Chasing The Wrong Benchmarks in ML

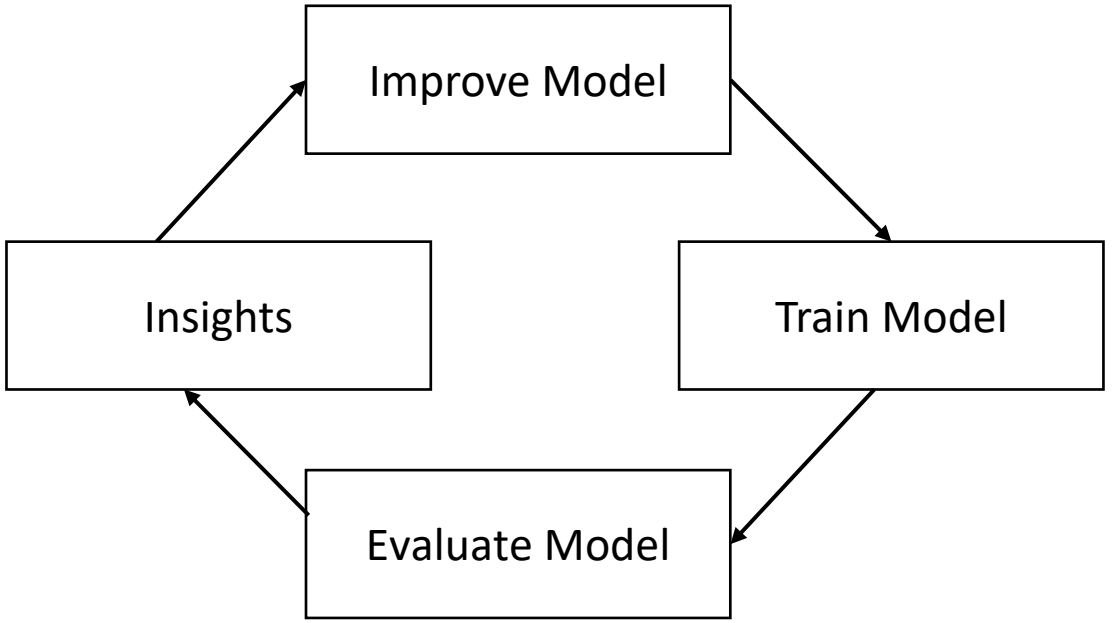
Nils Reimers
HuggingFace
Creator of Sentence-Transformers (www.SBERT.net)

Overview

- Why do we need benchmarks?
- What can happen when we chase the wrong benchmarks?
- Designing better benchmarks

Why Do We Need Benchmarks?

- “You can't improve what you don't measure.”



- What is the best model to solve a given task?

How does Publishing in ML Work?

1) Select a popular benchmark (e.g. CoNLL-2003 NER)

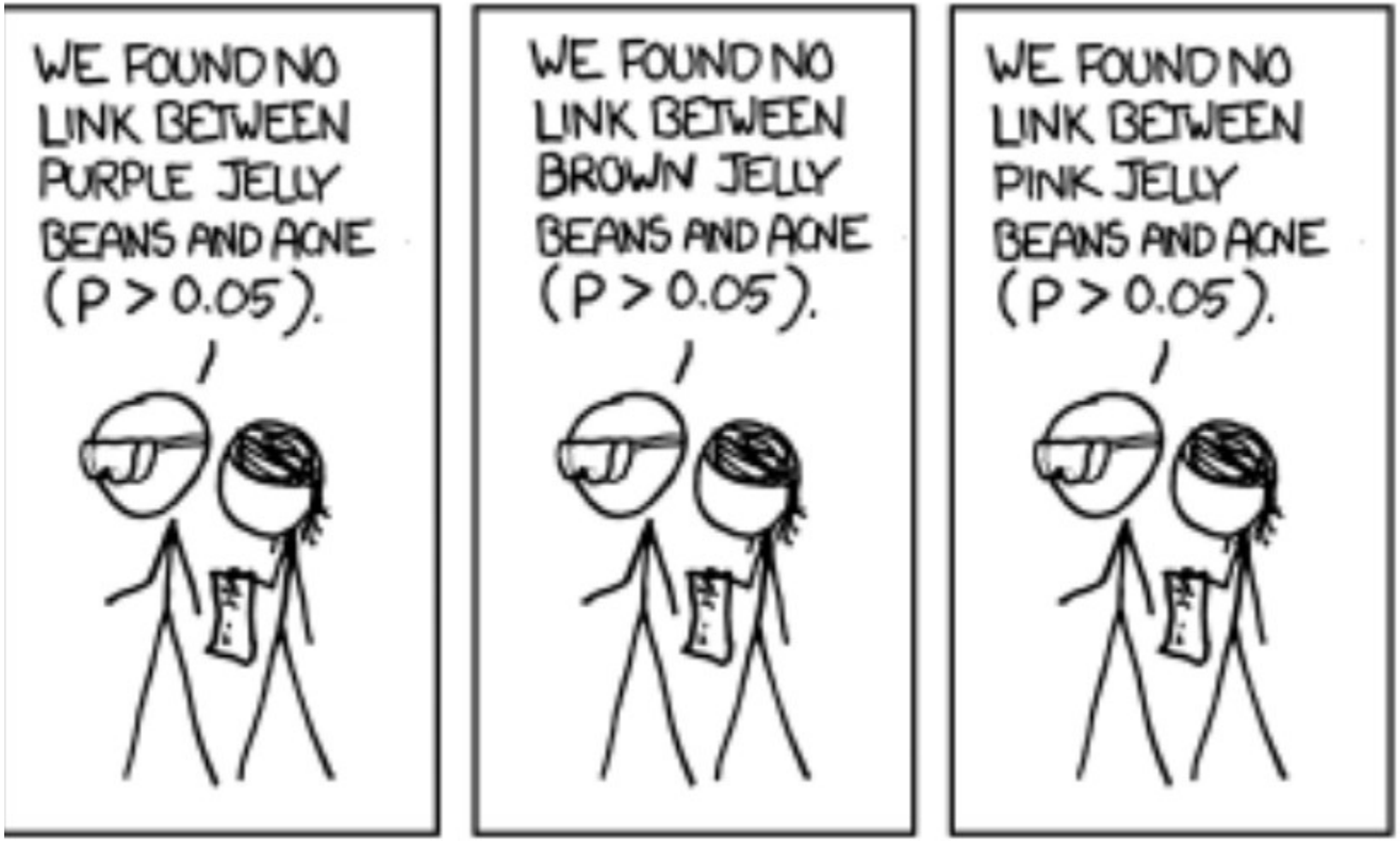
2) Check what is the state-of-the-art (e.g. 94.6 F1)

3) Run experiment
   - Did we improve? No => repeat again

4) Improved performance (e.g. **94.7** F1)
   - Publish paper
   - Mark your system in **bold numbers**

How Significant Are Your Results?

Question: Do Jelly Beans Cause Acne?



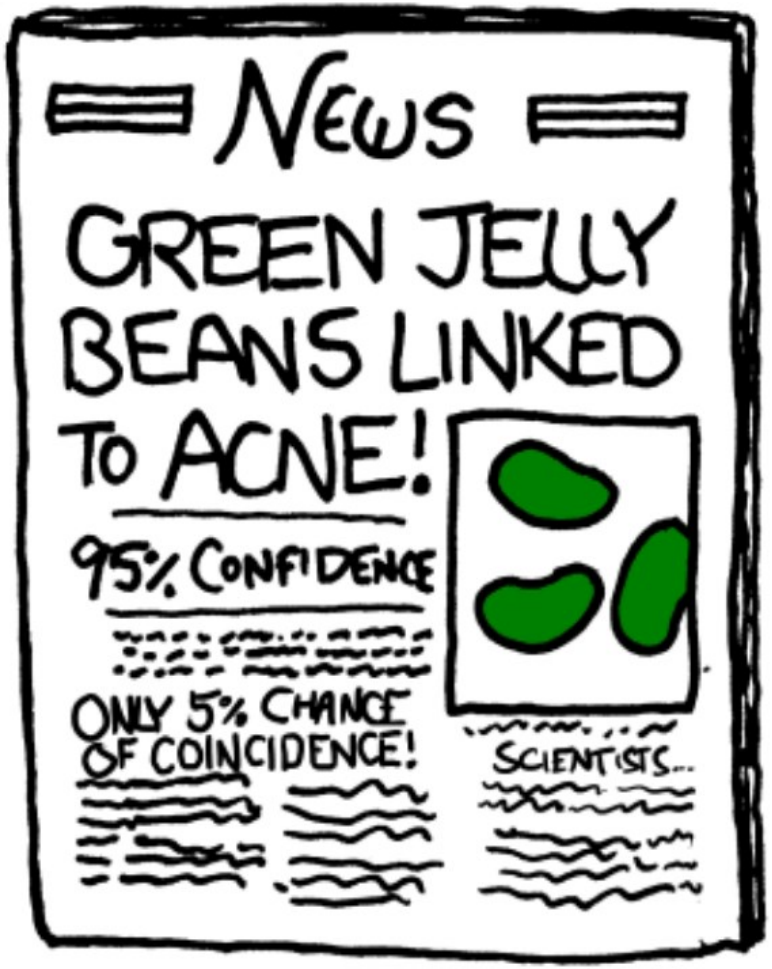
...

Img: https://xkcd.com/882/

How Significant Are Your Results?

Question: Do Jelly Beans Cause Acne?

...



Img: https://xkcd.com/882/

False Positives & Popularity of a Research Field

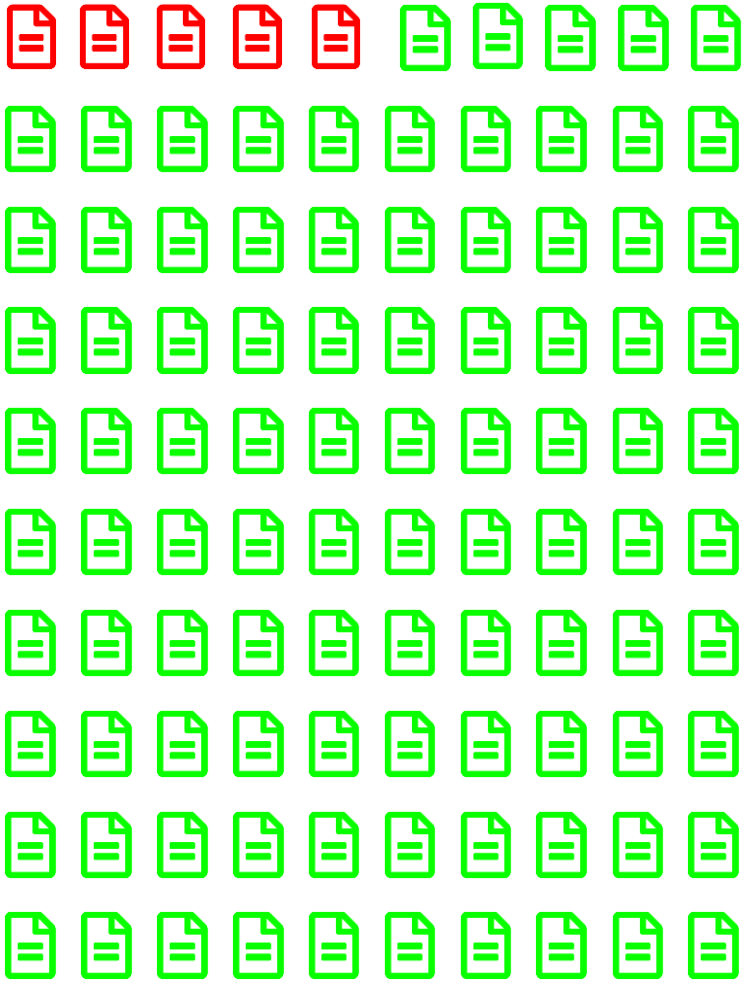
- A really popular field / benchmark (many papers)
  vs
- Unpopular field / benchmark (few papers)

- Which one has the higher scientific quality?
  - Fewer wrong research results?

- The curse of popularity:
  The more popular a field / benchmark => the less relevant the results

How much of the insights are wrong?

- Assume good scientific practices, with significance threshold 5%



The Curse of Popularity

- Assume 1000 hypothesis are tested in parallel
- 10% are true ("improve the model")
  - You can identify them in 80% of the cases
- Significance testing with p=0.05

	Significance Test True	**Significance Test False**
Hypothesis True	100*80% = **80** (True Positive)	20
Hypothesis False	900*5% = **45** (False Positive)	855

45/(45+80)=36% of published papers are wrong

The Curse of Popularity

- The older/more popular a field / benchmark:
  - Harder to find new (significant) insights
  - Major break throughs are found usually early
  - Higher number of experiments done

- Assume:
  - Only 1% of true hypothesis are left
  - 10k experiments
  - => 86% of published papers are wrong

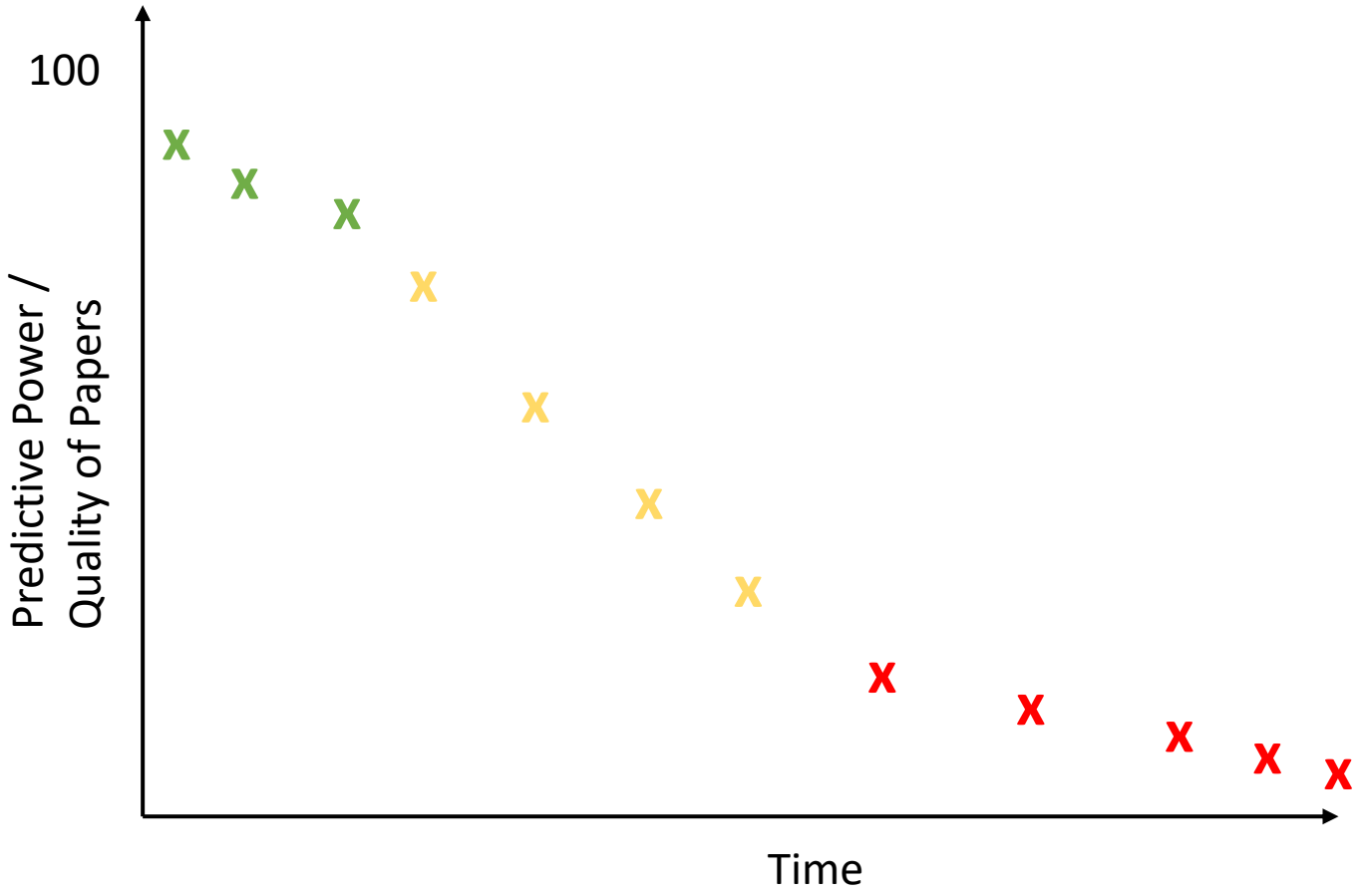
Issues with Reproducibility

- [1] Reproduction of 53 landmark studies in basic science of cancer
  - Only 6 were re-producible (47 were wrong??)
- [2] most published modifications to Transformers network do not improve performance

[1] https://www.economist.com/briefing/2013/10/18/trouble-at-the-lab, [2] https://arxiv.org/pdf/2102.11972.pdf

Predictive Power of a Benchmark

- Better on benchmark ⇔ better system?



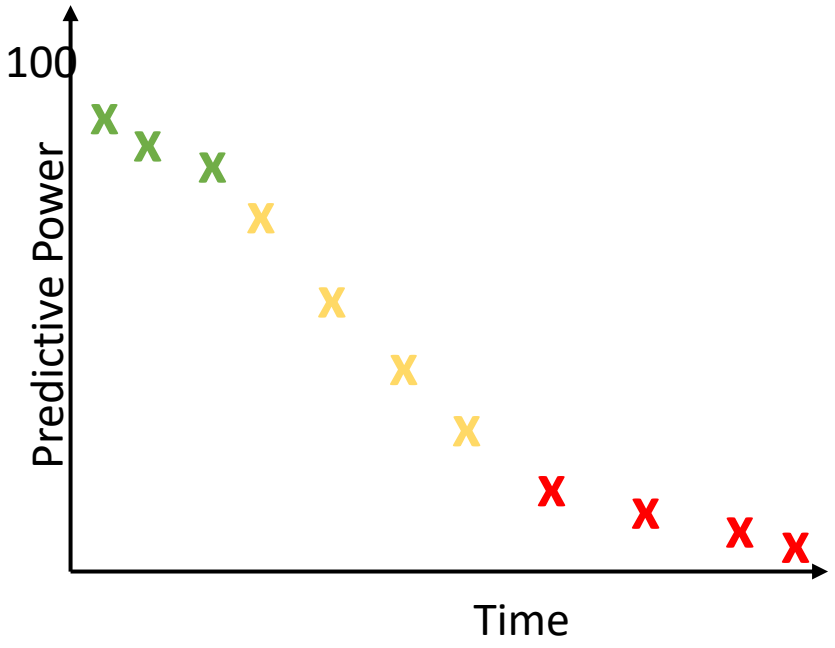
Many popular benchmarks have zero / low predictive power:

- CoNLL 2003
- GLUE
- STS
- MS MARCO
- SuperGLUE (?)

These are still heavily used

Disclaimer: My claim without a full proof

Why do Benchmarks loose their predictive power?



- Most impactful insights are discovered early
- Increasing number of experiments
- Change of language / data drift

- Systems are too good
  - Benchmarks treat the world as black & white (e.g. positive or negative sentiment)
  - But nearly all tasks are ambiguous to a certain degree
  - What is the human upper bound?
  - Performance beyond that: Fitting to annotation bias / errors
- Many measures are imperfect
  - Word overlap to evaluate machine translation
  - Can distinguish bad from good systems
  - Cannot distinguish between two really good systems

Do We Chase the Right Benchmarks?

A question that is seldom asked

OpenAI GPT-3 Embeddings

- In January, OpenAI publishes a GPT-3 embedding endpoint:
  - “new state-of-the-art results”
  - “impressive semantic search capabilities”

- Text embeddings are useful for many tasks:
  - Semantic Search
  - Clustering
  - De-duplication….
- Text are mapped to a dense vector space

- But how good are these embeddings really?

How OpenAI Evaluated

- Sentence Embeddings evaluated on SentEval

- SentEval benchmarks a tiny use-case for embeddings
  - Not a suitable benchmarks for more common tasks (search, clustering)
  - Most recent papers start to ignore this benchmark
    - Easier for OpenAI to achieve state-of-the-art

Evaluating on Good Benchmark

Model	Costs encoding 1M docs	Performance
OpenAI GPT-3 Embedding Models		
text-similarity-ada	$800	61.86
text-similarity-babbage	$1,200	62.62
text-similarity-curie	$6,000	62.39
text-similarity-davinci	$60,000	58.11
Google Embedding Models		
st5-base-1	$0.70	67.84
st5-large-1	$2.40	68.74
st5-3b-1	$6.80	69.23
universal-sentence-encoder-large-5	$0.35	64.51
Sentence-Transformers Model		
all-MiniLM-L6-v2	$0.12	68.06
all-MiniLM-L12-v1	$0.25	68.83
all-mpnet-base-v1	$0.70	69.98
all-roberta-large-v1	$2.40	70.23

- Evaluated on 14 diverse tasks for sentence embedding tasks
- Performance is worse than models from 2018

https://medium.com/@nils_reimers/openai-gpt-3-text-embeddings-really-a-new-state-of-the-art-in-dense-text-embeddings-6571fe3ec9d9

Evaluating for Semantic Search

Model	Performance on 11 IR datasets (nDCG@10)	Cost to encode 6M Wikipedia articles
OpenAI Models		
cpt-text S (Ada)	49.0	$17,000
cpt-text M (Babbage)	50.5	$25,000
cpt-text L (Curie)	50.9	$126,000
cpt-text XL (Davinci)	52.8	$1,260,000
Freely available models		
SpladeV2	52.7	$3

- For an 0.1 improvement
  - 400,000 times higher costs
  - 3000 times higher latency
  - 20 times more memory needed

https://medium.com/@nils_reimers/openai-gpt-3-text-embeddings-really-a-new-state-of-the-art-in-dense-text-embeddings-6571fe3ec9d9

Zillows $880 million loss debacle

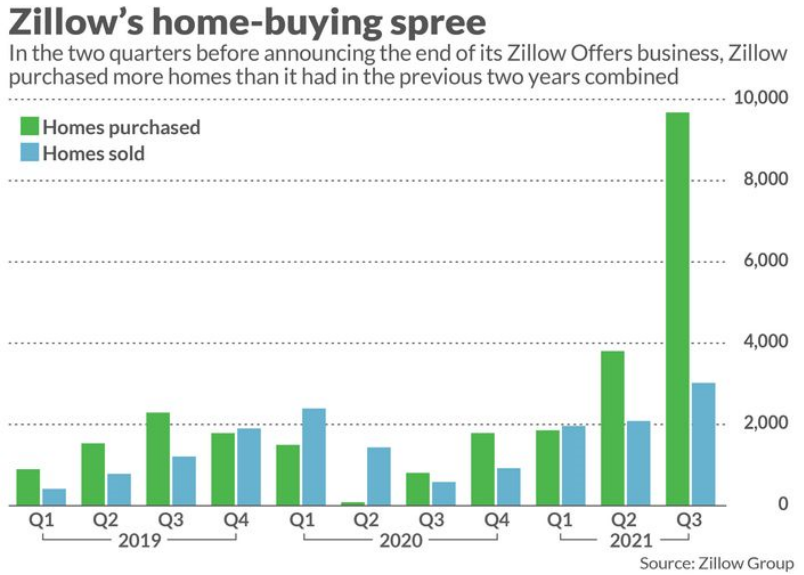
- Zestimate
  - Predict selling price for houses
  - Median absolute error below 5% of final sale price
- House flipping idea:
  - Selling a house is a slow & annoying process
  - Selling to Zillow is easy: Fill-in the form, get the offer
  - Zillow renovates & sells the house
  - Critical for Zillow: Don’t buy too expensive

- Zillow has lost $881 million in 2021 on this business
  - Bad model evaluation

https://www.wsj.com/articles/zillows-shuttered-home-flipping-business-lost-881-million-in-2021-11644529656

Why did Zillow loose money?

- Strategy tested over 3 years
  - Strategy was successful, but:
  - It was bull market - house prices quickly increased
  - Making money in a bull market is easy



- But then the housing mark cooled down in 2021
  - Making money in a side / bear market is a lot harder
  - Strategy no longer worked

Why did Zillow loose money?

- Adversarial market
  - On average, their price predictions were correct
  - Who sells to Zillow is not “the average”

-  Hidden Gem
  - Your house is great
  - Zestimate is far blow market value
  - You don’t sell to Zillow ✕

-  Houses with issues
  - Bad smell, noise, plumbing issue, …
  - Not reflected by Zestimate
  - Zestimate is higher than market value
  - You sell to Zillow ☺

How to Design Better Benchmarks?

Creating a Benchmark is Challenging

- Quality of a benchmark is extremely important
  - **Key component** to make progress

- Creating a benchmark is more challenging than improving models
  - Creation of benchmarks not well recognized
  - People often think: take some datasets -> select a measure -> done
  - A good benchmark must be critically iterated many times

- Spend a **large fraction** of your time on designing the benchmark

A Common Observation in Data Science Teams

-  Business Team
  - Here is our Excel file
  - As input we have columns A-D
  - The label in in column E
  - Build us a model!

-  Data Science Team
  - Great, let’s start & measure accuracy

- ❌ This is wrong on so many levels

A Good Benchmark

- 💻 What is the intended use-case?
  - Predicting a label, ranking of results, …
- 💰 Costs of Errors
  - Research treats all errors often with equal costs
  - In production this is seldom the case
- 🤷 Human upper bound
  - How good are humans in this task?
  - Business team doesn't have this info
  - Data science team cannot estimate it & doesn't want to ask business team
  - When creating a new dataset: Spend many cycles to improve human agreement
- 🚓 What else is important?
  - Inference speed
  - Robustness

A Good Benchmark

- 🦖 A benchmark must evolve
  - As models evolve, our benchmarks must evolve!
  - Stop using outdated benchmarks

- ⏳ Restrict number of submissions
  - The more experiments we run on a benchmark, the less likely we can trust the numbers
  - Only allow evaluation on test set in very rare cases!
  - Have a dev dataset for model development
  - If possible: use an “out-of-domain” test dataset

- ⏱ Temporal split
  - Test data should be the most recent, train data the oldest

A Good Benchmark

- Diversity
  - Don’t test only on one task / domain etc.

- Look for biases
  - What biases does your dataset have?
  - What biases does your benchmark has?
  - E.g. GLUE Benchmark:
    - Mostly sentence tasks
    - 7 out of 8 are sentence pair comparison tasks
    - Transformers networks are really strong on these tasks
    - See benchmark lottery: https://arxiv.org/abs/2107.07002

Summary

- Please think more about your benchmarks!
- Most research findings are irrelevant due to bad benchmarks
  - Quality of benchmark degrades with its popularity
- Ask frequently: Is this benchmark still good?
- We need constantly evolving benchmarks