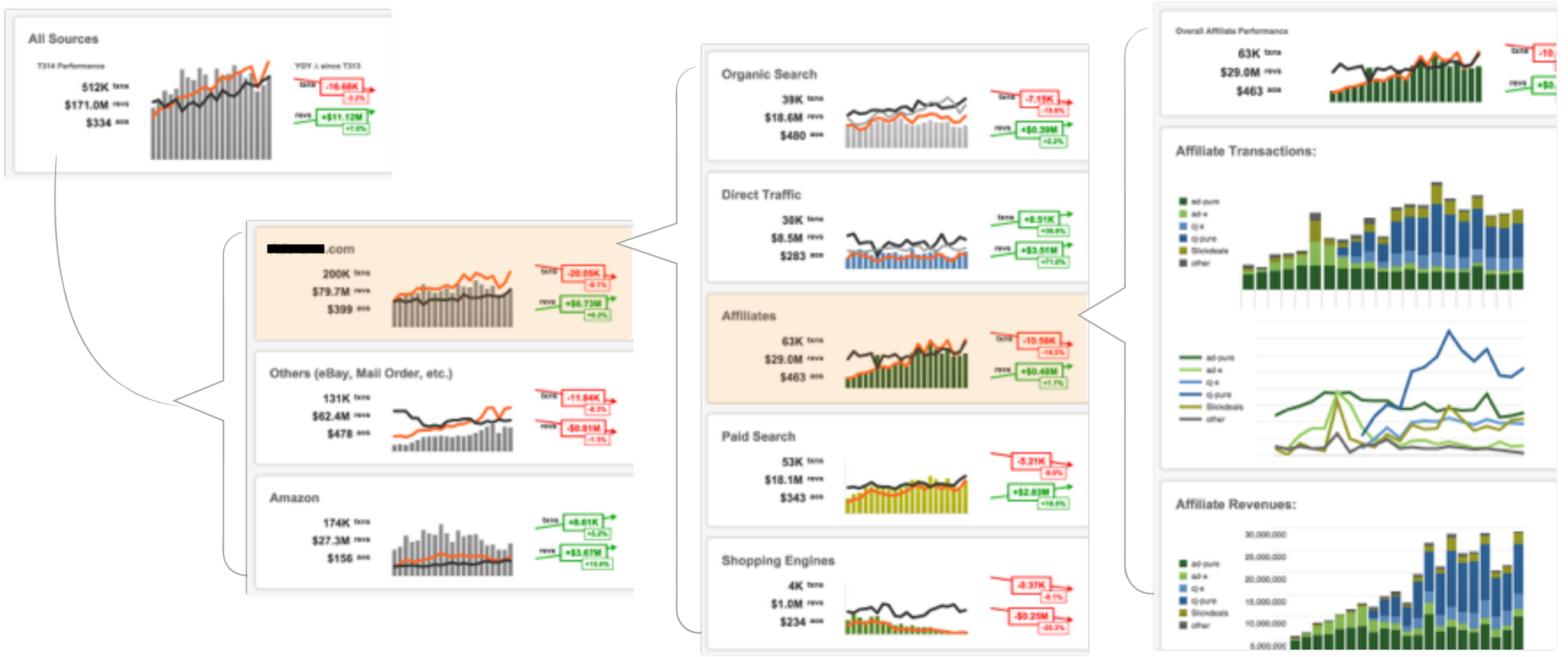
STACKED BARS





DATA VIZ TOOLS + TECHNIQUES

NESTED CHARTS + SMALL MULTIPLES



SPARKLINES

Tasks completed by team members

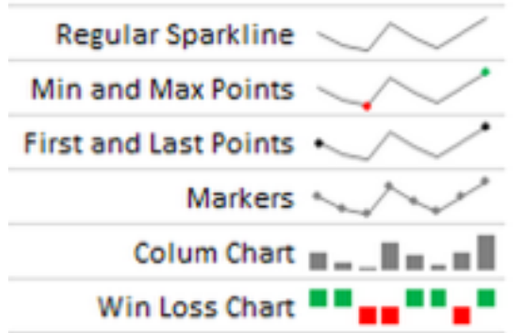
(last 26 weeks, YoY change shown in %s)

Team Member	Total Tasks Completed		w1	w2	w3	w25	w26
Julie		46%	13	15	19	11	19
John		45%	11	18	11	14	16
Jabba the hut		-20%	15	14	14	19	12
Johnson		6%	18	17	14	12	19
Jeremy		43%	14	20	10	12	20
Josh		-33%	15	12	19	11	10

Sum of Net Sales		Month				
Salesman	Region	Jan 07	Feb 07	Mar 07	Apr 07	May
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	502
Matt		4204	5886	3238	5634	477
		4512	5330	4052	3061	34
		6850	5277	4257	4901	64
Grand Total		56886	57884	51621	59781	624

http://chandoo.org/wp/2010/05/18/excel-sparklines-tutorial/

Types of Sparklines



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

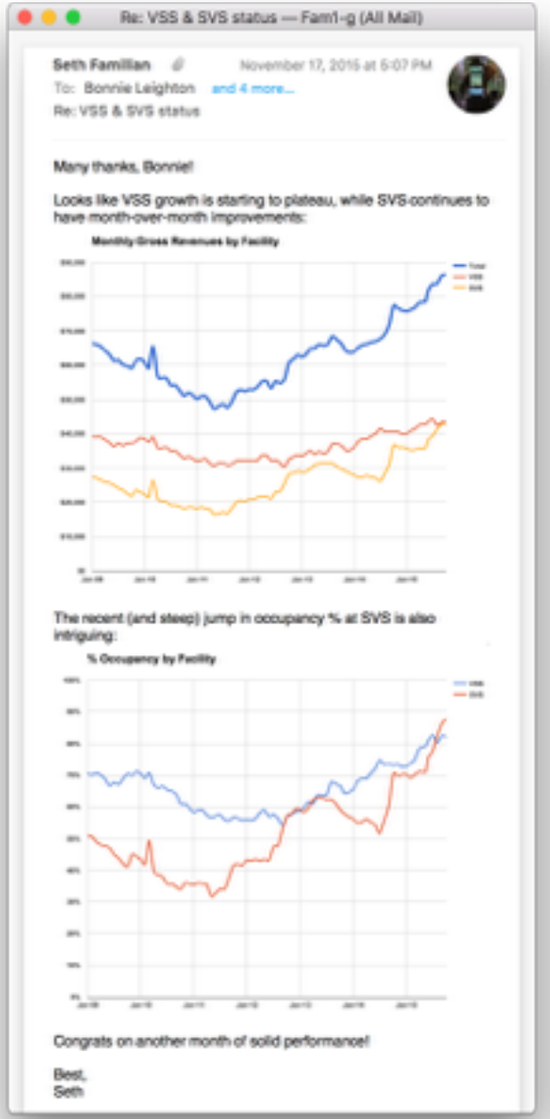
SPARKLINES

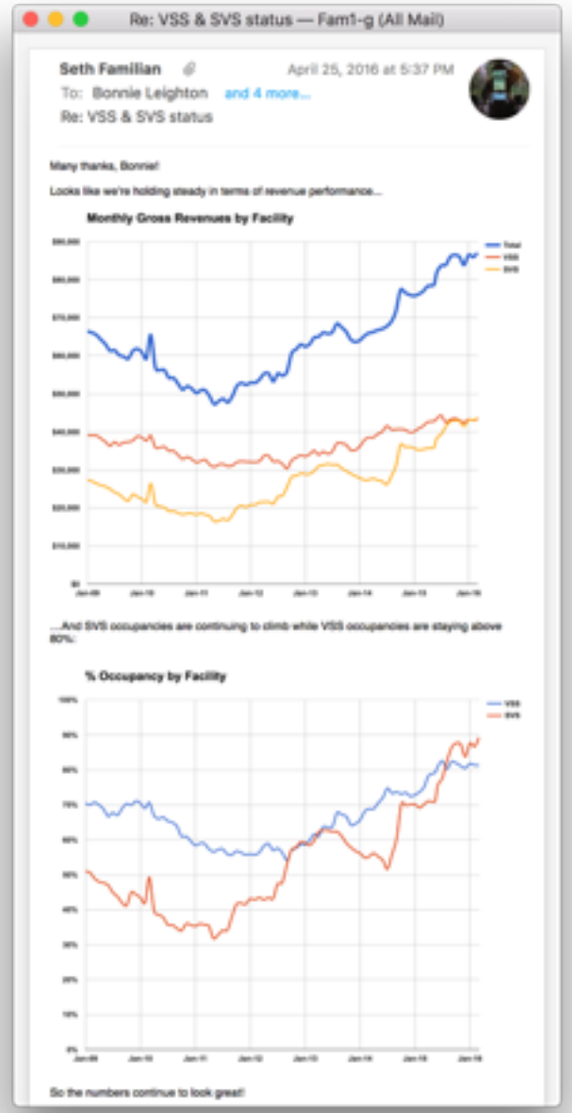
=SPARKLINE(B7:B36)

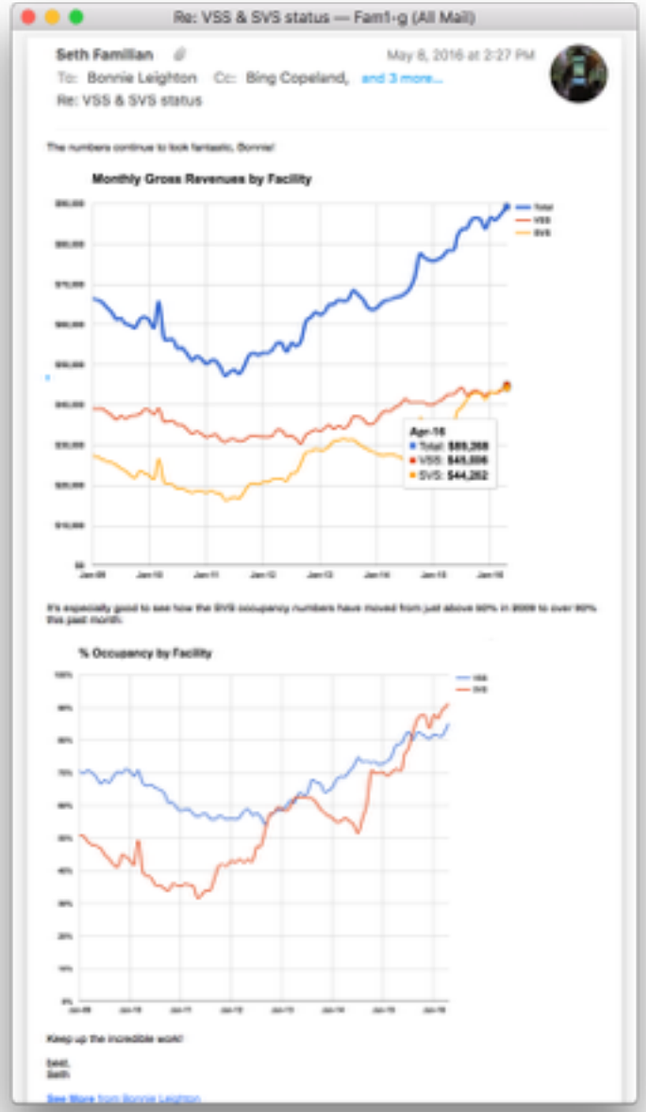
A	B	C	D	E	F	G	H	I	J	K	L
	SIGNUPS						VIRAL CONVERSIONS				COMPANIES
											
Past 30 Days	5,702	5,402	41	199	60		9,109	5,443	60%		1,747
	Total	Glip Viral	RC Viral	Glip Org.	RC Org.		Invites	Signups	Conv.		New Co's.
04-30-2016 0:00:00	268	41	0	53	174		37	41	111%		16
05-01-2016 0:00:00	329	79	0	71	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		304	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

http://chandoo.org/wp/2010/05/18/excel-sparklines-tutorial/

NARRATIVE!

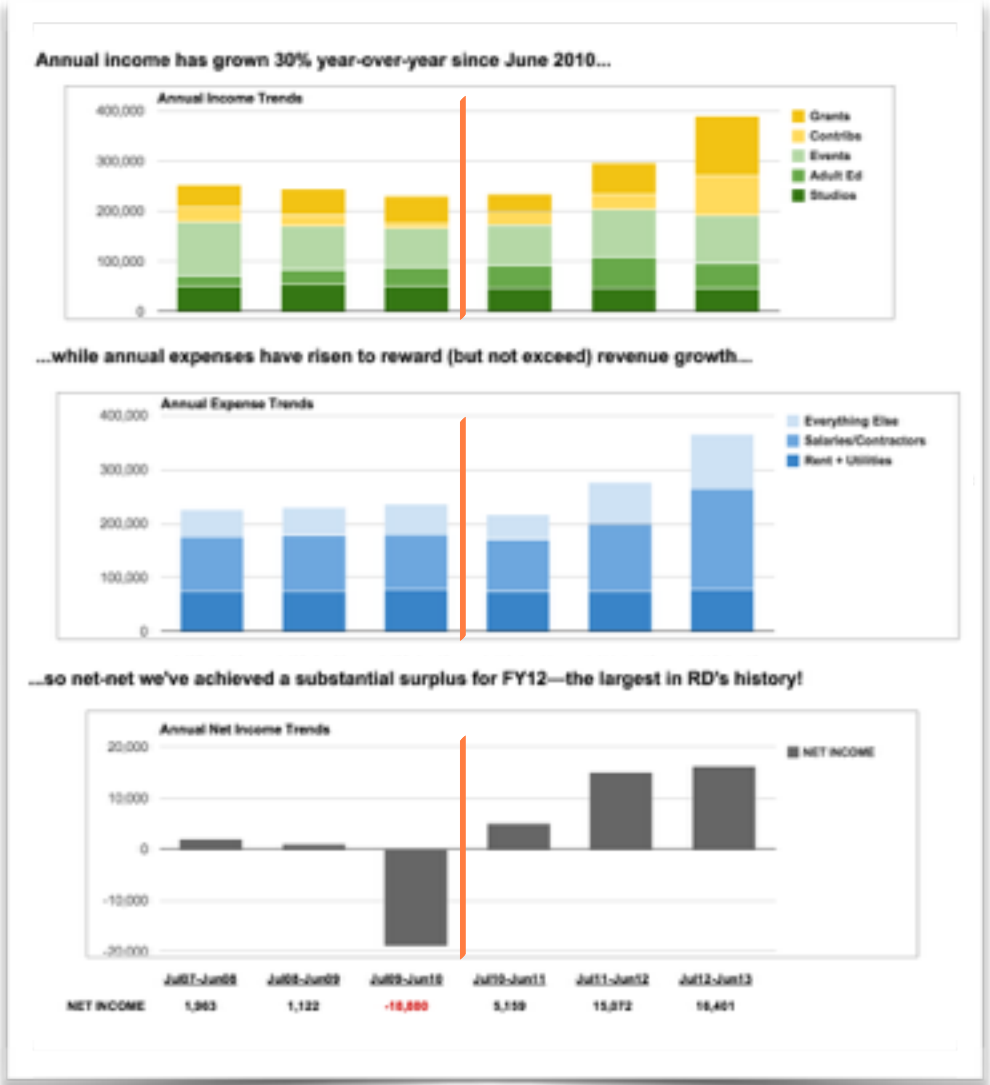

Re: VSS & SVS status — Fam1-g (All Mail)
Seth Familian
November 17, 2016 at 5:07 PM
To: Bonnie Leighton and 4 more...
Re: VSS & SVS status
Many thanks, Bonnie!
Looks like VSS growth is starting to plateau, while SVS continues to have month-over-month improvements:
Monthly Gross Revenues by Facility
The recent (and steep) jump in occupancy % at SVS is also intriguing:
% Occupancy by Facility
Congrats on another month of solid performance!
Best,
Seth


Re: VSS & SVS status — Fam1-g (All Mail)
Seth Familian
April 25, 2016 at 5:37 PM
To: Bonnie Leighton and 4 more...
Re: VSS & SVS status
Many thanks, Bonnie!
Looks like we're holding steady in terms of revenue performance...
Monthly Gross Revenues by Facility
...And SVS occupancies are continuing to climb while VSS occupancies are staying above 80%:
% Occupancy by Facility
So the numbers continue to look great!


Re: VSS & SVS status — Fam1-g (All Mail)
Seth Familian
May 8, 2016 at 2:27 PM
To: Bonnie Leighton Cc: Bing Copeland, and 3 more...
Re: VSS & SVS status
The numbers continue to look fantastic, Bonnie!
Monthly Gross Revenues by Facility
Apr-16
Total: $89,268
VSS: $45,024
SVS: $44,282
% Occupancy by Facility
Keep up the incredible work!
best,
Seth
See More from Bonnie Leighton

NARRATIVE!


Budget/Actual Combined
Income
TOTAL INCOME
Cost of Goods Sold
TOTAL COGS
GROSS PROFIT
Expense
TOTAL EXPENSES


Annual income has grown 30% year-over-year since June 2010...
Annual Income Trends
400,000
300,000
200,000
100,000
0
Grants
Contribs
Events
Adult Ed
Studios
...while annual expenses have risen to reward (but not exceed) revenue growth...
Annual Expense Trends
400,000
300,000
200,000
100,000
0
Everything Else
Salaries/Contractors
Rent + Utilities
...so net-net we've achieved a substantial surplus for FY12—the largest in RD's history!
Annual Net Income Trends
20,000
10,000
0
-10,000
-20,000
NET INCOME
Jul07-Jun08 Jul08-Jun09 Jul09-Jun10 Jul10-Jun11 Jul11-Jun12 Jul12-Jun13
NET INCOME 1,963 1,122 -18,880 5,159 15,072 16,401

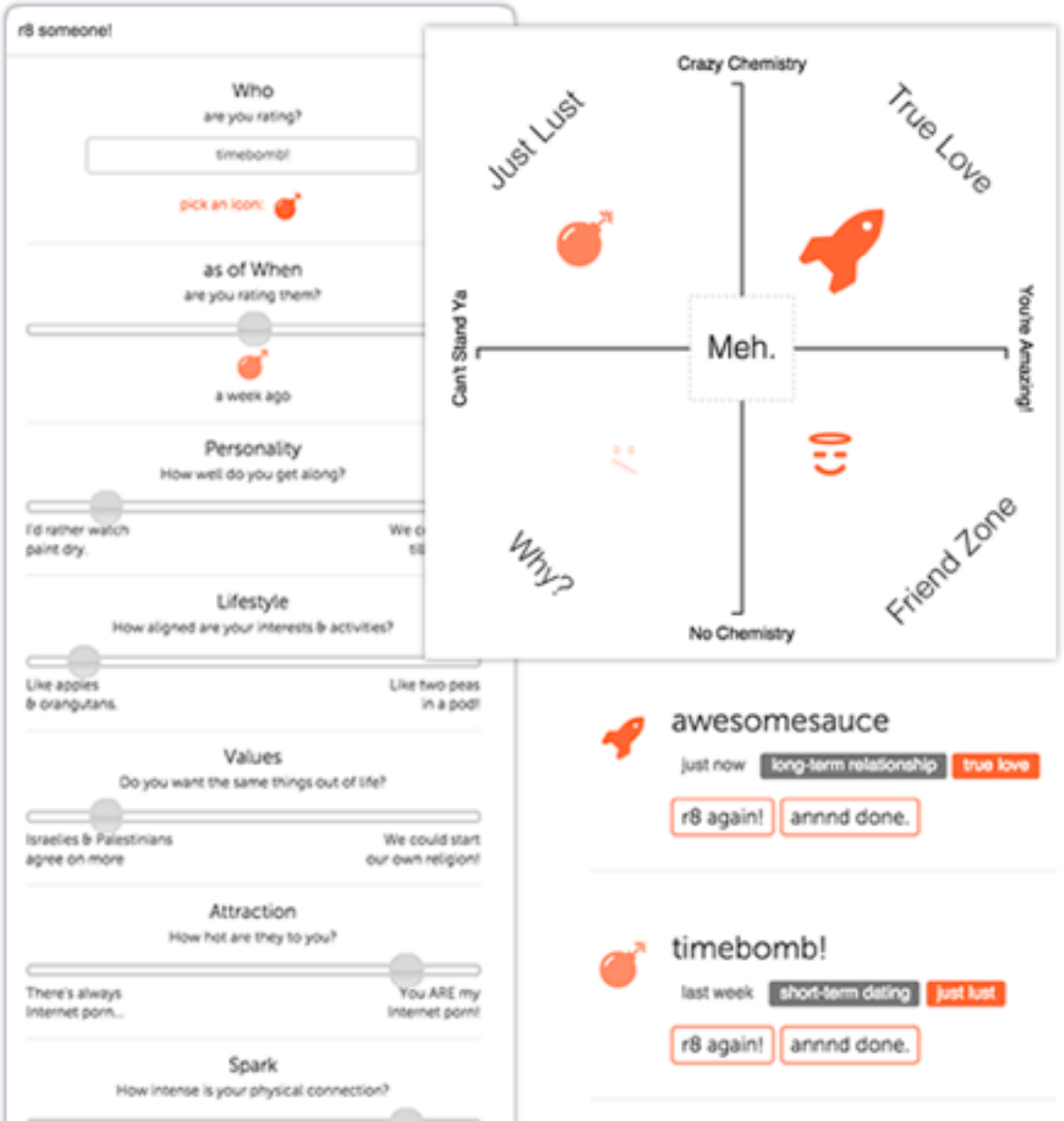
2x2s

GARTNER'S MAGIC QUADRANT

Source: Gartner (February 2014)

2x2s

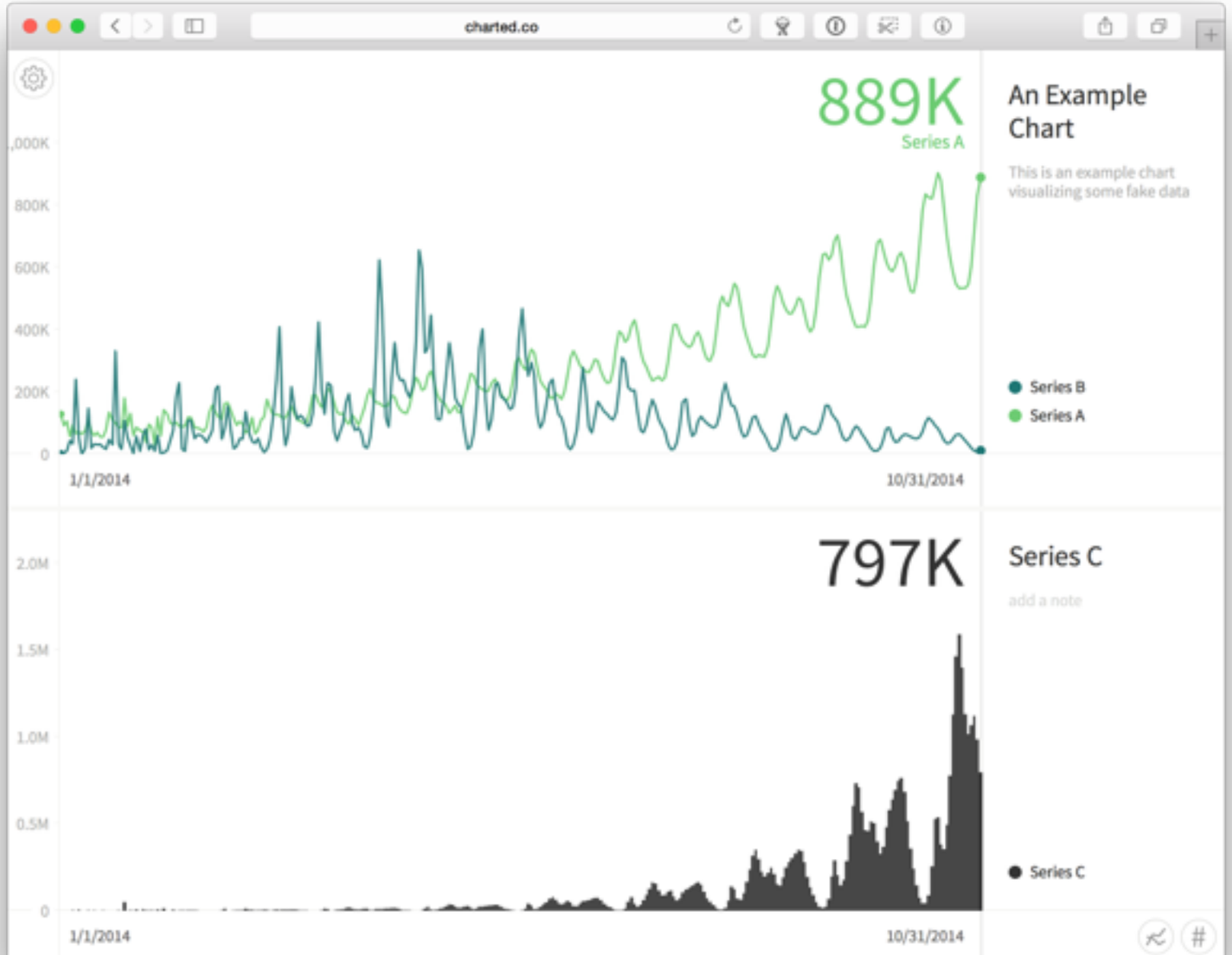

alpha
d
r
8
r
r
privately
visualize
your love life


r8 someone!
Who
are you rating?
timebomb!
pick an icon:
as of When
are you rating them?
a week ago
Personality
How well do you get along?
I'd rather watch paint dry.
Lifestyle
How aligned are your interests & activities?
Like apples & orangutans.
Like two peas in a pod!
Values
Do you want the same things out of life?
Israelies & Palestinians agree on more
We could start our own religion!
Attraction
How hot are they to you?
There's always internet porn...
You ARE my internet porn!
Spark
How intense is your physical connection?
Crazy Chemistry
Just Lust
True Love
Can't Stand Ya
Meh.
You're Amazing!
Why?
Friend Zone
No Chemistry
awesomesauce
just now
long-term relationship
true love
r8 again!
annnd done.
timebomb!
last week
short-term dating
just lust
r8 again!
annnd done.

CHARTED

FOR SUPER SIMPLE CHARTS

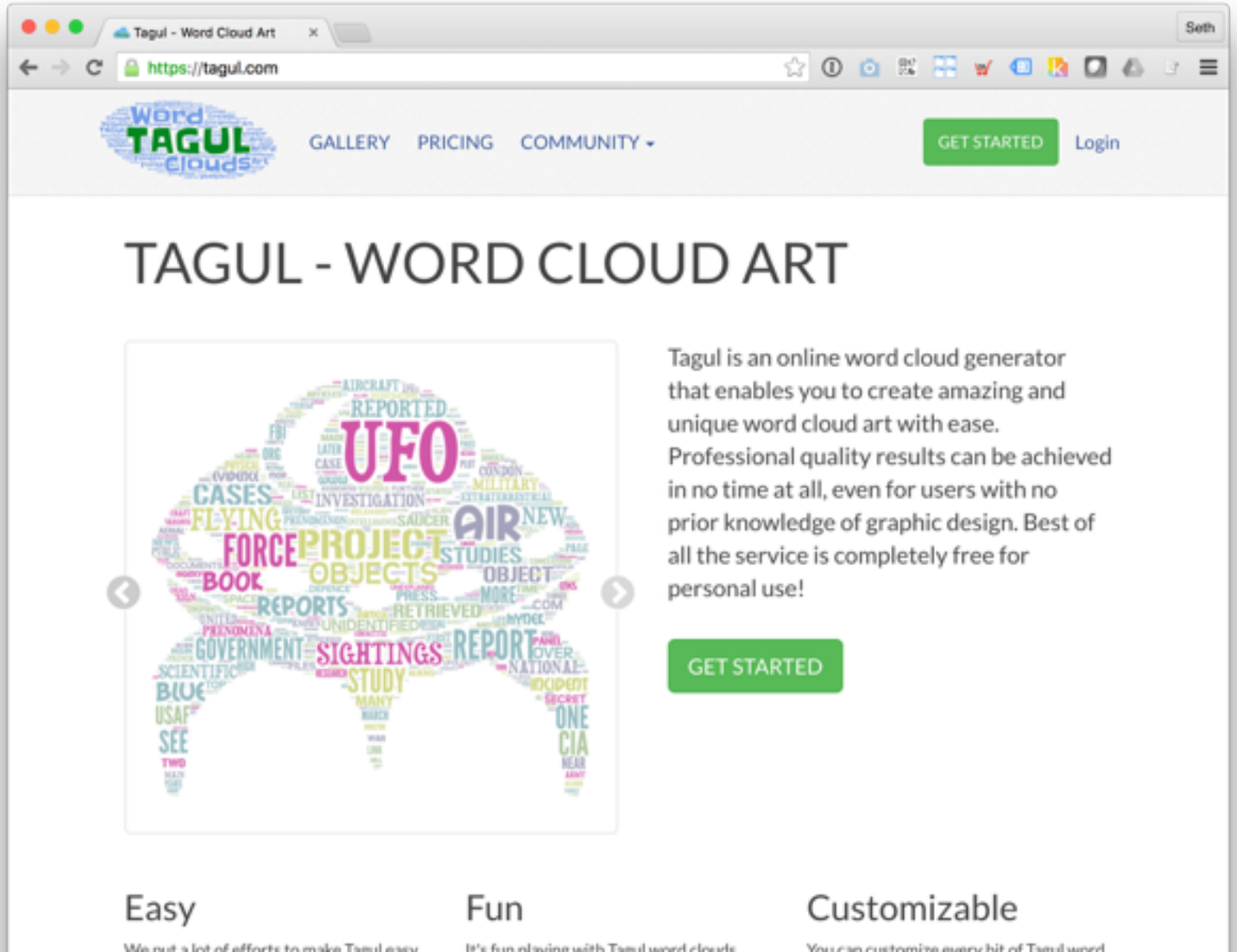
CHARTED.CO



TAGUL

FOR GORGEOUS WORD CLOUDS

TAGUL.COM

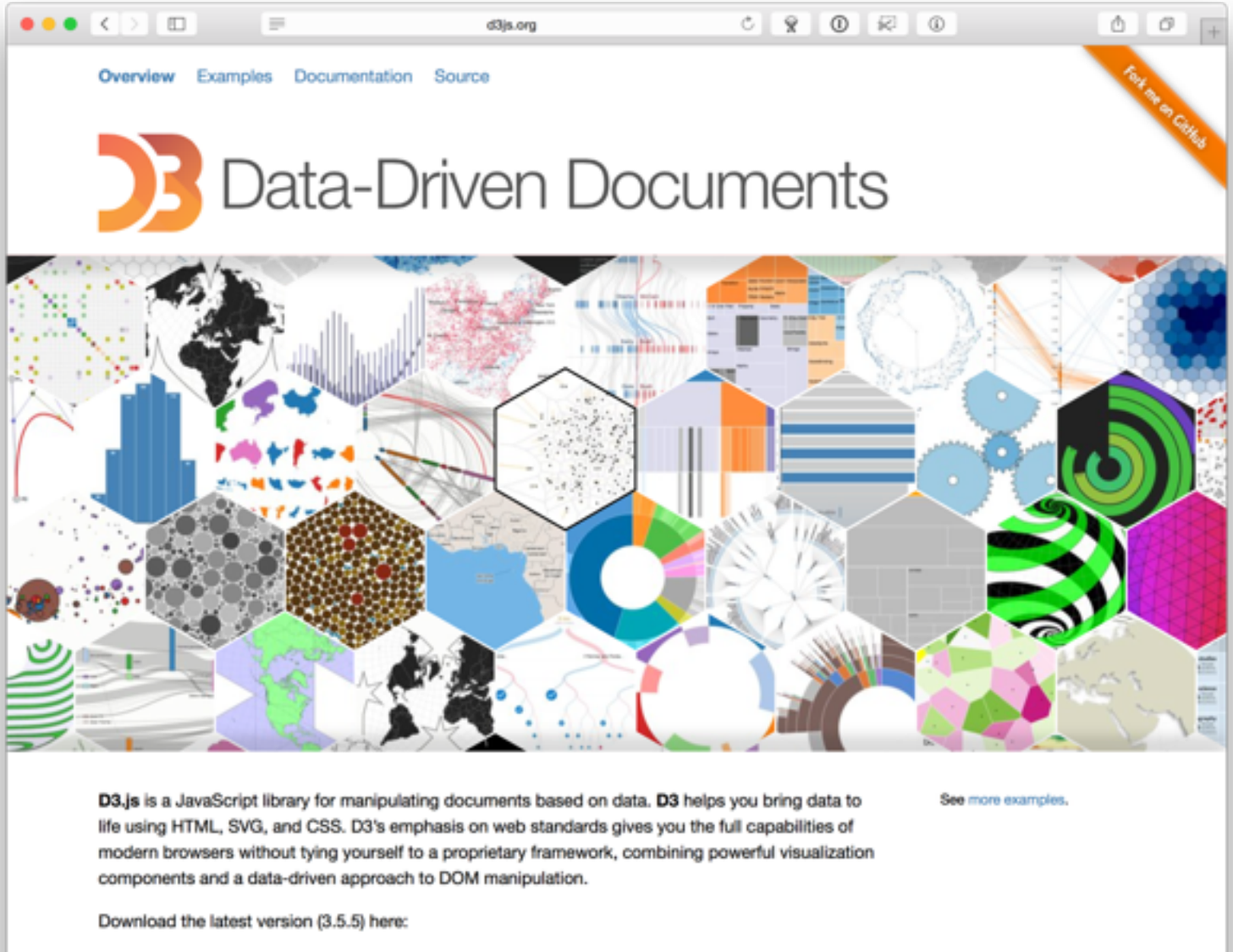


ADVANCED DATA VIZ TOOLS

D3.JS

FOR AMAZING(LY COMPLEX) DATA VIZ

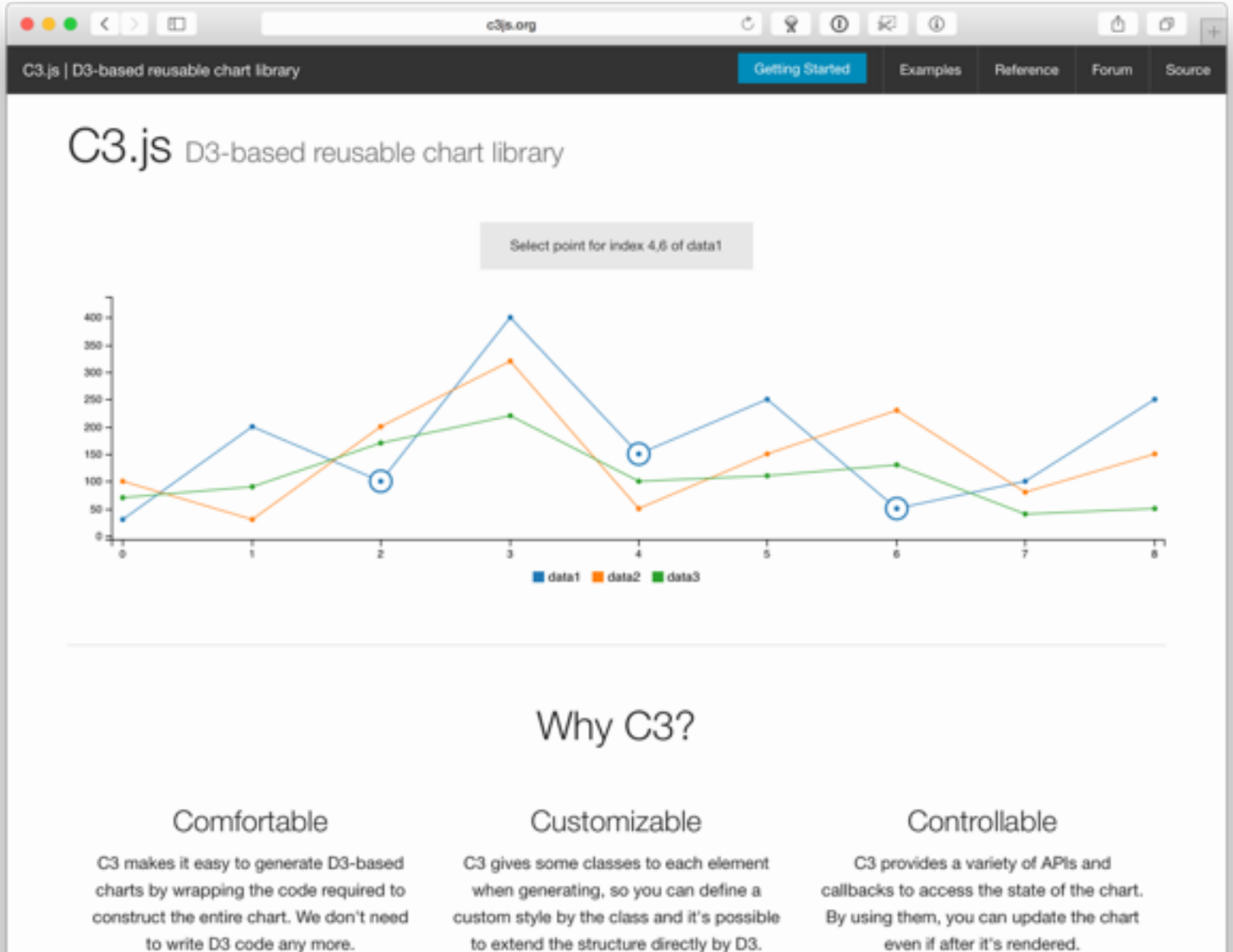
D3JS.ORG



C3.JS

D3 META-LIBRARY

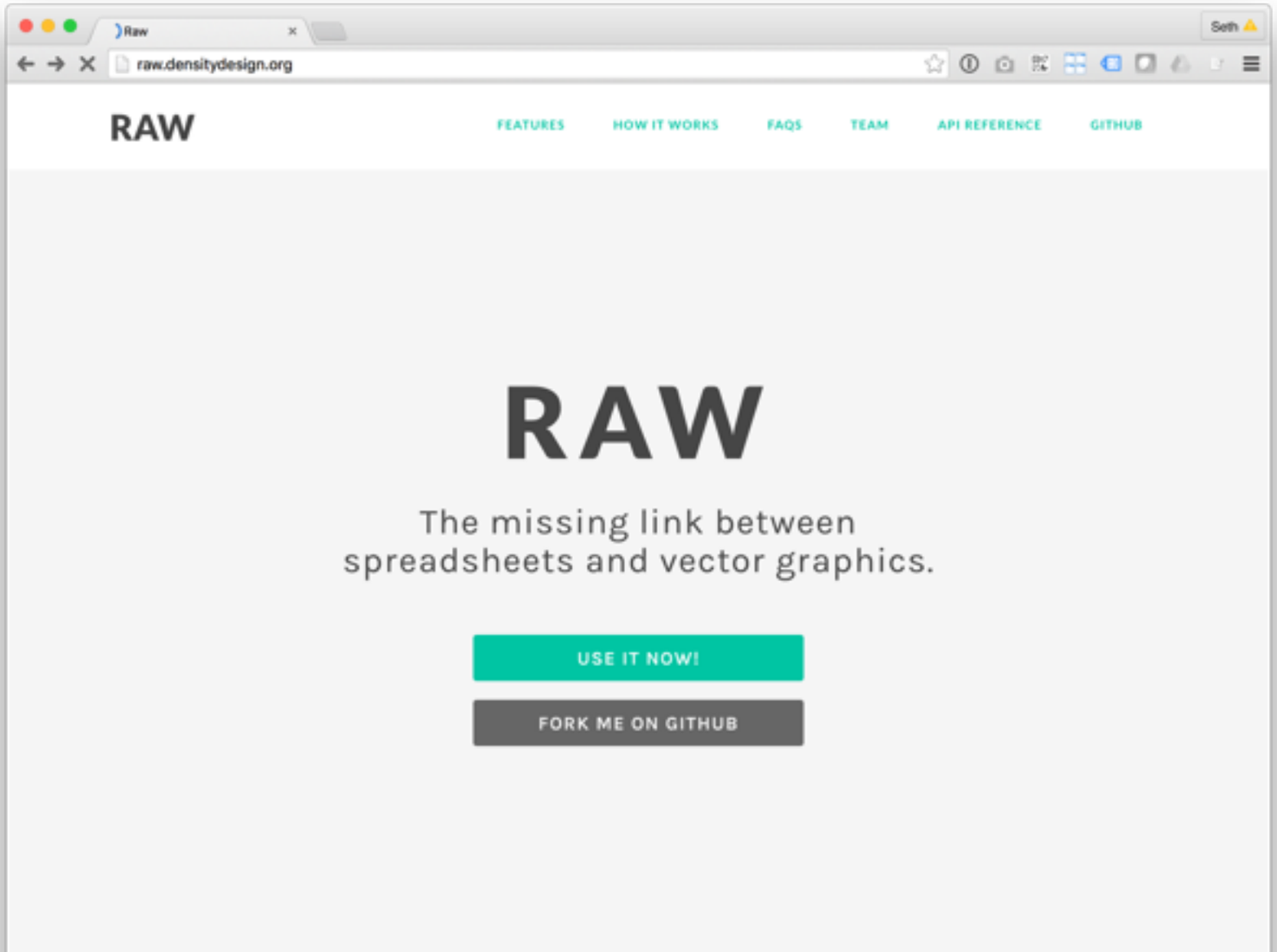
C3JS.ORG



RAW

D3 META-LIBRARY

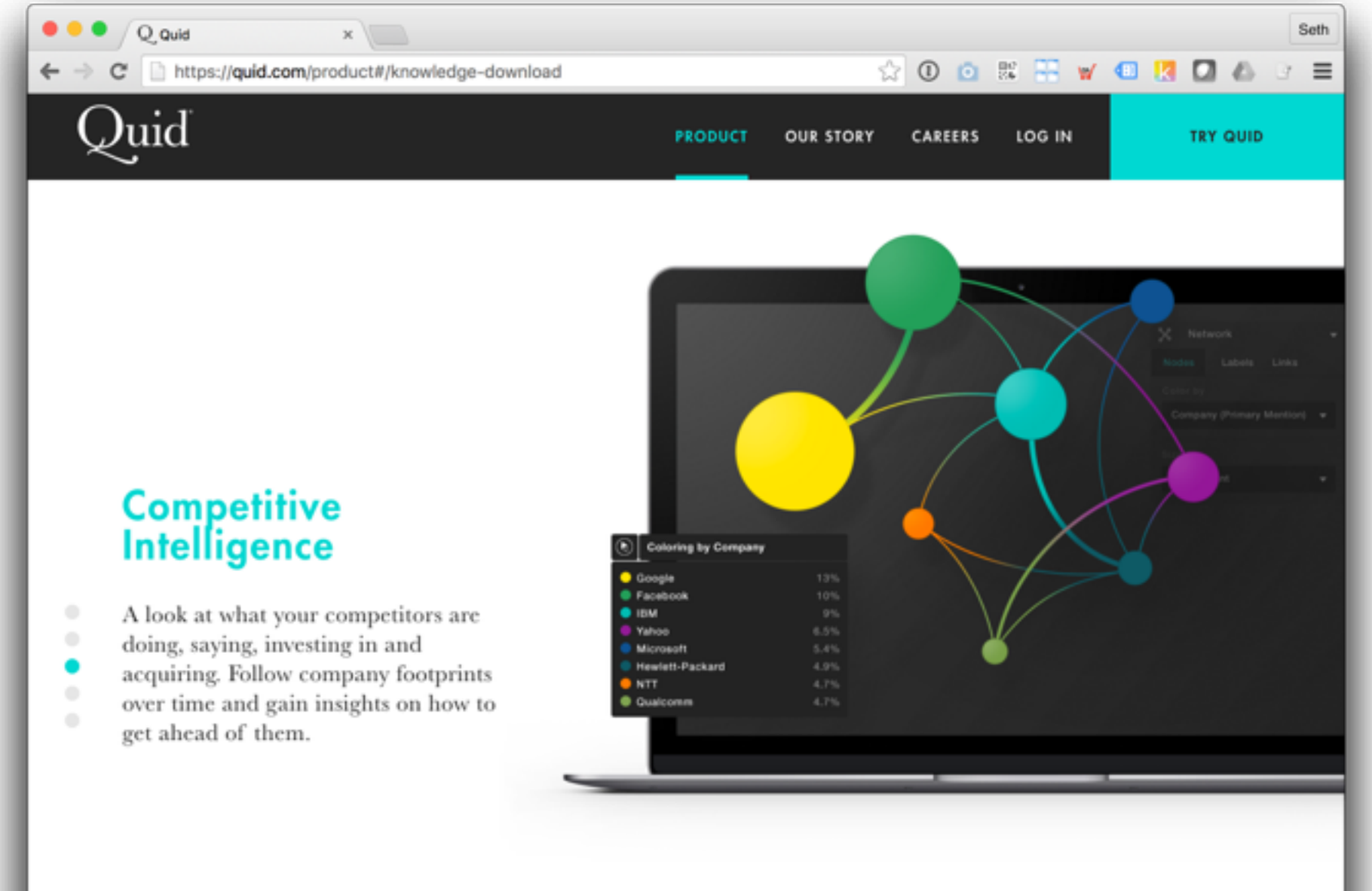
RAW.DENSITYDESIGN.ORG



QUID

FOR UNSTRUCTURED ANALYSIS

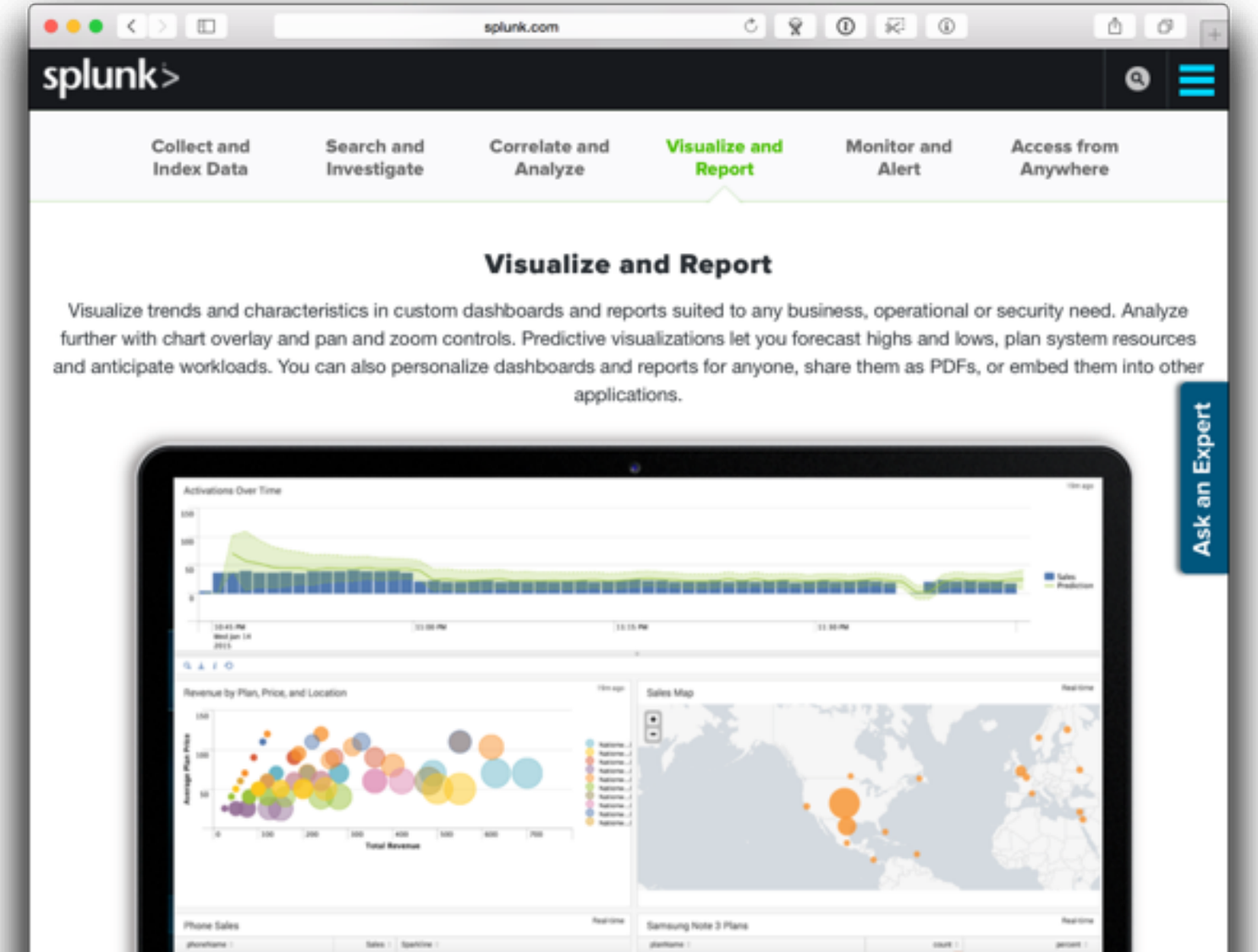
QUID.COM



SPLUNK

FOR ANY MACHINE DATA

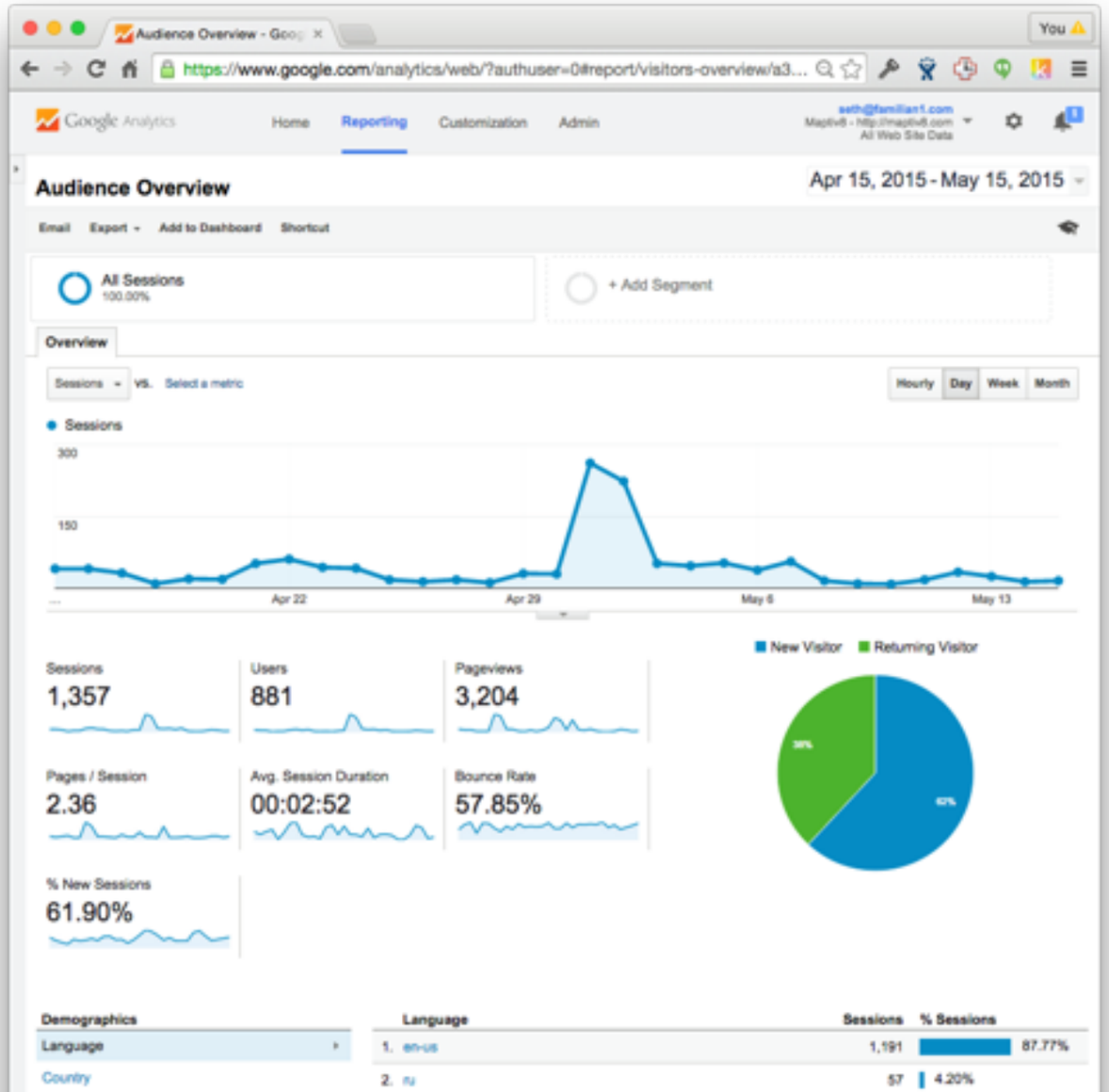
SPLUNK.COM



GOOGLE ANALYTICS

FOR WEBSITE TRAFFIC

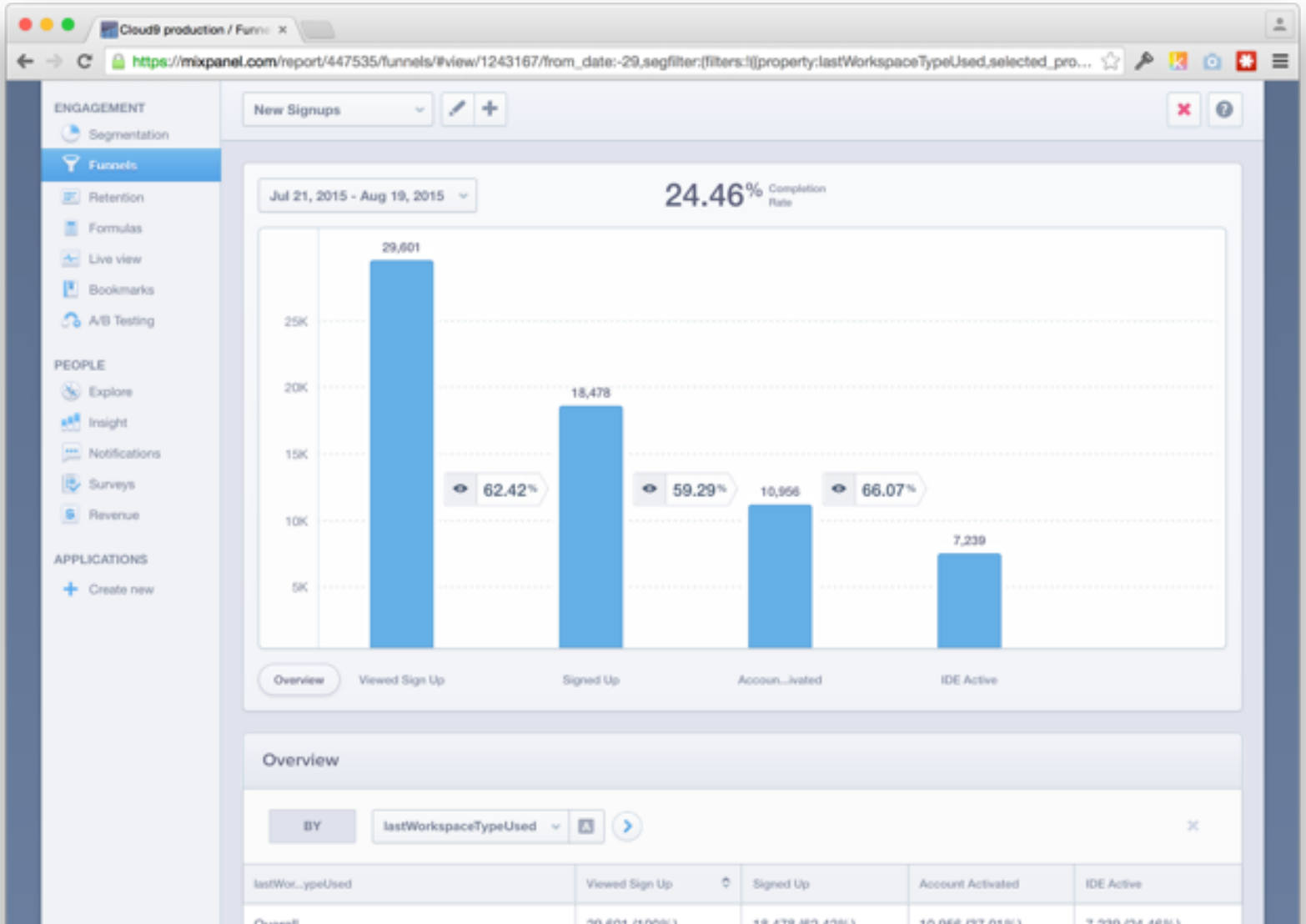
ANALYTICS.GOOGLE.COM



MIXPANEL

FOR USER-EVENT DATA

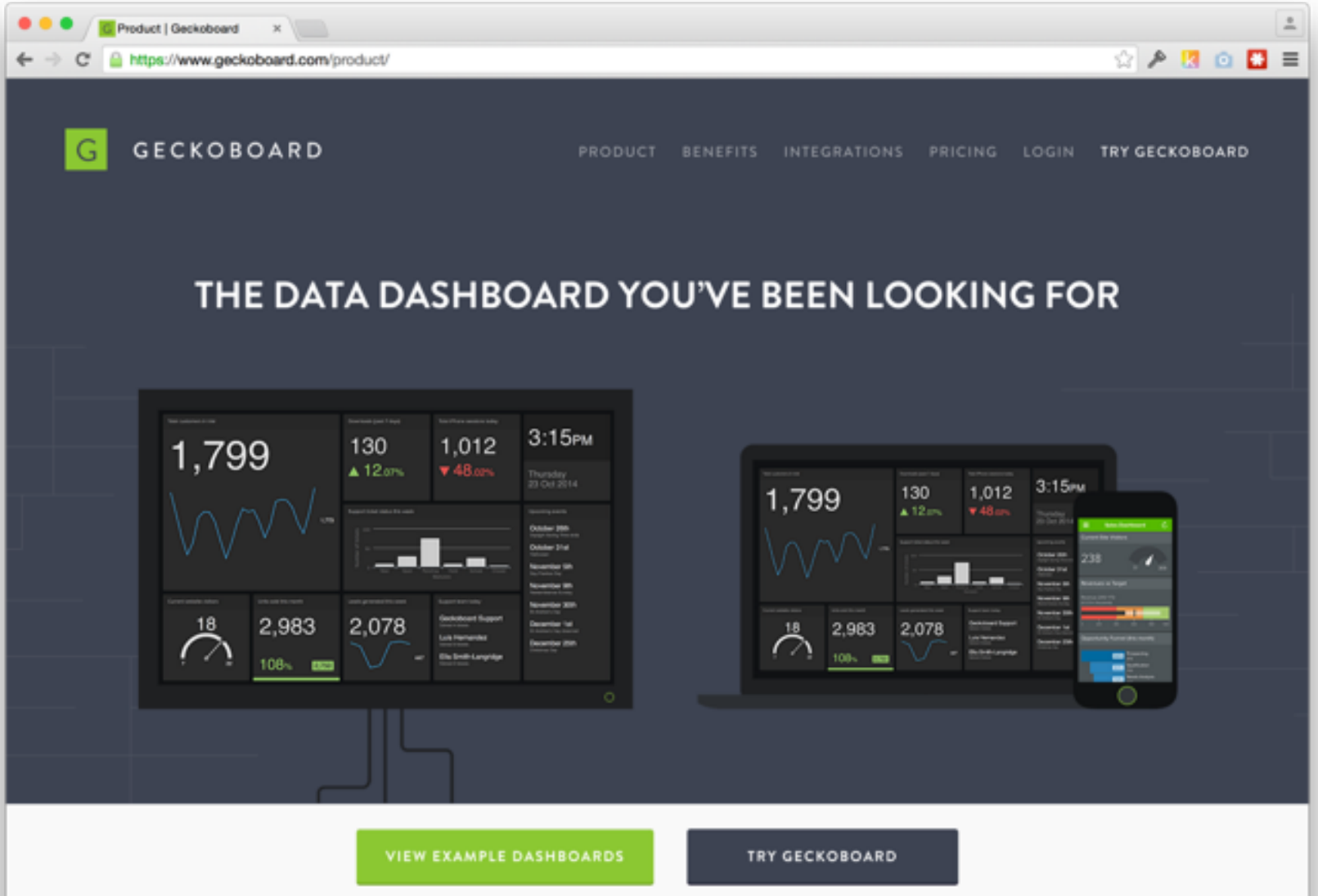
MIXPANEL.COM



GECKOBOARD

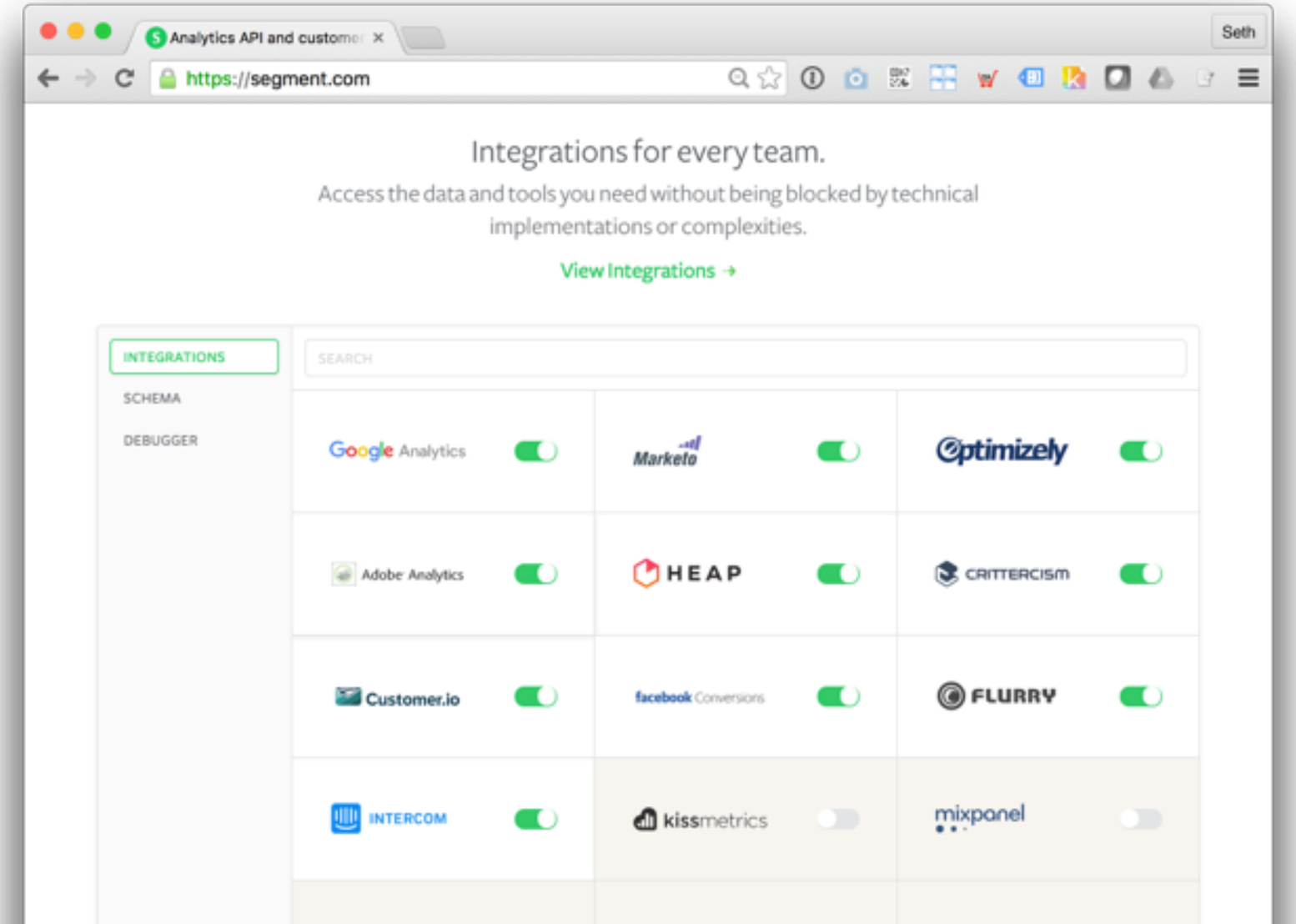
GECKOBOARD.COM

FOR AGGREGATING DATA SOURCES



SEGMENT

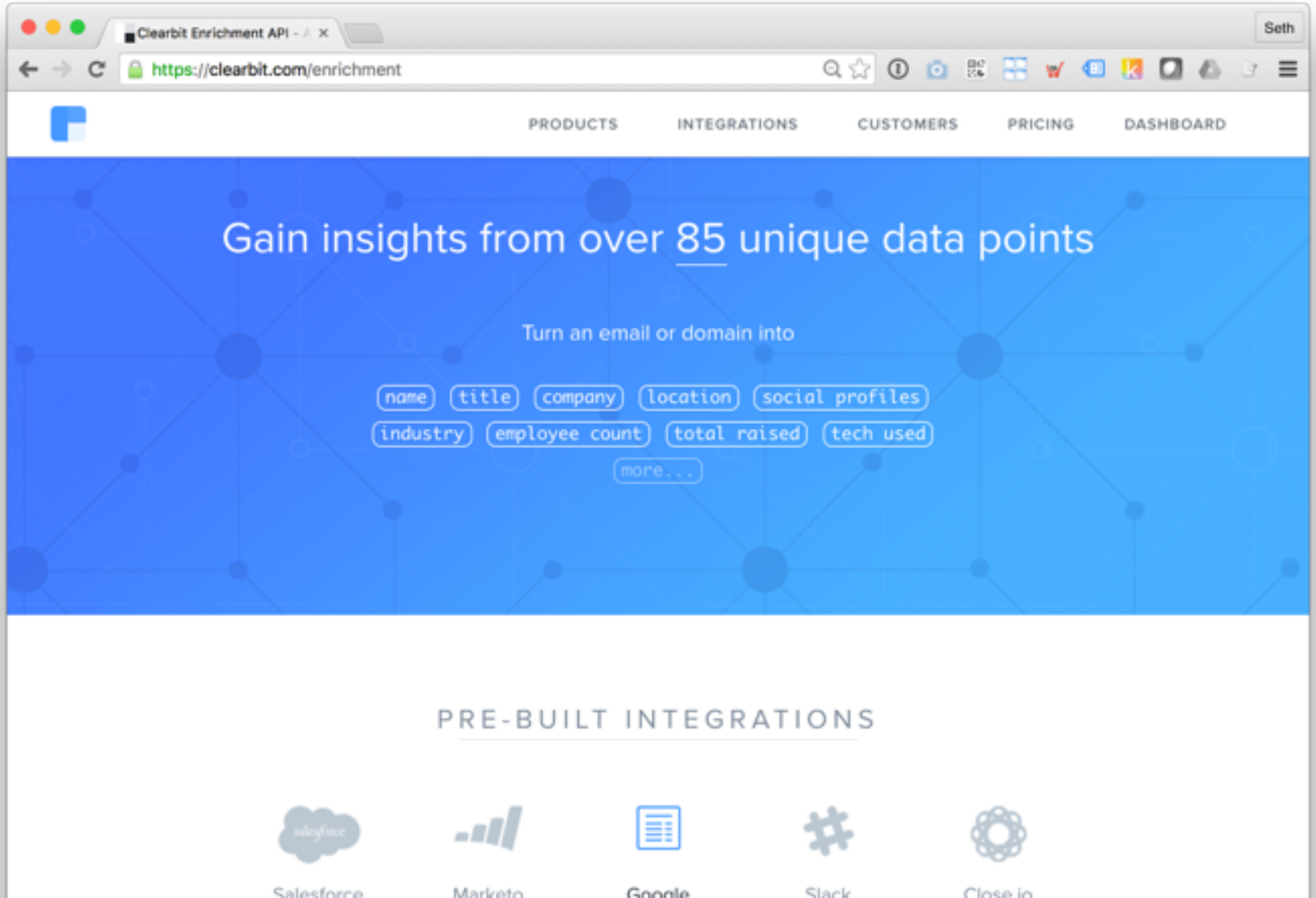
SEGMENT.COM



CLEARBIT

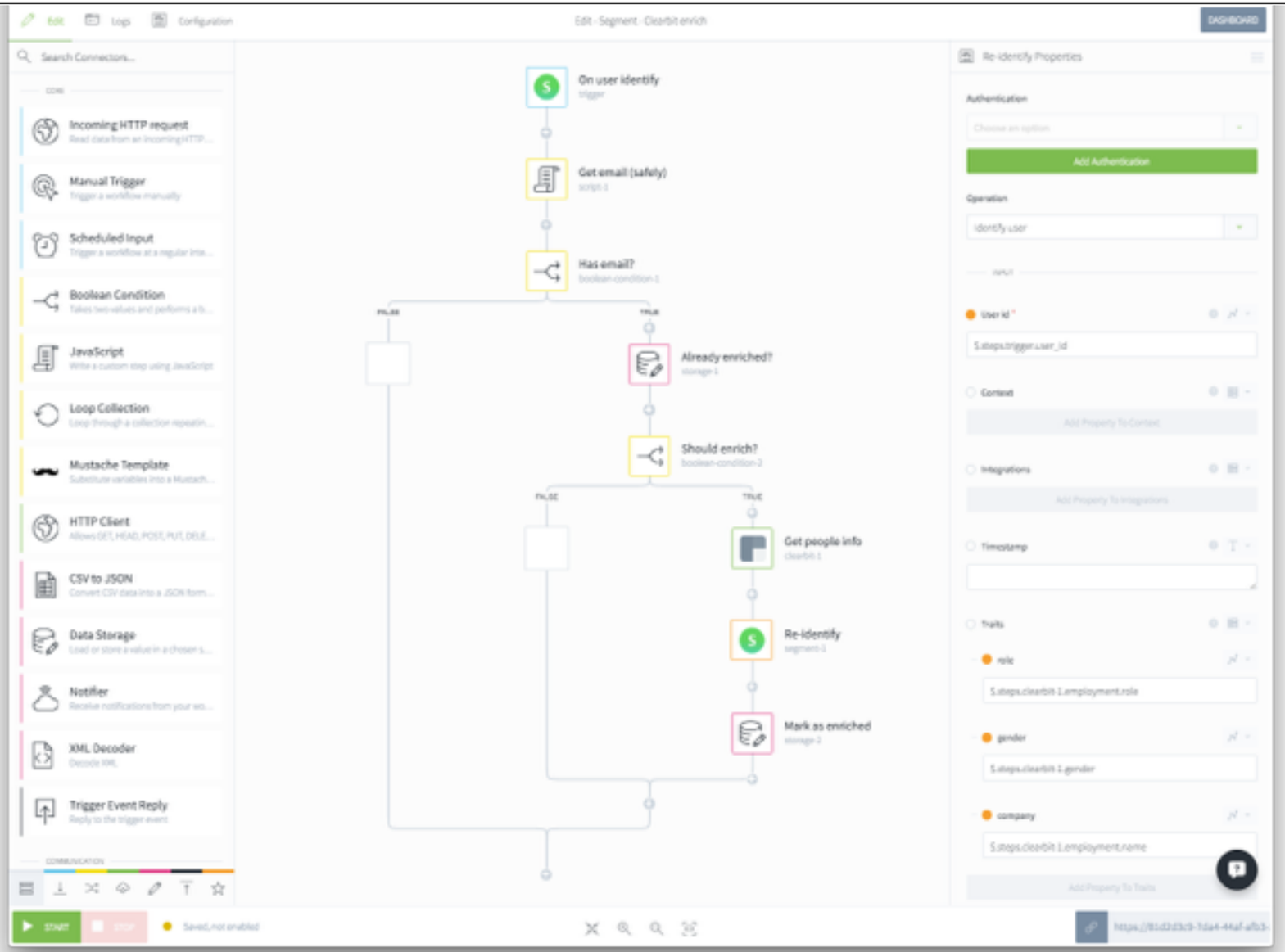
FOR DATA ENRICHMENT

CLEARBIT.COM



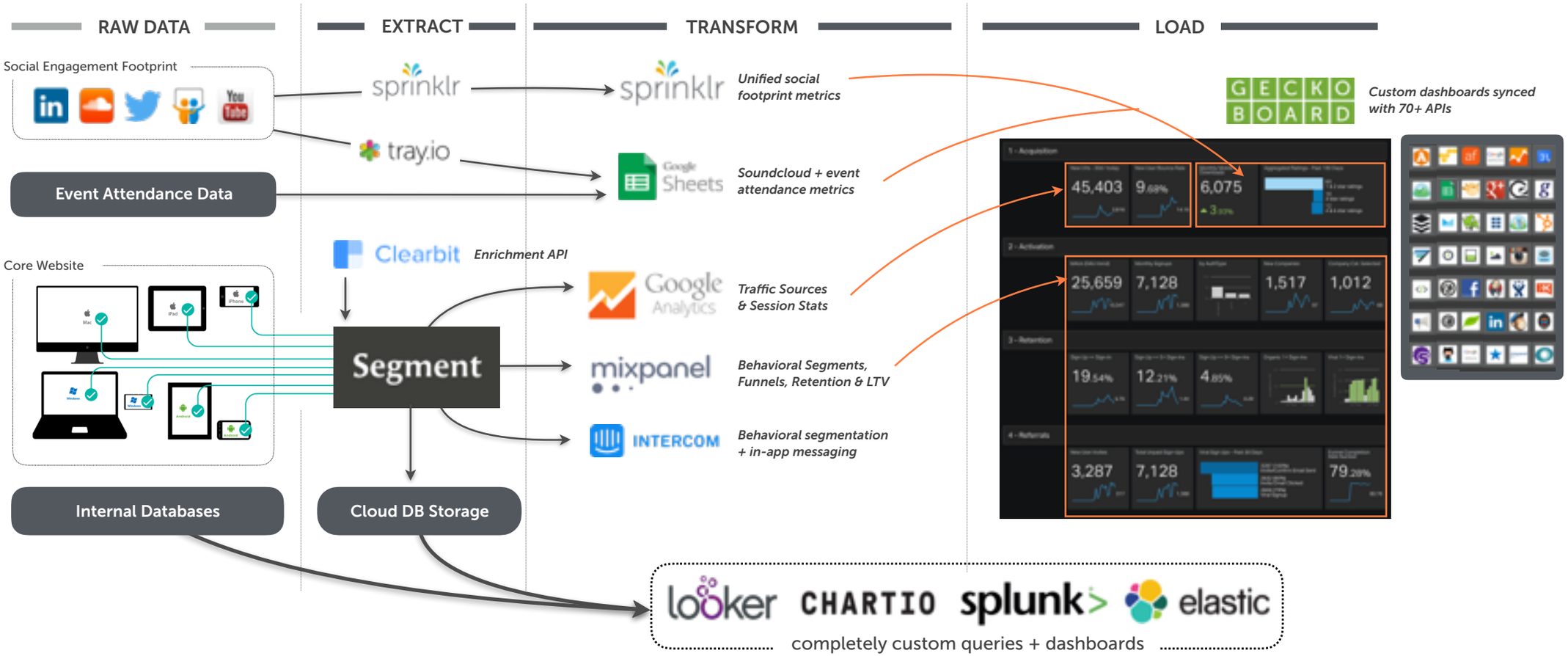
TRAY.IO

COMPLETELY CUSTOM FLOWS



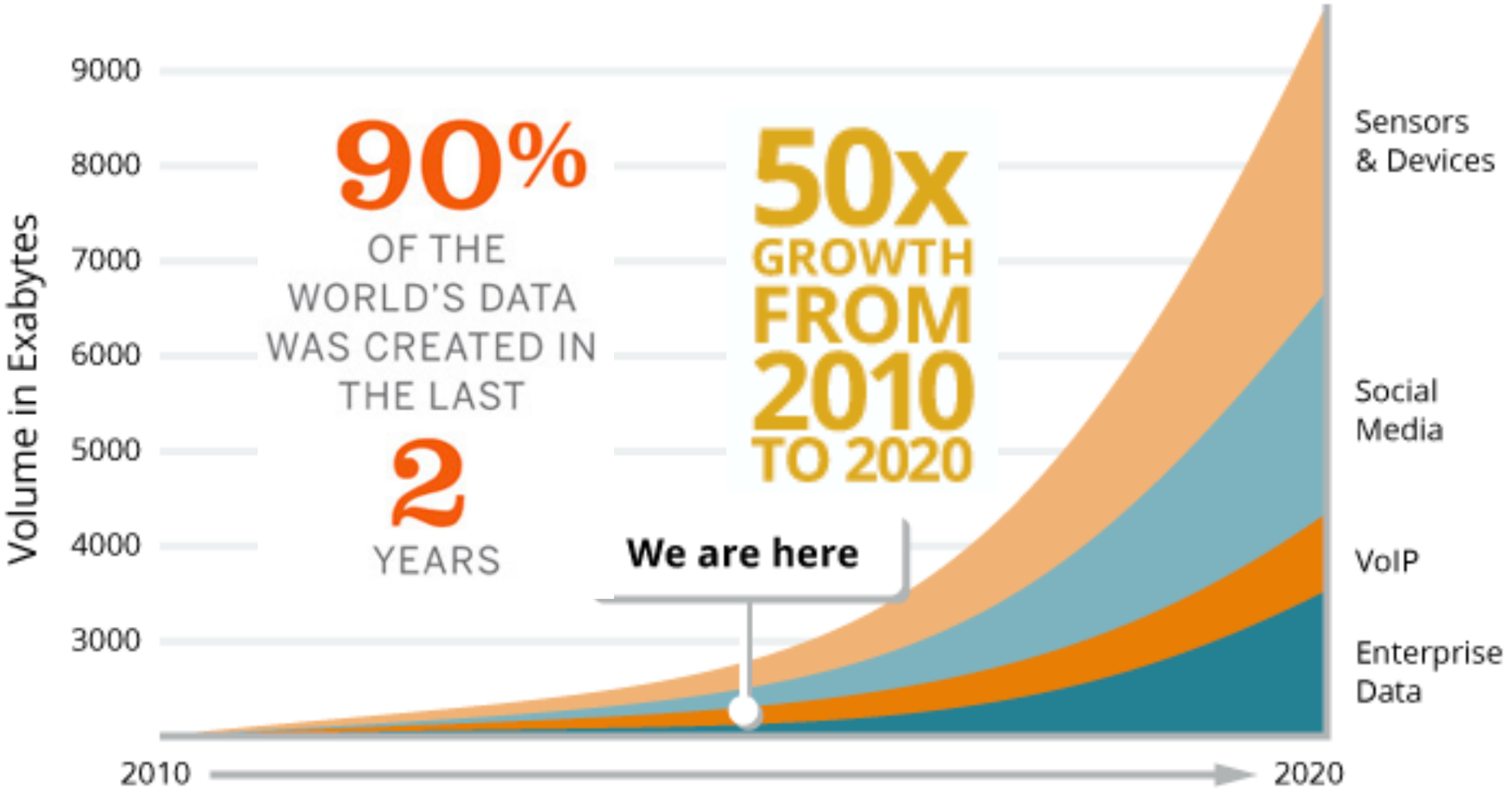
CONNECTING THE DOTS

WITH THE BIG DATA SAAS VALUE CHAIN!

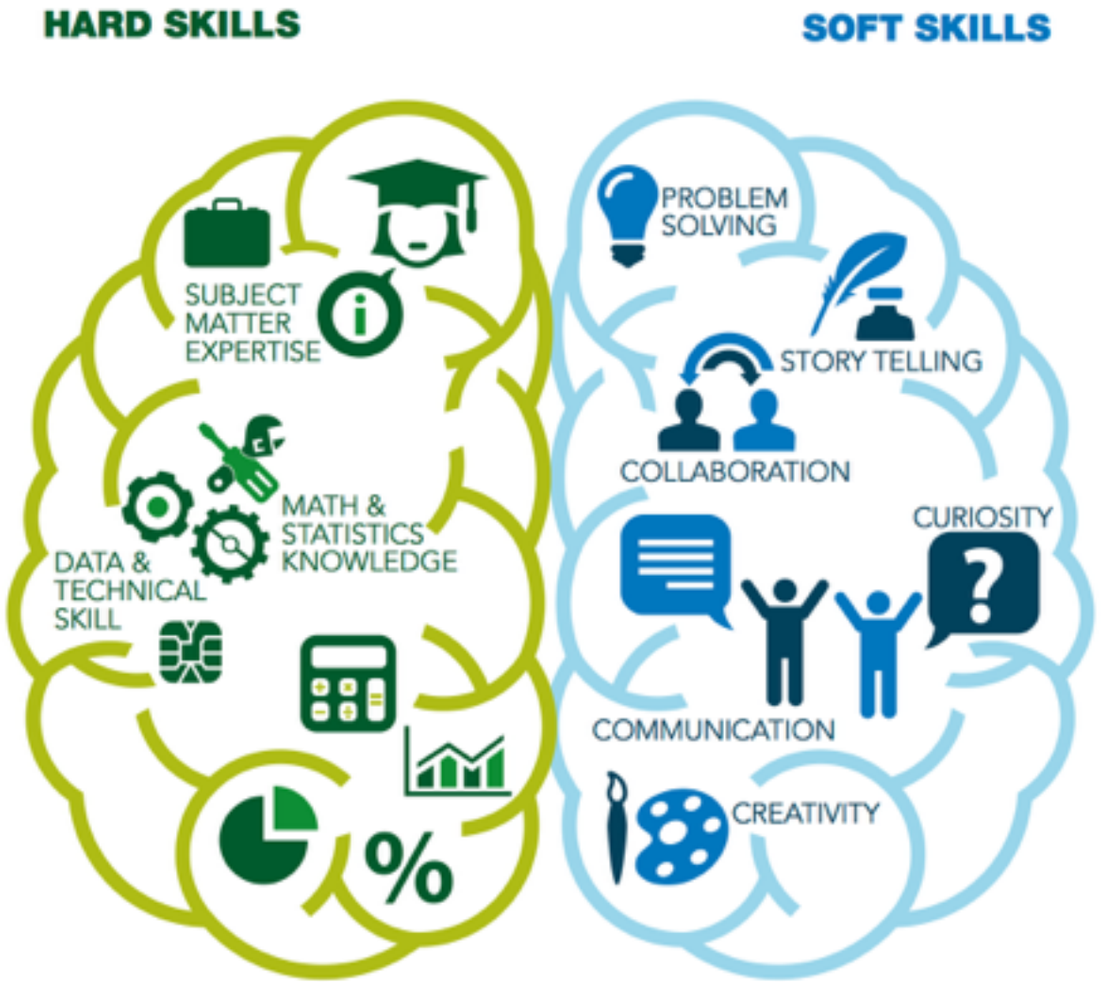


FINAL THOUGHTS

DATA OBESITY!



A NEW TYPE OF KNOWLEDGE WORKER



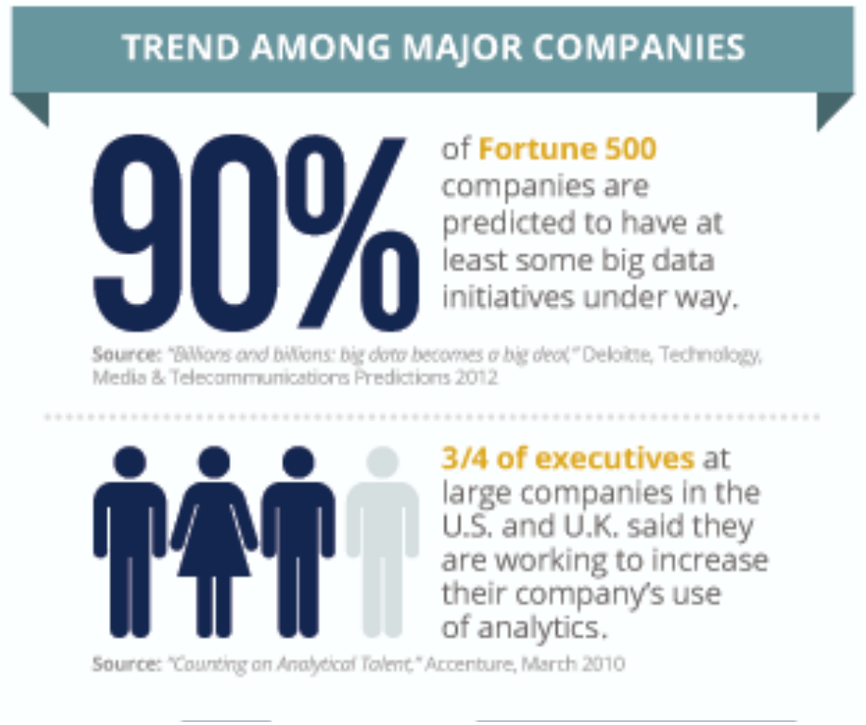
http://www.doclens.com/87922/think-issue-7-2014/

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

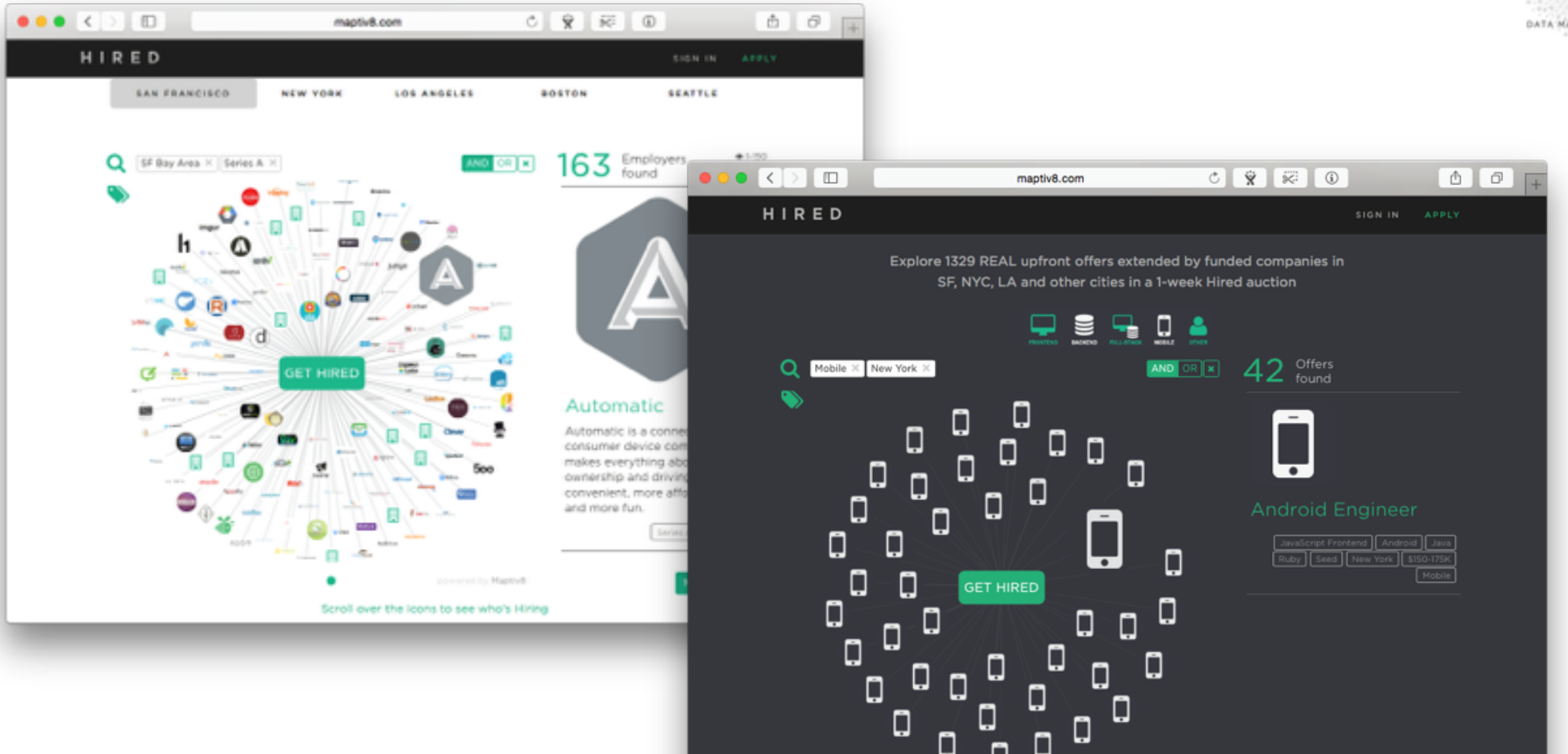


https://studentforce.wordpress.com/2013/09/21/umuc-big-data-revolution-is-here/

DATA AS INTERFACE

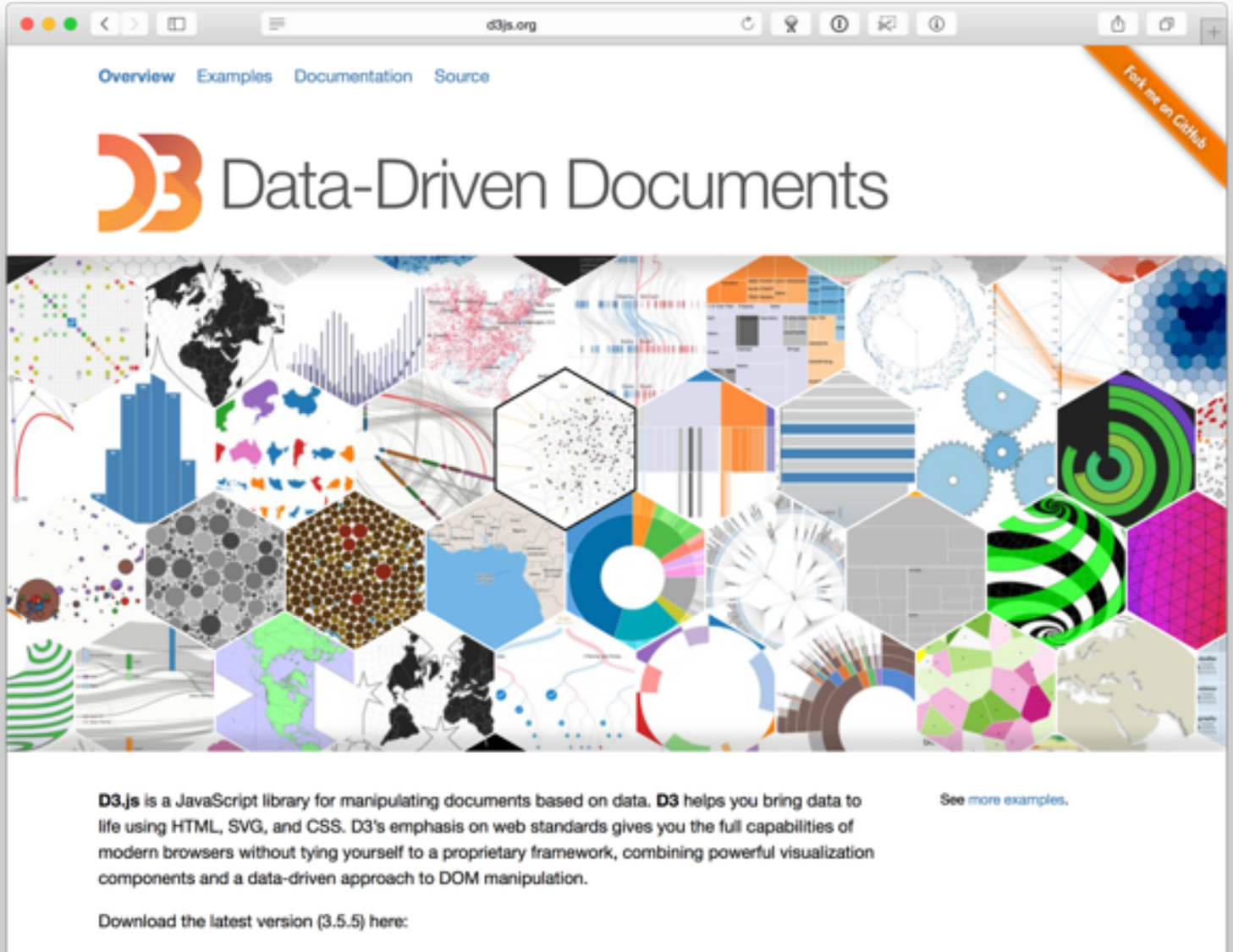
for HIRED

using Maptiv8
DATA MADE BEAUTIFUL



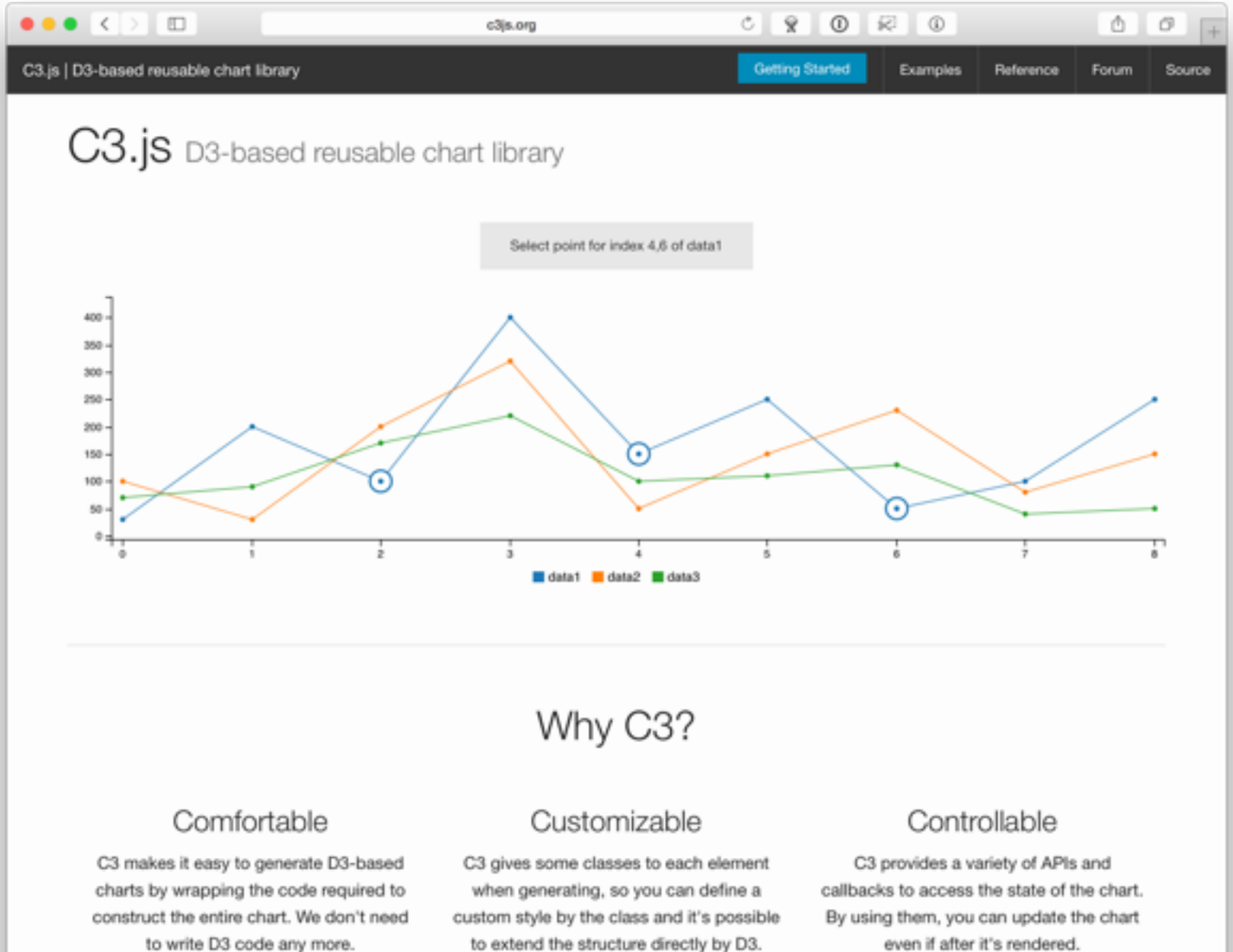
START HERE

D3JS.ORG



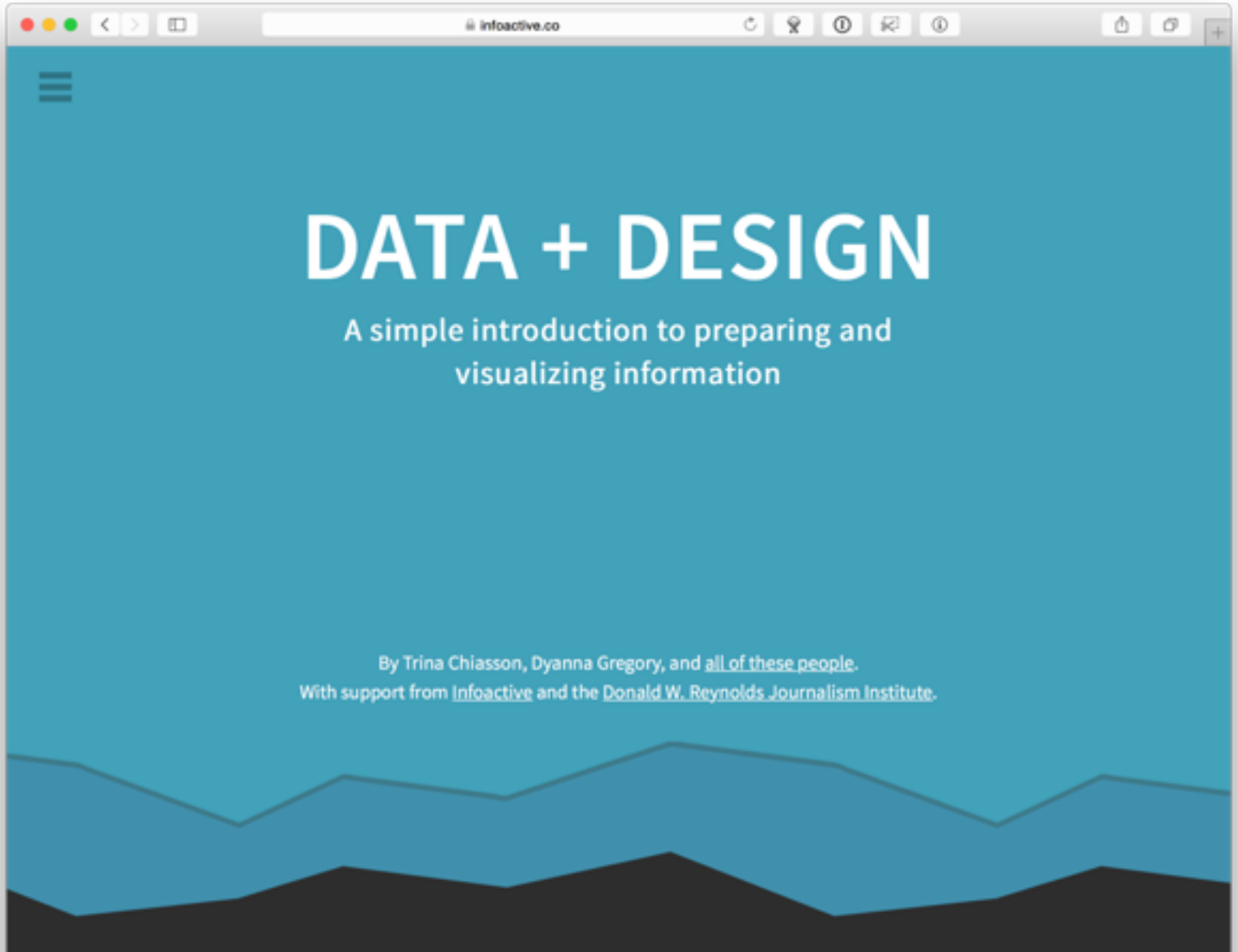
OR HERE

C3JS.ORG



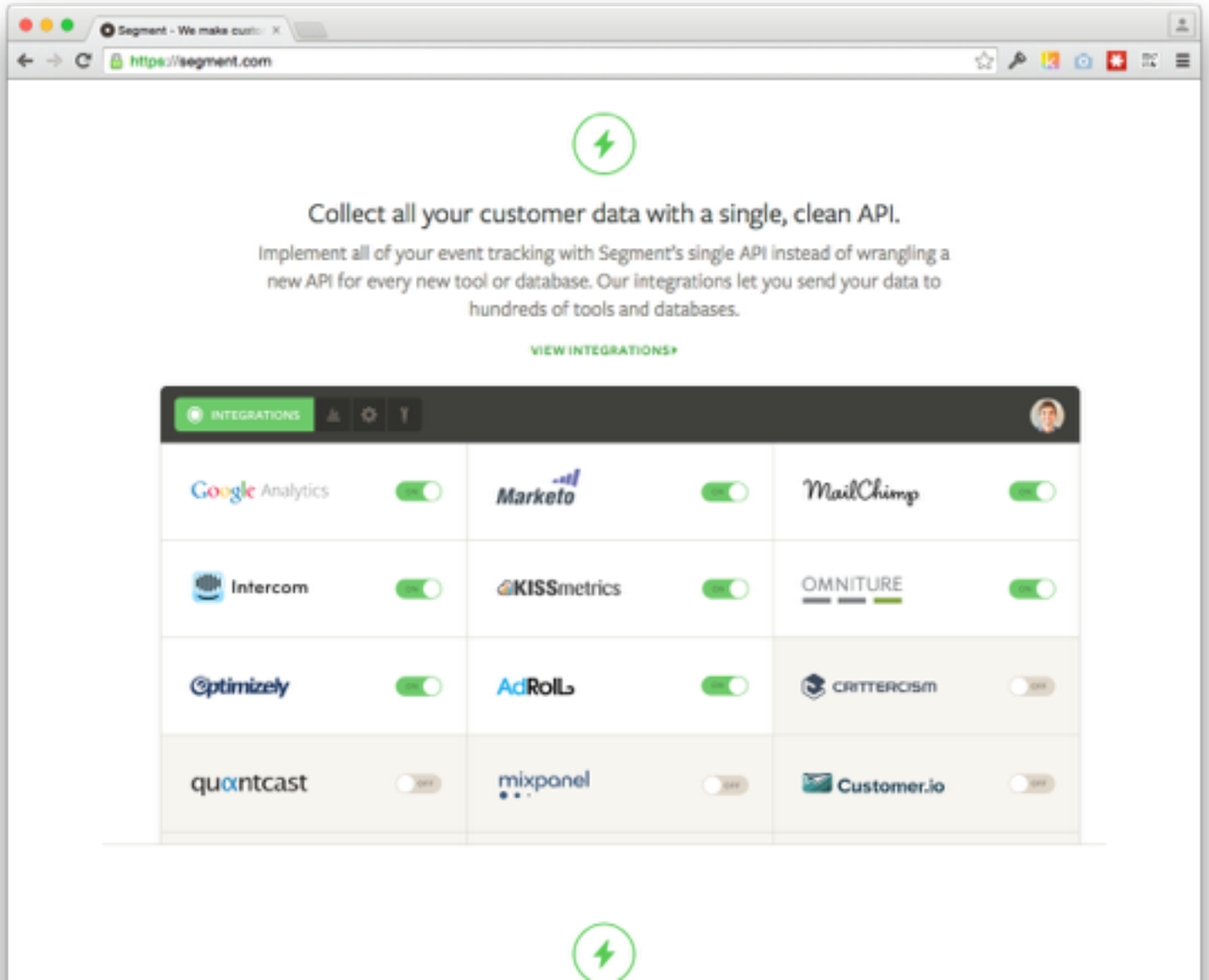
OR HERE

INFOACTIVE.CO/DATA-DESIGN



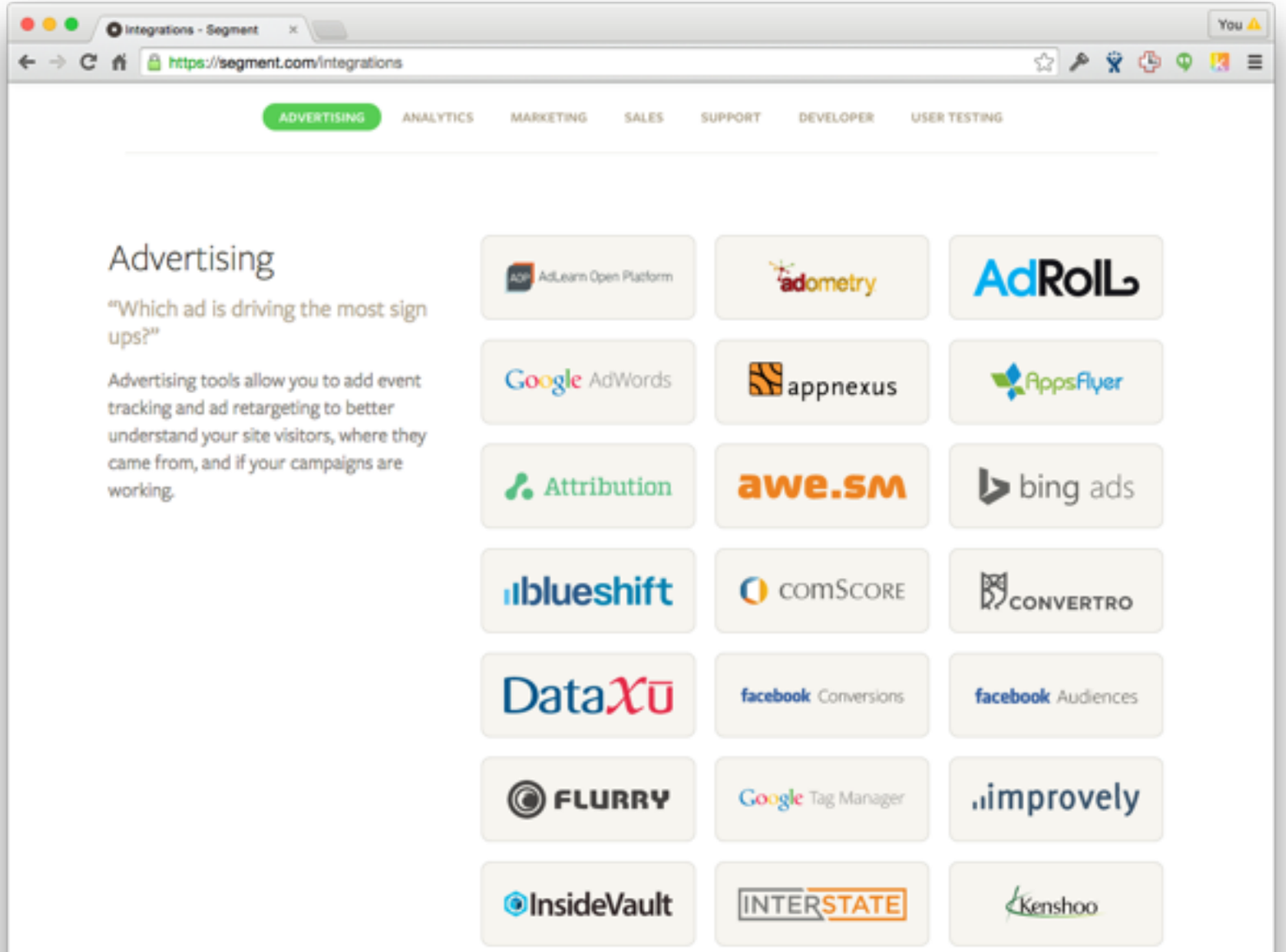
OR HERE

SEGMENT.COM



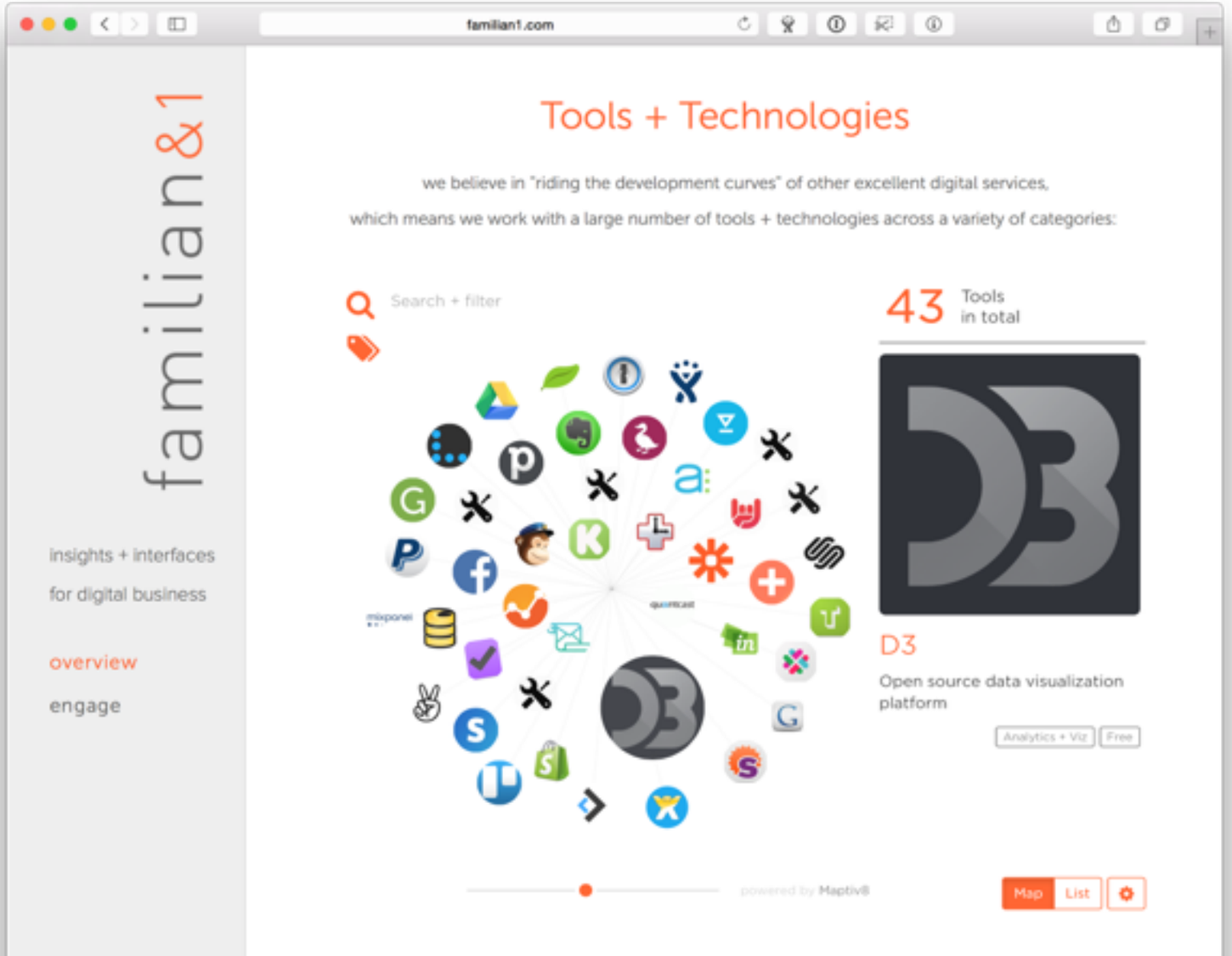
OR HERE

SEGMENT.COM/INTEGRATIONS



OR HERE

FAMILIAN1.COM



OR HERE


GENERAL ASSEMBLY
WORKING WITH BIG DATA
Seth Familian
Founder + Principal, Familian&1
GENERAL ASSEMBLY
VISUAL DESIGN WITH DATA
Seth Familian
Founder + Principal, Familian&1
GENERAL ASSEMBLY
DIGITAL TOOLS FOR BUSINESS
Seth Familian
Founder + Principal, Familian&1
FOLLOW ALONG!
http://bit.ly/ga-dt4b

HAVE FUN!

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

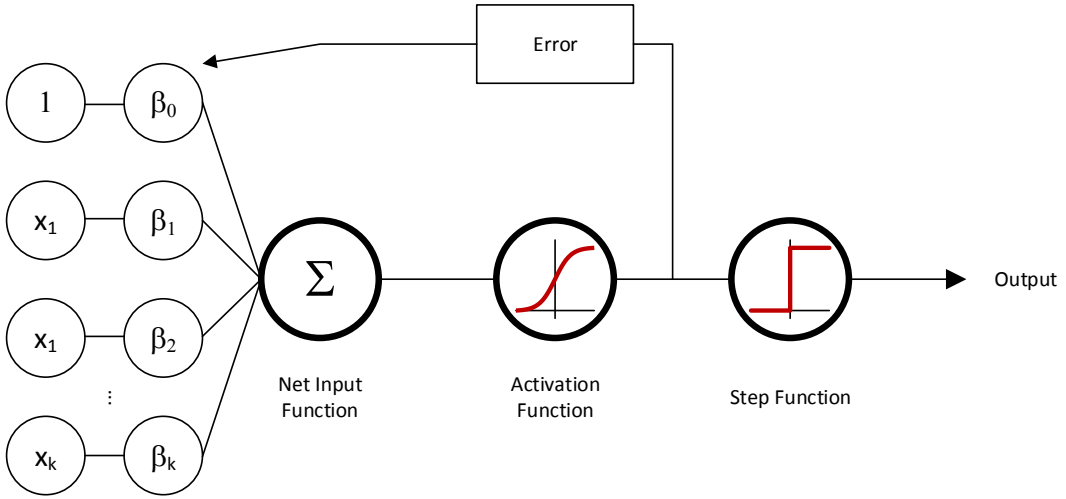
Review

Review

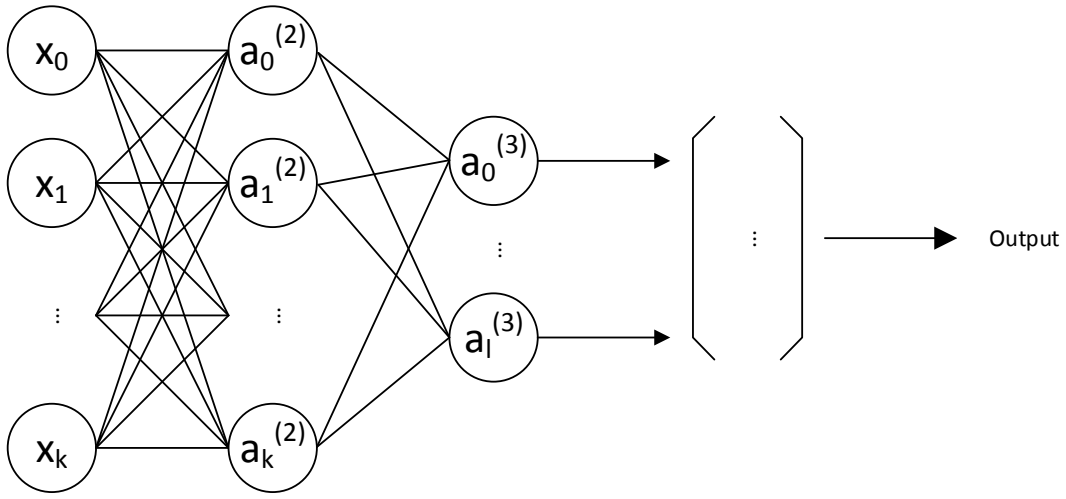
Logistic Regression

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

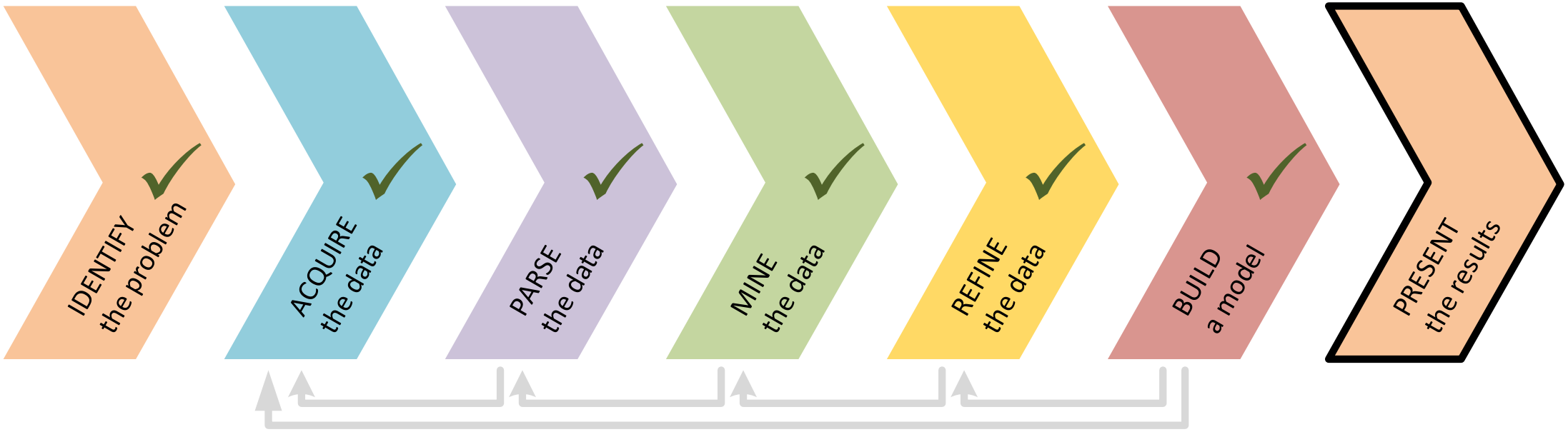
Outline

- 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 *pandas*	Statistics	Exploratory Data Analysis in *pandas*
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

❻ 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

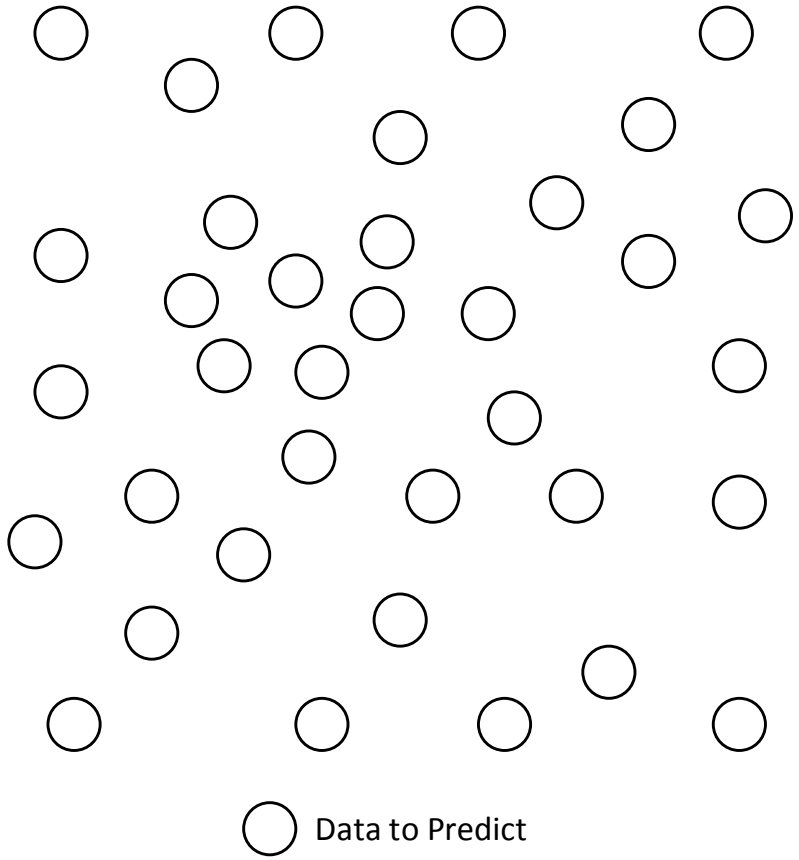
- Accuracy
  - How many observations that we predicted were correct? This is a value we'd want to increase (like R^2)

- 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)

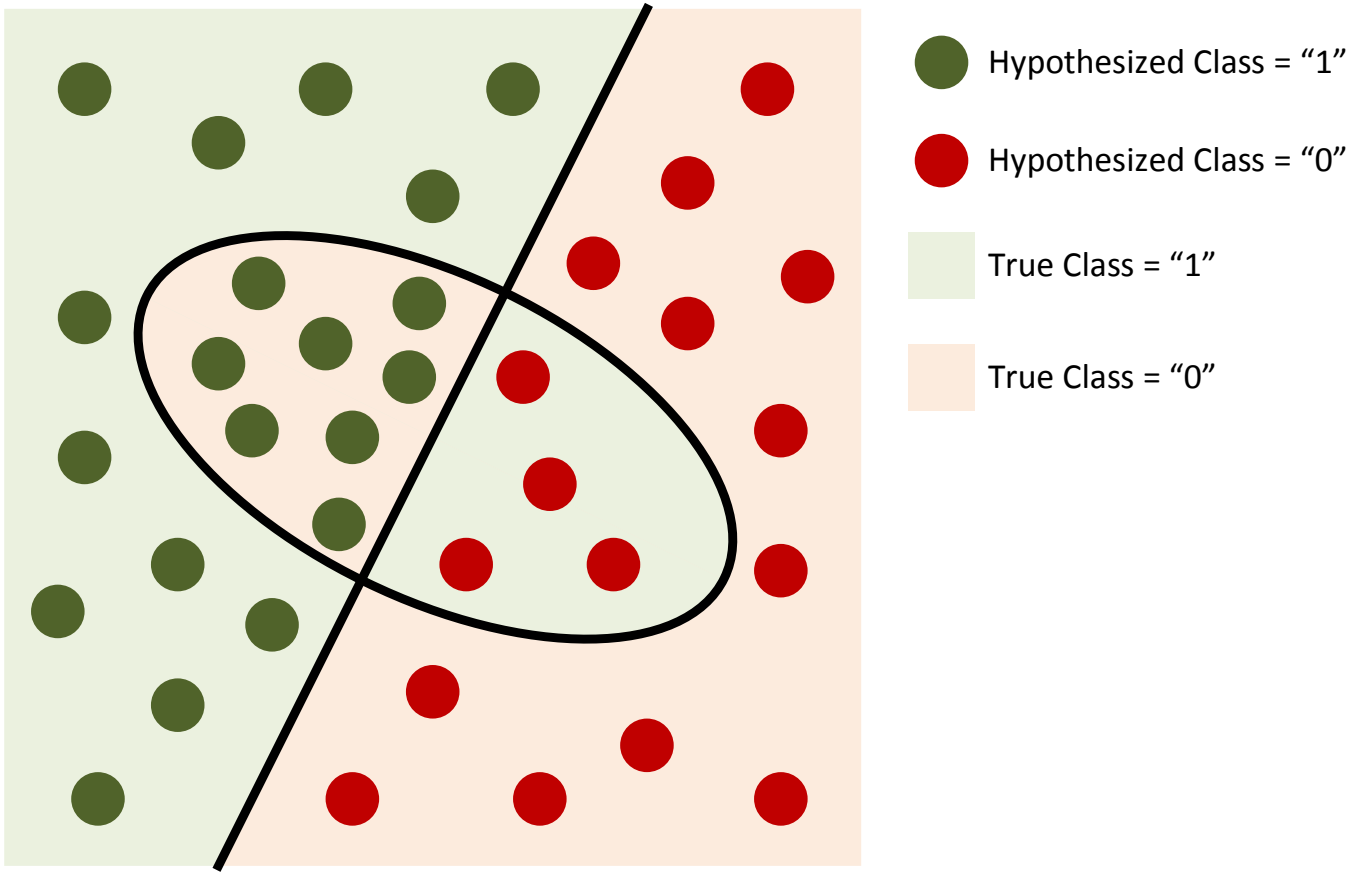
❻ Build a Model

Confusion Matrix

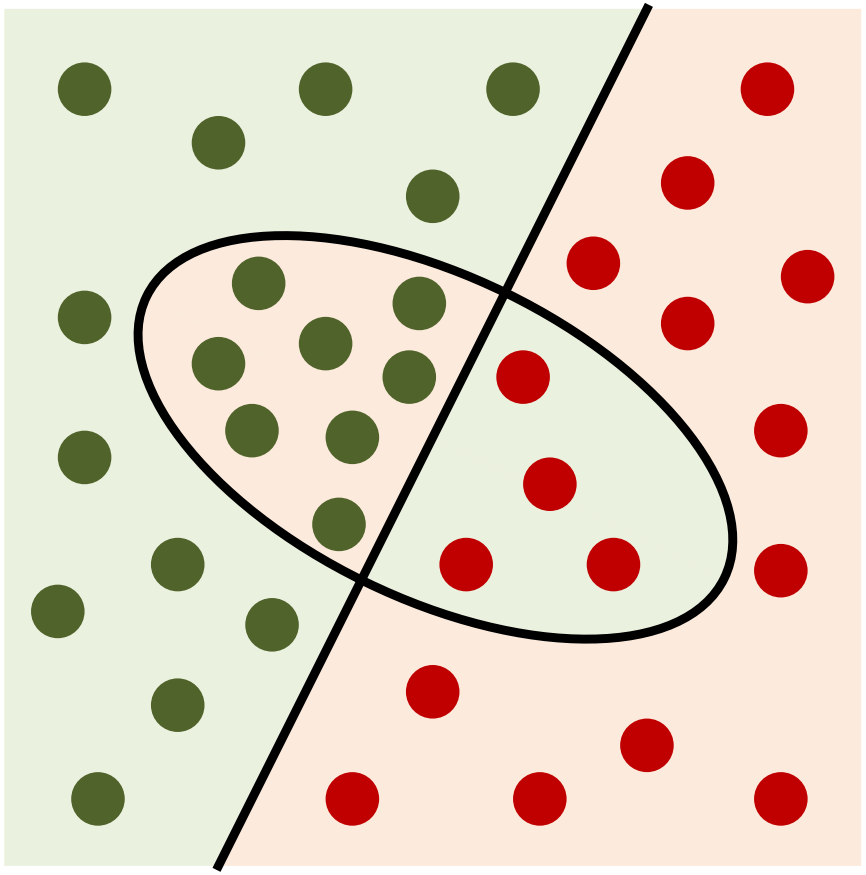
Stepping back | Let’s say we want to classify this data:



Putting hypothesized and true classes together, we get 4 possibilities



We can rearrange these 4 possibilities into a 2x2 table



Hypothesized Class = “1”

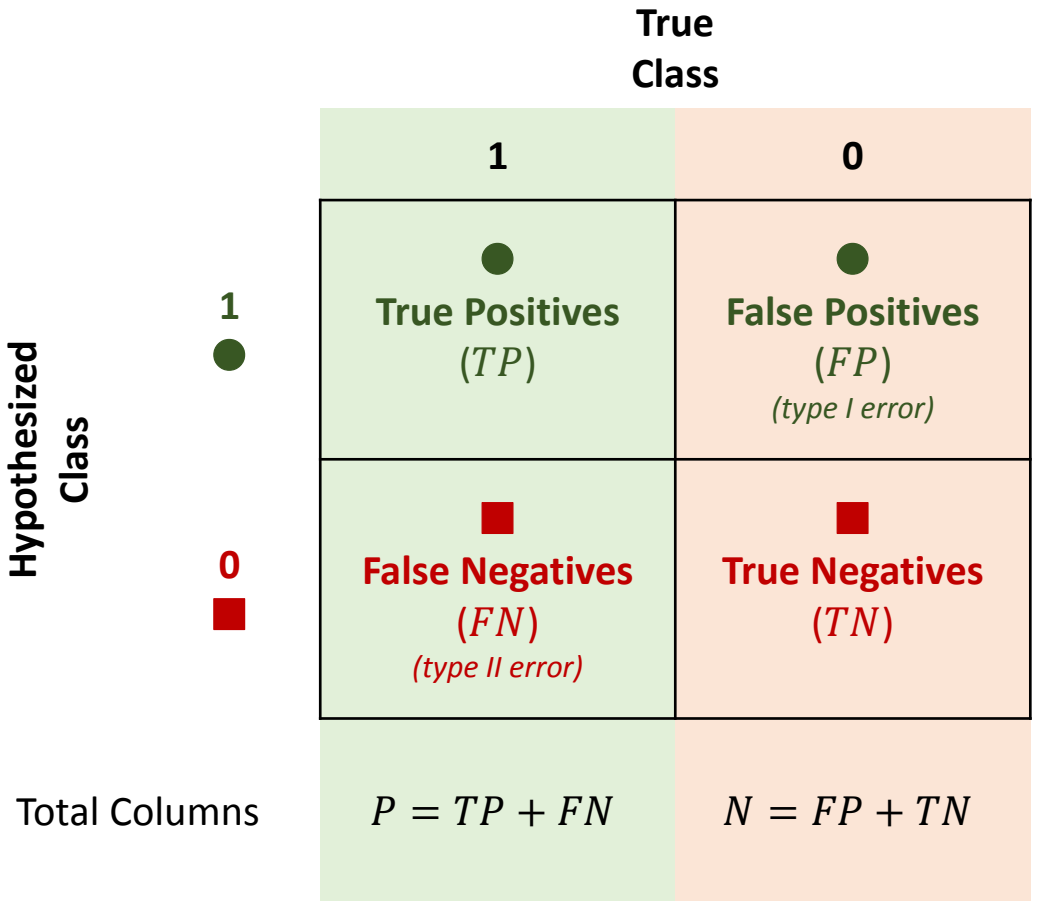
Hypothesized Class = “0”

True Class = “1”

True Class = “0”

True Positive	False Positive
False Negative	True Negative

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



- 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)

❻ 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 ($1 = Cancer$, $0 = Cancer\ free$), 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.)

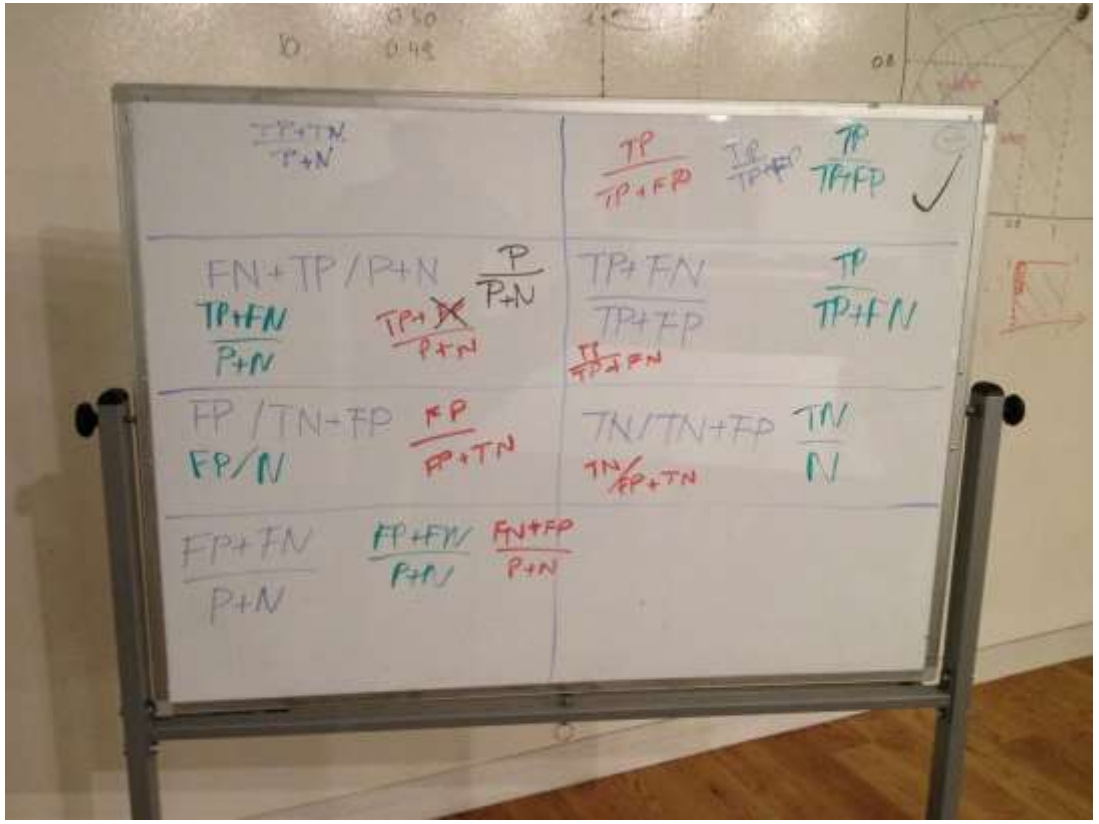
Hypothesized Class	True Class: 1	True Class: 0
1 ●	**True Positives** (TP)	**False Positives** (FP) *(type I error)*
0 ■	**False Negatives** (FN) *(type II error)*	**True Negatives** (TN)
Total Columns	$P = TP + FN$	$N = FP + TN$

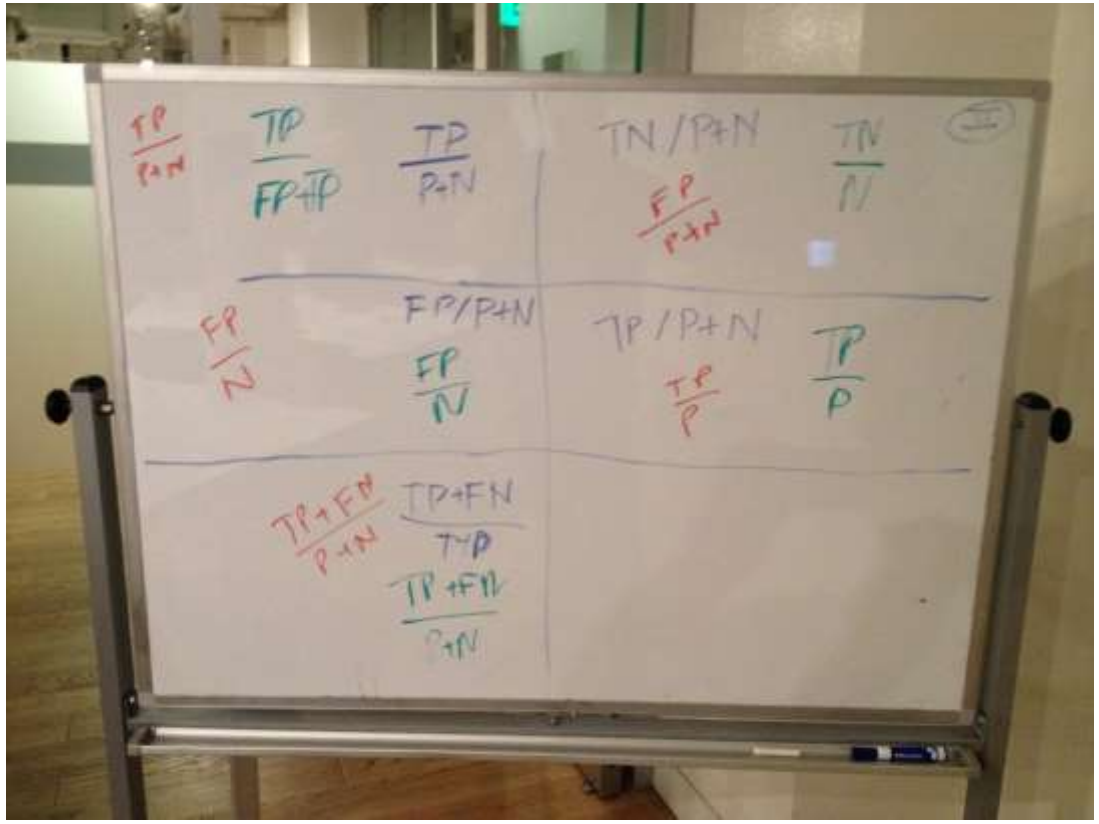
Question: Overall, how often is the classifier correct? Answer: $\frac{TP+TN}{P+N}$	*When the classifier predicts yes, how often is it correct?*
How often does the yes condition actually occur in our sample?	*When it's actually yes, how often does the classifier predict yes?*
When it's actually no, how often does the classifier predict yes?	*When it's actually no, how often does it predict no?*
Overall, how often is the classifier wrong?	

Activity | Interpreting the confusion matrix (cont.)

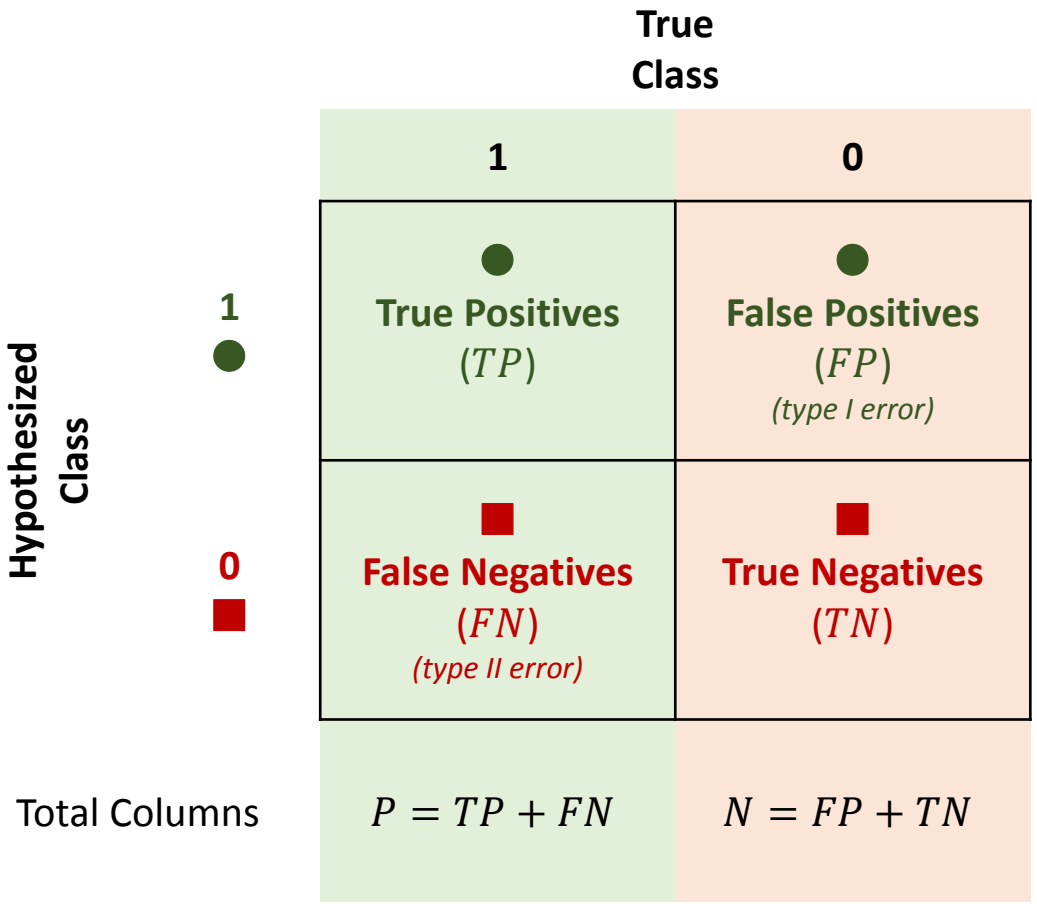
Hypothesized Class	True Class: Has Cancer	True Class: Doesn't have cancer	
Predict Cancer ●	● **True Positives** (TP)	● **False Positives** (FP) *(type I error)*	*How often is it correct when it identify patients with cancer?* / *How often does it correctly identify patients without cancer?*
Predict No Cancer ■	■ **False Negatives** (FN) *(type II error)*	■ **True Negatives** (TN)	*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	$P = TP + FN$	$N = FP + TN$	*How many patients have cancer?*

Activity | Interpreting the confusion matrix (cont.)



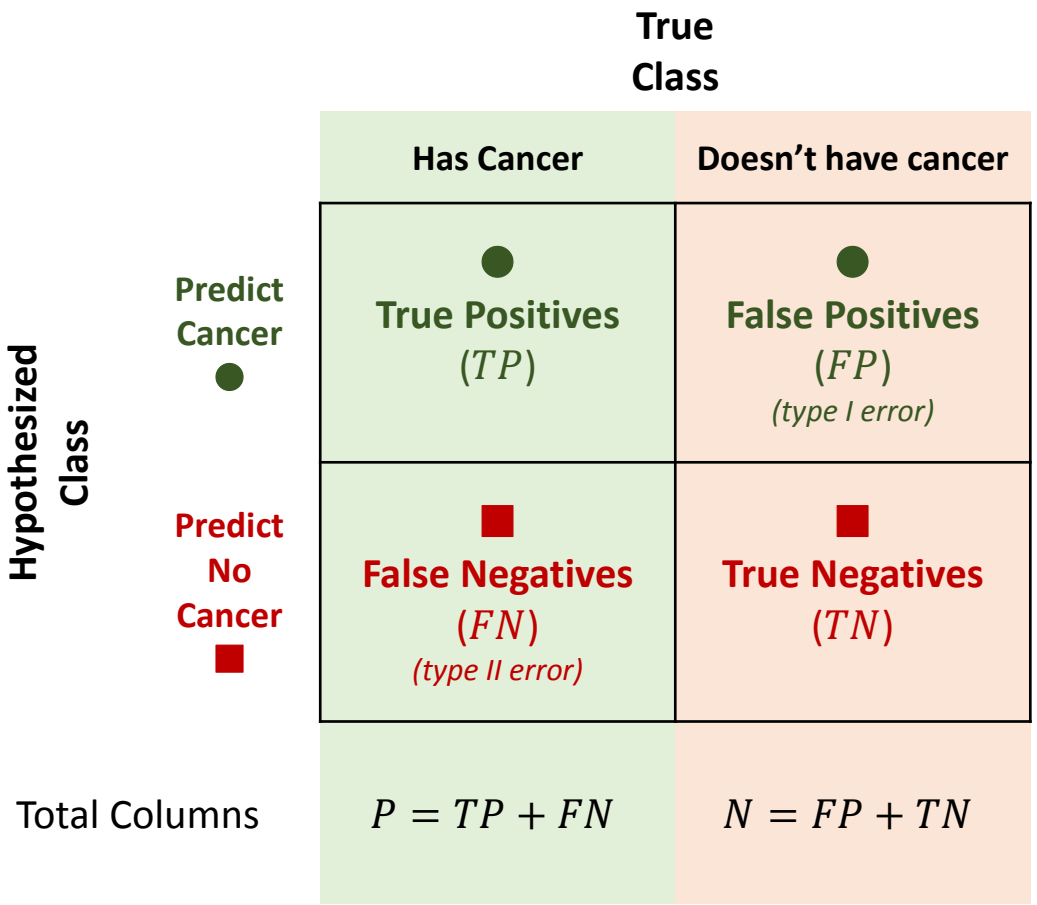


Activity | Interpreting the confusion matrix (cont.)



Question: Overall, how often is the classifier correct? Answer: $\frac{TP+TN}{P+N}$ *(accuracy)*	*When the classifier predicts yes, how often is it correct?* Answer: $\frac{TP}{TP+FP}$ *(precision)*
How often does the yes condition actually occur in our sample? Answer: $\frac{P}{P+N}$ *(prevalence)*	*When it's actually yes, how often does the classifier predict yes?* Answer: $\frac{TP}{P}$ *(TPR, sensitivity, recall)*
When it's actually no, how often does the classifier predict yes? Answer: $\frac{FP}{N}$ *(FPR, fall-out)*	*When it's actually no, how often does it predict no?* 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.)

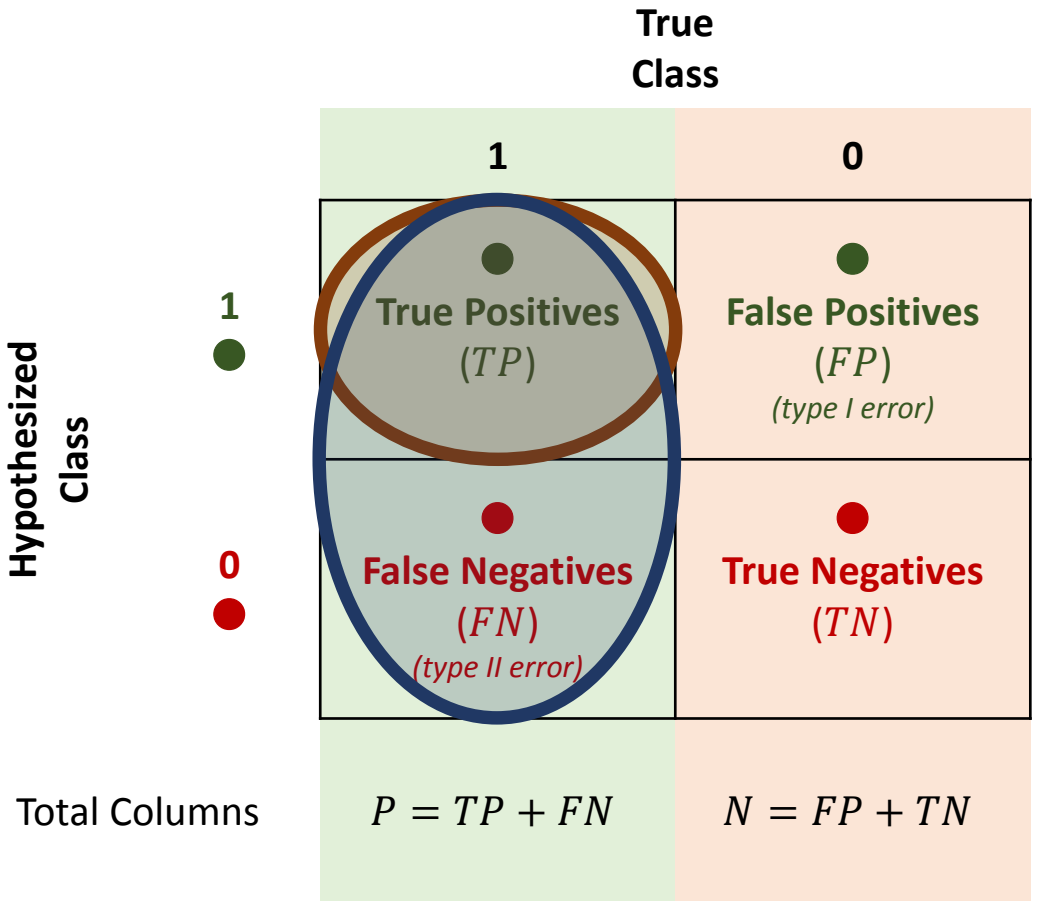


How often is it correct when it identify patients with cancer? Answer: $\frac{TP}{TP+FP}$ *(precision)*	*How often does it correctly identify patients without cancer?* Answer: $\frac{TN}{N}$ *(specificity)*
How often does it trigger a "false alarm" by saying a patient has cancer when they actually don't? Answer: $\frac{FP}{N}$ *(FPR, fall-out)*	*How often does it correctly identify patients with cancer?* Answer: $\frac{TP}{P}$ *(TPR, sensitivity, recall)*
How many patients have cancer? Answer: $\frac{P}{P+N}$ *(prevalence)*	

❻ Build a Model

True and False Positive Rates, ROC, and AUC

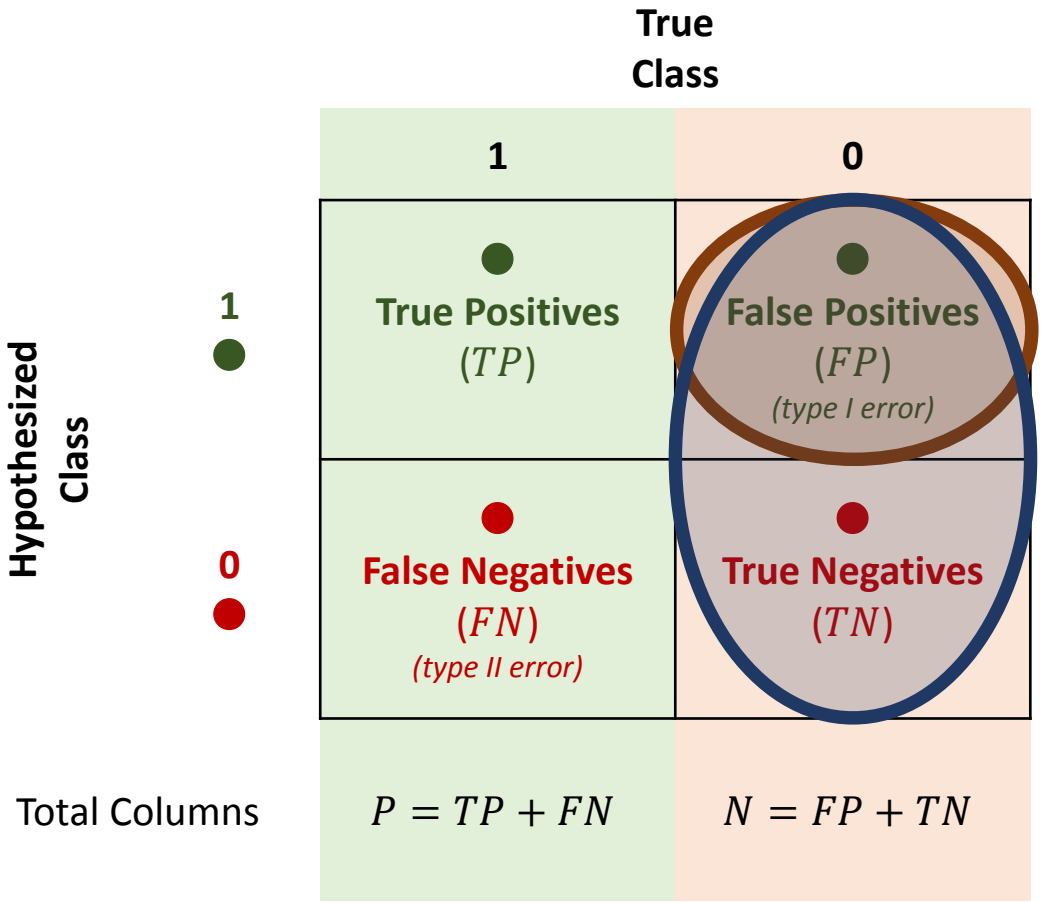
True Positive Rate, $TPR = \frac{TP}{P}$



- 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

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

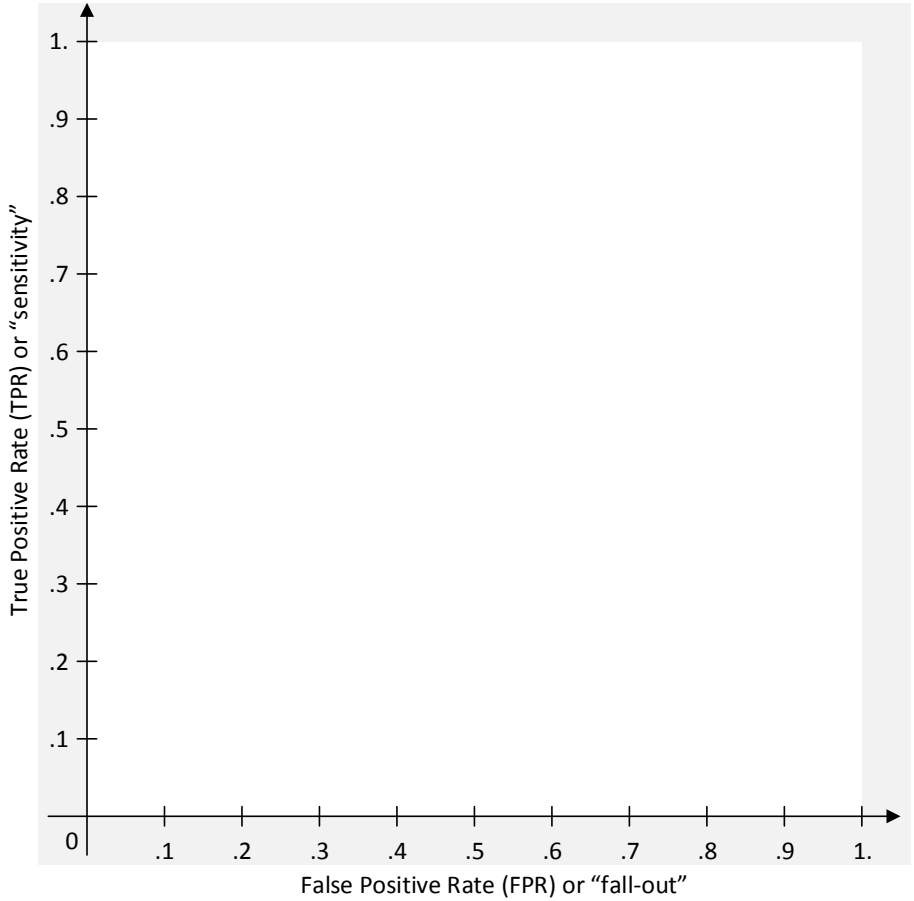
- We can split up the accuracy of each label by using true positive and false positive rates. Using them, we can get a much clearer picture of where predictions begin to fall apart
- A good classifier would have a true positive rate approaching 1, and a false positive rate approaching 0. In a binary 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

❻ Build a Model

ROC (receiver operating characteristic or relative operating characteristic) curve

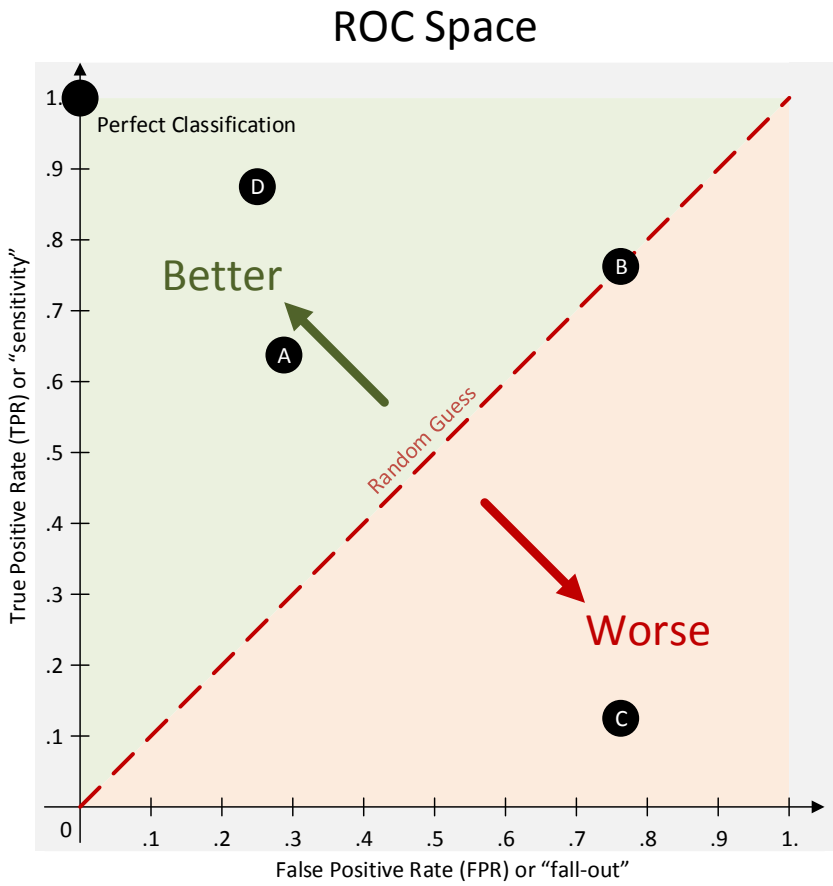
ROC (receiver operating characteristic) curve (a.k.a., relative operating characteristic curve)

- 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



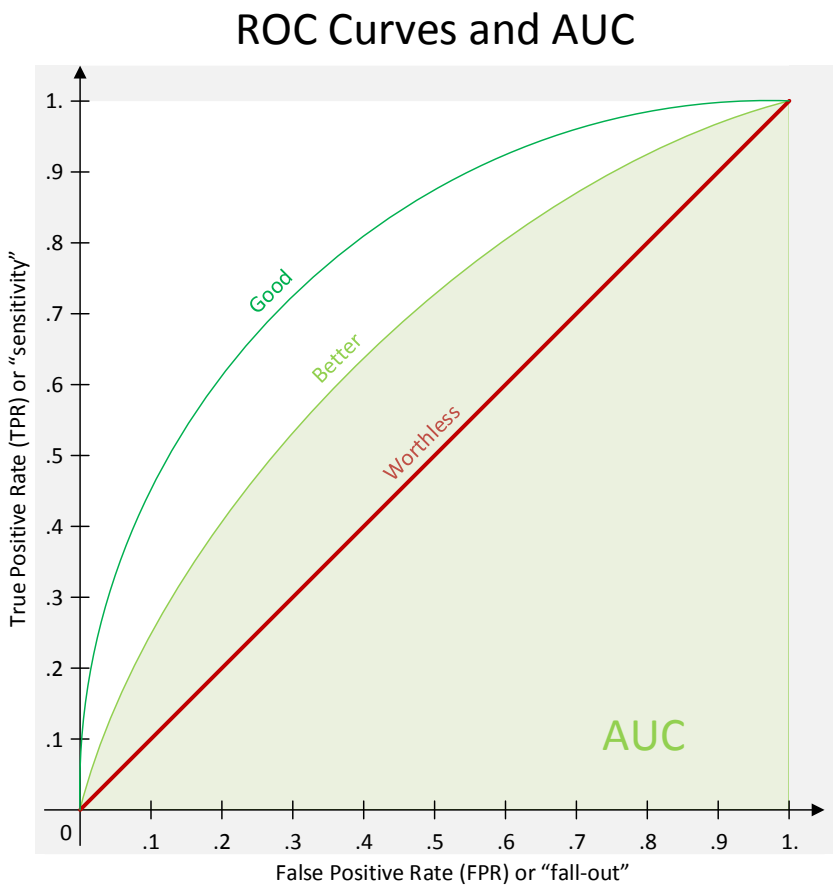
The ROC space demonstrates several things:

- 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

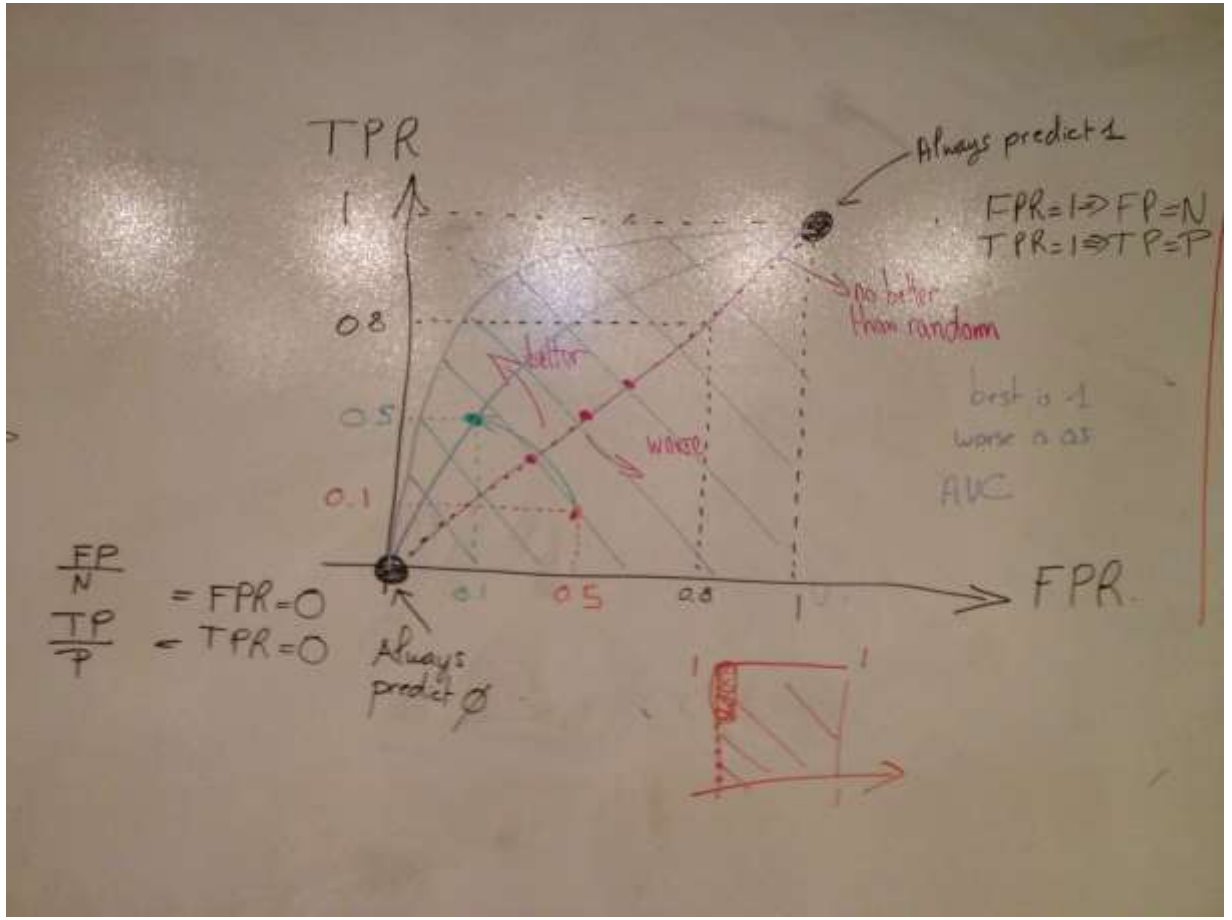


The ROC curves demonstrate several things:

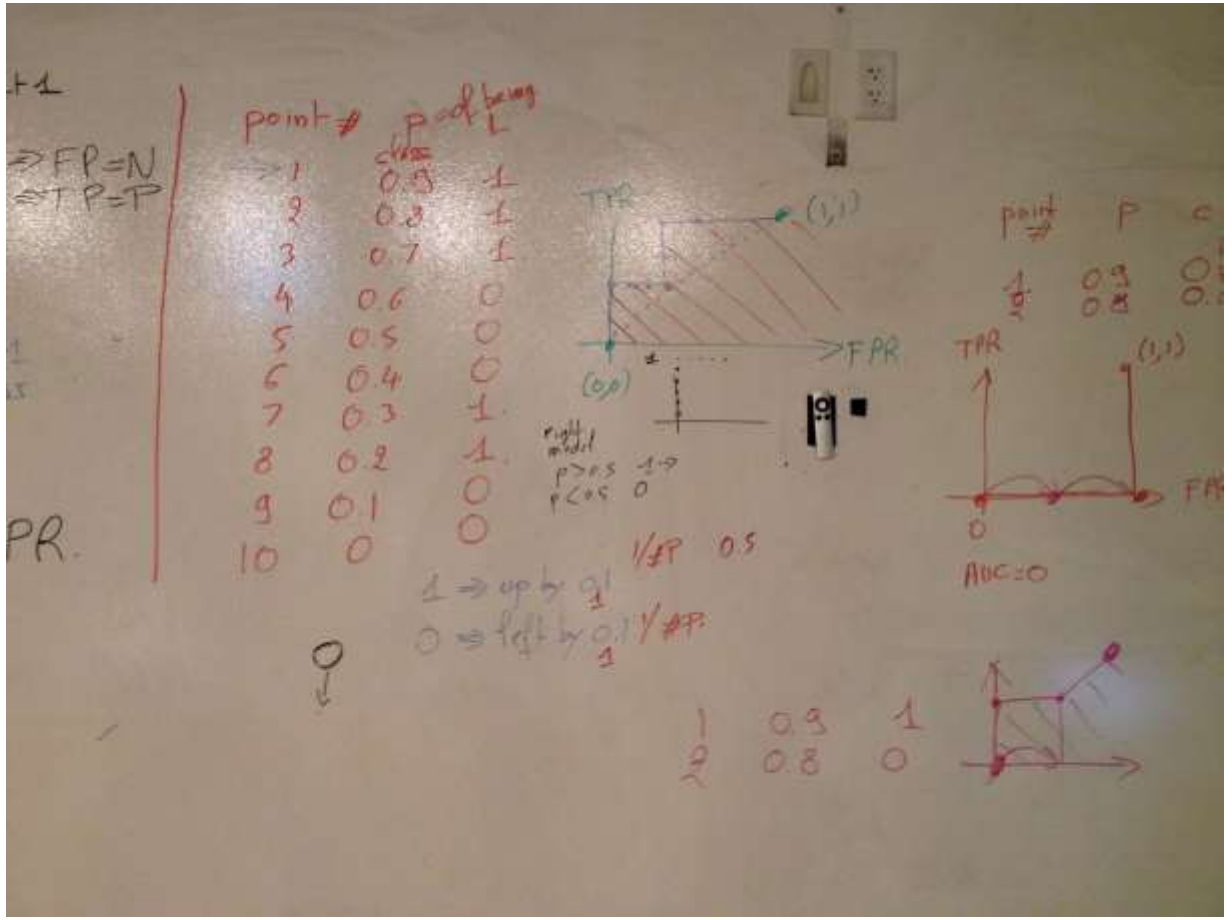
- The area under the curve (AUC) is a measure of classifier accuracy
  - The closer the **curve** follows the left-hand border and then the top border of the ROC space, the more accurate the classifier is
  - The closer the **curve** comes to the 45-degree diagonal of the ROC space, the less accurate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



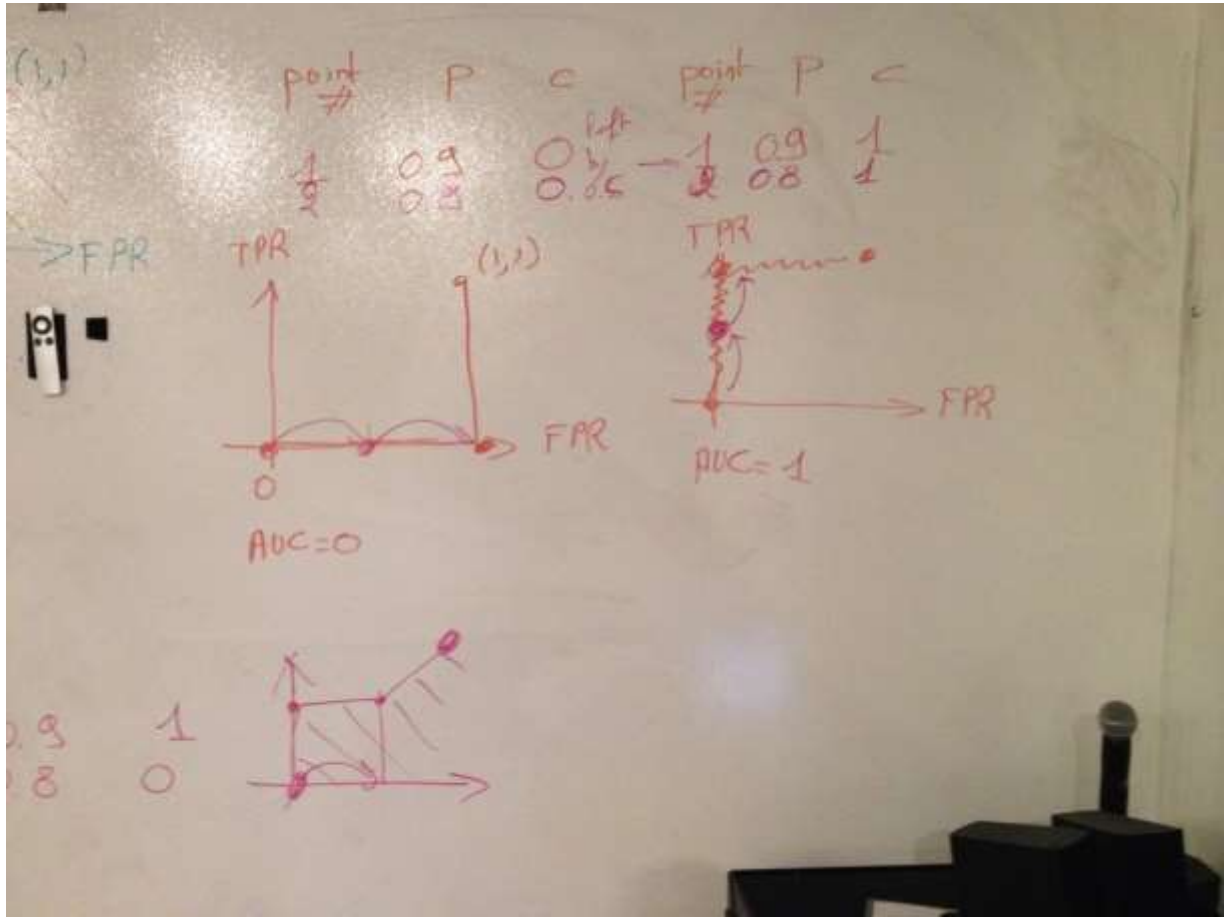
ROC and AUC


TPR
1
0.8
0.5
0.1
Always predict 1
FPR=1⇒FP=N
TPR=1⇒TP=P
no better than random
better
worse
best is 1
worse is 0.5
AUC
FP/N = FPR=0
TP/P = TPR=0
Always predict ∅
0.1
0.5
0.8
1
FPR

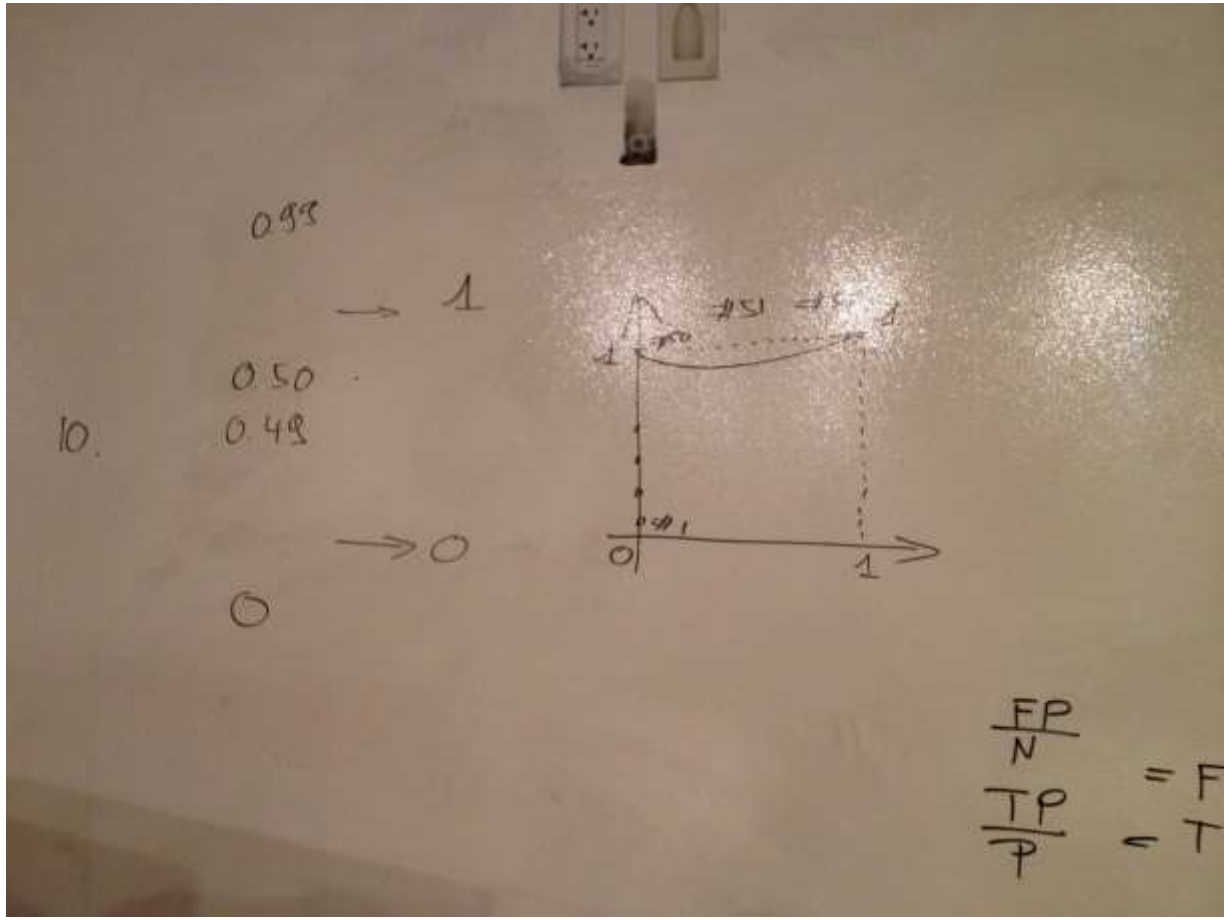
ROC and AUC (cont.)



ROC and AUC (cont.)



ROC and AUC (cont.)



❻ Build a Model

Codealong – Part A

ROC/AUC

❼ 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

❼ 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	What	How
‣ Who is my target audience for the visual?	‣ What do they already know about this project? What do they need to know?	‣ 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

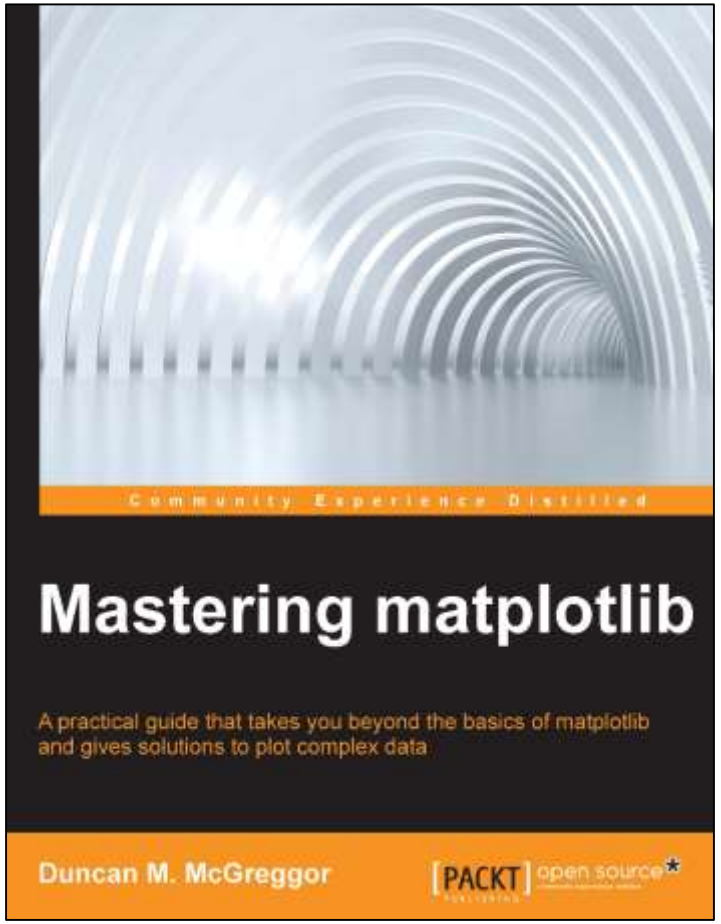
- Plots like this will also be useful when talking to your peers – other data scientists or analysts who are familiar with your project and interested in the progress you've made

❼ 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 ❺ Refine the Data and ❻ 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