The unreasonable effectiveness of mathematics, revisited

Big data and neuroscience

Jaime Gómez-Ramírez

Fundación Reina Sofia. Centre for Research in Neurodegenarative Diseases

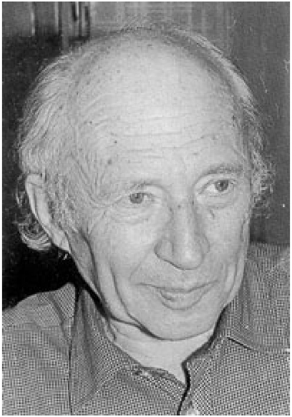
April 11 2018

Outline

The effectiveness of mathematics

Einstein: The most incomprehensible thing about the world is that is comprehensible

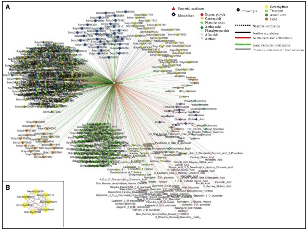
Wigner: The unreasonable effectiveness of mathematics



Gelfand: The Unreasonable Ineffectiveness of Mathematics in biology

heat loss in coffee

$$\frac{dQ}{dT} = As(T_{coffe} - T_{room})$$



The effectiveness of mathematics

- Wigner's 1960 essay *"the enormous usefulness of mathematics in natural science is something bordering on the mysterious"*

The effectiveness of mathematics

- Wigner's 1960 essay *"the enormous usefulness of mathematics in natural science is something bordering on the mysterious"*
- The typical interpretation of Wigner test is as follows:

  premise math concepts arise from aesthetic impulse in humans

The effectiveness of mathematics

- Wigner's 1960 essay *"the enormous usefulness of mathematics in natural science is something bordering on the mysterious"*
- The typical interpretation of Wigner test is as follows:

  premise math concepts arise from aesthetic impulse in humans

  premise is unreasonable to think that those same impulses are effective

The effectiveness of mathematics

- Wigner's 1960 essay *"the enormous usefulness of mathematics in natural science is something bordering on the mysterious"*
- The typical interpretation of Wigner test is as follows:

  premise math concepts arise from aesthetic impulse in humans

  premise is unreasonable to think that those same impulses are effective

observation nevertheless it so happens that they are effective

The effectiveness of mathematics

- Wigner's 1960 essay *"the enormous usefulness of mathematics in natural science is something bordering on the mysterious"*
- The typical interpretation of Wigner test is as follows:

  premise math concepts arise from aesthetic impulse in humans

  premise is unreasonable to think that those same impulses are effective

  observation nevertheless it so happens that they are effective

  consequence it follows that math concepts are unreasonably effective (assuming that the aesthetic premise as valid)

- Wigner's 1960 essay *"the enormous usefulness of mathematics in natural science is something bordering on the mysterious"*
- The typical interpretation of Wigner test is as follows:

  premise math concepts arise from aesthetic impulse in humans

  premise is unreasonable to think that those same impulses are effective

  observation nevertheless it so happens that they are effective

  consequence it follows that math concepts are unreasonably effective (assuming that the aesthetic premise as valid)

e.g imaginary numbers, tensor. Math concepts appear and propagate

- Wigner did seminal work on group theory applied to discover symmetry principles

The effectiveness of mathematics

- Wigner did seminal work on group theory applied to discover symmetry principles
- group theory replaced previous methods of analysis in quantum mechanics, group pest, finding invariants instead of seeking for explicit solution by calculus

The effectiveness of mathematics

- Wigner did seminal work on group theory applied to discover symmetry principles
- group theory replaced previous methods of analysis in quantum mechanics, group pest, finding invariants instead of seeking for explicit solution by calculus
- The goal of science is not to explain nature (the black box) but to explain the regularities in the behavior of the object *"Not the things in themselves but the* ***relationships*** *between the things. (Poincaré)*
- The search for causal explanation in terms of mathematical principles necessitates the belief of the mathematical structure of the universe the c-word

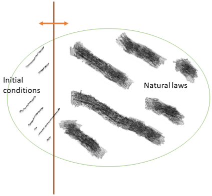
- We are "lucky" that regularities exist and that we can grasp them mathematically

The effectiveness of mathematics

- We are "lucky" that regularities exist and that we can grasp them mathematically
  - This is Newton's contribution and this is in essence why deep learning works

The effectiveness of mathematics

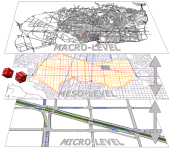
- We are "lucky" that regularities exist and that we can grasp them mathematically
  - This is Newton's contribution and this is in essence why deep learning works
- Regularities are invariant with respect to space and time. $A, B... \to X, Y....$ under T $T(A), T(B) \to T(X), T(Y)$
- Convolutional networks exploit image invariance to work (A *cat is a cat is a cat*)



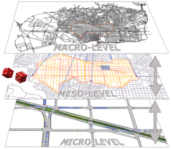
>>

- $t = \sqrt{\frac{2s}{g}}$
- What makes possible for us to disco
regularities is the division between initial conditions and regularities.
- Laws of nature are *IF initial conditions THEN event.*
- That's why causality is so hard, we need to include/exclude all possible combination of antecedents (initial conditions)

>>

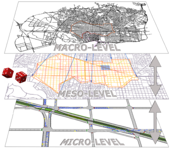


- Good doesn't play dice eg. stochastic brownian motion

>> 

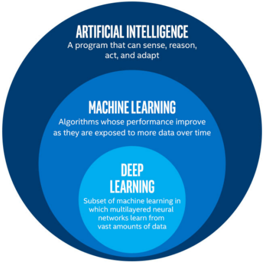
- Good doesn't play dice eg. stochastic brownian motion
- Our knowledge of nature contains 'a strange hierarchy' (Events we observed → Laws (regularities to discover) → Symmetry (invariance principles)

>>



- Good doesn't play dice eg. stochastic brownian motion
- Our knowledge of nature contains 'a strange hierarchy' (Events we observed $\rightarrow$ Laws (regularities to discover) $\rightarrow$ Symmetry (invariance principles)
- The future is always uncertain but nevertheless there are correlations - laws- that we can discover

AI, Machine Learning, Deep Learning



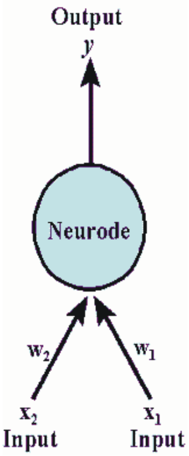
- AI
- Machine Learning
- ANN are non linear mapping systems whose functioning principles are vaguely based on the nervous systems of mammals
- Deep learning

>>

Data the most valuable asset and computation is a cheap commodity(information wants to be free)

Perceptron

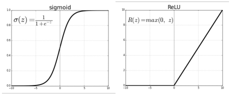
$$y = f(\sum_k w_x x_k) \qquad (1)$$



"A Logical Calculus of Ideas Immanent in Nervous Activity McCulloch, Pitts, 1943"
'If it doesn't rain ($x_1 w_1$) and homework done ($x_2 w_2$), go to the movies y (output)'

- neurons with a binary threshold activation function analogous to first order logic sentences
- By itself a neuron (or an ann) does very little but a sufficiently large network with appropriate structure and properly chosen weights can **approximate with arbitrary accuracy any function**

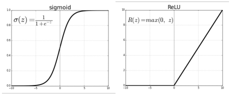
>>



- A perceptron is any feedforward network of nodes with responses like equation **??**.

$$y = f(\sum_k w_x x_k) = f(z) \tag{2}$$

Perceptron

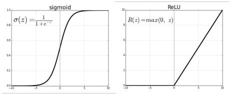


- A perceptron is any feedforward network of nodes with responses like equation **??**.

$$y = f(\sum_k w_x x_k) = f(z) \tag{2}$$

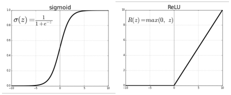
- In general, f is bounded nondecreasing nonlinear *squeezing* function function, eg. the sigmoid

$$f(z) = \frac{1}{1+e^{-z}}, f'(z) = \frac{e^{-}z}{(1+e^{-z})^2}$$



- Other choices are the tanh, step function and more recently the relu function .

$$y = ReLU(z) = max(0, z), y' = 1, z > 0$$

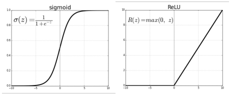


- Other choices are the tanh, step function and more recently the relu function .

$$y = ReLU(z) = max(0, z), y' = 1, z > 0$$

- ReLu works better, faster (gradient constant), y'(0) approximated $y = ln(1.0 + e^x)$

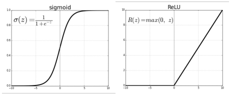
Perceptron



- Other choices are the tanh, step function and more recently the relu function .

$$y = ReLU(z) = max(0, z), y' = 1, z > 0$$

- ReLu works better, faster (gradient constant), y'(0) approximated $y = ln(1.0 + e^x)$
- Reduced likelihood of gradient to vanish



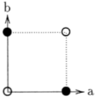
- Other choices are the tanh, step function and more recently the relu function .

$$y = ReLU(z) = max(0, z), y' = 1, z > 0$$

- ReLu works better, faster (gradient constant), y'(0) approximated $y = ln(1.0 + e^x)$
- Reduced likelihood of gradient to vanish
- Sparsity produced when $z \leq 0$, sigmoids on the other hand tend to represent more dense representations

What can and can't perceptrons do?

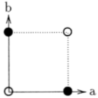
a	b	XOR(a,b)
0	0	0
0	1	1
1	0	1
1	1	0



- (Single-layer) perceptrons can correctly classify only data sets that are linearly separable (they can be separated by a hyperplane)

What can and can't perceptrons do?

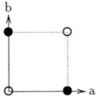
a	b	XOR(a,b)
0	0	0
0	1	1
1	0	1
1	1	0



- (Single-layer) perceptrons can correctly classify only data sets that are linearly separable (they can be separated by a hyperplane)
- The XOR function is famously non linearly separable and this is very important because many classification problems are not linearly separable.

What can and can't perceptrons do?

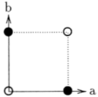
a	b	XOR(a,b)
0	0	0
0	1	1
1	0	1
1	1	0



- There are 2^{2^d} boolean functions of d boolean input variables and only $O(2^{2^2})$ are linearly separable.
    - For d=2 14/16 are linearly separable (XOR and its complement are the exceptions), but for $d = 4$, only 1882/65536 are linearly separable.

What can and can't perceptrons do?

a	b	XOR(a,b)
0	0	0
0	1	1
1	0	1
1	1	0



- There are 2^{2^d} boolean functions of d boolean input variables and only $O(2^{2^2})$ are linearly separable.
  - For d=2 14/16 are linearly separable (XOR and its complement are the exceptions), but for $d = 4$, only 1882/65536 are linearly separable.
- Although at that time it was known that multilayer networks were more powerful than single layer ones, the learning algorithms for multilayer architectures were not known

Deep networks

- ANN learn by example and use backpropagation

Deep networks

- ANN learn by example and use backpropagation
- If data are well-behaved it will learn not only the training examples but also the underlying relationships

Deep networks

- ANN learn by example and use backpropagation
- If data are well-behaved it will learn not only the training examples but also the underlying relationships
- ANN are adaptive and self-repairing, also has some fault tolerance due to its redundant parallel structure (dense connectivity makes it resilient to minor damage, graceful degradation)

Deep networks

- ANN learn by example and use backpropagation
- If data are well-behaved it will learn not only the training examples but also the underlying relationships
- ANN are adaptive and self-repairing, also has some fault tolerance due to its redundant parallel structure (dense connectivity makes it resilient to minor damage, graceful degradation)
- Units within a layer are independent so they can be be evaluated simultaneously eg. network with 2,000 nodes in two layers will produce a response in 2 time steps rather than in 2,000 steps if each neuron required to be processes serially (dependent)

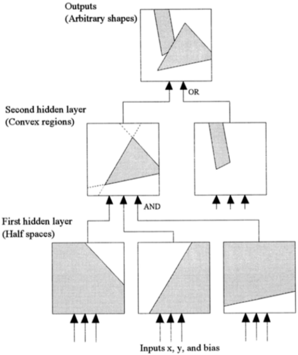
Deep networks

- ANN learn by example and use backpropagation
- If data are well-behaved it will learn not only the training examples but also the underlying relationships
- ANN are adaptive and self-repairing, also has some fault tolerance due to its redundant parallel structure (dense connectivity makes it resilient to minor damage, graceful degradation)
- Units within a layer are independent so they can be be evaluated simultaneously eg. network with 2,000 nodes in two layers will produce a response in 2 time steps rather than in 2,000 steps if each neuron required to be processes serially (dependent)
- Until the advent of GPUs this advantage were not fully exploited by computers

Table: ANN versus real nervous system

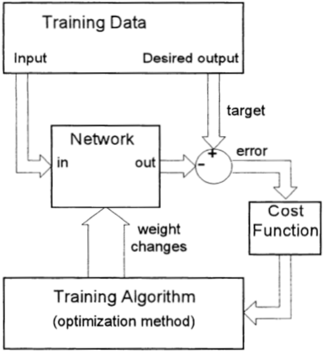
MLP	Nervous System
feedforward	recurrent
dense(fullyconnected)	sparse(local)
$O(10^{2,3,4})$	$O(10^{10}), O(10^{15})$
static	dynamic:spike trains, synchronization, fatigue

A frame



Outputs
(Arbitrary shapes)
OR
Second hidden layer
(Convex regions)
AND
First hidden layer
(Half spaces)
Inputs x, y, and bias

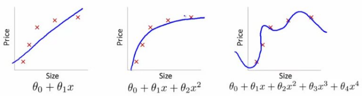
>>



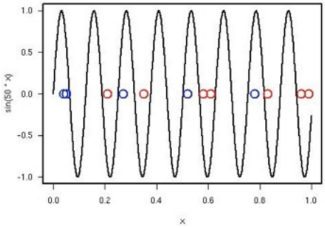
Training Data
Input
Desired output
target
Network
in
out
error
Cost
Function
weight
changes
Training Algorithm
(optimization method)

>>

Why MLP is better than one layer?

>> 

$y = mx$ is a system with one parameter, m, what kind of datasets can separate? only the linearly separable ones

>> 

$y = sin(kx)$, also has one parameter, the frequency k, but can separate any arbitrary distribution of points in the x-axis

Universality of MLP

- Any bounded function can be approximated with arbitrary accuracy if enough hidden units are available -multilayer perceptrons are universal approximators

Universality of MLP

- Any bounded function can be approximated with arbitrary accuracy if enough hidden units are available -multilayer perceptrons are universal approximators
- How many layers do we need for this astounding property (universal approximators)? Kolmogorov showed that one hidden layer is sufficient

Universality of MLP

- Any bounded function can be approximated with arbitrary accuracy if enough hidden units are available -multilayer perceptrons are universal approximators
- How many layers do we need for this astounding property (universal approximators)? Kolmogorov showed that one hidden layer is sufficient
    - Any continuous function with n variables to a m-dimensional output can be implemented by a network with one hiddden layer

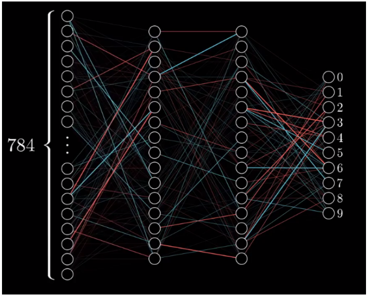
Universality of MLP

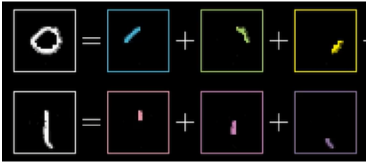
- Any bounded function can be approximated with arbitrary accuracy if enough hidden units are available -multilayer perceptrons are universal approximators
- How many layers do we need for this astounding property (universal approximators)? Kolmogorov showed that one hidden layer is sufficient
    - Any continuous function with n variables to a m-dimensional output can be implemented by a network with one hiddden layer
- Unfortunately the proof is not constructive, that is, it does not tell how the weights should be chosen to produce such a function

How important is the universality of MLP?

- Is it universal approximation a rare property? Not really, many other systems such as polynomials, trigonometric polynomials (eg Fourier series), wavelets, kernel regression systems (svm) have also universal properties

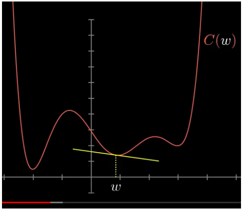
Architecture

>> 

>> 

First layer detects the edges, and the second has the abstract concept of loop and straight lines, this is actually the hope of having a layer structure and it works because what Wigner already said

Gradient descent



Cost $C(w)$, the gradient $\frac{dC(w)}{dw} = 0$ (huge column vector with $784 + 16 * 16 + 16 * 10 + 16 + 16 + 10$ dimensions).
The negative of the gradient which is the direction of the steepest increase gives the direction to take to decrease the error(cost) more quickly

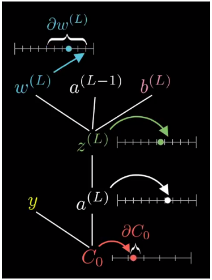
Backprop

The method to calculate the gradient vector, which tells you which direction to take and how step the step is

1. compute ∇C
2. take step in $-\nabla C$ direction
3. repeat

Learning is finding the minimizing the weight function.
Backprop is the algo used in gradient descent.
Learning is 'just' finding the right weights and biases.

Backprop in action, chain rule

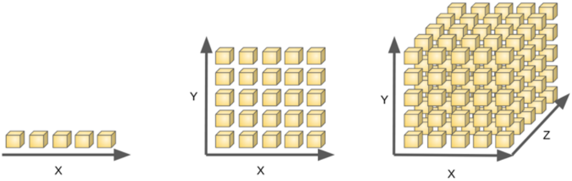


The cost of one training example is $C_0 = (a^L - y)^2$, the last activation is $a^L = \sigma(w_L a^{L-1} + b^L) = \sigma(z^L)$

How sensitive is the Cost function to small changes in the weight?

- $\frac{\partial C_0}{\partial w^L} = \frac{\partial z^L}{\partial w^L}\frac{\partial a^L}{\partial z^L}\frac{\partial C_0}{\partial a^L}$
- $\frac{\partial C_0}{\partial a^L} = 2(a^L - y)$, $\frac{\partial a^L}{\partial z^L} = \sigma'(z^L)$, $\frac{\partial z^L}{\partial w^L} = a^{L-1}$
- Average over all training examples $\frac{\partial C}{\partial w^L} = \frac{1}{n}\sum_{k=0}^{n-1}\frac{\partial C_k}{\partial w^L}$
- $\nabla C = [\frac{\partial C}{\partial w^1}, \frac{\partial C}{\partial b^1}, \dots, \frac{\partial C}{\partial w^L}]$

Curse of dimensionality



- Curse of dimensionality refers to the apparent intractability of systematically searching through a high-dimensional space
- As n get bigger it gets harder and harder to sample all the boxes, with n dimensions each allowing for m states, we will have m^n possible combinations

Blessing of dimensionality

- In MLP approximation error decreases with the number of training samples *error* $O(1/sqrtN)$ and also with the number of hidden units *error* $O(1/M)$ and unlike other systems, eg polynomials this is independent of the input size and avoid the curse of dimensionality problem.
- From these results we can build bounds, for example

$$N > O(Mp/\epsilon) \tag{3}$$

  where N is the number of samples, M the hidden nodes, p input dimension (Mp number of parameters) and ϵ the desired approximation error.
- More layers is better and do not harm

Bias variance trade off

Bias–variance tradeoff is the problem of simultaneously minimizing two sources of error in the estimand. The bias-variance decomposition:

$$MSE = E((\hat{\theta}-\theta)^2) = E(\hat{\theta}-\theta)^2 + Var(\hat{\theta}) + = (Bias(\theta))^2 + Var(\theta) \quad (4)$$

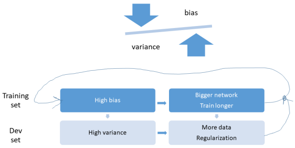
The bias/variance trade off in deep learning is not exactly a trade off it can be tackle algorithmically

Table: Bias variance

high var	**high bias**	**high bias and var**	**low bias and var**
2%	15%	15%	0.5%
11%	15%	30%	1%

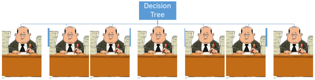
you don't have the dialectical tension one thing or the other but in the table we have 4 cases rather than a trade off and luckily we can take action that fit every case.

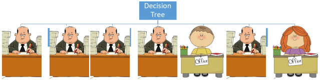
Bias variance trade off



- A bigger network will improve your fitting without hurting the variance problem, with the caveat that you regularize properly.
- Before we couldn't make better one without hurting the other, now we can get both better.

Ensemble models



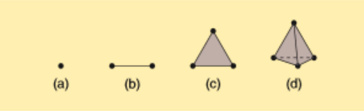


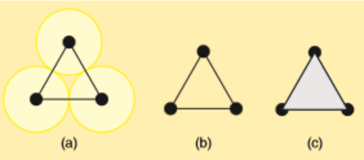
- Idea: you don't want an organization with all the same('good') you may want to introduce variability
- decision trees are grown by introducing a random element, eg at each node choose randomly the features to split the node
- Random forest (randomly constructed trees), each voting for a class, Bagging: boosting + aggregation.
- Great predictors but interpretability is obscured by the complexity of the model -accuracy generally requires more complex prediction methods-

>> 

topology is concerned with the properties of space that are preserved under continuous deformations: stretching, crumpling and bending, but not tearing or gluing

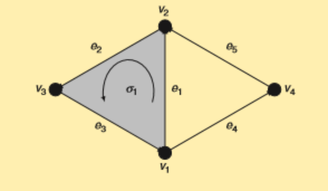
- Topology is an intermediate analysis medium that focuses on coarse structures.
- Why to use topology over Big data?
  - It studies the invariants of continuous formations of the shape of data -resistant to threshold selection problem-
  - It allows measures of shape (clumps, holes and voids) which are invariant across scales

>> 

>> 

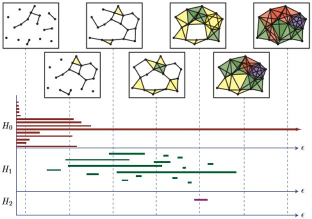
- Edges in a graph capture dyadic relationships.
- Graphs can't capture high order relationships but simplicial complex can
- A simplicial complex is a generalized graph consisting on vertices, edges, triangles and simplices of higher dimension glue together.

Persistent homology

>> 

- $C_0(X) =< v_1, v_2, v_3, v_4 >$, $C_1(X) =< e_1, e_2, e_3, e_4, e_5 >$, $C_2(X) = \sigma_1$

- boundary operator $\rho : C_1(X) \to C_0(X)$, $\rho_2 : C_2(X) \to C_1(X)$ when applied to an edge it yields a difference of vertices, higher order operator to act on triangles (2-simplices),
- Loop is when we have



>>

- $e_1 + e_2 + e_3$ is obtained as the image of
triangle σ_1 under the map ρ_2, whereas i
the image of a triangle, in other words,
$Im(\rho_2) = \{y \in C_1, \exists x \in C_2(X), \rho_2(x) =$
then $e_1 + e_2 + e_3 \in Im(\rho_2)$ and
$e_1 + e_5 + e_4 \notin Im(\rho_2)$.
- The 1-D homology is the quotient space
$H_1(X) = [Ker(\rho_1)/Im(\rho_2)]$

$$H_i(X) = \frac{Ker(\rho_i)}{Im(\rho_{i+1}}$$

Exploring the alpha desynchronization hypothesis in resting state networks with intracranial electroencephalography and wiring cost estimates

Jaime Gómez-Ramírez, Shelagh Freedman, Diego Mateos, José Luis Pérez Velázquez & Taufik A. Valiante

>> Scientific Reports 7, Article number: 15670 | Received: 27 March 2017

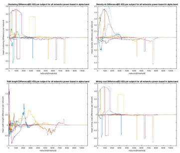
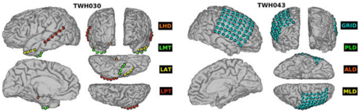


Figure 1



>>

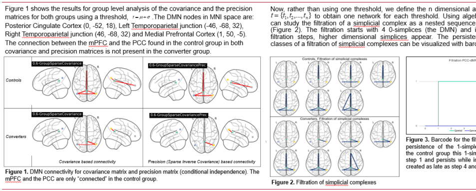


Figure 1 shows the results for group level analysis of the covariance and the precision matrices for both groups using a threshold, $t = \mu + \sigma$. The DMN nodes in MNI space are: Posterior Cingulate Cortex (0, -52, 18), Left Temporoparietal junction (-46, -68, 32), Right Temporoparietal junction (46, -68, 32) and Medial Prefrontal Cortex (1, 50, -5). The connection between the mPFC and the PCC found in the control group in both covariance and precision matrices is not present in the converter group.

Figure 1. DMN connectivity for covariance matrix and precision matrix (conditional independence). The mPFC and the PCC are only "connected" in the control group.

Now, rather than using one threshold, we define the n dimensional a $t = \{t_1, t_2, \ldots, t_n\}$ to obtain one network for each threshold. Using algel can study the filtration of a simplicial complex as a nested sequence (Figure 2). The filtration starts with 4 0-simplices (the DMN) and i filtration steps, higher dimensional simplices appear. The persiste classes of a filtration of simplicial complexes can be visualized with barc

Figure 2. Filtration of simplicial complexes

Figure 3. Barcode for the fi persistence of the 1-simple the control group this 1-sin step 1 and persists while i created as late as step 4 and

Conclusions

- With enough imagination a classifier(regression) can be useful to solve a large a number of problems
- Deep learning works because there is structure in the world but we don't know why because we don't know anything about the initial conditions

> laws of nature are precise beyond anything reasonable; we know virtually nothing about the initial conditions (Wigner)

- There are other ways to reduce complexity in big data while preserving maximal intrinsic information -computational topology
- Occam's dilemma (*lex parsimoniae*): accuracy generally requires more complex prediction methods, simple and interpretable functions do not make the most accurate predictions
- The curse of dimensionality can be a blessing

Thanks!