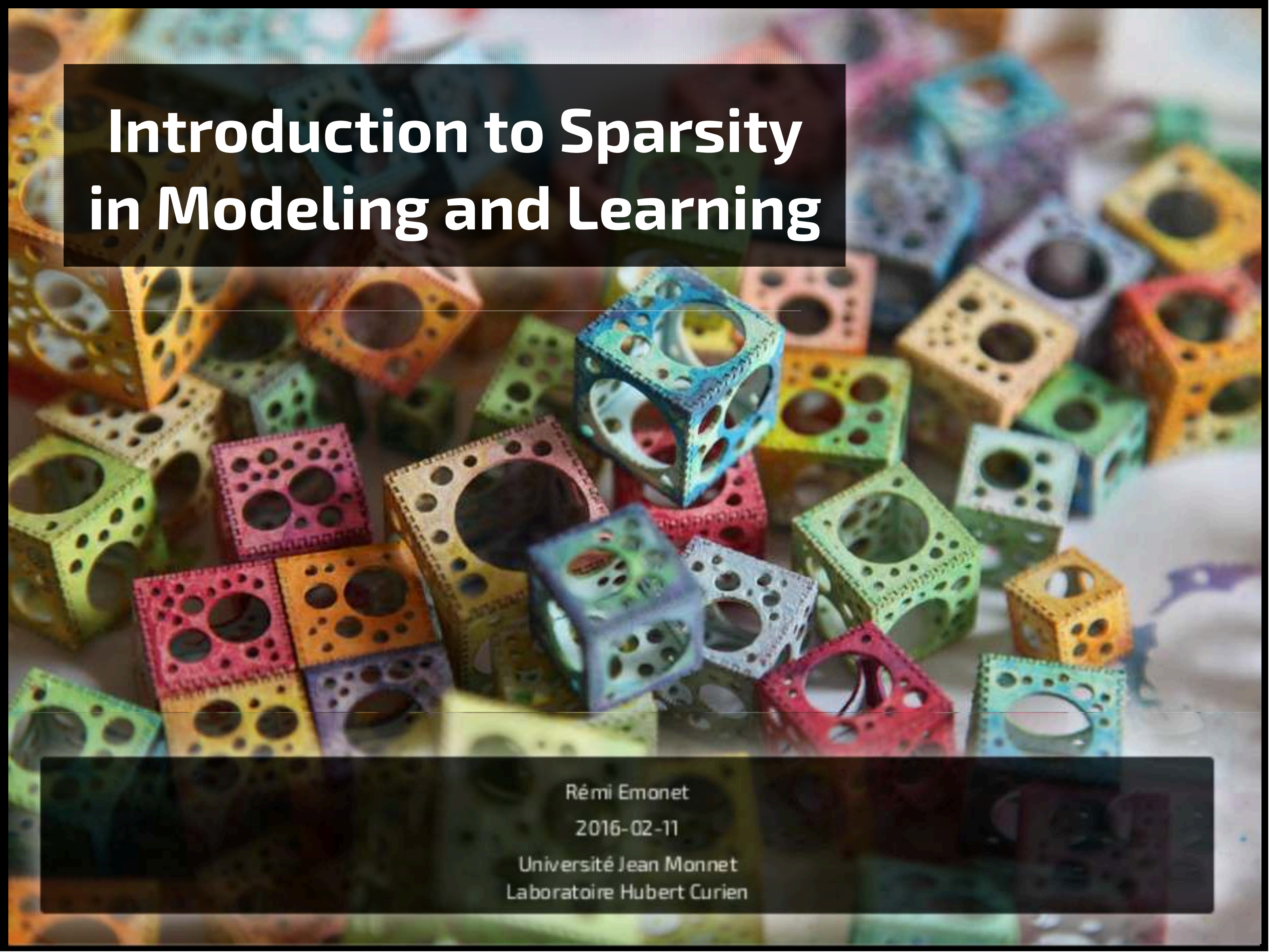


Introduction to Sparsity in Modeling and Learning

Rémi Emonet

2016-02-11

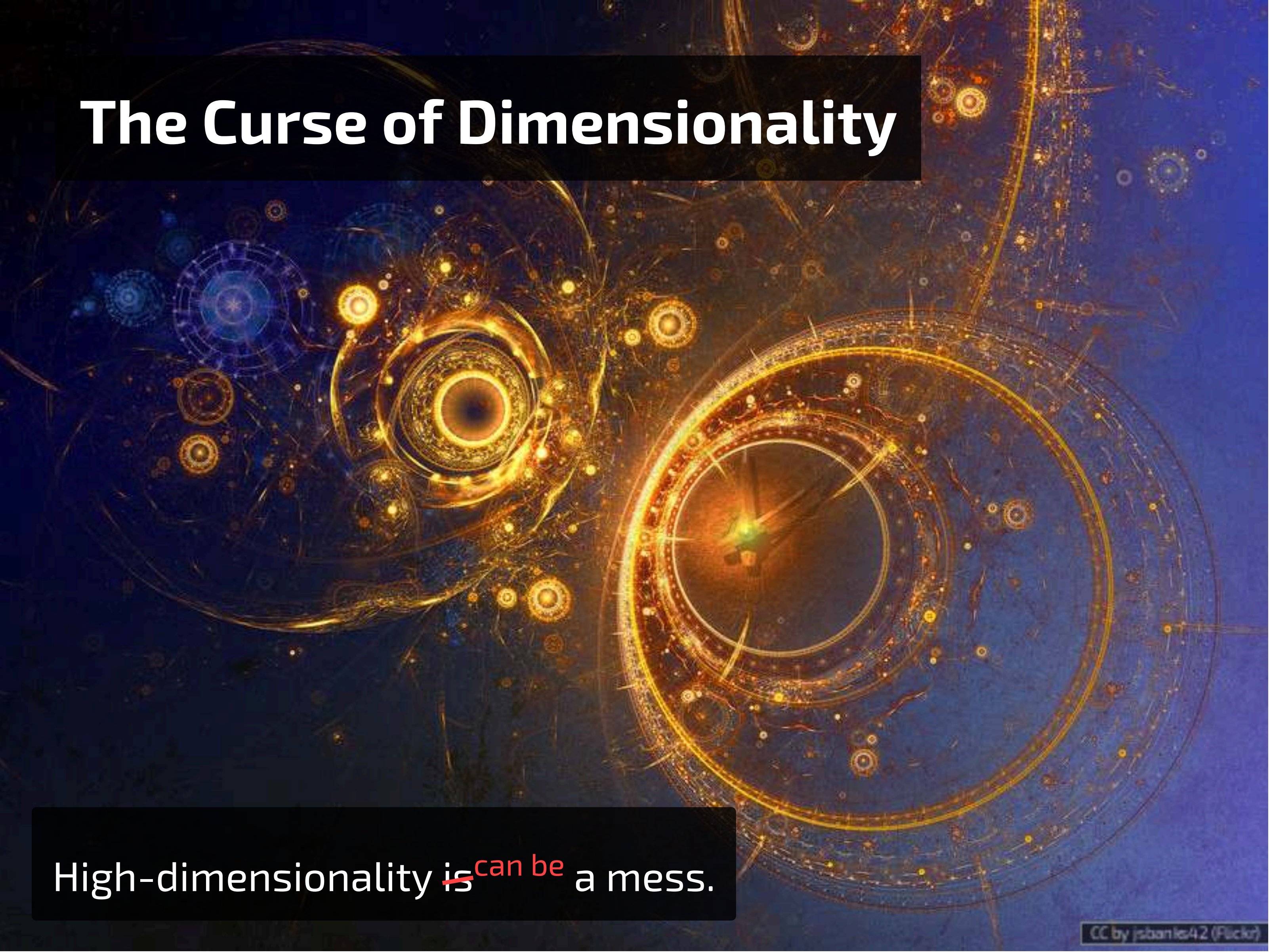
Université Jean Monnet
Laboratoire Hubert Curien

Introduction to Sparsity in Modeling and Learning

- The Curse of Dimensionality
- Ockham's Razor
- Notions of Simplicity
- Conclusion

The Curse of Dimensionality

The Curse of Dimensionality



High-dimensionality ~~is~~ can be a mess.

CC by jsbanks42 (Flickr)

What is this Curse Anyway?

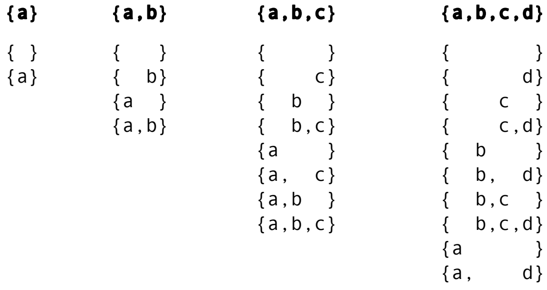
- Some definition:

  > *Various phenomena that arise*
  > *when analyzing and organizing data*
  > *in high-dimensional spaces.*

- Term coined by Richard E. Bellman
  - 1920 – 1984
  - dynamic programming
  - differential equations
  - shortest path
- What is (not) the cause?
  - not an intrinsic property of the data
  - depends on the representation
  - depends on how data is analyzed

Combinatorial Explosion

- Suppose
  - you have d entities
  - each can be in 2 states
- Then
  - there are 2^d combinations to consider/test/evaluate
- Happens when considering
  - all possible subsets of a set (2^d)
  - all permutations of a list $(d!)$
  - all affectations of entities to labels (k^d, with k labels)

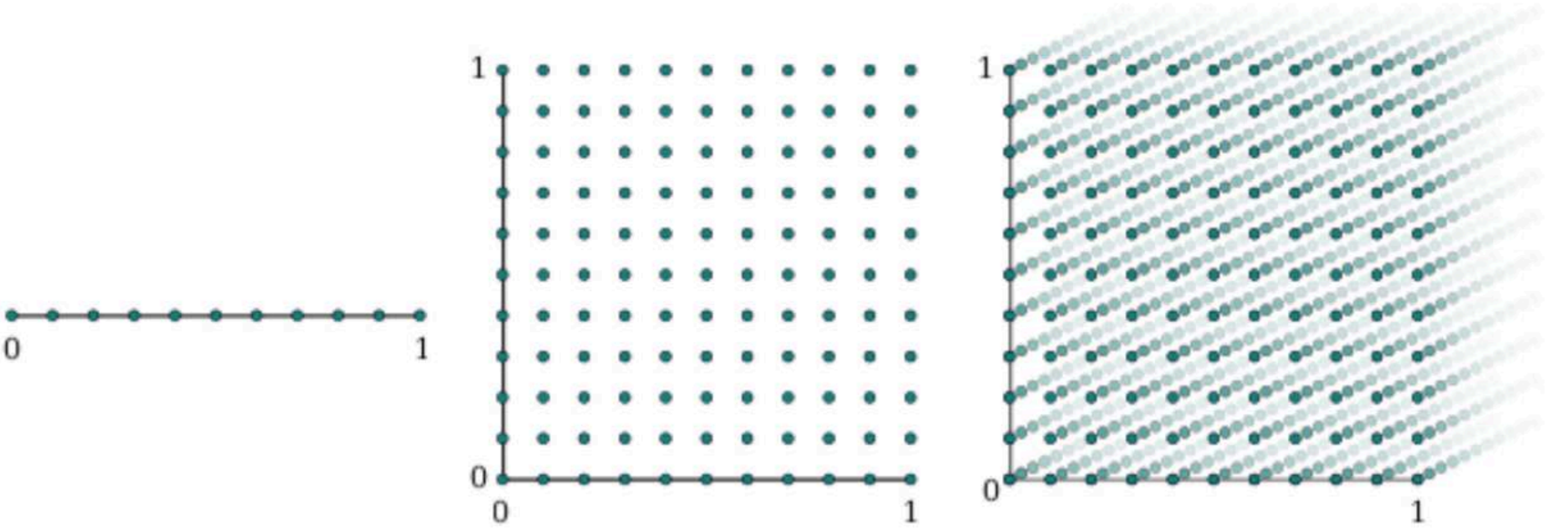


```
{a}       {a,b}        {a,b,c}          {a,b,c,d}

{ }       {   }        {     }          {       }
{a}       {  b}        {    c}          {      d}
          {a  }        {  b  }          {    c  }
          {a,b}        {  b,c}          {    c,d}
                       {a    }          {  b    }
                       {a,  c}          {  b,  d}
                       {a,b  }          {  b,c  }
                       {a,b,c}          {  b,c,d}
                                        {a      }
                                        {a,    d}
```

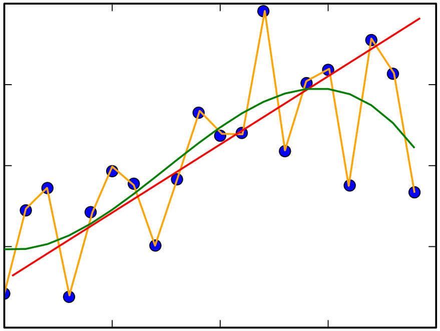
Regular Space Coverage

- Analogous to combinatorial explosion, in continuous spaces
- Happens when considering
    - histograms
    - density estimation
    - anomaly detection
    - ...



In Modeling and Learning

- The world is complicated
  - state with a huge number of variables (dimensions)
  - possibly noisy observations
  - e.g. a 1M-pixel image has 3 million dimensions



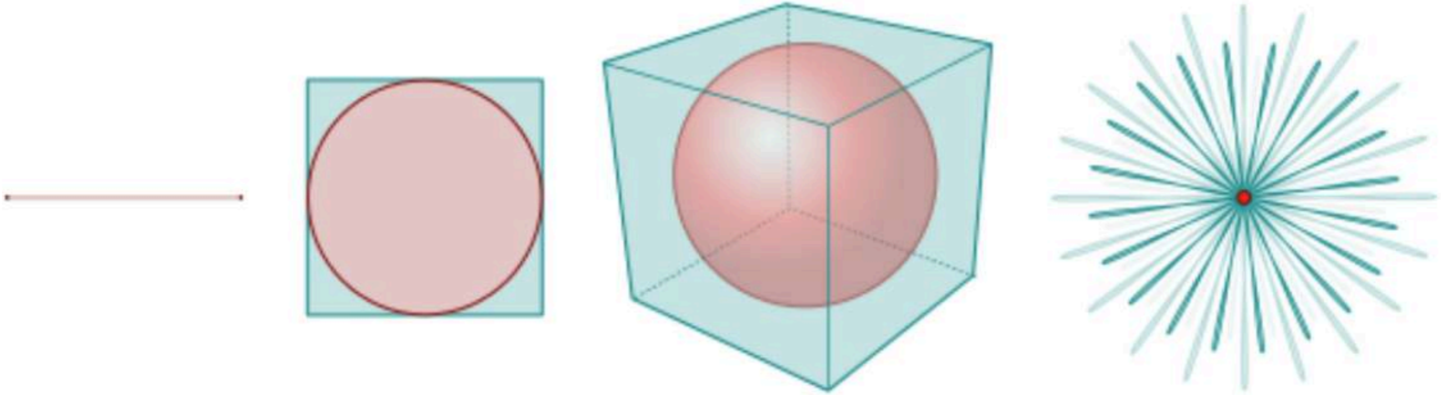
- Learning would need observations for each state
  - it would require too many examples
  - need for an “interpolation” procedure, to avoid overfitting
- Hughes phenomenon, 1968 paper (which is wrong, it seems)

  given a (small) number of training samples, additional feature measurements may reduce the performance of a statistical classifier

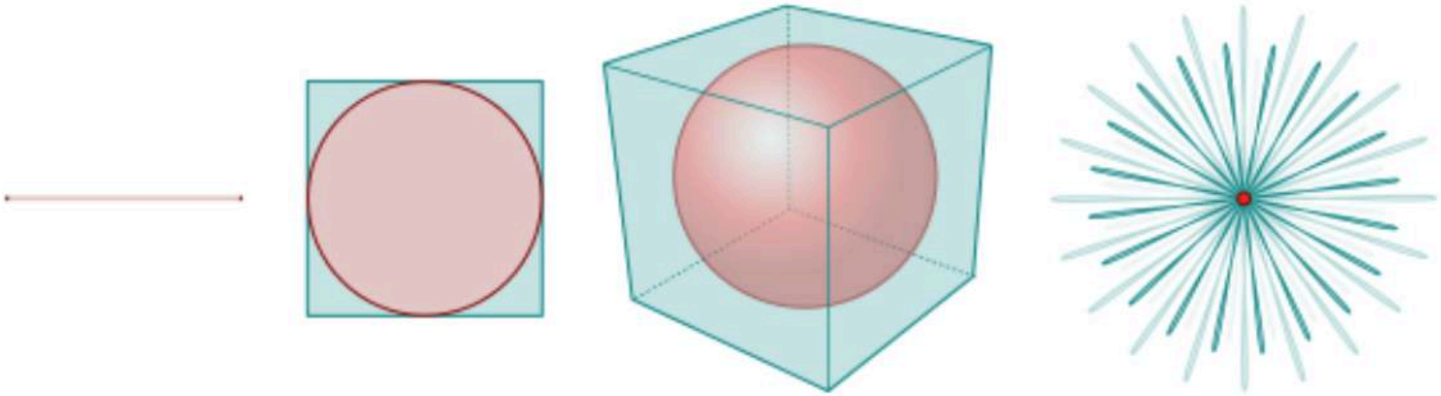
A Focus on Distances/Volumes

- Considering a d dimensional space
- About volumes
  - volume of the cube: $C_d(r) = (2r)^d$
  - volume of a sphere with radius r: $S_d(r) = \dfrac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} r^d$

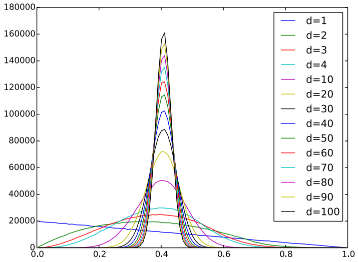
    (Γ is the continuous generalization of the factorial)
  - ratio: $\dfrac{S_d(r)}{C_d(r)} \rightarrow 0$ (linked to space coverage)



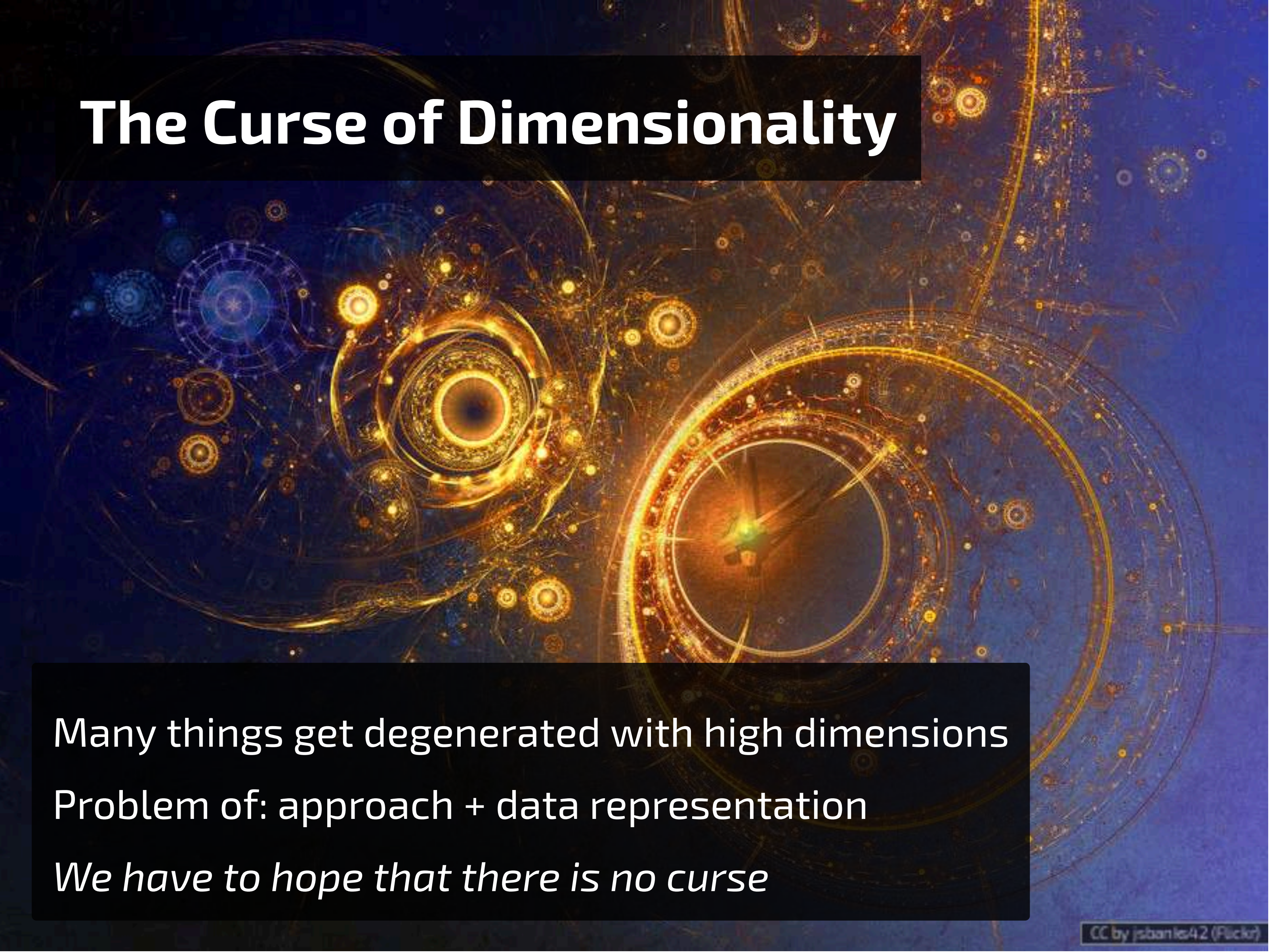
A Focus on Distances/Volumes (cont'd)



- About distances
  - average (euclidean) distance between two random points?
  - everything becomes almost **as** “far”
- Happens when considering
  - radial distributions (multivariate normal, etc)
  - k-nearest neighbors (hubiness problem)
  - other distance-based algorithms



The Curse of Dimensionality



Many things get degenerated with high dimensions

Problem of: approach + data representation

We have to hope that there is no curse

CC by jsbanks42 (Flickr)

- The Curse of Dimensionality
- Ockham's Razor
- Notions of Simplicity
- Conclusion

UNIVERSITÉ JEAN MONNET SAINT-ÉTIENNE

LABORATOIRE HUBERT CURIEN UMR • CNRS • 5516 • SAINT-ETIENNE

Ockham's Razor

Shave unnecessary assumptions.

CC by Fuxcinator (Flickr)

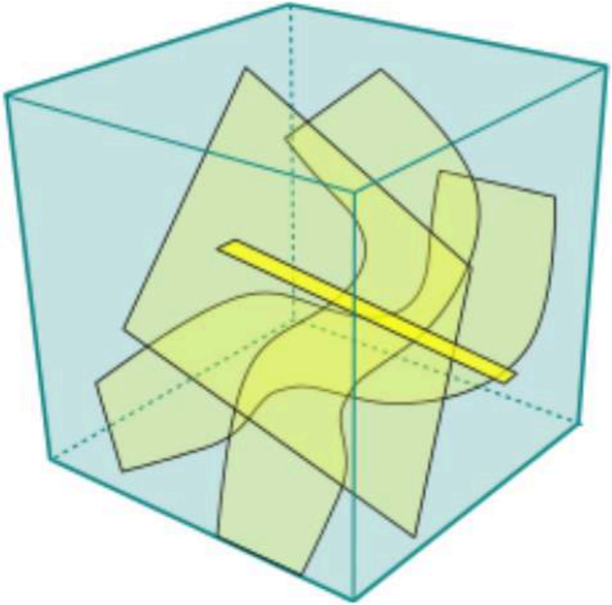
Ockham's Razor

- Term from 1852, in reference to Ockham (XIVth)
- *lex parsimoniae*, law of parsimony
- *Prefer the simplest hypothesis that fits the data.*
- Formulations by Ockham, but also earlier and later
- More a concept than a rule
    - simplicity
    - parsimony
    - elegance
    - shortness of explanation
    - shortness of program (Kolmogorov complexity)
    - falsifiability (sciencific method)
- According to Jürgen Schmidhuber, *the appropriate mathematical theory of Occam's razor already exists, namely, Solomonoff's theory of optimal inductive inference.*

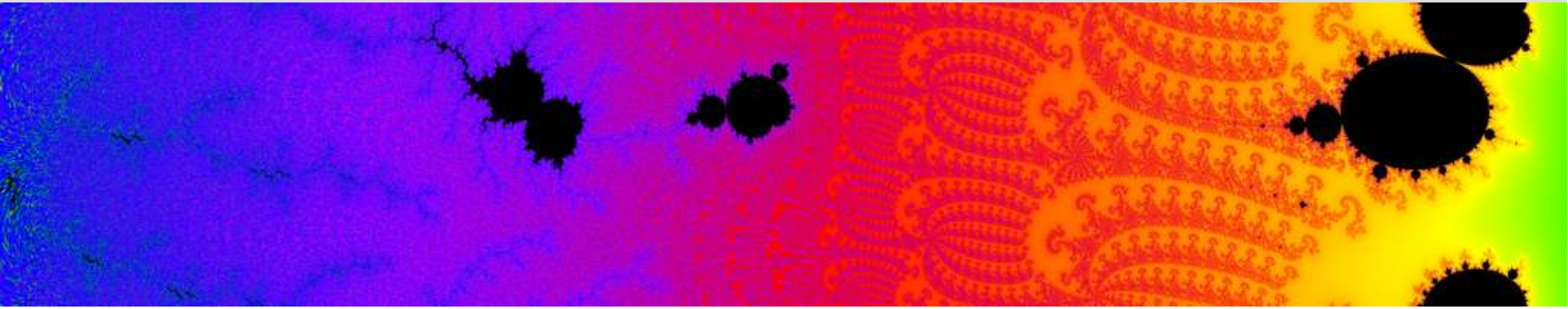
Notions of Simplicity

Simplicity of Data: subspaces

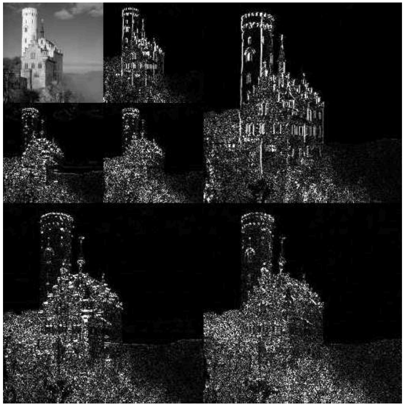
- Data might be high-dimensional, but we have hope
  - that there is a organization or regularity in the high-dimensionality
  - that we can guess it
  - or, that we can learn/find it
- Approaches: dimensionality reduction, manifold learning
  - PCA, kPCA, *PCA, SOM, Isomap, GPLVM, LLE, NMF, …



Simplicity of Data: compressibility



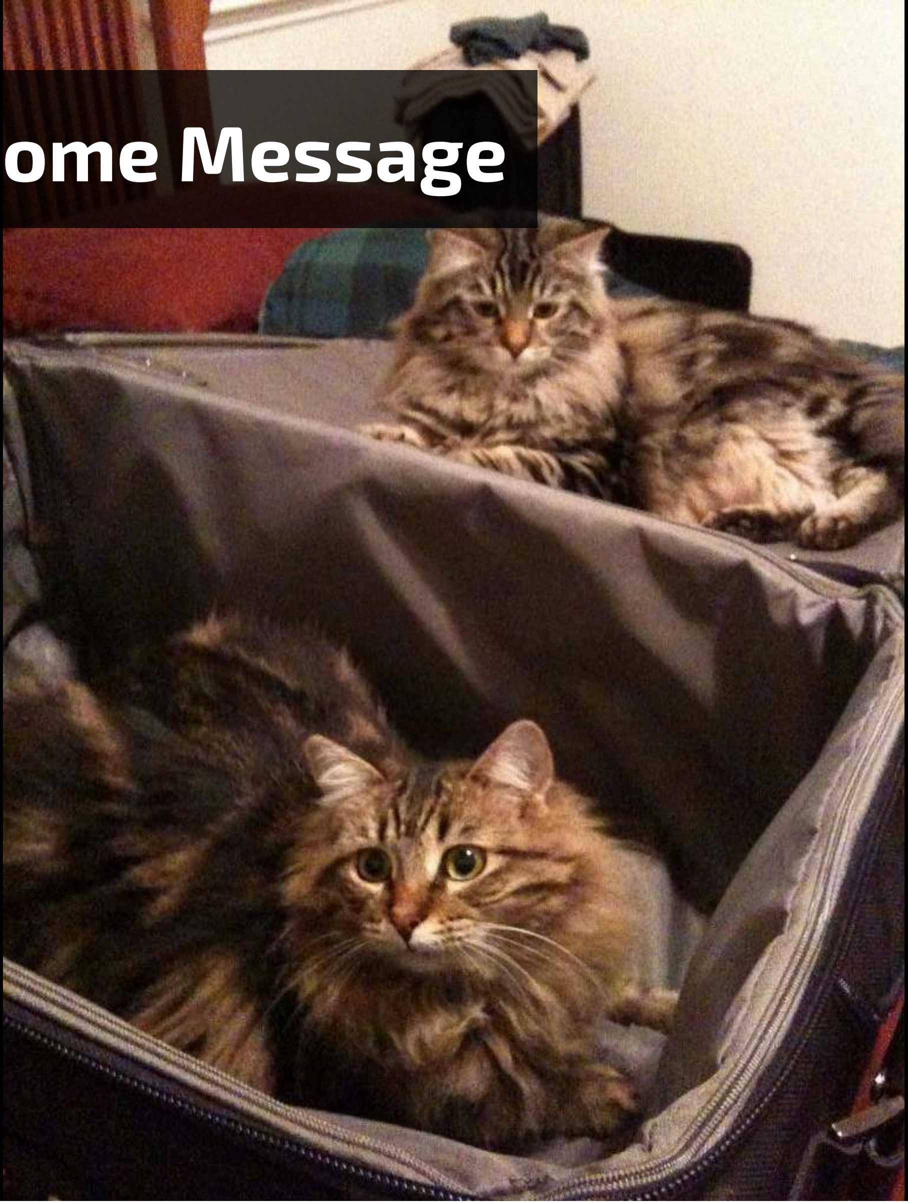
- Idea
  - data can be high dimensional but compressible
  - i.e., there exist a compact representation
- Program that generates the data (Kolmogorov complexity)
- Sparse representations
  - wavelets (jpeg), fourier transform
  - sparse coding, representation learning
- Minimum description length
  - size of the “code” + size of the encoded data



Simplicity of Models: information criteria

- Used to select a model
- Penalizes by the number k of *free parameters*
  - AIC (Aikake Information Criterion)
    - penalizes the Negative-Log-Likelihood by k
  - BIC (Bayesian IC)
    - penalizes the NLL by $k \log(n)$ (for n observations)
  - BPIC (Bayesian Predictive IC)
  - DIC (Deviance IC)
  - FIC (Focused IC)
  - Hannan-Quinn IC
  - TIC (Takeuchi IC)
- Sparsity of the parameter vector ($l0$ norm)
  - penalizes the number of non-zero parameters

Take-home Message



CC by someToast (Flickr)

Thank You!

Questions?

CC by tech na logic (Flickr)