Planning E-commerce search relevance work

MICES

June, 2024

Obligatory Bio Slide

👋🏻 Hi I’m Doug
(@softwaredoug everywhere)

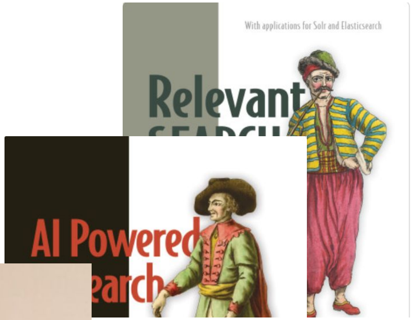
Long-time search enthusiast... Not yet (never?) an expert

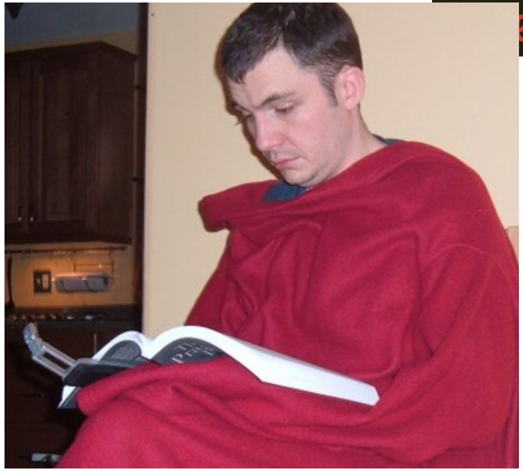
I wrote some search books, did some open source

I work at Reddit

I worked at Shopify & OpenSource Connections in search

I blog here: http://softwaredoug.com





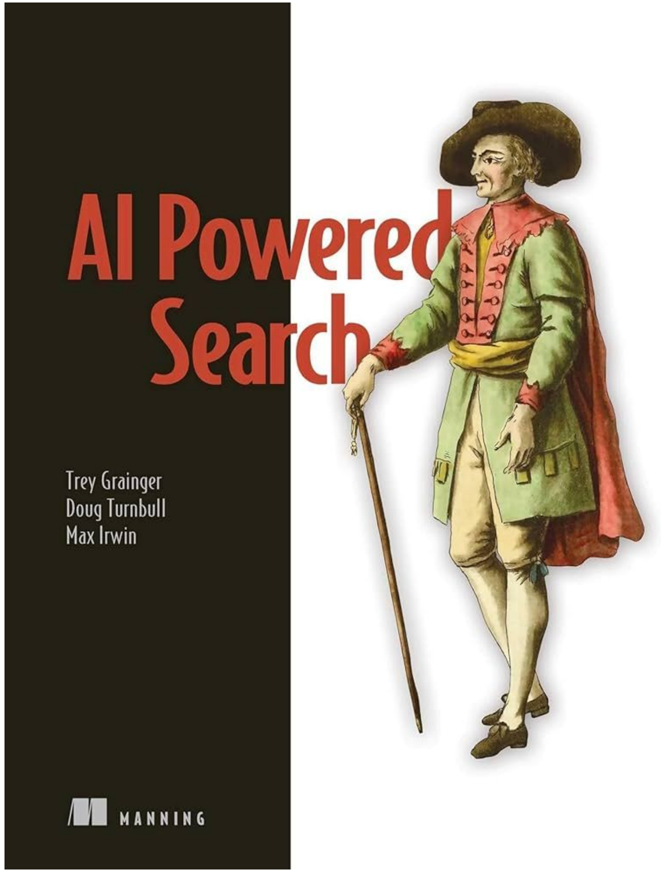
:itme:

AI Powered Search

https://aipoweredsearch.com/

Being published soon!

Join the community!



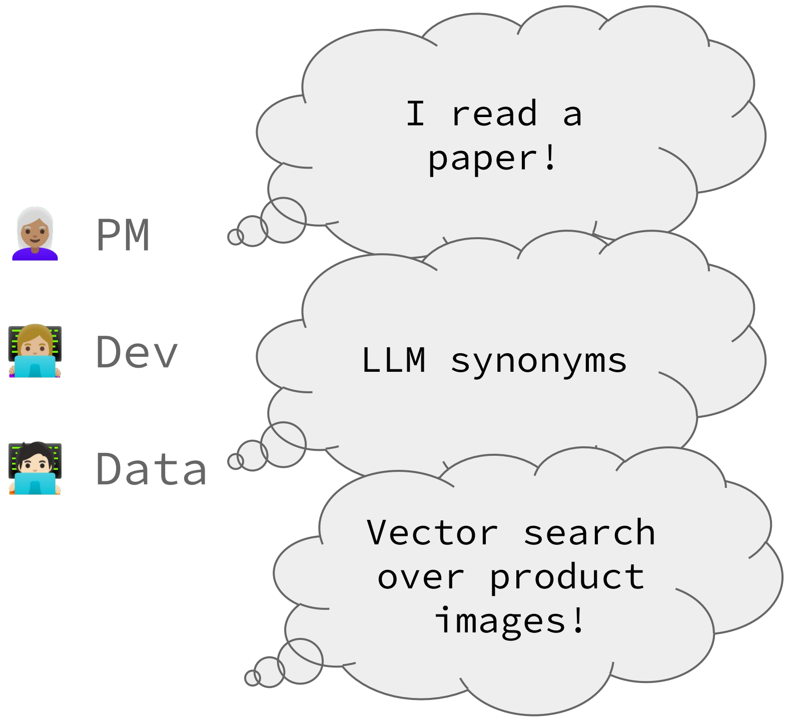
reddit

Who’s this talk for?

- Talk is good for search teams collaborating across DS, ML, Eng, PM
- Especially a DS trying to understand Eng and vice-versa
- Anyone trying to see the future, without needing to do the work!

The current planning chaos

Every quarter, so many options!



My gut* says

Q1 Plan	Launch Date
LLM Synonyms	Next Quarter!

* Because my CEO, Director, etc likes “AI”

Sometime later... 📉

PM

Dev

Data

I read a paper!

LLMs for synonyms!

Vector search over product images!

We pull ideas out of a hat!

..A/B test months later...

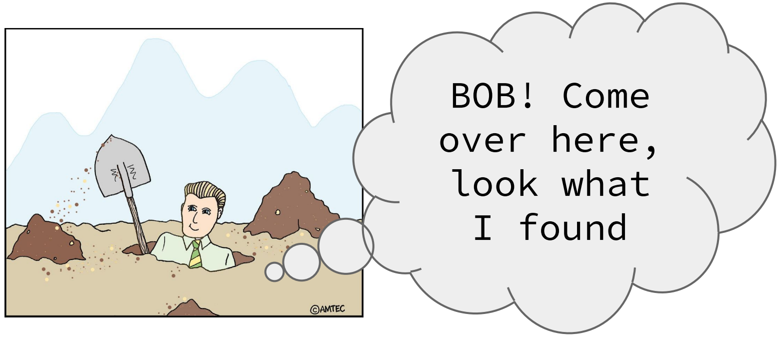
Sunk cost trap

I can make this work!
©AMTEC

...even after failure...
We keep digging that hole to save face

(image from Amtec Photos)
https://www.flickr.com/photos/141761303@N08/36547675260/in/photostream/

Opportunity cost



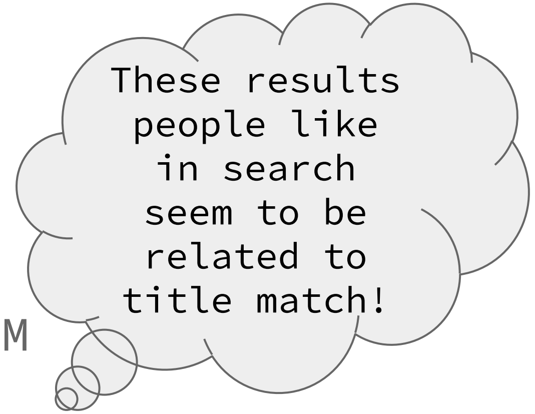
(image from Amtec Photos)
https://www.flickr.com/photos/141761303@N08/36547675260/in/photostream/

Opportunity cost

(image from Amtec Photos)
https://www.flickr.com/photos/141761303@N08/36547675260/in/photostream/

Usual Solutions

Simple correlation studies


These results people like in search seem to be related to title match!

👩🏽‍🦳 PM

👩🏼‍💻 Dev

🧑🏻‍💻 Data

Query	Click-Thru Rate	Add To Cart Rate
shoes	0.09	0.05
running shoes	0.11	0.04
…	…	…
ipad	0.21	0.11

Correlation != Causation

Gap analysis

What's wrong with the weakest queries?

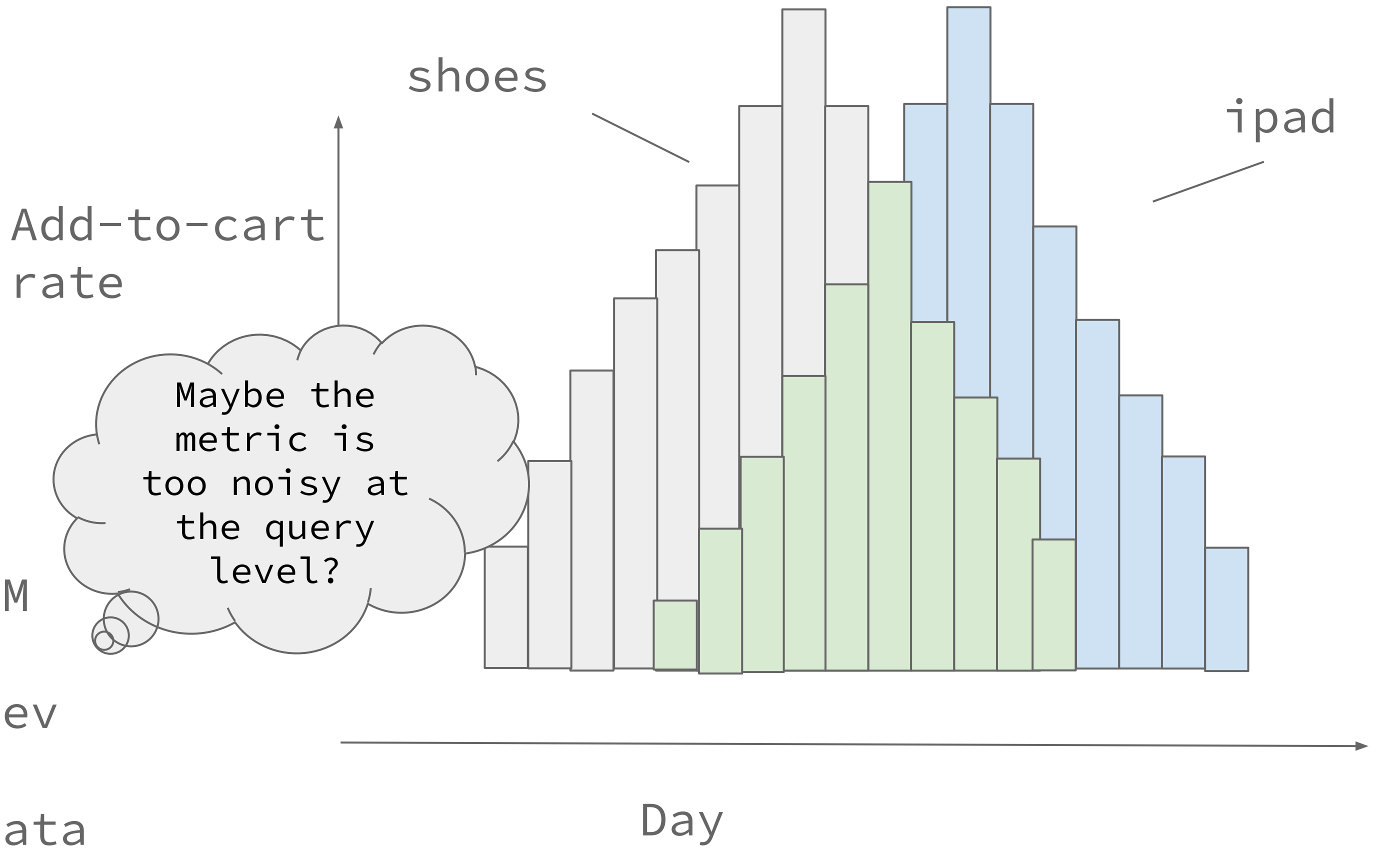
👩🏽‍🦳 PM

👩🏼‍💻 Dev

🧑🏻‍💻 Data

Query	Click-Thru Rate	Add To Cart Rate
shoes	0.09	0.05
running shoes	0.11	0.04
…	…	…
ipad	0.21	0.11

Beware gap analysis on noisy data



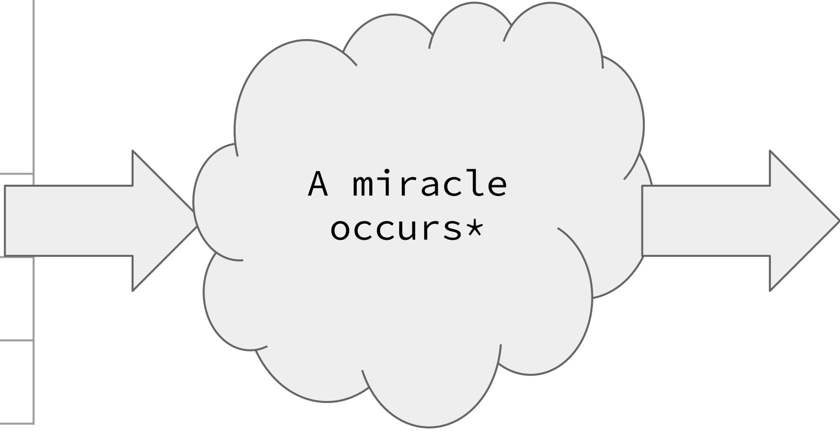
👩🏽‍🦳 PM

👩🏼‍💻 Dev

🧑🏻‍💻 Data

We can make a judgment list for eval

Query	Product Id	30 day num clicks	30 day total impressions	30 day add to carts
shoes	1234	34	150	2
shoes	5678	32	110	4
shoes	8989	5	400	0



* References: Click models for web search
Rene Kriegler Haystack
AI Powered Search CH 11+12

Judgment List

Query	Product Id	Grade (0-1)
shoes	1234	0.7
shoes	5678	0.8
shoes	8989	0.1

Relevance of product for query

Judgment List… to eval search

Query	Product Id	Grade (0-1)
shoes	1234	0.7
shoes	5678	0.8
shoes	8989	0.1

👩‍⚖️ Which search for ‘shoes’ is better?

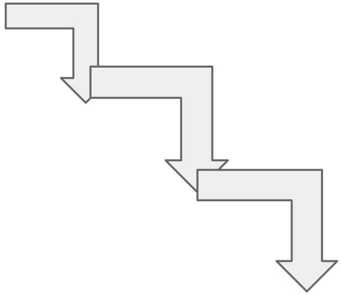
Rank	Product Id	Grade (0-1)
1	5678	0.8
2	8989	0.1
3	1234	0.7

Solution 1 ranking of ‘shoes’

Rank	Product Id	Grade (0-1)
1	8989	0.1
2	5678	0.8
3	1234	0.7

Solution 2 ranking of ‘shoes’

Now we could check an idea before launch



jupyter

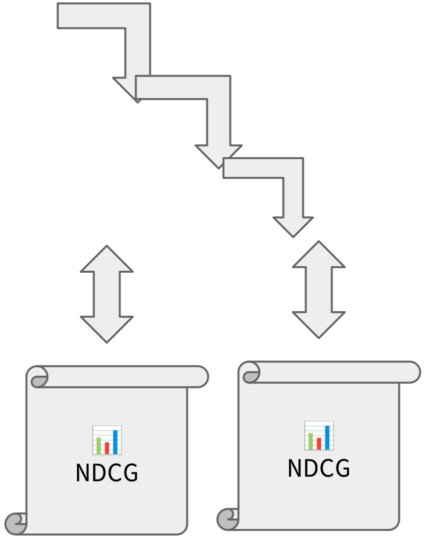
📊NDCG, etc

PM gives 👍 to launch

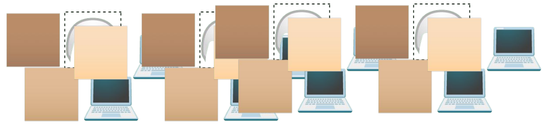
SHIP IT!

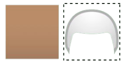
Offline testing during…

Smarter teams offline test while building



…DS+Dev fine tune…





Keep going? 🤙🏽

What we want!!

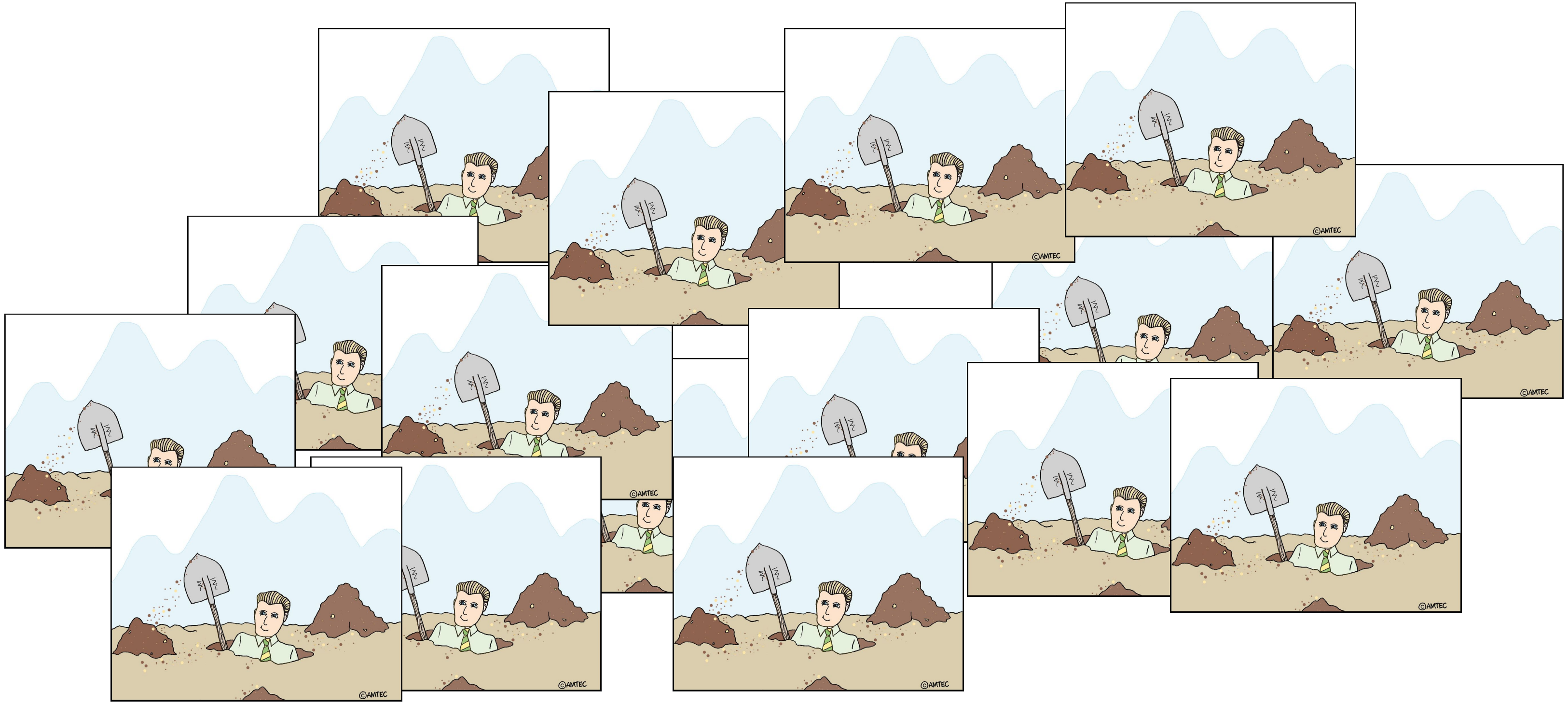
… We want to Measure BEFORE committing to project…

- 👩🏽‍🦳 PM
- 👩🏼‍💻 Dev
- 🧑🏻‍💻 Data

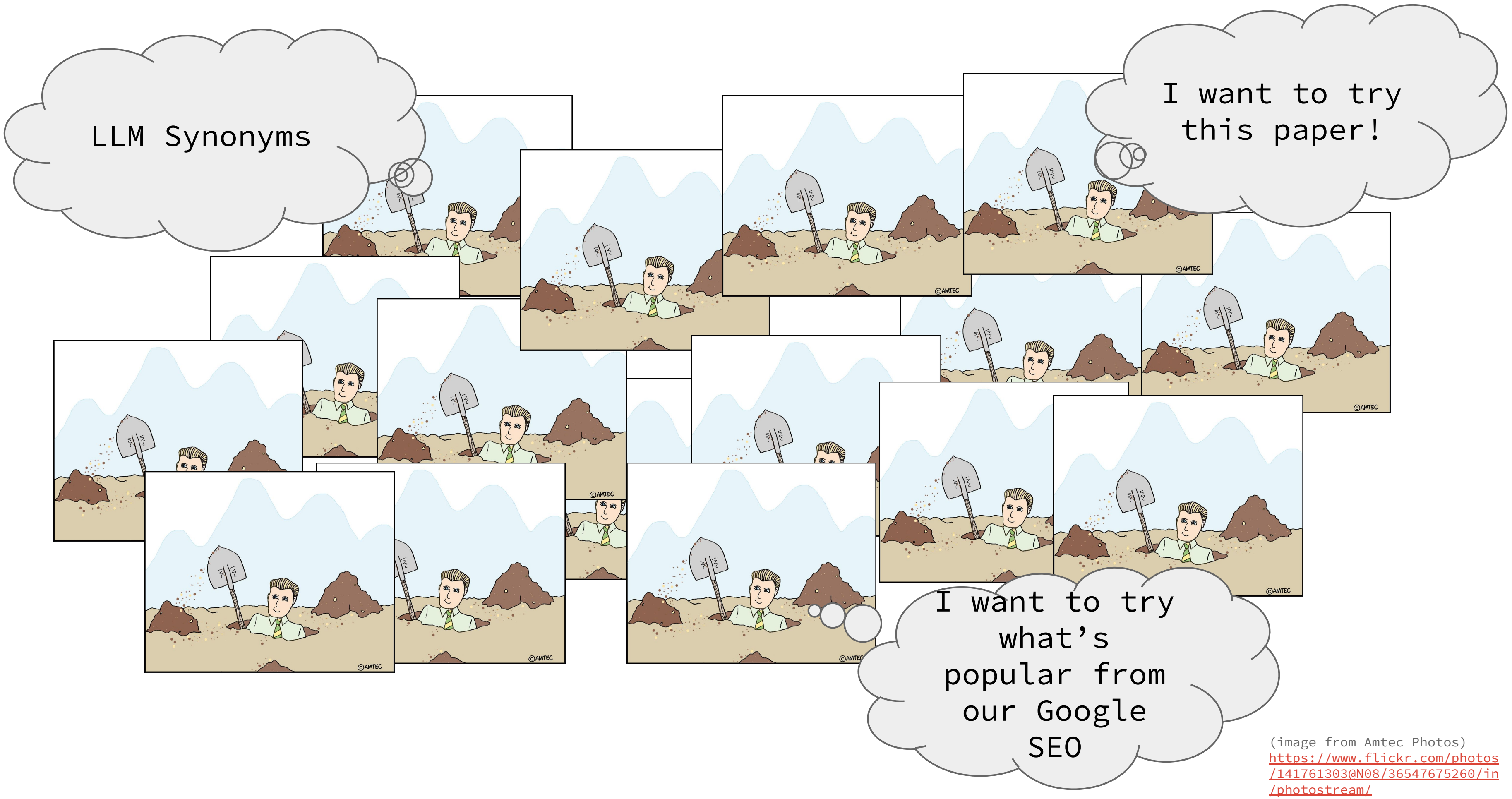
We're all asking ourselves:

- How do we justify a change before planning?
- How do we predict outcomes?
- How do we manage stakeholder expectations?

We want to try ALL THE THINGS!

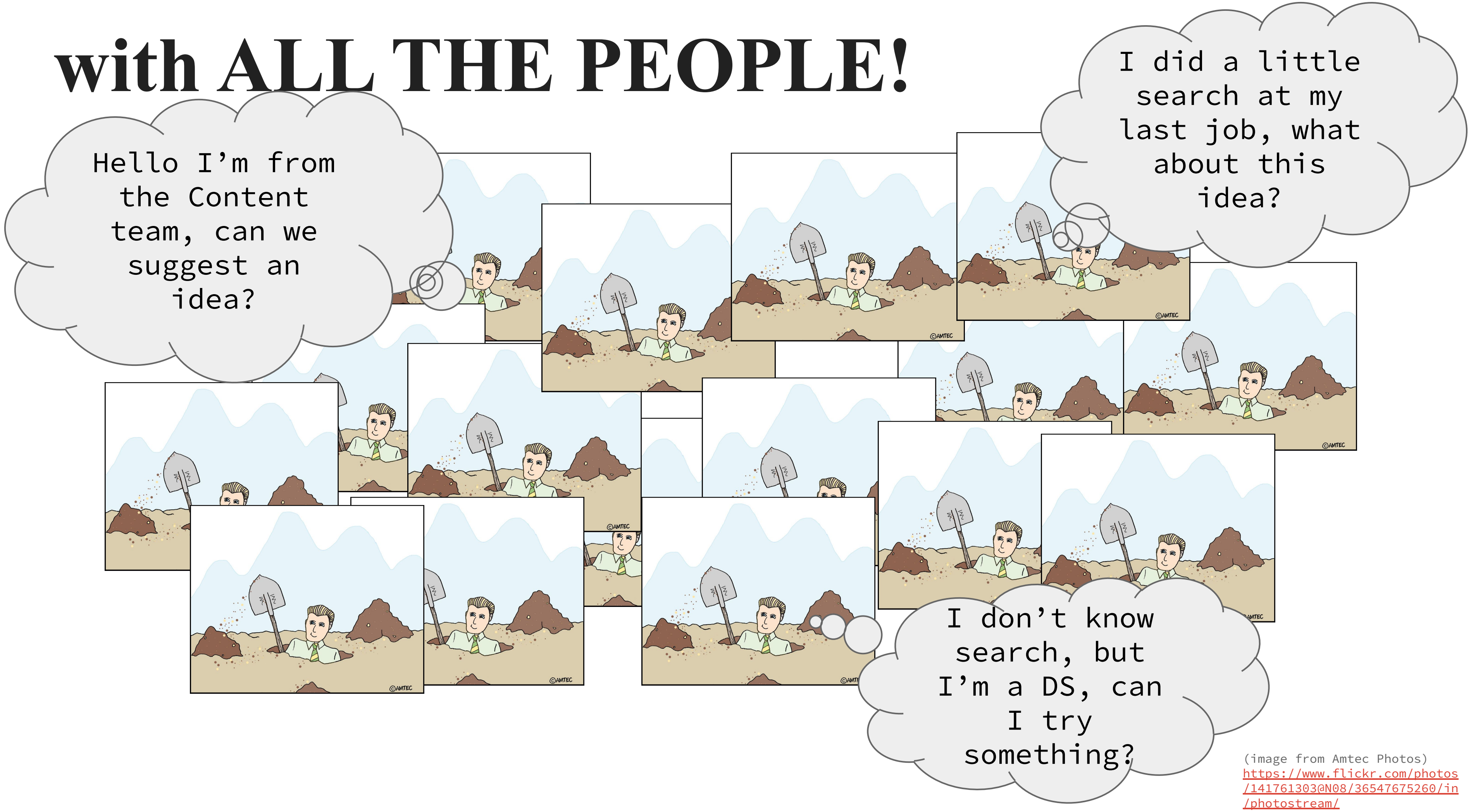


(image from Amtec Photos)
https://www.flickr.com/photos/141761303@N08/36547675260/in/photostream/


LLM Synonyms
I want to try this paper!
I want to try what's popular from our Google SEO

(image from Amtec Photos)
https://www.flickr.com/photos/141761303@N08/36547675260/in/photostream/

with ALL THE PEOPLE!


Hello I’m from the Content team, can we suggest an idea?
I did a little search at my last job, what about this idea?
I don’t know search, but I’m a DS, can I try something?
©AMTEC

(image from Amtec Photos)
https://www.flickr.com/photos/141761303@N08/36547675260/in/photostream/

We want shallow / fast feedback not slow / accurate

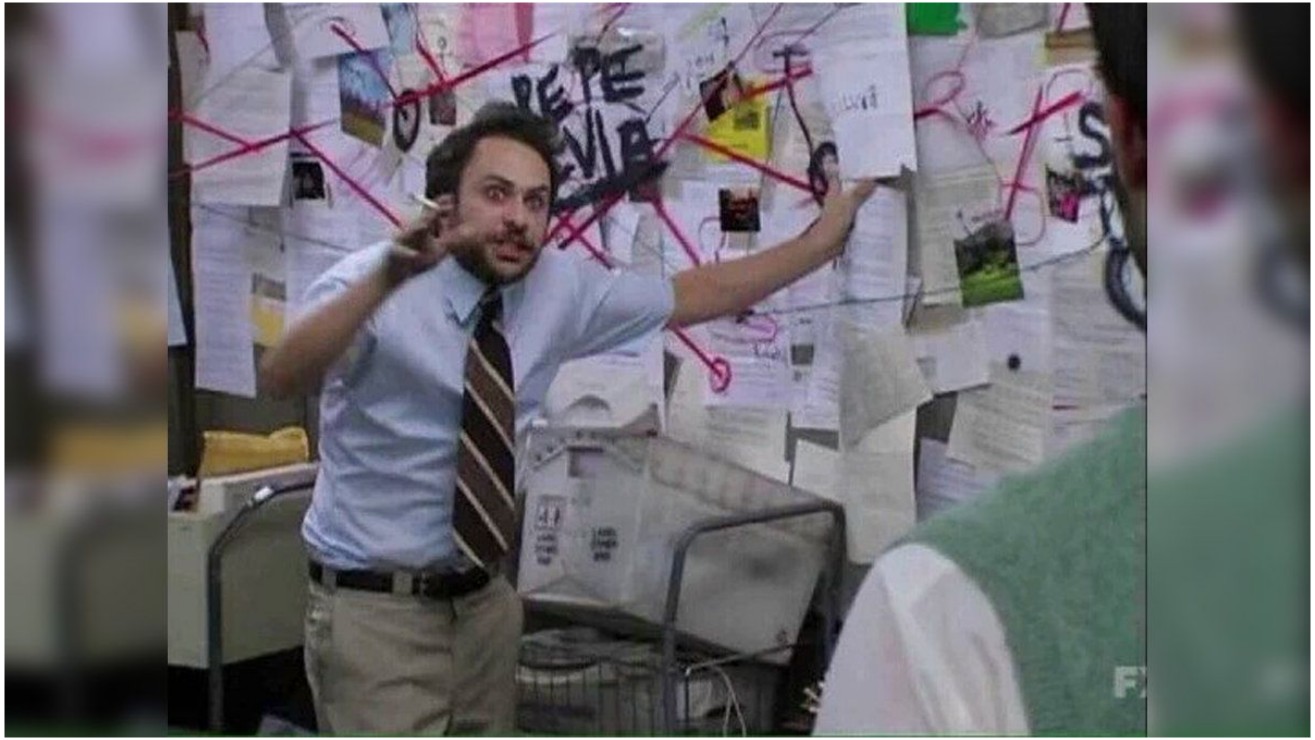
We need to get an idea ***fast*** if an idea is even worth trying.

NOT a super good/accurate methodology

4 hrs, not 4 weeks

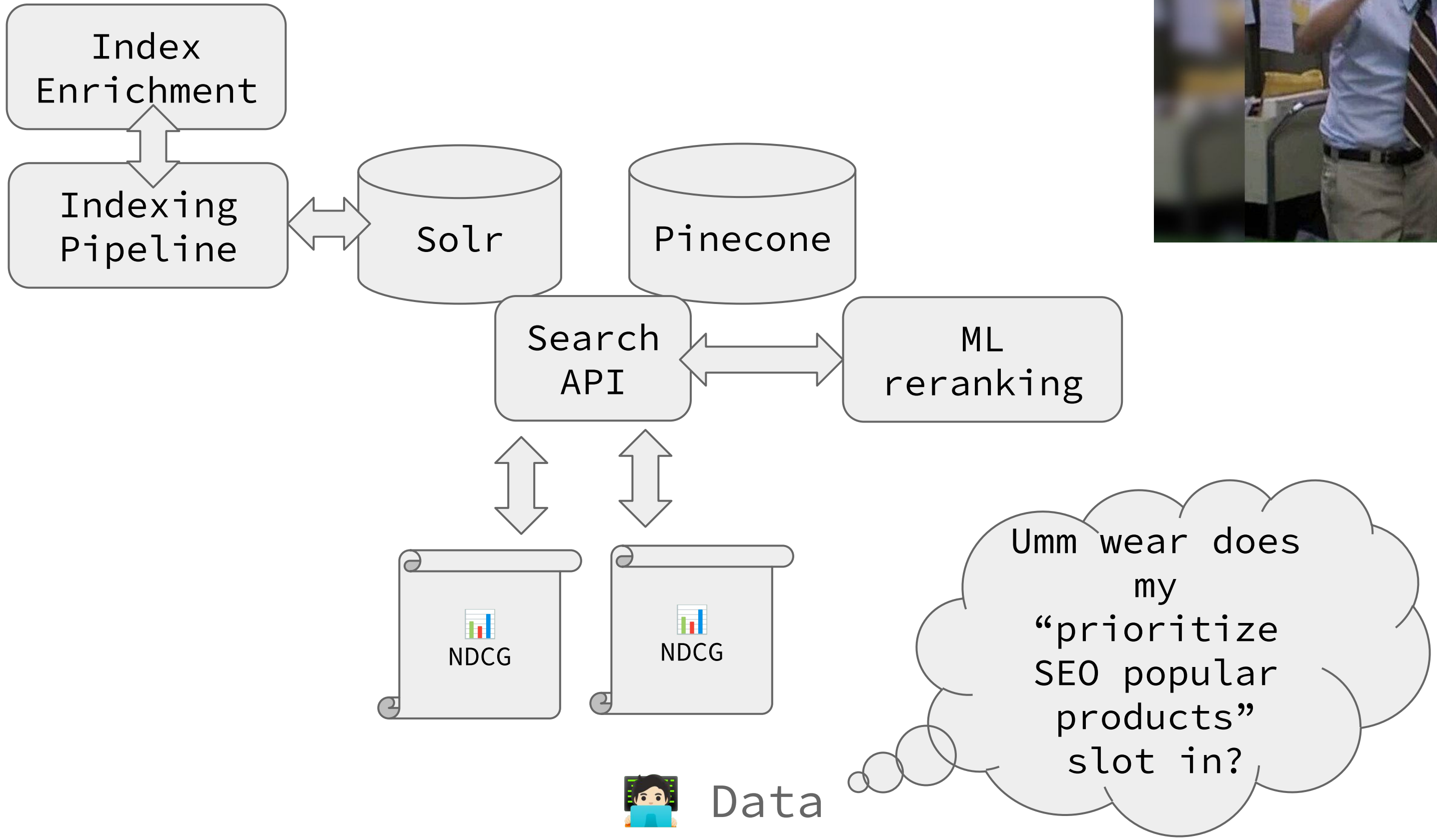
(image from Amtec Photos) https://www.flickr.com/photos/141761303@N08/36547675260/in/photostream/

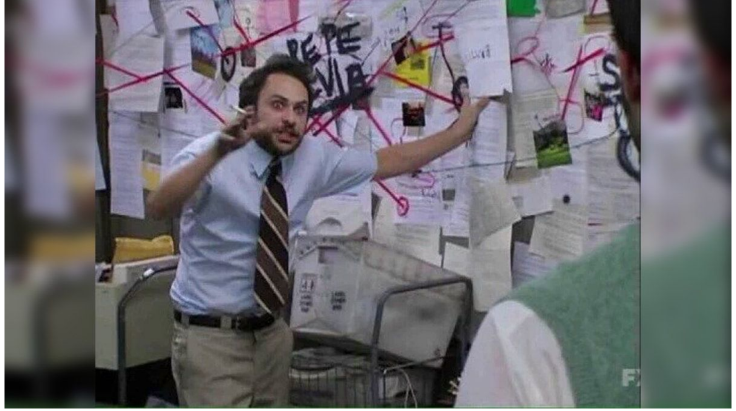
But there’s a problem

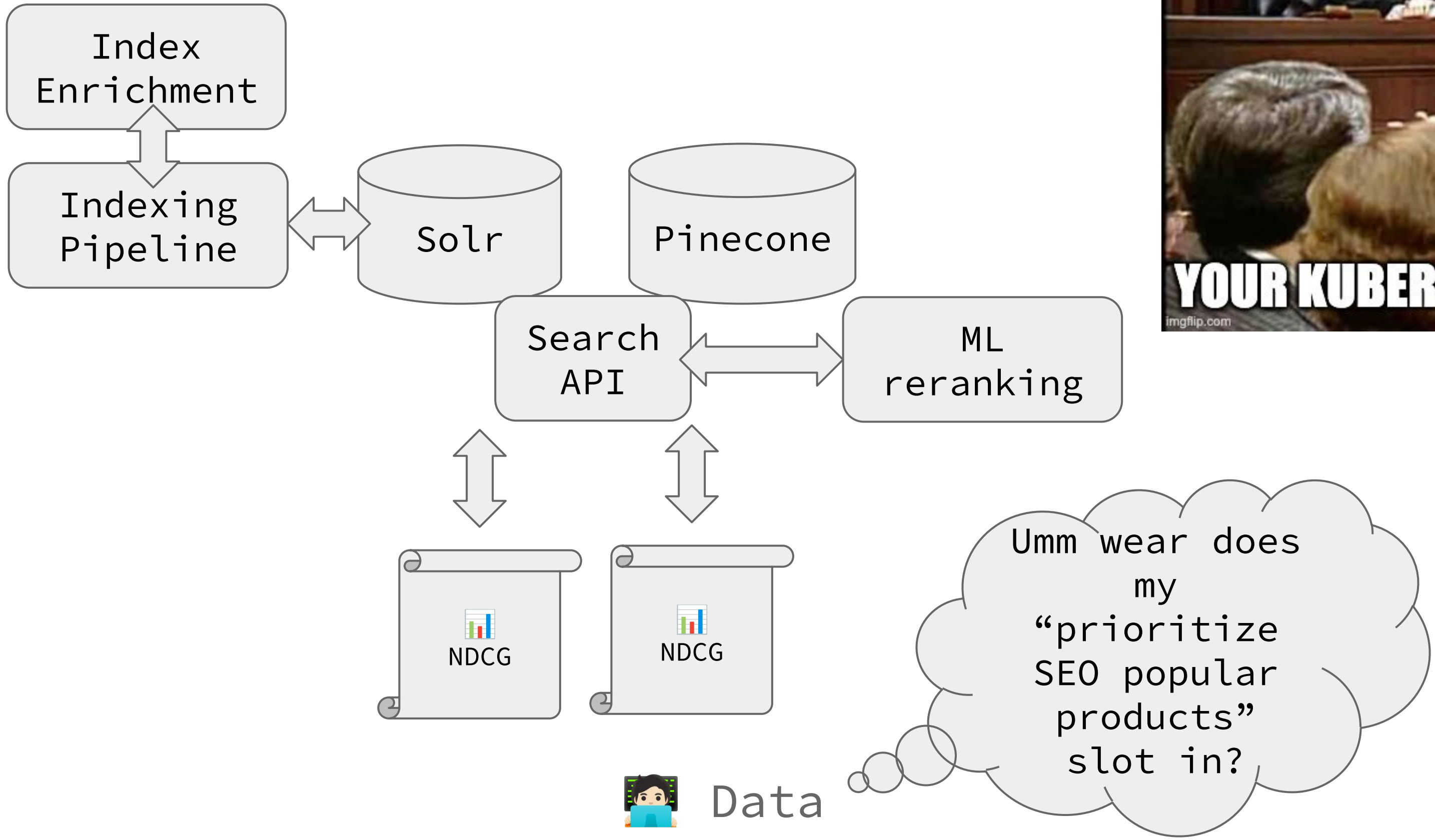


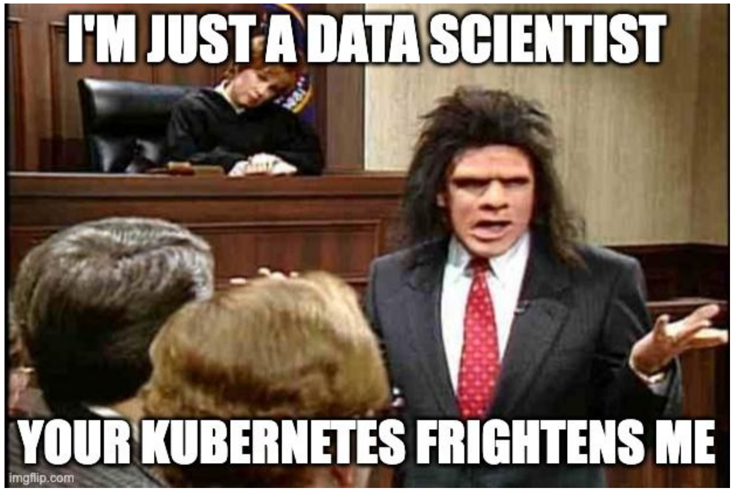
Our engineer describing how search actually works

The search system

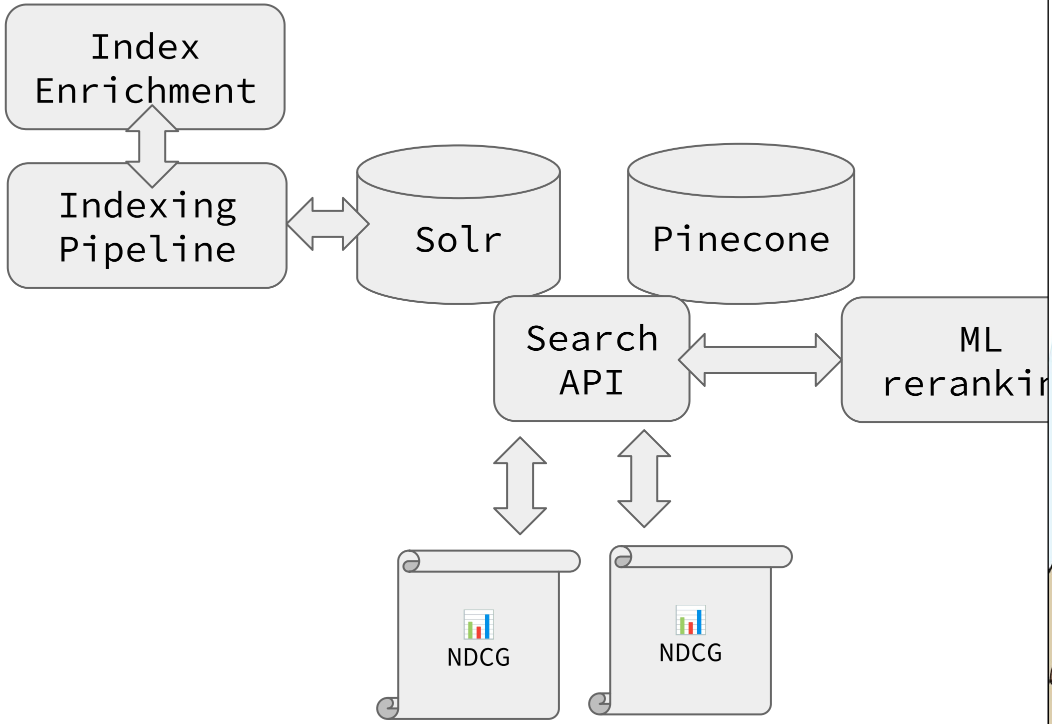


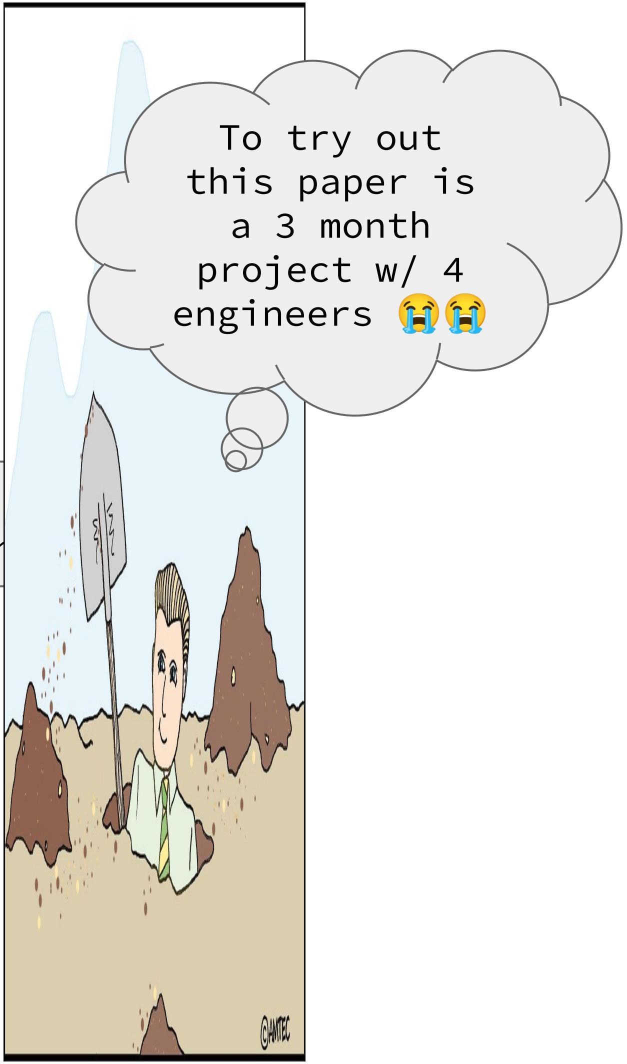



Index Enrichment
Indexing Pipeline
Solr
Pinecone
Search API
ML reranking
📊 NDCG
📊 NDCG
🧑🏻‍💻 Data
Umm wear does my “prioritize SEO popular products” slot in?


I'M JUST A DATA SCIENTIST
YOUR KUBERNETES FRIGHTENS ME
imgflip.com

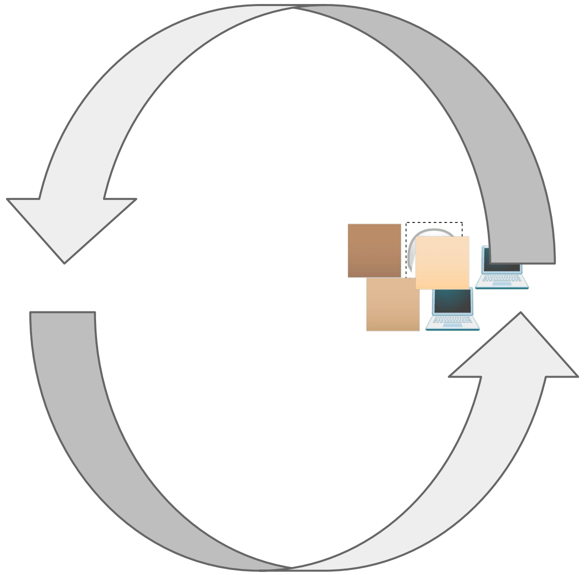
Every hole got deeper




To try out this paper is a 3 month project w/ 4 engineers 😭😭
©AMTEC

Stuck in a paradox

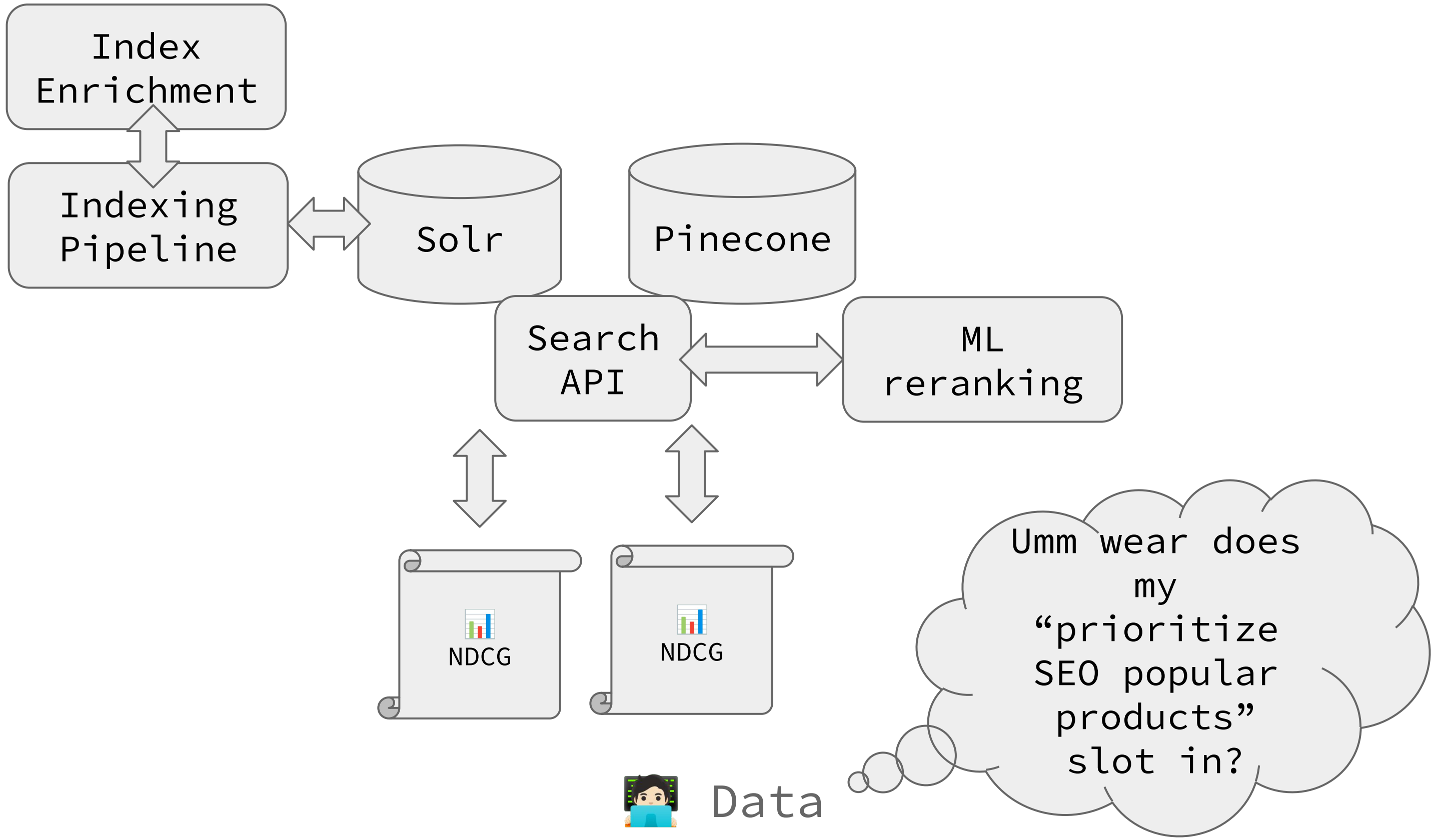
Can’t **do** without **planning**



Can’t **plan** without **building**

Overcoming the paradox

Step 1: know this is all just math



Ranking function

```
score = f(f1, f2, f3, … fn)

Features:

f1 - some BM25 score of the query on some
field

f2 - if query mentions electronics, 1,
otherwise 0

f3 - dimension[121] of query embedding

...
```

Unbury the ranking function from infra

Final ranking produced by your API

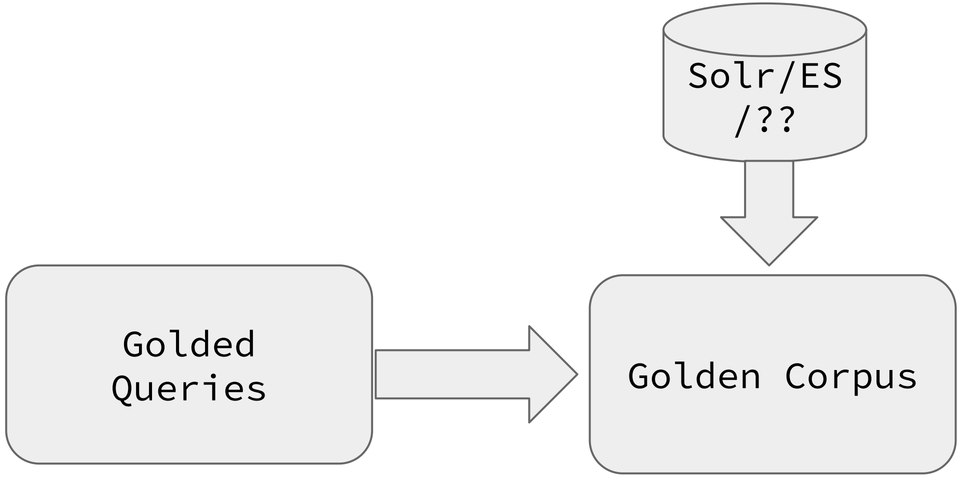
```
score = f(f1, f2, f3, … fn)

Features:

f1 - some BM25 score of the query on some
field

f2 - if query mentions electronics, 1,
otherwise 0

f3 - dimension[121] of query embedding

...
```

How you query Elasticsearch

Some complicated function in your API

Lookup in the vector

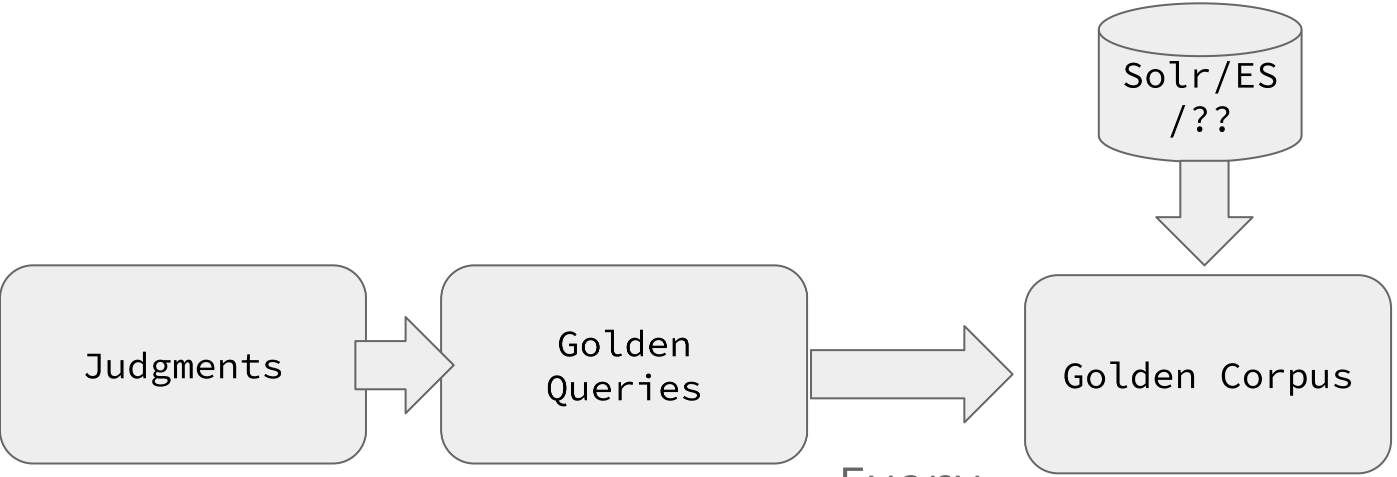
Play with a new feature



Every
labeled
result

A downsample of every
query’s labeled result
for quick prototyping in
a ‘kaggle style’

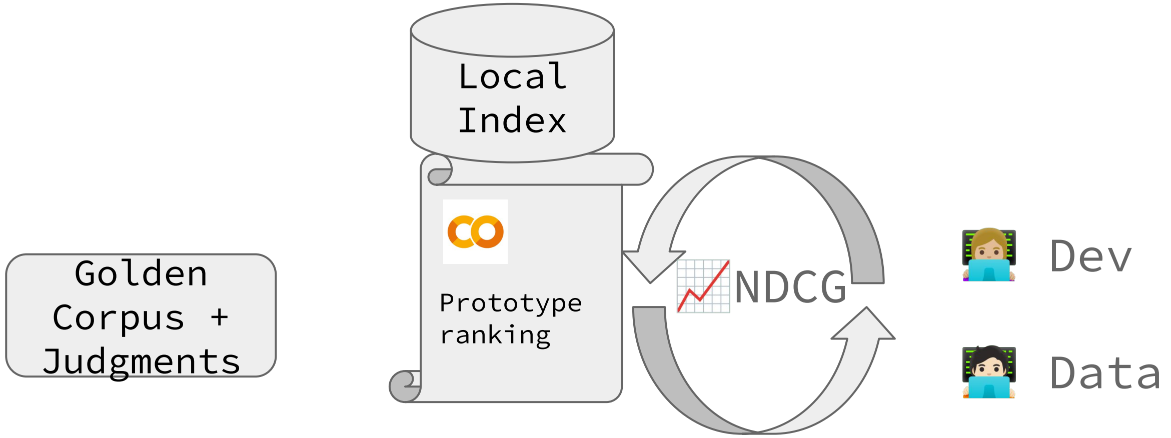
Baseline



Every labeled result

A downsample of every query’s labeled resul
for quick prototyping
a ‘kaggle style’

Colab notebook or something



Let's try an example!

Colab Notebook

What happens when we add fn+1

```
score = f(f1, f2, f3, … fn, fn+1)

Features:

…

fn+1 - that cool thing you read about in
some paper
```

Answer these questions

1. Does **fn+1** add information not already present?

2. Does **fn+1** add information that helps distinguish relevant from irrelevant?

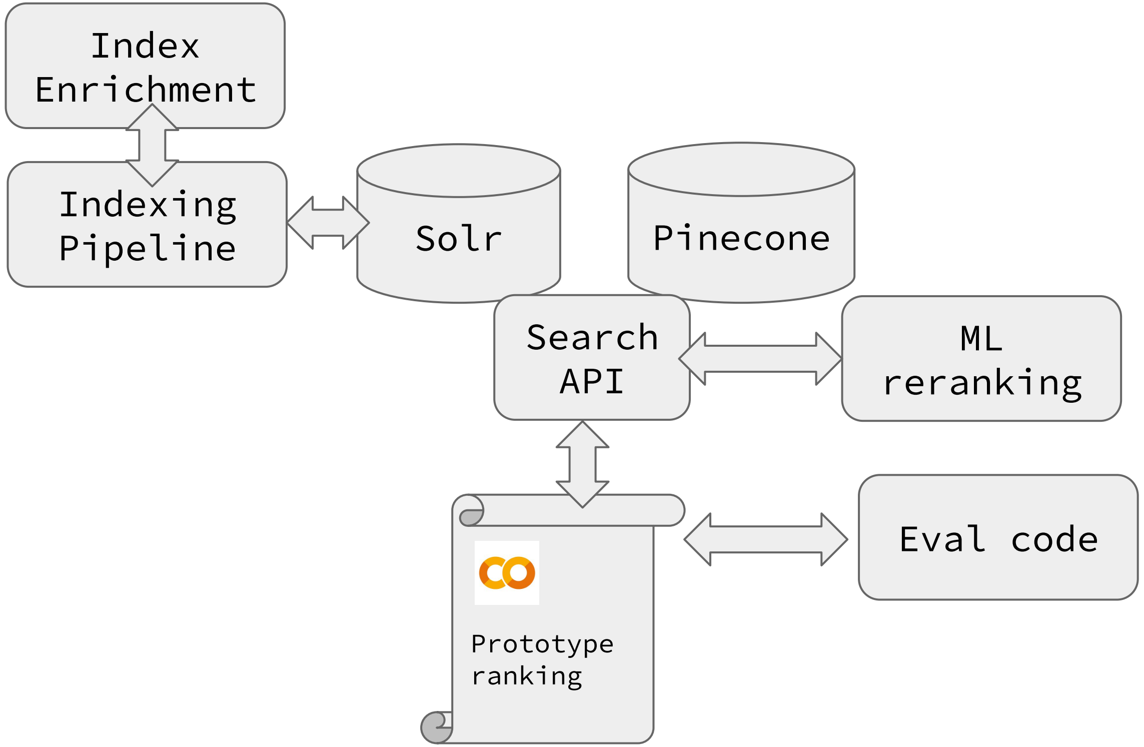
Let's try an example!

Feature Exploration:
[Baseline]() | [Add Phrase Search]()

Make your code prototypable

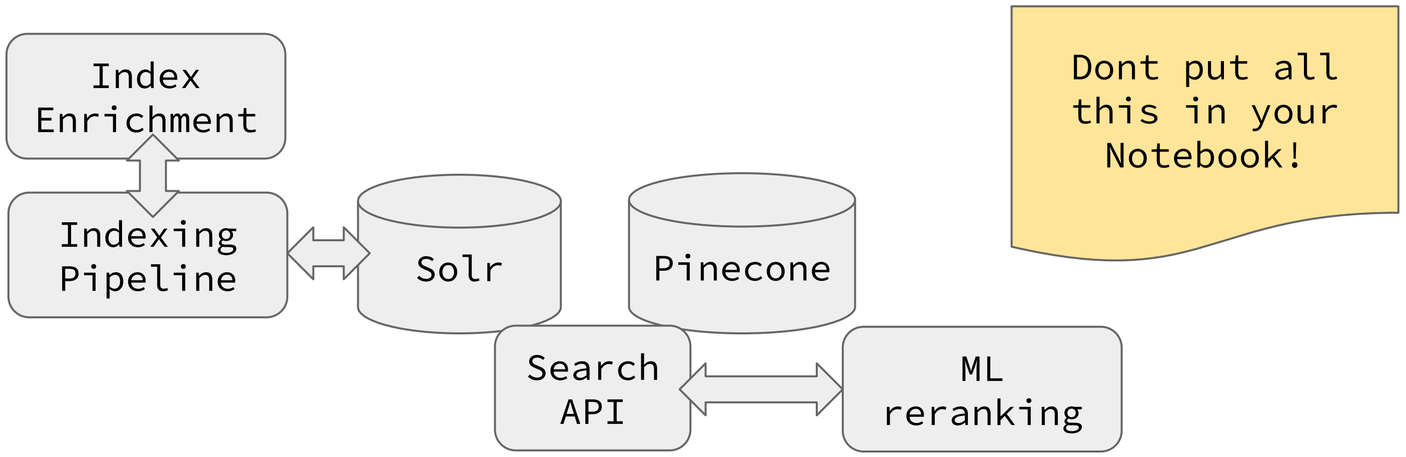
Too much notebook code is… hell

Not-tested, global state, etc



“I don’t like notebooks”
Joel Grus

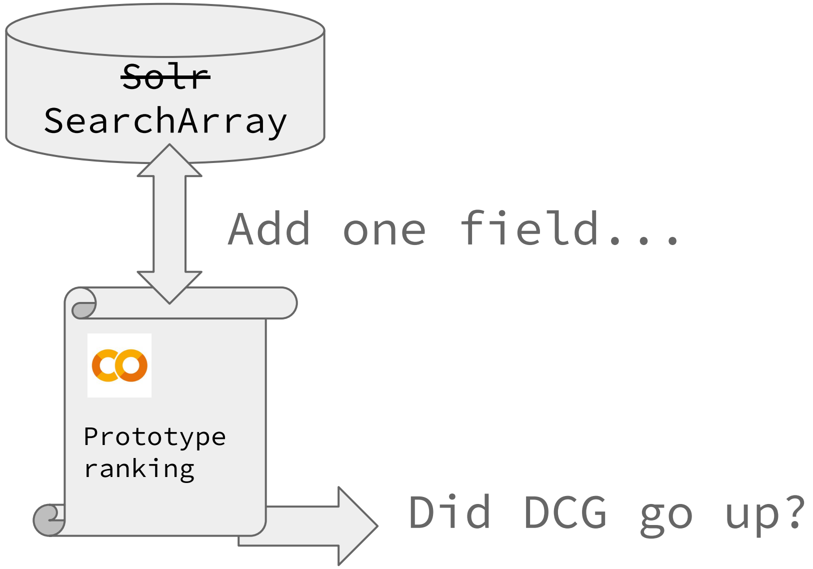
Don’t reimplement search in notebook



(Just create an OK-ish baseline to get quick signal)

Actual goal: *just get a signal!* Not be 100% accurate

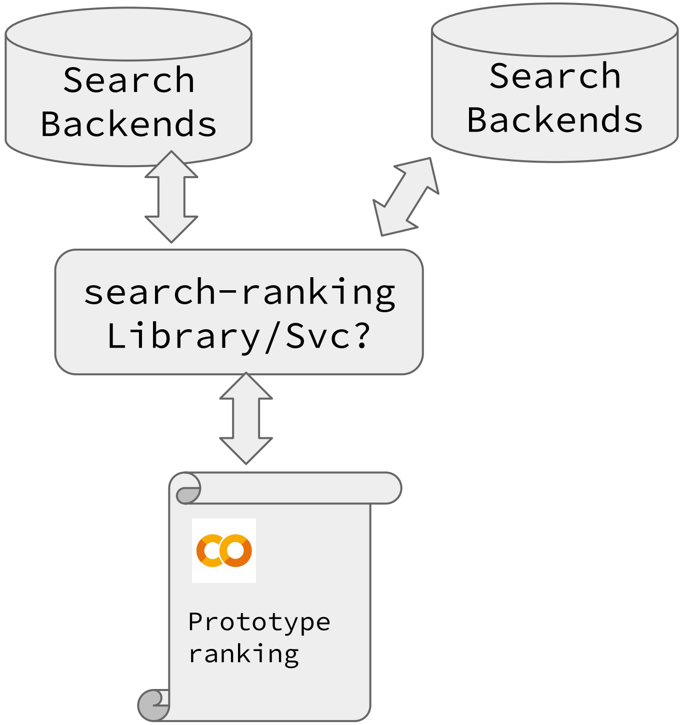
(but make in notebook-able)



Wrap ranking smarts in a library

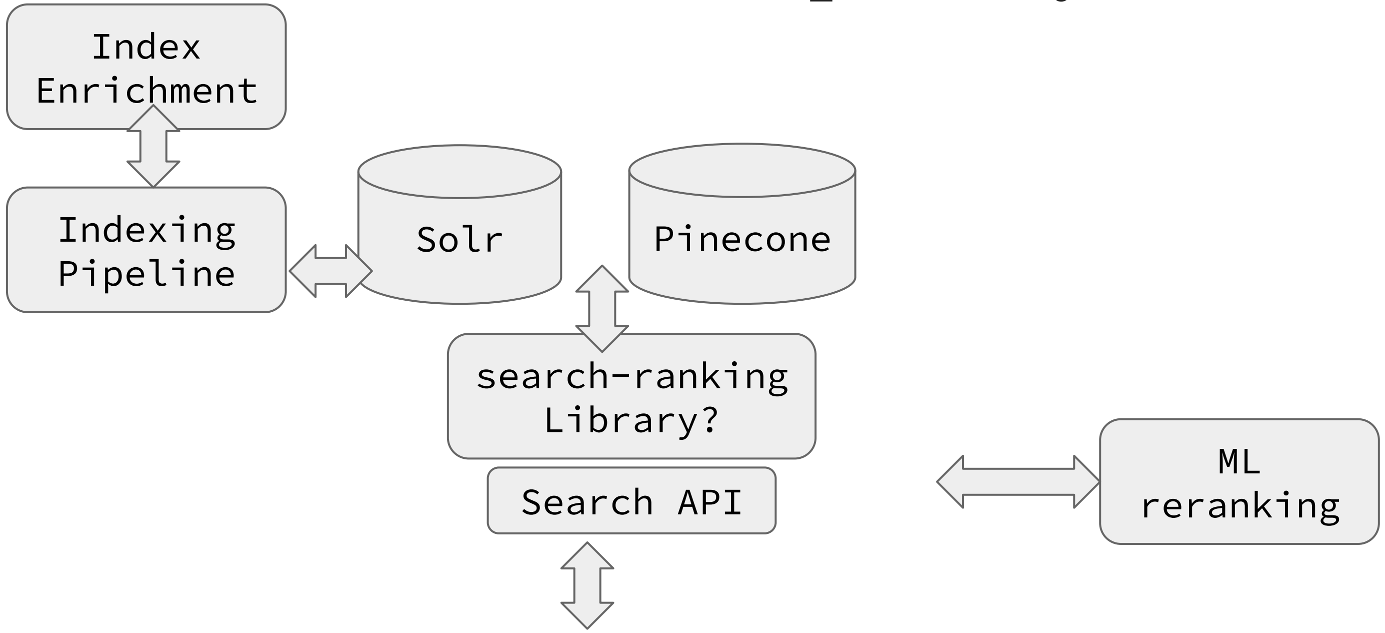
Can work w/ local (colab) or prod backends

Can run as just a library used to execute searches



pip install search-ranking

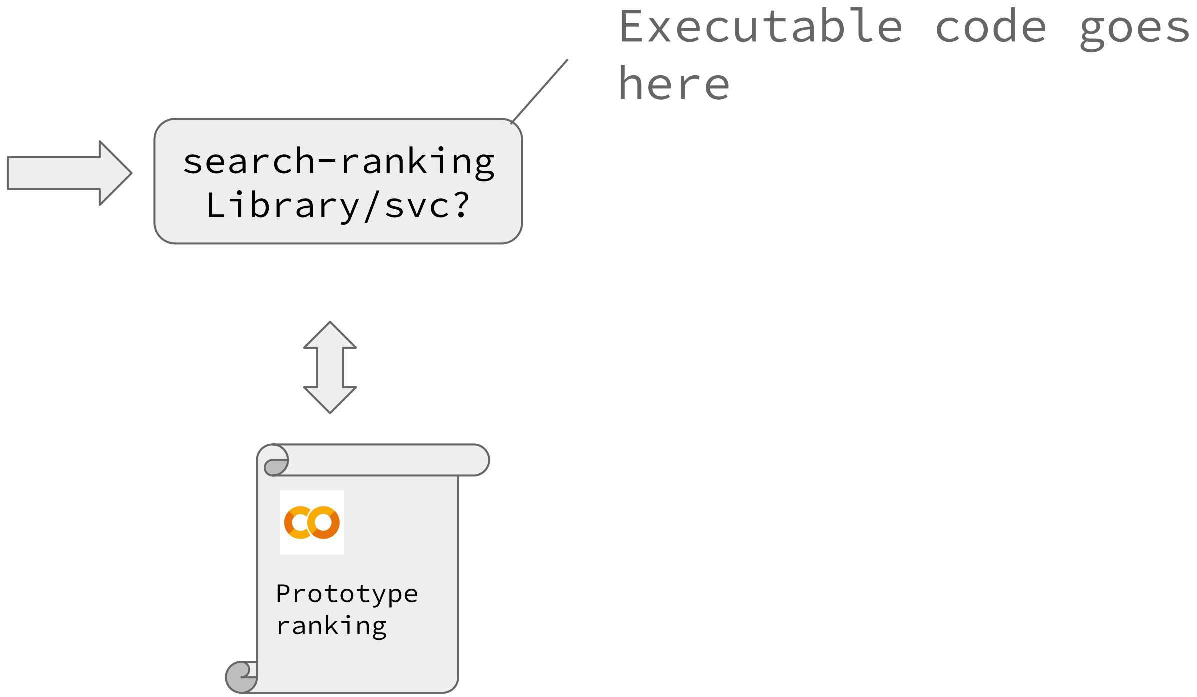
And use those smarts in prod by search API



Make it declarative

Configurable Strategy over executable code:

```
{
  “stages”:
  [
    {“name”:
“parse_entities”,
    “file”: “foo.txt”},
    {“name”: “call_search”,
     “params”: {
        “qf”: “title desc”
        “bq”: …
     }
```



```
pip install search-ranking
```

Make it clear to the whole team

Configurable
Strategy over executable
code:

```
{
  “stages”:
  [
    {“name”: “parse_entities”,
     “file”: “foo.txt”},
    {“name”: “call_search”,
     “params”: {
        “qf”: “title desc”
        “bq”: …
     },
    {“name”: “rerank”,
     “params”: {
        “depth”: 1000
        “model”: “foo”
```

We all see
how search
works - it’s
not magic!

👩🏼‍💻 Dev

👩🏽‍🦳 PM

🧑🏻‍💻 Data

Arrange, Act, Assert - but notebooks

1. Declare system state (ie config)
2. Run the queries
3. Measure the relevance state

Search configs, expected state of the system, etc

(Reference: ARRANGE, ACT, ASSERT)

Arrange, Act, Assert - but notebooks

1. Declare system state (ie config)
2. Run the queries
3. Measure the relevance state

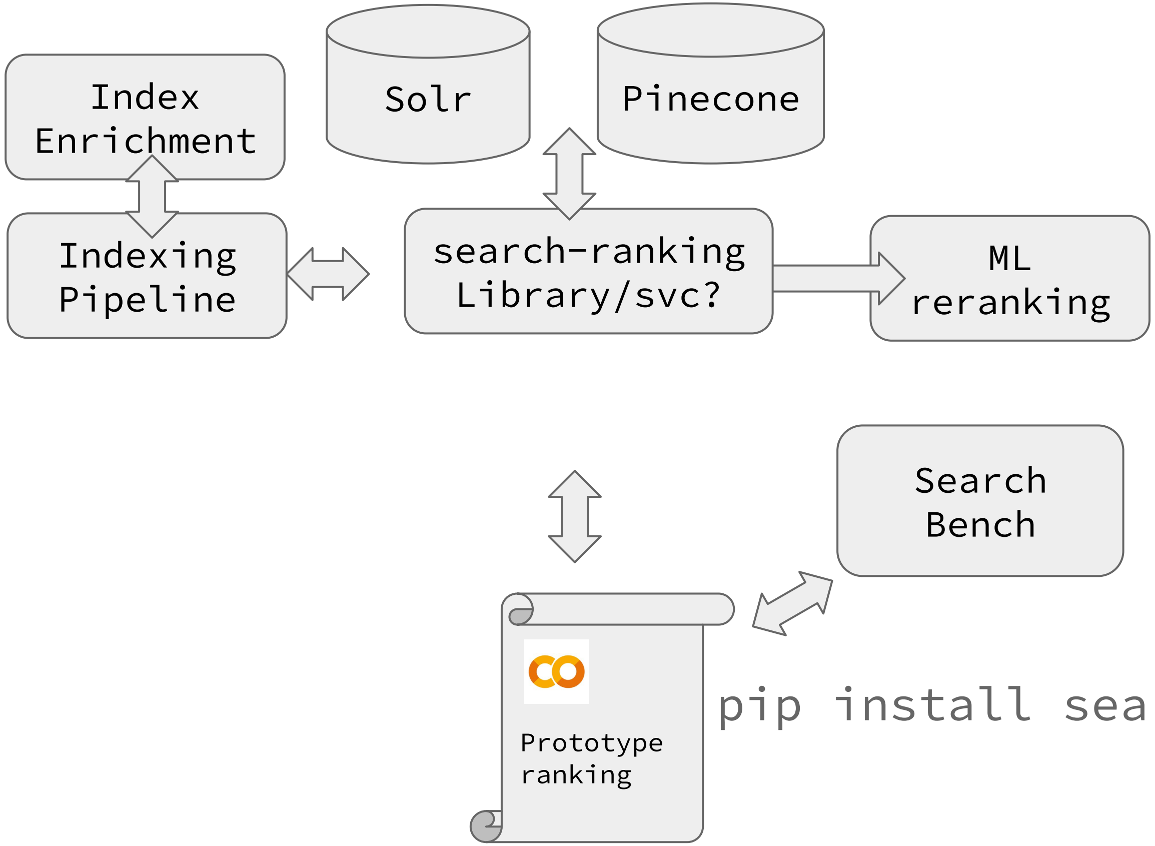
Run 1K queries

Arrange, Act, Assert - but notebooks

1. Declare system state (ie config)
2. Run the queries
3. Measure the relevance state

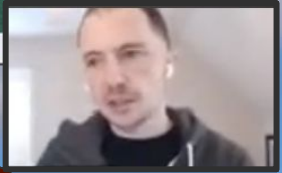
Run reports, compute DCG, etc

And push measurement code to library



```
pip install search-bench
```

If you have questions i might answer them i might not haha



Slide by ian turnbull… again. (if you havent thrown rotten fruit and stuff at my dad please do it now)

Examples of rotten fruit

🍒🪰🍒🪰🍒🪰🍒🪰🍒🪰🍒🪰🍒🪰🍒🪰🍒🪰🍒
🪰🍒🪰