



# TensorFlow

## Milestone Project 2

### SkimLit



# Where can you get help?

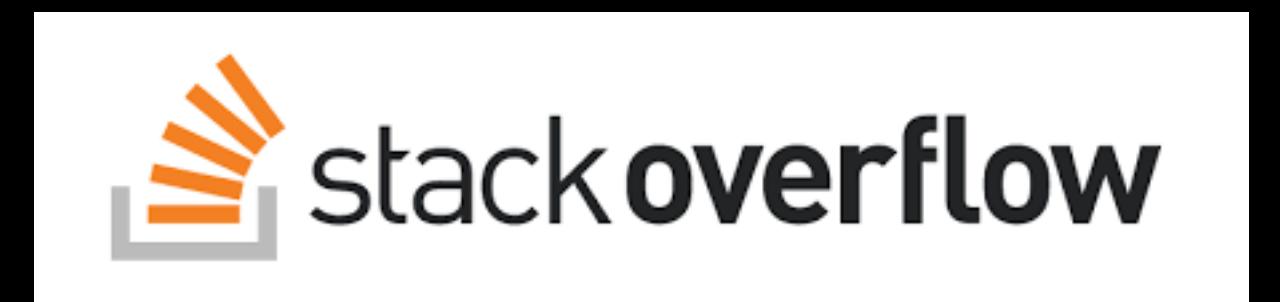
- Follow along with the code



- Try it for yourself

"If in doubt, run the code"

- Press SHIFT + CMD + SPACE to read the docstring



```
def get_lines(filename):
    """this is the second line of the docstring
    reads filename (a text file) and returns the lines of text as a list.
    Args:
        filename: a string containing the target filename.
    Returns:
        A list of strings with one string per line from the target filename.
    Example:
        ["this is the first line of filename",
         "this is the second line of filename",
         "..."]
```

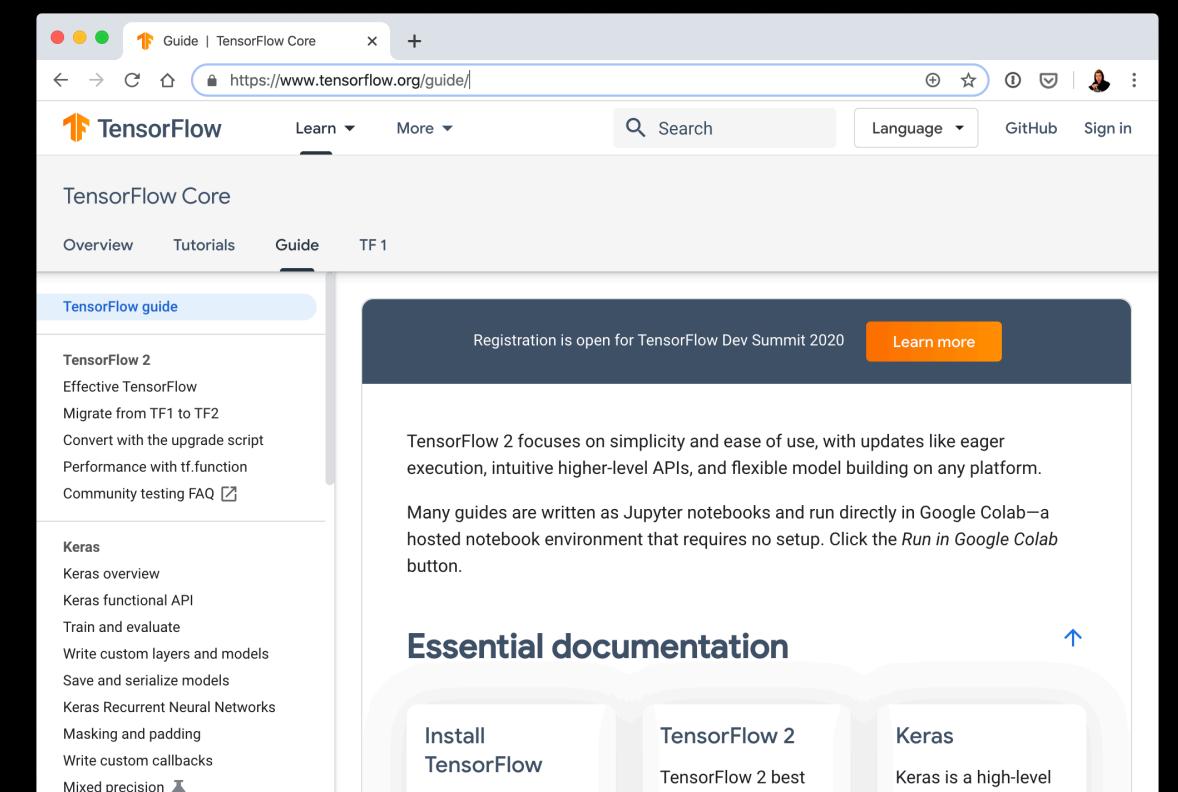
- Search for it



- Try again

- Ask (don't forget the Discord chat!)

(yes, including the “dumb”  
questions)



# Example NLP problems

and NLU... (natural language understanding)

“What tags should this article have?”

The screenshot shows the Wikipedia article on "Deep learning". Below the main content, there is a sidebar with a section titled "Machine learning" which includes "Representation learning" and "Artificial intelligence". A handwritten note "(multiple label options per sample)" is overlaid on this sidebar.

Classification

The screenshot shows the Google Translate interface translating the Latin phrase "acta non verba" from Latin to English. The English translation is "actions not words". A handwritten note "(multiple label options per sample)" is overlaid on the sidebar of the translate interface.

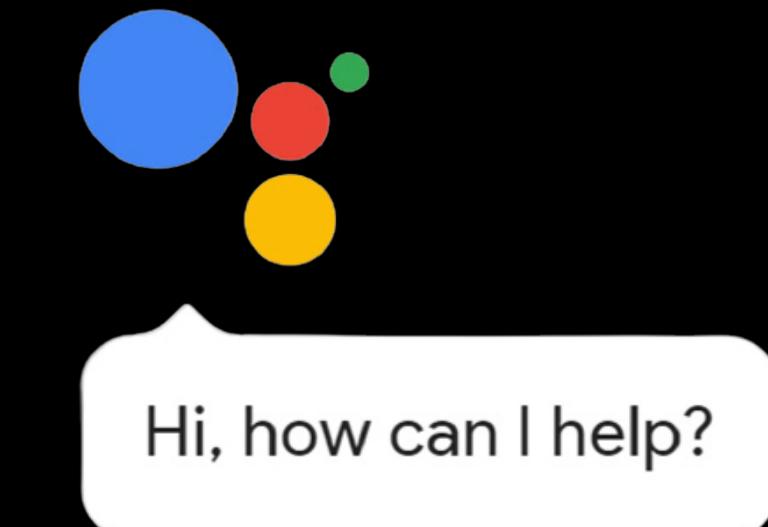
Machine Translation

These are also referred to as sequence problems

The screenshot shows a text generation example from William Shakespeare's "Pandarus". The text reads: "Alas, I think he shall be come approached and the day When little strain would be attain'd into being never fed, And who is but a chain and subjects of his death, I should not sleep." A handwritten note "(multiple label options per sample)" is overlaid on the sidebar of the text generation interface.

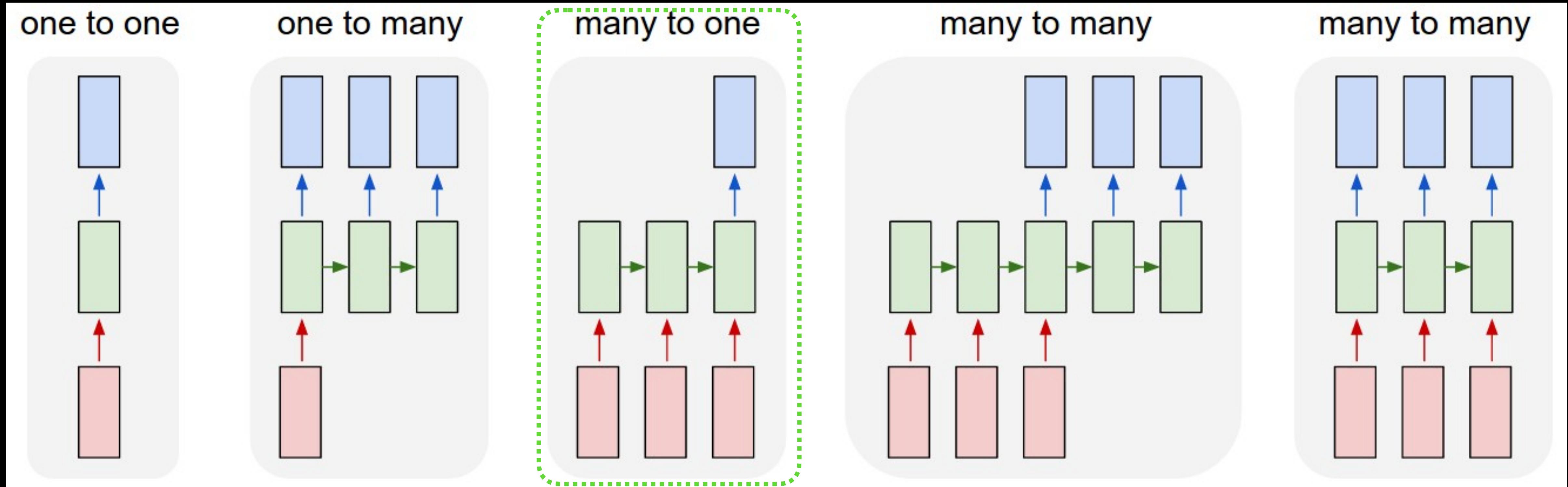
Source: <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Text Generation



Voice Assistants

# Sequence problems



Source: <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Let's make PubMed abstracts  
easier to read!

# What we're doing

Replicating  
this paper

The screenshot shows a web browser displaying an arXiv.org page. The title of the page is "arXiv.org > cs > arXiv:1710.06071". The main content is an academic paper titled "PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts" by Franck Dernoncourt and Ji Young Lee. The abstract discusses the creation of a new dataset based on PubMed for sequential sentence classification, consisting of approximately 200,000 abstracts of randomized controlled trials. The paper is submitted on 17 Oct 2017. The right sidebar contains sections for "Download" (PDF, Other formats), "Current browse context" (cs.CL), "Change to browse by" (links to cs, cs.AI, stat, stat.ML), "References & Citations" (NASA ADS, Google Scholar, Semantic Scholar), and "DBLP – CS Bibliography" (listing, bibtex, links to Franck Dernoncourt and Ji Young Lee). The footer of the page includes a link to "Export Bibtex Citation".

[1710.06071] PubMed 200k RC

arXiv.org /abs/1710.06071

Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv.org > cs > arXiv:1710.06071

Search... All fields Search

Help | Advanced Search

Computer Science > Computation and Language

[Submitted on 17 Oct 2017]

## PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts

Franck Dernoncourt, Ji Young Lee

We present PubMed 200k RCT, a new dataset based on PubMed for sequential sentence classification. The dataset consists of approximately 200,000 abstracts of randomized controlled trials, totaling 2.3 million sentences. Each sentence of each abstract is labeled with their role in the abstract using one of the following classes: background, objective, method, result, or conclusion. The purpose of releasing this dataset is twofold. First, the majority of datasets for sequential short-text classification (i.e., classification of short texts that appear in sequences) are small: we hope that releasing a new large dataset will help develop more accurate algorithms for this task. Second, from an application perspective, researchers need better tools to efficiently skim through the literature. Automatically classifying each sentence in an abstract would help researchers read abstracts more efficiently, especially in fields where abstracts may be long, such as the medical field.

Comments: Accepted as a conference paper at IJCNLP 2017

Subjects: Computation and Language (cs.CL); Artificial Intelligence (cs.AI); Machine Learning (stat.ML)

Cite as: arXiv:1710.06071 [cs.CL]  
(or arXiv:1710.06071v1 [cs.CL] for this version)

Download:

- PDF
- Other formats (license)

Current browse context: cs.CL

< prev | next >

new | recent | 1710

Change to browse by:

cs

cs.AI

stat

stat.ML

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

DBLP – CS Bibliography

listing | bibtex

Franck Dernoncourt  
Ji Young Lee

Export Bibtex Citation

Source: <https://arxiv.org/abs/1710.06071>

```
59 Machine Learning Engineer*
60 -----
61
62 1. Download a paper
63 2. Implement it
64 3. Keep doing this until you have skills
```

- George Hotz

\*Machine Learning Engineering also involves building infrastructure around your models/  
data preprocessing steps

# What we're doing

## Nutritional psychiatry: the present state of the evidence

Wolfgang Marx <sup>1</sup>, Genevieve Moseley <sup>2</sup>, Michael Berk <sup>2</sup>, Felice Jacka <sup>2</sup>

Affiliations + expand

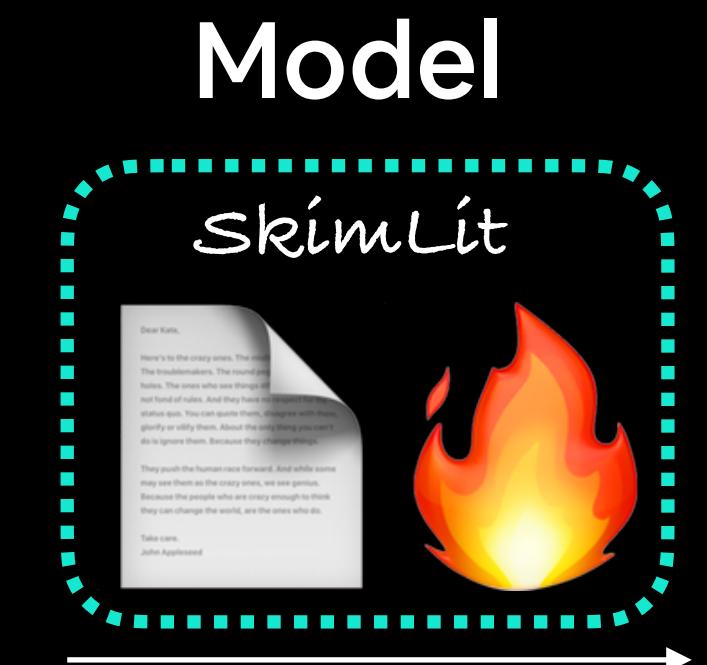
PMID: 28942748 DOI: [10.1017/S0029665117002026](https://doi.org/10.1017/S0029665117002026)

### Abstract

Mental illness, including depression, anxiety and bipolar disorder, accounts for a significant proportion of global disability and poses a substantial social, economic and health burden. Treatment is presently dominated by pharmacotherapy, such as antidepressants, and psychotherapy, such as cognitive behavioural therapy; however, such treatments avert less than half of the disease burden, suggesting that additional strategies are needed to prevent and treat mental disorders. There are now consistent mechanistic, observational and interventional data to suggest diet quality may be a modifiable risk factor for mental illness. This review provides an overview of the nutritional psychiatry field. It includes a discussion of the neurobiological mechanisms likely modulated by diet, the use of dietary and nutraceutical interventions in mental disorders, and recommendations for further research. Potential biological pathways related to mental disorders include inflammation, oxidative stress, the gut microbiome, epigenetic modifications and neuroplasticity. Consistent epidemiological evidence, particularly for depression, suggests an association between measures of diet quality and mental health, across multiple populations and age groups; these do not appear to be explained by other demographic, lifestyle factors or reverse causality. Our recently published intervention trial provides preliminary clinical evidence that dietary interventions in clinically diagnosed populations are feasible and can provide significant clinical benefit. Furthermore, nutraceuticals including n-3 fatty acids, folate, S-adenosylmethionine, N-acetyl cysteine and probiotics, among others, are promising avenues for future research. Continued research is now required to investigate the efficacy of intervention studies in large cohorts and within clinically relevant populations, particularly in patients with schizophrenia, bipolar and anxiety disorders.

Source: <https://pubmed.ncbi.nlm.nih.gov/28942748/>

Harder to read



## Considerations for a surgical RCT for diffuse low-grade glioma: a survey

Alireza Mansouri <sup>1</sup>, Karanbir Brar <sup>2</sup>, Michael D Cusimano <sup>3</sup>

Affiliations + expand

PMID: 32537182 PMCID: PMC7274180 (available on 2021-06-01) DOI: [10.1093/nop/npz058](https://doi.org/10.1093/nop/npz058)

### Abstract

**Background:** Diffuse low-grade gliomas (DLGGs) are heterogeneous tumors that inevitably differentiate into malignant entities, leading to disability and death. Recently, a shift toward up-front maximal safe resection of DLGGs has been favored. However, this transition is not supported by randomized controlled trial (RCT) data. Here, we sought to survey the neuro-oncology community on considerations for a surgical RCT for DLGGs.

**Methods:** A 21-question survey focusing on a surgical RCT for DLGGs was developed and validated by 2 neurosurgeons. A sample case of a patient for whom management might be debatable was presented to gather additional insight. The survey was disseminated to members of the Society for Neuro-Oncology (SNO) and responses were collected from March 16 to July 10, 2018.

**Results:** A total of 131 responses were collected. Sixty-three of 117 (54%) respondents thought an RCT would not be ethical, 39 of 117 (33%) would consider participating, and 56 of 117 (48%) believed an RCT would be valuable for determining the differing roles of biopsy, surgery, and observation. This was exemplified by an evenly distributed selection of the latter management options for our sample case. Eighty-three of 120 (69.2%) respondents did not believe in equipoise for DLGG patients. Quality of life and overall survival were deemed equally important end points for a putative RCT.

**Conclusions:** Based on our survey, it is evident that management of certain DLGG patients is not well defined and an RCT may be justified. As with any surgical RCT, logistic challenges are anticipated. Robust patient-relevant end points and standardization of perioperative adjuncts are necessary if a surgical RCT is undertaken.

Source: <https://pubmed.ncbi.nlm.nih.gov/32537182/>

Easier to read  
(Skimmable)

# Raw abstract text (numbers have been replaced with "@")

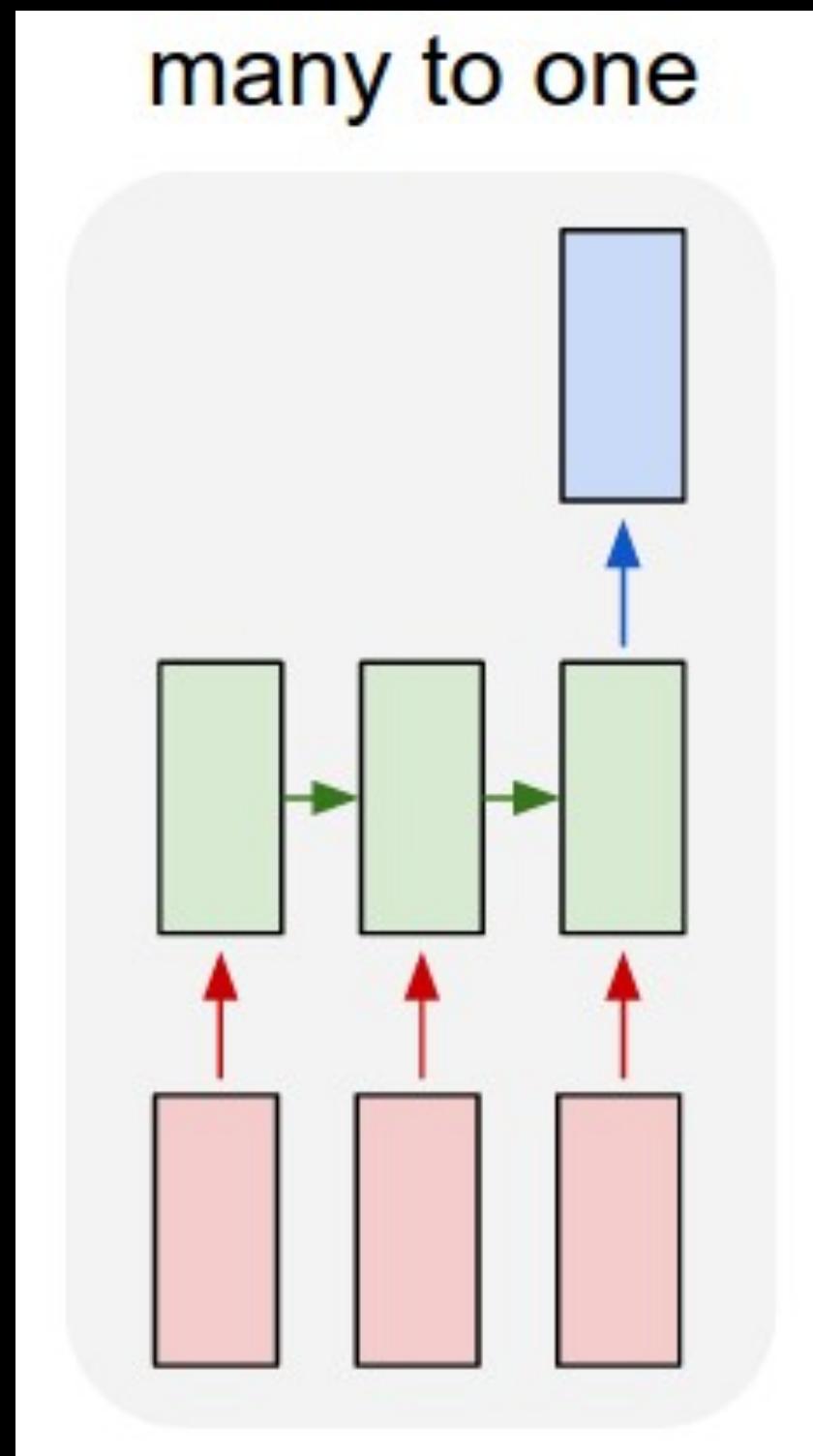
To investigate the efficacy of @ weeks of daily low-dose oral prednisolone in improving pain , mobility , and systemic low-grade inflammation in the short term and whether the effect would be sustained at @ weeks in older adults with moderate to severe knee osteoarthritis ( OA ). A total of @ patients with primary knee OA were randomized @:@ ; @ received @ mg/day of prednisolone and @ received placebo for @ weeks. Outcome measures included pain reduction and improvement in function scores and systemic inflammation markers. Pain was assessed using the visual analog pain scale ( @-@ mm ). Secondary outcome measures included the Western Ontario and McMaster Universities Osteoarthritis Index scores , patient global assessment ( PGA ) of the severity of knee OA , and @-min walk distance ( @MWD ), Serum levels of interleukin @ ( IL-@ ) , IL-@ , tumor necrosis factor ( TNF ) - , and high-sensitivity C-reactive protein ( hsCRP ) were measured. There was a clinically relevant reduction in the intervention group compared to the placebo group for knee pain , physical function , PGA , and @MWD at @ weeks. The mean difference between treatment arms ( @ % CI ) was @ ( @-@ @ ) , p < @ ; @ ( @-@ @ ) , p < @ ; @ ( @-@ @ ) , p < @ ; and @ ( @-@ @ ) , p < @ , respectively. Further , there was a clinically relevant reduction in the serum levels of IL-@ , IL-@ , TNF - , and hsCRP at @ weeks in the intervention group when compared to the placebo group. These differences remained significant at @ weeks. The Outcome Measures in Rheumatology Clinical Trials-Osteoarthritis Research Society International responder rate was @ % in the intervention group and @ % in the placebo group ( p < @ ). Low-dose oral prednisolone had both a short-term and a longer sustained effect resulting in less knee pain , better physical function , and attenuation of systemic inflammation in older patients with knee OA ( ClinicalTrials.gov identifier NCT@ ).

↓ Model SkimLit 🔥 ↓

```
[###24293578\n',
'OBJECTIVE\tTo investigate the efficacy of @ weeks of daily low-dose oral prednisolone in improving pain , mobility ,
and systemic low-grade inflammation in the short term and whether the effect would be sustained at @ weeks in
older adults with moderate to severe knee osteoarthritis ( OA ) .\n',
'METHODS\tA total of @ patients with primary knee OA were randomized @:@ ; @ received @ mg/day of prednisolone and @ received
placebo for @ weeks .\n',
'METHODS\tOutcome measures included pain reduction and improvement in function scores and systemic inflammation markers .\n',
'METHODS\tPain was assessed using the visual analog pain scale ( @-@ mm ) .\n',
'METHODS\tSecondary outcome measures included the Western Ontario and McMaster Universities Osteoarthritis Index scores ,
patient global assessment ( PGA ) of the severity of knee OA , and @-min walk distance ( @MWD ) .\n',
'METHODS\tSerum levels of interleukin @ ( IL-@ ) , IL-@ , tumor necrosis factor ( TNF ) - , and high-sensitivity C-reactive
protein ( hsCRP ) were measured .\n',
'RESULTS\tThere was a clinically relevant reduction in the intervention group compared to the placebo group for knee pain ,
physical function , PGA , and @MWD at @ weeks .\n',
'RESULTS\tThe mean difference between treatment arms ( @ % CI ) was @ ( @-@ @ ) , p < @ ; @ ( @-@ @ ) , p < @ ; @ ( @-@ @ ) ,
p < @ ; and @ ( @-@ @ ) , p < @ , respectively .\n',
'RESULTS\tFurther , there was a clinically relevant reduction in the serum levels of IL-@ , IL-@ , TNF - , and hsCRP at @
weeks in the intervention group when compared to the placebo group .\n',
'RESULTS\tThese differences remained significant at @ weeks .\n',
'RESULTS\tThe Outcome Measures in Rheumatology Clinical Trials-Osteoarthritis Research Society International responder
rate was @ % in the intervention group and @ % in the placebo group ( p < @ ) .\n',
'CONCLUSIONS\tLow-dose oral prednisolone had both a short-term and a longer sustained effect resulting in less knee pain ,
better physical function , and attenuation of systemic inflammation in older patients with knee OA
( ClinicalTrials.gov identifier NCT@ ) .\n']
```

# Model predicts topic of each sequence

# SkimLit sequence problem: many to one



Source: <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Input

Output

To investigate the efficacy of @ weeks of daily low-dose oral prednisolone in improving pain , mobility , and systemic low-grade inflammation in the short term and whether the effect would be sustained at @ week in older adults with moderate to severe knee osteoarthritis ( OA ).

**OBJECTIVE**

# What we're going to cover

(broadly)

- Downloading a text dataset (PubMed 200K RCT)
- Writing a **preprocessing function** for our text data
- Setting up multiple modelling experiments with **different levels of embeddings**
- Building a **multimodal model** to take in different sources of data
  - Replicating the model powering <https://arxiv.org/abs/1710.06071>
  - Finding the most wrong prediction examples

(we'll be cooking up lots of code!)

How:



# NLP inputs and outputs



“Is this Tweet for a disaster or not?”



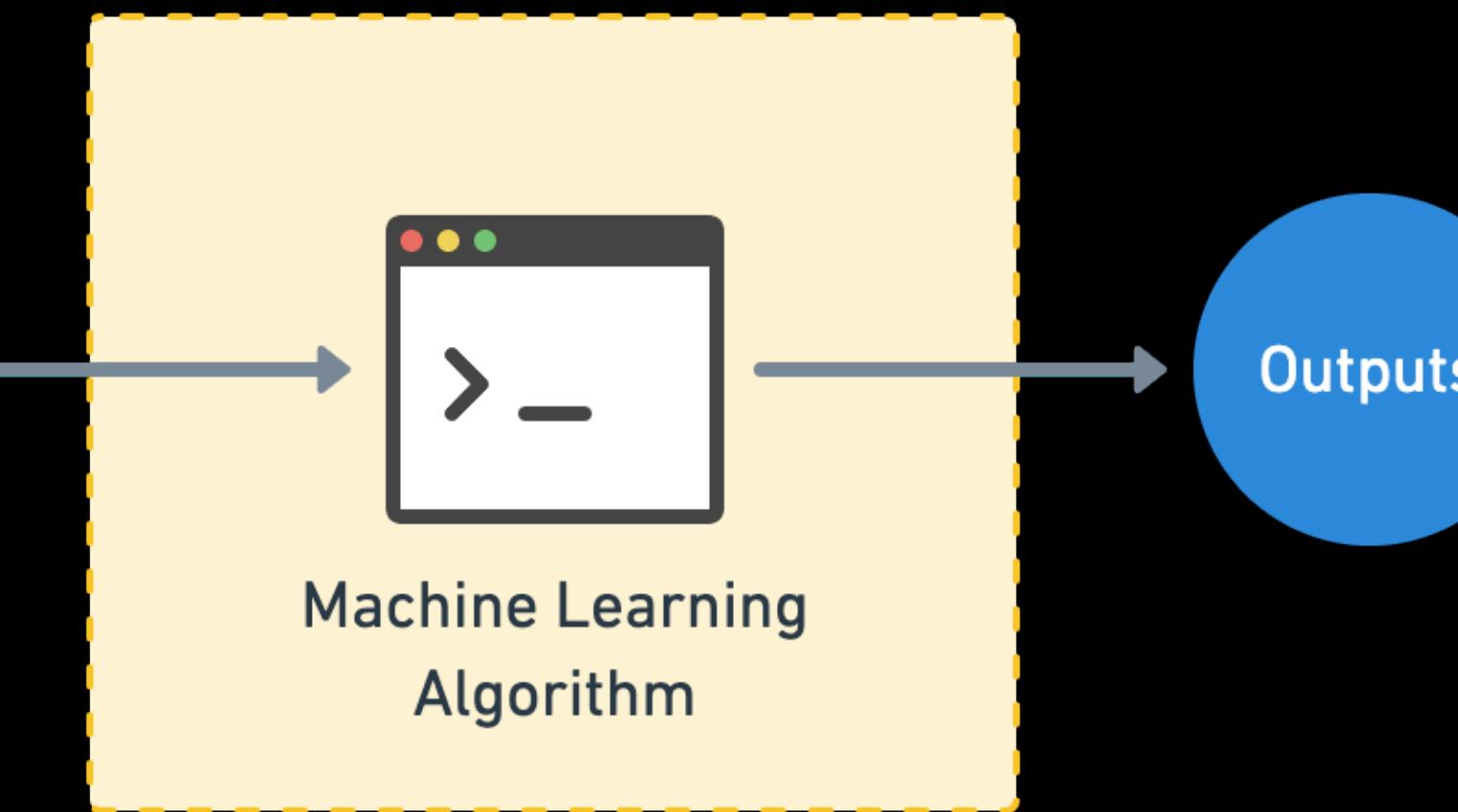
Diaster

Not Diaster

Actual output

$[[0.22, 0.98, 0.02...],$   
 $[0.09, 0.55, 0.87...], \rightarrow$   
 $[0.53, 0.81, 0.79...],$   
 $\dots,$

**Numerical encoding**  
(Tokenization + Embedding)



(often already exists, if not,  
you can build one)

$[[0.97, 0.03],$   
 $[0.81, 0.19],$   
 $\dots,$

**Predicted output**

(comes from looking at lots  
of these)

# SkimLit inputs and outputs

To investigate the efficacy of @ weeks of daily low-dose oral prednisolone in improving pain , mobility , and systemic low-grade inflammation in the short term and whether → the effect would be sustained at @ week in older adults with moderate to sever knee osteoarthritis ( OA ).

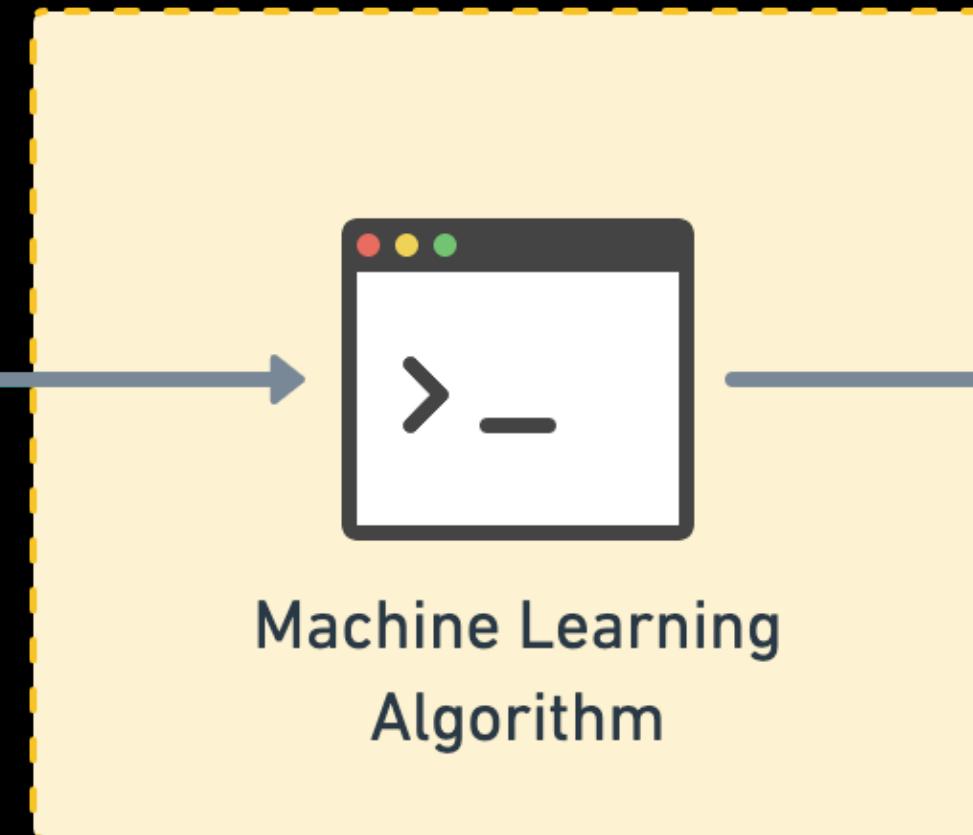
“What section should this sentence belong to?”

[ [ 0.22, 0.98, 0.02... ] ,  
[ 0.09, 0.55, 0.87... ] , →  
[ 0.53, 0.81, 0.79... ] ,  
...,

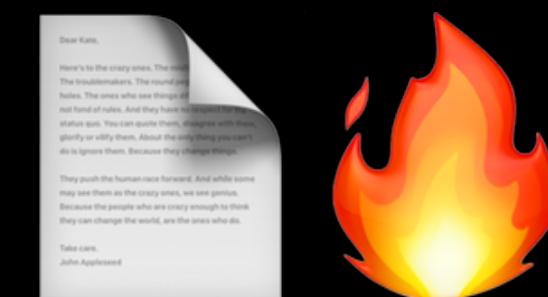
## Numerical encoding

(Tokenization + Embedding)

Inputs



(often already exists, if not,  
you can build one)



SkimLit



BACKGROUND  
OBJECTIVE  
METHODS  
RESULTS  
CONCLUSION

Actual output

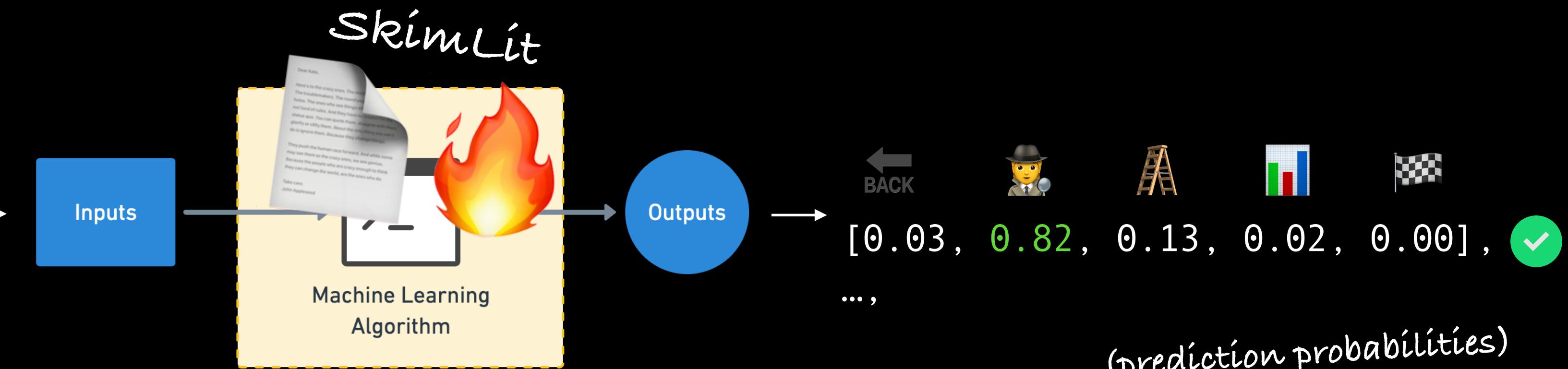
BACK ← BACKPACKER 🔧📊🏁 → [ 0.03, 0.82, 0.13, 0.02, 0.00 ] ,  
...,

## Predicted output

(comes from looking at lots  
of these)

# SkimLit input and output shapes

To investigate the efficacy of @ weeks of daily low-dose oral prednisolone in improving pain , mobility , and systemic low-grade inflammation in the short term and whether the effect would be sustained at @ week in older adults with moderate to sever knee osteoarthritis ( OA ).



[batch\_size, embedding\_size]  
(gets represented as a tensor/embedding)

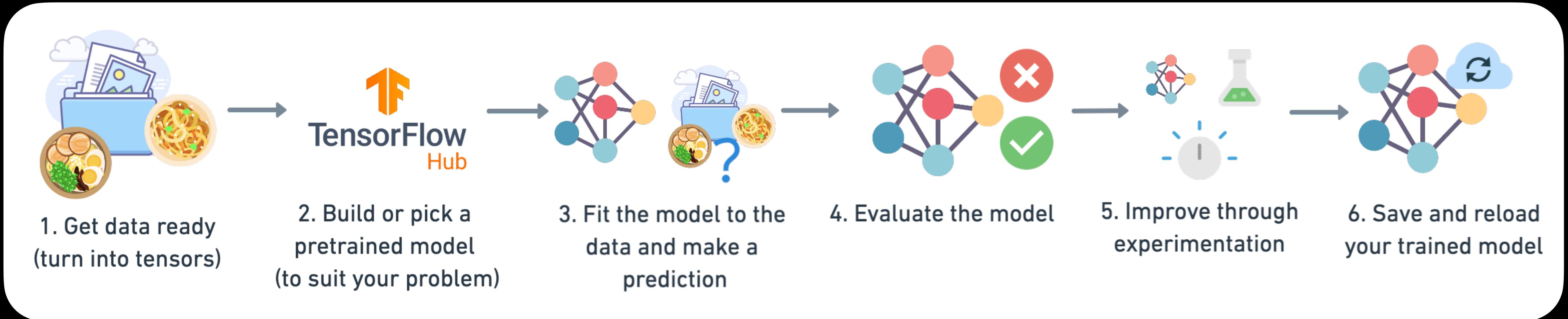
Shape = [None, 512]  
or  
Shape = [32, 512]

(32 is a very common batch size)

These will vary depending on the problem you're working on/what embedding style you use.

Shape = [5]

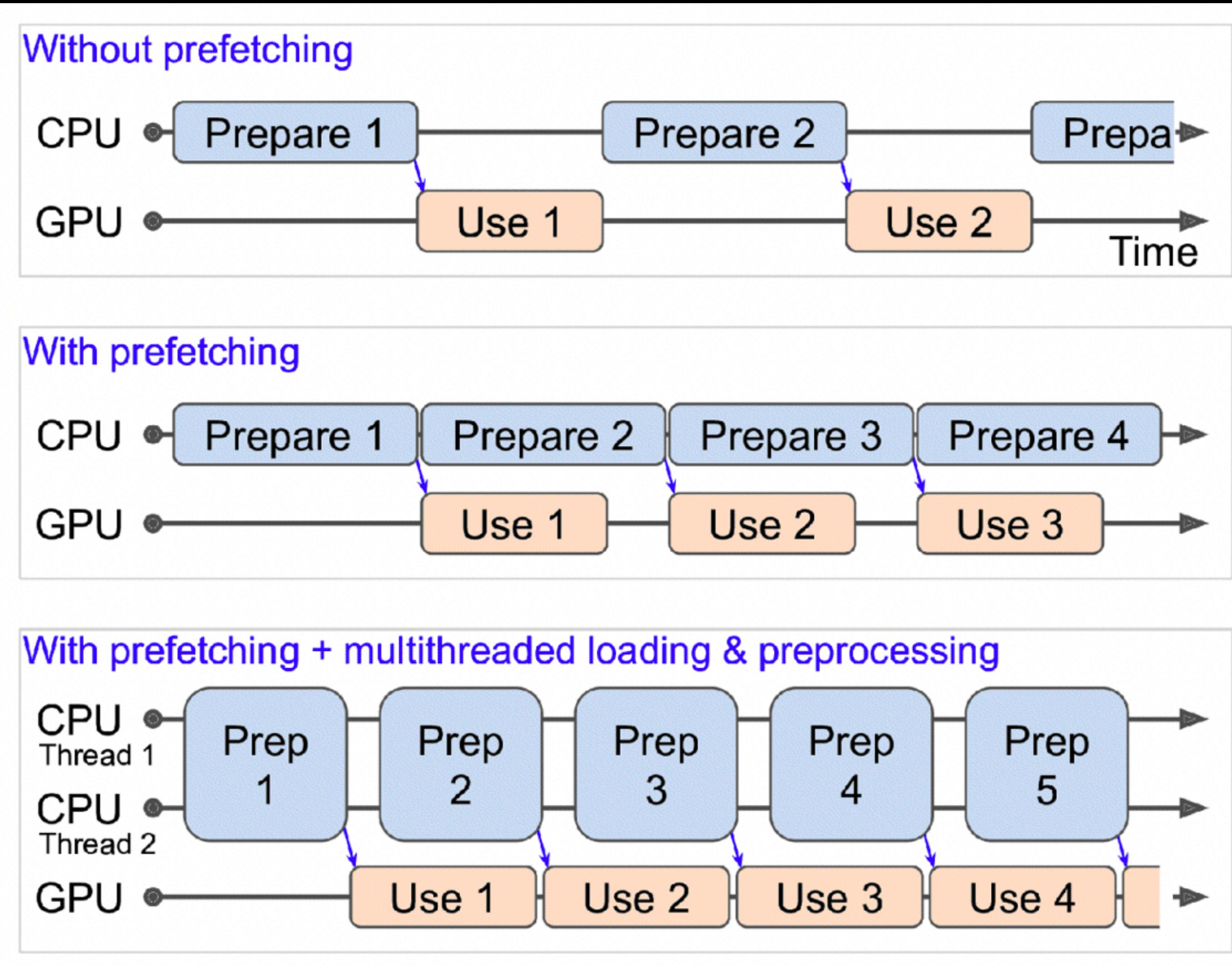
# Steps in modelling with TensorFlow



1. Turn all data into numbers (neural networks can't handle text/natural language)
2. Make sure all of your tensors are the right shape (pad sequences which don't fit)

Let's code!

# Prefetching



Source: Page 422 of [Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow Book](#) by Aurélien Géron

# Tokenization vs Embedding

I love TensorFlow

**Tokenization** — straight mapping from token to number (can be modelled but quickly gets too big)

**Embedding** — richer representation of relationships between tokens (can limit size + can be learned)



$l = 0$   
love = 1  
TensorFlow = 2

[ [1, 0, 0],  
[0, 1, 0],    One-hot  
[0, 0, 1],    Encoding  
...,

[ [0.492, 0.005, 0.019],  
[0.060, 0.233, 0.899],    Embedding  
[0.741, 0.983, 0.567],  
...,

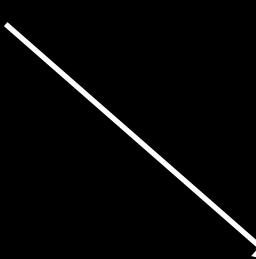
# Token vs Character-level Embedding

[I, love, TensorFlow]



[[0.492, 0.005, 0.019], Each token  
[0.060, 0.233, 0.899], gets turned  
[0.741, 0.983, 0.567], into a feature  
..., vector

[I, , I, O, V, E, , T, E, N, S, O, R, F, I, O, W]

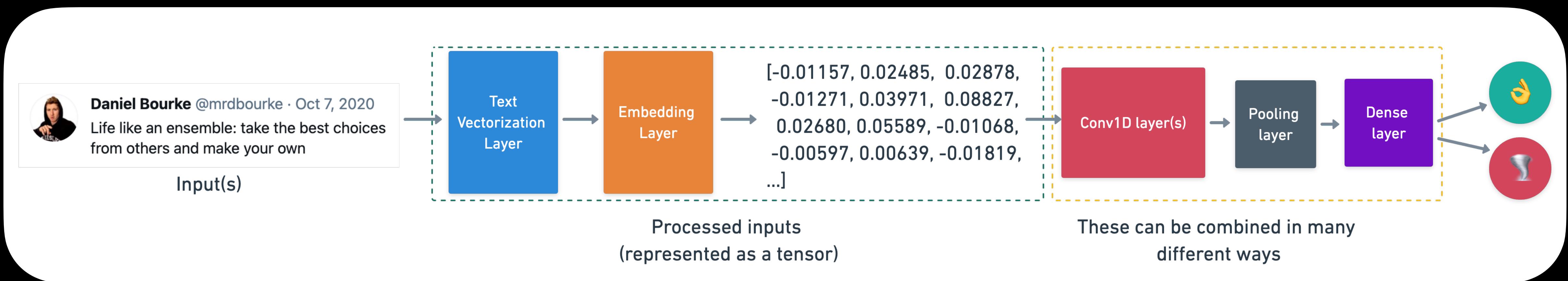


[[0.692, 0.235, 0.088], Each character  
[0.009, 0.956, 0.343], gets turned  
[0.122, 0.454, 0.596], into a feature  
..., vector

# Architecture of a Sequence Conv1D Model

(coloured block edition)

## Conv1D Sequence model



(Model 1)

# Model we're building (USE\* feature extractor)

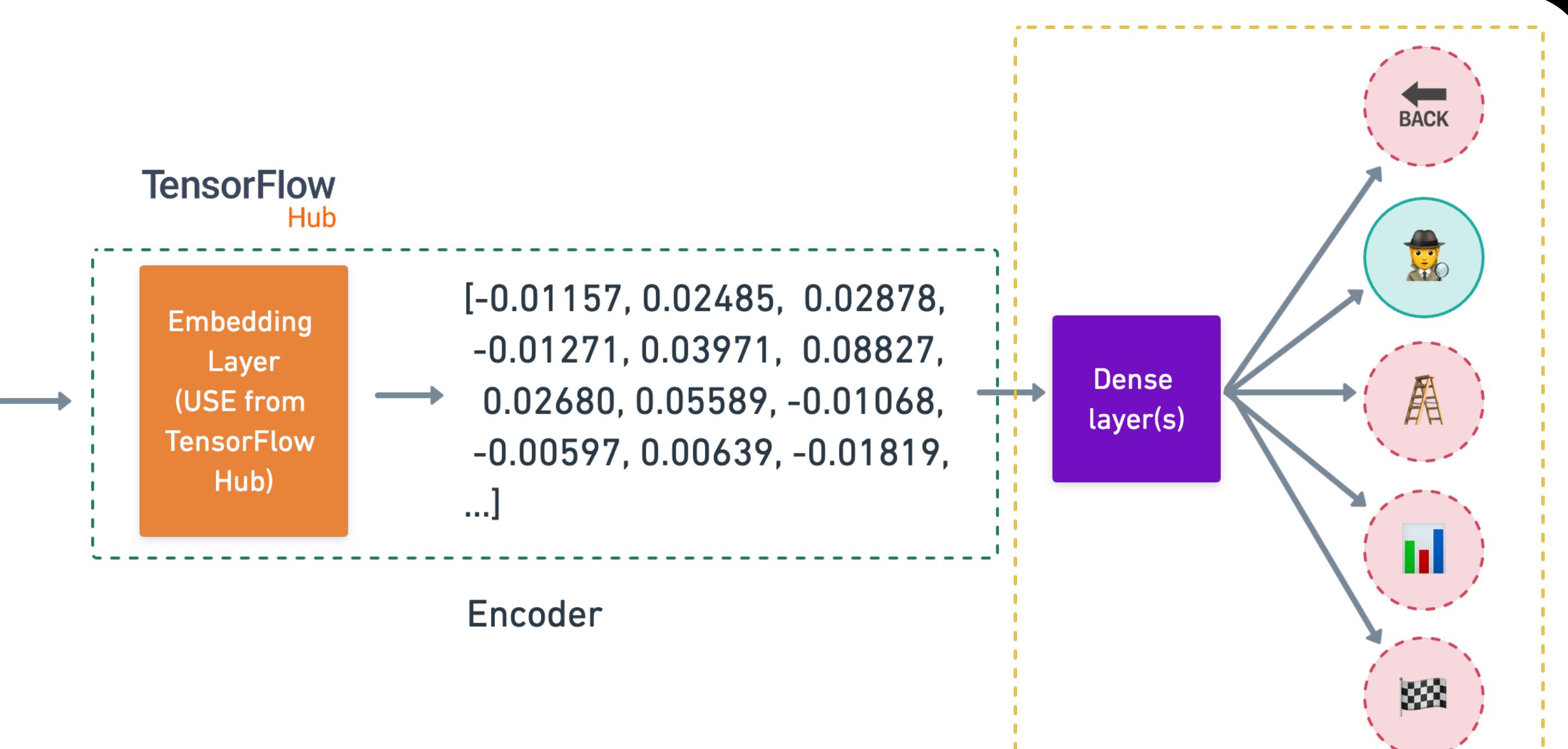
(Model 2)

\*USE = Universal Sentence Encoder

Source: <https://tfhub.dev/google/universal-sentence-encoder/4>

```
['###24293578\n',
'OBJECTIVE\tTo investigate the efficacy of @ weeks of daily low-dose oral prednisolone in improving pain , mobility ,
and systemic low-grade inflammation in the short term and whether the effect would be sustained at @ weeks in
older adults with moderate to severe knee osteoarthritis ( OA ) .\n',
'METHODS\tA total of @ patients with primary knee OA were randomized @@ ; @ received @ mg/day of prednisolone and @ received
placebo for @ weeks .\n',
'METHODS\tOutcome measures included pain reduction and improvement in function scores and systemic inflammation markers .\n',
'METHODS\tPain was assessed using the visual analog pain scale ( @-@ mm ) .\n',
'METHODS\tSecondary outcome measures included the Western Ontario and McMaster Universities Osteoarthritis Index scores ,
patient global assessment ( PGA ) of the severity of knee OA , and @-min walk distance ( @MWD ) .\n',
'METHODS\tSerum levels of interleukin @ ( IL-@ ) , IL-@ , tumor necrosis factor ( TNF ) - , and high-sensitivity C-reactive
protein ( hsCRP ) were measured .\n',
'RESULTS\tThere was a clinically relevant reduction in the intervention group compared to the placebo group for knee pain ,
physical function , PGA , and @MWD at @ weeks .\n',
'RESULTS\tThe mean difference between treatment arms ( @ % CI ) was @ ( @-@ @ ) , p < @ ; @ ( @-@ @ ) , p < @ ; @ ( @-@ @ ) ,
p < @ ; and @ ( @-@ @ ) , p < @ , respectively .\n',
'RESULTS\tFurther , there was a clinically relevant reduction in the serum levels of IL-@ , IL-@ , TNF - , and hsCRP at @
weeks in the intervention group when compared to the placebo group .\n',
'RESULTS\tThese differences remained significant at @ weeks .\n',
'RESULTS\tThe Outcome Measures in Rheumatology Clinical Trials-Osteoarthritis Research Society International responder
rate was @ % in the intervention group and @ % in the placebo group ( p < @ ) .\n',
'CONCLUSIONS\tLow-dose oral prednisolone had both a short-term and a longer sustained effect resulting in less knee pain ,
better physical function , and attenuation of systemic inflammation in older patients with knee OA
( ClinicalTrials.gov identifier NCT@ ) .\n']
```

Input(s)



(Encodes sequences into  
numerical representation)

(Decodes sequences into  
desired output)

# Experiments we're running

Experiment	Model
0	Naive Bayes with TF-IDF encoder (baseline)
1	Conv1D with token embeddings
2	TensorFlow Hub Pretrained Feature Extractor
3	Conv1D with character embeddings
4	Pretrained token embeddings (same as 2) + character embeddings (same as 3)
5	Pretrained token embeddings + character embeddings + positional embeddings

# Feature Engineering

- Taking **non-obvious features** from the data and encoding them numerically to help our model learn
- How can we add extra sources of data to our model?

Line numbers 1  
2  
:  
:  
Total lines 12

```
[ '###24293578\n',
  'OBJECTIVE\tTo investigate the efficacy of @ weeks of daily low-dose oral prednisolone in improving pain , mobility ,
  and systemic low-grade inflammation in the short term and whether the effect would be sustained at @ weeks in
  older adults with moderate to severe knee osteoarthritis ( OA ) .\n',
  'METHODS\tA total of @ patients with primary knee OA were randomized @:@ ; @ received @ mg/day of prednisolone and @ received
  placebo for @ weeks .\n',
  'METHODS\tOutcome measures included pain reduction and improvement in function scores and systemic inflammation markers .\n',
  'METHODS\tPain was assessed using the visual analog pain scale ( @-@ mm ) .\n',
  'METHODS\tSecondary outcome measures included the Western Ontario and McMaster Universities Osteoarthritis Index scores ,
  patient global assessment ( PGA ) of the severity of knee OA , and @-min walk distance ( @MWD ) .\n',
  'METHODS\tSerum levels of interleukin @ ( IL-@ ) , IL-@ , tumor necrosis factor ( TNF ) - , and high-sensitivity C-reactive
  protein ( hsCRP ) were measured .\n',
  'RESULTS\tThere was a clinically relevant reduction in the intervention group compared to the placebo group for knee pain ,
  physical function , PGA , and @MWD at @ weeks .\n',
  'RESULTS\tThe mean difference between treatment arms ( @ % CI ) was @ ( @-@ @ ) , p < @ ; @ ( @-@ @ ) , p < @ ; @ ( @-@ @ ) ,
  p < @ ; and @ ( @-@ @ ) , p < @ , respectively .\n',
  'RESULTS\tFurther , there was a clinically relevant reduction in the serum levels of IL-@ , IL-@ , TNF - , and hsCRP at @
  weeks in the intervention group when compared to the placebo group .\n',
  'RESULTS\tThese differences remained significant at @ weeks .\n',
  'RESULTS\tThe Outcome Measures in Rheumatology Clinical Trials-Osteoarthritis Research Society International responder
  rate was @ % in the intervention group and @ % in the placebo group ( p < @ ) .\n',
  'CONCLUSIONS\tLow-dose oral prednisolone had both a short-term and a longer sustained effect resulting in less knee pain ,
  better physical function , and attenuation of systemic inflammation in older patients with knee OA
  ( ClinicalTrials.gov identifier NCT@ ) .\n']
```

# Data augmentation is a form of feature engineering

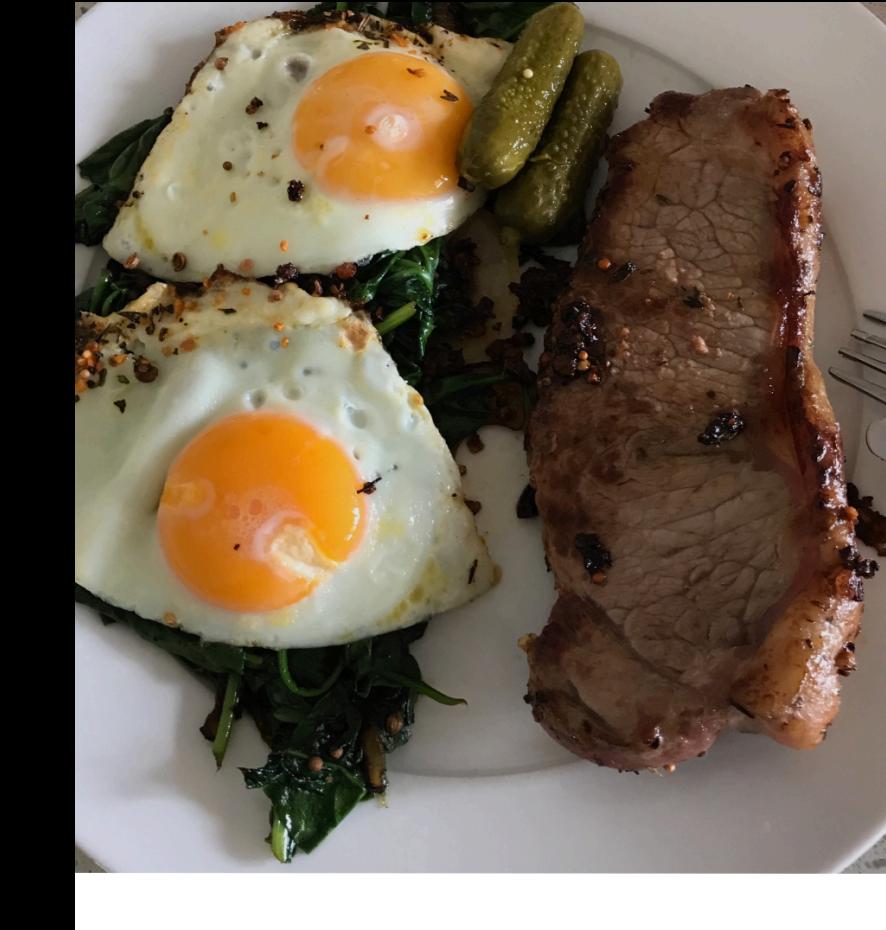
Looking at the same image but from different perspective(s)\*.



Original



Rotate



Shift



Zoom

\*Note: There are many more different kinds of data augmentation such as, cropping, replacing, shearing. This slide only demonstrates a few.

# Engineered features need to be available at test time

## Nutritional psychiatry: the present state of the evidence

Wolfgang Marx <sup>1</sup>, Genevieve Moseley <sup>2</sup>, Michael Berk <sup>2</sup>, Felice Jacka <sup>2</sup>

Affiliations + expand

PMID: 28942748 DOI: [10.1017/S0029665117002026](https://doi.org/10.1017/S0029665117002026)

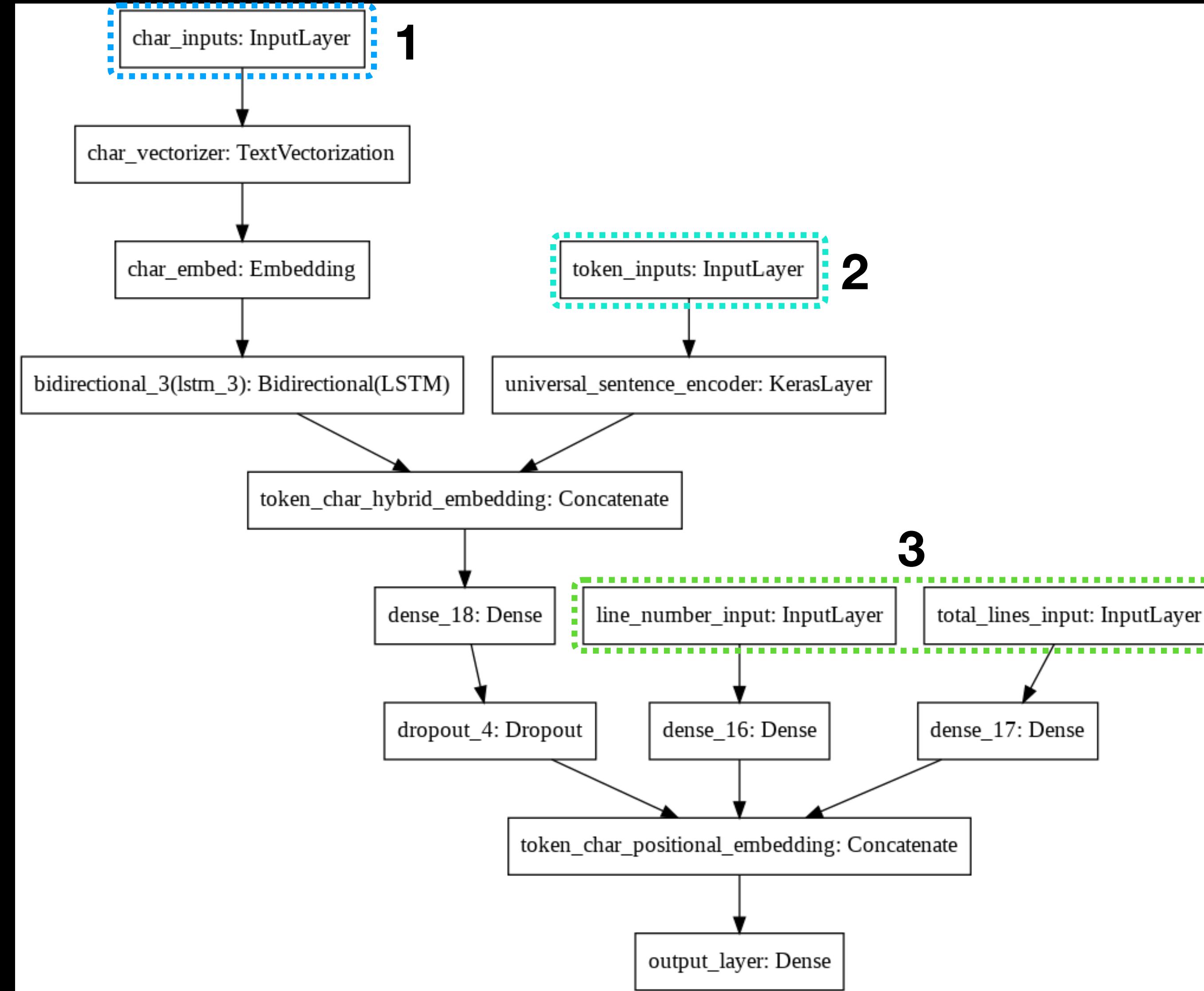
### Abstract

**1**Mental illness, including depression, anxiety and bipolar disorder, accounts for a significant proportion of global disability and poses a substantial social, economic and health burden. **2**Treatment is presently dominated by pharmacotherapy, such as antidepressants, and psychotherapy, such as cognitive behavioural therapy; however, such treatments avert less than half of the disease burden, suggesting that additional strategies are needed to prevent and treat mental disorders.**3**There are now consistent mechanistic, observational and interventional data to suggest diet quality may be a modifiable risk factor for mental illness.**4**This review provides an overview of the nutritional psychiatry field. It includes a discussion of the neurobiological **5** mechanisms likely modulated by diet, the use of dietary and nutraceutical interventions in mental disorders, and recommendations for further research. Potential biological pathways related to **6** mental disorders include inflammation, oxidative stress, the gut microbiome, epigenetic modifications and neuroplasticity. Consistent epidemiological evidence, particularly for depression, suggests an association between measures of diet quality and mental health, across multiple populations and age groups; these do not appear to be explained by other demographic, lifestyle factors or reverse causality. Our recently published intervention trial provides preliminary clinical **8** evidence that dietary interventions in clinically diagnosed populations are feasible and can provide significant clinical benefit. Furthermore, nutraceuticals including n-3 fatty acids, folate, S- **9** adenosylmethionine, N-acetyl cysteine and probiotics, among others, are promising avenues for future research. Continued research is now required to investigate the efficacy of intervention **10** studies in large cohorts and within clinically relevant populations, particularly in patients with schizophrenia, bipolar and anxiety disorders.

- Are the features we've engineered available at test time?
  - Line numbers
  - Total lines

# Model we're building (tribrid embeddings)

## 1. Character Embeddings



## 2. Token Embeddings

## 3. Feature Embeddings

(some common)

# Classification evaluation methods

Key: **tp** = True Positive, **tn** = True Negative, **fp** = False Positive, **fn** = False Negative

Metric Name	Metric Forumla	Code	When to use
Accuracy	<b>Accuracy</b> = $\frac{tp + tn}{tp + tn + fp + fn}$	<code>tf.keras.metrics.Accuracy()</code> or <code>sklearn.metrics.accuracy_score()</code>	Default metric for classification problems. Not the best for imbalanced classes.
Precision	<b>Precision</b> = $\frac{tp}{tp + fp}$	<code>tf.keras.metrics.Precision()</code> or <code>sklearn.metrics.precision_score()</code>	Higher precision leads to less false positives.
Recall	<b>Recall</b> = $\frac{tp}{tp + fn}$	<code>tf.keras.metrics.Recall()</code> or <code>sklearn.metrics.recall_score()</code>	Higher recall leads to less false negatives.
F1-score	<b>F1-score</b> = $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$	<code>sklearn.metrics.f1_score()</code>	Combination of precision and recall, usually a good overall metric for a classification model.
Confusion matrix	NA	Custom function or <code>sklearn.metrics.confusion_matrix()</code>	When comparing predictions to truth labels to see where model gets confused. Can be hard to use with large numbers of classes.

# Improving a model

(from a model's perspective)

```
# 1. Create the model (specified to your problem)
model = tf.keras.Sequential([
    tf.keras.layers.Flatten(input_shape=(28, 28)),
    tf.keras.layers.Dense(4, activation="relu"),
    tf.keras.layers.Dense(10, activation="softmax")
])

# 2. Compile the model
model.compile(loss=tf.keras.losses.BinaryCrossentropy(),
              optimizer=tf.keras.optimizers.Adam(lr=0.001),
              metrics=[ "accuracy" ])

# 3. Fit the model
model.fit(x_train_subset, y_train_subset, epochs=5)
```

Smaller model

```
# 1. Create the model (specified to your problem)
model = tf.keras.Sequential([
    tf.keras.layers.Flatten(input_shape=(28, 28)),
    tf.keras.layers.Dense(100, activation="relu"),
    tf.keras.layers.Dense(100, activation="relu"),
    tf.keras.layers.Dense(100, activation="relu"),
    tf.keras.layers.Dense(10, activation="softmax")
])

# 2. Compile the model
model.compile(loss=tf.keras.losses.BinaryCrossentropy(),
              optimizer=tf.keras.optimizers.Adam(lr=0.0001),
              metrics=[ "accuracy" ])

# 3. Fit the model
model.fit(x_train_full, y_train_full, epochs=100)
```

Larger model

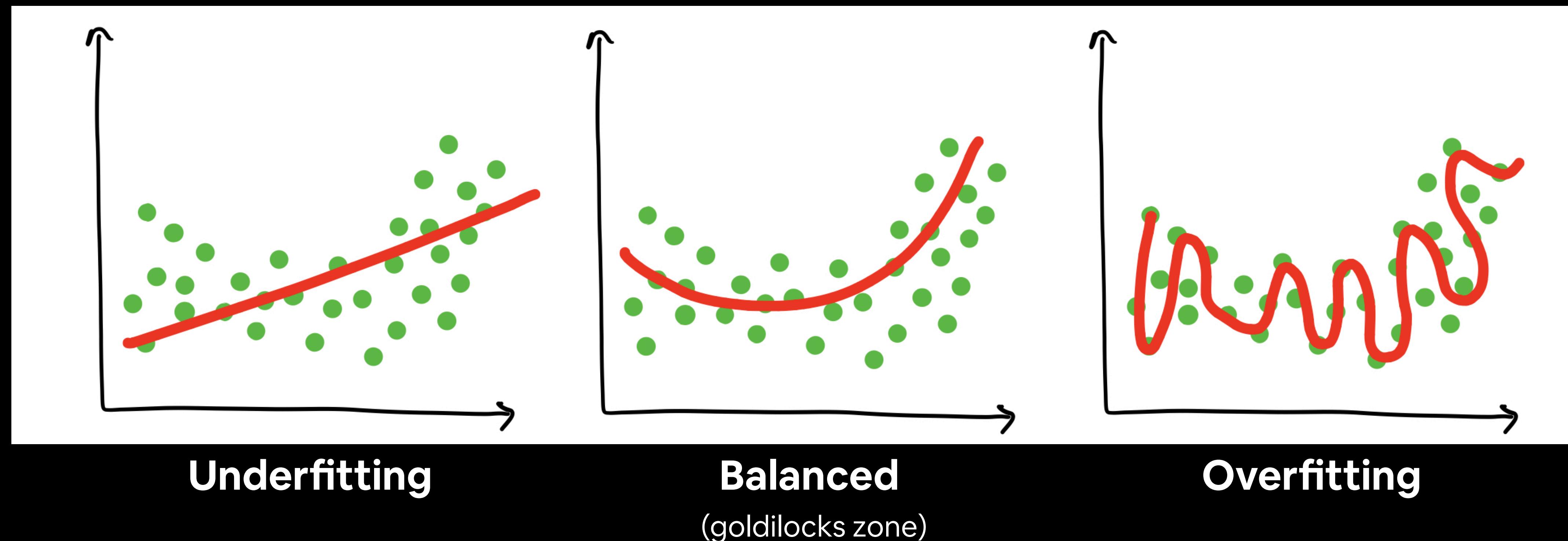
## Common ways to improve a deep model:

- Adding layers
- Increase the number of hidden units
- Change the activation functions
- Change the optimization function
- Change the learning rate      (because you can alter each of these, they're hyperparameters)
- Fitting on more data
- Fitting for longer

# What is overfitting?

**Overfitting** — when a model over learns patterns in a particular dataset and isn't able to generalise to unseen data.

For example, a student who studies the course materials too hard and then isn't able to perform well on the final exam. Or tries to put their knowledge into practice at the workplace and finds what they learned has nothing to do with the real world.



# Improving a model

(from a data perspective)

## Method to improve a model (reduce overfitting)

### What does it do?

More data

Gives a model more of a chance to learn patterns between samples (e.g. if a model is performing poorly on images of pizza, show it more images of pizza).

Data augmentation (usually for images)

Increase the diversity of your training dataset without collecting more data (e.g. take your photos of pizza and randomly rotate them 30°). Increased diversity forces a model to learn more generalisation patterns.

Better data (**feature engineering**)

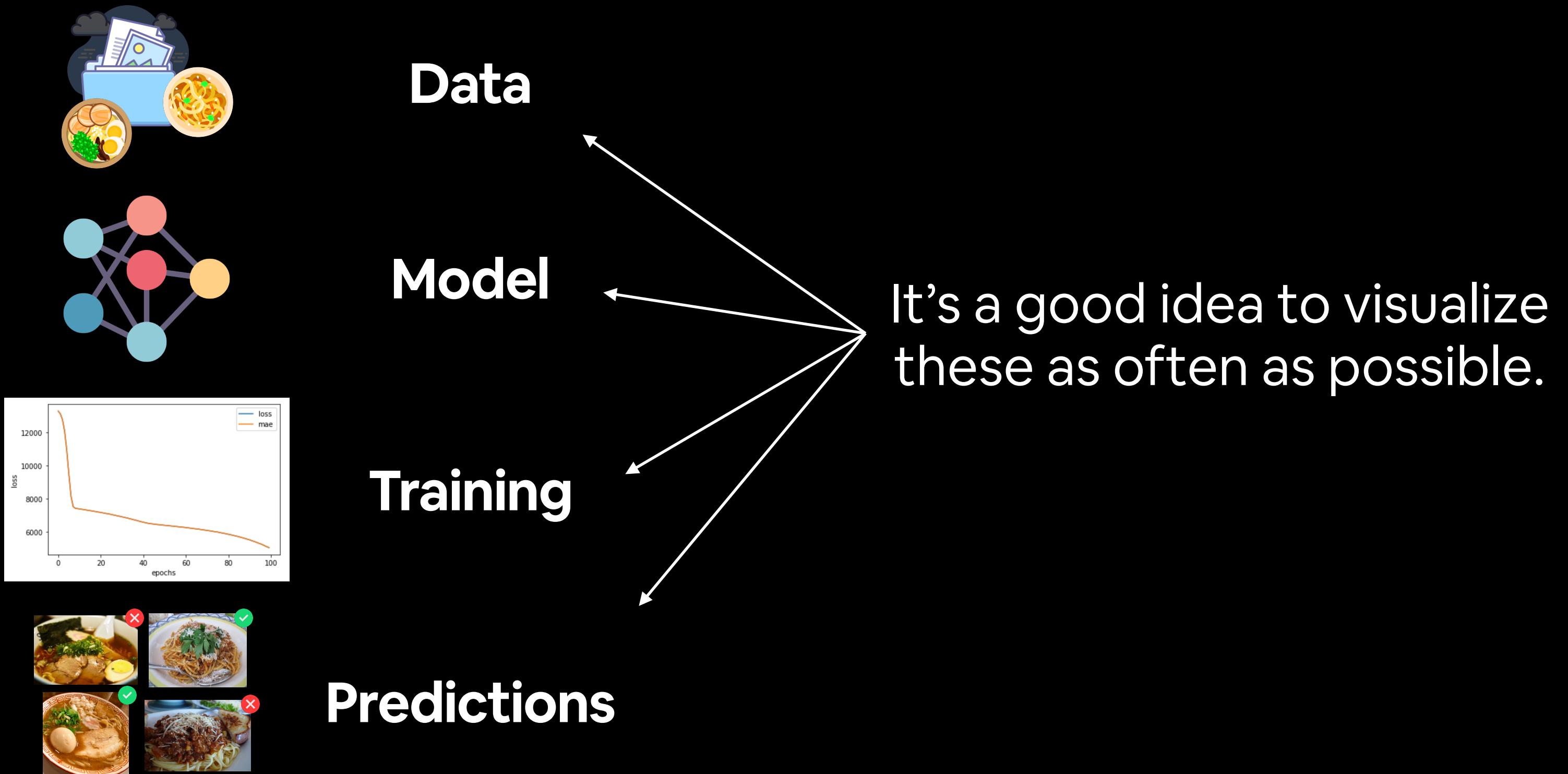
Not all data samples are created equally. Removing poor samples from or adding better samples/**engineering new features** in your dataset can improve your model's performance.

Use transfer learning

Take a model's pre-learned patterns from one problem and tweak them to suit your own problem. For example, take a model trained on pictures of cars to recognise pictures of trucks.

# The machine learning explorer's motto

“Visualize, visualize, visualize”



# The machine learning practitioner's motto

“Experiment, experiment, experiment”



*(try lots of things and see what  
tastes good)*