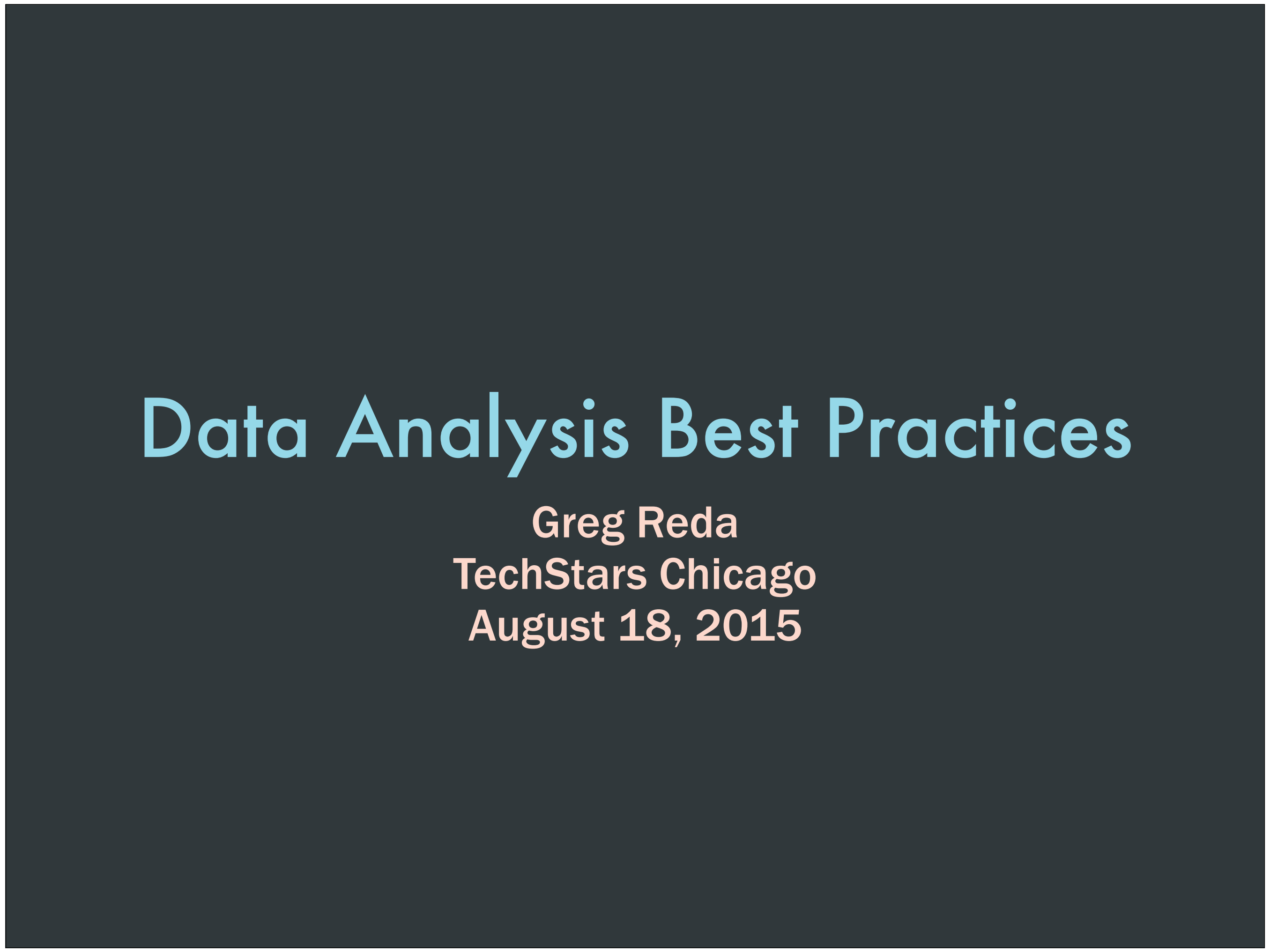


Data Analysis Best Practices

Greg Reda
TechStars Chicago
August 18, 2015

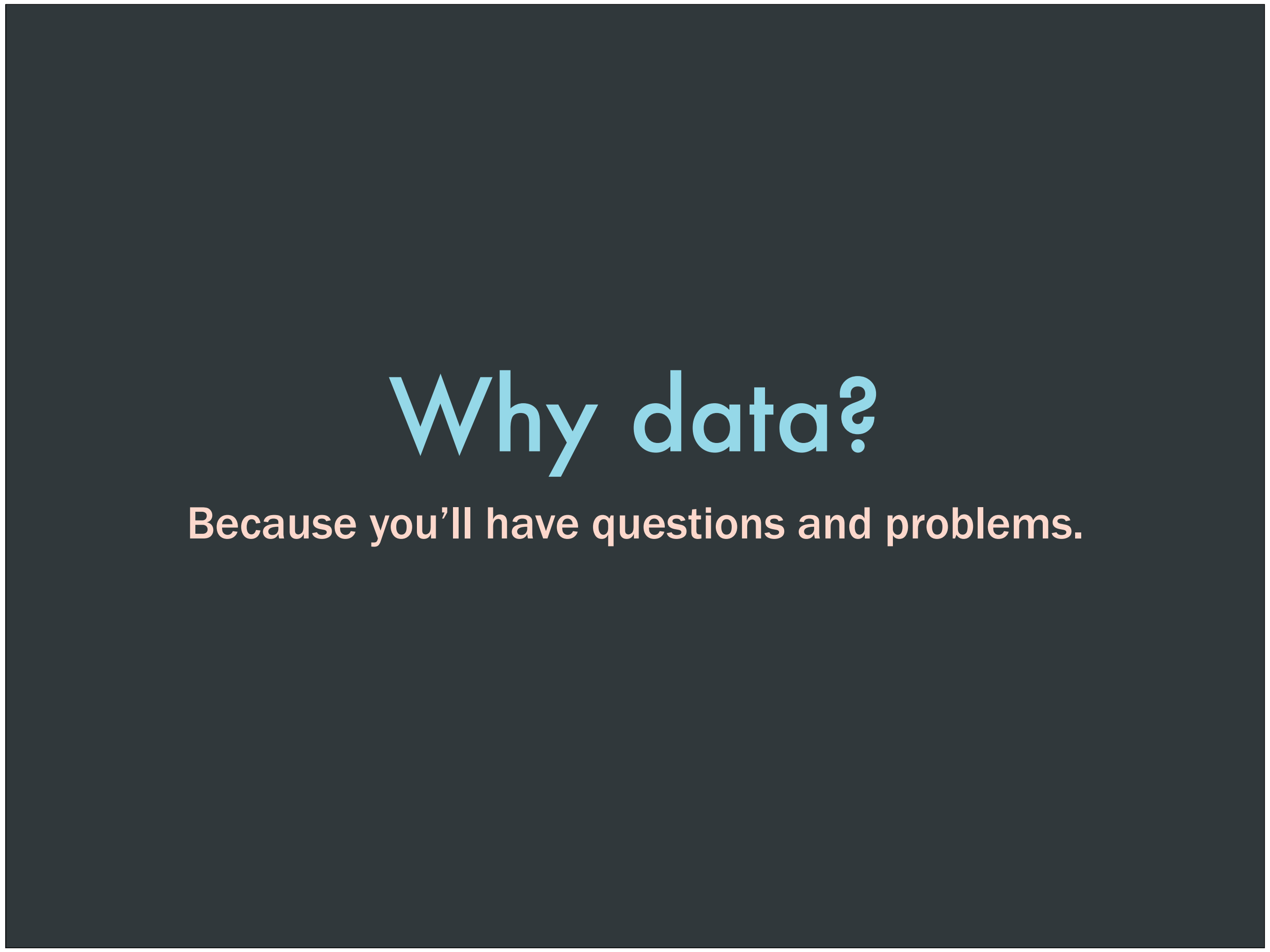
Apologies if this talk is a little disjointed. It’s basically a previous talk I’ve given and a blog post I’ve written mashed together for the first time.

Hi. I’m Greg Reda.

🏠 gregreda.com

🐦 @gjreda

🐙 gjreda



Why data?

Because you'll have questions and problems.

How else will you know how you're doing? Whether you're making progress?

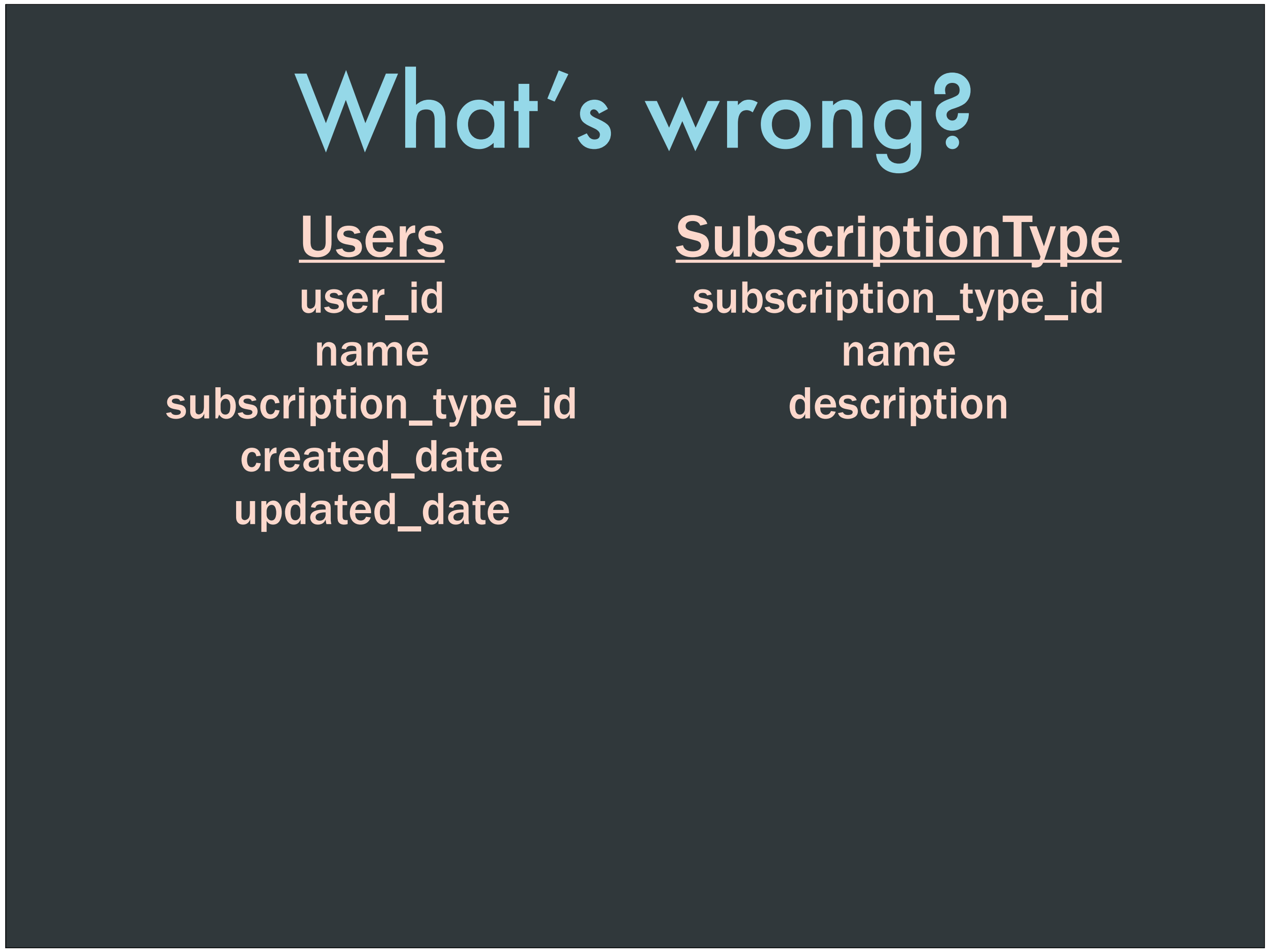
And because your board's going to want it. Lots of it.

Collection

Think up front. It all starts here.

Step one is making sure you’re collecting what you need - you can’t analyze what you don’t have. Think up front about what’s important to your business? What’s your customer’s funnel / workflow? You need to capture that. Is it joinable?

Akouba = IP -> Fraud. Persist what you need, but adhere to privacy concerns.



What’s wrong?

Users

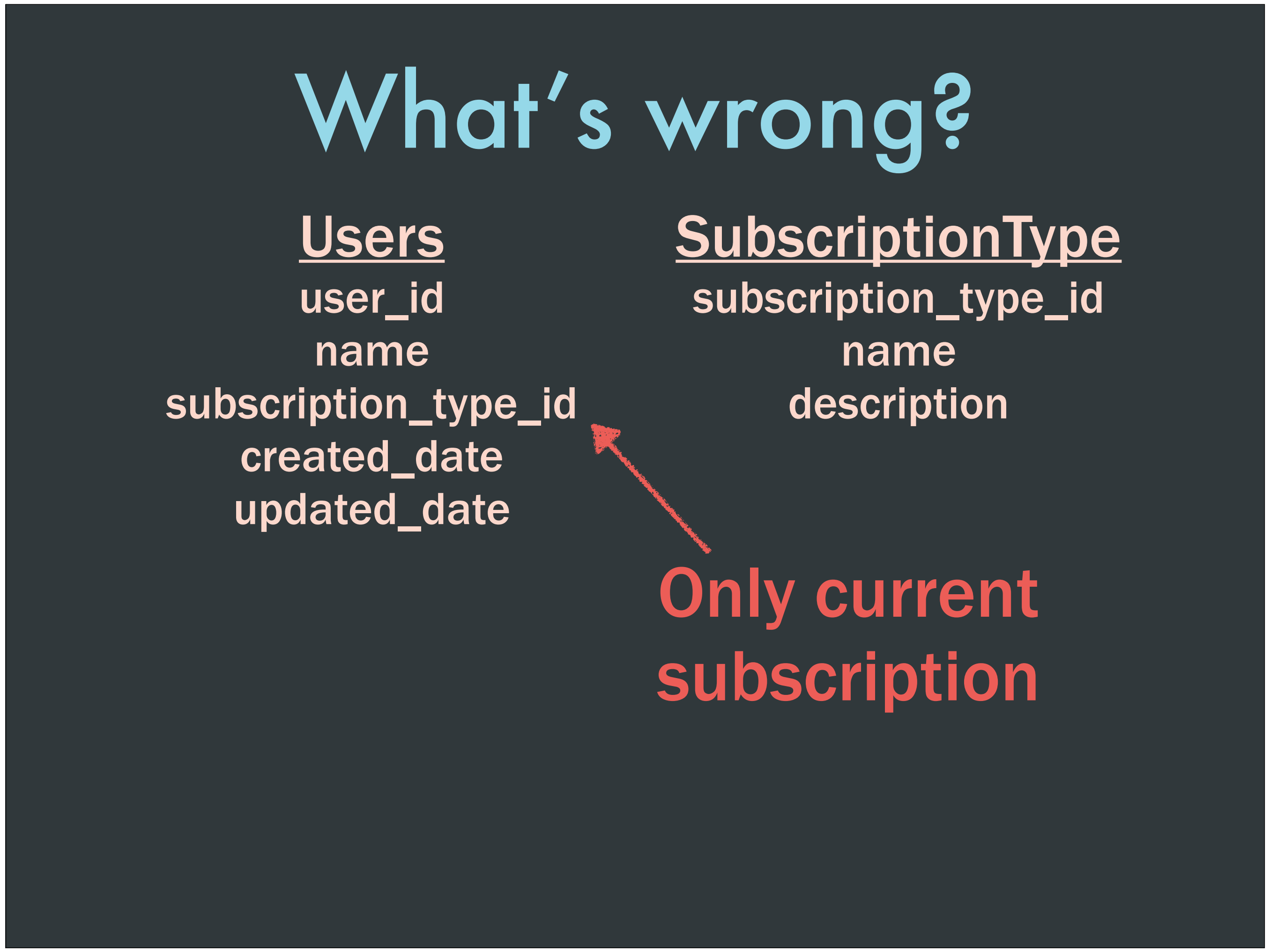
user_id
name
subscription_type_id
created_date
updated_date

SubscriptionType

subscription_type_id
name
description

Database fields that capture “state” = not great.

"How many restaurants were live two weeks ago?”



What's wrong?

Users

user_id
name
subscription_type_id
created_date
updated_date

SubscriptionType

subscription_type_id
name
description

Only current subscription

Database fields that capture "state" = not great.

"How many restaurants were live two weeks ago?"

Implicit Data Loss

Database fields capturing “state.”

Better

Users

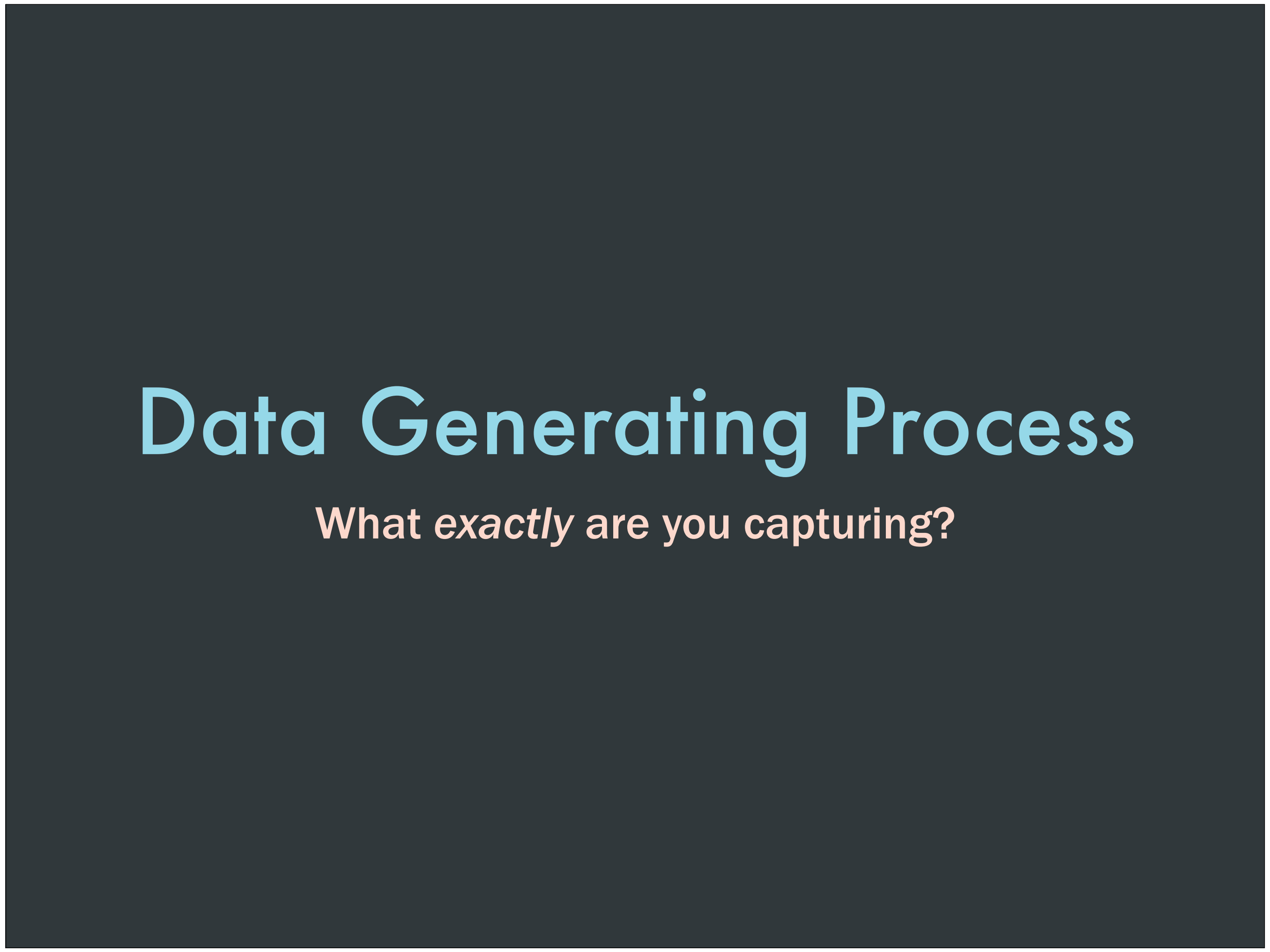
user_id
name
subscription_type_id
created_date
updated_date

SubscriptionType

subscription_type_id
name
description

SubscriptionHistory

old_subscription_type_id
new_subscription_type_id
created_date



Data Generating Process

What *exactly* are you capturing?

Not only is important to capture data, but you need to understand what ***exactly*** you are capturing. Does the event fire on page load, or when they activate the form? Is a new row inserted at the ***start*** of a purchase, or the ***completion*** of a purchase.

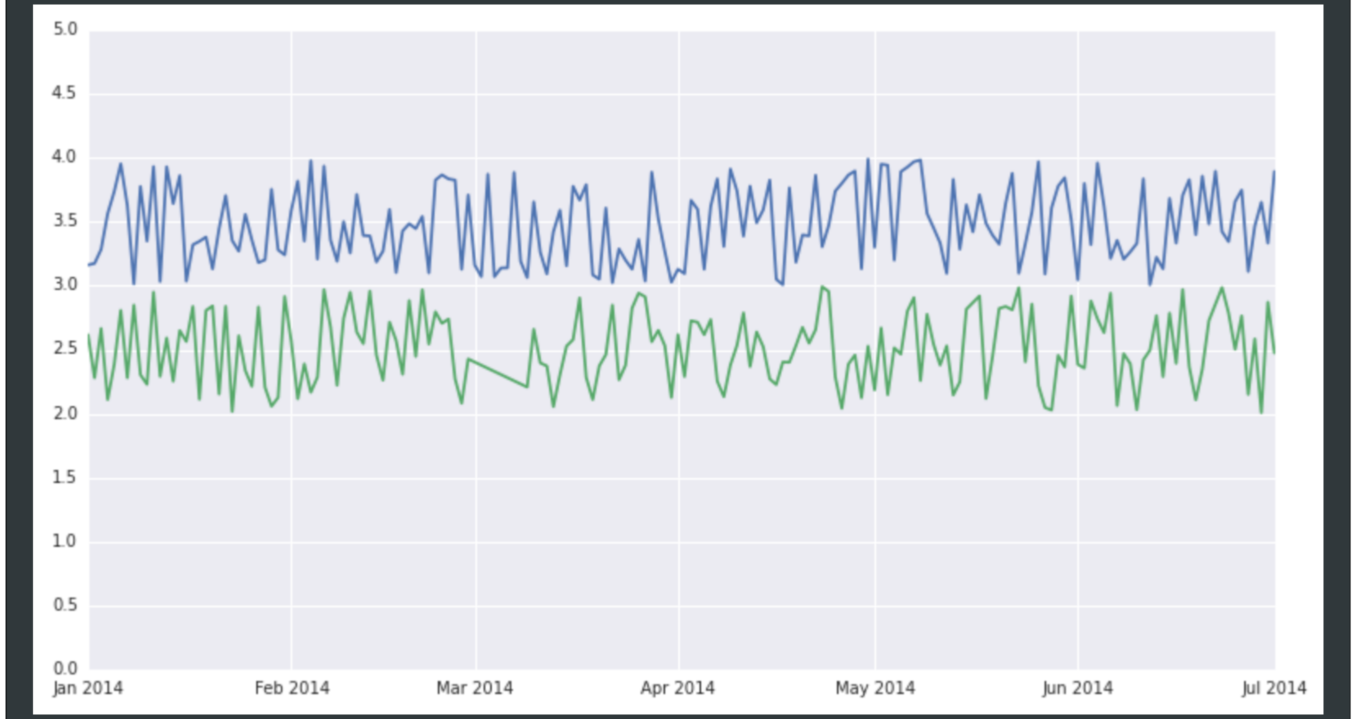
Inspection

Is it all there?

Congrats - you have some data and now want to answer some questions about it. Your first step should be to inspect it.

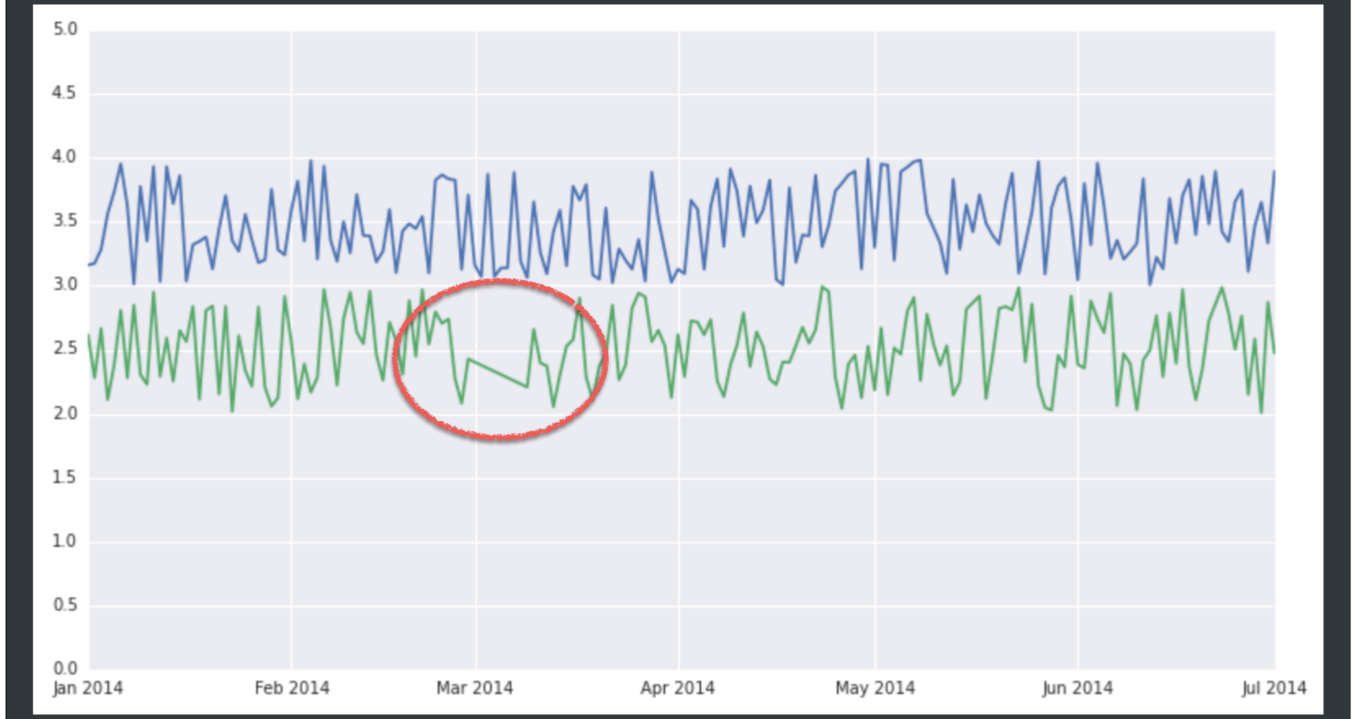
Is it all there? How many data points? What are some summary stats?

What's wrong?



Example: bozo log files. Spent a good amount of time on the analysis, but we were missing days of data.

What's wrong?

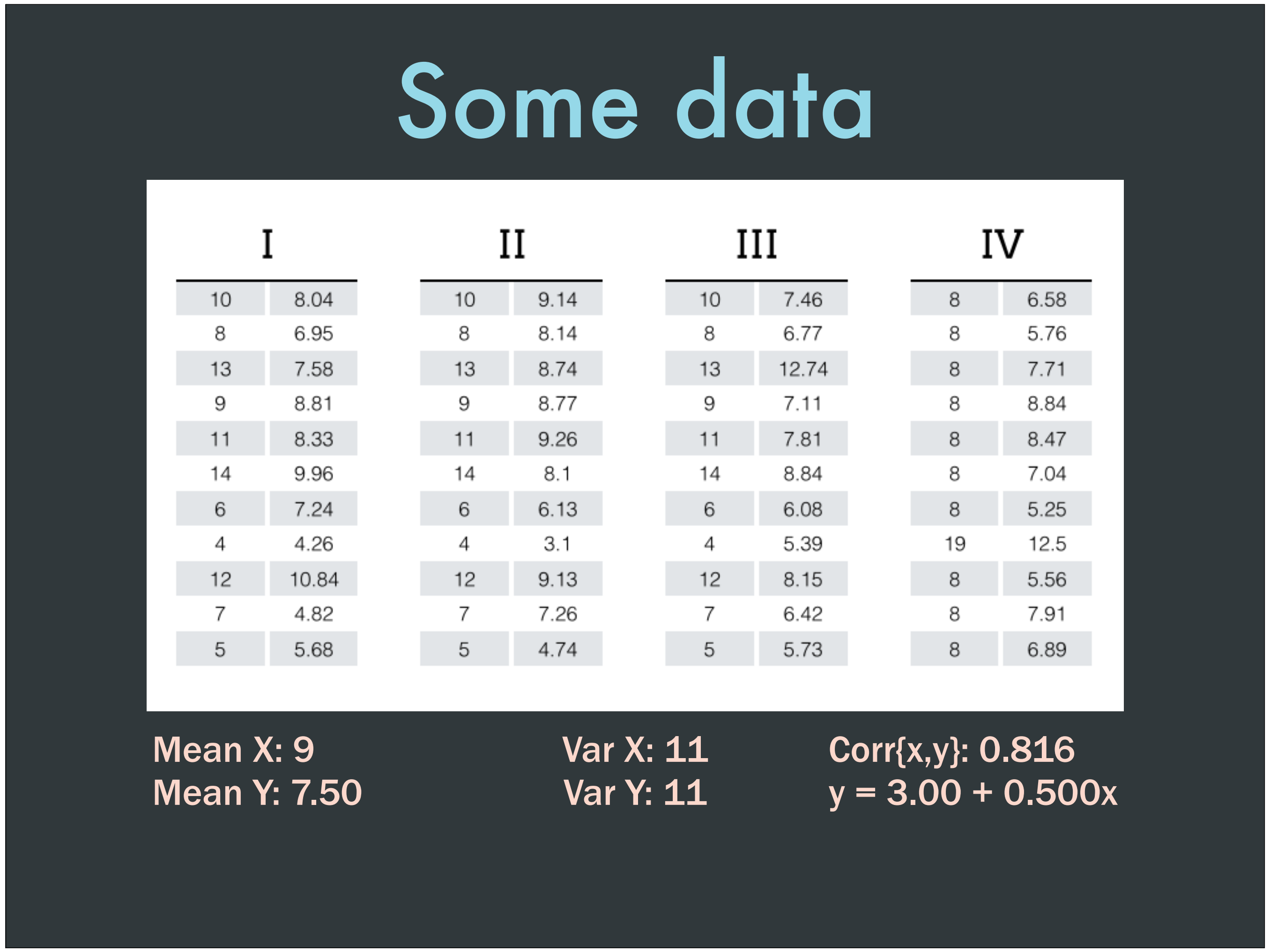

5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Jan 2014
Feb 2014
Mar 2014
Apr 2014
May 2014
Jun 2014
Jul 2014

Summary Statistics

Good.

But not enough.

Summary statistics - we all remember these from grade school. Things like the mean, median, variance, and standard deviation are all good, but by summarizing, we are essentially *throwing away data.*



Some data

I		II		III		IV	
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

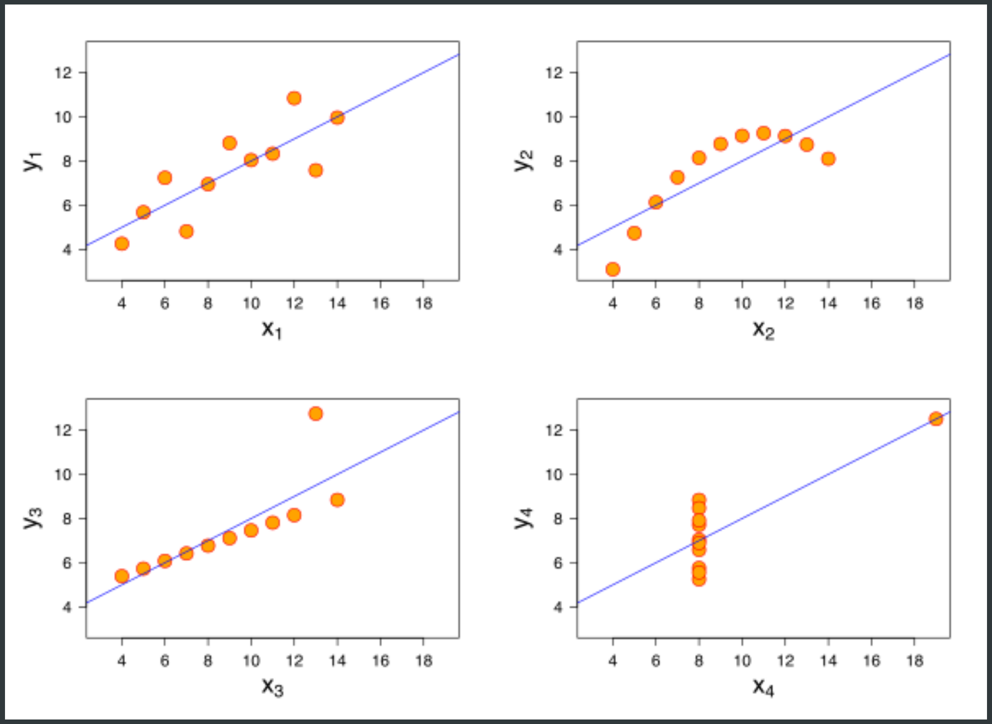
Mean X: 9
Mean Y: 7.50

Var X: 11
Var Y: 11

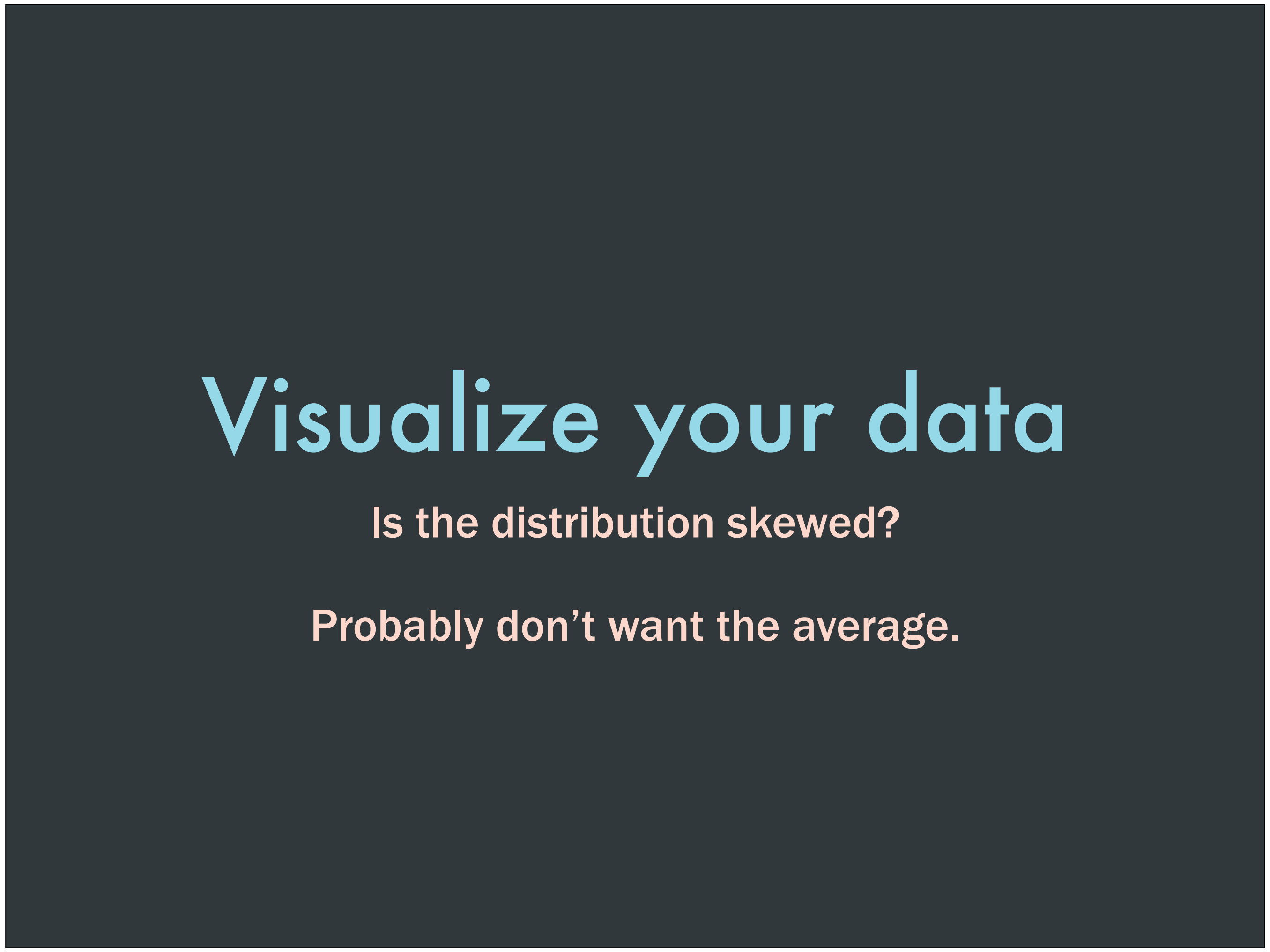
Corr{x,y}: 0.816
y = 3.00 + 0.500x

Here we have four groups of data. Let’s say we wanted to make some generalizations about each population. We can grab the summary statistics and move on, right? Wrong.

Anscombe's Quartet



All have the same mean, variance, and linear model. You ***have*** to look at your data.



Visualize your data

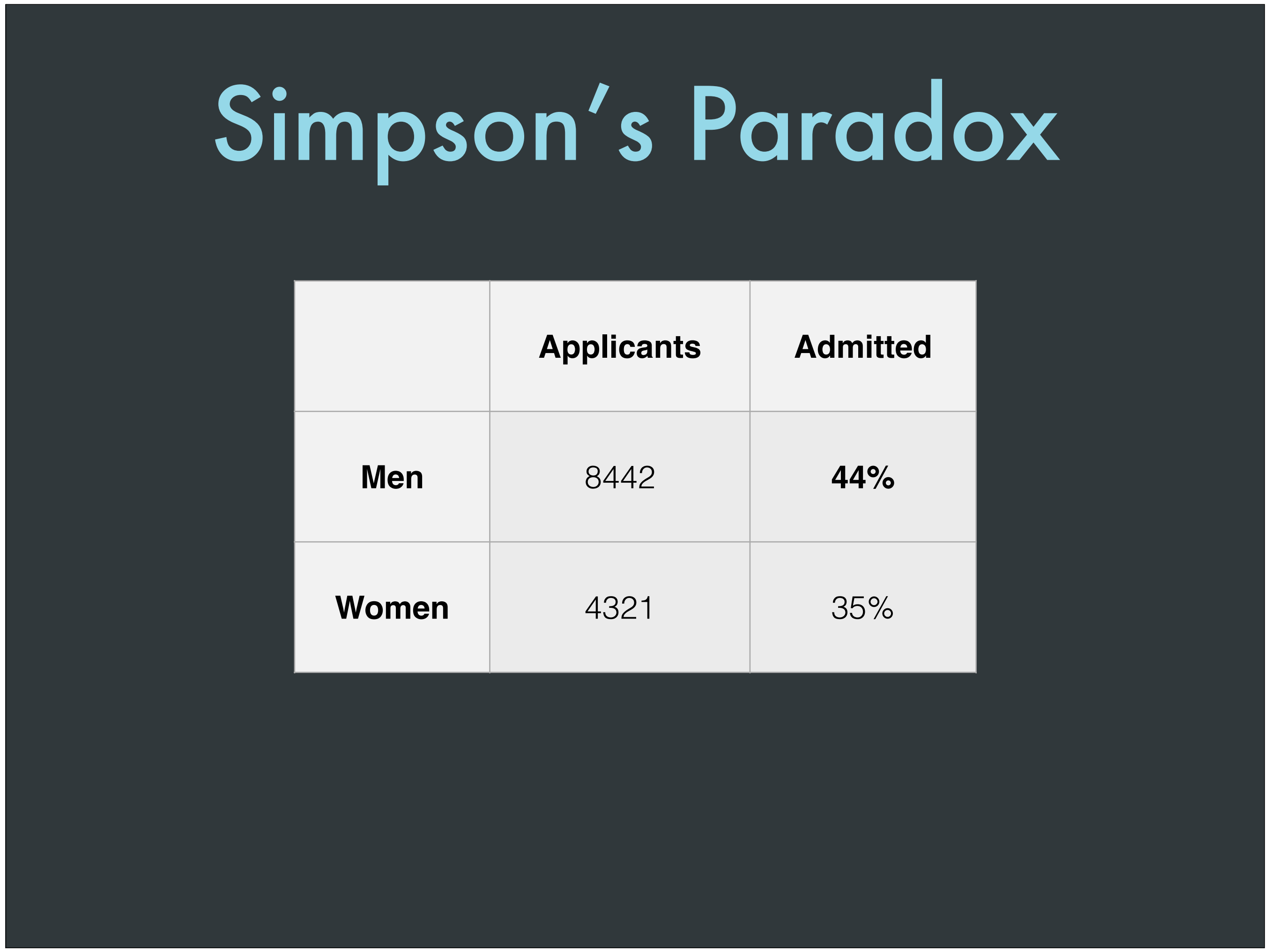
Is the distribution skewed?

Probably don’t want the average.

Distributions are also important here. Highly skewed? You don’t want the mean.

Simpson’s Paradox

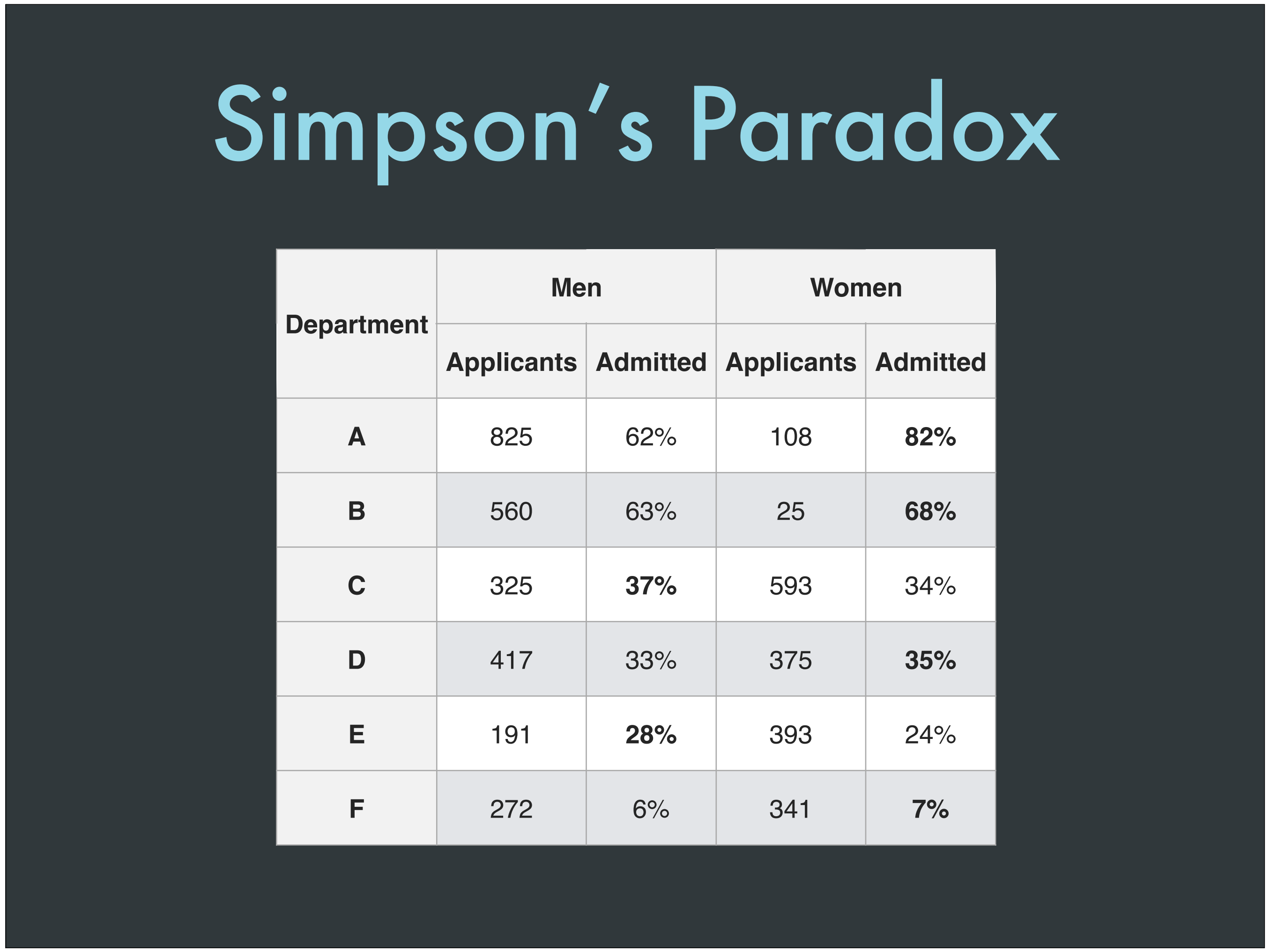
More summary stats hate.



Simpson's Paradox

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Berkeley gets sued due to an alleged gender bias in their graduate school admissions.



Simpson’s Paradox

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

When faceting the statistics by department, there seemed to be in bias at all - women actually had a statistically significant bias *in their favor* for some departments.

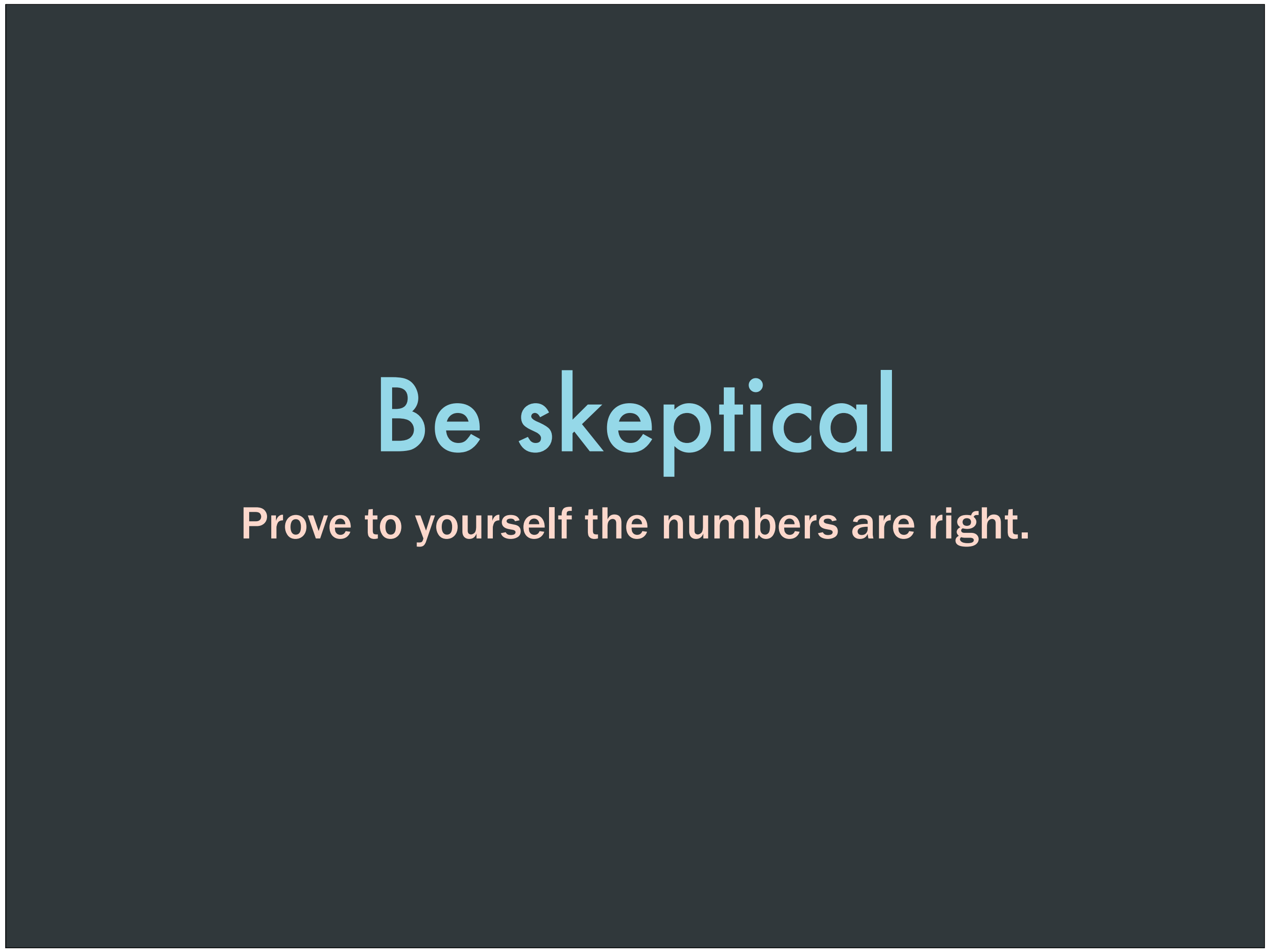
The reality was that women applied to more competitive departments.

Facet all the things!

Split by dimensions.

The same can apply to your key metrics - there will be nuisance to them. Users behave differently depending on their location or age. Markets grow differently. Certain industries will love your product more than others.

Always look at your metrics by multiple dimensions. You'll understand your users better, spot trends easier, and diagnose problems quicker.



Be skeptical

Prove to yourself the numbers are right.

Assume you’re wrong until proven otherwise. Be skeptical. Try to recreate your analysis another way in order to validate it further.

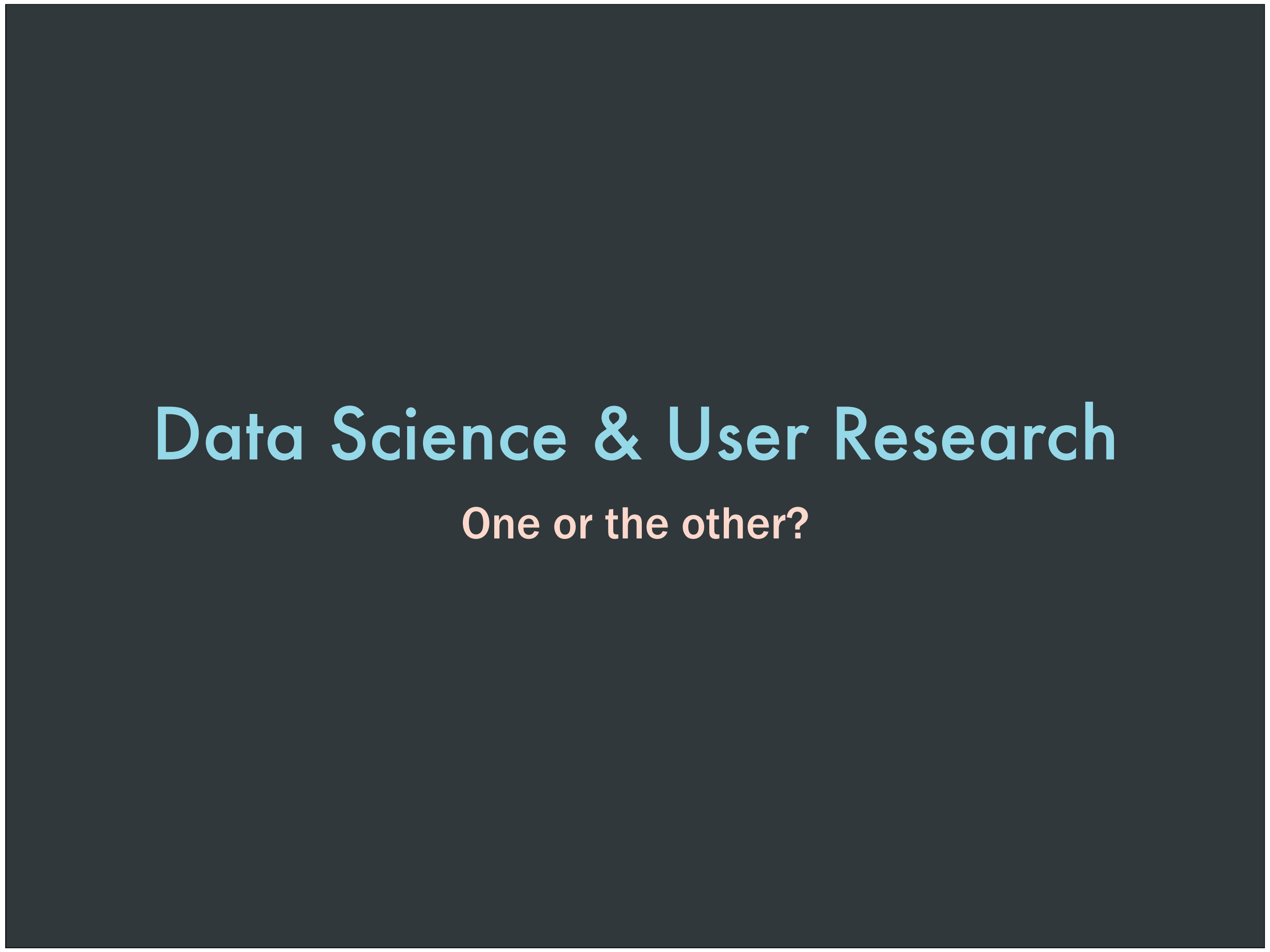
Think like a trial lawyer.

Data as a product

Close the feedback loop.

Data that people will actually pay for (Food Genius).

Don’t just look backwards - Better search rankings. Advertising. Personalization.



Data Science & User Research

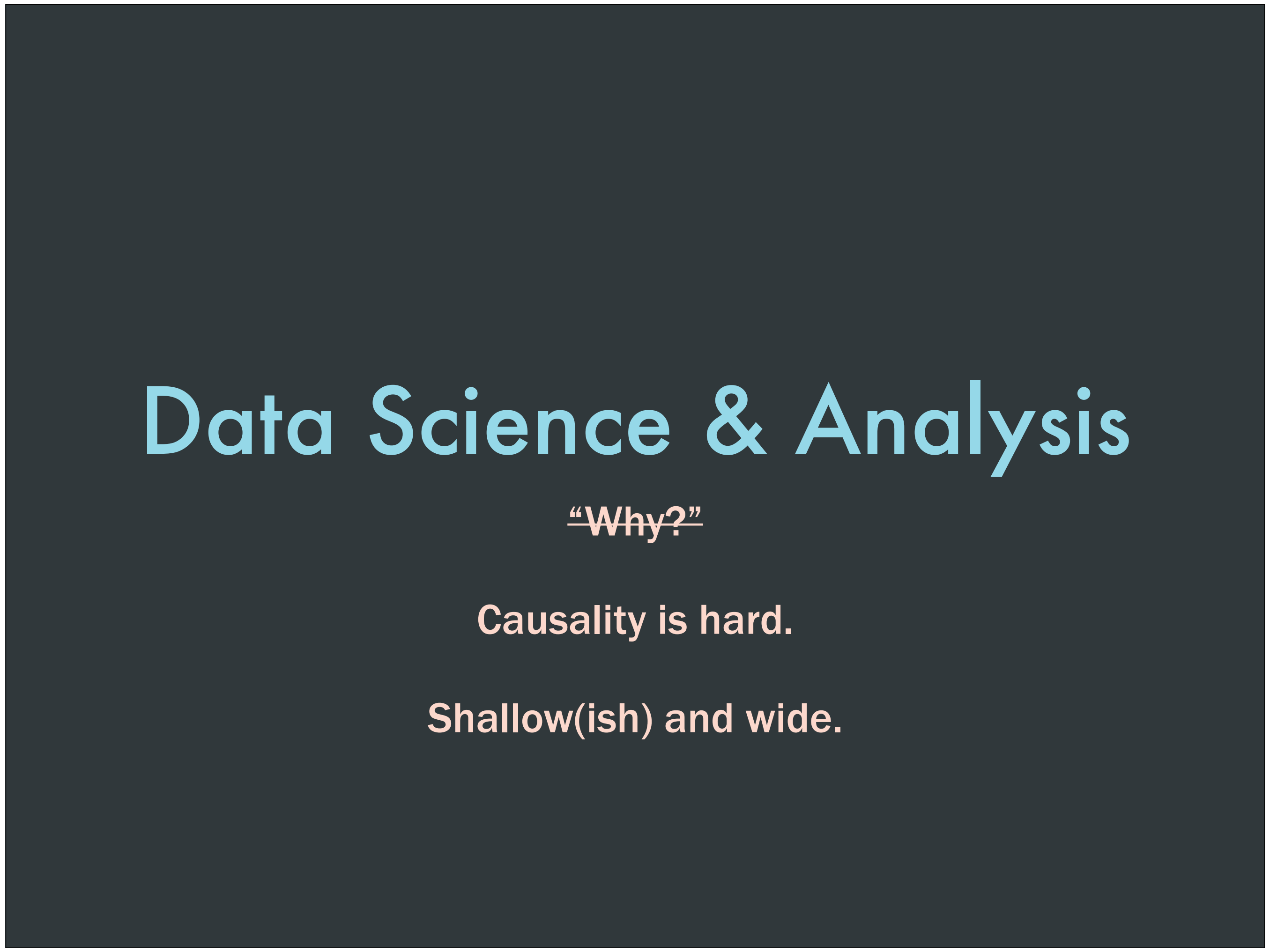
One or the other?

No. Do both.

User Research Studies

Narrow(ish) but deep.

They’re narrow. User research is great at many things - particularly at giving us a deep understanding about our users, their environment, their behaviors, etc. How many users can we do a study with 10? 20? This won’t give us an understanding of all of our users, but it will allow us to generalize about them - to potentially even create personas of our users.



Data Science & Analysis

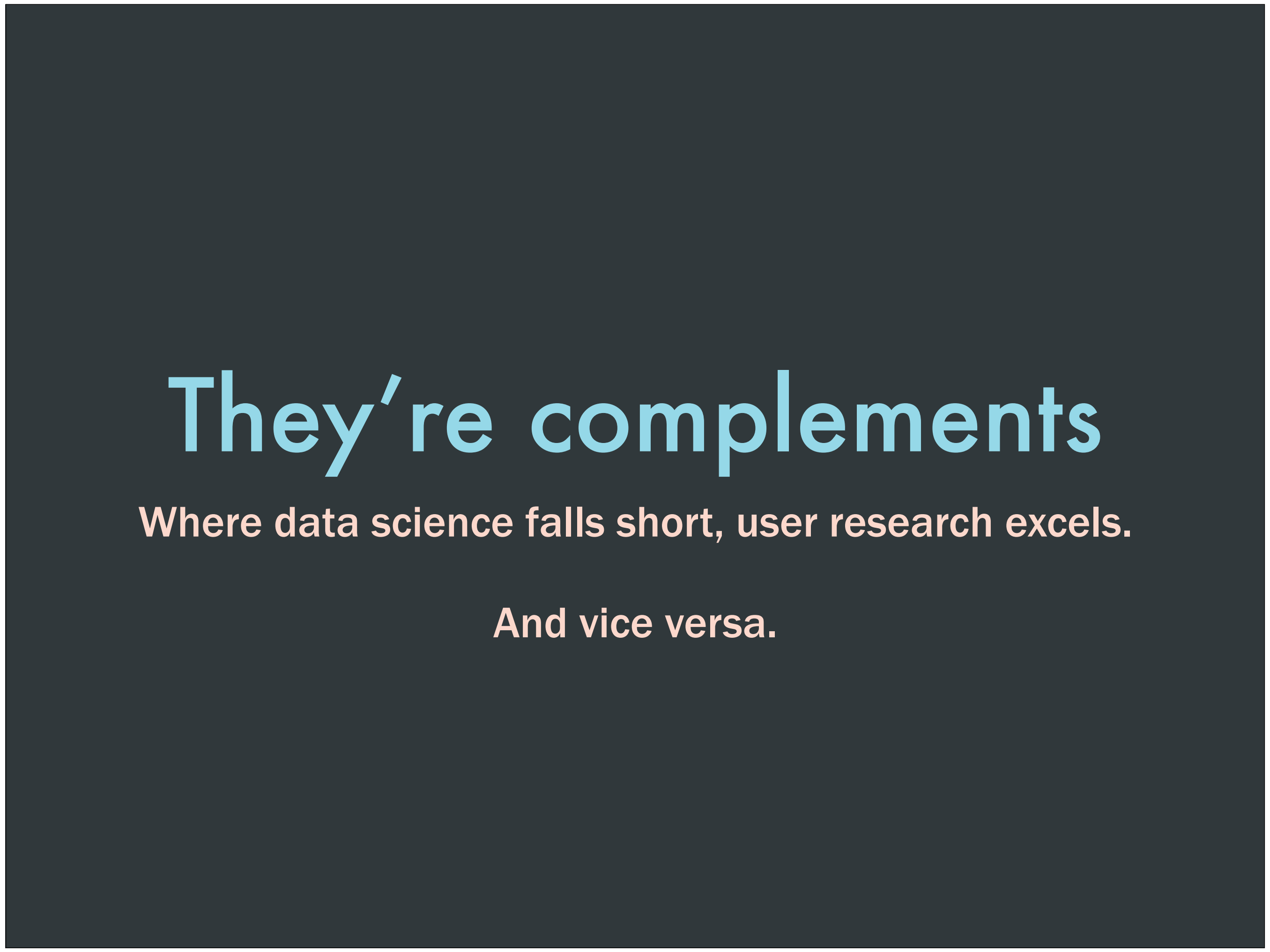
~~“Why?”~~

Causality is hard.

Shallow(ish) and wide.

Data science suffers from the opposite problem - sample size generally isn’t a problem for us. We can look at a TBs of data without too much difficulty.

But causality is hard. Data is great at telling us many things - who, what, where, when - but the one thing it often can’t tell us is _why_. Why is something happening?

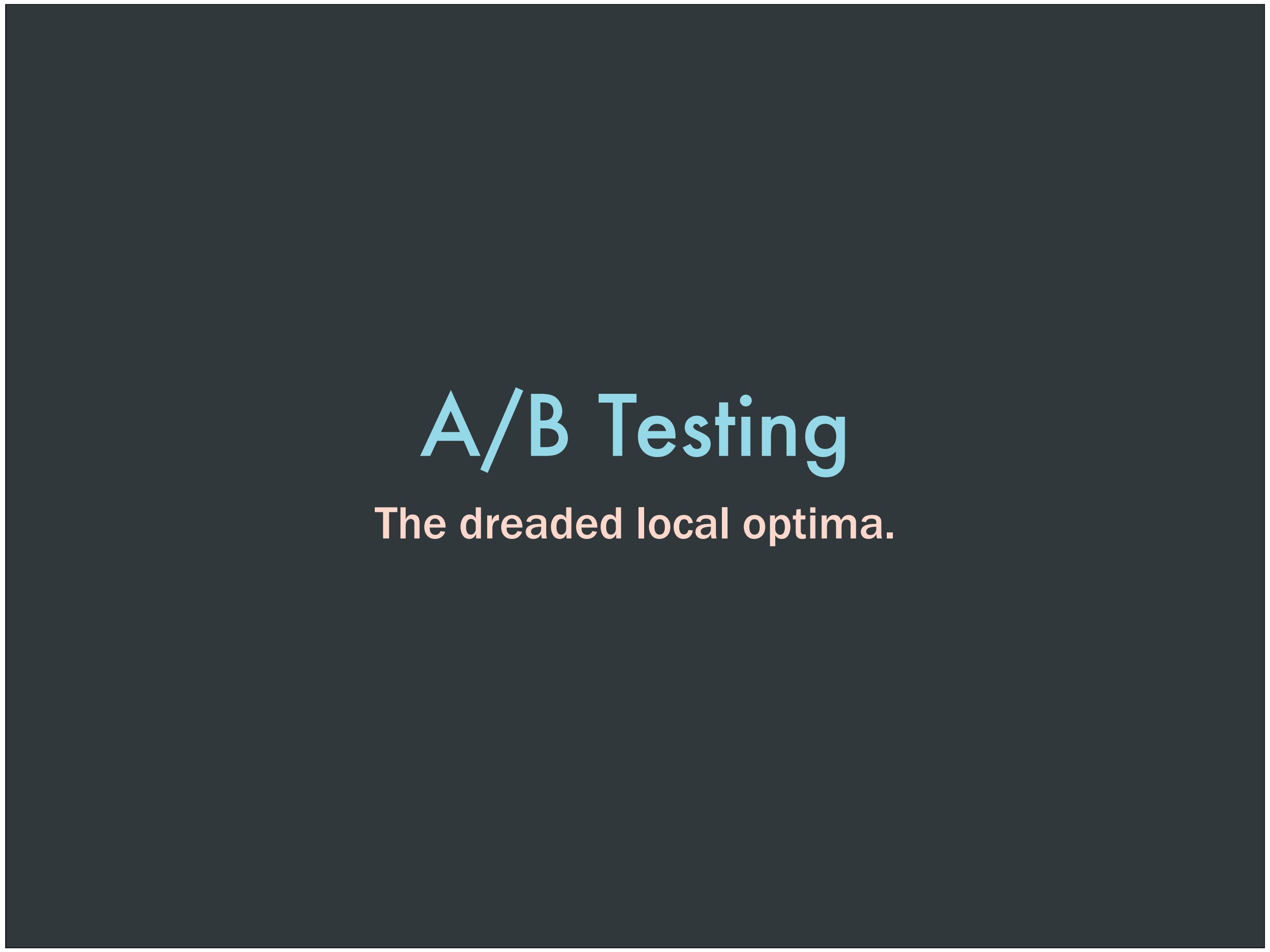


They're complements

Where data science falls short, user research excels.

And vice versa.

Despite things often appearing as quantitative vs qualitative, the fact is data science and user research are complements - they balance each other out.

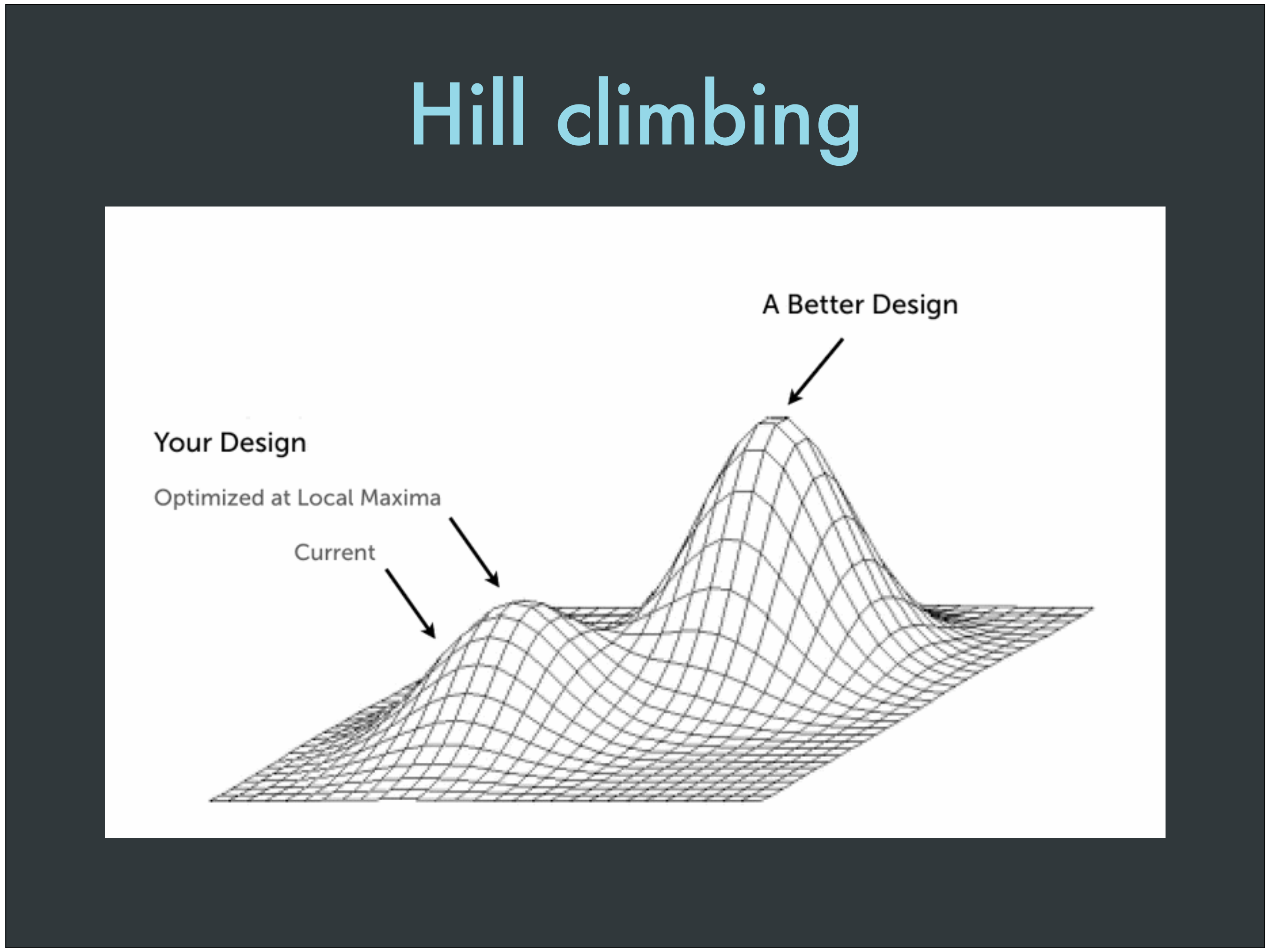


A/B Testing

The dreaded local optima.

Does everyone know what A/B testing is?

With A/B testing, we have even more opportunity for collaboration with our coworkers in design and engineering, but if we’re not careful, we find ourselves at the dreaded local optima.

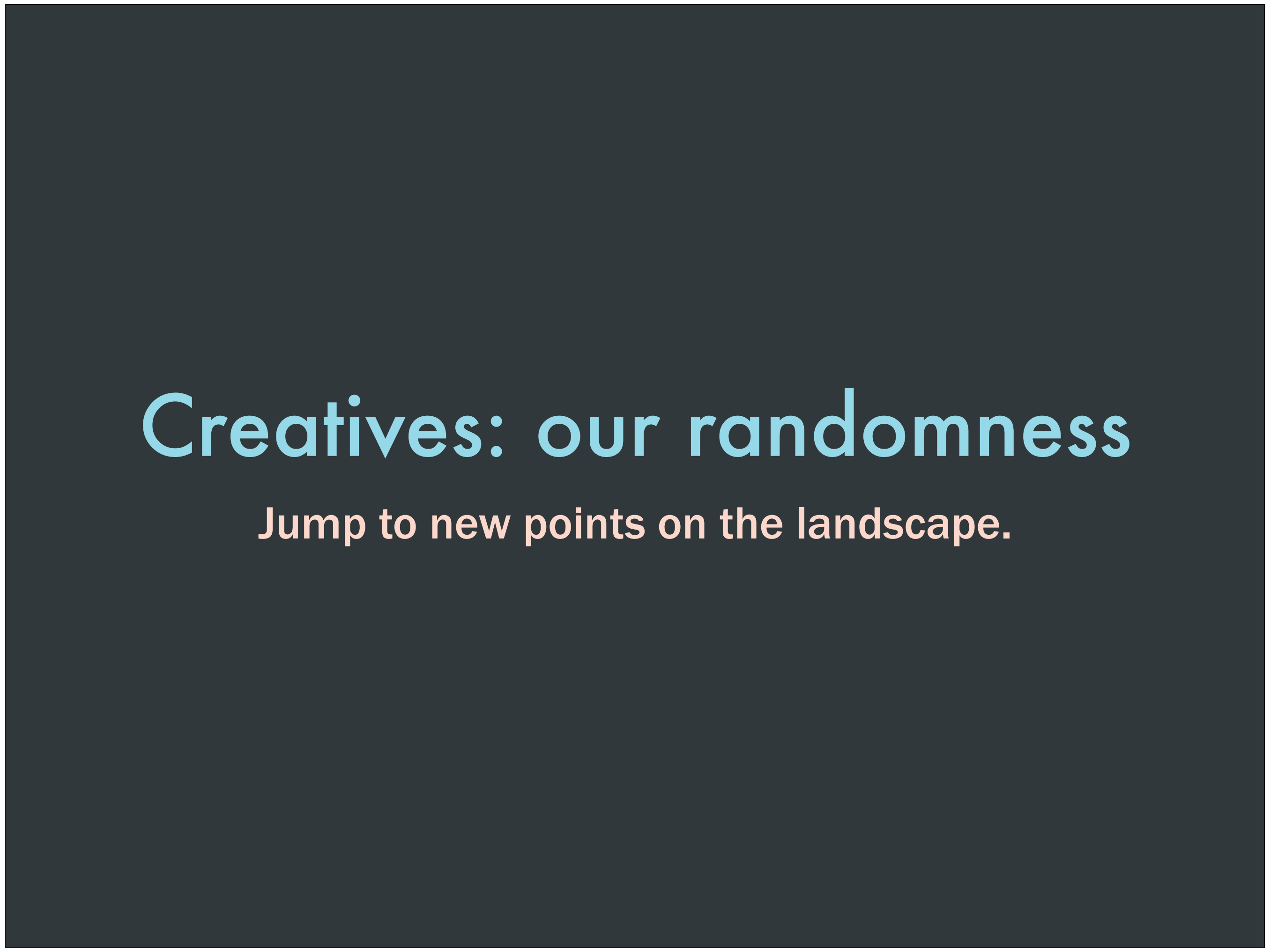


Hill climbing

It’s akin to hill climbing. If we are driven by data, we’re at risk of hitting local maxima because we don’t have a great understanding of where we are on the landscape.

How can we approximate that we’re not at a local maxima though? How do we know?

We jump around with some randomness. Essentially, you randomly move to a new position on the map and if it’s higher than your previous position … start climbing (and jumping around more).



Creatives: our randomness

Jump to new points on the landscape.

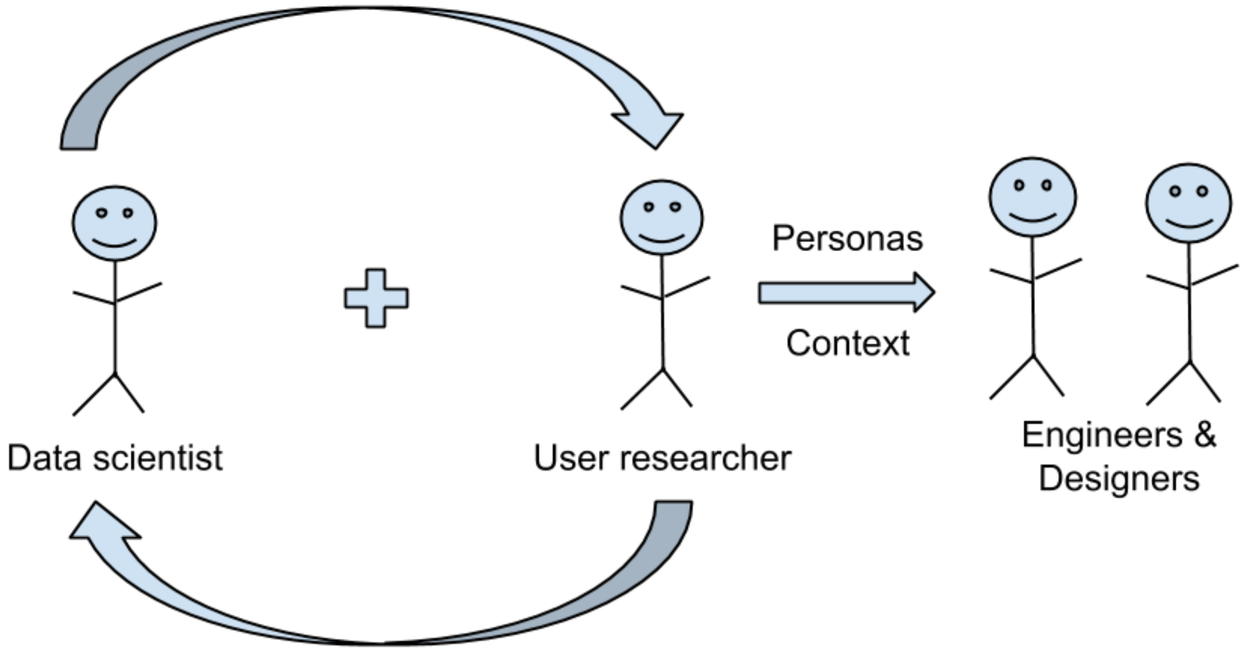
Designers are our randomness - For our product to reach new heights, we need to take chances - and inject some randomness into it. If we continue to optimize the same thing, we’ll run out of hill to climb. We need to evaluate new designs holistically; not just incremental changes - these new designs are akin to jumping around the landscape, and by jumping around, we learn more about the landscape and its topology.

Data + UX + Design

More feedback loops.

Multidisciplinary teams of data scientists, user researchers, designers, and engineers, we’re able to better inform everyone in the process, and create better experiences for our users.

It probably ends up looking something like this …



Data scientist

User researcher

Personas

Context

Engineers & Designers

- Designers love this.
- See world through users’ eyes - personas
- Data points about their users, coupled with behaviors, feelings, deep understanding from user research.
- Potential for tailored experiences; why “one-size-fits-all?”

Don’t rely on data alone

“No one got fired for hiring IBM.”

Clients sometimes just want to “do what the data says.” Akin to IBM.

To actually reach new heights, create new great products, you need to use all available information AND take risks.

Try new things. Just make sure you collect data when you do it.

Summary

1. Collection: Think up front.
2. How was this data generated?
3. Inspection: Is the data as you’d expect?
4. Summary statistics are your friend …
5. But visualizing your data is a must.
6. Facet your data: metrics across dimensions.
7. Be skeptical.
8. Data as a product: feed it back in.
9. Quant + Qual = complements.
10. Be data-informed, not data-driven.