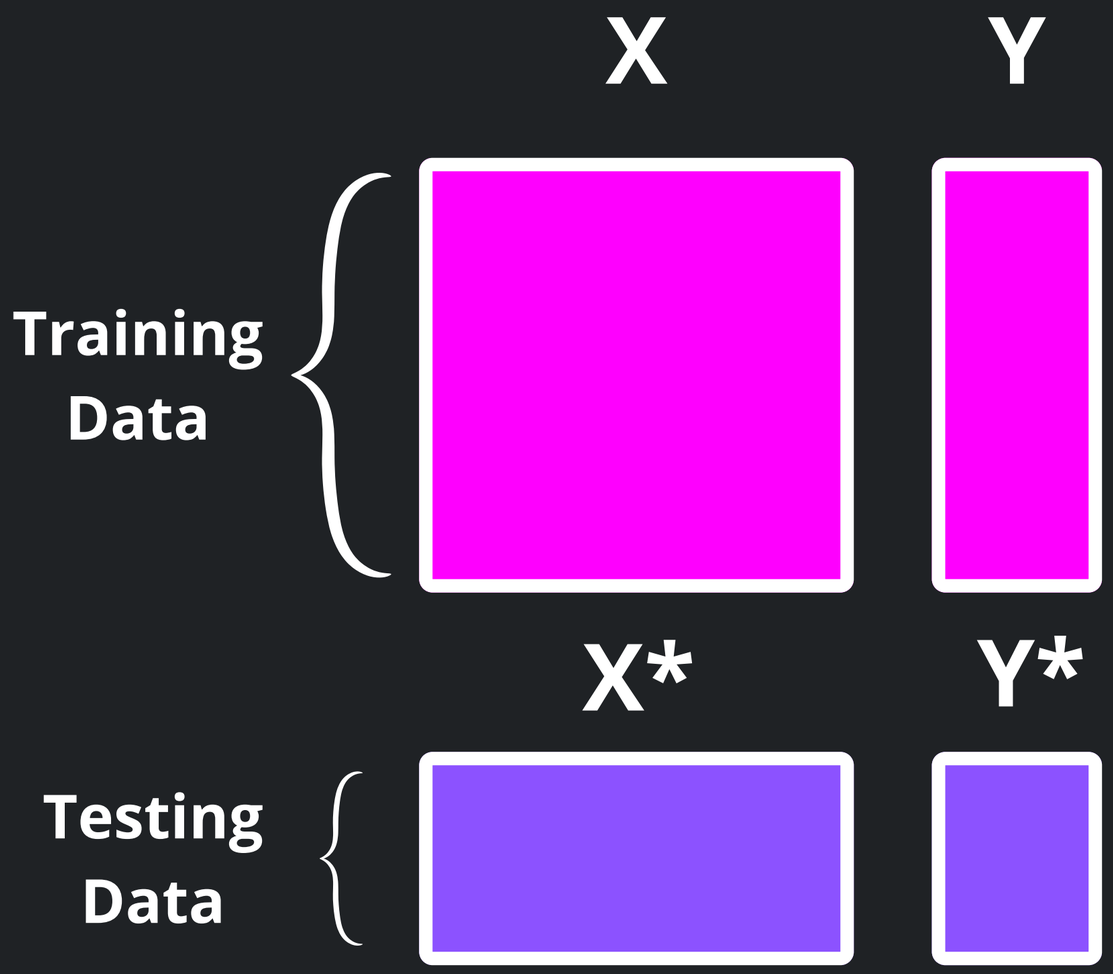
Anote

Label a few, we label the rest

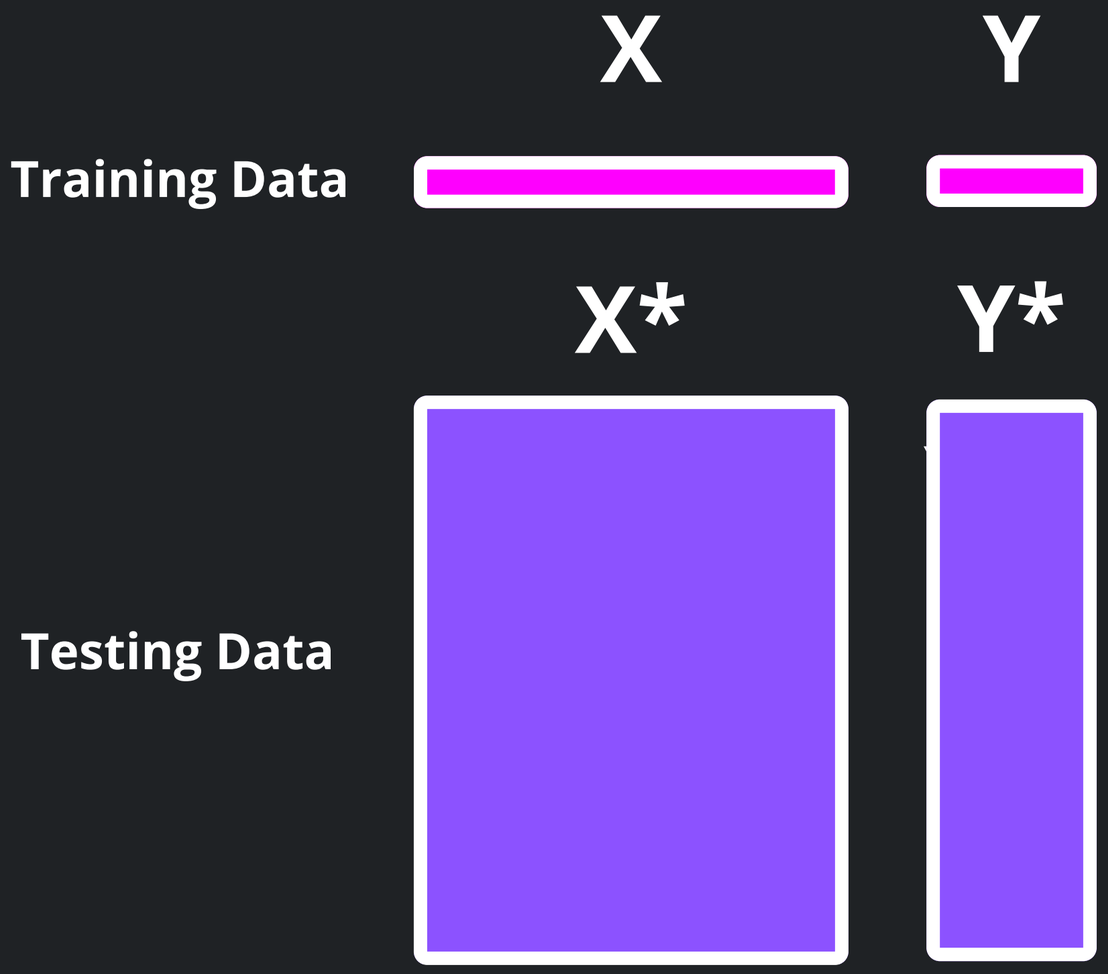
Email: vidranatan@gmail.com
LinkedIn: https://www.linkedin.com/in/natanvidra/

How modern AI/ML works?



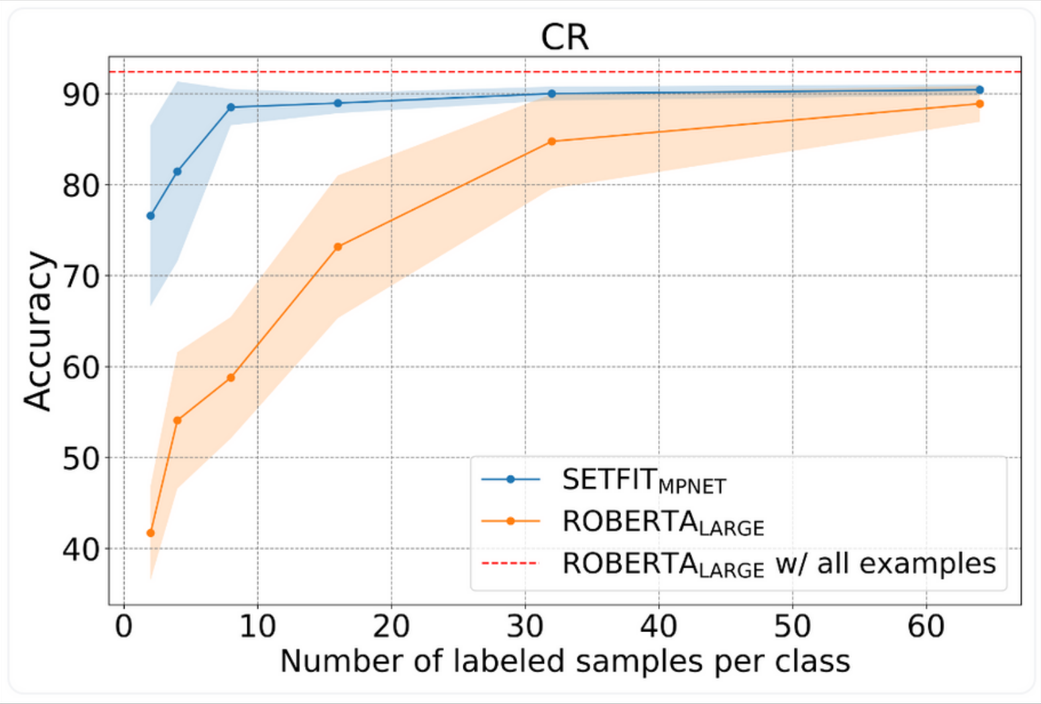
In traditional supervised learning settings, there are millions of rows of training data (X) and corresponding labels (Y). Given new testing data (X*), the goal is to predict a corresponding label (Y*)

How few shot learning works?

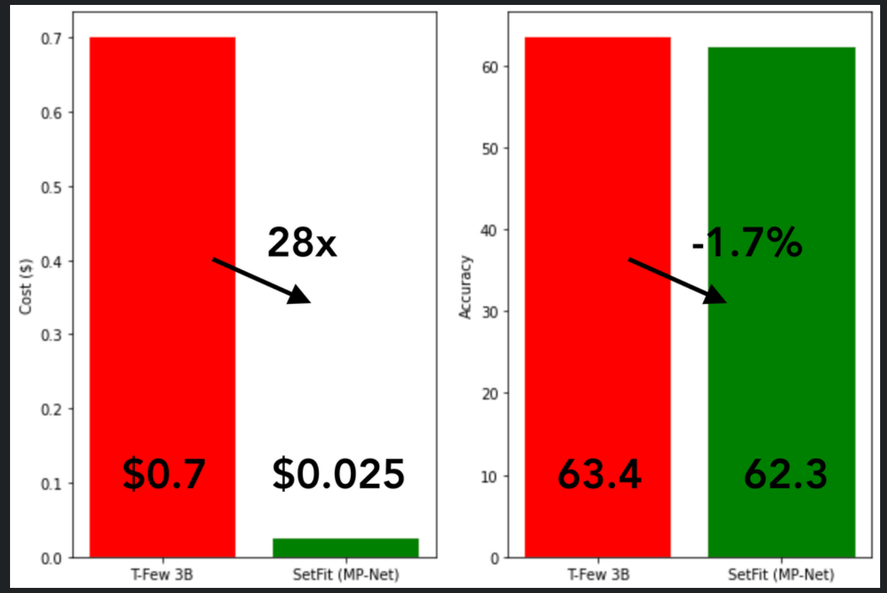


In few shot learning, the goal is to predict label (Y*) with just a few rows of training data (X) and corresponding labels (Y). This requires novel foundation models.

Setfit: Few Shot Learning in practice



Rank	Method	Accuracy	Model Size
2	T-Few	75.8	11B
4	Human Baseline	73.5	N/A
6	SetFit (Roberta Large)	71.3	355M
9	PET	69.6	235M
11	SetFit (MP-Net)	66.9	110M
12	GPT-3	62.7	175 B

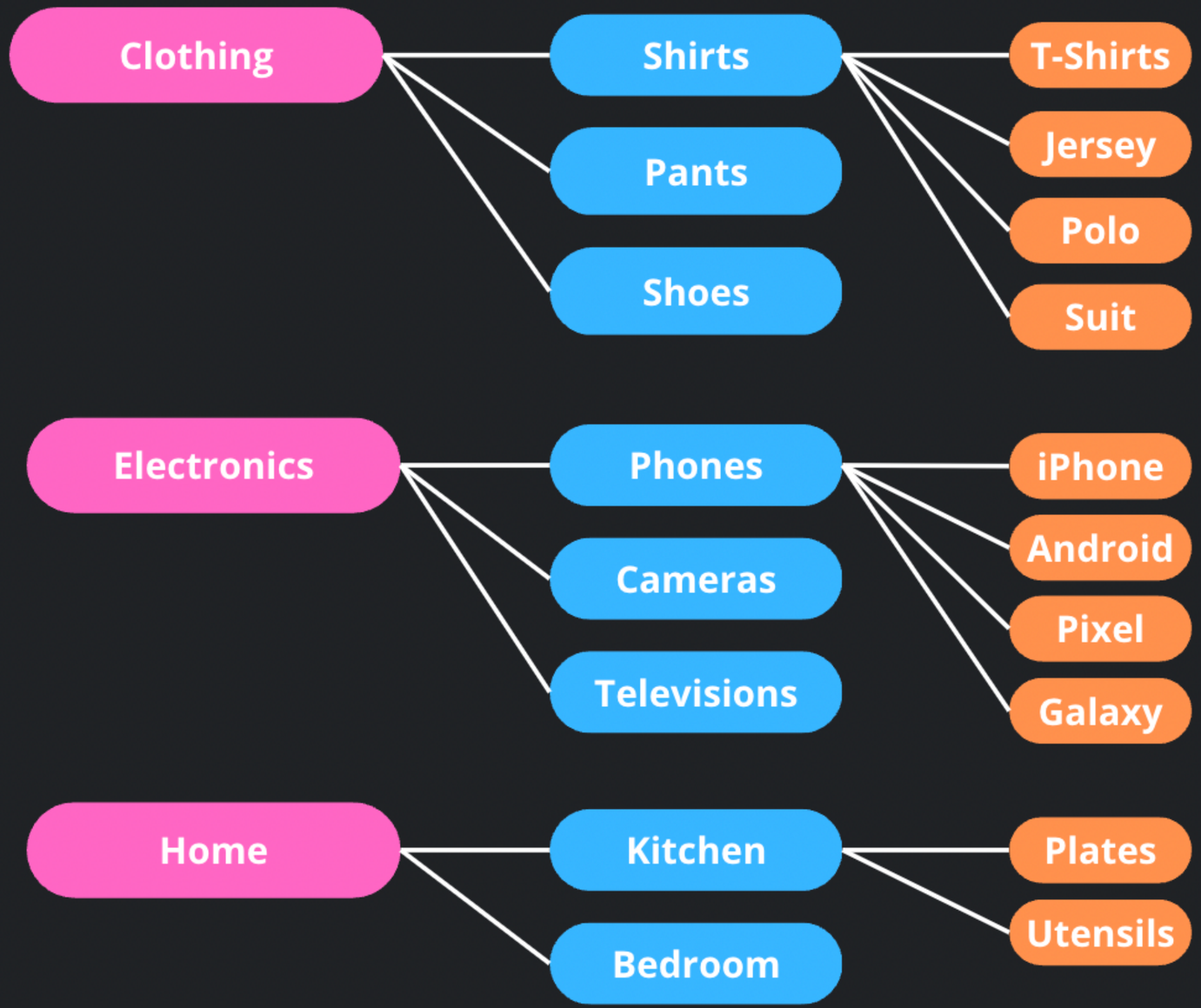


Results:

After just a few labels, you are able to get a high degree of accuracy for certain, specific tasks of simple text classification across many domains and verticals. Results are benchmarked on SST-5, Emotions, AmazonCF, Enron Spam, and Ag News

Can be done: Manual Data Labeling	Can be done: Few Shot Learning	Can not be done: Few Shot Learning	Example of Task
Text Classification	Simple Text Classification	Text Classification	Spam vs Not Spam in Emails
Document Labeling	Simple Document Labeling	Document Labeling	Good vs. Bad Resume
Named Entity Recognition	Simple Sentiment Analysis	Named Entity Recognition	Identifying PII/PHI in Medical Documents
Sentiment Analysis		Sentiment Analysis	Positive vs. Negative Amazon Reviews
Co-referencing		Co-referencing	Research Paper Citations of overlapping linked terms
Part of Speech Tagging		Part of Speech Tagging	Identifying Nouns, Verbs and Adjectives in articles
Entity Linkage		Entity Linkage	Wikipedia Citations of key terms related to eachother
Text Summarization		Text Summarization	Converting 100 page agreements to 2 paragraphs
Conversational AI		Conversational AI	Understanding intent of patient to chatbot
Speaker Diarization		Speaker Diarization	Understanding who is saying what in novel
Entity Disambiguation		Entity Disambiguation	Bank vs. river bank

Few Shot Hierarchical Classification



Input Prompt:

I am at the Apple Store getting a cellphone. I admire Steve Job's vision of the iPhone. How can I get this electronic device at a good price?

Category Prediction

Clothing: Prob = .15
Electronics: Prob = .75
Home: Prob = .1

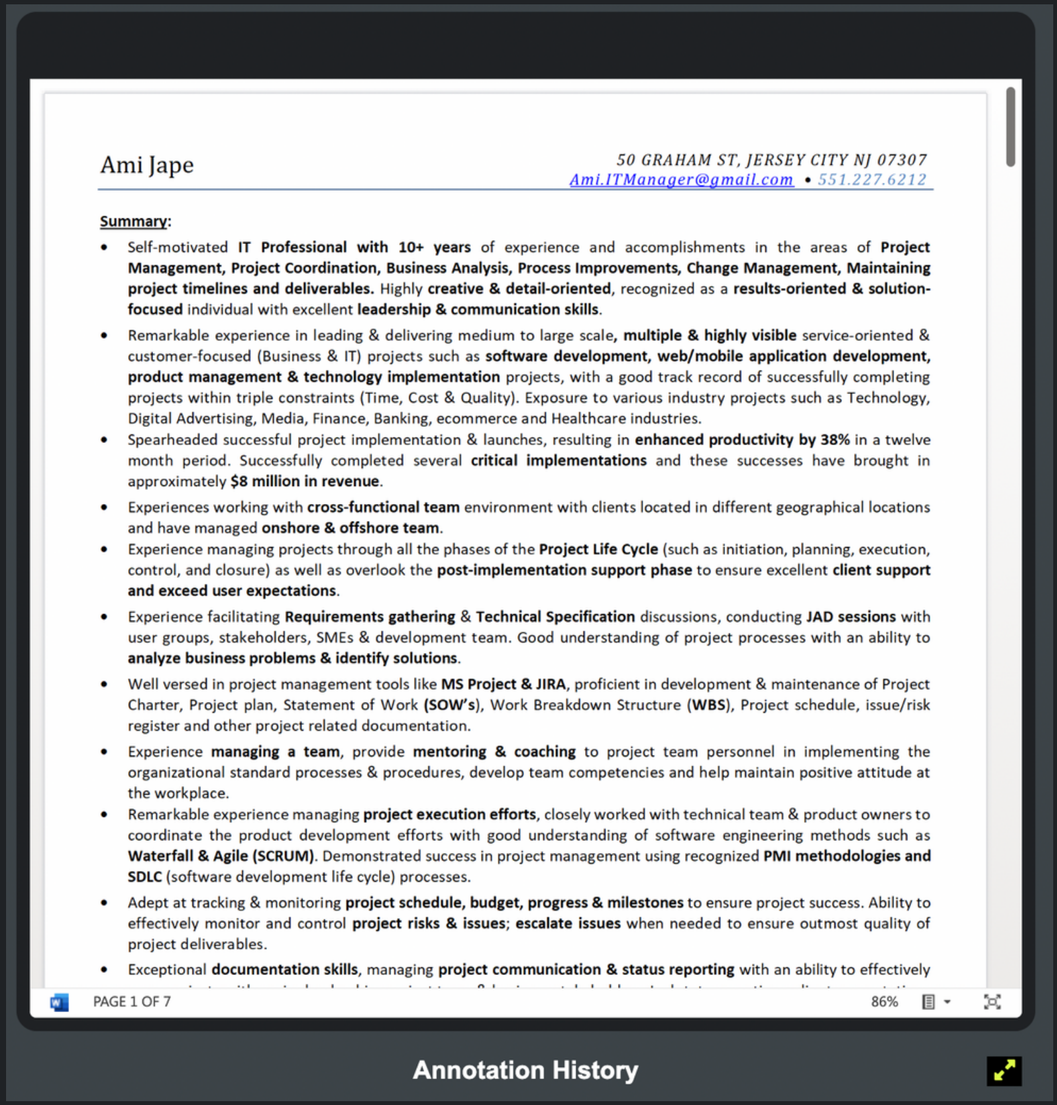
Sub-Category Prediction

Phones: Prob = .8(.75)
Cameras: Prob = .15(.75)
Televisions: Prob = .05(.75)

Sub-Sub-Category Prediction

iPhone: Prob = .65(.8)(.75)
Android: Prob = .2(.8)(.75)
Pixel: Prob = .1(.8)(.75)
Galaxy: Prob = .05(.8)(.75)

Semi-Structured Prompts



Question 1: Summarize this document

- This is a resume about Ami Jape's work experience

Question 2: Do so with formal language in a bulleted list

- Ami Jape has 10+ years of experience
- Ami Jape is strong at managing teams

Question 3: Does this candidate have software engineering experience?

Yes | No | N/A

Confirm

Model Centric AI

Try different models on static datasets

Text Classification

Text	Label
I like bananas	**Fruit**
Broccoli is not good	**Vegetable**
We like tomatoes and potatoes	**Vegetable**

Named Entity Recognition

Entity_Text	Entity_Label
I like bananas	I like bananas **[FRUIT]**
Broccoli is not good	Broccoli **[VEGETABLE]** is not good
We like tomatoes and potatoes	We like tomatoes **[FRUIT]** and potatoes **[VEGETABLE]**

Fruits: {Apple, Orange}
Vegetable: {Zucchini, Spinach}

In model-centric AI, the training data is treated as a fixed input. Your training data is something you download as a static file. New iterations of your project result from changes to the model.

Data Centric AI

Try different datasets on static model

Text Classification

Text	Label
I like bananas	**Fruit**
Broccoli is not good	**Vegetable**
We like tomatoes and cake	**Dessert**
Cookies and milk are great	**Dessert**

Named Entity Recognition

Entity_Text	Entity_Label
I like bananas	I like bananas **[FRUIT]**
Broccoli is not good	Broccoli **[VEGETABLE]** is not good
We like tomatoes and cake	We like tomatoes **[FRUIT]** and cake **[DESSERT]**
Cookies and milk are great	Cookies **[DESSERT]** and milk **[BEVERAGE]** are great

In data-centric AI, data quality and quantity is increasingly the key to successful results. Teams spend more time on labeling, managing, slicing, augmenting, and curating the data, with the model itself relatively more fixed. In data centric AI, you programmatically iterate on your training data.

Fruits: {Apple, Orange}
Vegetable: {Zucchini, Spinach}
Dessert: {Chocolate, Ice Cream}
Beverage: {Water, Juice}

The *"Scale"*-ability Issue

Manual labeling:
As we label data, new classes of data are added due to changing business requirements. Oftentimes, this requires manually relabeling all of the data again, from scratch.

Programmatic labeling:
A solution that works in a specific vertical (i.e. finance) may not translate well to a different vertical (i.e. healthcare). Many labeling applications can't be solved programmatically

Less than **1/1000th** of data that businesses require to be labeled is currently labeled today. As Carl Sagan said, we are just a pale blue dot.

A New Way of Doing AI

Try different datasets and different models in conjunction.

Text Classification

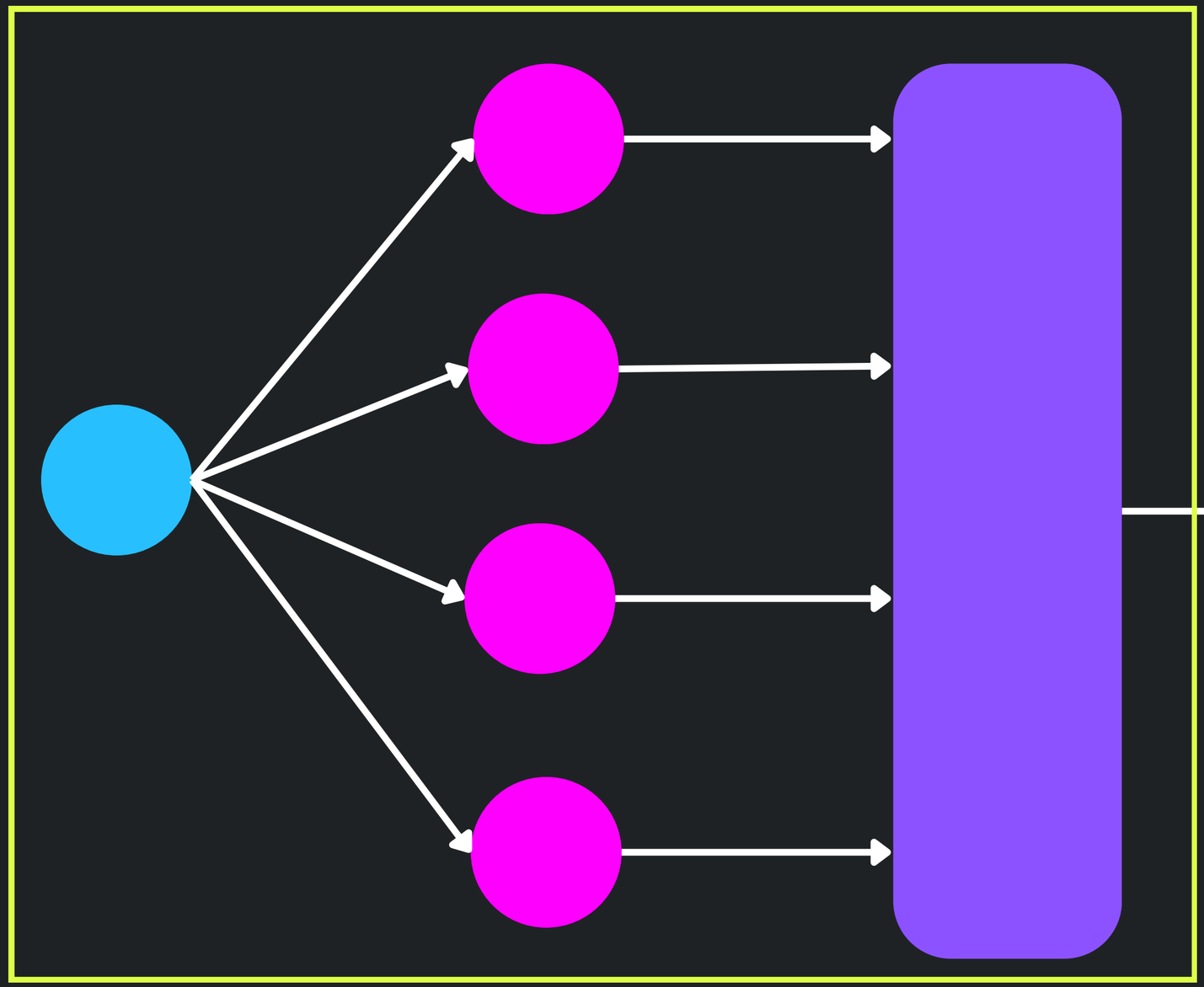
Text	Label
I like bananas	**Fruit**
Broccoli is not good	**Vegetable**
We like tomatoes and cake	**Dessert**
Cookies and milk are great	**Dessert**

Named Entity Recognition

Entity_Text	Entity_Label
I like bananas	I like bananas **[FRUIT]**
Broccoli is not good	Broccoli **[VEGETABLE]** is not good
We like tomatoes and cake	We like tomatoes **[FRUIT]** and cake **[DESSERT]**
Cookies and milk are great	Cookies **[DESSERT]** and milk **[BEVERAGE]** are great

As we iterate on our data, we also iterate on our ensemble model
The ensemble model output is a function of the data input
As we label more, the ensemble model learns which local model is best

Enabling a Scalable Data Labeling Solution



Ensemble Model API Endpoint, Triggered on:

- Annotation
- Important Feature
- Category Added

Each "weak model" runs asynchronously, though to the user they appear to run instantaneously

Model merger joins all "weak learner" models, weighs better performing models higher over time

What applications can this be applied to?

Legal Tech - Classifying Terms in Legal Agreements

A tech-based Law Firm looks over thousands of investor term sheets and acquisition agreements for startups. In each long contract, there is important information that lawyers are paid ~$1,000 per hour to identify, detect, and note to startups. For many startups, going to law firms can be extremely expensive, and well outside of their budget, but these startups still need legal counsel on specific documents.

Research - Co-referencing in Research Papers

A research institute at a top-tier university wanted to understand relationships between phenotypes of species in research papers. They needed a tool to go through research papers and annotate co-references, named entities and key words to build this model. Their data labeling process was relatively manual, as existing labeling tool did not work well and satisfy their use case.

Miscellaneous - Unsubscribe From Emails

A large organization has employees that constantly receive annoying emails which they want to unsubscribe from. The organization mentioned that each employee spends on average 15 minutes a day sifting through and deleting irrelevant emails to their mailbox. The organization wanted to build a tool where employees could unsubscribe from a few emails, and the AI would automatically recommend the remaining emails from which employees should unsubscribe from. That would save employees time, and keep their mailbox clean.

Finance - Document Information Extraction

A large finance company has a team of 20-30 data annotators (analysts) out of Princeton, NJ, in charge of extracting key points from financial documents. These documents are normally in the form of PDFs, and some of the information exists in tables and pictures. The companies research team has looked into AI techniques such as OCR (Optical Character Recognition) to extract this information from documents, though the information extraction is predominantly still done manually.

Education - Writing Proposals

A technical team was looking for funding for research projects, and oftentimes needed to write proposals. However, when writing proposals, it oftentimes was unclear whether what employees were writing was beneficial content for their proposal being accepted. They needed labeled data describing whether certain sentences helped or hurt employees chance's of their proposal being accepted.

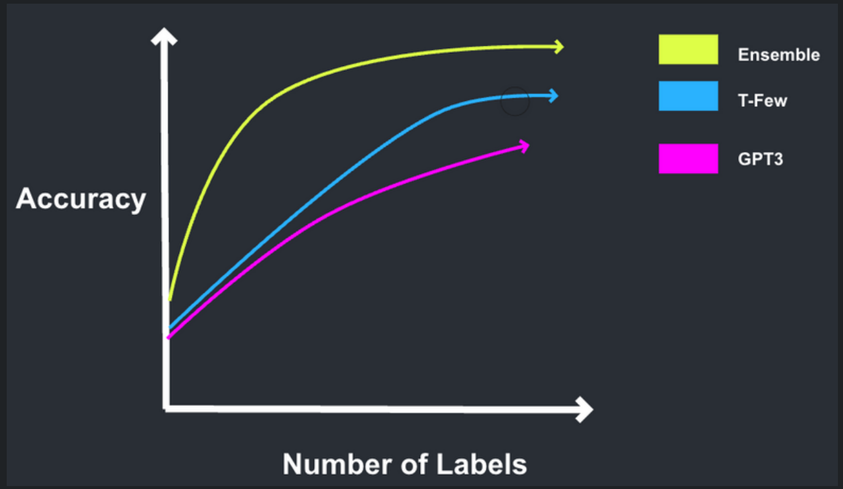
Healthcare - Classifying Cancer

A large hospital in New York goes through radiology reports to classify certain types of cancer present in the report. They attempted to build an AI text classification model for this task, using BERT, but struggled with an imbalanced dataset (lack of labeled training data). Regarding their data distribution, the team had 250 classes of cancer that were present relatively infrequently / only appeared a small fraction of the time, whereas the most prominent cancer types occurred very frequently in the reports. Because of this, the team looked into heuristic models to train their system, as well as alternative data labeling mechanisms.

Non Profit - Identifying Violent Tweets

A non-profit that was looking to analyze social media data from twitter for public safety purposes in New York City. They wanted to build a model that scrapes data off of twitter to analyze if a specific tweet in the New York City area is violent. To do this, they need annotated data from the tweets to see whether specific tweets correspond to "violent" or "not violent", that way they could train their AI model.

Few Shot Learning



After you label just a few row of data (the edge cases), we label the rest, using state of the art transformer models

Programmatic Labeling

Heuristics such as key words, entities and regex expressions are fed into the model

Decomposition

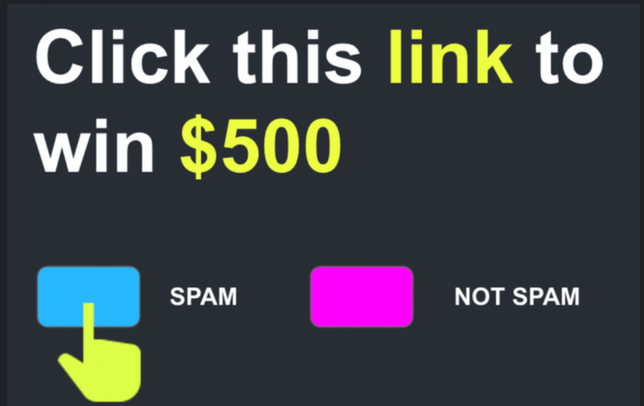
Convert unstructured text data to a spreadsheet

Synchronous

Index	Text	Label	Probability
0	Click this link to win $500	SPAM	.98
1	This is a normal sentence	NOT SPAM	.89
2	You just won a free iPhone	SPAM	.93
3	Hope you are doing well	NOT SPAM	.87
4	Enter your credit card info	SPAM	.97

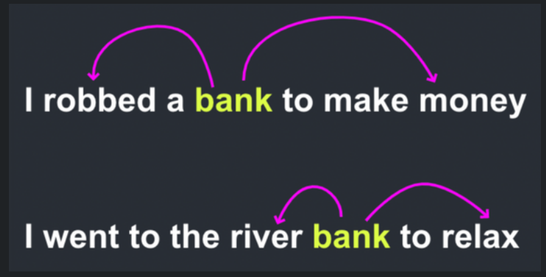
Get your data labeled in real time, and receive immediate feedback on the model's performance as you label data

Human in the Loop



We actively learn from subject matter experts, who provide input into our model

Contextual



Our large language models extract the context of words in sentences, not just the individual words themselves