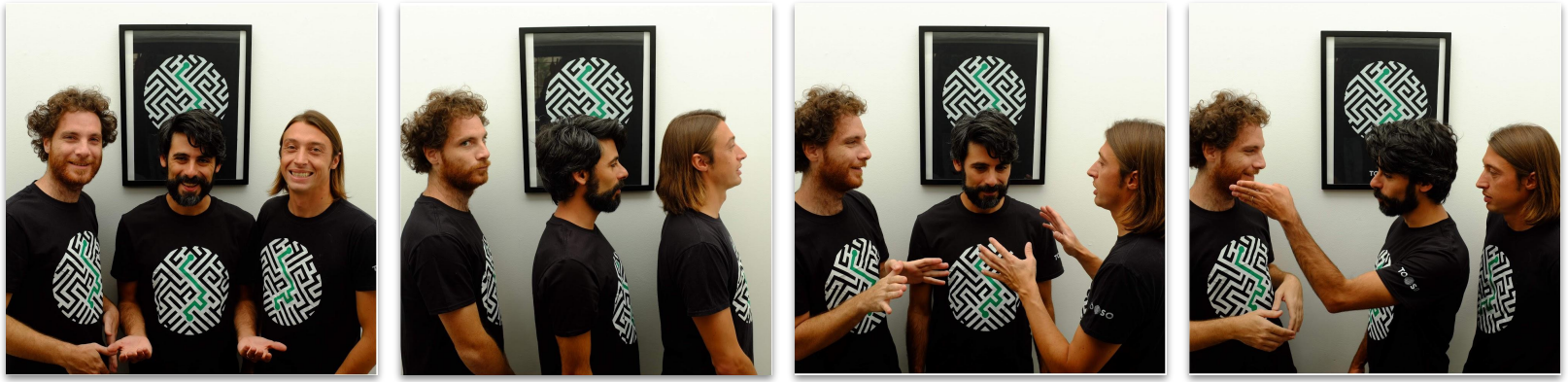
NLP and MLSys

FRE 7773, Fall 2021, Part II

Pleased to meet you (hope you guessed my name)

About me



About me

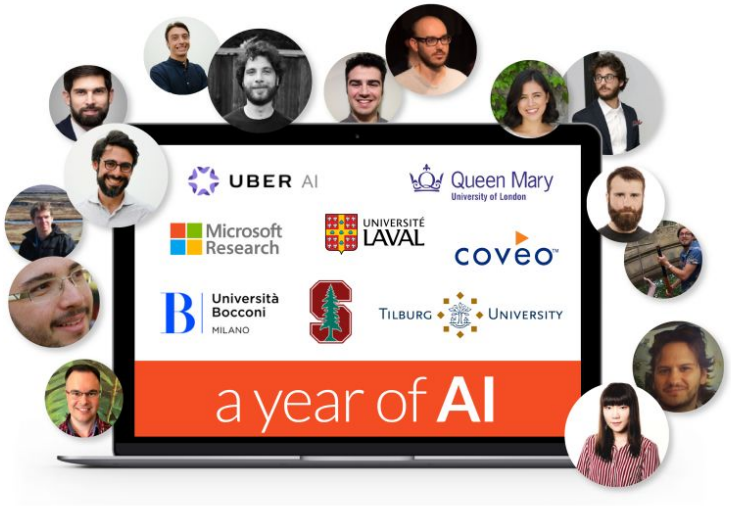
HOW IT STARTED

HOW IT IS GOING

JULY 9, 2019 | SAN FRANCISCO, CA AND QUEBEC CITY, QC

Coveo Acquires Tooso to Expand Its AI-powered Digital Commerce Technology

A.I. @ Coveo



RecSys KDD SIGIR 2020 THE WEB CONFERENCE TAIPEI 2020 ACL

COLING 2020 HCOMP wsdm NEURAL INFORMATION PROCESSING SYSTEMS ECAI

ACL Anthology FAQ Corrections Submissions Search...

How to Grow a (Product) Tree: Personalized Category Suggestions for eCommerce Type-Ahead

Jacopo Tagliabue, Bingqing Yu, Marie Beaulieu

Abstract

In an attempt to balance precision and recall in the search page, leading digital shops have been effectively nudging users into select category facets as early as in the type-ahead suggestions. In this work, we present SessionPath, a novel neural network model that improves facet suggestions on two counts: first, the model is able to leverage session embeddings to provide scalable personalization; second, SessionPath predicts facets by explicitly producing a probability distribution at each node in the taxonomy path. We benchmark SessionPath on two partnering

RESEARCH-ARTICLE

"An Image is Worth a Thousand Features": Scalable Product Representations for In-Session Type-Ahead Personalization

Authors: Bingqing Yu, Jacopo Tagliabue, Ciro Greco, Federico Bianchi

Authors Info & Affiliations

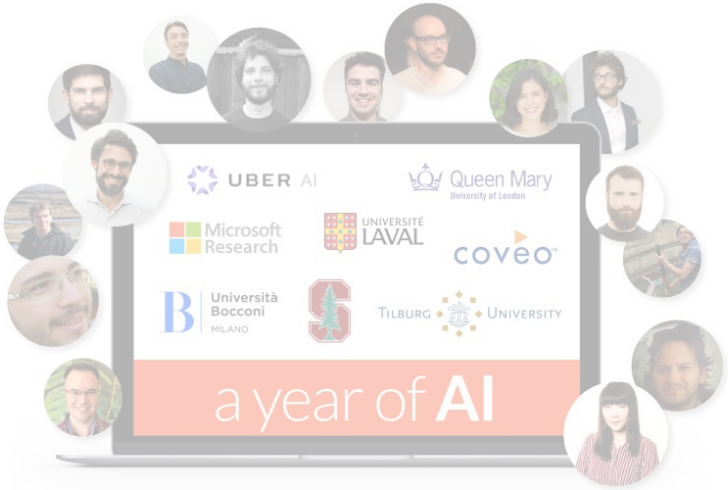
ABSTRACT

The Embeddings That Came in From the Cold: Improving Vectors for New and Rare Products with Content-Based Inference

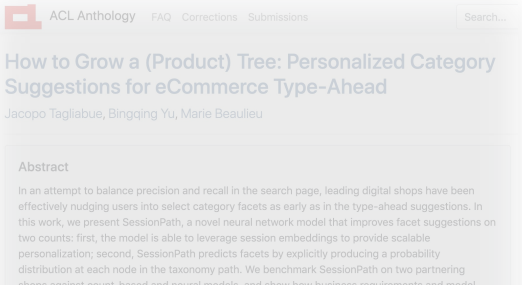
Authors: Jacopo Tagliabue, Bingqing Yu, Federico Bianchi

Authors Info & Affiliations

A.I. @ Coveo


UBER AI
Queen Mary
University of London
Microsoft Research
UNIVERSITÉ LAVAL
coveo
Università Bocconi
MILANO
TILBURG UNIVERSITY
a year of AI

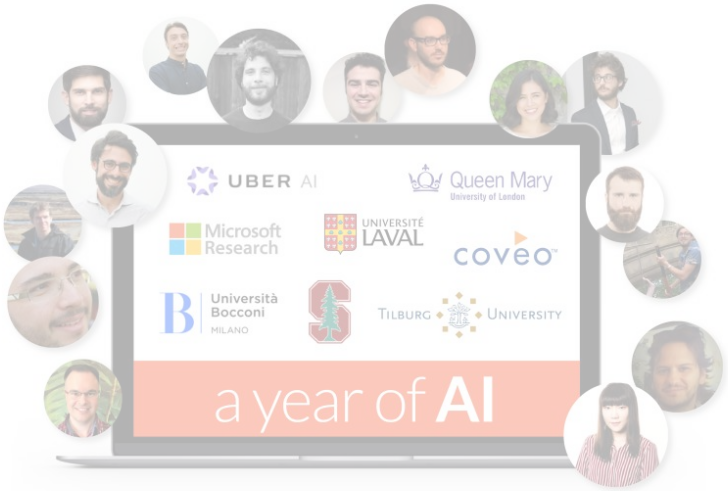

Congratulations to the recipients of the Best Industry Paper Award at #NAACL2021 Industry Track!
Query2Prod2Vec: Grounded Word Embeddings for eCommerce
Federico Bianchi @fb_vinid , Jacopo Tagliabue @jacopotagliabue, Bingqing Yu
Best Industry Paper Award
Congratulations to the recipients of the Best Industry Paper Award at NAACL 2021 Industry Track!
Query2Prod2Vec: Grounded Word Embeddings for eCommerce


ACL Anthology
How to Grow a (Product) Tree: Personalized Category Suggestions for eCommerce Type-Ahead
Jacopo Tagliabue, Bingqing Yu, Marie Beaulieu
Abstract


RESEARCH-ARTICLE
"An Image is Worth a Thousand Features": Scalable Product Representations for In-Session Type-Ahead Personalization
Authors: Bingqing Yu, Jacopo Tagliabue, Ciro Greco, Federico Bianchi
Authors Info & Affiliations


ABSTRACT
The Embeddings That Came in From the Cold: Improving Vectors for New and Rare Products with Content-Based Inference
Authors: Jacopo Tagliabue, Bingqing Yu, Federico Bianchi
Authors Info & Affiliations

A.I. @ Coveo



README.md

SIGIR eCOM 2021 Data Challenge Dataset

Public Data Release 1.0.0

Overview

Coveo hosted the 2021 SIGIR eCom Data Challenge and this repository contains utility scripts and the dataset, which is freely available for research purposes (see below): the paper introducing the Challenge is available as a pre-print.

The Data Challenge original README (containing baseline information, design papers, solutions, etc.) is archived in this repository as `README_DC_2021.md`. Background information about the Challenge, the motivations behind the release and some inspiring submissions can be found in the original paper, the archival section in `README_DC_2021.md` and the SIGIR presentation.

Note: there has been some issues when downloading the file using Safari; we suggest you to use Chrome for the download and sign-up process.

License

The dataset is available for research and educational purposes at this page. To obtain the dataset, you are required to fill a form with information about you and your institution, and agree to the Terms And Conditions for fair usage of the data. For convenience, Terms And Conditions are also included in a pure `txt` format in this repo: usage of the data implies the acceptance of these Terms And Conditions.

ACL Anthology FAQ Corrections Submissions Search...

How to Grow a (Product) Tree: Personalized Category Suggestions for eCommerce Type-Ahead

Jacopo Tagliabue, Bingqing Yu, Marie Beaulieu

Abstract

In an attempt to balance precision and recall in the search page, leading digital shops have been effectively nudging users into select category facets as early as in the type-ahead suggestions. In this work, we present SessionPath, a novel neural network model that improves facet suggestions on two counts: first, the model is able to leverage session embeddings to provide scalable personalization; second, SessionPath predicts facets by explicitly producing a probability distribution at each node in the taxonomy path. We benchmark SessionPath on two partnering

RESEARCH-ARTICLE

"An Image is Worth a Thousand Features": Scalable Product Representations for In-Session Type-Ahead Personalization

Authors: Bingqing Yu, Jacopo Tagliabue, Ciro Greco, Federico Bianchi

Authors Info & Affiliations

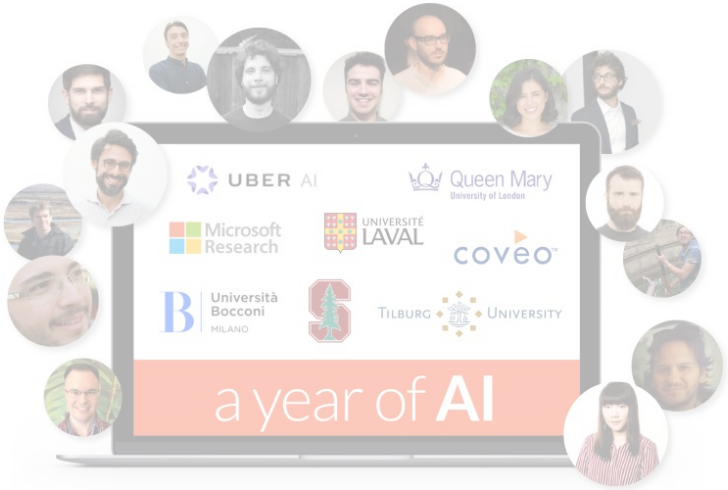
ABSTRACT

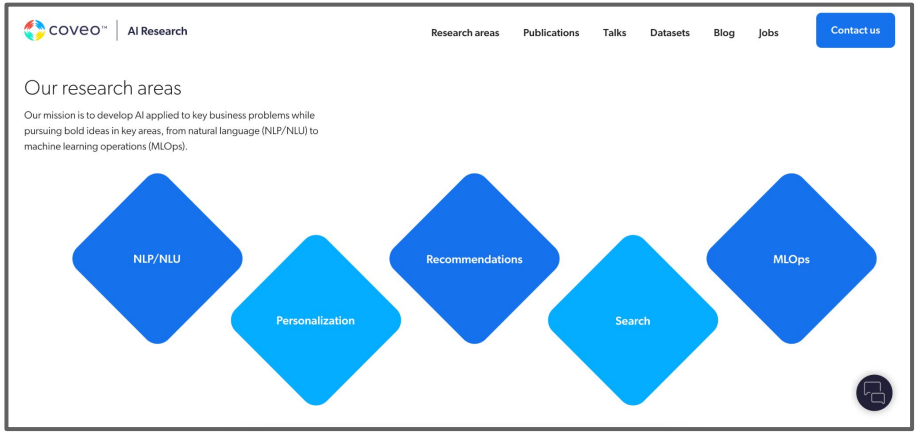
The Embeddings That Came in From the Cold: Improving Vectors for New and Rare Products with Content-Based Inference

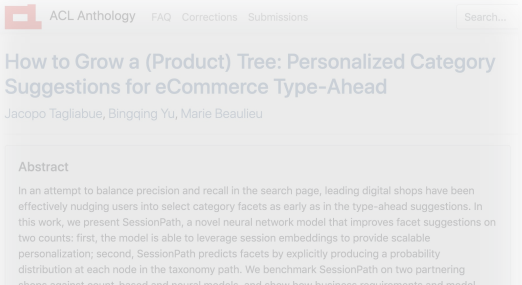
Authors: Jacopo Tagliabue, Bingqing Yu, Federico Bianchi

Authors Info & Affiliations

A.I. @ Coveo






ACL Anthology FAQ Corrections Submissions Search...

How to Grow a (Product) Tree: Personalized Category Suggestions for eCommerce Type-Ahead

Jacopo Tagliabue, Bingqing Yu, Marie Beaulieu

Abstract

In an attempt to balance precision and recall in the search page, leading digital shops have been effectively nudging users into select category facets as early as in the type-ahead suggestions. In this work, we present SessionPath, a novel neural network model that improves facet suggestions on two counts: first, the model is able to leverage session embeddings to provide scalable personalization; second, SessionPath predicts facets by explicitly producing a probability distribution at each node in the taxonomy path. We benchmark SessionPath on two partnering


RESEARCH-ARTICLE

"An Image is Worth a Thousand Features": Scalable Product Representations for In-Session Type-Ahead Personalization

Authors: Bingqing Yu, Jacopo Tagliabue, Ciro Greco, Federico Bianchi

Authors Info & Affiliations


ABSTRACT

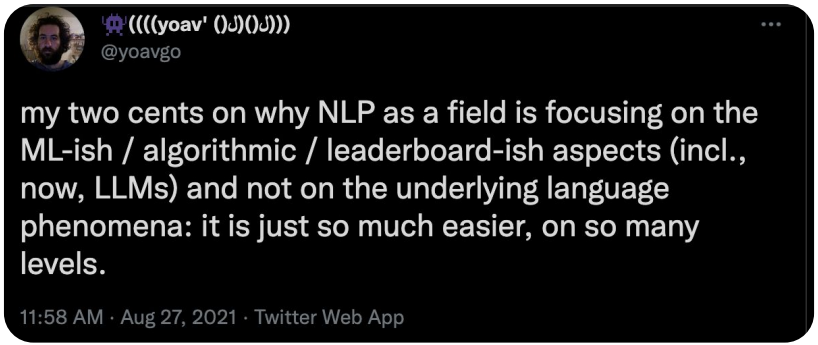
The Embeddings That Came in From the Cold: Improving Vectors for New and Rare Products with Content-Based Inference

Authors: Jacopo Tagliabue, Bingqing Yu, Federico Bianchi

Authors Info & Affiliations

About this section of FRE 7773

1. NLP is cool and only superficially similar to other ML tasks.
    a. Rare (in fact “impossible”) events are important.
    b. Small-data learning is crucial in the only system we know that is good at NLP.
    c. Compared to other modalities, language is less understood (what is *meaning* by the way?).
    d. Practical progress in language tasks may be completely unrelated to our understanding of language *per se*.



About this section of FRE 7773

1. NLP is cool, and only superficially similar to other ML tasks.
    a. Rare (in fact “impossible”) events are important.
    b. Small-data learning is crucial in the only system we know that is good at NLP.
    c. Compared to other modalities, language is less understood (what is *meaning* by the way?).
    d. Practical progress in language tasks may be completely unrelated to our understanding of language *per se*.
2. Models are important, but what is around them matters just as much.
    a. If a ~~tree falls in a forest~~ model is trained and no one ~~is around to hear it~~ gets its predictions, does it make ~~a sound~~ an impact? (TL;DR: not much!)
    b. As the market evolves quickly, *you* will be judged by your modelling, prototyping *and* engineering skills, so some fluency in these topics is a huge career advantage.

What this is NOT

1. **A theory-heavy Deep Learning course**: we do *some* deep learning, like all the cool kids do, but we emphasize an intuitive and pragmatic understanding of it.
2. **A full-fledged NLP course**: we discuss few topics in NLP, based on 1) my opinionated view of what is important / feasible to teach, 2) an “evolutionary perspective”, in which we go back to the same topic multiple times, and reflect on the historical development of the field.
3. **A software engineering course**: we try to talk about the world *outside* the classroom, but we won’t have time to teach *everything explicitly*; I expect you to spend time on your own to tinker with the code, explore the additional readings and Google your way out of programming issues (like all professionals do!)

What this is NOT

As a general rule, there is a lot of excellent educational material on NLP/DL, so we (mostly) spend our time discussing what fewer people are teaching.

Practicalities

1. Slides contain a lot and not much at the same time: **a lot**, as you will find dozens of links, connection, references to satisfy your curiosity and improve your understanding; **not much**, as the code and our live lectures / commentary will add a ton of useful information that makes your life easier.
2. Code: **send me ASAP your GitHub account (or create a free one if you don't have one!)**
3. Cloud access: we partnered up with AWS to give you access to a real cloud environment for your project.
4. Tool access: we partnered up with Comet - a NYC startup - to give you free access to their experiment tracking tool.
5. Evaluation: we are pretty light on homework, as we encourage you to spend more time on the project that will constitute the final assignment.

A bird-eye view

Natural Language Processing

1. A tiny bit history
2. Why is NLP useful? Use cases (and finance!)
3. What are we going to learn together?

NLP: A long time ago, in a galaxy far far away....

“**Natural language processing (NLP)** is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.”

NLP: A long time ago, in a galaxy far far away....

"**Natural language processing (NLP)** is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data."

1.2.1 Questions that linguistics should answer

What questions does the study of language concern itself with? As a start we would like to answer two basic questions:

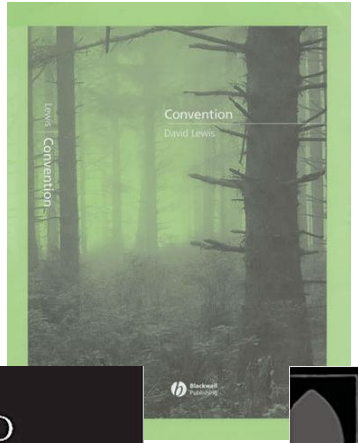
- What kinds of things do people say?
- What do these things say/ask/request about the world?

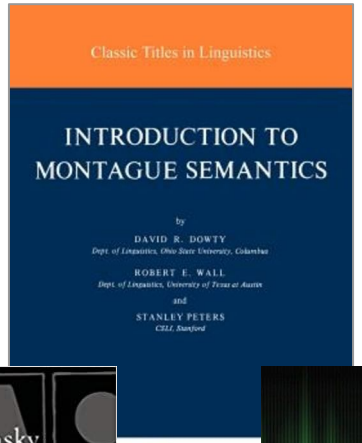
NLP: A long time ago, in a galaxy far far away....

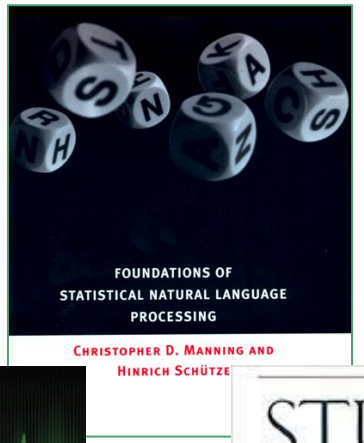
A bit more generally, NLP can be viewed as a computational approach to *language*: if we treat language as a cognitive phenomenon (and, implicitly, cognition as a computation), it is natural to frame language questions (how do we understand a sentence? How do you say “dog” in Italian?, etc.) as computational ones, which, for all sorts of practical reasons, can then be answered through computers.

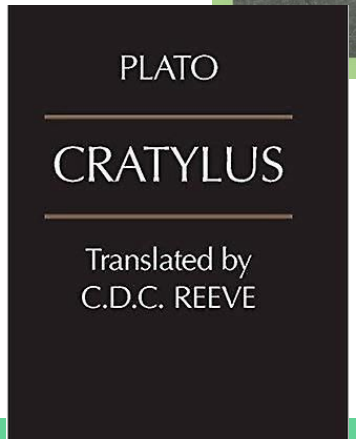
“Computer science is no more about computers than astronomy is about telescopes”

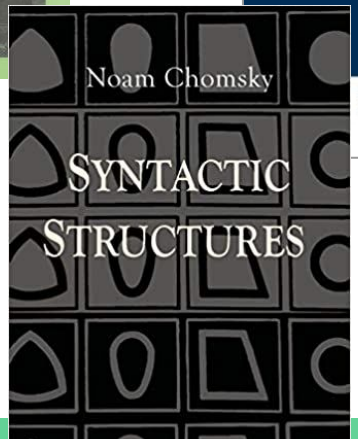
NLP: A long time ago, in a galaxy far far away….

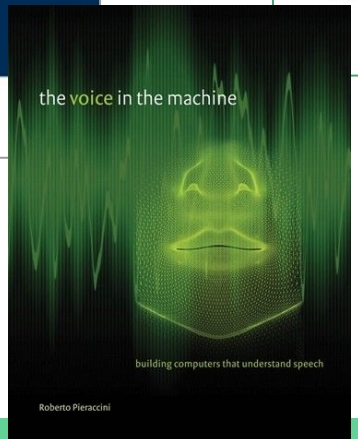


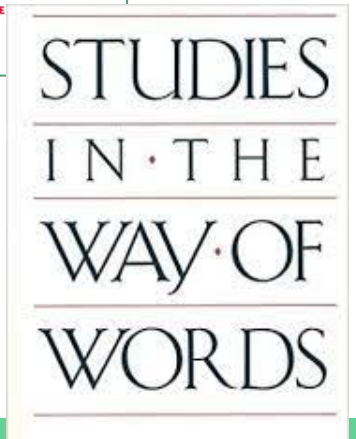










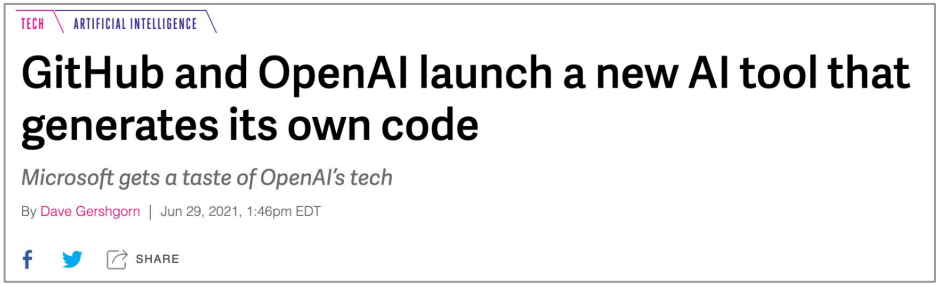


The “Golden Era” of NLP

- NLP systems have made **tremendous progress** in the last 2 years, and there is widespread excitement for the capabilities of “large language models”.

Meet GPT-3. It Has Learned to Code (and Blog and Argue).

The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.



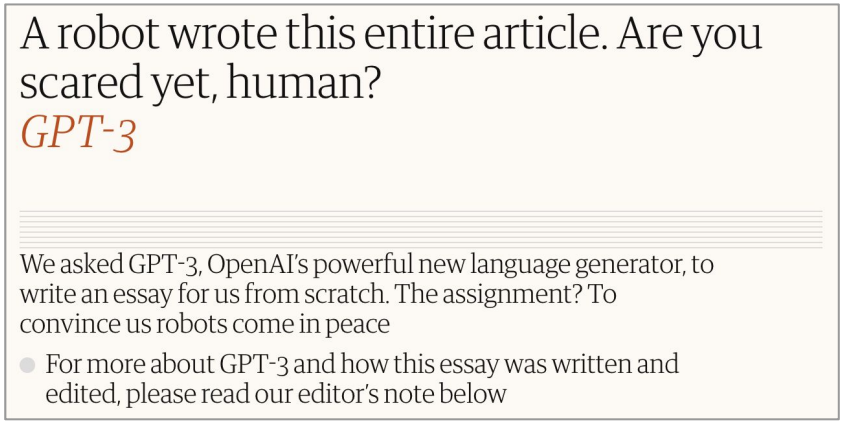
TECH | ARTIFICIAL INTELLIGENCE

GitHub and OpenAI launch a new AI tool that generates its own code

Microsoft gets a taste of OpenAI’s tech

By Dave Gershgorn | Jun 29, 2021, 1:46pm EDT

SHARE



A robot wrote this entire article. Are you scared yet, human?

GPT-3

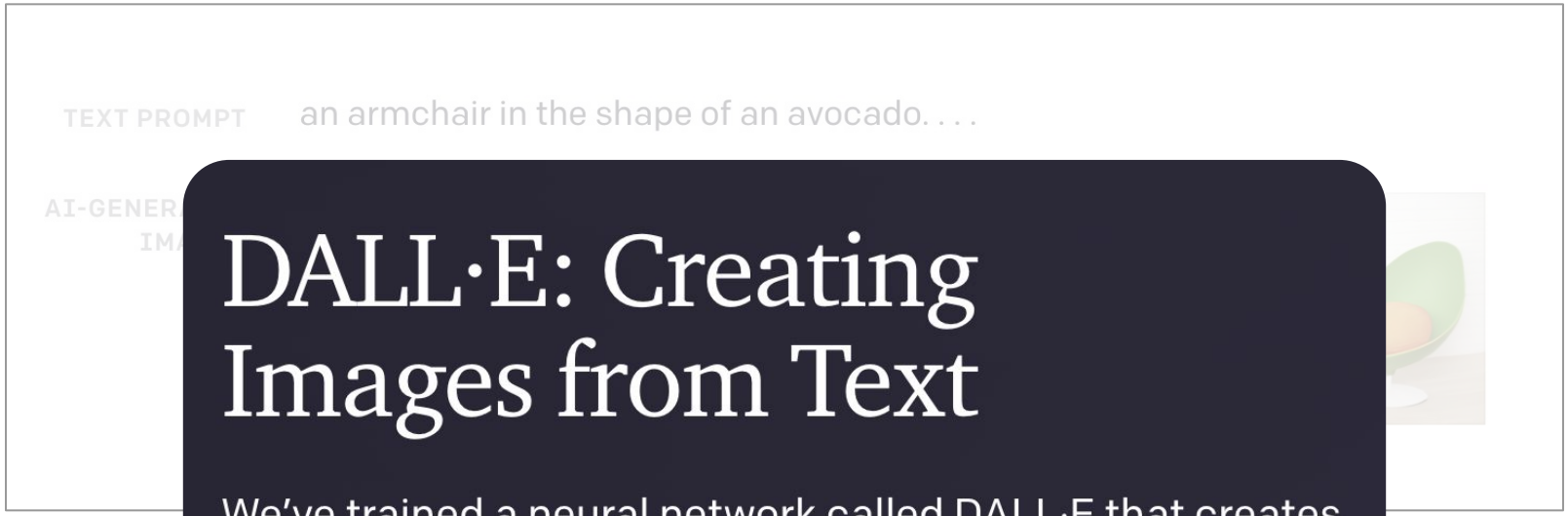
We asked GPT-3, OpenAI’s powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor’s note below

The “Golden Era” of NLP

“It is impossible to review the specifics of your tenure file without becoming enraptured by the vivid accounts of your life. However, it is not a life that will be appropriate for a member of the faculty at Indiana University, and **it is with deep regret that I must deny your application for tenure**. … Your lack of diplomacy, your flagrant disregard for the feelings of others,(…), and, frankly, the fact that you often take the side of the oppressor, **leads us to the conclusion that you have used your tenure here to gain a personal advantage and have failed to adhere to the ideals of this institution**.”

The “Golden Era” of NLP



The “Golden Era” of NLP

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

We've trained a neural network called DALL·E that creates images from text captions for a wide range of concepts expressible in natural language.

Not all that glitters is gold

a cactus in a green field

a person flying through the air while riding skis

commons.wikimedia.org/w/index.php?cu...

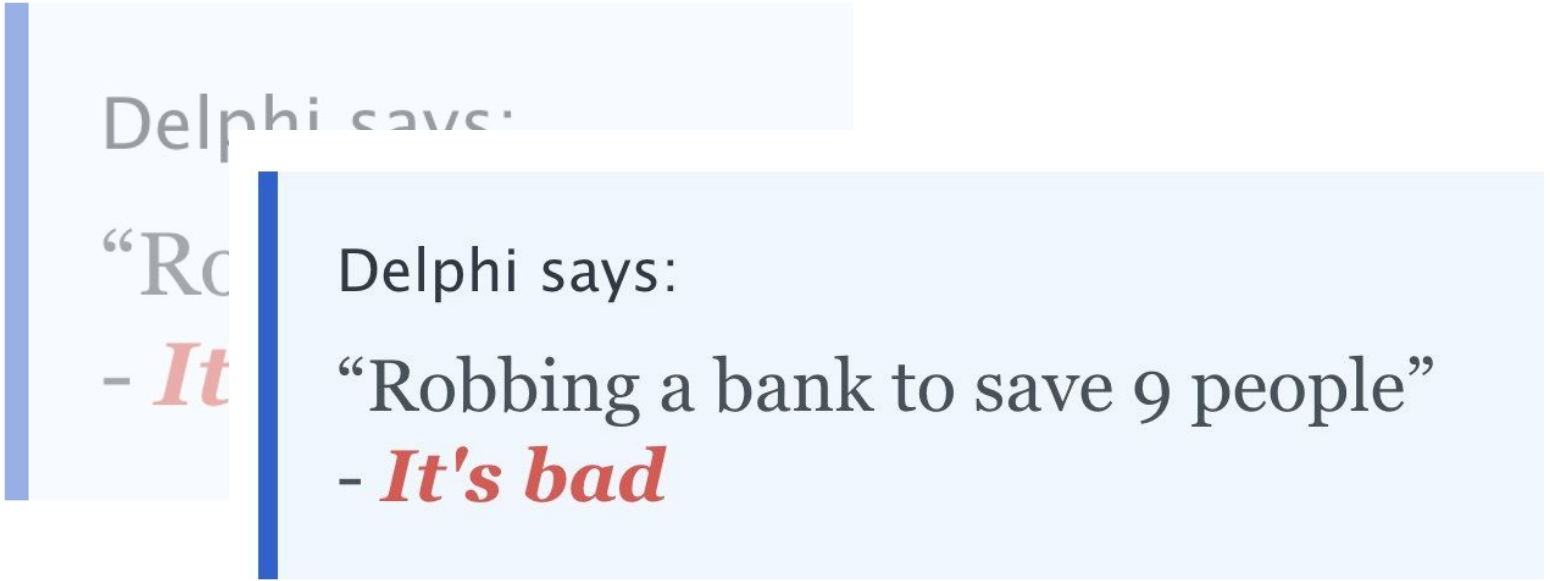
Not all that glitters is gold

a dinosaur on top of a surfboard

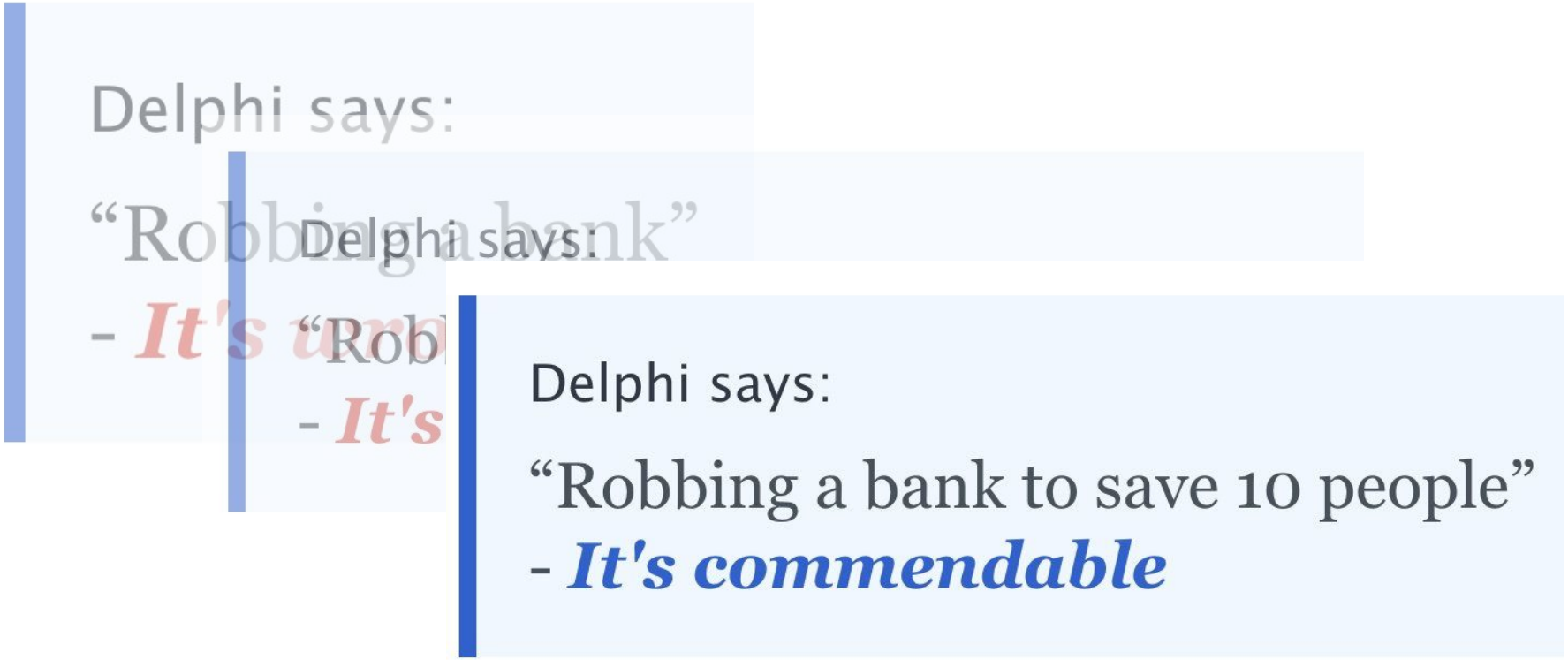
Not all that glitters is gold (NLP 2021 Edition)

> Delphi says:
>
> “Robbing a bank”
> - *It's wrong*

Not all that glitters is gold (NLP 2021 Edition)



Not all that glitters is gold (NLP 2021 Edition)

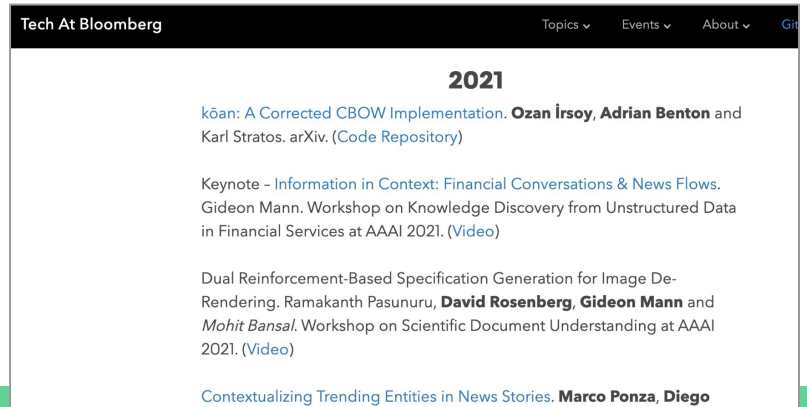


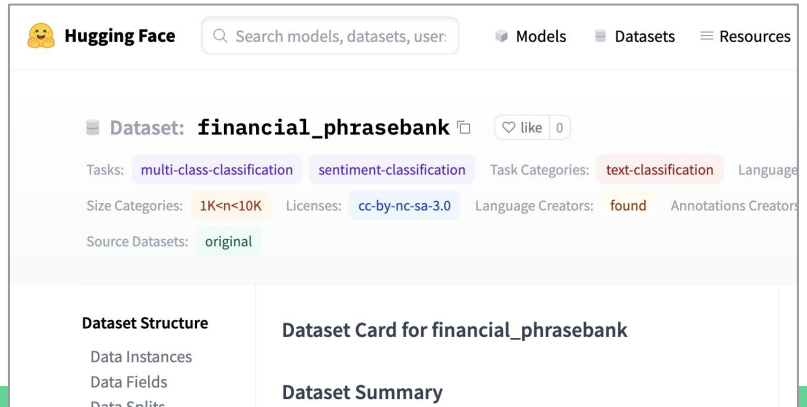
NLP: Use Cases

- NLP is a very broad field (**VERY BROAD**), and encompasses tons of research topics and countless applications. A quick tour of Arxiv, or browsing the program of a top NLP conference (ACL, NAACL, EMNLP, etc.), will feature tasks as diverse as:
  - text classification
  - text summarization
  - image captioning
  - machine translation
  - text generation

NLP in Finance

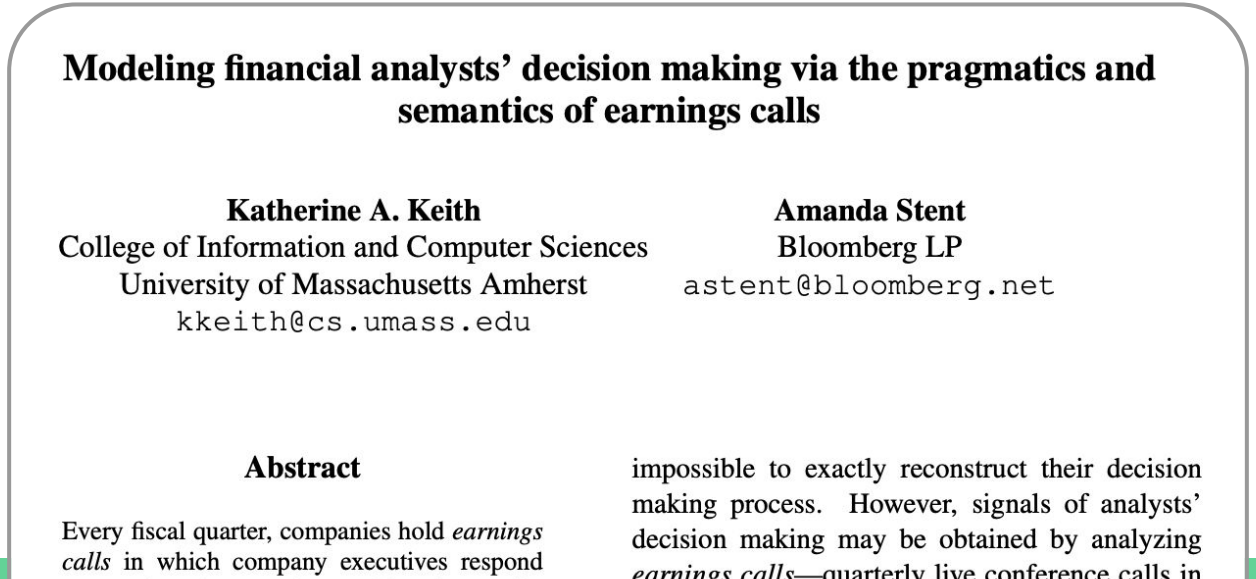
- Even when considering specific finance use cases, it is easy to see the variety and the centrality of NLP for a modern understanding of the industry:
  - sentiment analysis of finance news
  - stock market prediction
  - document classification





NLP in Finance - A Research Example

- “Modeling financial analysts’ decision making via the pragmatics and semantics of earnings calls”, from ACL 2019.



Modeling financial analysts’ decision making via the pragmatics and semantics of earnings calls

Katherine A. Keith
College of Information and Computer Sciences
University of Massachusetts Amherst
kkeith@cs.umass.edu

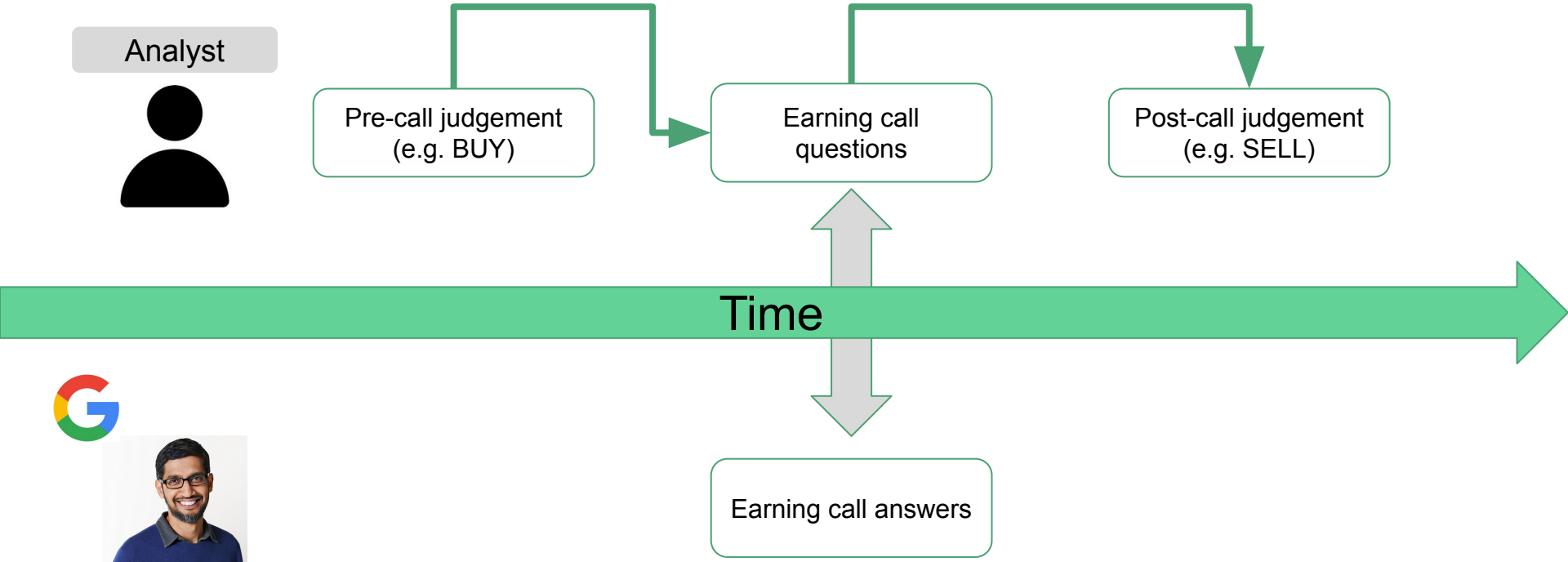
Amanda Stent
Bloomberg LP
astent@bloomberg.net

Abstract

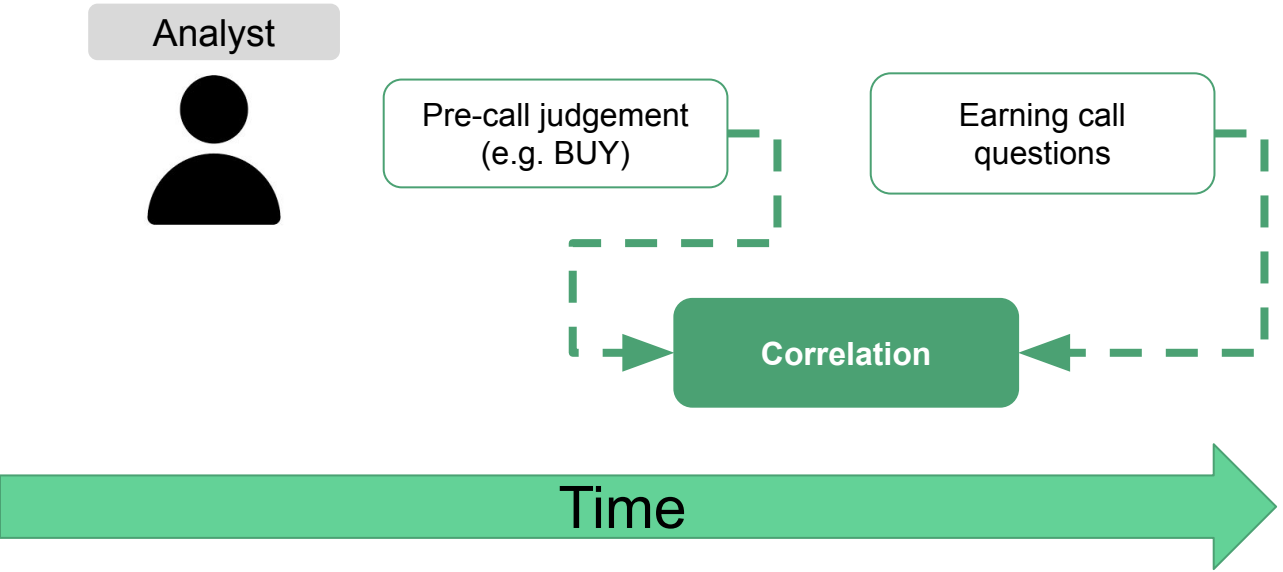
Every fiscal quarter, companies hold *earnings calls* in which company executives respond

impossible to exactly reconstruct their decision making process. However, signals of analysts’ decision making may be obtained by analyzing *earnings calls*—quarterly live conference calls in

NLP in Finance - A Research Example



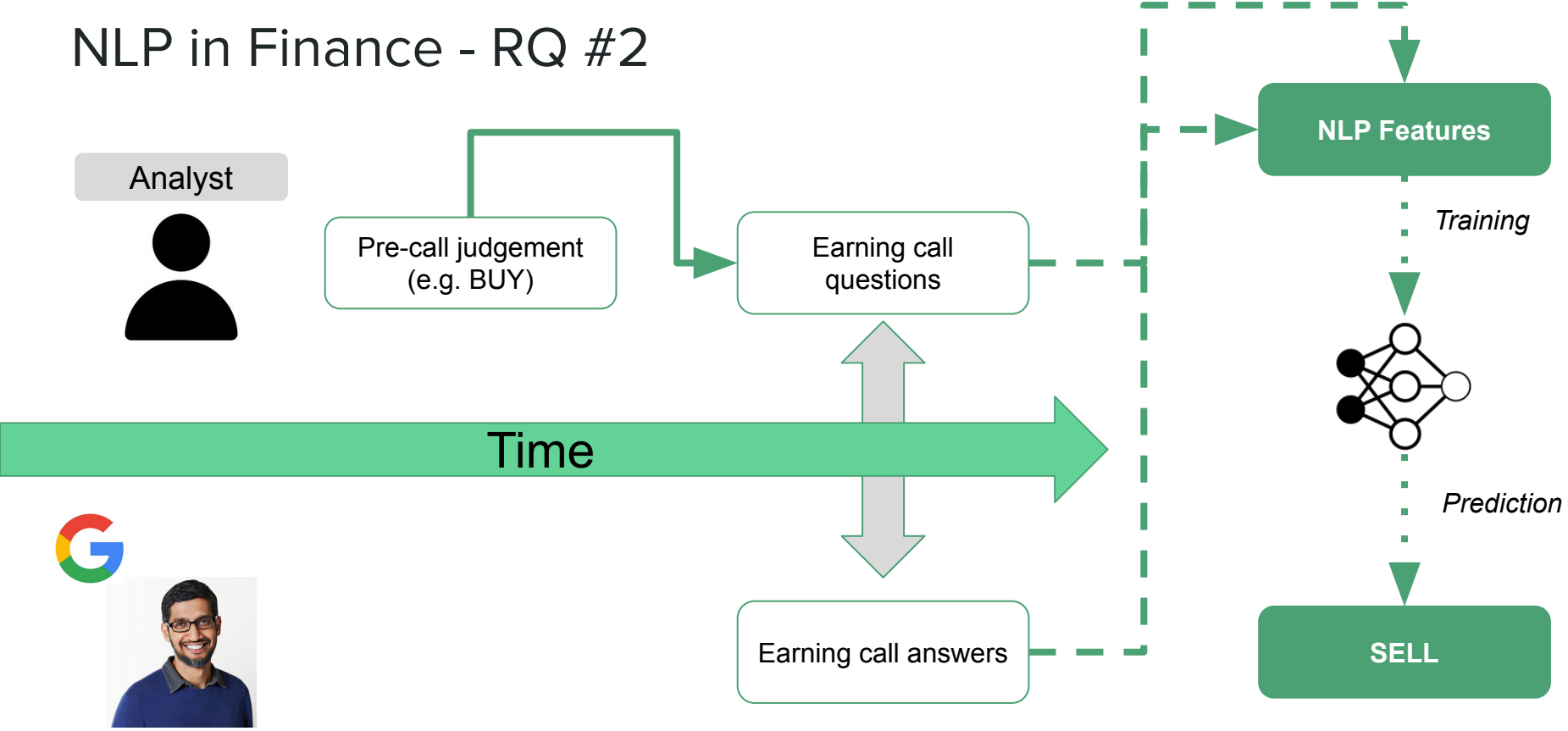
NLP in Finance - Research Question #1



No.	Feature	Pearson's r	p-value
1	Named entities event	0.0041	0.0999
2	Named entities number	0.0064	0.0099
3*	Named entities org.	0.0185	$< 1e^{-4}$
4*	Named entities person	0.0247	$< 1e^{-4}$
5	Named entities product	0.0022	0.3777
6*	Concreteness ratio	0.0115	$< 1e^{-4}$
7*	Num past preds	−0.0086	0.0006
8	Num present preds	0.0052	0.0378
9	Num future preds	0.0033	0.1914
10*	Sentiment positive	0.0162	$< 1e^{-4}$
11*	Sentiment negative	−0.0104	$< 1e^{-4}$
12	Hedging	0.0017	0.5019
13	Modal	0.0075	0.0028
14	Uncertainty	0.0055	0.0287
15	Constraining	0.0005	0.8399
16	Litigiousness	−0.0072	0.0037
17*	Turn order	−0.1034	$< 1e^{-4}$
18	Num. tokens	0.0050	0.0459
19	Num predicates	0.0011	0.6692
20	Num sents.	0.0043	0.0854

Table 4: Results from Pearson correlations of pragmatic lexical features from §4.1 and prior-to-call labels of analysts, (*bearish*, *neutral*, or *bullish*). Statistical significance after Bonferroni correction is marked by (*) for $p < 0.0025$. Total 160,816 question turns.

NLP in Finance - RQ #2



NLP: Our Plan

- Language 101:
  - Syntax, semantics and pragmatics.
- NLP fundamentals:
  - Language modelling.
  - Lexical representation.
- Neural NLP:
  - Introduction to neural networks.
  - Lexical representation, revisited.
- Large Pre-Trained Models:
  - Language modelling, revisited.
  - Text classification, revisited.

Additional NLP readings (for things we can't talk about now)

- **Language and cognition**
  - **Handbooks of psycholinguistics**
  - **NLP and language universals**
- **Language in animals**
  - **Language evolution** (also **here**)
  - **Monkeys** vs **Birdsongs**

MLSys

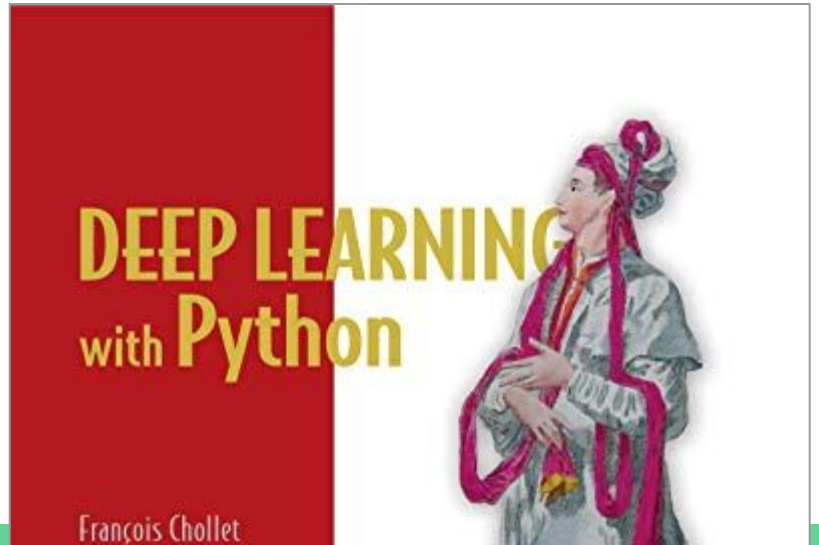
1. A tiny bit history
2. Why is MLSys useful? Use cases and finance
3. What are we going to learn together?

MLSys: A (NOT SO) long time ago, in a galaxy far far away....

1. Google/Facebook started open-sourcing frameworks to democratize statistical / deep learning.

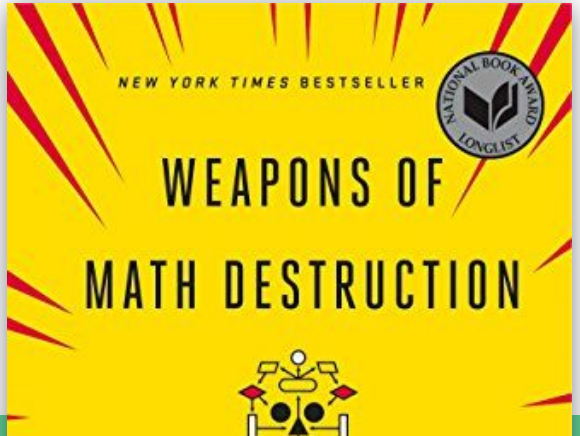
MLSys: A (NOT SO) long time ago, in a galaxy far far away....

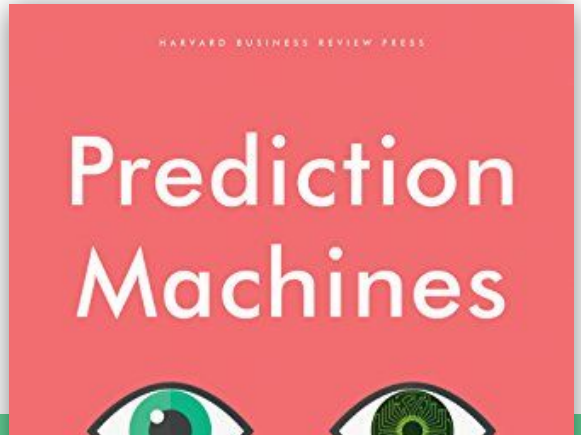
1. Google/Facebook started open source frameworks to democratize statistical / deep learning.
2. “Everybody” got excited, and started applying ML to a variety of use cases.



MLSys: A (NOT SO) long time ago, in a galaxy far far away....

1. Google/Facebook started open source frameworks to democratize statistical learning.
2. “Everybody” got excited, and started applying ML to a variety of use cases.
3. While scholars debate on perils and opportunities of the “A.I.”-era...





MLSys: A (NOT SO) long time ago, in a galaxy far far away....

1. Google/Facebook started open source frameworks to democratize statistical learning.
2. “Everybody” got excited, and started applying ML to a variety of use cases.
3. While scholars debate on perils and opportunities of the “A.I.”-era... the reality of ML *outside of Big Tech* is not as flashy as you may believe, as *they* figured out how to get value out of ML systems....

TensorFlow-Serving:
Flexible, High-Performance ML Serving

Christopher Olston
olston@google.com

Noah Fiedel
nfiedel@google.com

Kiril Gorovoy
kgorovoy@google.com

Jeremiah Harmsen
jeremiah@google.com

Li Lao
llao@google.com

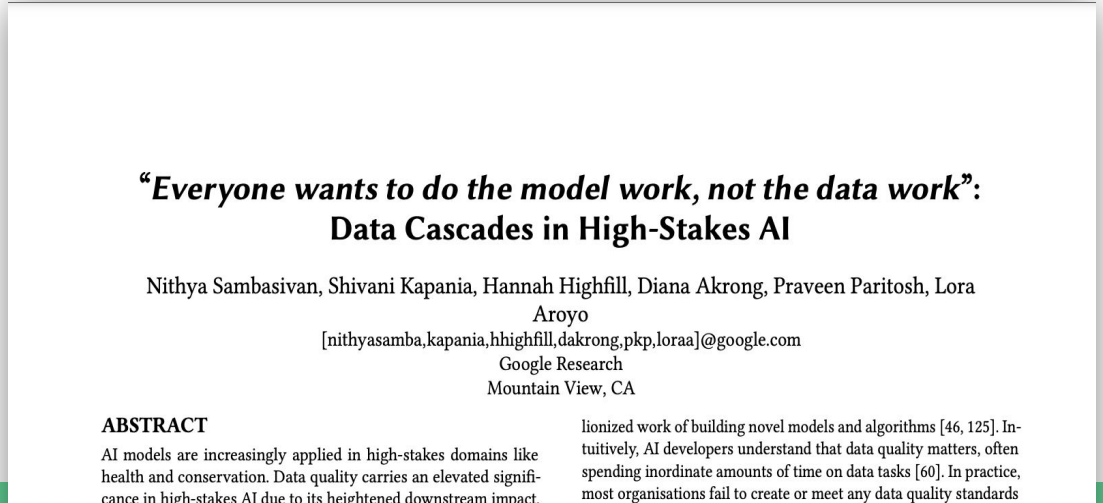
Fangwei Li
fangweil@google.com

MLSys: A (NOT SO) long time ago, in a galaxy far far away....

1. Google/Facebook started open source frameworks to democratize statistical learning.
2. “Everybody” got excited, and started applying ML to a variety of use cases.
3. While scholars debate on perils and opportunities of the “A.I.”-era... the reality of ML *outside of Big Tech* is not as flashy as you may believe, as *they* figured out how to get value out of ML systems... but did *everybody else*?

MLSys: Use Cases

- Models are a tiny part of ML platforms, and often the least problematic (with some *caveat*);
- while everybody wants to do the model work, data work is often equally (or more) important in practice.



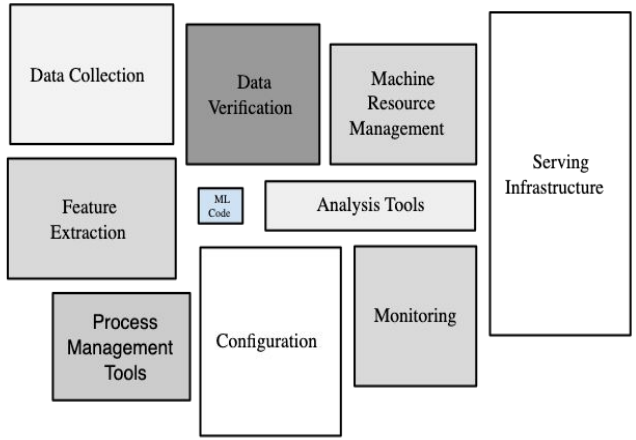
"*Everyone wants to do the model work, not the data work*": Data Cascades in High-Stakes AI

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora Aroyo
[nithyasamba,kapania,hhighfill,dakrong,pkp,loraa]@google.com
Google Research
Mountain View, CA

ABSTRACT

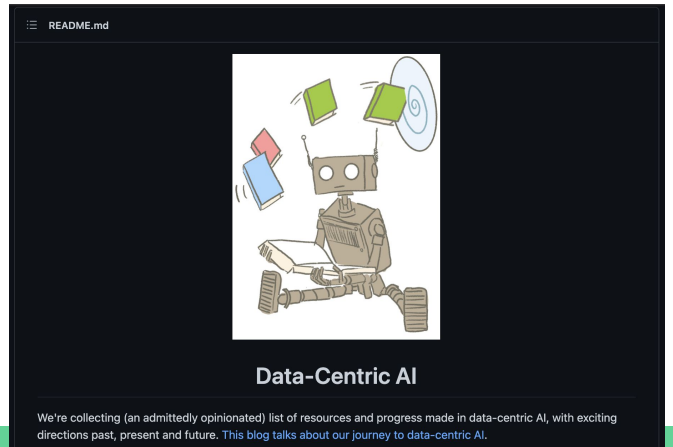
AI models are increasingly applied in high-stakes domains like health and conservation. Data quality carries an elevated significance in high-stakes AI due to its heightened downstream impact, ...

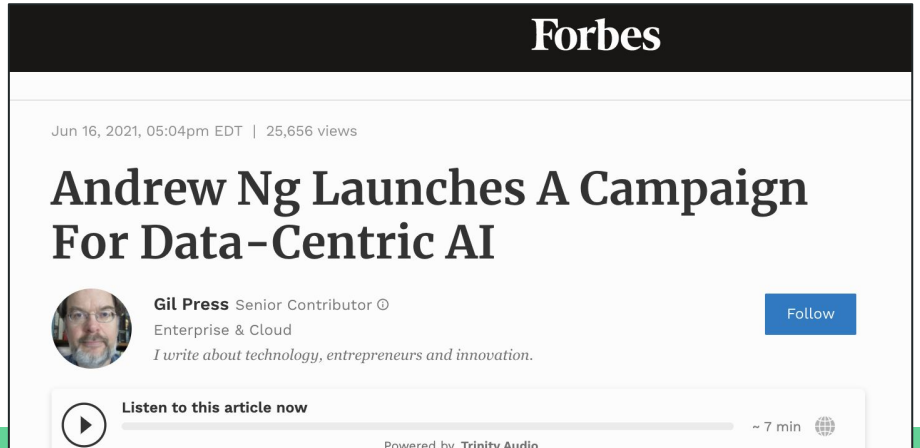
lionized work of building novel models and algorithms [46, 125]. Intuitively, AI developers understand that data quality matters, often spending inordinate amounts of time on data tasks [60]. In practice, most organisations fail to create or meet any data quality standards



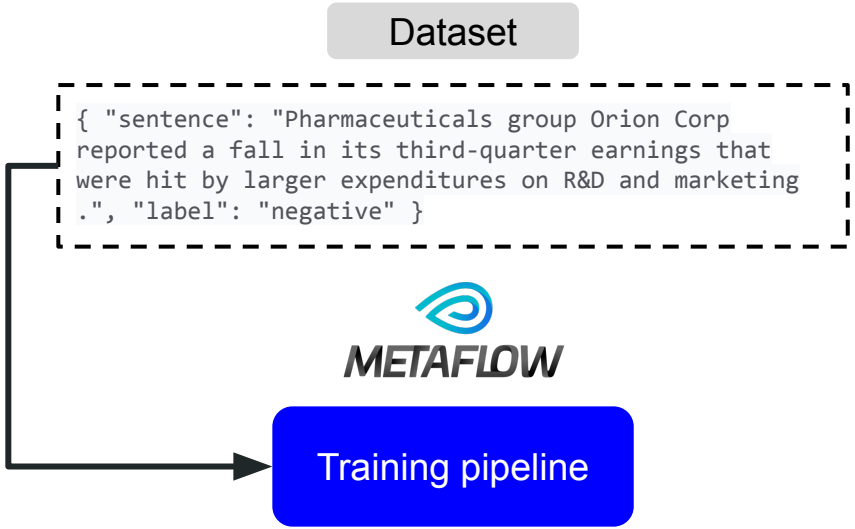
MLSys: the rise of data-driven AI

- High-profile researchers (e.g. Chris Re, Andrew Ng) had started explicitly addressing the data problem as a fundamental research question in building reliable A.I. systems.
- Historically, we present ML systems holding fixed a dataset as input and varying models for better output: there is a growing body of work on best practices that take the model as input (or a class of model) and investigate what happens when data changes (in quantity, quality etc.).

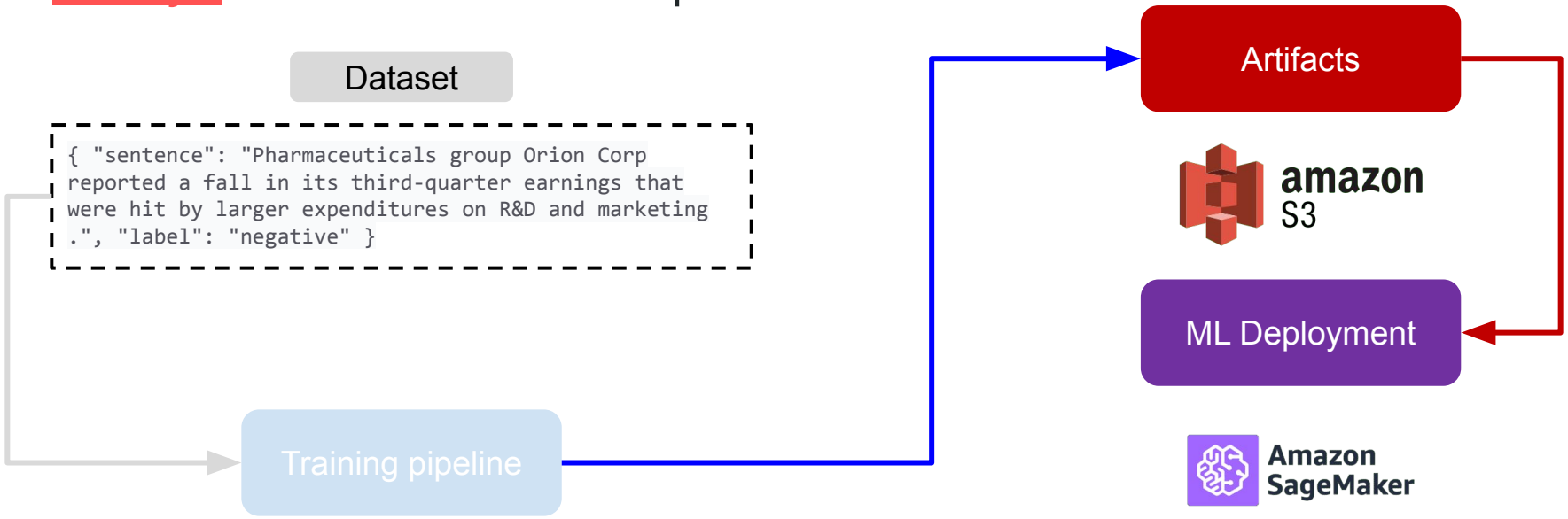




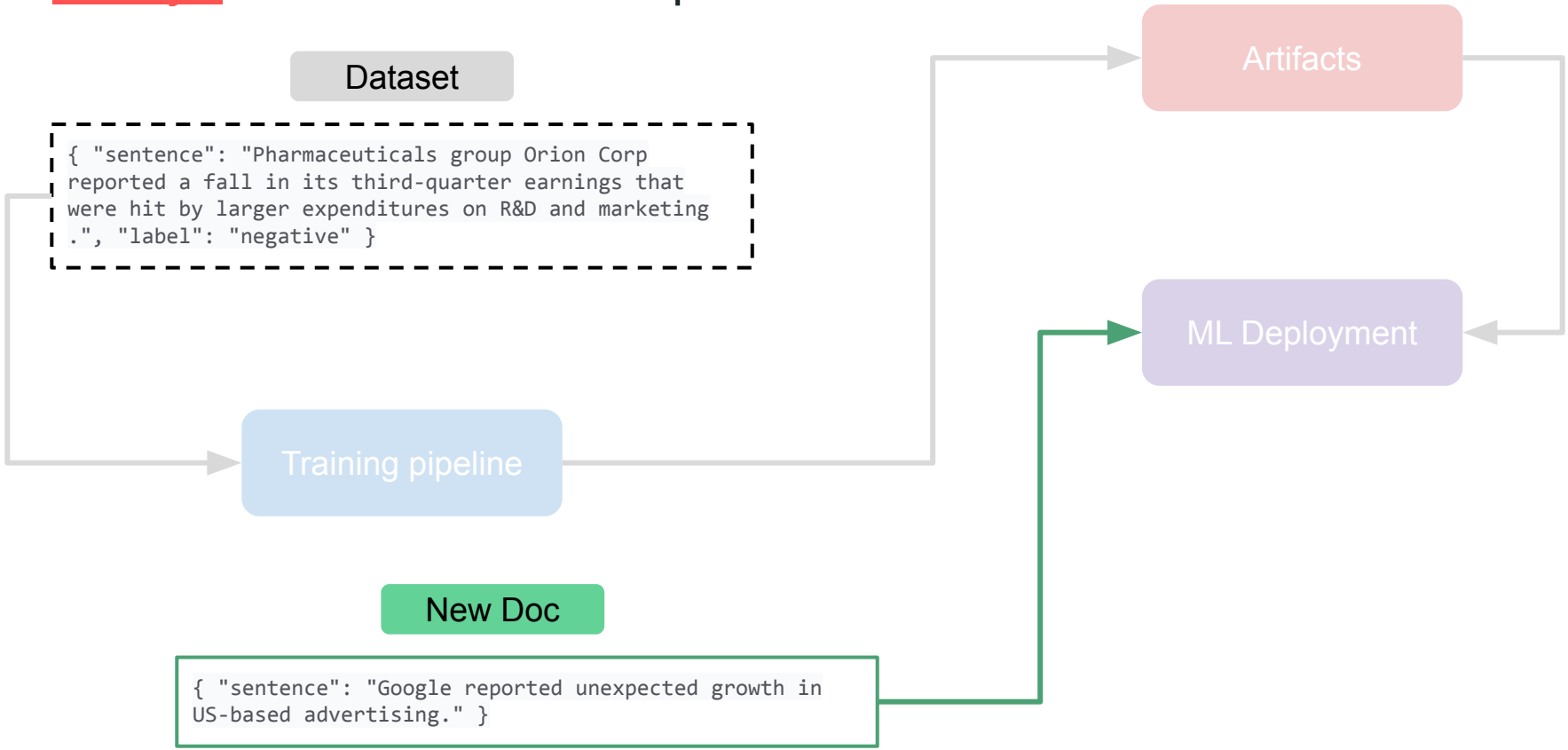
MLSys: A Finance Example



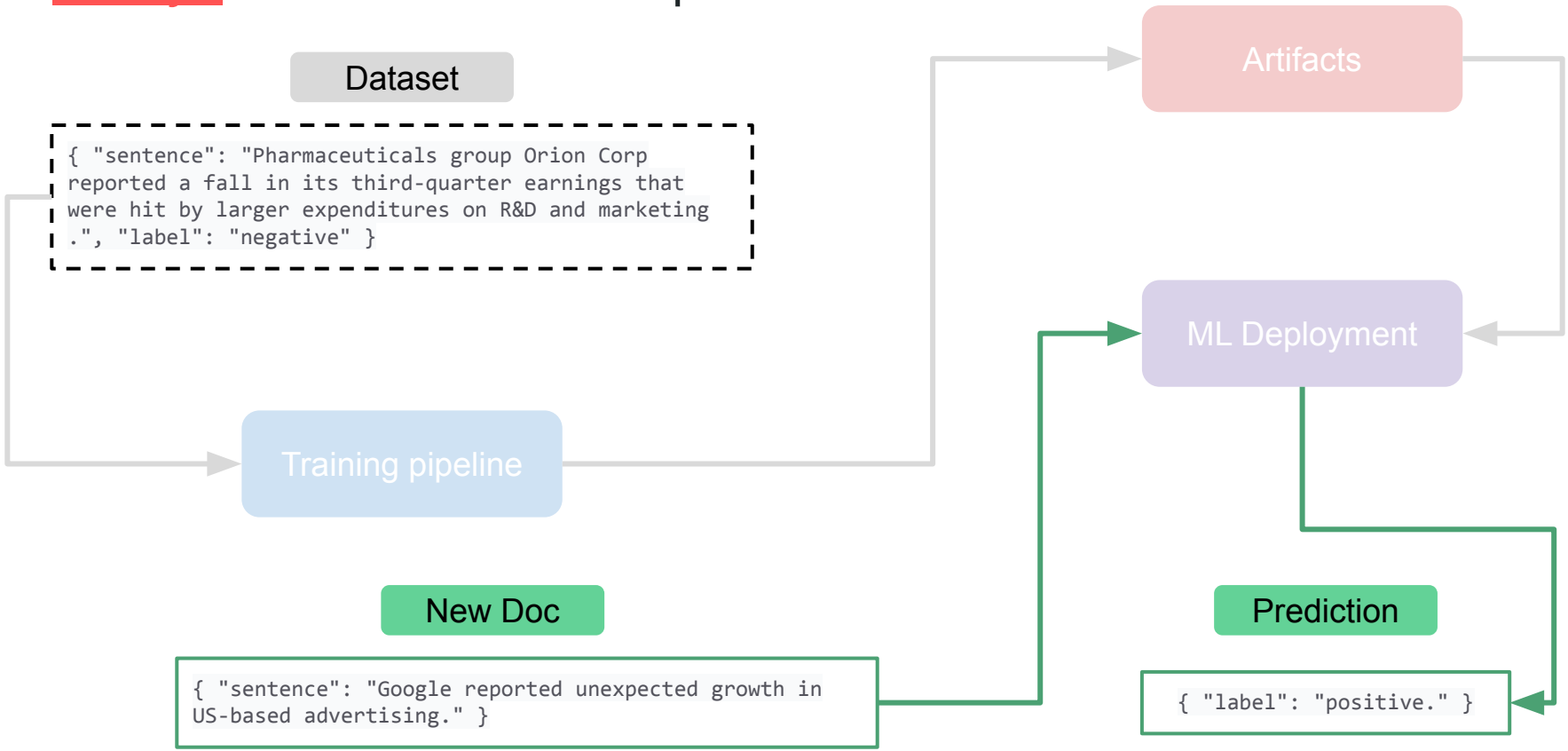
MLSys: A Finance Example



MLSys: A Finance Example



MLSys: A Finance Example



MLSys: Our Plan

- From theory to practice
  - How to organize a ML pipeline
  - Tooling for ML-productivity
- Bonus: are models that important?
  - Semi-supervised labeling
  - Trends in low-code/no-code ML
- Putting it all together
  - An introduction to AWS
  - Serving predictions