Today I want to talk about a subject that doesn't get covered enough.

Why good models fail

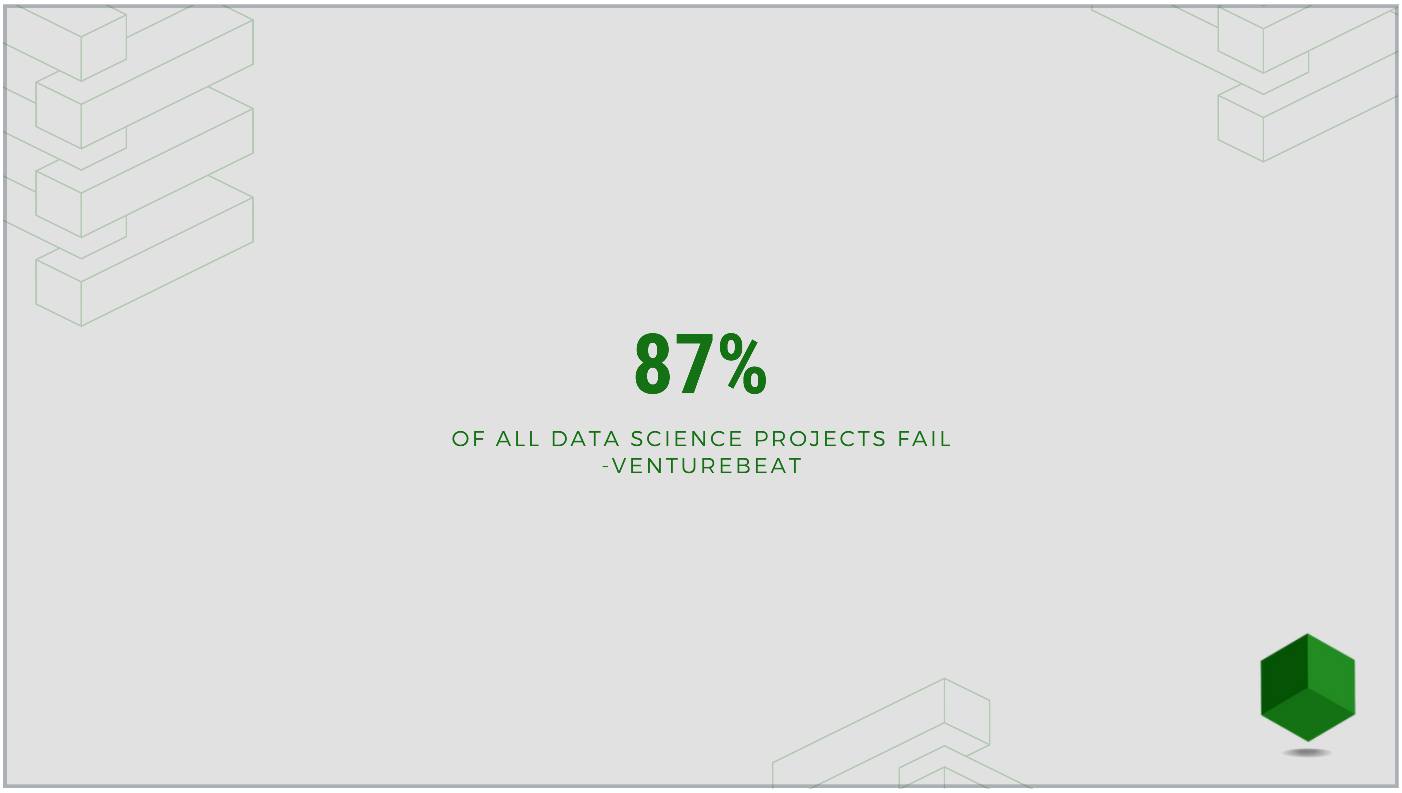
To be clear, there are many many reasons why DS models fail, but I argue that these can be traced back to 3 primary sources.

Ambiguity

Uncertainty

Bad Science

For simplicity, I'm talking about models that are deployed correctly. Otherwise we'd have to throw a few more reasons in here on why models fail. Another talk for another day.



Who knows if this is accurate but it lines up with my experiences.

I've worked on way more projects than those i put in production.

I can count on one hand how many models I've put into production.

And looking back it was because of 3 main reasons.

AMBIGUITY

WHEN REQUIREMENTS ARE HARD TO DEFINE

TOPIC MODELING for the contact center.

We had just discovered topic modeling and wanted to use it.

There was no appetite for it because we didn't know:

- What we were trying to solve
- Who the user was (rep? caller?)
- What was valuable to the user or to leadership

As with most junior DSs, I didn't let that stop me from pulling data and building a solution anyways.

We had no clue what questions to even ask.

The reality is, the DSs are given a very high level objective and not much more.

Then it's up to the DSs to ask a ton of wrong questions that hopefully begin to lead to the right ones (or at least spark the right ones from the business).

Then the business begins to discover what it wants.

Wheels turned, we did some fancy analysis. Project failed.

AMBIGUITY OCCURS WHEN WE DON'T KNOW WHAT QUESTIONS TO ASK

HOW TO NAVIGATE **AMBIGUITY** IN DATA SCIENCE?

data science is at the intersection of data, business, and tech

we are responsible for bringing all these domains together into a cohesive and coherent analysis

find your gaps early and fill them

- resist the urge to open a coding env
- ask tons of questions
- get comfortable not knowing things

I got better at data science not because i was better at ML, or coding, or math.

I got better at data science when i started to learn the business.

Becoming better at ML, data eng, software, math, stats, will not make you a better DS as much as becoming a SME will. (assuming a base level in all these things).

once you know the questions you're answering

research, present, feedback

shorten feedback cycles to ensure constant alignment with the business

you will be asked to "uncover insights" and "find patterns", "analyze trends"

unless you are a SME this isn't feasible.

instead get areas of study from the business

is this important?

research, present, get feedback

repeat

Always defer to the experts.

Your role is the builder, their role is the surveyor

i have no clue where to put the building

On the TOPIC MODELING project, we were trying to force a solution on a problem we didn't understand.

Had we consulted an expert first we would've been better off.

Can be a process expert, domain expert, systems expert, data expert. will need each at different times.

defer to them!

4. EXPLORE WITHOUT BUILDING

UNDERSTAND THE PROBLEM FULLY, THEN START DESIGNING

Building before understanding the problem leads to serious status quo bias.

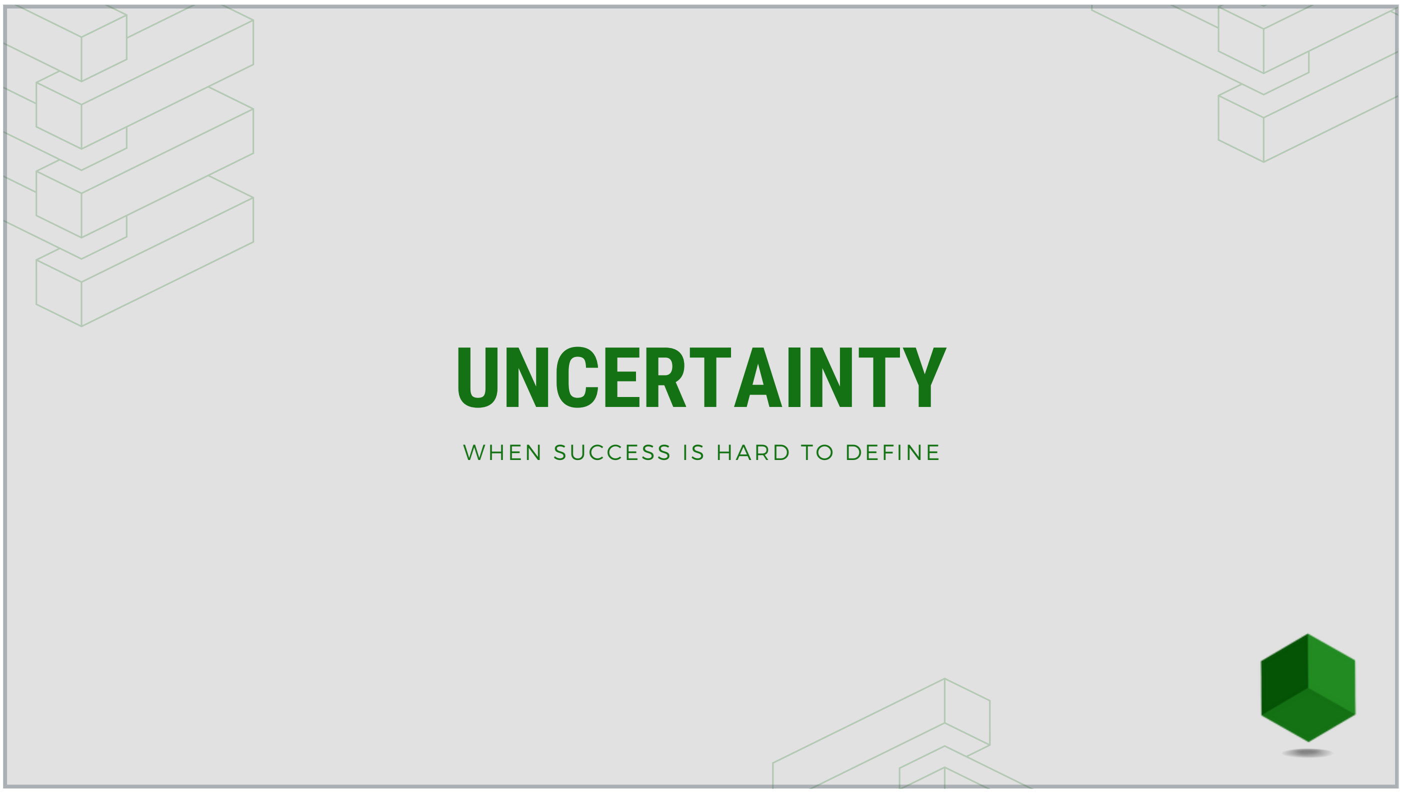
status quo bias - let's keep going this route since we've already started. it can be costly to pivot!

Fully understand the problem then start designing the solution

In the TOPIC MODELING example, we started with the solution, and totally missed the problem.

Be prepared to throw everything away after the POC, even if successful

The only thing you take from a POC to production is what you learned



This is hard for an org because usually the DS project you're working on is brand new to the org and they don't know how measure success.

EXAMPLE.

One of the models i put in production a few years ago (and still is) was ATTORNEY REP.

The model was good, performed well.

To measure success

We split the predictions into two groups—test, control.

But we didn't define these groups identically.

First we compared likely AR to all claims instead of just those similar to the likely AR claims. This lead to us proclaiming a MUCH higher savings number.

We didn't know how to define success.

We got called out.

HOW TO NAVIGATE **UNCERTAINTY** IN DATA SCIENCE?

1. UNDERSTAND THE PROCESS

THE FULL LIFECYCLE OF A RECORD

For ATTORNEY REP, we needed to dive deeper into the process to learn how these claims were handled.

Understand the full lifecycle of a record.

You're predicting at some level (group, individual, multiple preds per level, etc).

Understand the data at that level.

your model can be perfect at the group level, but if the business needs it at an individual level it's useless.

How important is time (time-to-prediction, time-after-prediction)? Contact after prediction? This is huge when humans are using your ML, eve more important to understand the process.

How does a record get created, edited, aggregated?

2. UNDERSTAND THE KPIS

THE SOURCE, TARGET, CALCULATION, STAKEHOLDERS

ASKUNUM

AskUnum AI resolving easier tickets but leaving the harder ones which caused avg time-to-close to increase.

How to prevent reps from looking bad?

Increased resolution time was actually expected, and potentially a good sign.

3. UNDERSTAND THE DOLLARS

SAVINGS/REVENUE CALCULATIONS MUST BE DONE METHODICALLY

It can be very tempting to throw out numbers -- 10 million, 100 million

or even have a shallow link between accuracy and revenue that tracks with it linearly or whatever.

these calculations can get very complicated.

headcount, hours worked, per click, avg cost per unit, cost of something broken down by different dimensions, etc

When presenting to leadership, especially finance, having this calculation be air tight is a good way to build trust.

Understand how the process impacts the KPIs

How the KPIs impact the dollars

And that your savings/revenue calculations need to be air-tight

which brings us to our 3rd reason why good models fail.

incorrectly measuring success is spurred by bad science

BAD SCIENCE

WHEN VALUE IS MIS-MEASURED

1. DATA LEAKAGE

TRAINING INFO SHOULD NOT LEAK INTO TESTING BUT OFTEN DOES

through data

through hyperparameter tuning

through embedding layers

through aggregate functions before cross validation splits

target leakage

this inflates performance metrics

ATTORNEY REP (EMBEDDINGS)

2. TRAINING-SERVING SKEW

TRAINING PIPELINES SHOULD MATCH INFERENCE PIPELINES

the production data is often sent through many layers of cleaning, aggregating, etc before it gets to down stream data stores like an analytics database or semantic layer.

locate your data and process experts and have them review your code/data

API contracts

3. INCORRECT METRICS

INAPPROPRIATE METRICS ARE THE HALLMARK OF BAD SCIENCE

Accuracy for imbalanced data

p-values for decision making

r^2 for a performance metric

BIGGEST STATISTICS MISTAKES IN DATA SCIENCE

treating points estimates as truth, or confidence intervals.

i've been many situations where an ML vendor came in boasting of a mediocre model but then totally relied upon that model's confidence intervals

not calculating uncertainty

parametric tests are used when data follow a particular distribution (normal).

to quote Leigh, "most real data don't fit those models!"

basing decisions around p-values

check assumptions

check for leakage, train-serving skew, etc

correlation does not imply causation

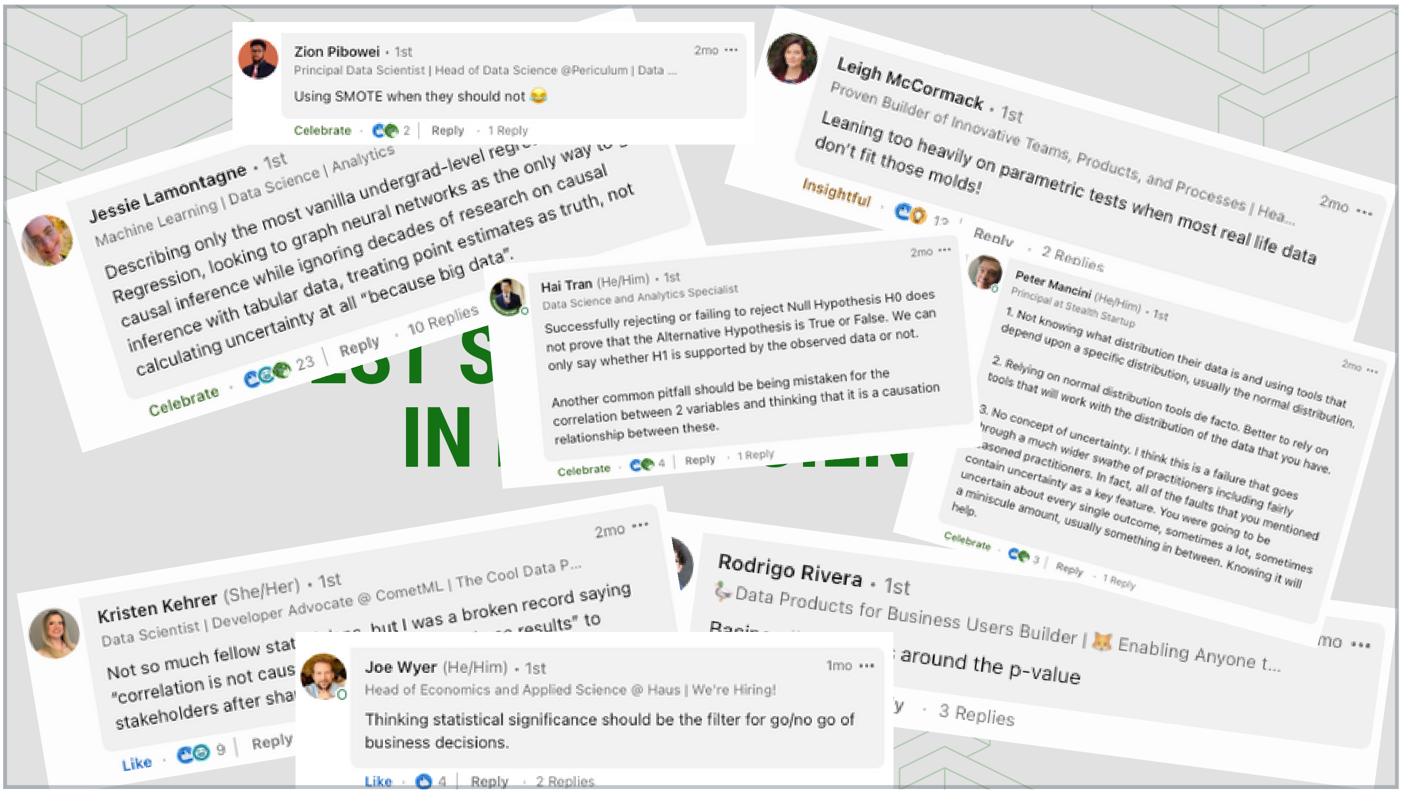
not knowing distributions

using algorithms that assume normal distributions

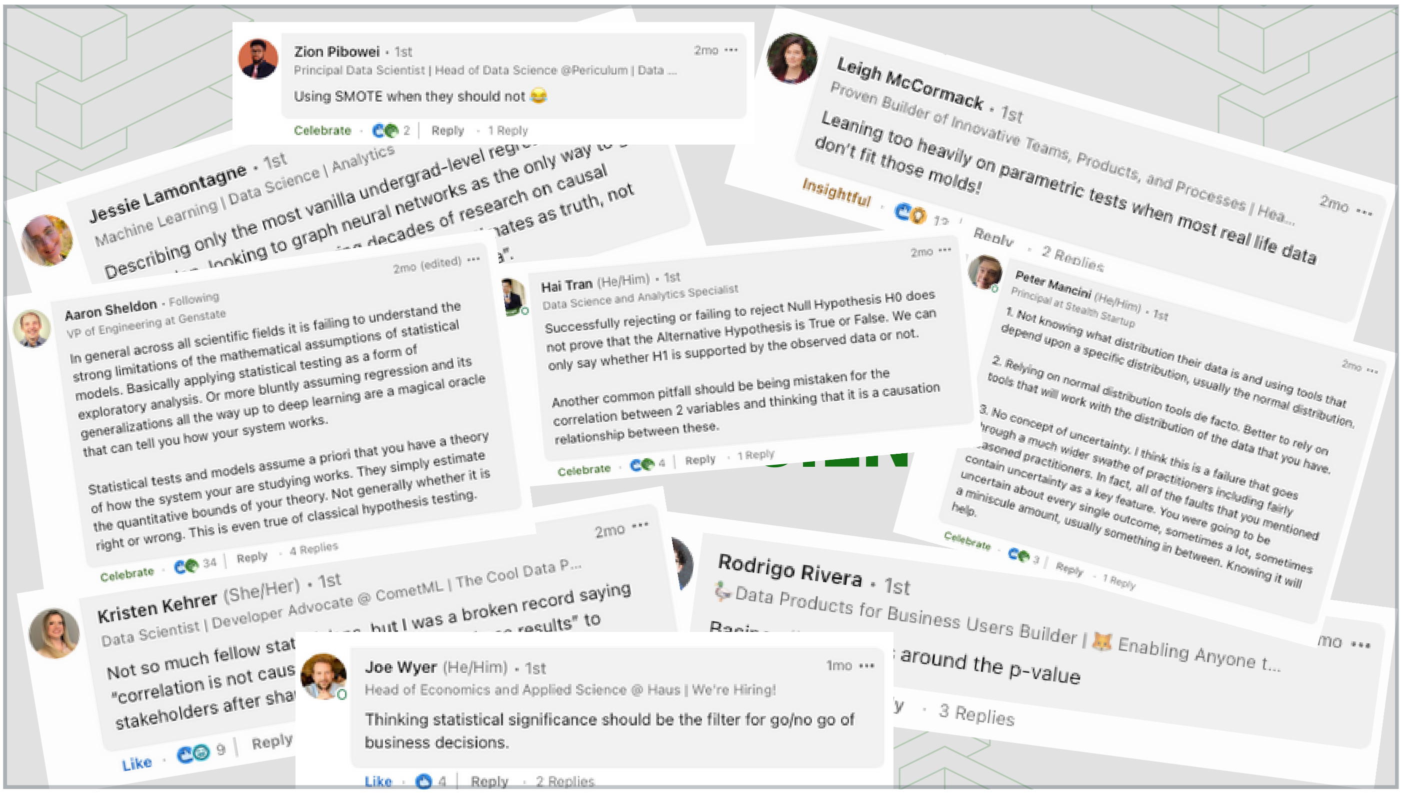
uncertainty

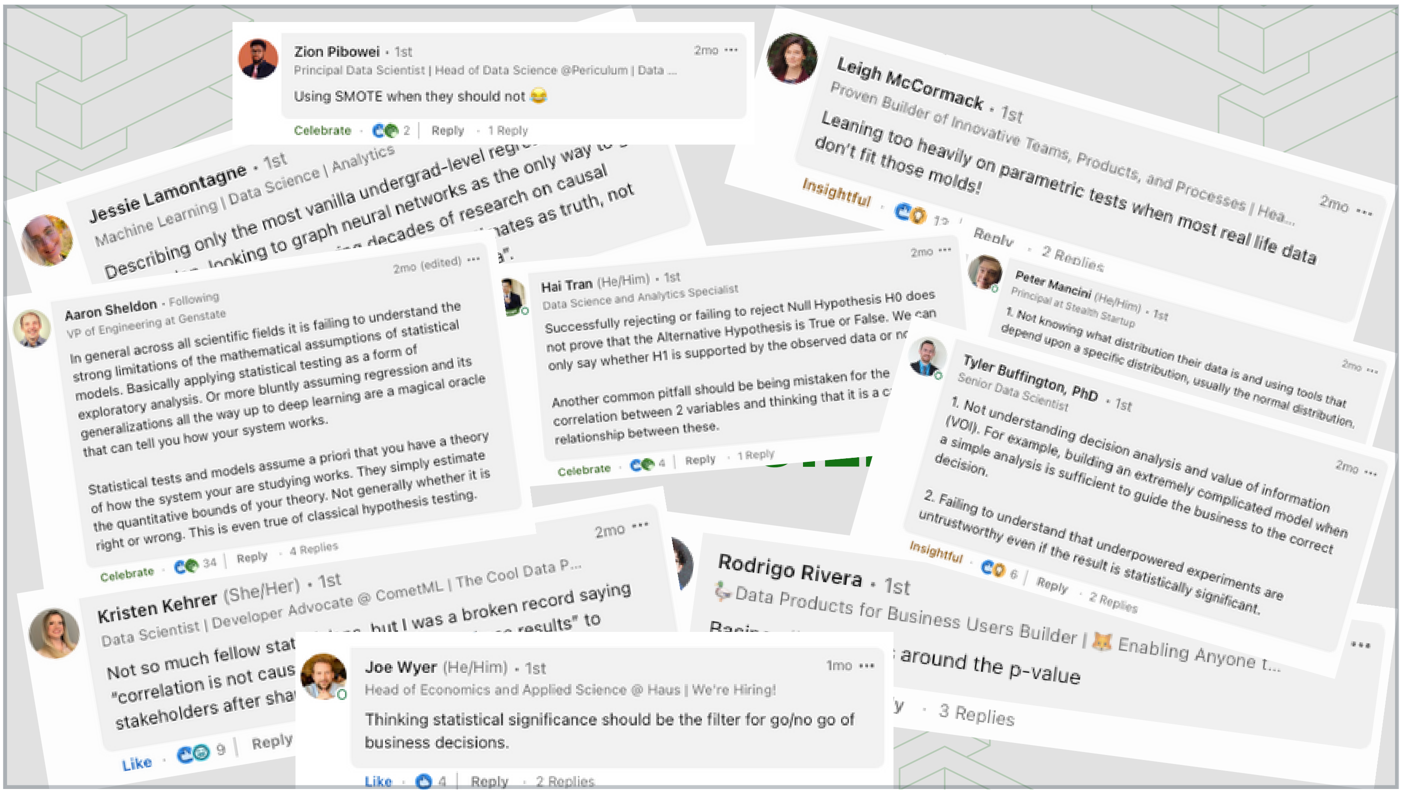
thinking statistical significance implies success or go/no-go


Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
Using SMOTE when they should not
Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics
Regression, looking to graph neural networks as the only way to
causal inference while ignoring decades of research on causal
inference with tabular data, treating point estimates as truth, not
calculating uncertainty at all "because big data".
Leigh McCormack · 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
Leaning too heavily on parametric tests when most real life data don't fit those molds!
Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.
2. Relying on normal distribution tools de facto. Better to rely on tools that will work with the distribution of the data that you have.
Rodrigo Rivera · 1st
Data Products for Business Users Builder | Enabling Anyone t...
around the p-value
Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate @ CometML | The Cool Data P...
Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
Thinking statistical significance should be the filter for go/no go of business decisions.



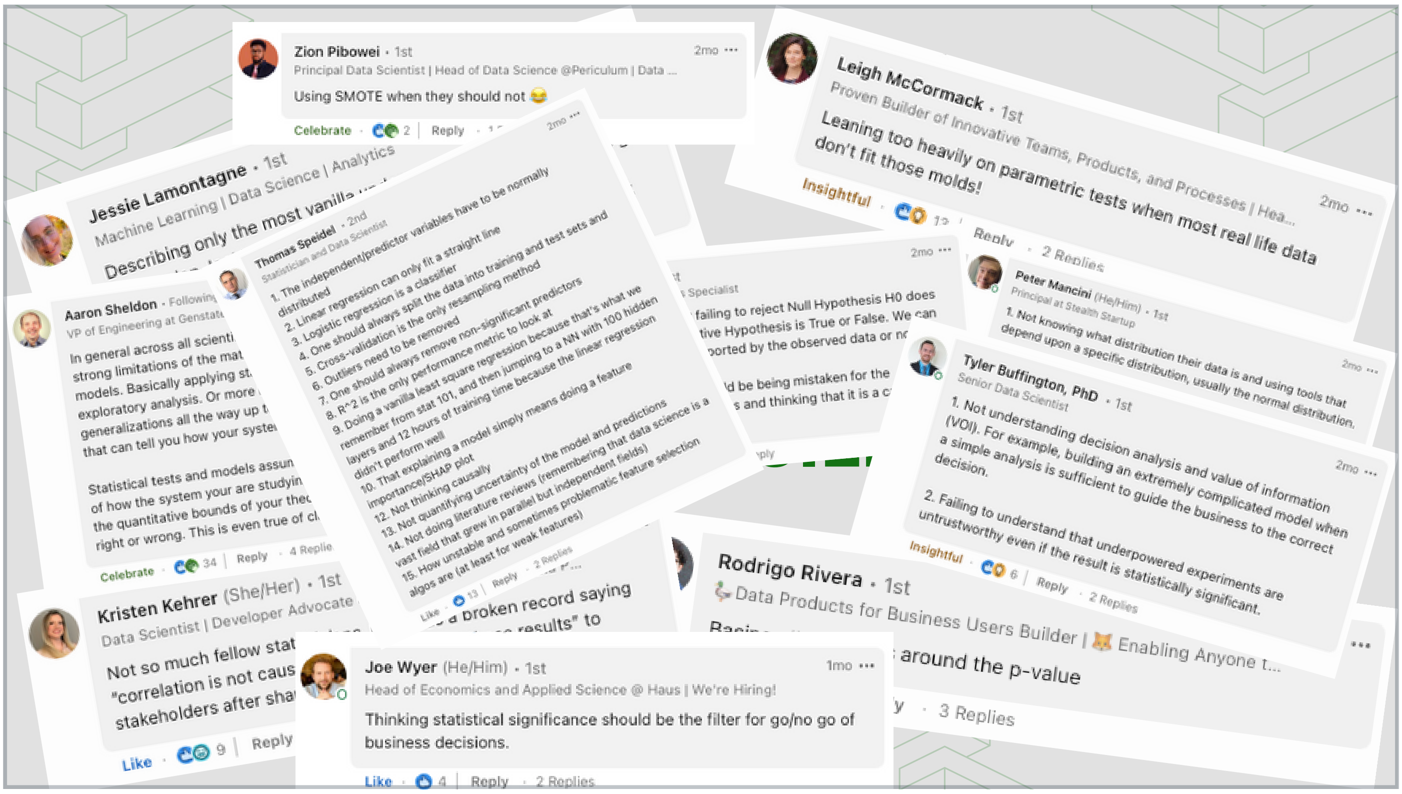
hypothesis testing


Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
Using SMOTE when they should not 😂
Celebrate · 2 | Reply · 1 Reply
Leigh McCormack · 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
Leaning too heavily on parametric tests when most real life data don't fit those molds!
Insightful
2 Replies
Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics
Aaron Sheldon · Following
VP of Engineering at Genstate
2mo (edited)
In general across all scientific fields it is failing to understand the strong limitations of the mathematical assumptions of statistical models. Basically applying statistical testing as a form of exploratory analysis. Or more bluntly assuming regression and its generalizations all the way up to deep learning are a magical oracle that can tell you how your system works.
Statistical tests and models assume a priori that you have a theory of how the system your are studying works. They simply estimate the quantitative bounds of your theory. Not generally whether it is right or wrong. This is even true of classical hypothesis testing.
Celebrate · 34 | Reply · 4 Replies
Hai Tran (He/Him) · 1st
Data Science and Analytics Specialist
2mo
Successfully rejecting or failing to reject Null Hypothesis H0 does not prove that the Alternative Hypothesis is True or False. We can only say whether H1 is supported by the observed data or not.
Another common pitfall should be being mistaken for the correlation between 2 variables and thinking that it is a causation relationship between these.
Celebrate · 4 | Reply · 1 Reply
Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
2mo
1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.
2. Relying on normal distribution tools de facto. Better to rely on tools that will work with the distribution of the data that you have.
Celebrate · 3 | Reply · 1 Reply
Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate @ CometML | The Cool Data P...
2mo
Like · 9 | Reply
Rodrigo Rivera · 1st
Data Products for Business Users Builder | 🦊 Enabling Anyone t...
around the p-value
3 Replies
Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
1mo
Thinking statistical significance should be the filter for go/no go of business decisions.
Like · 4 | Reply · 2 Replies



An underpowered study does not have a sufficiently large sample size to answer the research question of interest

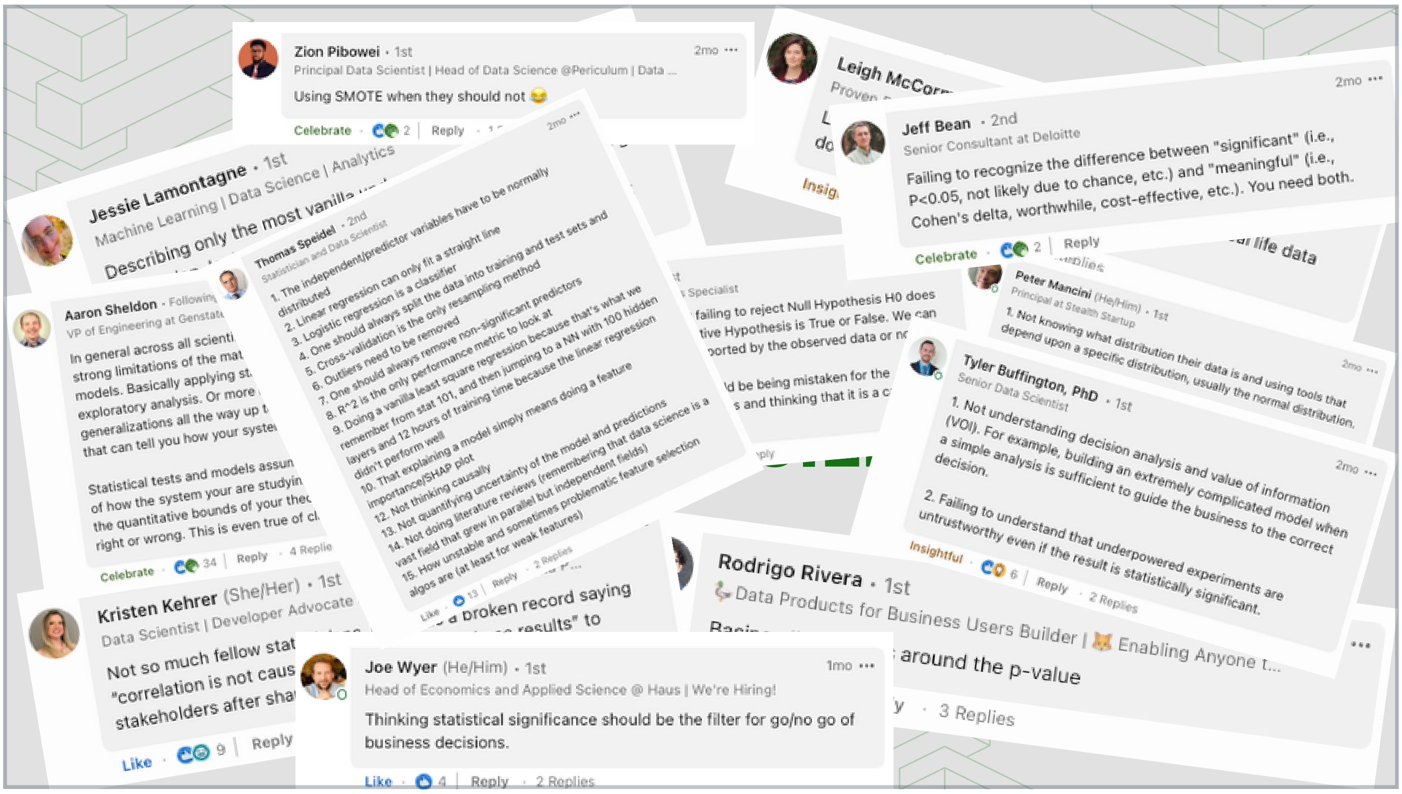
not knowing how to determine a sufficient sample size



outliers dont always need to be removed.

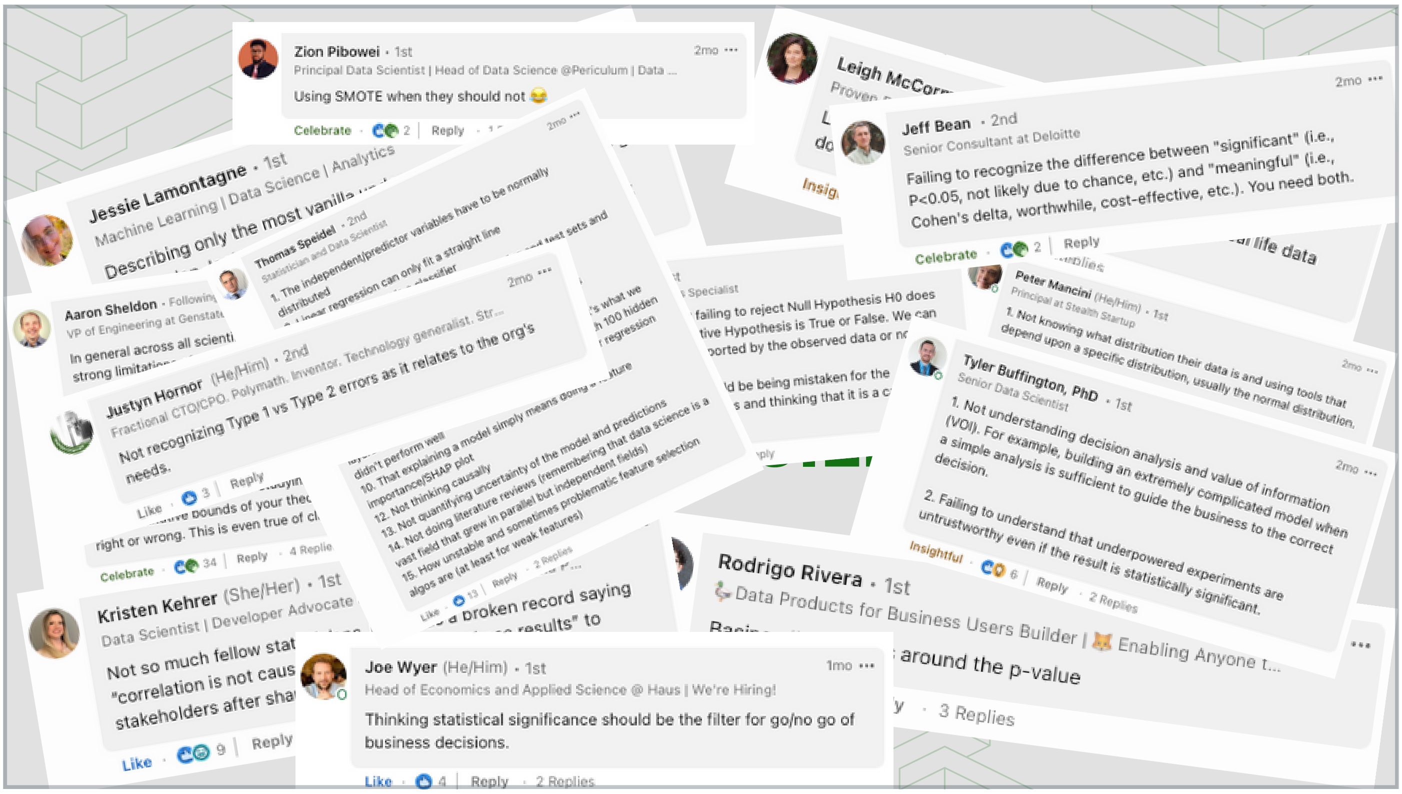
r^2

feature importance doesn't provide model interpretability



there's a huge difference between significant and meaningful

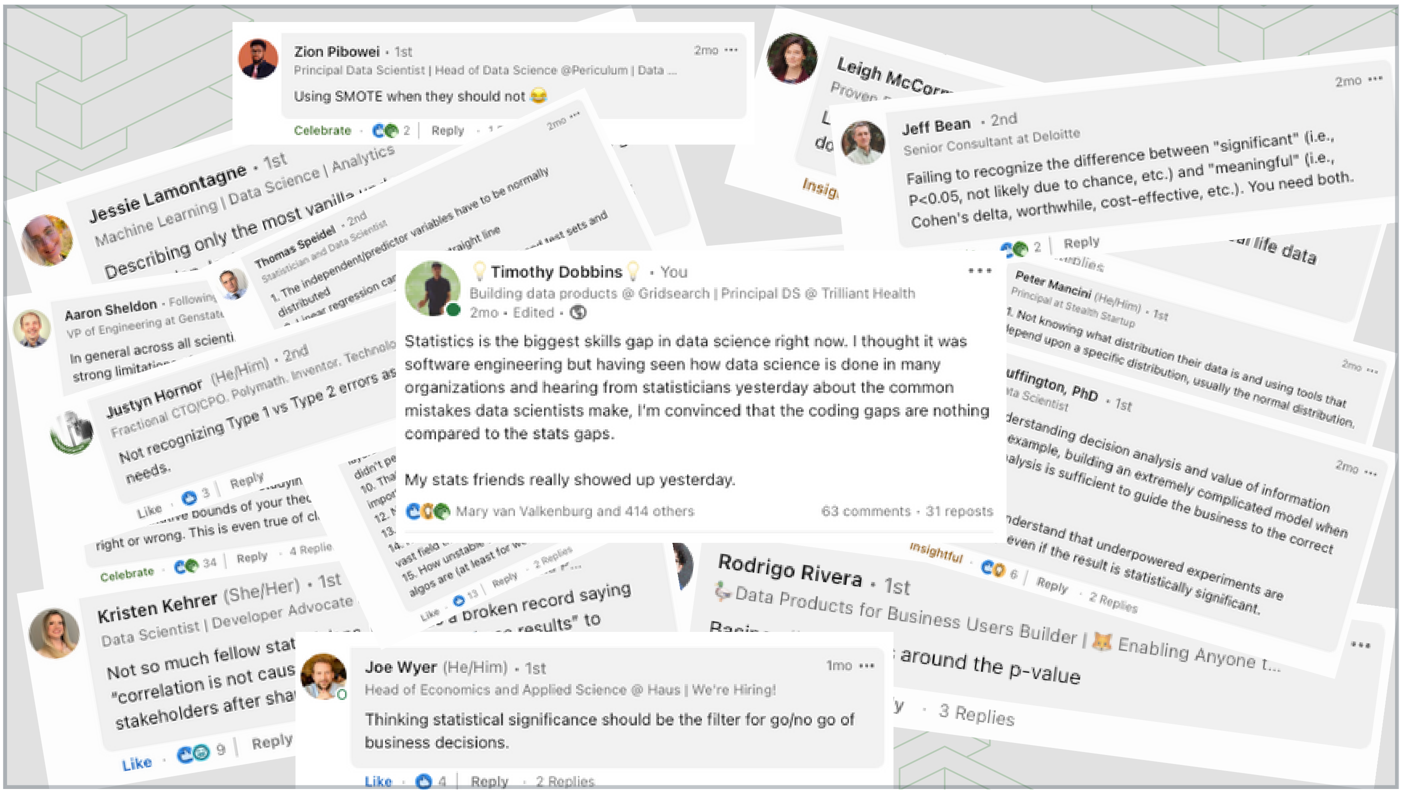
for something to be meaningful, you have to take into account cost, appetite, ability, architecture, constraints, etc



false negatives

false positives

which one to minimize?


Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
Using SMOTE when they should not
Jeff Bean · 2nd
Senior Consultant at Deloitte
Failing to recognize the difference between "significant" (i.e., P<0.05, not likely due to chance, etc.) and "meaningful" (i.e., Cohen's delta, worthwhile, cost-effective, etc.). You need both.
Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics
Timothy Dobbins · You
Building data products @ Gridsearch | Principal DS @ Trilliant Health
2mo · Edited
Statistics is the biggest skills gap in data science right now. I thought it was software engineering but having seen how data science is done in many organizations and hearing from statisticians yesterday about the common mistakes data scientists make, I'm convinced that the coding gaps are nothing compared to the stats gaps.
My stats friends really showed up yesterday.
Mary van Valkenburg and 414 others
63 comments · 31 reposts
Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
Aaron Sheldon
VP of Engineering at Genstat
Justyn Hornor (He/Him) · 2nd
Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate
Rodrigo Rivera · 1st
Data Products for Business Users Builder | Enabling Anyone t...
around the p-value
Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
Thinking statistical significance should be the filter for go/no go of business decisions.

REQUIREMENTS ARE HARD TO DEFINE BECAUSE OF **AMBIGUITY**

SUCCESS IS HARD TO DEFINE BECAUSE OF **UNCERTAINTY**

VALUE IS MIS-MEASURED BECAUSE OF **BAD SCIENCE**

AMBIGUITY OCCURS WHEN WE DON'T KNOW WHAT QUESTIONS TO ASK

UNCERTAINTY OCCURS WHEN WE DON'T KNOW HOW TO VALIDATE THE ANSWERS

BAD SCIENCE OCCURS WHEN WE INCORRECTLY VALIDATE THE ANSWERS.

THANK YOU!

LET'S CONNECT