# Machine Learning for Economists: Part 1 – Curse of Dimensionality

Michal Andrle
International Monetary Fund

Washington, D.C.,
October, 2018

# Disclaimer #1:

**The views expressed herein are those of the authors and should not be attributed to the International Monetary Fund, its Executive Board, or its management.**

# CURSE OF DIMENSIONALITY

# Curse of Dimensionality (1)

As number of dimensions in the problem increases, things get less intuitive...

1. **Overfitting issues**
   With enough dimensions, almost everybody is an outlier...
   `Prob(you=female,you=Greek,you=play harp, you=IMF econ) = ?`

2. **Computational issues**

**Curse of dimensionality** can make the **BIG DATA** often quite **SMALL**, as the effective no. of data points for some cases is small

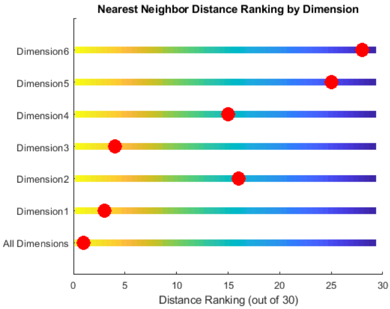
A few things are common, most things are rare (language, movie ratings, ...)

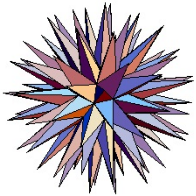
# Curse of Dimensionality (1b)

k-Nearest Neighbor modeling is flexible and can work really well in low-dimensional problems. . .

It can break down in high dimensions.

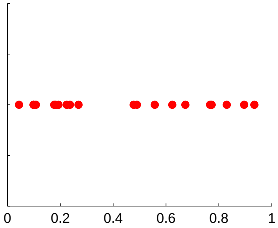
Your nearest neighbor can be on the opposite side of spectrum along some dimensions. . .

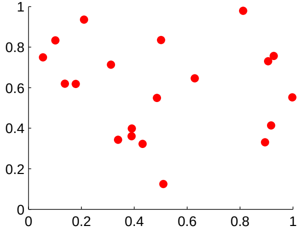


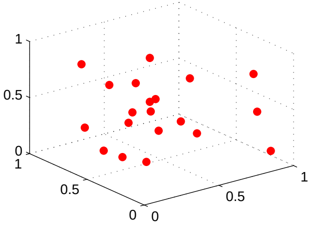



# Curve of Dimensionality (2)

If $N_1 = 20$ is **dense** for $d = 1$, you need $N_2 = 400$, $N_3 = 8000$, …



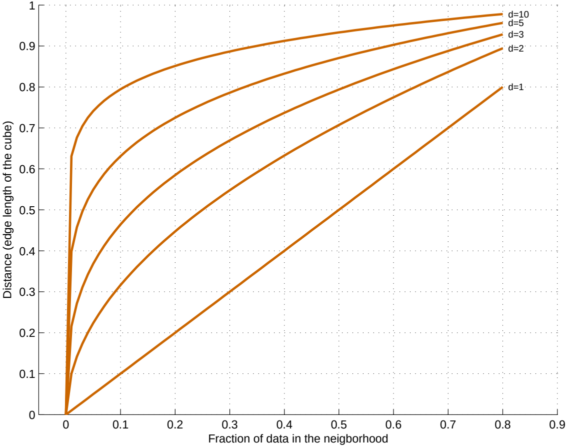



# Curse of Dimensionality (3a)

Searching for a **nearest neighbor** in uniformly dist. *d*-dim unit hypercube?

With 10 dimensions, to find 10% of nearest neighbors, you must "travel" through 80% of the cube's edges... Not very **local**, is it?



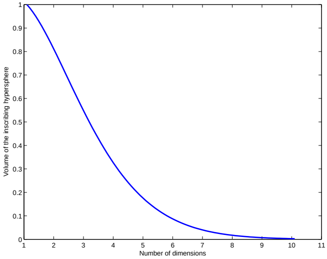

Follows Hastie et al. 2009

# Curse of Dimensionality (3b)

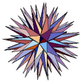
Our **intuition** betrays us tremendously in high-dimensions!

For a high-dim unit-radius sphere:

- **Almost all data live in the corners of the hyper cube**
- Almost all volume of high-dim sphere is contained in a thin slice
- There is essentially no interior volume
- As the number of dim increases, the volume of the sphere goes to zero
- . . .




If with 10 dimensions most data live in its 1024 corners, again, how do you do find your **nearest neighbors**?!



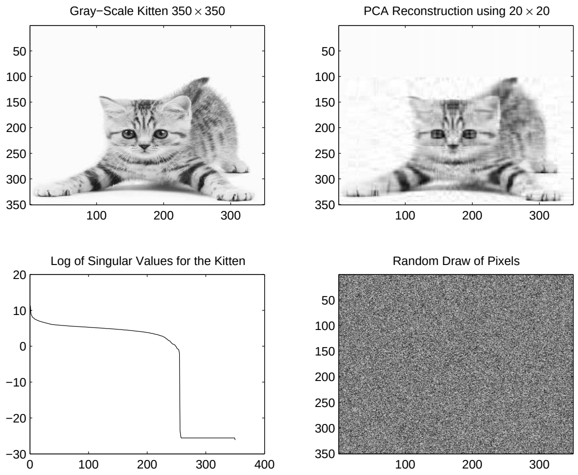
# Curse of Dimensionality (3)

To avoid overfitting, learning algorithms impose enough a priori structure (**regularization**)

**Manifold hypothesis:**
Real data—text, sounds, images—often live in a portion of the $R^D$ space that is effectively smaller than $D$ (**manifold learning**)

# Curse of Dimensionality (4)

…kittens seem to like living on a small manifold!



**Thank you for your patience. . .**