# Advanced Data Science

## Lecture 9 : Statistical Learning Outlook

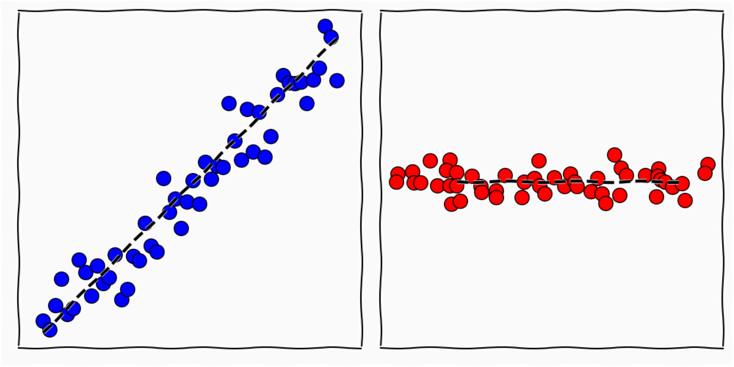
Carl Henrik Ek - che29@cam.ac.uk

16th of November, 2022

http://carlhenrik.com

# Introduction

# Linear Dimensionality Reduction



# PCA and MDS

**Principal Component Analysis** diagonalises a $D \times D$ matrix

# PCA and MDS

**Principal Component Analysis** diagonalises a $D \times D$ matrix

- finds a geometrical representation where the covariance is diagonal

# PCA and MDS

**Principal Component Analysis** diagonalises a $D \times D$ matrix

- finds a geometrical representation where the covariance is diagonal

**Multi-Dimensional-Scaling** diagonalises a $N \times N$ matrix

# PCA and MDS

**Principal Component Analysis** diagonalises a $D \times D$ matrix

- finds a geometrical representation where the covariance is diagonal

**Multi-Dimensional-Scaling** diagonalises a $N \times N$ matrix

- finds a geometrical representation that "matches" a distance matrix

# PCA and MDS

**Principal Component Analysis** diagonalises a $D \times D$ matrix

- finds a geometrical representation where the covariance is diagonal

**Multi-Dimensional-Scaling** diagonalises a $N \times N$ matrix

- finds a geometrical representation that "matches" a distance matrix
- equivivalent to PCA with euclidian distance

# PCA and MDS

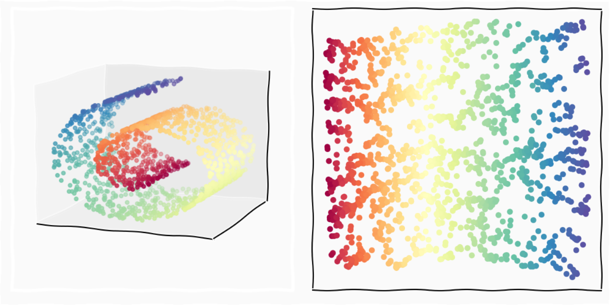
**Principal Component Analysis** diagonalises a $D \times D$ matrix

- finds a geometrical representation where the covariance is diagonal

**Multi-Dimensional-Scaling** diagonalises a $N \times N$ matrix

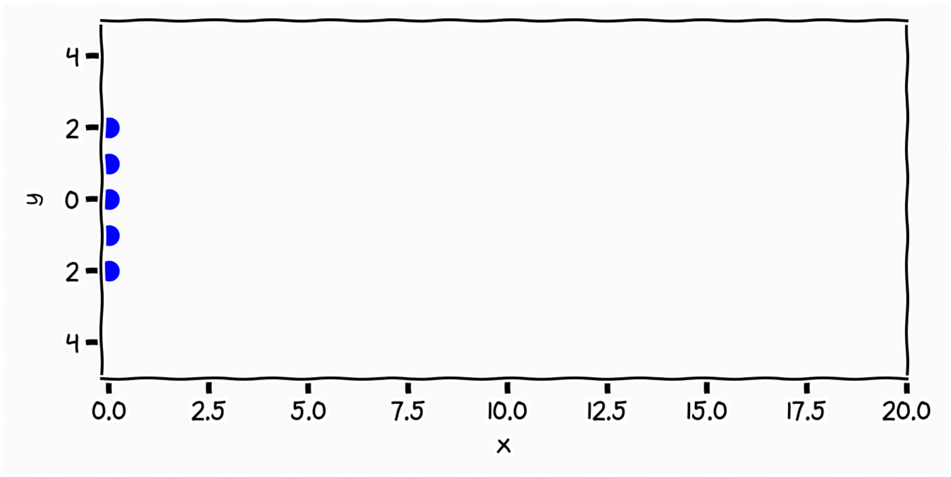
- finds a geometrical representation that "matches" a distance matrix
- equivivalent to PCA with euclidian distance
- can be non-linearised with a non-linear distance measure

# Generative Model

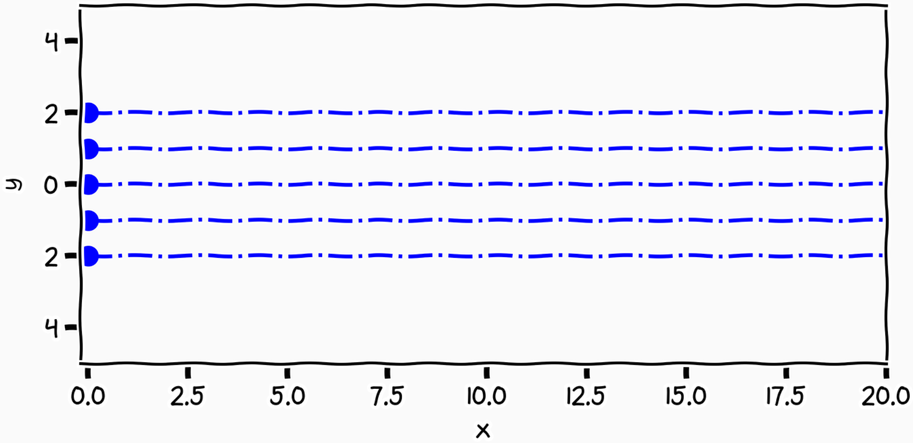


$$\mathbf{y}_i = f(\mathbf{x}_i)$$

# Unsupervised Learning



# Unsupervised Learning



# Ill posed

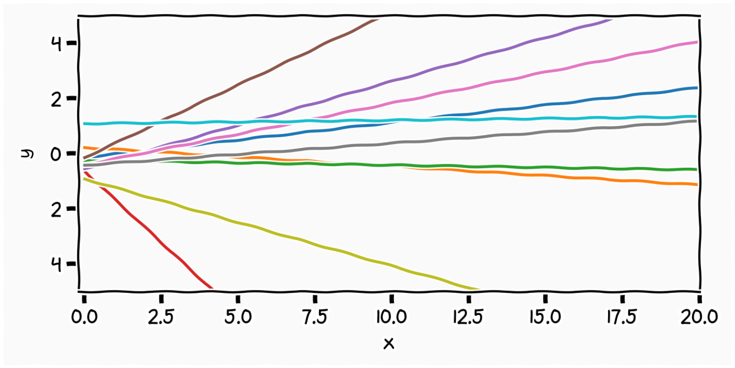
- This problem is very ill-posed
- We have to encode a preference towards the solution that we want

# Generalised Linear Model: Learning

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmax}_{\boldsymbol{\beta}} \prod_{i=1}^{N} p(y_i \mid \boldsymbol{\beta}, \mathbf{x}_i)$$

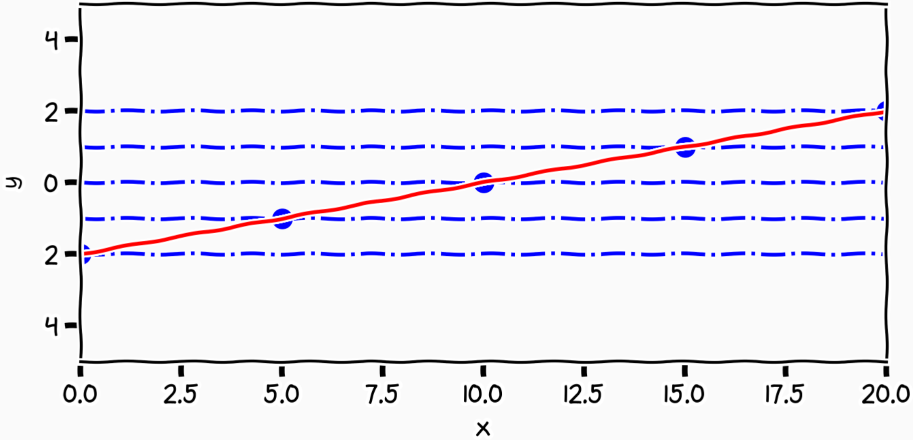
# Generalised Linear Model: Learning

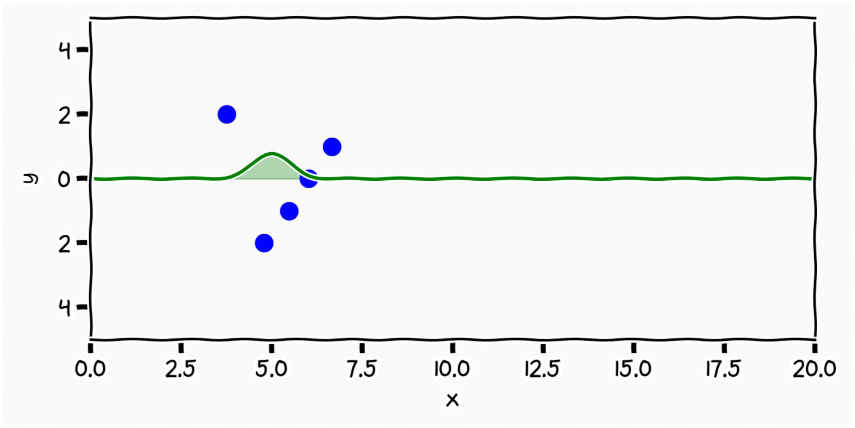
$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(y_i \mid \boldsymbol{\beta}, \mathbf{x}_i) + \lambda \left( \sum_{j=1}^{d} \beta_j^p \right)^{\frac{1}{p}}$$



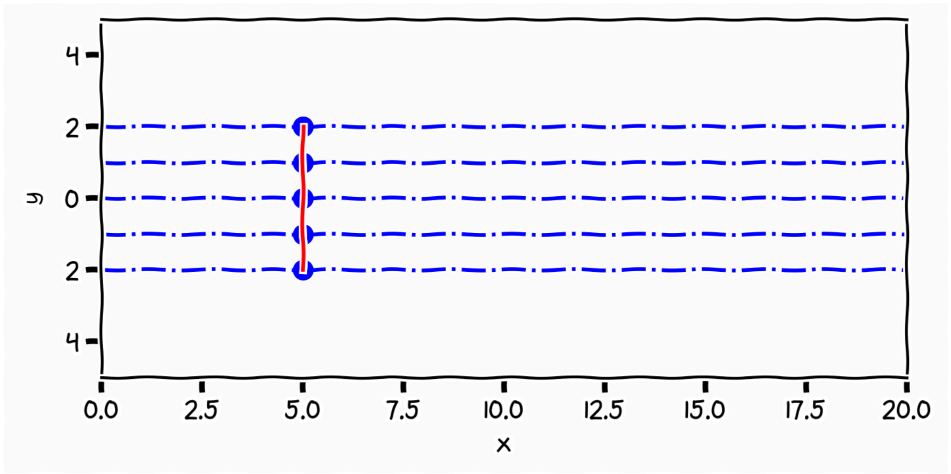
$$y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

$$p(\mathbf{w}) \sim \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha\mathbf{I})$$

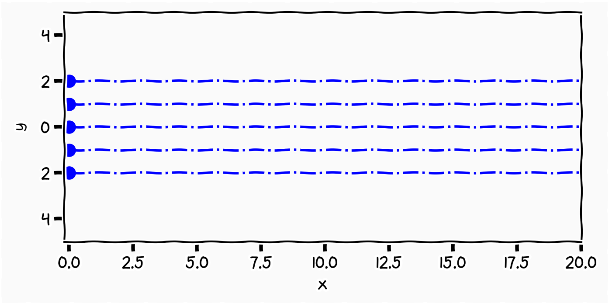

4
2
0
2
4
y
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
x



$$p(\mathbf{X}) \sim \mathcal{N}(\mathbf{X} \mid \mathbf{0}, \alpha_2 \mathbf{I})$$

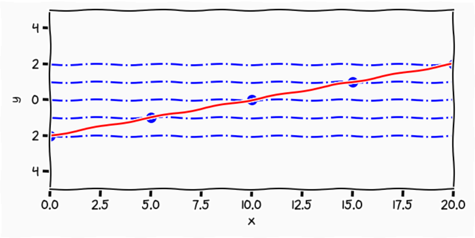


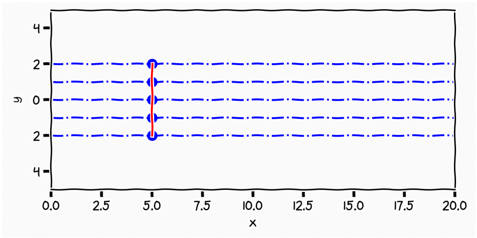
$$p(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \alpha_2 \mathbf{I})$$



$$\{\hat{\mathbf{X}}, \hat{\mathbf{w}}\} = \underset{\hat{\mathbf{X}},\hat{\mathbf{w}}}{\operatorname{argmax}} \left( \underbrace{\mathcal{L}(\mathbf{Y}, \mathbf{X}, \mathbf{w})}_{\log p(\mathbf{Y}|\mathbf{X},\mathbf{w})} + \gamma_1 \log p(\mathbf{w}) + \gamma_2 \log p(\mathbf{X}) \right)$$

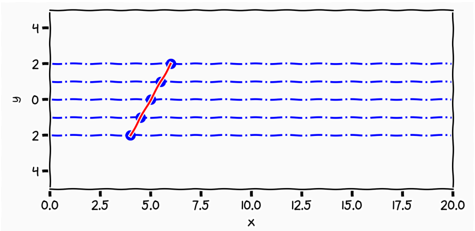
# Unsupervised Learning





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

$$p(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \alpha_2\mathbf{I})$$



# Today

**Statistical learning** machine learning is inherently ill-posed

# Today

**Statistical learning** machine learning is inherently ill-posed

**Interpretation** our "predictions" can only ever be interpreted in light of the knowledge we put in

# Today

**Statistical learning** machine learning is inherently ill-posed

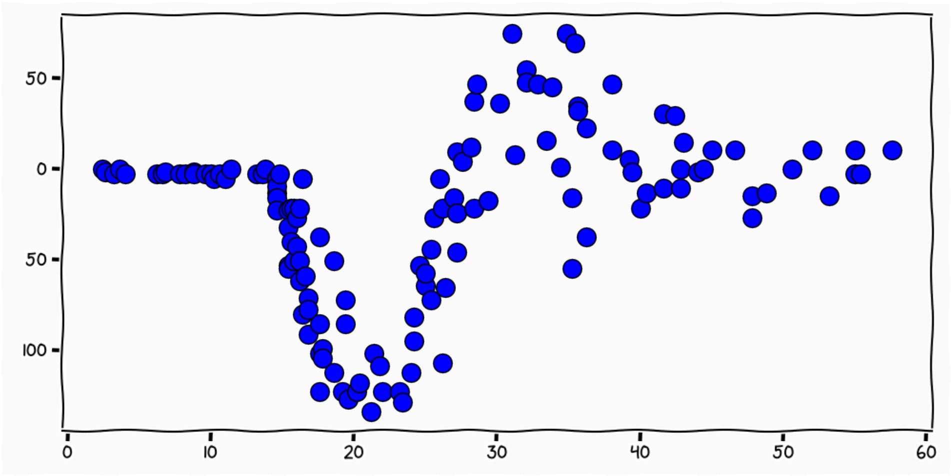
**Interpretation** our "predictions" can only ever be interpreted in light of the knowledge we put in

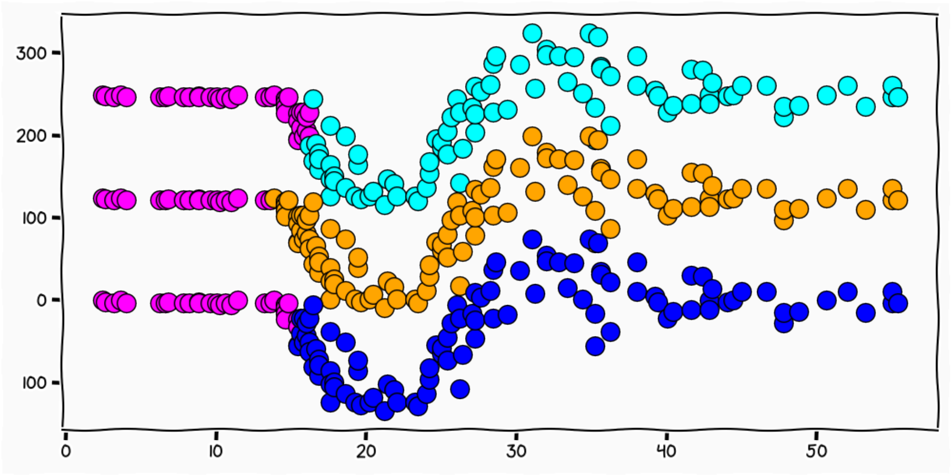
**Knowledge** how can we incorporate knowledge in a principled manner

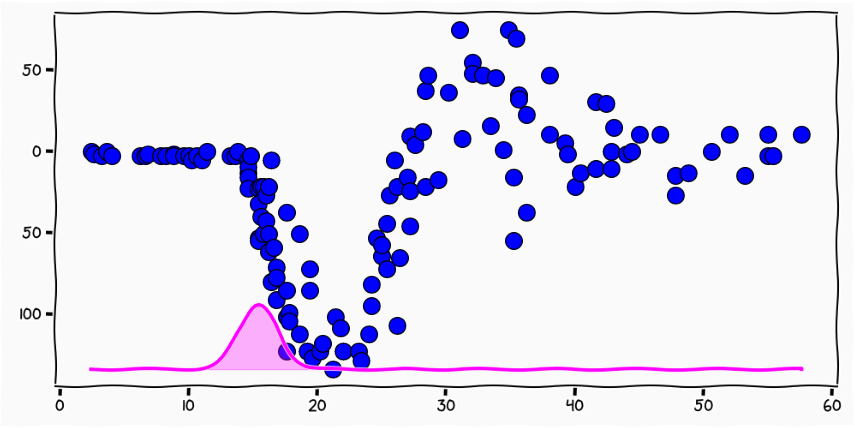
# Knowledge/Assumption in the Data Science Pipeline

| | |
|---|---|
| **access** | what data did I acquire |
| **assess** | how did I prepare/treat the data |
| **address** | which model to choose, how did I set the parameters of the model |

# Learning


50
0
50
100
0
10
20
30
40
50
60


300
200
100
0
100
0
10
20
30
40
50



$$p(t) \sim \mathcal{N}(15, 1.5)$$

# Variables

## Deterministic Variable

Code

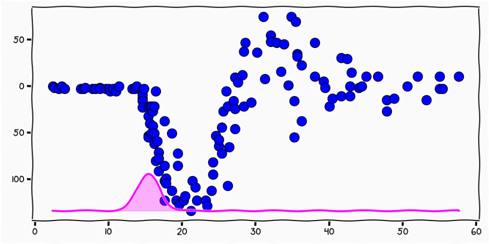
```
int x = 3;
float y = 3.14;
```

## Stochastic Variable

$$x \sim p(x)$$

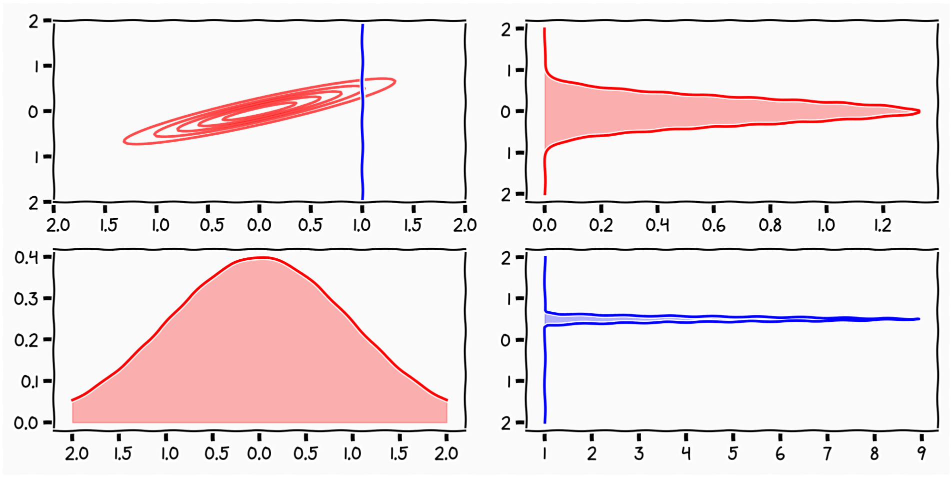
$$y \sim \mathcal{N}(0, 1)$$

# Encoding Knowledge



$$\tilde{f}(x) = \int f(x,t)p(t)\mathrm{d}t$$

# Basic Probabilities



# Rules of Probablity

Sum Rule

$$p(x) = \sum_{\forall y \in \mathcal{Y}} p(x, y)$$

Product Rule

$$p(x, y) = p(x \mid y) p(y)$$

# Marginalisation

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) p(\theta) \mathrm{d}\theta$$

# Marginalisation

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) p(\theta) \mathrm{d}\theta$$

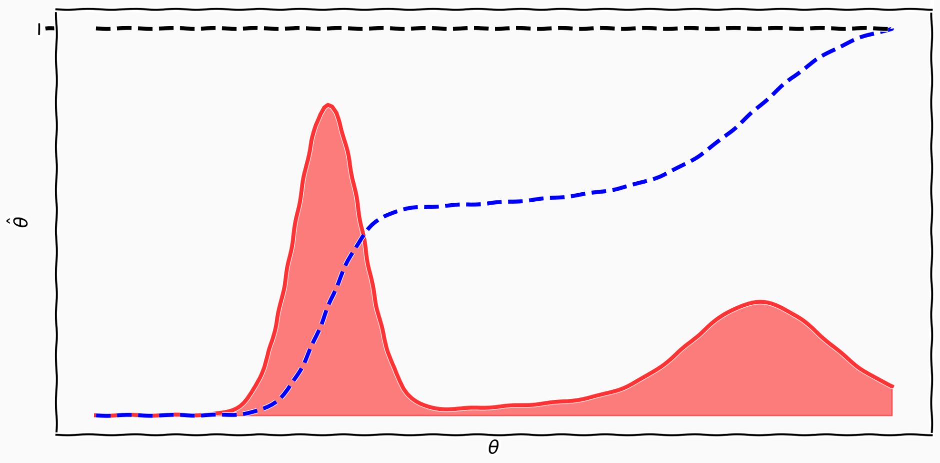
# Marginalisation

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) p(\theta) \mathrm{d}\theta$$

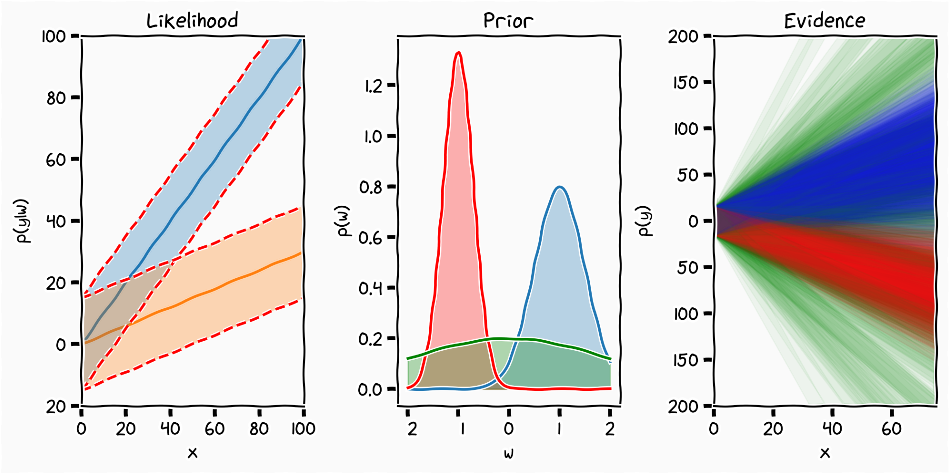
# Marginalisation

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta)\, \underbrace{p(\theta)\mathrm{d}\theta}_{\mathrm{d}\hat{\theta}}$$

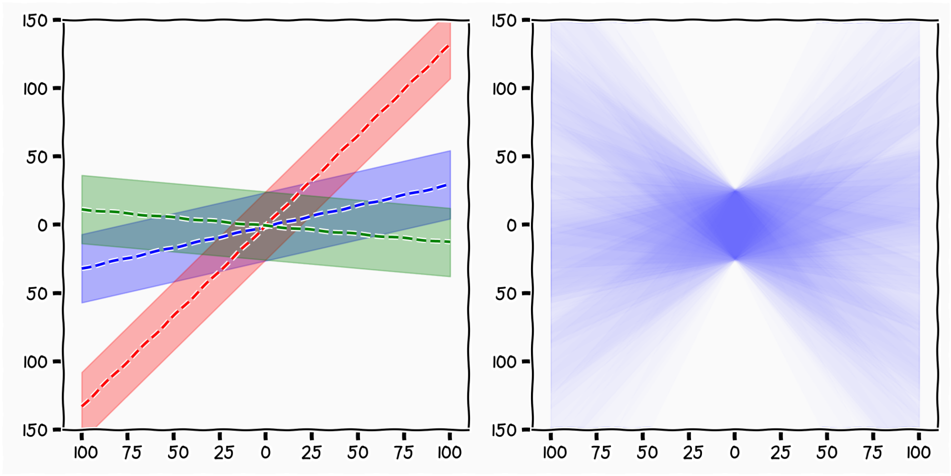
# Marginalisation



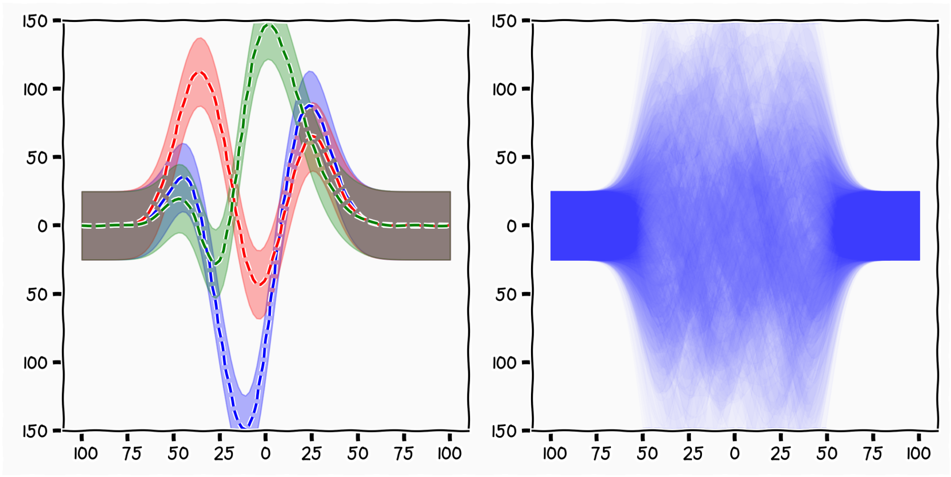
# Marginalisation



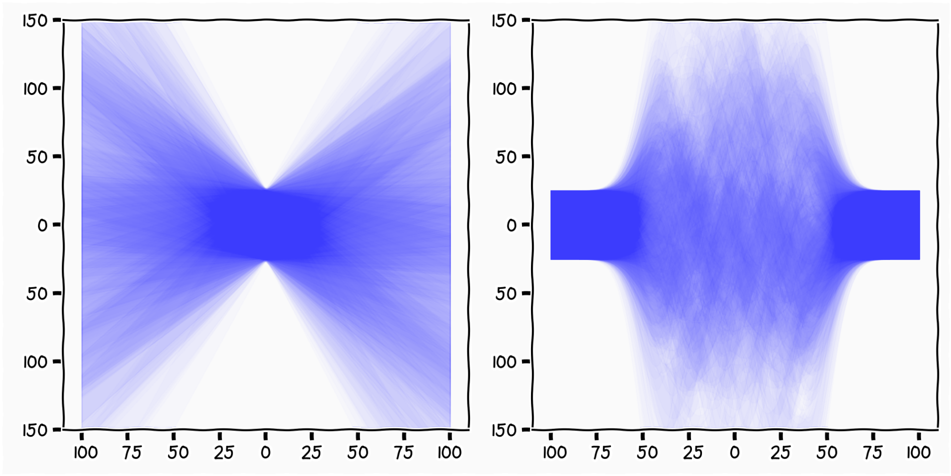
# Marginalisation Model Linear



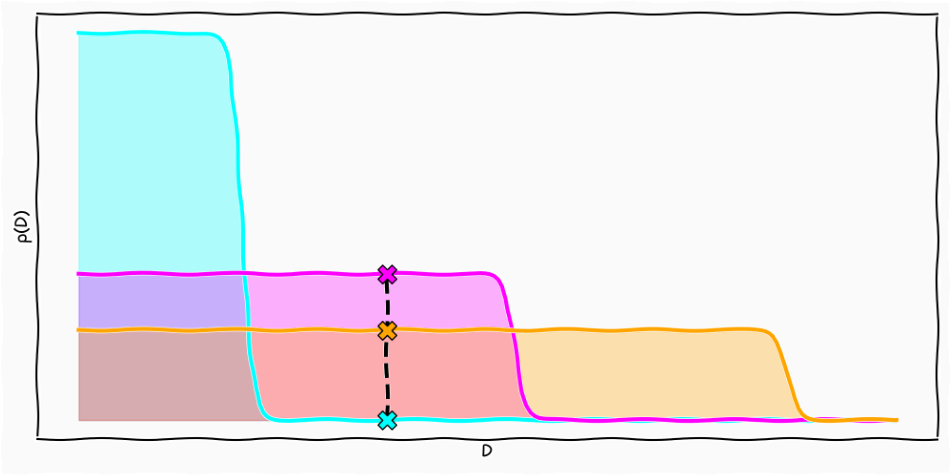
# Marginalisation Model Basis



# Marginalisation Model



# Model Selection [Mackay, 1991]



3?

# Bayes' "Rule"

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

# Bayes' "Rule"

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

# Bayes' "Rule"

$$\begin{aligned} p(x,y) &= p(y|x)p(x) \\ p(x,y) &= p(x|y)p(y) \\ p(x|y)p(y) &= p(y|x)p(x) \end{aligned}$$

# Bayes' "Rule"

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

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

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

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

# Bayes' "Rule"

$$
\begin{aligned}
p(x, y) &= p(y|x)p(x) \\
p(x, y) &= p(x|y)p(y) \\
p(x|y)p(y) &= p(y|x)p(x) \\
p(x|y) &= \frac{p(y|x)p(x)}{p(y)} \\
&= \frac{p(y|x)p(x)}{\sum_x p(y|x)p(x)}
\end{aligned}
$$

# Bayes' Rule Semantics

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}$$

**Likelihood** for a specific parameter setting how does the observation manifest itself

**Prior** what do I believe/know about the parameters

**Evidence** what is the probability of a specific set of data

**Posterior** what is the probability for different parameter settings given a set of data

# Regularisation

**Ad-hoc Regularisation** maximum likelihood or regularised error

$$\{\hat{\mathbf{X}}, \hat{\mathbf{w}}\} = \operatorname*{argmax}_{\hat{\mathbf{X}}, \hat{\mathbf{w}}} \left( \underbrace{\mathcal{L}(\mathbf{Y}, \mathbf{X}, \mathbf{w})}_{\log p(\mathbf{Y}|\mathbf{X},\mathbf{w})} + \gamma_1 \log p(\mathbf{w}) + \gamma_2 \log p(\mathbf{X}) \right)$$

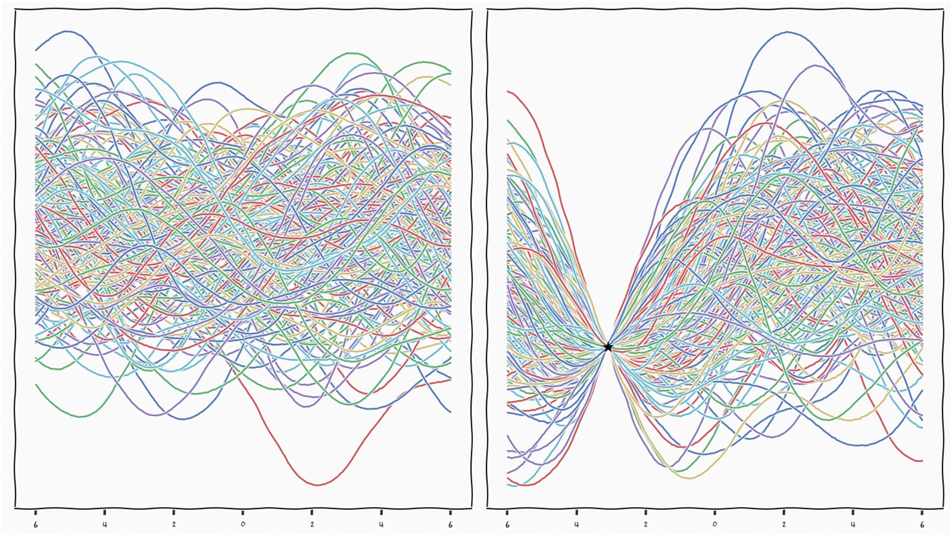
**Principled Regularisation** posterior distribution

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta) p(\theta)}{\int p(\mathcal{D} \mid \theta) p(\theta) \mathrm{d}\theta}$$

# The importance of Integration

> *Integration is a key step in inference, where it is encountered when averaging over the many states of the world consistent with observed data. Indeed, a provocative Bayesian view is that integration is the single challenge separating us from systems that fully automate statistics. More speculatively still, such systems may even exhibit artificial intelligence (ai)*
>
> *– Universal Artificial Intelligence - M. Hutter*

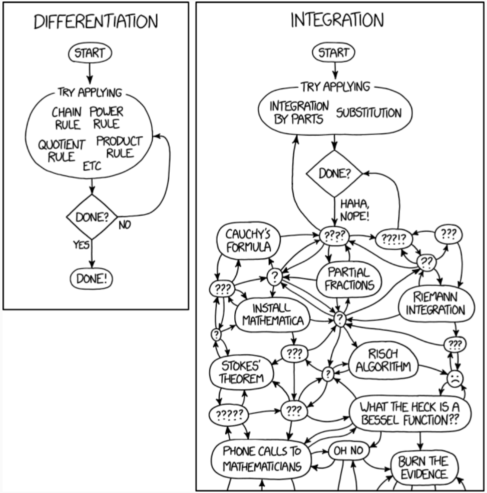
# The challenge of Marginalisation[1]



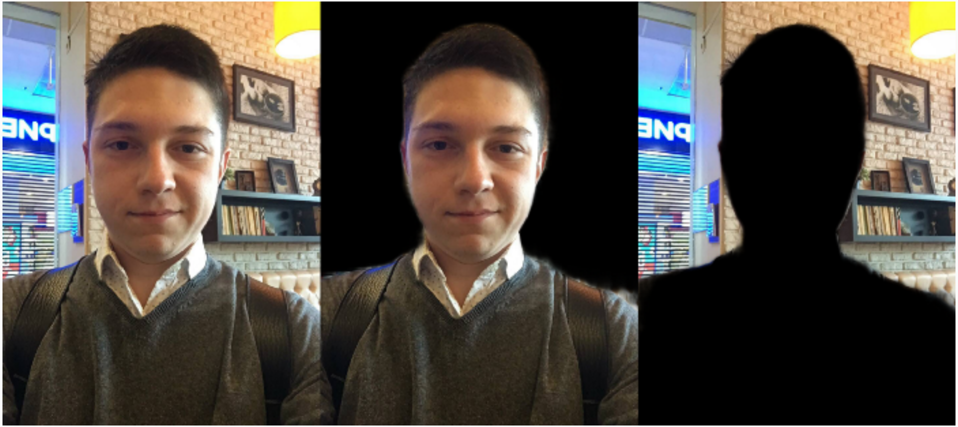
[1] or machine learning as a whole

# Laplace Integration

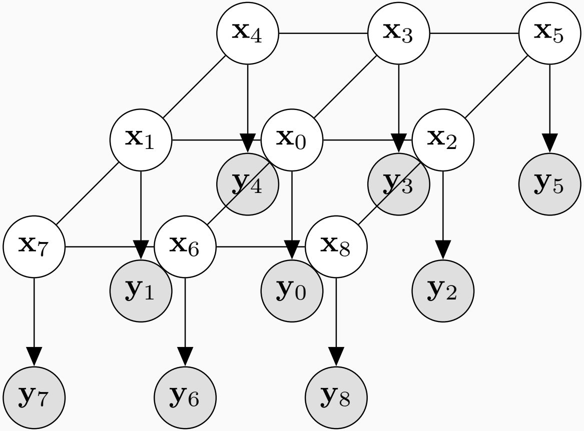
*"Nature laughs at the difficulties of integrations"*
*– Simon Laplace*


DIFFERENTIATION
START
TRY APPLYING
CHAIN RULE
POWER RULE
QUOTIENT RULE
PRODUCT RULE
ETC
DONE?
NO
YES
DONE!
INTEGRATION
START
TRY APPLYING
INTEGRATION BY PARTS
SUBSTITUTION
DONE?
HAHA, NOPE!
CAUCHY'S FORMULA
????
???!?
???
??
?
???
PARTIAL FRACTIONS
RIEMANN INTEGRATION
INSTALL MATHEMATICA
?
?
???
???
RISCH ALGORITHM
STOKES' THEOREM
?
?????
???
WHAT THE HECK IS A BESSEL FUNCTION??
PHONE CALLS TO MATHEMATICIANS
OH NO
BURN THE EVIDENCE

# Image Segmentation



# Markov Random Field



# Markov Random Field

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \sum_{i}^{N} p(\mathbf{y}|\mathbf{x}_i)p(\mathbf{x}_i)$$

- $\mathbf{x}_i$ is a specific binary images

AT&T
9:42 AM
SMS
Text
Tuesday
9
Calendar
Photos
Camera
YouTube
Stocks
Maps
73°
Weather
Clock
Calculator
Notes
Settings
iTunes
App Store
Phone
Mail
Safari
iPod
iPhone
8GB

# Number of terms i

2290593203500326442498254071102 8779924646158308390547680551234
5054431338510774037915738775865 8057318635099533562444284837656
6408900340661545734126916095393 4651531316272895970961099648619
5486636741656944283948869330648 4701733713508133208092688099524
0707971539803921050200955733579 4366205566676730638553849508752
9677470990968153918788613785751 3890052212385415364000233552517
9230941551480812783648467474496 1578781252261713953420063416790
7552057630497077601674681891226 1453204962575441115371836944715
6895505073882545721273943517481 6507334054019330445298798029650
8746618030728963410359112463410 9184832439049686890853942279882

# Number of terms ii

965540636137098078969750475941 6746133102362814600105499829189 2

885044803396603840787819652704 4715747436853386831577880020356 2

147412103415587157296801980525 1898240972502308488120023873650 0

202728357227524884496348873647 1394352603191284822724882619046 4

847696594892838239669305251912 4168772517553390869295245378359 8

283702354351658853691637104648 9422031070150882793338052642997 9

259981580192092290389815887171 2892609715338272913453162186531 3

978608581541705515982751534447 1332632503478183677651370310036 0

979388975857537790830350106677 6654831199960534747537034342674 3

825340005381099786418727660970 8209309038066394442278969691365 4

890020232228508254497953096787 0630443700983384921773149302167 4

# Number of terms iii

2550624871750833859476679189509 5680602732346712939153259990811

4893913032842065037601973054196 1524092173016464047938013691439

6671843203605981118777513627755 7250792266837423597968228683403

4089138475154767372727122932222 8878852083218796660305975797728

8778298768646815994259957325408 8749600987758158350339985951647

5121708697580746029473842801833 8592485796034133919973077413533

6869491956368516611377674237208 1780419191068702807890339161440

9912666138730775266005780452422 5302437317858452782485229505751

3761093944464722805553911771716 4315059230286413698788578331540

1782239495790781650110059887274 5959467831004471989549305375741

9073809906471822251882514747849 0657161167548497523333968812279

# Number of terms iv

4911475119965635459462447339289 7828672753085721621023943443062

0144907278084466853892944205719 8697060107876495003418069047901

8142025673307261276950347320181 6461274039931292984401423199725

4340930170763466037725337419662 9143599599348813527131013125346

3508530232037816302115328138866 8643014293963947674718567131663

5043595580465472543695170605663 2361702749907044372801683830358

6991365299464326205642839343150 4053504888101754720253838078891

9253939272110382634932825138554 3816977282386956487514065578882

3474751813846542682825520838131 0069117625217360239526199430454

3464350338428593031654513507976 7510717638042435127189839307791

2093765743451201386745554882022 4148073627378623609980111113076

# Number of terms v

0640189547044207203761774747082 0243516866198003957569584101060

8046613562965001201466456771415 5778664863093617634553900426210

9110167208910075825348801584001 7224071067971558665492397885347

6607256313817084019127947685341 8537351879721277733449450507730

3189505040470344922506903873556 9656865708529073446623478695245

6543122517479114466613670208736 0842313671545657762822696089905

6802168279902278674508669673834 7816102210900054189076993778672

7705964820658607375143364171301 1744511704016132334906338900377

1777472580944833242545989973822 5646744609738390155521757096422

2619375692340966923479020630115 9076383049447801135255878205328

2752643299087648267991015324907 4963538068771014944040060242262

# Number of terms vi

3804497742682401904233153226013 9373317250133351983527123955504
2292211010517136771541981666250 0131430427440349387764312765762
4870317305687566284108475166000 1324414350620739304183073837766
8972502903711649967733818943578 9237255328232566165426546313829
1135999395862937 6

# Numbers

- Possible black and white 3 Megapixel images

$$2^{3145728}$$

# Numbers

- Possible black and white 3 Megapixel images

$$2^{3145728}$$

- Number of atoms in the universe

$$10^{80} \approx (2^{\frac{10}{3}})^{80} \approx 2^{267}$$

# Numbers

- Possible black and white 3 Megapixel images

$$2^{3145728}$$

- Number of atoms in the universe

$$10^{80} \approx (2^{\frac{10}{3}})^{80} \approx 2^{267}$$

- Age of the universe in seconds

$$4.35 \cdot 10^{17} \approx 2^{59}$$

# Intractability

- Computational intractability: there are too many states to sum over (image segmentation)
- Analytic: no closed form exists for the distribution (unsupervised learning)

# Intractability

- Computational intractability: there are too many states to sum over (image segmentation)
- Analytic: no closed form exists for the distribution (unsupervised learning)
- The double annoyance: machine learning is not just ill-posed, the computations needed for making it well posed is intractable

# The No-Free Lunch Theorem

- There exists no universal learner

# The No-Free Lunch Theorem

- There exists no universal learner
- For every learner there exist a task on which it fails

# The No-Free Lunch Theorem

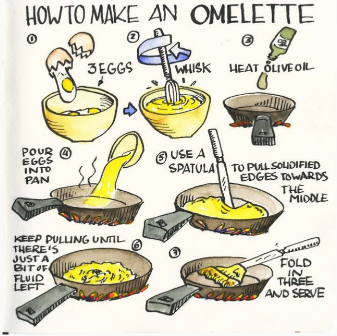
- There exists no universal learner
- For every learner there exist a task on which it fails
- Every algorithm that learns something useful does so by assumptions

# The No-Free Lunch Theorem

- There exists no universal learner
- For every learner there exist a task on which it fails
- Every algorithm that learns something useful does so by assumptions
- *There is no free lunch algorithm*

# Every Algorithm Does this

# Explicit vs. Tacit Knowledge

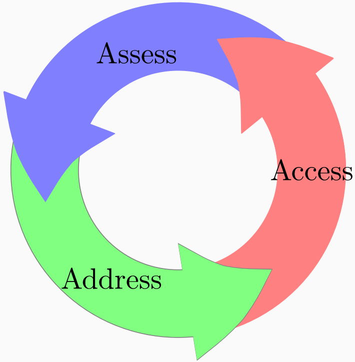


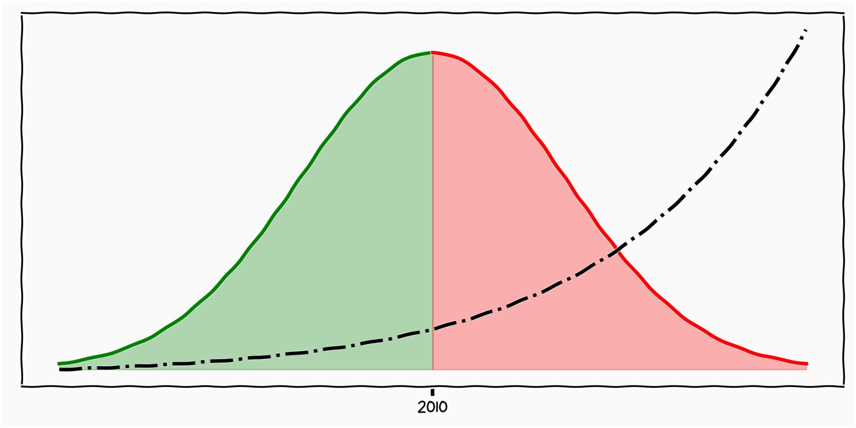
# Dangers of misattribution

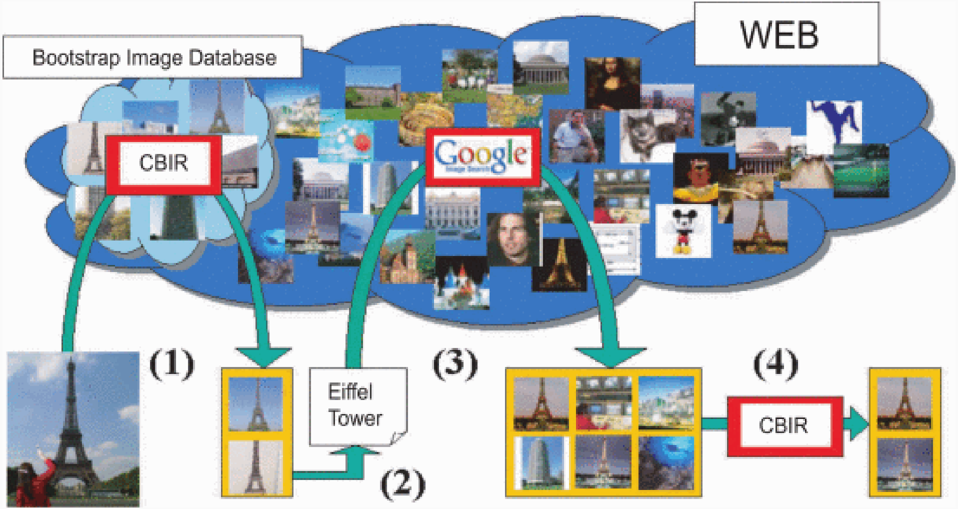
# Data centric thinking

*"You need to put Machine Learning in the context of data (and humans)"*

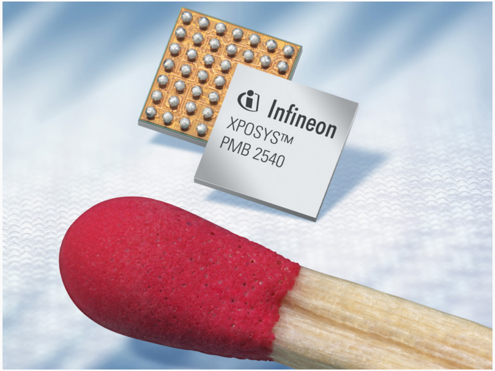
# The Datascience Loop




2010



Bootstrap Image Database
WEB
CBIR
Google
(1)
Eiffel Tower
(2)
(3)
(4)
CBIR


Infineon
XPOSYS™
PMB 2540

# Summary

# Summary

- Machine learning problems are inherently ill-posed

# Summary

- Machine learning problems are inherently ill-posed
- We need to introduce knowledge/assumptions

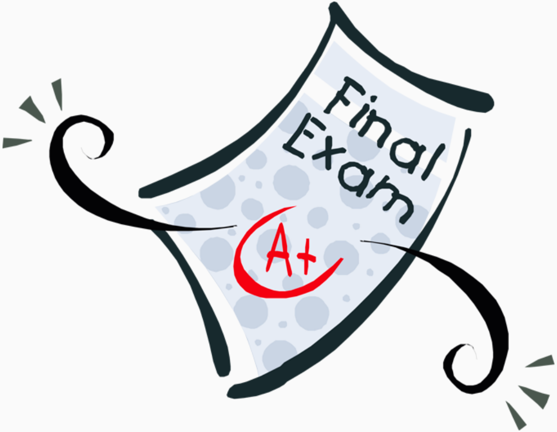
# Summary

- Machine learning problems are inherently ill-posed
- We need to introduce knowledge/assumptions
- The results can only be interpreted in light of the knowledge/assumptions

# Summary

- Machine learning problems are inherently ill-posed
- We need to introduce knowledge/assumptions
- The results can only be interpreted in light of the knowledge/assumptions
- Use methods that you can explain as you need to communicate with domain experts

# Final Submission 2



# How

*Alongside your implementation you will provide a short repository overview describing how you have implemented the different parts of the project and where you have placed those parts in your code repository. You will submit your code alongside a version of this notebook that will allow your examiner to understand and reconstruct the thinking behind your analysis.*

# What are we looking for?

*Remember the notebook you create should tell a story, any code that is not critical to that story can safely be placed into the associated analysis library and imported for use (structured as given in the Fynesse template)*

# What are we not looking for?

**Lack of narrative** why are you doing what you are doing?

# What are we not looking for?

**Lack of narrative** why are you doing what you are doing?

**Spaghetti code** encapsulate code, clean up code

# What are we not looking for?

**Lack of narrative** why are you doing what you are doing?

**Spaghetti code** encapsulate code, clean up code

**The perfect prediction** what does this even mean?

# What are we not looking for?

**Lack of narrative** why are you doing what you are doing?

**Spaghetti code** encapsulate code, clean up code

**The perfect prediction** what does this even mean?

**ML Ninjas** we will not give additional marks for "advanced methods" eof