

Machine Learning

Summary

Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

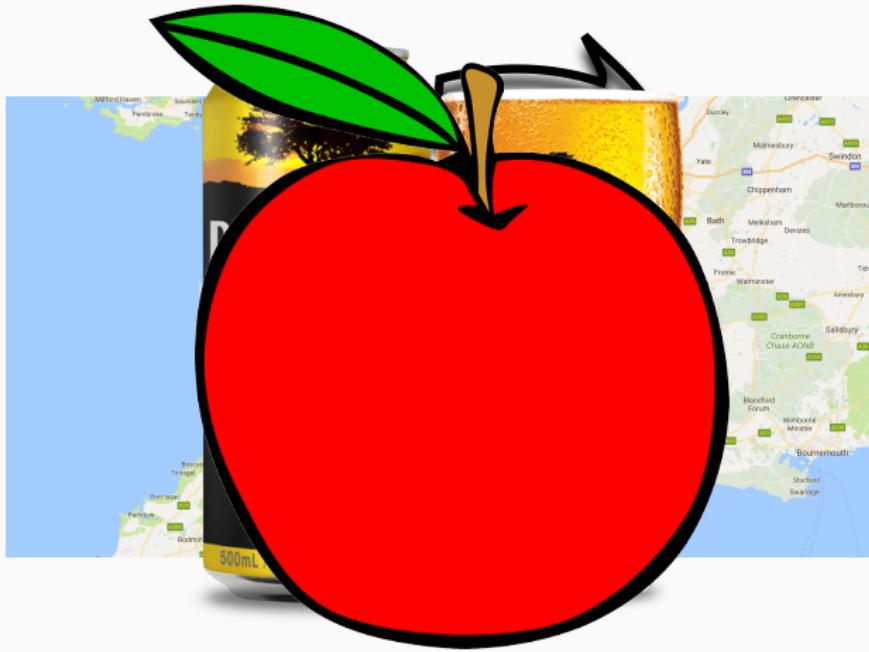
December 17, 2018

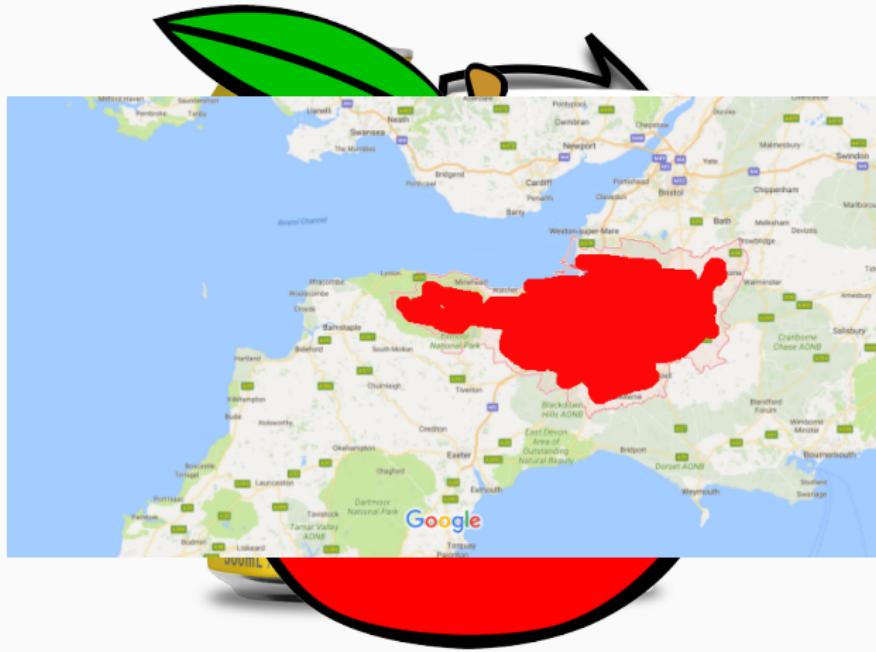
<http://www.carlhenrik.com>

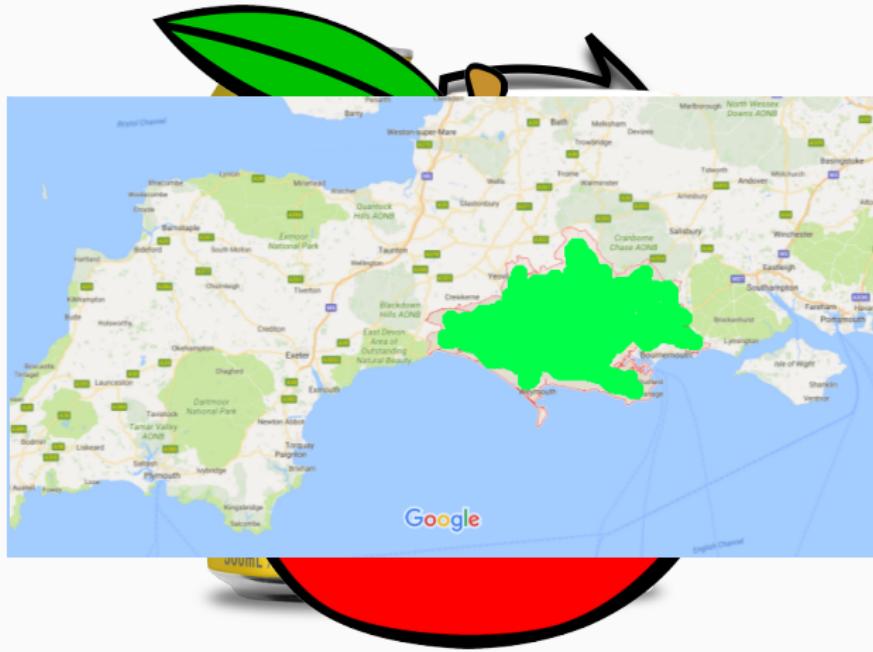


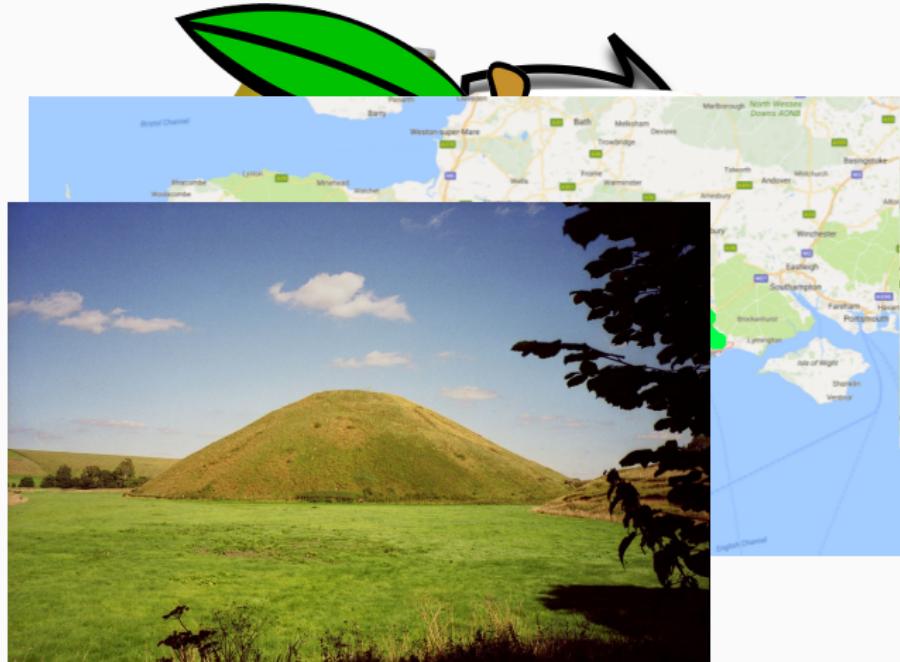


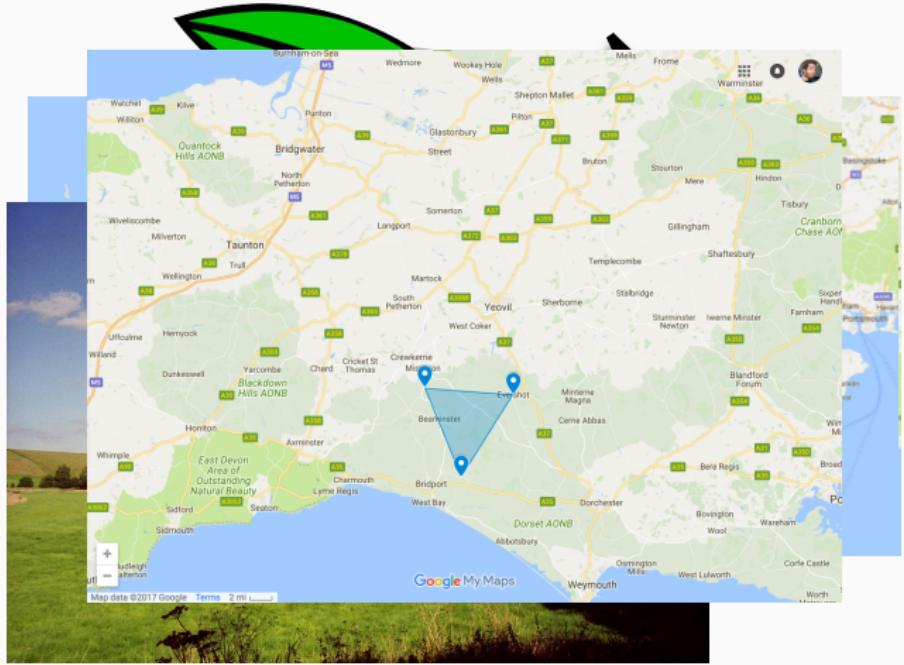




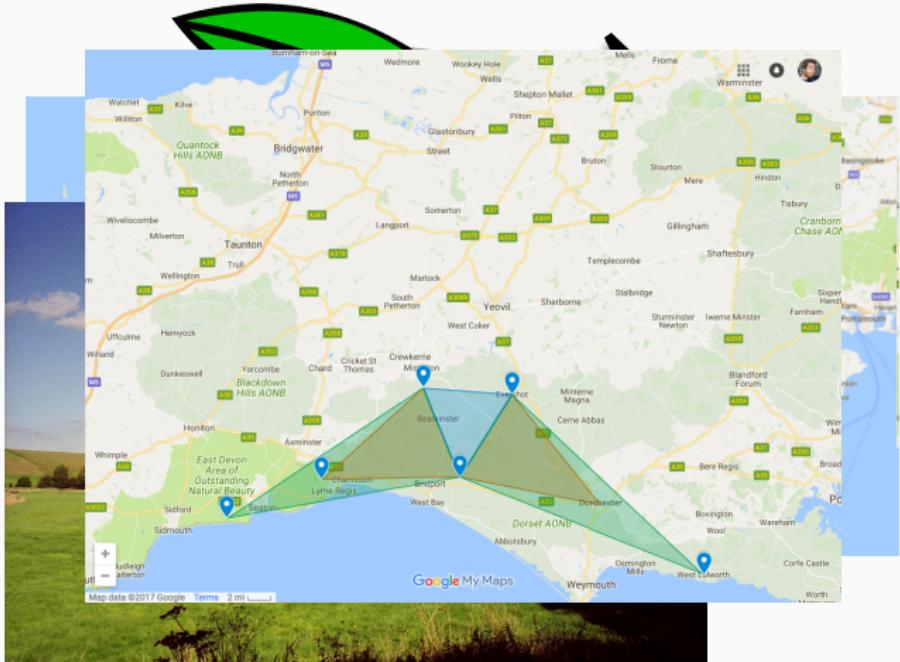




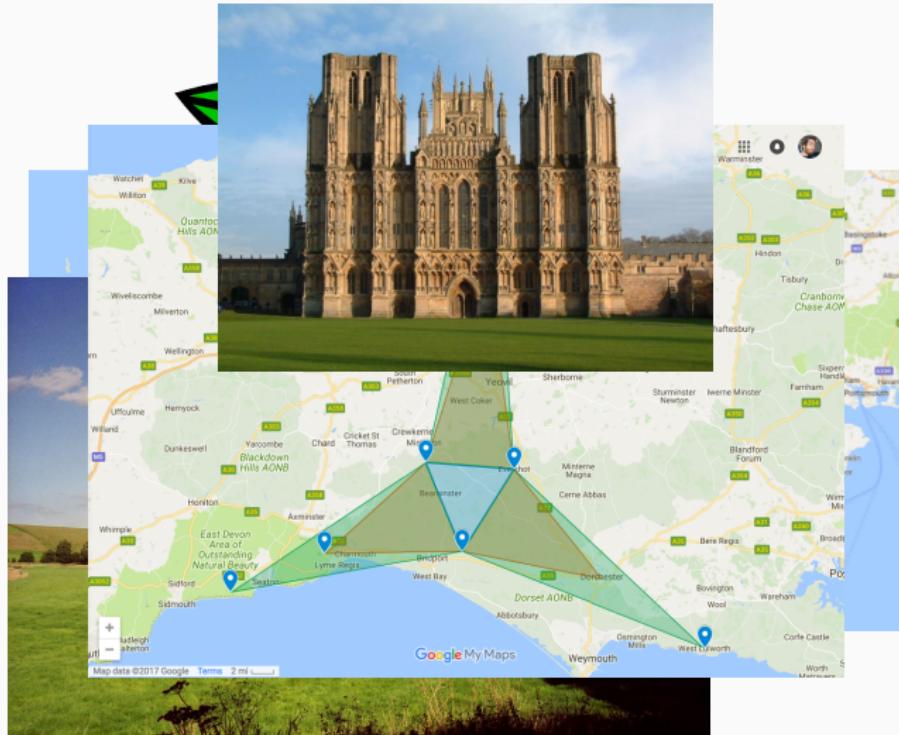


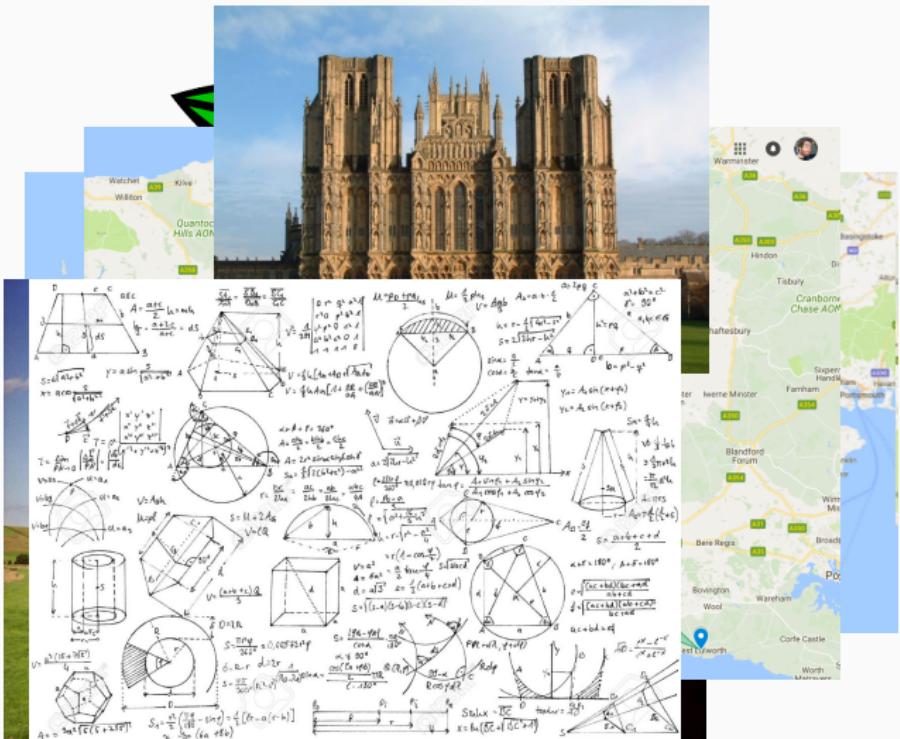


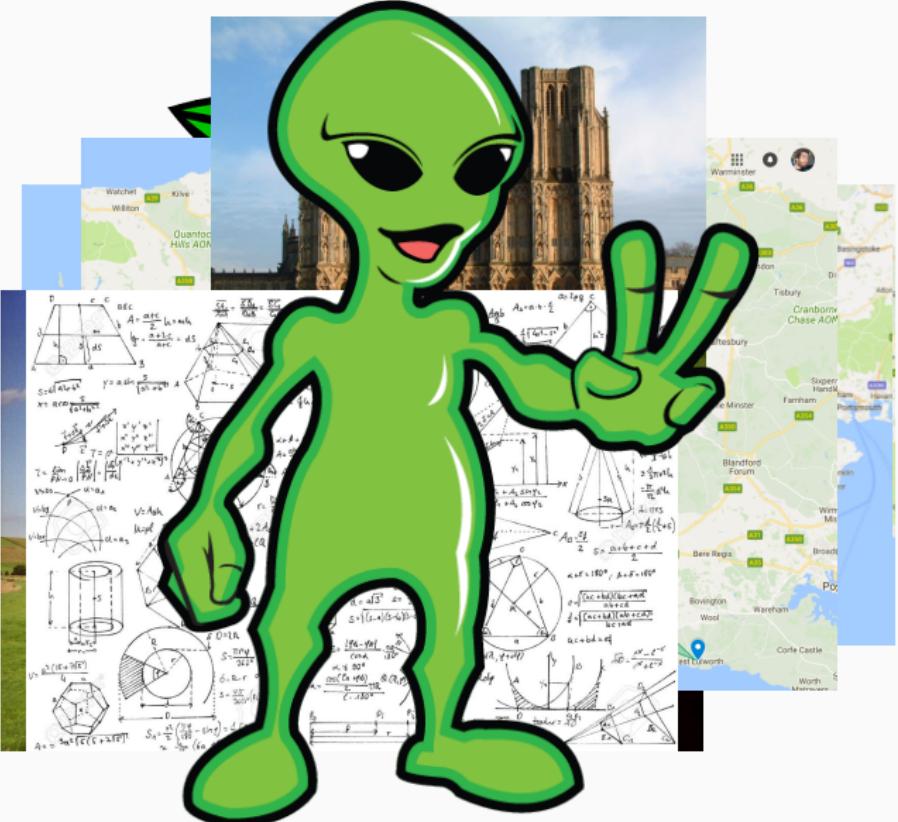










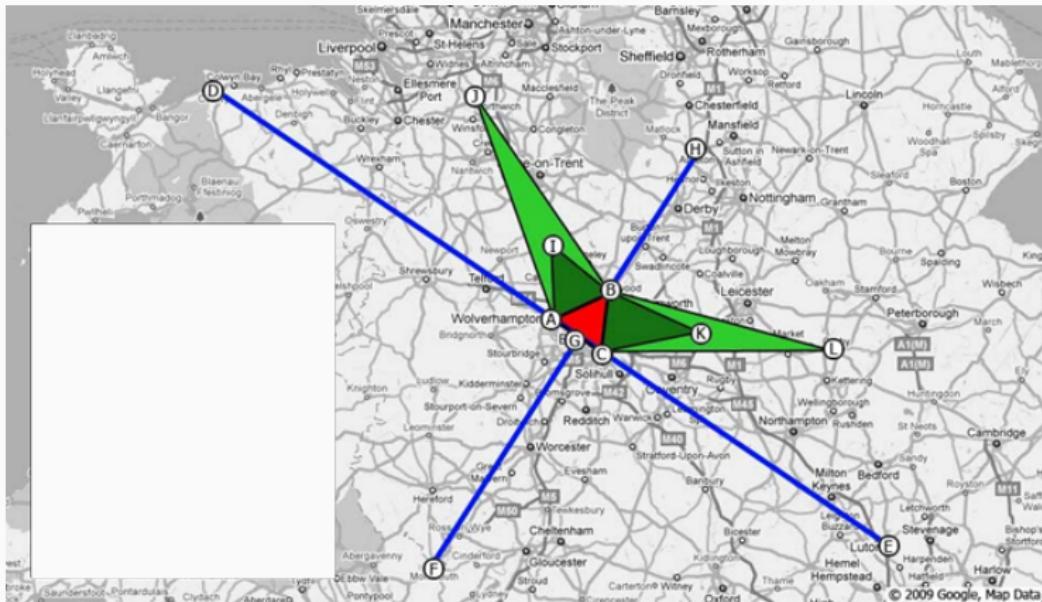


Tom Brooks



"Brooks has proved, he explains, that there were keen mathematicians here 5,000 years ago, millennia before the Greeks invented geometry." He does not rule out extraterrestrial help.

– The Guardian





"We know so little about the ancient Woolworths stores," he explains, "but we do still know their locations. I thought that if we analysed the sites we could learn more about what life was like in 2008 and how these people went about buying cheap kitchen accessories and discount CDs" — The Guardian¹

¹ <https://www.badscience.net/>

Laplace Demon [1]



Laplace Demon [1]

Laplace's Demon [1]

An intelligence which at a given instant knew all the forces acting in nature and the position of every object in the universe - if endowed with a brain sufficiently vast to make all necessary calculations - could describe with a single formula the motions of the largest astronomical bodies and those of the smallest atoms. To such an intelligence, nothing would be uncertain; the future, like the past, would be an open book.

Laplace Demon [1]

All these efforts in the search for truth tend to lead the mind continuously towards the intelligence we have just mentioned, although it will always remain infinitely distant from this intelligence.

Assumptions



Assumptions



Assumptions



Zero Sum Games



All models are wrong but some are useful

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

– George Box

All models are wrong but some are useful

"truth . . . is much too complicated to allow anything but approximations"

– John von Neumann

– no models are true — not even the Newtonian laws. When you construct a model you leave out all the details which you, with the knowledge at your disposal, consider inessential. . . . Models should not be true, but it is important that they are applicable, and whether they are applicable for any given purpose must of course be investigated. This also means that a model is never accepted finally, only on trial.

– George Rasch

No Free Lunch

- There are infinitely many explanations that explain a set of data

It's naive to believe in truth

- Very few (if any) things we know are true, we can only ever believe they are true

Good Assumptions

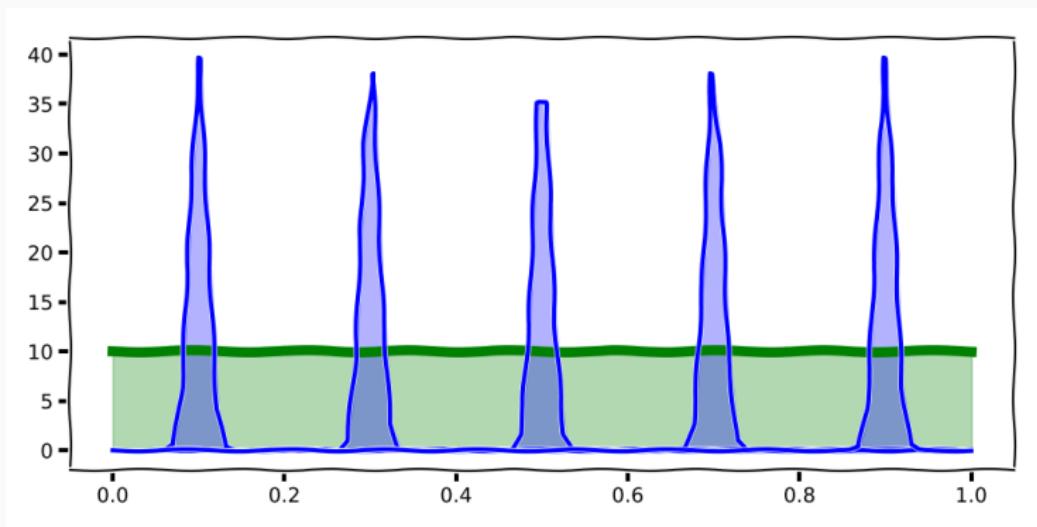
- A useful explanation is something that helps me solve a problem, independent if it is true or not

Ed Jaynes [2]



It was our use of probability theory as logic that has enabled us to do so easily what was impossible for those who thought of probability as a physical phenomenon associated with “randomness”. Quite the opposite; we have thought of probability distributions as carriers of information.

Assumptions



$$p(x) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Probability theory

- Distributions
- This provides a language

$$p(x) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Probability theory

- Distributions
- This provides a language
- *The language we have to formulate our assumptions and knowledge with*

$$p(z, y, x) = p(z|y)p(y|x)p(x)$$

Models

- How can we put together probability distributions to merge several different assumptions
- Examples of different priors, parametric, non-parametric etc.

$$p(z, y, x) = p(z|y)p(y|x)p(x)$$

Models

- How can we put together probability distributions to merge several different assumptions
- Examples of different priors, parametric, non-parametric etc.
- *How can we factorise a probability distribution such that we get as low entropy model as possible*

$$p(x|y) = \frac{p(y, x)}{p(y)} \approx q(x)$$

Inference

- Inductive reasoning means we can disprove our "theory"
- Learning implies interpreting data through our assumptions as the posterior distribution
- The posterior is often intractable, how can we approximate it

$$p(x|y) = \frac{p(y, x)}{p(y)} \approx q(x)$$

Inference

- Inductive reasoning means we can disprove our "theory"
- Learning implies interpreting data through our assumptions as the posterior distribution
- The posterior is often intractable, how can we approximate it
- *How can we reach an intractable posterior*

The other stuff

- Bayesian Optimisation
- Reinforcement Learning and decisions
- Neural networks
- Topic Models

The other stuff

- Bayesian Optimisation
- Reinforcement Learning and decisions
- Neural networks
- Topic Models
- *Examples of things that you can do and ways to think*

Ellen Key



*"Knowledge is what is left when you have forgotten
everything that you have learned"*

– Ellen Key

My aim

1. Kill the AI hype

My aim

1. Kill the AI hype
2. Identify what it fundamentally takes to learn

My aim

1. Kill the AI hype
2. Identify what it fundamentally takes to learn
3. Provide a strict and unified language of communication

Code

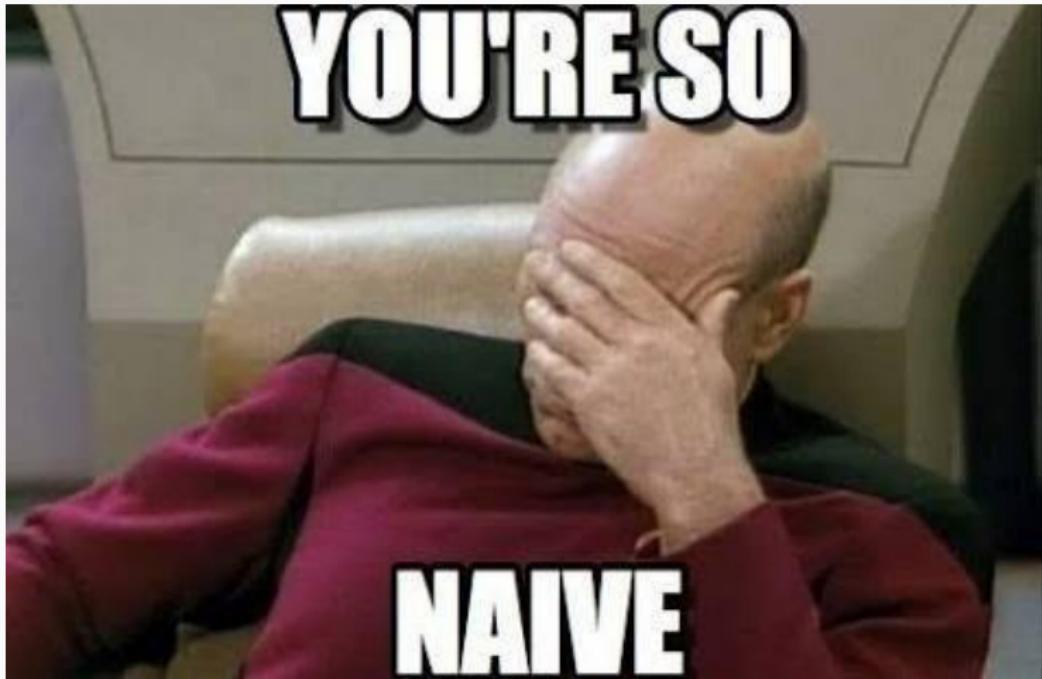
```
import tensorflow as tf

def cnn_model_fn(features, labels, mode):
    """Model function for CNN."""
    input_layer = tf.reshape(features["x"],
                           [-1, 28, 28, 1])

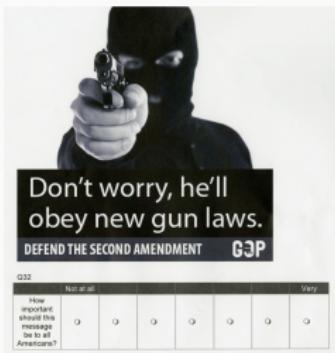
    # Convolutional Layer #1
    conv1 = tf.layers.conv2d(inputs=input_layer,
                           filters=32, kernel_size=[5, 5],
                           padding="same",
                           activation=tf.nn.relu)
```



CONSUMER
CREATOR



If we do not understand





- 700 000 000h watched every day
- 70% of all views are click-throughs of recommendations
- **Youtube the great radicaliser** - Zeynep Tufekci, New York Times

Why

AI today is described in breathless terms as computer algorithms that use silicon incarnations of our organic brains to learn and reason about the world, intelligent superhumans rapidly making their creators obsolete. The reality could not be further from the truth. As deep learning moves from the lab into production use in mission critical fields from medicine to driverless cars, we must recognize its very real limitations as nothing more than a pile of software code and statistics, rather than the learning and thinking intelligences we describe them as.

– *Kalev Leetaru - Forbes 15th of December 2018*

Why

.. a physicist does not proclaim that their analysis software is alive and thinking its own thoughts about the universe. They list the algorithm they used to analyze their dataset and talk about how and why it surfaced a new finding.

– Kalev Leetaru - Forbes 15th of December 2018

Right of Explanation²

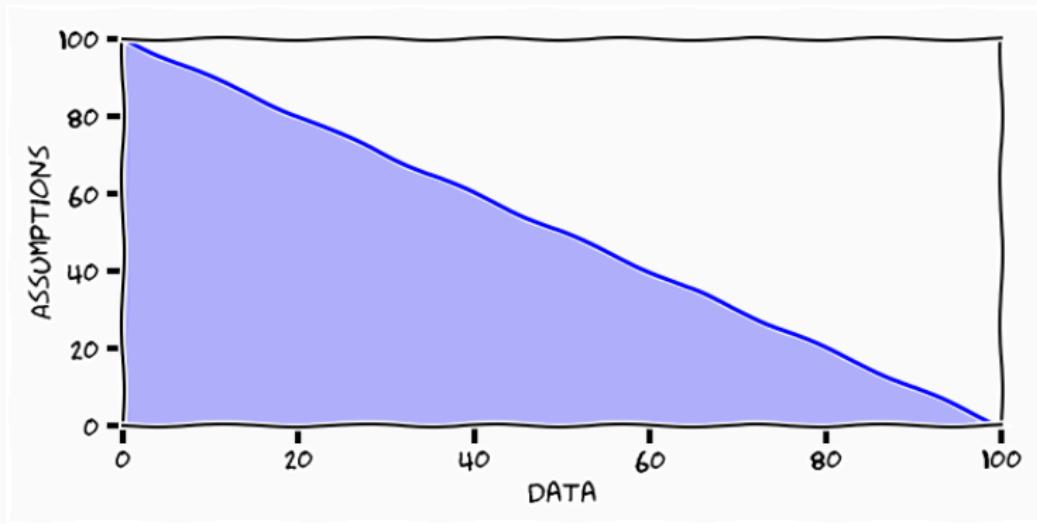


Article 22. Automated individual decision making, including profiling

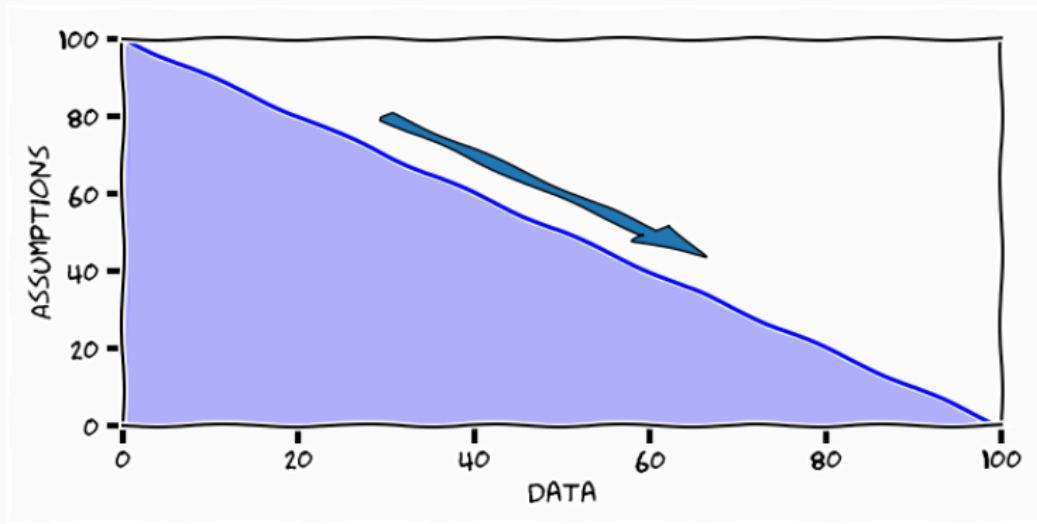
The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

²Parliament and Council of the European Union (2016). General Data Protection Regulation.

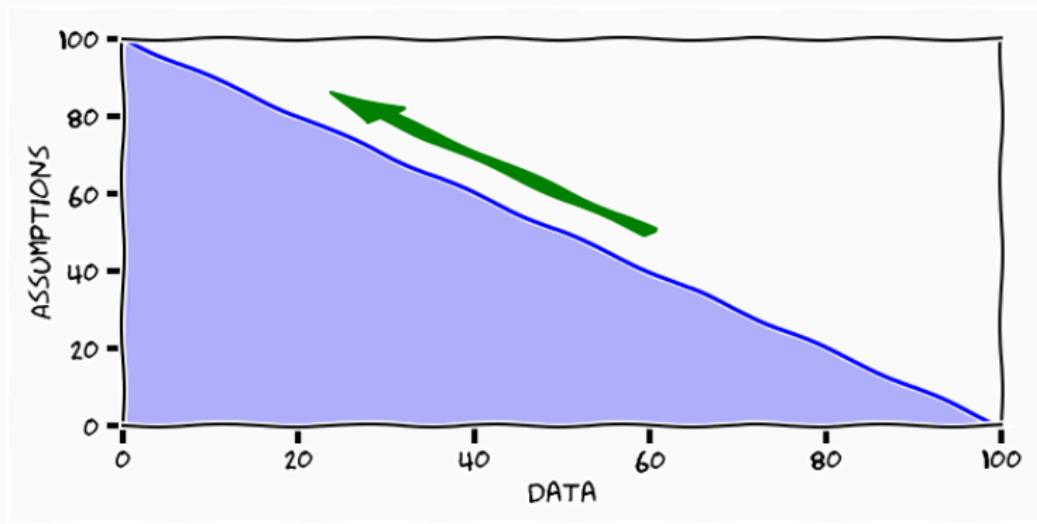
Machine Learning



Machine Learning



Machine Learning



What to do next?

Data



Methods



- always remember and think about the trade-off between principle and what it allows you to do

Communication



- A machine learner does not know anything but can do everything if we can communicate

Quotes

"Don't just sit on the runway and hope someone will come along and push the airplane. It simply won't happen. Change your attitude and gain some altitude."



Quotes

"Remember, new "environment friendly" lightbulbs can cause cancer. Be careful— the idiots who came up with this stuff don't care."

– Donald Trump (President of the United States, yes seriously)

Quotes

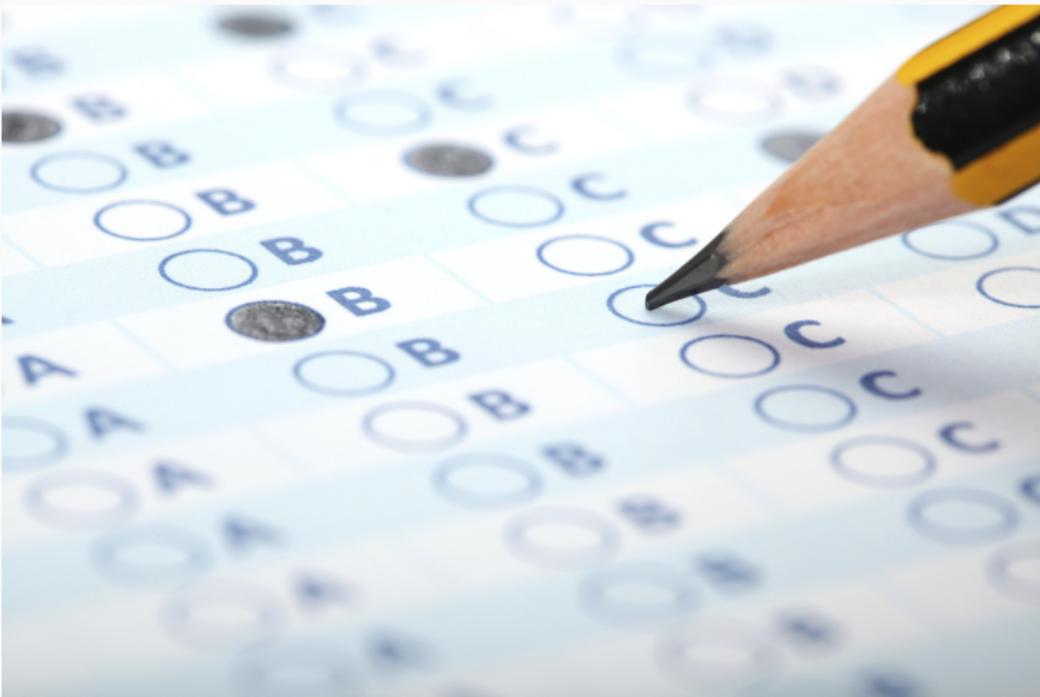
"I've said if Ivanka weren't my daughter, perhaps I'd be dating her."
– Donald Trump (President of the United States, yes seriously)

The leader of the free world



Exam

Exam



Exam

- 20 Questions
- Multiple choice
- Conceptual understanding

Material

- Assignments
 - read through your own reports and the notes, what did you actually do? What where the key conceptual ideas.

Material

- Assignments
 - read through your own reports and the notes, what did you actually do? What where the key conceptual ideas.
- Summary document
 - main thing is the conceptual ideas

Material

- Assignments
 - read through your own reports and the notes, what did you actually do? What were the key conceptual ideas.
- Summary document
 - main thing is the conceptual ideas
- Lecture notes
 - summarise each lecture to yourself, there are a few things in each lecture, understand them but skip the details
 - topic model lecture summarises a lot of part I and II

Question 1

*The conjugate prior to the **scale** θ of a Weibull distribution with known shape β is a Inverse-Gamma distribution. Which functional form will the posterior distribution over the scale θ take?*

Answer 1

Definition (Conjugate Prior)

In Bayesian probability theory, if the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

Question 2

Can any function be a kernel function?

Answer 2

$$||\phi(x_i) - \phi(x_j)||_2^2$$

Answer 2

$$||\phi(x_i) - \phi(x_j)||_2^2 = (\phi(x_i) - \phi(x_j))^T(\phi(x_i) - \phi(x_j))$$

Answer 2

$$\begin{aligned} \|\phi(x_i) - \phi(x_j)\|_2^2 &= (\phi(x_i) - \phi(x_j))^T (\phi(x_i) - \phi(x_j)) \\ &= \phi(x_i)^T \phi(x_i) + \phi(x_j)^T \phi(x_j) - 2\phi(x_i)^T \phi(x_j) \end{aligned}$$

Answer 2

$$\begin{aligned} \|\phi(x_i) - \phi(x_j)\|_2^2 &= (\phi(x_i) - \phi(x_j))^T (\phi(x_i) - \phi(x_j)) \\ &= \phi(x_i)^T \phi(x_i) + \phi(x_j)^T \phi(x_j) - 2\phi(x_i)^T \phi(x_j) \\ &= k(x_i, x_i) + k(x_j, x_j) - 2k(x_i, x_j) \end{aligned}$$

Answer 2

$$\begin{aligned} \|\phi(x_i) - \phi(x_j)\|_2^2 &= (\phi(x_i) - \phi(x_j))^T(\phi(x_i) - \phi(x_j)) \\ &= \phi(x_i)^T\phi(x_i) + \phi(x_j)^T\phi(x_j) - 2\phi(x_i)^T\phi(x_j) \\ &= k(x_i, x_i) + k(x_j, x_j) - 2k(x_i, x_j) \end{aligned}$$

- The kernel function **needs** to represent an inner-product in a metric space, i.e. the triangle inequality needs to hold

Question 3

In Bayesian optimisation we use a surrogate model to optimise an intractable function.

- *If we design an acquisition function that does not explore sufficiently, is our optimisation likely to become more or less dependent on initialisation?*
- *If our function explores inefficiently what is the likely effect on the optimisation to be?*

Question 3

In Bayesian optimisation we use a surrogate model to optimise an intractable function.

- *If we design an acquisition function that does not explore sufficiently, is our optimisation likely to become more or less dependent on initialisation?*
 - More dependent on a good initialisation
- *If our function explores inefficiently what is the likely effect on the optimisation to be?*

Question 3

In Bayesian optimisation we use a surrogate model to optimise an intractable function.

- *If we design an acquisition function that does not explore sufficiently, is our optimisation likely to become more or less dependent on initialisation?*
 - More dependent on a good initialisation
- *If our function explores inefficiently what is the likely effect on the optimisation to be?*
 - The optimisation is likely to require more iterations

Question 4

Marginalisation is the process of integrating out our belief in a variable. This is useful because the uncertainty in the variable disappears.

Answer 4

Wrong, very wrong, we take the expectation of the variable so we include all our knowledge in the variable

Question 5

When using a mean-field approximation to a variational distribution, what is our main assumption?

Answer 5

$$q(\mathbf{x}) = \prod_i^D q_i(x_i) \approx p(\mathbf{x}|\mathbf{y})$$

That we model the marginals of the posterior distribution rather than the joint.

Summary

- Go through the assignment, what where the actual things that you needed to think of, the choices that enabled you to solve the tasks, the assumptions that you did

Summary

- Go through the assignment, what where the actual things that you needed to think of, the choices that enabled you to solve the tasks, the assumptions that you did
 - are you clear that linear regression and PCA are **exactly** the same thing?

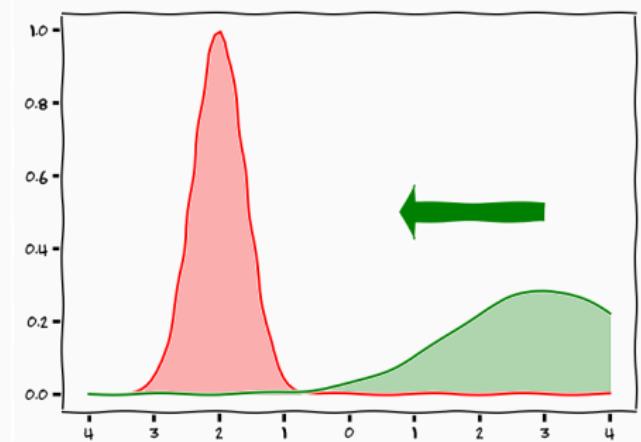
Summary

- Go through the assignment, what where the actual things that you needed to think of, the choices that enabled you to solve the tasks, the assumptions that you did
 - are you clear that linear regression and PCA are **exactly** the same thing?
- Go through the lecture notes/summary document, what is the key conceptual messages for each lecture, understand these

Summary

- Go through the assignment, what where the actual things that you needed to think of, the choices that enabled you to solve the tasks, the assumptions that you did
 - are you clear that linear regression and PCA are **exactly** the same thing?
- Go through the lecture notes/summary document, what is the key conceptual messages for each lecture, understand these
- 2h exam, 20 questions, i.e. 6 minutes per questions, if it is complicated, you are on the wrong track

Back yourself



$$p(\text{ML}|\text{COMS30007}) = \frac{p(\text{COMS30007}|\text{ML})p(\text{ML})}{p(\text{COMS30007})}$$

Course Review

Fill out the course review on SAFE

eof

References

 Pierre Simon Laplace.

A philosophical essay on probabilities, 1814.

 E T Jaynes.

Probability theory: The logic of science.

Cambridge university press, June 2003.