HIGH-DIMENSIONAL DATA AND THEIR PROJECTIONS

AP

DATA SCIENCE CONTEXT

- Dataset: n points in a d-dimensional space:
- essentially, a $n \times d$ matrix of floats
- For $d > 3$ and growing, several practical problems

1-HOT ENCODINGS RAISE DIMENSIONALITY

Feature vector **x** | Target y

	User				Movie					Other Movies rated					Time	Last Movie rated					Target y	
	A	B	C	...	TI	NH	SW	ST	...	TI	NH	SW	ST	...	Time	TI	NH	SW	ST	...		
$\mathbf{x}_1$	1	0	0	...	1	0	0	0	...	0.3	0.3	0.3	0	...	13	0	0	0	0	...	5	y_1
$\mathbf{x}_2$	1	0	0	...	0	1	0	0	...	0.3	0.3	0.3	0	...	14	1	0	0	0	...	3	y_2
$\mathbf{x}_3$	1	0	0	...	0	0	1	0	...	0.3	0.3	0.3	0	...	16	0	1	0	0	...	1	y_3
$\mathbf{x}_4$	0	1	0	...	0	0	1	0	...	0	0	0.5	0.5	...	5	0	0	0	0	...	4	y_4
$\mathbf{x}_5$	0	1	0	...	0	0	0	1	...	0	0	0.5	0.5	...	8	0	0	1	0	...	5	y_5
$\mathbf{x}_6$	0	0	1	...	1	0	0	0	...	0.5	0	0.5	0	...	9	0	0	0	0	...	1	y_6
$\mathbf{x}_7$	0	0	1	...	0	0	1	0	...	0.5	0	0.5	0	...	12	1	0	0	0	...	5	y_7

Fig. 1. Example (from Rendle [2010]) for representing a recommender problem with real valued feature vectors $\mathbf{x}$. Every row represents a feature vector $\mathbf{x}_i$ with its corresponding target y_i. For easier interpretation, the features are grouped into indicators for the active user (blue), active item (red), other movies rated by the same user (orange), the time in months (green), and the last movie rated (brown).

HOW TO SEE DIMENSIONS

data points are row vectors

	$\mathbf{X_1}$	$\mathbf{X_2}$	…	$\mathbf{X_d}$
$\mathbf{x_1}$	x_{11}	x_{12}	…	x_{1d}
…	…	…	…	…
$\mathbf{x_n}$	x_{n1}	x_{n2}	…	x_{nd}

ISSUES

- visualization is hard, we need projection. Which?
- decision-making is impaired by the need of chosing which dimensions to operate on
- **sensitivity analyis** or causal analysis: which dimension affects others?

ISSUES WITH HIGH-DIM. DATA

I: A FALSE SENSE OF SPARSITY

adding dimensions makes points seems further apart:

Name	Type	Degrees
Chianti	Red	12.5
Grenache	Rose	12
Bordeaux	Red	12.5
Cannonau	Red	13.5

d(Chianti, Bordeaux) = 0

let type differences count for 1:

d(red, rose) = 1

take the alcohol strengh as integer tenths-of-degree: d(12, 12.5) = 5

d(Chianti, Grenache) = $\sqrt{1^2 + 5^2} = 5.1$

Adding further dimensions make points seem further from each other

NOT CLOSE ANYMORE?

Name	Type	Degrees	Grape	Year
Chianti	Red	12.5	Sangiovese	2016
Grenache	Rose	12	Grenache	2011
Bordeaux	Red	12.5		2009
Cannonau	Red	13.5	Grenache	2015

d(Chianti, Bordeaux) > 7

d(Chianti, Grenache) > $\sqrt{5^2 + 1^2 + 5^2} = 7.14$

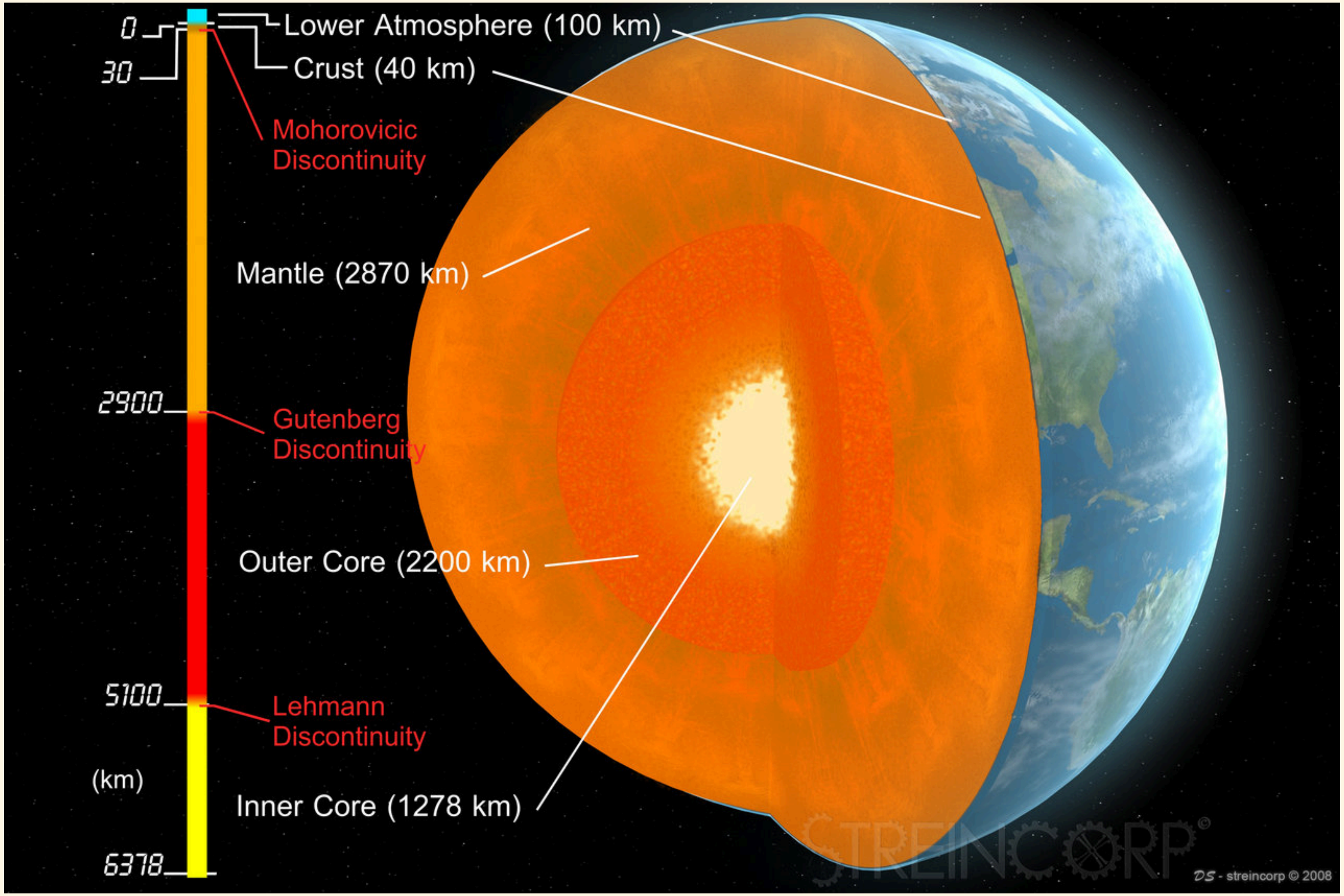
II: THE COLLAPSING ON THE SURFACE

Bodies have most of their mass distributed close to the surface (even under uniform density)

the *outer* orange is twice as big, but how much more juice will it give?

- for d=3, $vol = \frac{4}{3}\pi r^3$.
- With 50% radius, vol. is only $\frac{1}{8} = 12.5\%$

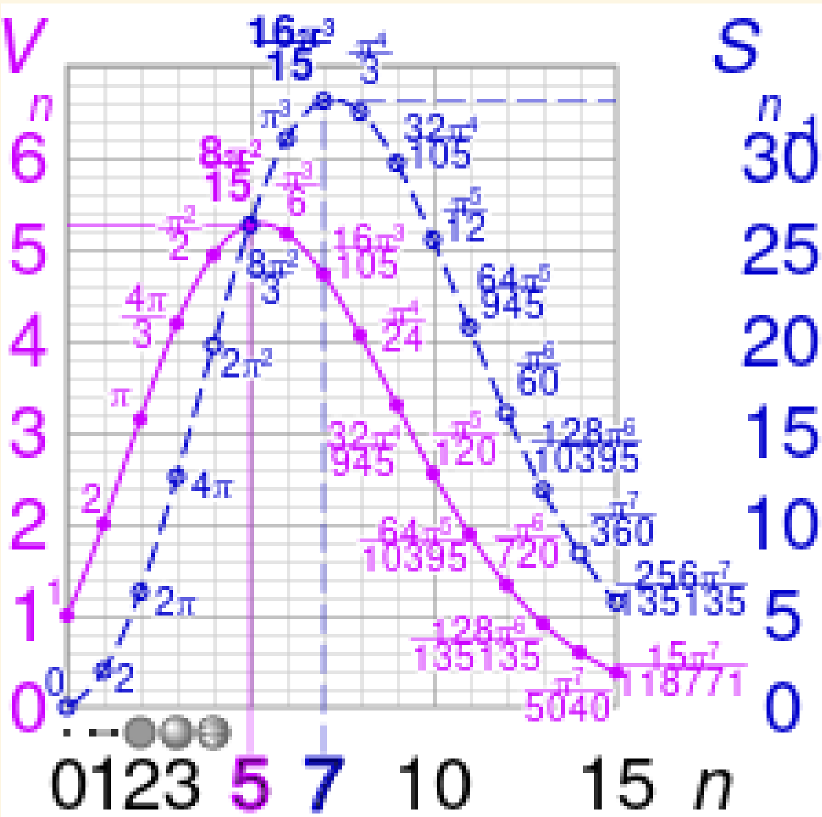
POSSIBLY MISGUIDING



The most volume (and thus weight) is in the external ring (the equators)

COUNTER-INTUITIVE PROPERTIES

At a fixed radius (r=1), raising dimensionality *above 5* in fact decreases the volume.



Hyperballs deflate.

Geometry is not what we experienced in $d \leq 3$.

THE CURSE OF DIMENSIONALITY

Volume will concentrate near the surface: most points will look as if they are at a uniform distance from each other

- distance-based similarity fails

CONSEQUENCES

ADDING DIMENSIONS APPARENTLY INCREASES SPARSITY

Deceiving as a chance to get a clean-cut segmentation of the data, as we did with Iris

In high dimension, all points tend to be at the same distance from each other

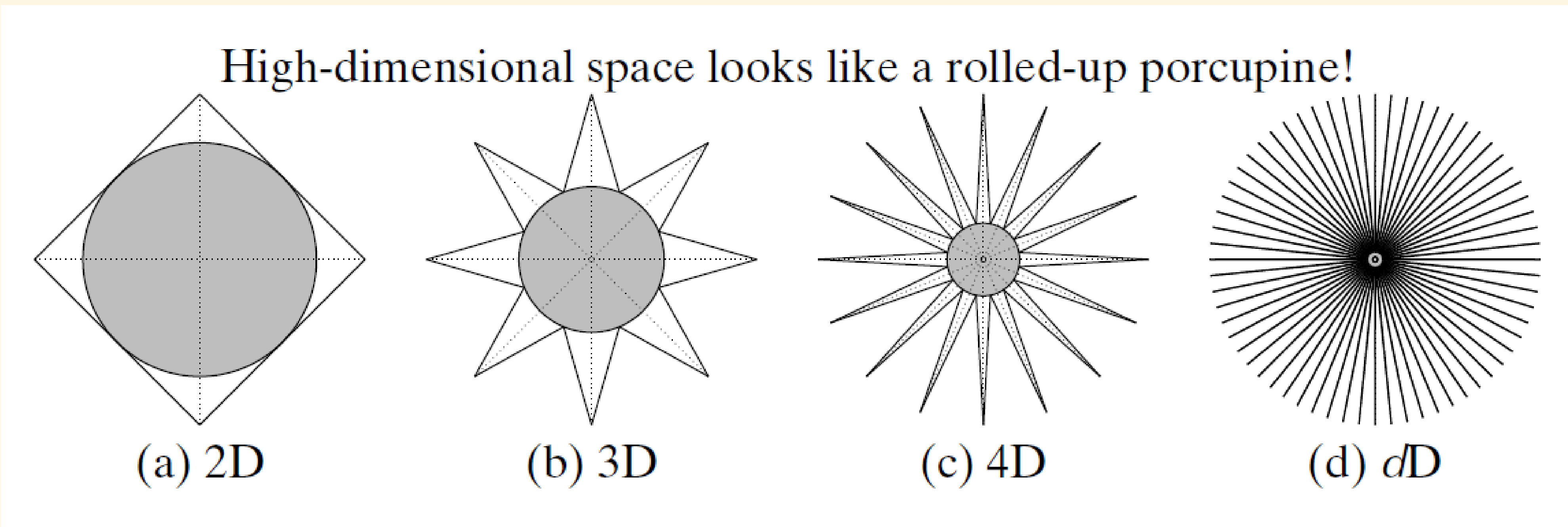
Exp: generate a set of random points in D^n, compute Frobenius norms: very little variance.

bye bye clustering algorithms, e.g., k-NN.

THE PORCUPINE

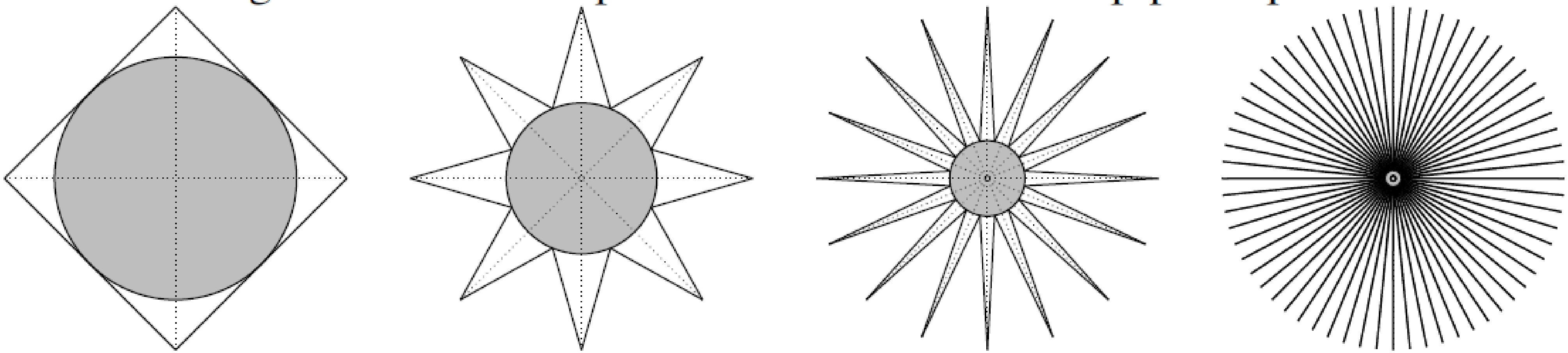
At high dimensions,

- all diagonals strangely become orthogonal to the axes
- points distributed along a diagonal gets ``*compressed down*” to the origin of axes.



WHERE ARE ALL MY DATA POINTS?

High-dimensional space looks like a rolled-up porcupine!



(a) 2D (b) 3D (c) 4D (d) *d*D