Data Science:
Principles and Practice

Lecture 8

Guy Emerson

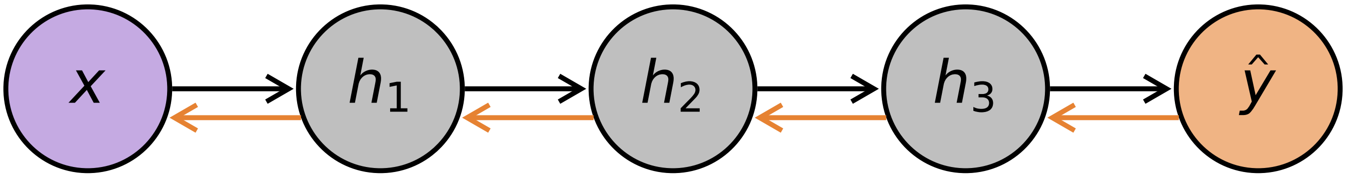
Today's Lecture

- Writing code
- Significance testing
- Ethics

Last lecture was more about the principles; this lecture is more about the practice.

Backpropagation

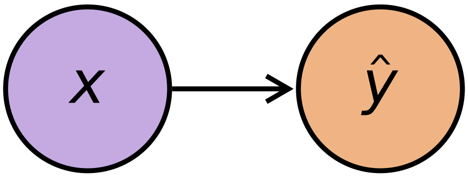
Forward pass



Backward pass
(calculate gradients with chain rule)

We saw this last lecture. We'll now look at this in more detail.

Backpropagation



$$\hat{y} = \sigma(w.x)$$
$$\mathcal{L} = (\hat{y} - y)^2$$

Here is a very simple neural model (only one layer), with a simple loss function (square error).

Backpropagation

$$\begin{tikzcd}
 & w \arrow[d] & & y \arrow[d] & \\
x \arrow[r] & h_1 \arrow[r] & h_2 \arrow[r] & h_3 \arrow[r] & \mathcal{L} \\
 & =w.x & =\sigma(h_1) & =h_2-y & =h_3^2
\end{tikzcd}$$

$$\frac{d\mathcal{L}}{dw_i} = \frac{d\mathcal{L}}{dh_3}\frac{dh_3}{dh_2}\frac{dh_2}{dh_1}\frac{dh_1}{dw_i}$$

$$= 2h_3\,1\,\sigma(h_1)(1-\sigma(h_1))\,x_i$$

In this graph, the computation is broken down into smaller steps. We're calculating the loss L based on the input x, network parameters (weights) w, and desired output y.

We would like to optimise the loss with respect to each network parameter. The gradient can be calculated using the chain rule. As long as each operation in the computation graph has a known derivative, we can calculate each term in the chain rule. Notice how we need to use the intermediate values (h_1 and h_3).

(For $\sigma(x) = (1+e^{-x})^{-1}$, it is a simple exercise to find the derivative. It can be written in the above form for convenience.)

Backpropagation

- Need to store computation graph
- Need to store intermediate values

The previous slide illustrates why these are necessary.

Autograd

- NumPy with automatic differentiation

We have already seen NumPy earlier in this course, so Autograd is useful to learn about automatic differentiation without having to also learn a new API.

Autograd

```
from autograd import numpy, scipy, grad


def forward(x, w):
    return scipy.special.expit(numpy.dot(x, w))

def loss_fn(x, y, w):
    return (forward(x, w) - y)**2


calculate_w_grad = grad(loss_fn, 2)


w = numpy.random.standard_normal(size=3)
x = numpy.array([.1,.3,.7])
y = 0
w_grad = calculate_w_grad(x, y, w)
```

numpy is imported from autograd. This gives us the usual numpy objects, but with extra book-keeping (remember we need the computation graph) so that we can automatically calculate derivatives.

To implement our previous example, we can write the functions forward and loss_fn in a normal NumPy/SciPy way. Note that expit is another name for the sigmoid/logistic function.

Autograd also gives us the function grad, which takes a function as input (the function to be differentiated) and returns a function as output (the derivative). The second argument specifies what we are differentiating with respect to – in this case, w is argument number 2 of loss_fn.

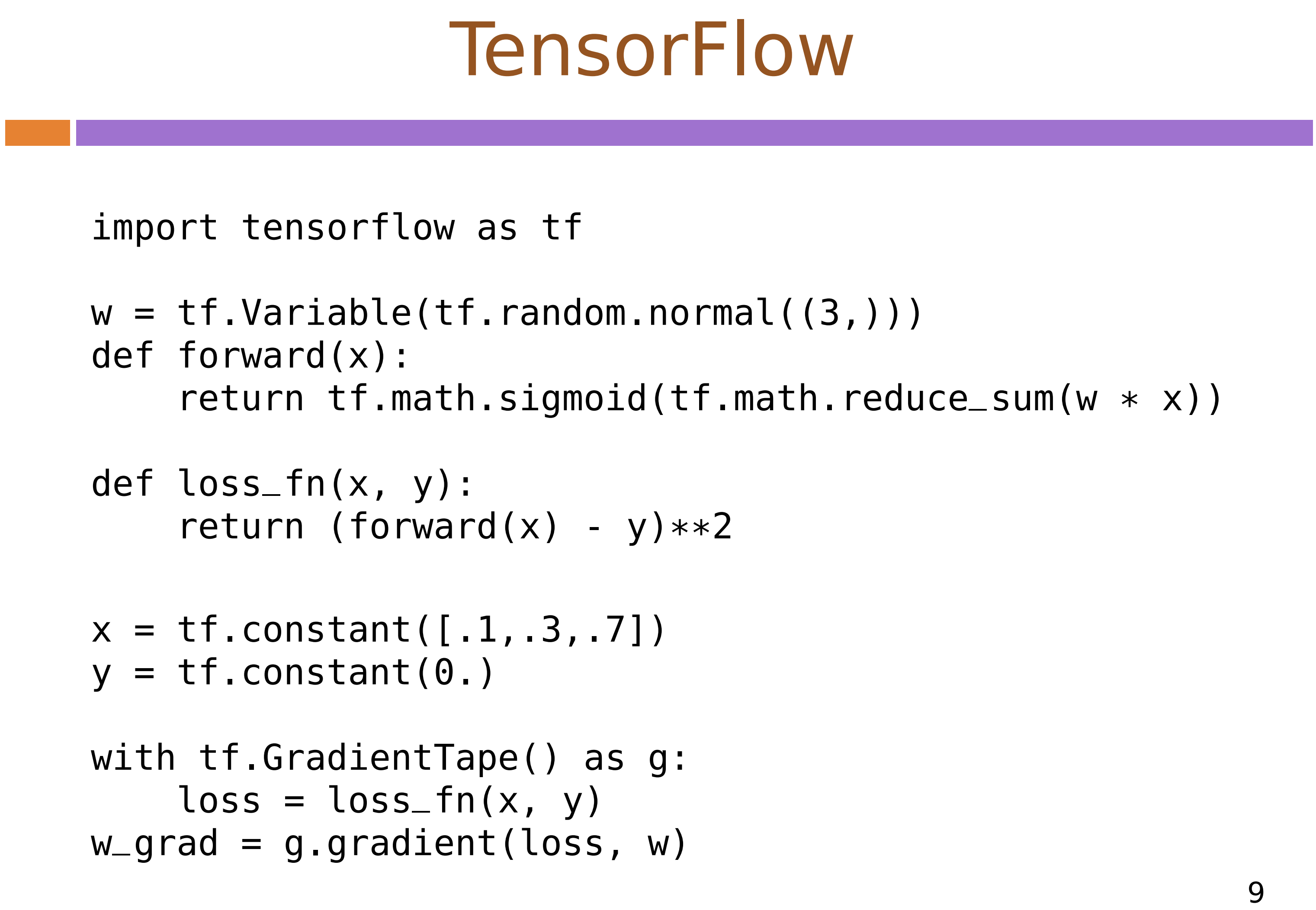
We can now calculate the gradient, given values for w, x and y.

TensorFlow

- Automatic differentiation
- Compilation for speed
- Range of architectures
- Range of training algorithms

In practice, there are other libraries which can do more than just calculate derivatives.

Apart from TensorFlow, another popular library is PyTorch.



TensorFlow

```
import tensorflow as tf

w = tf.Variable(tf.random.normal((3,)))
def forward(x):
    return tf.math.sigmoid(tf.math.reduce_sum(w * x))

def loss_fn(x, y):
    return (forward(x) - y)**2


x = tf.constant([.1,.3,.7])
y = tf.constant(0.)

with tf.GradientTape() as g:
    loss = loss_fn(x, y)
w_grad = g.gradient(loss, w)
```

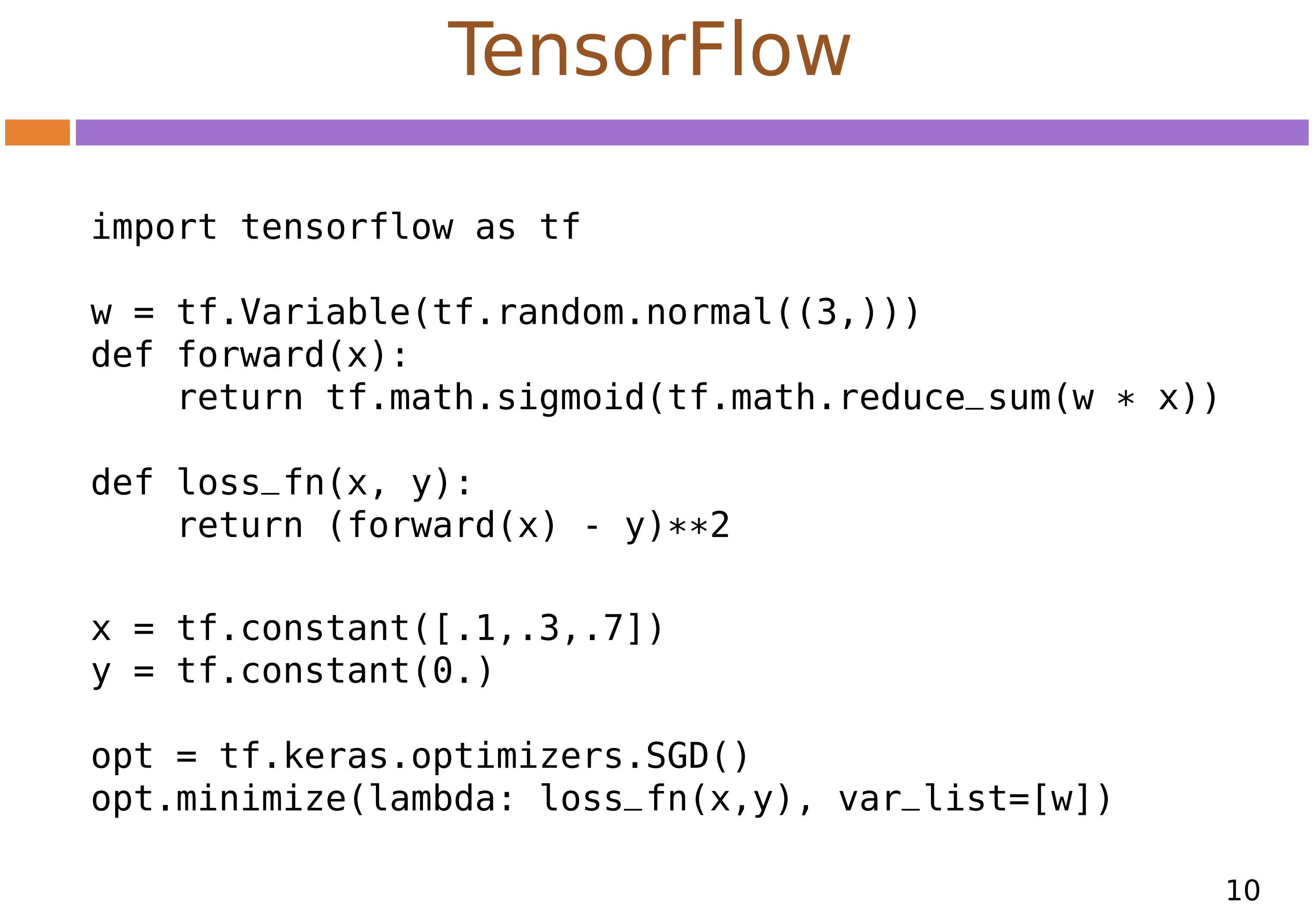
This implements our previous example, in TensorFlow instead of Autograd.

We specify that `w` is a variable, i.e. something that we would like to optimise. It doesn't need to be passed as an argument, but can be used implicitly, as shown above.

`*` does an elementwise product (mapping two vectors to a new vector), and then `reduce_sum` sums the values to give a scalar.

We specify `x` and `y` as constants, since they are data, which we can't change.

To set up the book-keeping to do automatic differentiation, we can use a `GradientTape` object.



TensorFlow

```
import tensorflow as tf

w = tf.Variable(tf.random.normal((3,)))
def forward(x):
    return tf.math.sigmoid(tf.math.reduce_sum(w * x))

def loss_fn(x, y):
    return (forward(x) - y)**2


x = tf.constant([.1,.3,.7])
y = tf.constant(0.)

opt = tf.keras.optimizers.SGD()
opt.minimize(lambda: loss_fn(x,y), var_list=[w])
```

In practice, we don’t need to explicitly write code to calculate gradients. We want the gradients in order to optimise the model parameters, and TensorFlow makes this easy for us.

We can instead use an optimizer object (here, using stochastic gradient descent), simply telling the optimizer what the loss is (this must be a function taking no arguments), and the parameters are (a list of variables). Running the `minimize` method automatically does a forward pass and a backward pass, and makes an update to the parameters.

Summary

- Backpropagation:
  - Store computation graph
  - Store intermediate values
- Software packages:
  - Automatic backpropagation
  - Automatic compilation
  - Pre-defined architectures
  - Pre-defined training algorithms

Neural Network Research

- Emphasis on empirical performance
- Large number of architectures,
  Large number of hyperparameters
- Datasets re-used many times

→ Easy to get inflated results

The large number of architectures and hyperparameters means that testing significance is important (or should be!) – if we test a large number of models, some are bound to perform better than others, just by random variation.

The re-use of datasets means that the field as a whole may be overfitting, even if each individual researcher is not.

Significance Testing

Dror et al. (2018) survey of NLP papers:

	ACL 2017		TACL 2017	
Total papers	196		37	
Experimental papers	180		33	
– reporting significance	63	(35%)	18	(55%)
– correctly	36	(20%)	15	(45%)

Dror et al. (2018) survey ACL and TACL papers from 2017, and give recommendations for significance testing. `http://aclweb.org/anthology/P18-1128`

Of the papers that report significance incorrectly, some use an inappropriate test (6 ACL papers), and some do not state what test they used (21 ACL papers and 3 TACL papers).

The vast majority of papers are experimental, but significance testing is not the norm!

p-Values

- Probability the result would be at least this extreme, under the null hypothesis

NOT:

- Probability the null hypothesis is true

Most data scientists have heard of p-values, but they are often misunderstood!

(In 2018, there was a mistake in the practical notes for the Part IA course MLRD!)

There are more details on significance testing in the notes for the Part IB course Foundations of Data Science. The following slides give a summary and some practical suggestions.

Statistical Significance Testing

- Decide on a **null hypothesis**
- Decide on a **test statistic**
- Decide on a **threshold**
- **Significance level**: probability of incorrectly rejecting null hypothesis (assuming null hypothesis)
- **Power**: probability of correctly rejecting null hypothesis (assuming alternative hypothesis)

The null hypothesis formalises the idea that the method doesn’t work (e.g. it’s no better than the baseline). The test statistic summarises the results in single number. (How the null hypothesis and test statistic are defined will depend on the task.) If the observed test statistic is too extreme (beyond some threshold), we reject the null hypothesis.

A p-value is a way to re-express the test statistic in terms of a probability. Rather than using the observed test statistic itself, we can calculate the probability that the statistic would be at least as extreme as observed.

In data science, the term “significant” should be reserved for statistical significance. Using the term loosely is bad practice.

Parametric Tests

- Test statistic follows known distribution (with known parameters)
- Paired Student's t-test:
  - Paired samples (test datapoints)
  - Scores normally distributed
  - Null hypothesis: same mean
  - Test statistic: $t = \frac{\sqrt{n}}{s_D}\bar{x}_D$
  - "Student's t-distribution with $n-1$ degrees of freedom"

The paired Student's t-test is a parametric test, because it assumes that the scores are normally distributed. It is useful when comparing the results of two systems on the same data.

$\bar{x}_D$ is the average difference between the scores of the two systems.
s_D is the standard deviation of the differences between scores.
n is the number of datapoints.

We have to divide by the observed standard deviation, because we don't know what the standard deviation should be. The resulting distribution is called Student's t-distribution, and it looks a bit like the normal distribution. The details aren't important here – this is a standard test, available in any reasonable statistics package.

Nonparametric Tests

- No assumptions about distribution
- Sign test:
  - Paired samples (test datapoints)
  - System A better or system B better
  - Null hypothesis: equal chance
  - Test statistic: n
  - Binomial distribution

The sign test is an example of a nonparametric test. It is useful when comparing two systems, when we don't know the distribution of scores – here, we simply look at which system is better.

n is the number of times system A is better than system B.

(In the case of ties, we can evenly split the ties between the two systems, or we can discard them. Discarding them gives a more powerful test – see *power* on slide 15. An alternative is the trinomial test, which includes the ties as a third outcome.)

Compared to a parametric test, a nonparametric test is more general (it doesn't make assumptions), but less powerful.

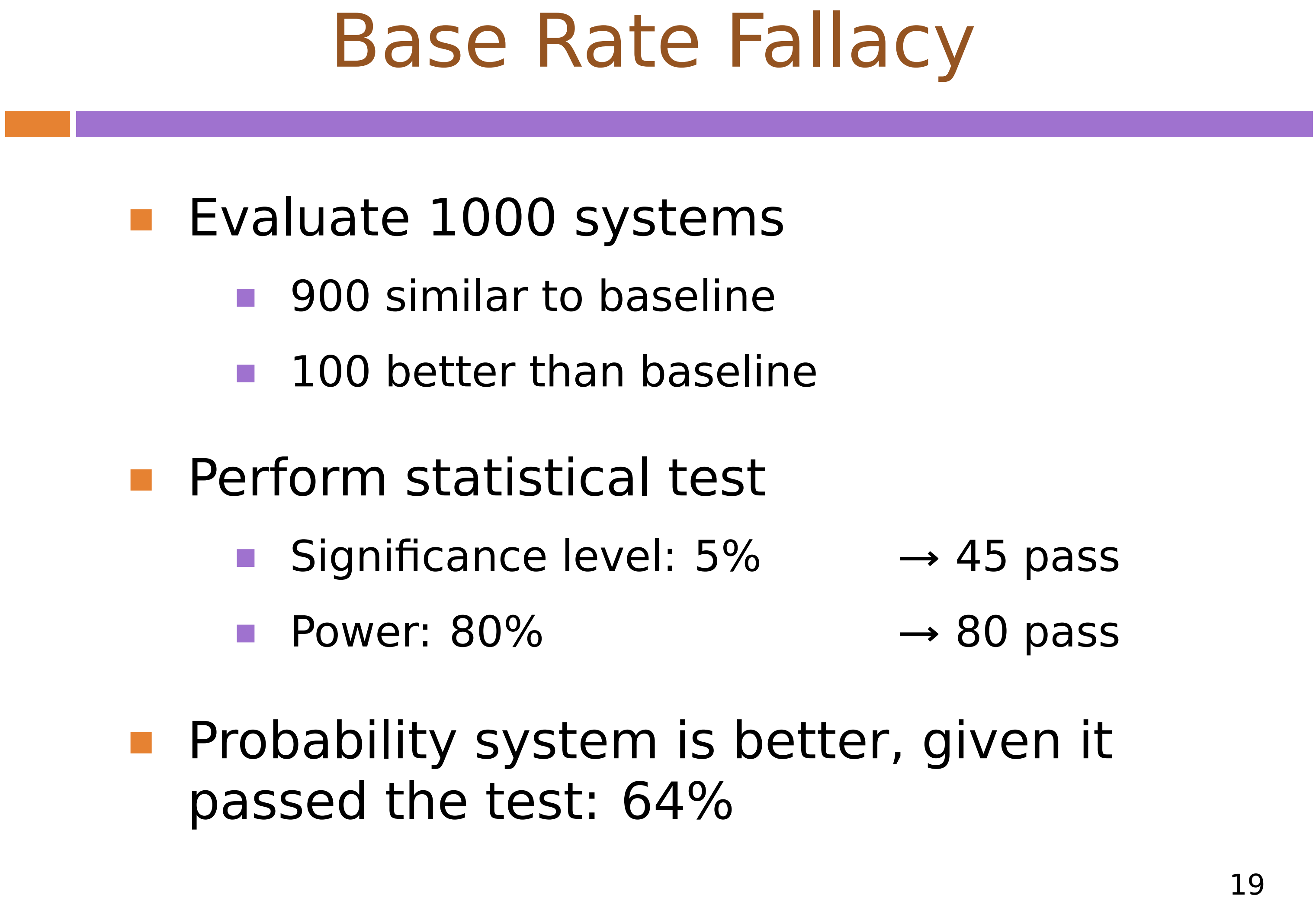
Multiple Tests

- If we test many systems, we expect some will pass
- Bonferroni correction:
  - Replace nominal significance level
  - $\alpha \mapsto \frac{\alpha}{m}$

α is the desired significance level, for all tests combined.
m is the number of systems being tested.
$\frac{\alpha}{m}$ is the significance level that should be used for each individual test.

Further reading:

`https://xkcd.com/882/`



Base Rate Fallacy

- Evaluate 1000 systems
  - 900 similar to baseline
  - 100 better than baseline
- Perform statistical test
  - Significance level: 5% → 45 pass
  - Power: 80% → 80 pass
- Probability system is better, given it passed the test: 64%

19

The base rate fallacy shows why the misunderstanding about p-values is so dangerous.

Here, the probability that the system is better than the baseline, given that it passed the test, is only 64%. This is much lower than 95%!

The reason for this is the *base rate*, the proportion of tested systems that are actually better.

Base Rate Fallacy

- Evaluate 1000 systems
  - 960 similar to baseline
  - 40 better than baseline
- Perform statistical test
  - Significance level: 5% → 48 pass
  - Power: 80% → 32 pass
- Probability system is better, given it passed the test: 40%

If we reduce the base rate to 4%, the the probability that the system is better than the baseline, given that it passed the test, drops to only 40%.

Base Rate Fallacy

- Evaluate 1000 systems
  - 1000 similar to baseline
  - 0 better than baseline
- Perform statistical test
  - Significance level: 5% → 50 pass
  - Power: 80% → 0 pass
- Probability system is better, given it passed the test: 0%

In the extreme case, a base rate of 0 means that all passes are just due to random variation. This is referred to as the “look elsewhere effect” or “look everywhere effect” – if you keep looking for long enough, you will find something that appears significant.

This is not just a toy problem, but a common problem in scientific research. For example, see Ioannidis (2005) “Why Most Published Research Findings are False” `https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124`

It’s a problem in data science, particularly when many variants of a model are tested – unless we have good reason to believe that some variants should perform better, the base rate could be very low.

Effect Size

- A significant difference may not be a large difference
- e.g. a coin toss
  - Coins not perfectly symmetric
  - Probability of heads not exactly 50%
  - Difference so small we don’t care

In many cases, we don’t just care about finding a difference – we want to find a large difference.

For example, if system A performs marginally better than system B, but takes a lot longer to run, or is less interpretable, we might prefer system B anyway.

Publication Bias

- Hard to publish negative results...
- Authors may hide failed experiments

Publication bias means that we get a skewed view of results. Remember that when we make multiple tests, we need to correct for this, e.g. using the Bonferroni correction. However, if negative results are not published, we don't get to see how many experiments were run.

This becomes more serious when publication bias leads to authors changing how they try to present their work.

In your coursework (e.g. Part II project), you don't need to get positive results! Try to run your experiments carefully, and report whatever you find.

Summary of Significance Testing

- Significance testing is important but underused in deep learning
- Choice of test:
  - Parametric (e.g. paired Student’s t-test)
  - Nonparametric (e.g. sign test)
  - Multiple tests (e.g. Bonferroni correction)
- Be careful:
  - Base rate fallacy
  - Effect size
  - Publication bias

Ethics in Data Science

- Task
- Data

What if this goes wrong?

- Model
- Training

Most research

Broadly speaking, we can break up any data science problem into these four parts.

If the task is poorly defined, or there is a problem with the data, any model is going to struggle, no matter how it’s trained.

And if something goes wrong, what happens if we use the trained system in a real-world application?

Caruana et al. (2015)

- Task: Predict death from pneumonia
- Pattern in data: asthma reduces risk
- Real reason: asthma patients sent to Intensive Care Unit, reducing risk
- Shallow models (e.g. logistic regression) → can identify and fix such problems

An example from healthcare, to demonstrate the problem: this example is serious and uncontroversial. Patients with a high risk of death would be treated in the hospital, while patients with a low risk would be treated as outpatients. If a high-risk patient is mistakenly sent home, and then they die, this is a serious mistake.

Caruana et al. (2015) show how a real pattern in the data is that having asthma correlates with lower risk – despite asthma and pneumonia both being lung conditions. It turned out that this was because the asthma patients were given intensive care, and improved as a result of that care. Here, there is a bias in the dataset that we don't want in the trained model. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.704.9327&rep=rep1&type=pdf

Bias

- Bias (statistics):
  expected value differs from true value
- Bias (law):
  unfair or undesirable prejudice

There is nothing ethically wrong with statistical bias – it’s just a technical definition.

However, “bias” in the legal sense (or just the usual day-to-day sense) is a problem.

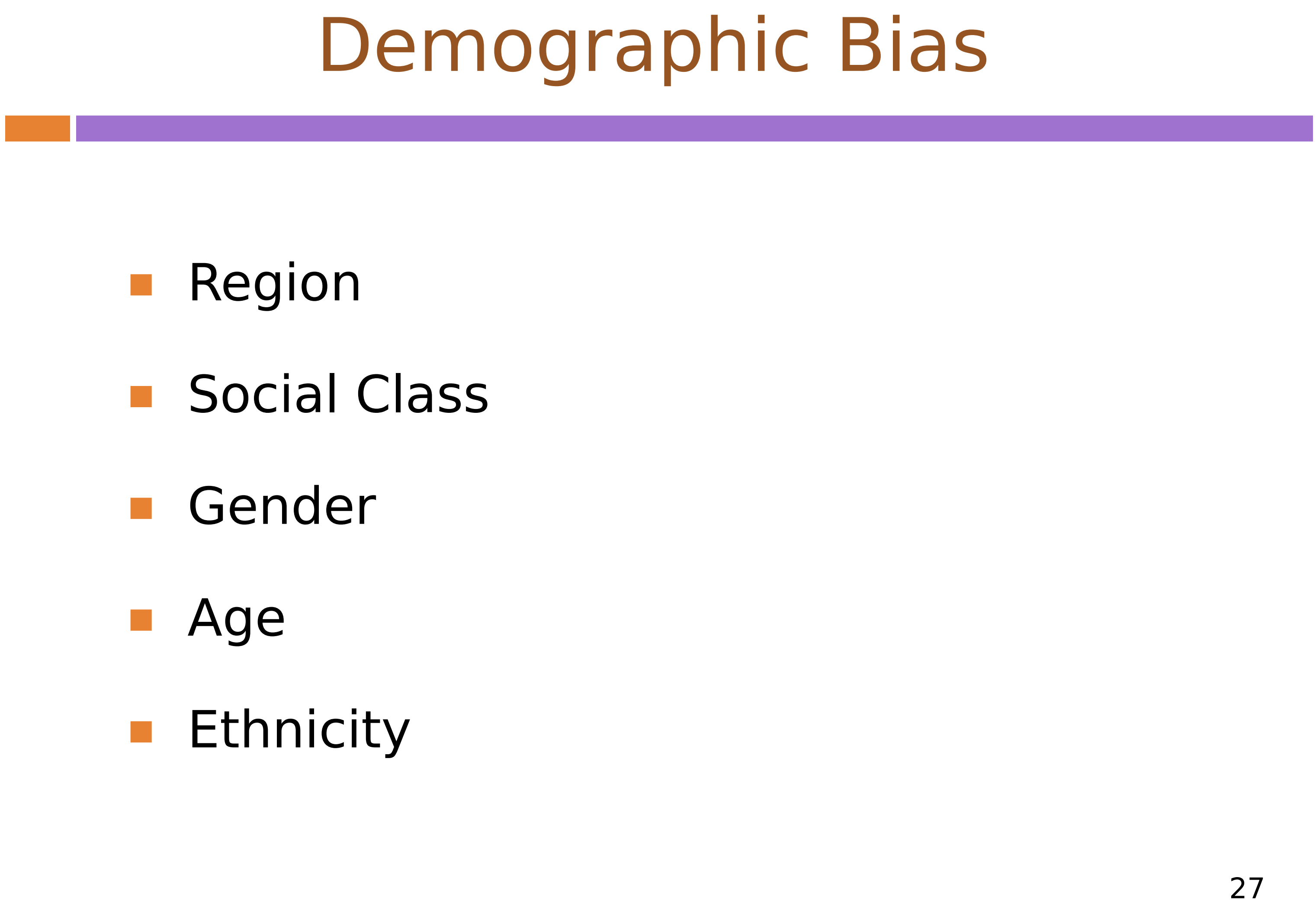
Bias

> “Bias is a social issue first,
> and a technical issue second.”
>
> (Crawford, 2017)

Kate Crawford (2017) “The Trouble With Bias”, NIPS Keynote Lecture

`https://www.youtube.com/watch?v=ggzWIipKraM`

Many machine learning researchers prefer to work on technical issues, but the social issues are still there. Social issues are important for any real-world application.



Demographic Bias

- Region
- Social Class
- Gender
- Age
- Ethnicity

This is an important type of social bias to be aware of: treating different demographic groups differently and unfairly.

Jørgensen et al. (2015)

- Many NLP tools trained on newspaper text (e.g. Penn Treebank)
- Test POS-taggers on Twitter data, incl. African-American Vernacular English:

Group	Stanf.	Gate	Ark
AAVE	.614	.791	.775
non-AAVE	**.745**	**.833**	.779

(significant differences in bold)

This is an example of demographic bias in NLP, where NLP tools are trained on the language of a particular demographic group.

For two of the taggers, the effect size is substantial.

For AAVE, POS-tagging is far from a solved problem!

The Gate and Ark taggers have been adapted for Twitter, while the Stanford tagger is not (but is often treated as a standard tool).

`http://aclweb.org/anthology/W15-4302`

Decision Making

- The Guardian (2017):
  "Computer says no: Irish vet fails oral English test needed to stay in Australia"
- Bias in training data
  vs. bias in decisions

Guardian article:
`https://www.theguardian.com/australia-news/2017/aug/08/computer-says-no-irish-vet-fails-oral-english-test-needed-to-stay-in-australia`

Follow-up Guardian article:
`https://www.theguardian.com/australia-news/2017/aug/10/outsmarting-the-computer-the-secret-to-passing-australias-english-proficiency-test`

This is a newspaper article, not a research article, so there's no comparison between English speakers from different countries. However, it illustrates the point that machine learning models are being used in practice, sometimes without carefully considering how they might go wrong.

Regardless of whether this particular system fails on Irish accents, this is a plausible scenario. In a real-world application, we need to make sure that a bias in the training data (such as not having any Irish accents) doesn't result in biased decisions (such as rejecting visas for people with Irish accents).

Privacy

- Collecting and analysing personal data requires consent
- Personal data must be stored securely
- Anonymising personal data is hard

A final ethical issue regards personal data. Even if there is no bias in the data, there are other reasons to be careful when we work with personal data.

Privacy

- Nouwens et al. (2020): “our empirical survey of CMPs [cookie banners] illustrates the extent to which illegal practices prevail”

Getting consent needs to be done carefully.

With the GDPR, many websites now have cookie banners asking for consent. However, many websites aren’t doing this properly!

`https://arxiv.org/abs/2001.02479`

Privacy

- Narayanan and Shmatikov (2007), on the Netflix Prize dataset: “Using the Internet Movie Database as background knowledge, we successfully identified known users”
- Four users sued Netflix

Anonymising personal data needs to be done carefully.

Just removing names often isn’t enough. If a person can be identified based on some part of the data, the data isn’t anonymous – and then the rest of the data might reveal sensitive information about that person.

Narayanan and Shmatikov (2007): `https://arxiv.org/abs/cs/0610105`

News article on the lawsuit: `https://www.wired.com/2010/03/netflix-cancels-contest/`

Summary of Ethics

- Bias in:
  - Training data
  - Model predictions
  - Real-world decisions
- Personal data
  - Consent to use of data
  - Access to data

What We’ve Covered

- Writing code
  - Backpropagation
  - Software packages
- Statistical Significance
  - Student’s t-test, Sign test
  - Base rate fallacy, Bonferroni correction
- Ethics
  - Bias
  - Privacy