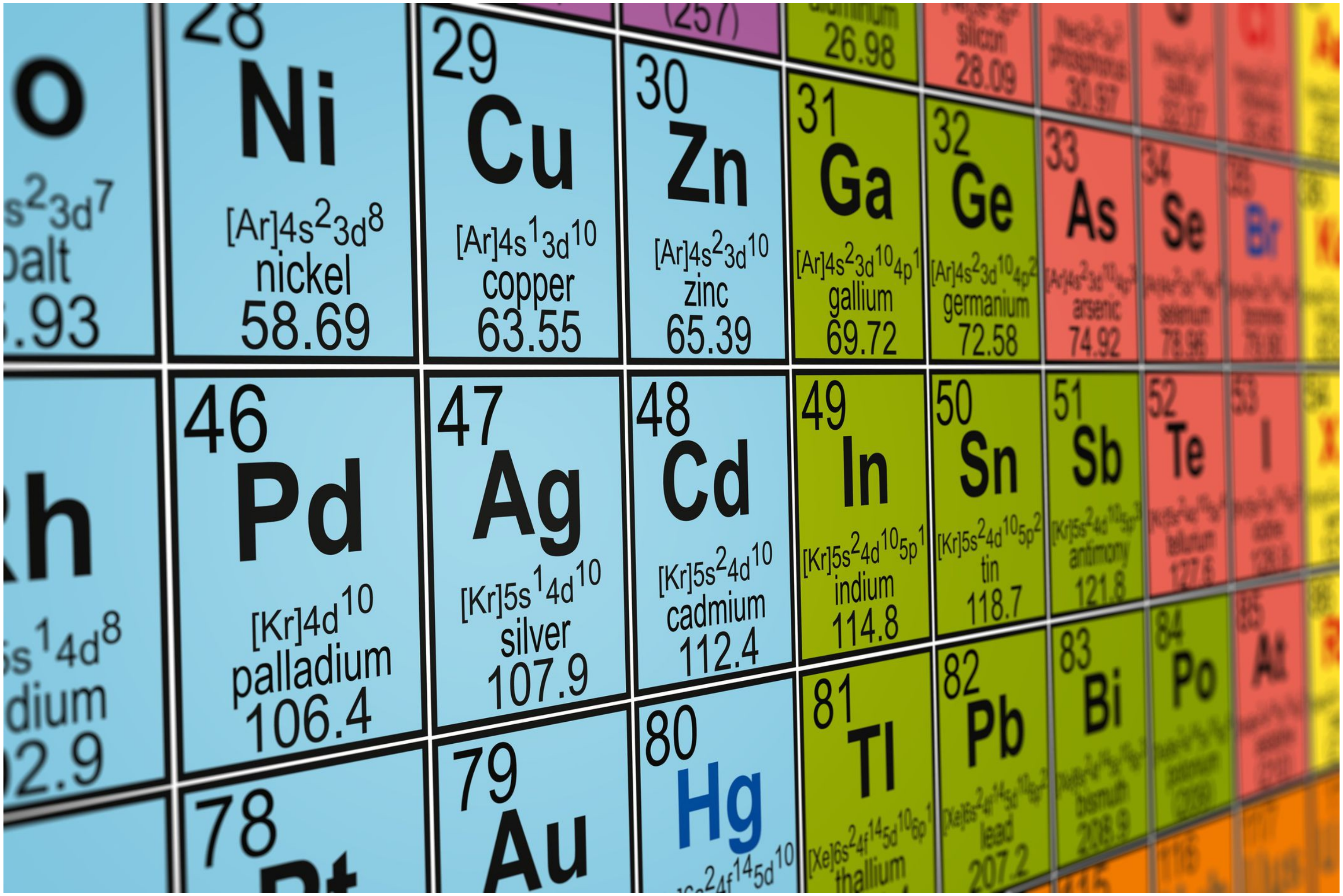
DIMENSIONALITY & VISUALIZATION

– Multidimensional Scaling and Related Approaches –

Moontae Lee

INFO3300: Data-driven Web Pages (4/20/2018)


28
Ni
[Ar]4s2 3d8
nickel
58.69
29
Cu
[Ar]4s1 3d10
copper
63.55
30
Zn
[Ar]4s2 3d10
zinc
65.39
26.98
28.09
31
Ga
[Ar]4s2 3d10 4p1
gallium
69.72
32
Ge
[Ar]4s2 3d10 4p2
germanium
72.58
33
As
arsenic
74.92
34
Se
46
Pd
[Kr]4d10
palladium
106.4
47
Ag
[Kr]5s1 4d10
silver
107.9
48
Cd
[Kr]5s2 4d10
cadmium
112.4
49
In
[Kr]5s2 4d10 5p1
indium
114.8
50
Sn
[Kr]5s2 4d10 5p2
tin
118.7
51
Sb
antimony
121.8
52
Te
I
78
79
Au
80
Hg
81
Tl
thallium
82
Pb
lead
207.2
83
Bi
84
Po
At

1	Hydrogen	H	1.00794
2	Helium	He	4.002602
3	Lithium	Li	6.941
4	Beryllium	Be	9.012182
5	Boron	B	10.811
6	Carbon	C	12.011
7	Nitrogen	N	14.00674
8	Oxygen	O	15.9994
9	Fluorine	F	18.9984032
10	Neon	Ne	20.1797
11	Sodium	Na	22.989768
12	Magnesium	Mg	24.3050
13	Aluminum	Al	26.981539
14	Silicon	Si	28.0855
15	Phosphorus	P	30.973762
16	Sulfur	S	32.066
17	Chlorine	Cl	35.4527
18	Argon	Ar	39.948
19	Potassium	K	39.0983
20	Calcium	Ca	40.078
21	Scandlum	Sc	44.955910
22	Titanium	Ti	47.867
23	Vanadium	V	50.9415
24	Chromium	Cr	51.9961
25	Manganese	Mn	54.93805
26	Iron	Fe	55.845
27	Cobalt	Co	58.93320
28	Nickel	Ni	58.6934
29	Copper	Cu	63.546
30	Zinc	Zn	65.39
31	Gallium	Ga	69.723
32	Germanium	Ge	72.61

55	Cesium	Cs	132.90543
56	Barium	Ba	137.327
57	Lanthanum	La	138.9055
58	Cerium	Ce	140.115
59	Praseodymium	Pr	140.90765
60	Neodymium	Nd	144.24
61	Promethium	Pm	144.9127*
62	Samarium	Sm	150.36
63	Europium	Eu	151.965
64	Gadolinium	Gd	157.25
65	Terbium	Tb	158.92534
66	Dysprosium	Dy	162.50
67	Holmium	Ho	164.93032
68	Erbium	Er	167.26
69	Thulium	Tm	168.93421
70	Ytterbium	Yb	173.04
71	Lutetium	Lu	174.967
72	Hafnium	Hf	178.49
73	Tantalum	Ta	180.9479
74	Tungsten	W	183.84
75	Rhenium	Re	186.207
76	Osmium	Os	190.23
77	Iridium	Ir	192.217
78	Platinum	Pt	195.08
79	Gold	Au	196.96654
80	Mercury	Hg	200.59
81	Thallium	Tl	204.3833
82	Lead	Pb	207.2
83	Bismuth	Bi	208.98037
84	Polonium	Po	208.9824*
85	Astatine	At	209.9871*
86	Radon	Rn	222.0176*

Atomic Number → 1

Symbol → H

1.008 ← Atomic Mass

Hydrogen ← Name

1 H 1.008 Hydrogen																	2 He 4.002602 Helium
3 Li 6.94 Lithium	4 Be 9.0121831 Beryllium											5 B 10.81 Boron	6 C 12.011 Carbon	7 N 14.007 Nitrogen	8 O 15.999 Oxygen	9 F 18.998403163 Fluorine	10 Ne 20.1797 Neon
11 Na 22.98976928 Sodium	12 Mg 24.305 Magnesium											13 Al 26.9815385 Aluminium	14 Si 28.085 Silicon	15 P 30.973761998 Phosphorus	16 S 32.06 Sulfur	17 Cl 35.45 Chlorine	18 Ar 39.948 Argon
19 K 39.0983 Potassium	20 Ca 40.078 Calcium	21 Sc 44.955908 Scandium	22 Ti 47.867 Titanium	23 V 50.9415 Vanadium	24 Cr 51.9961 Chromium	25 Mn 54.938044 Manganese	26 Fe 55.845 Iron	27 Co 58.933194 Cobalt	28 Ni 58.6934 Nickel	29 Cu 63.546 Copper	30 Zn 65.38 Zinc	31 Ga 69.723 Gallium	32 Ge 72.630 Germanium	33 As 74.921595 Arsenic	34 Se 78.971 Selenium	35 Br 79.904 Bromine	36 Kr 83.798 Krypton
37 Rb 85.4678 Rubidium	38 Sr 87.62 Strontium	39 Y 88.90584 Yttrium	40 Zr 91.224 Zirconium	41 Nb 92.90637 Niobium	42 Mo 95.95 Molybdenum	43 Tc 98 Technetium	44 Ru 101.07 Ruthenium	45 Rh 102.90550 Rhodium	46 Pd 106.42 Palladium	47 Ag 107.8682 Silver	48 Cd 112.414 Cadmium	49 In 114.818 Indium	50 Sn 118.710 Tin	51 Sb 121.760 Antimony	52 Te 127.60 Tellurium	53 I 126.90447 Iodine	54 Xe 131.293 Xenon
55 Cs 132.90545196 Caesium	56 Ba 137.327 Barium	57/71	72 Hf 178.49 Hafnium	73 Ta 180.94788 Tantalum	74 W 183.84 Tungsten	75 Re 186.207 Rhenium	76 Os 190.23 Osmium	77 Ir 192.217 Iridium	78 Pt 195.084 Platinum	79 Au 196.966569 Gold	80 Hg 200.592 Mercury	81 Tl 204.38 Thallium	82 Pb 207.2 Lead	83 Bi 208.98040 Bismuth	84 Po 209 Polonium	85 At 210 Astatine	86 Rn 222 Radon
87 Fr 223 Francium	88 Ra 226 Radium	89/103	104 Rf 267 Rutherfordium	105 Db 268 Dubnium	106 Sg 269 Seaborgium	107 Bh 270 Bohrium	108 Hs 269 Hassium	109 Mt 278 Meitnerium	110 Ds 281 Darmstadtium	111 Rg 281 Roentgenium	112 Cn 285 Copernicium	113 Uut 286 Ununtrium	114 Fl 289 Flerovium	115 Uup 289 Ununpentium	116 Lv 293 Livermorium	117 Uus 294 Ununseptium	118 Uuo 294 Ununoctium

Lanthanide Series	57 La 138.90547 Lanthanum	58 Ce 140.116 Cerium	59 Pr 140.90766 Praseodymium	60 Nd 144.242 Neodymium	61 Pm 145 Promethium	62 Sm 150.36 Samarium	63 Eu 151.964 Europium	64 Gd 157.25 Gadolinium	65 Tb 158.92535 Terbium	66 Dy 162.500 Dysprosium	67 Ho 164.93033 Holmium	68 Er 167.259 Erbium	69 Tm 168.93422 Thulium	70 Yb 173.054 Ytterbium	71 Lu 174.9668 Lutetium
Actinide Series	89 Ac 227 Actinium	90 Th 232.0377 Thorium	91 Pa 231.03588 Protactinium	92 U 238.02891 Uranium	93 Np 237 Neptunium	94 Pu 244 Plutonium	95 Am 243 Americium	96 Cm 247 Curium	97 Bk 247 Berkelium	98 Cf 251 Californium	99 Es 252 Einsteinium	100 Fm 257 Fermium	101 Md 258 Mendelevium	102 No 259 Nobelium	103 Lr 266 Lawrencium

Better visualization?

Atomic #
Symbol
Name
Weight

C Solid
Hg Liquid
H Gas
Rf Unknown

Metals: Alkali metals | Alkaline earth metals | Lanthanoids (Lanthanides) | Actinoids (Actinides) | Transition metals | Post-transition metals
Metalloids
Nonmetals: Other nonmetals | Noble gases

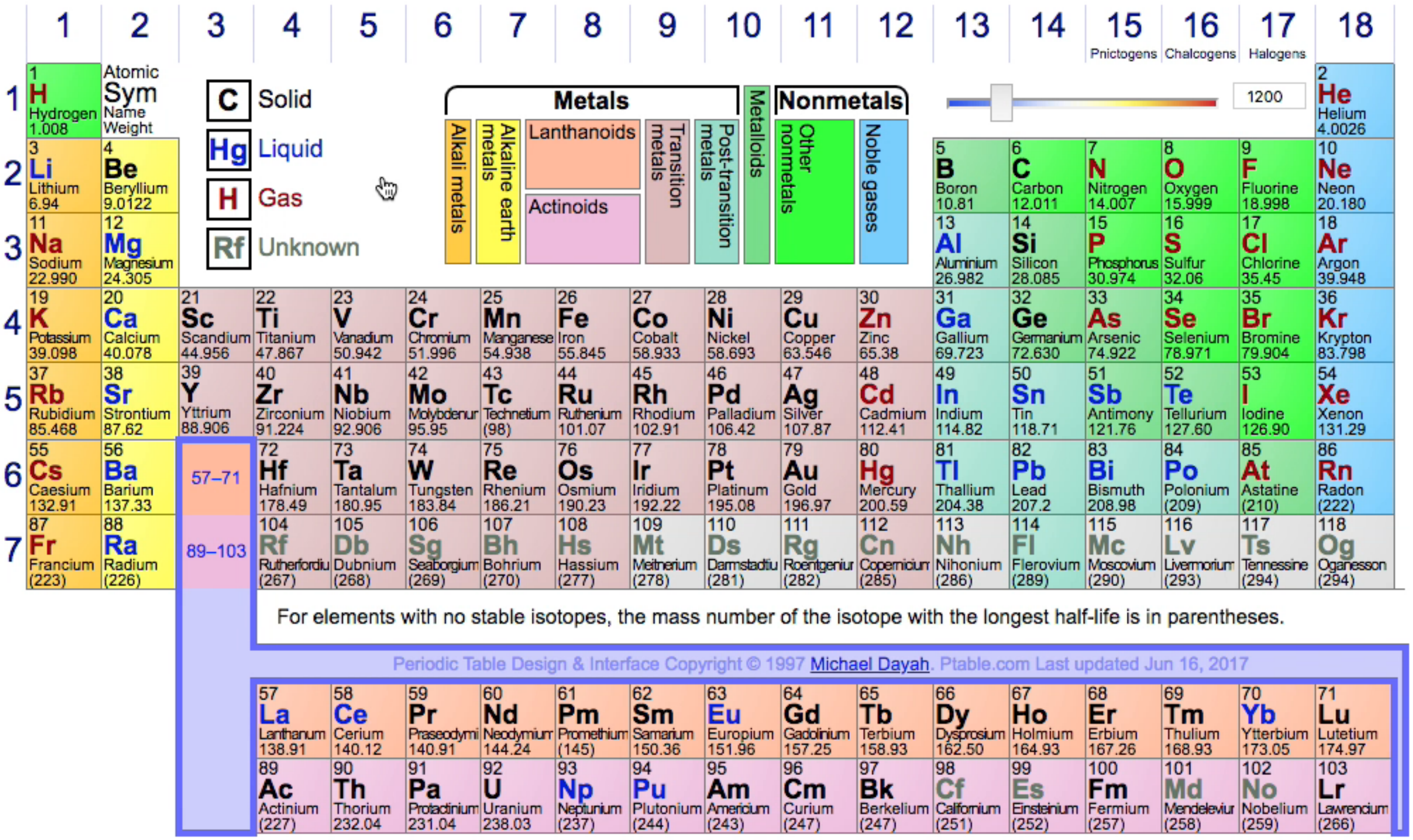
Pnictogens (15) · Chalcogens (16) · Halogens (17)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H Hydrogen 1.008																	2 He Helium 4.0026
2	3 Li Lithium 6.94	4 Be Beryllium 9.0122											5 B Boron 10.81	6 C Carbon 12.011	7 N Nitrogen 14.007	8 O Oxygen 15.999	9 F Fluorine 18.998	10 Ne Neon 20.180
3	11 Na Sodium 22.990	12 Mg Magnesium 24.305											13 Al Aluminium 26.982	14 Si Silicon 28.085	15 P Phosphorus 30.974	16 S Sulfur 32.06	17 Cl Chlorine 35.45	18 Ar Argon 39.948
4	19 K Potassium 39.098	20 Ca Calcium 40.078	21 Sc Scandium 44.956	22 Ti Titanium 47.867	23 V Vanadium 50.942	24 Cr Chromium 51.996	25 Mn Manganese 54.938	26 Fe Iron 55.845	27 Co Cobalt 58.933	28 Ni Nickel 58.693	29 Cu Copper 63.546	30 Zn Zinc 65.38	31 Ga Gallium 69.723	32 Ge Germanium 72.630	33 As Arsenic 74.922	34 Se Selenium 78.971	35 Br Bromine 79.904	36 Kr Krypton 83.798
5	37 Rb Rubidium 85.468	38 Sr Strontium 87.62	39 Y Yttrium 88.906	40 Zr Zirconium 91.224	41 Nb Niobium 92.906	42 Mo Molybdenum 95.95	43 Tc Technetium (98)	44 Ru Ruthenium 101.07	45 Rh Rhodium 102.91	46 Pd Palladium 106.42	47 Ag Silver 107.87	48 Cd Cadmium 112.41	49 In Indium 114.82	50 Sn Tin 118.71	51 Sb Antimony 121.76	52 Te Tellurium 127.60	53 I Iodine 126.90	54 Xe Xenon 131.29
6	55 Cs Caesium 132.91	56 Ba Barium 137.33	57–71	72 Hf Hafnium 178.49	73 Ta Tantalum 180.95	74 W Tungsten 183.84	75 Re Rhenium 186.21	76 Os Osmium 190.23	77 Ir Iridium 192.22	78 Pt Platinum 195.08	79 Au Gold 196.97	80 Hg Mercury 200.59	81 Tl Thallium 204.38	82 Pb Lead 207.2	83 Bi Bismuth 208.98	84 Po Polonium (209)	85 At Astatine (210)	86 Rn Radon (222)
7	87 Fr Francium (223)	88 Ra Radium (226)	89–103	104 Rf Rutherfordium (267)	105 Db Dubnium (268)	106 Sg Seaborgium (269)	107 Bh Bohrium (270)	108 Hs Hassium (277)	109 Mt Meitnerium (278)	110 Ds Darmstadtium (281)	111 Rg Roentgenium (282)	112 Cn Copernicium (285)	113 Nh Nihonium (286)	114 Fl Flerovium (289)	115 Mc Moscovium (290)	116 Lv Livermorium (293)	117 Ts Tennessine (294)	118 Og Oganesson (294)

For elements with no stable isotopes, the mass number of the isotope with the longest half-life is in parentheses.

6	57 La Lanthanum 138.91	58 Ce Cerium 140.12	59 Pr Praseodymium 140.91	60 Nd Neodymium 144.24	61 Pm Promethium (145)	62 Sm Samarium 150.36	63 Eu Europium 151.96	64 Gd Gadolinium 157.25	65 Tb Terbium 158.93	66 Dy Dysprosium 162.50	67 Ho Holmium 164.93	68 Er Erbium 167.26	69 Tm Thulium 168.93	70 Yb Ytterbium 173.05	71 Lu Lutetium 174.97

What about dynamic?



Implications

Good visualization helps better understanding large complex data.

Implications

Good visualization helps better understanding large complex data.

succinct

Implications

Good visualization helps better understanding large complex data.

succinct

transparent

Implications

Good visualization helps better understanding large complex data.

succinct

transparent

comparative

Implications

Good visualization helps better understanding large complex data.

succinct | transparent | comparative | coherent

Even better

Good visualization helps better understanding large complex data.

succinct | transparent | comparative | coherent

beautiful | creative

Failure modes

If you tried very hard but frustrated,

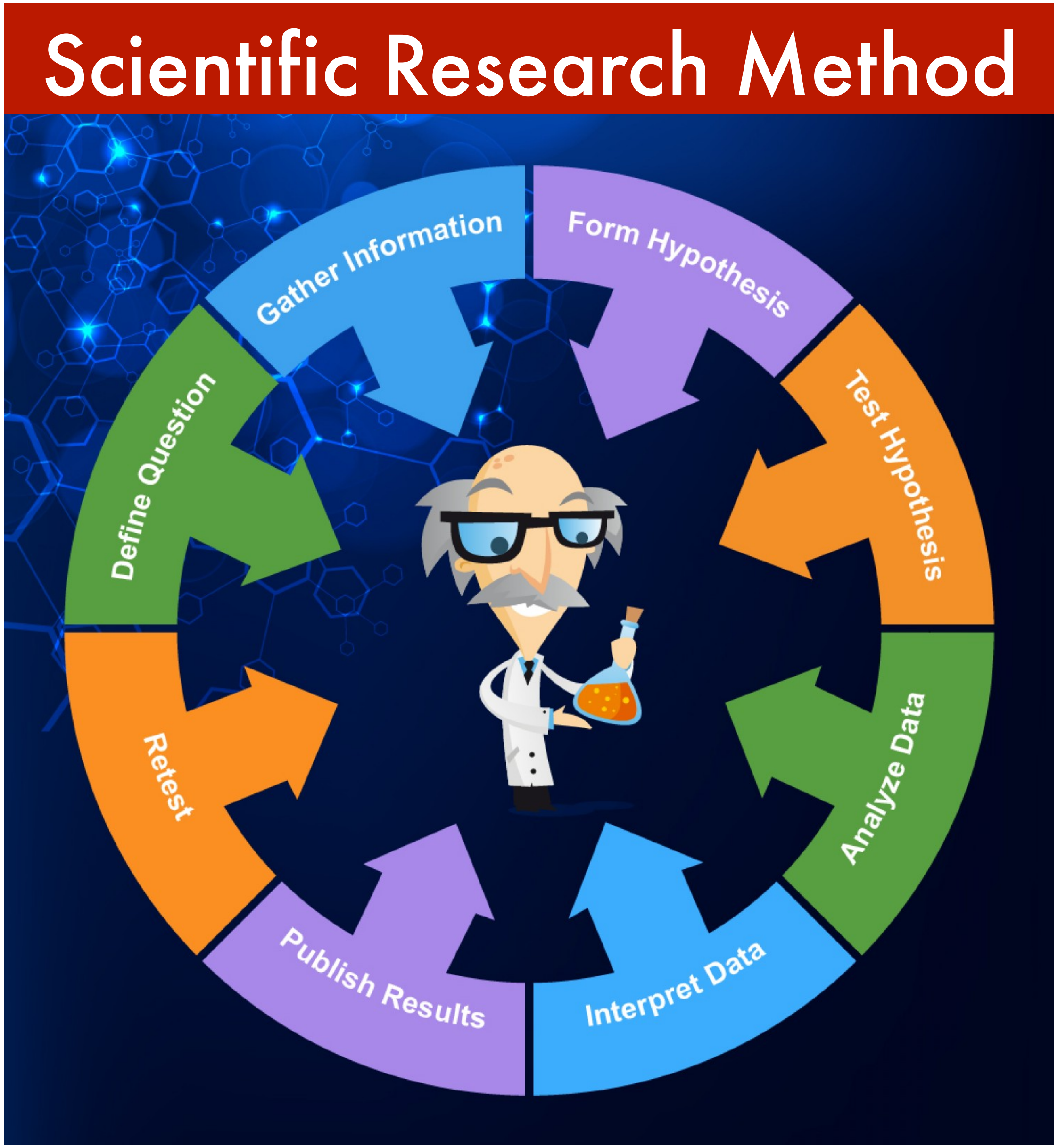
Failure modes

If you tried very hard but frustrated, then check

Readability: too much content? busy?

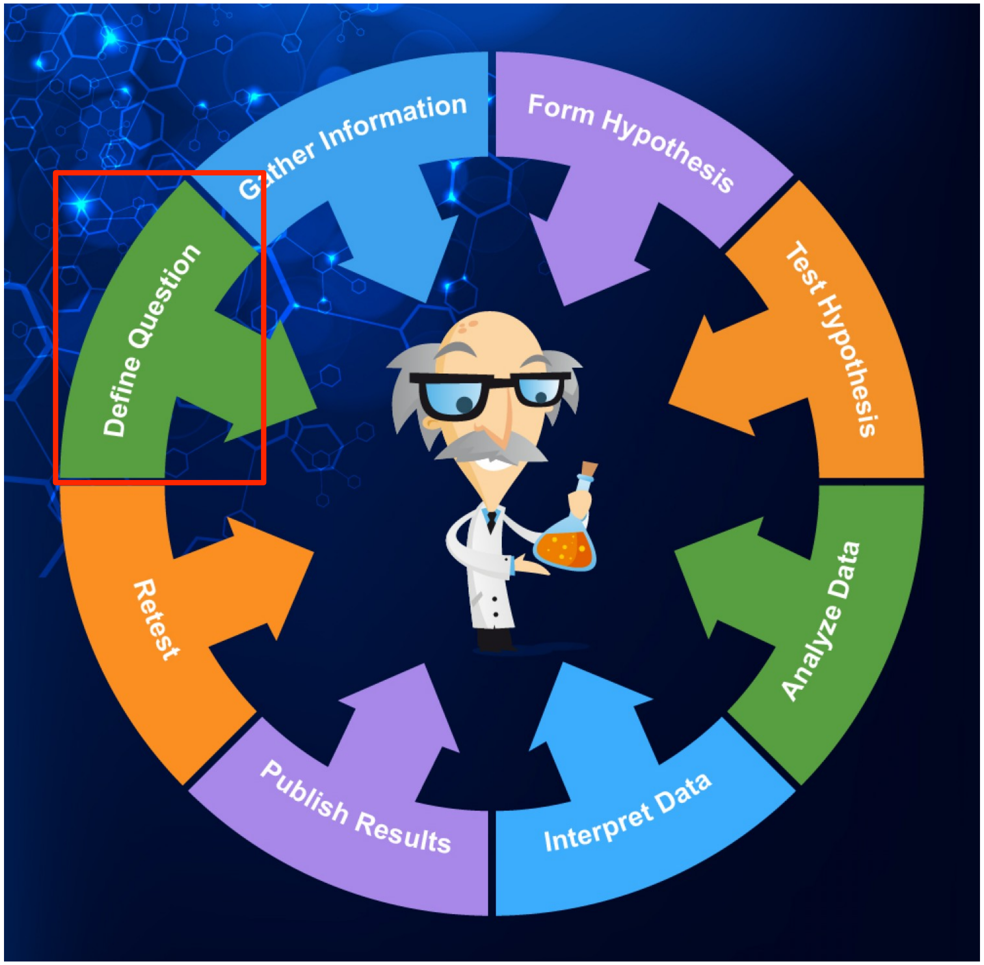
Usability: inform useful knowledge?

Accessibility, Creativity, and Complexity


Scientific Research Method
Gather Information
Form Hypothesis
Test Hypothesis
Analyze Data
Interpret Data
Publish Results
Retest
Define Question

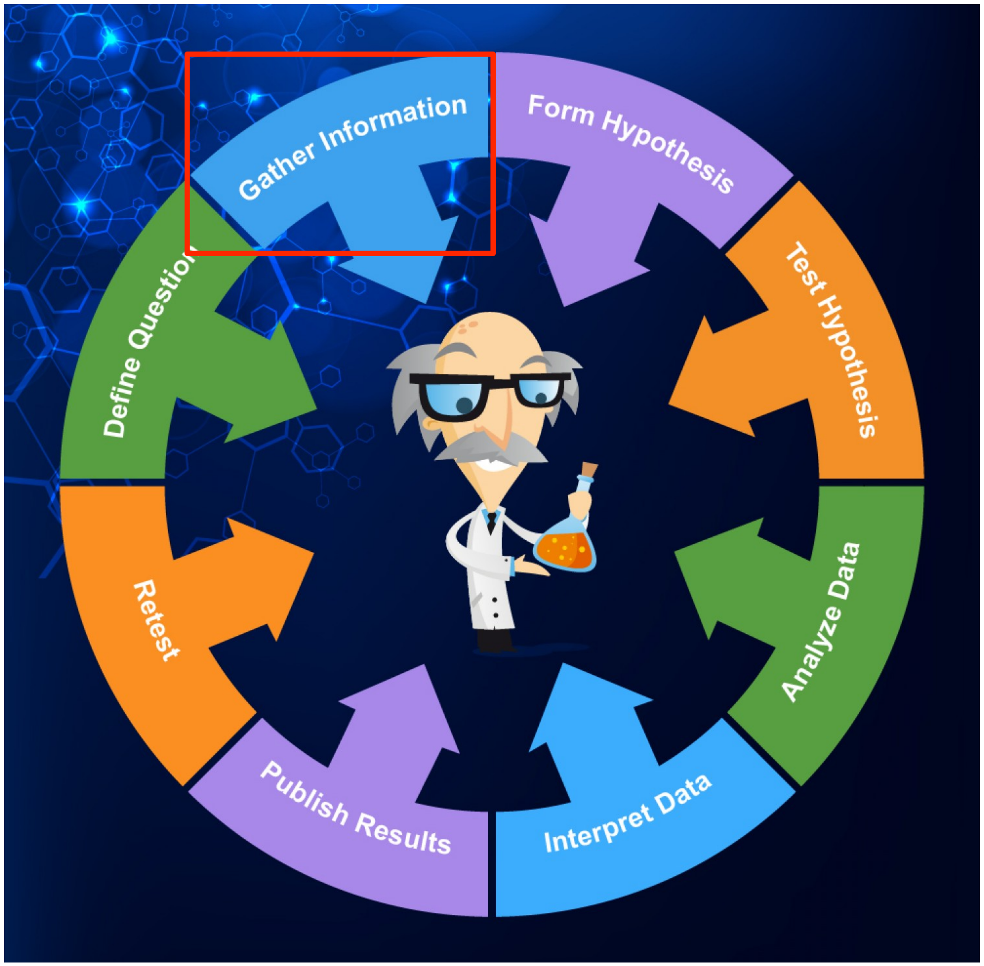
David's mindmap

David’s mindmap



Based on your observations.

David’s mindmap



Based on your observations.

By picking the data of your interest.

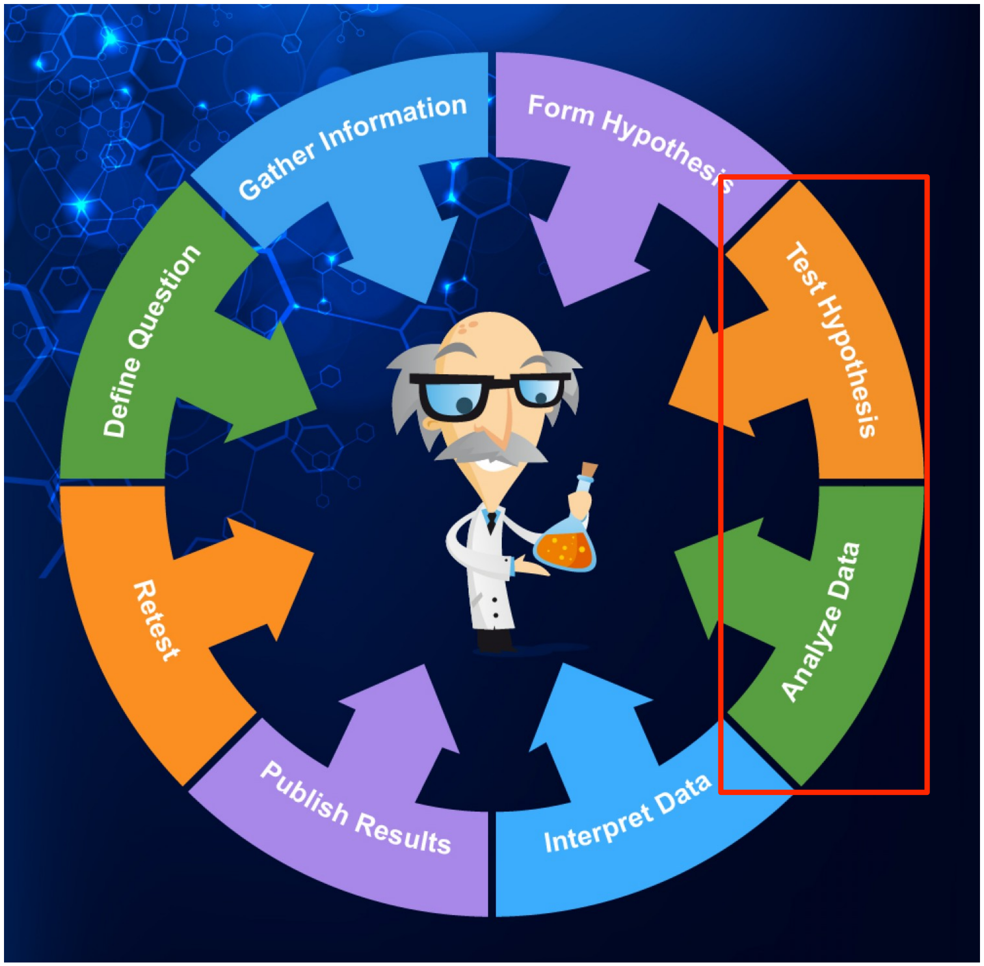
David's mindmap

Based on your observations.

By picking the data of your interest.

That your data can answer.

David's mindmap



Based on your observations.

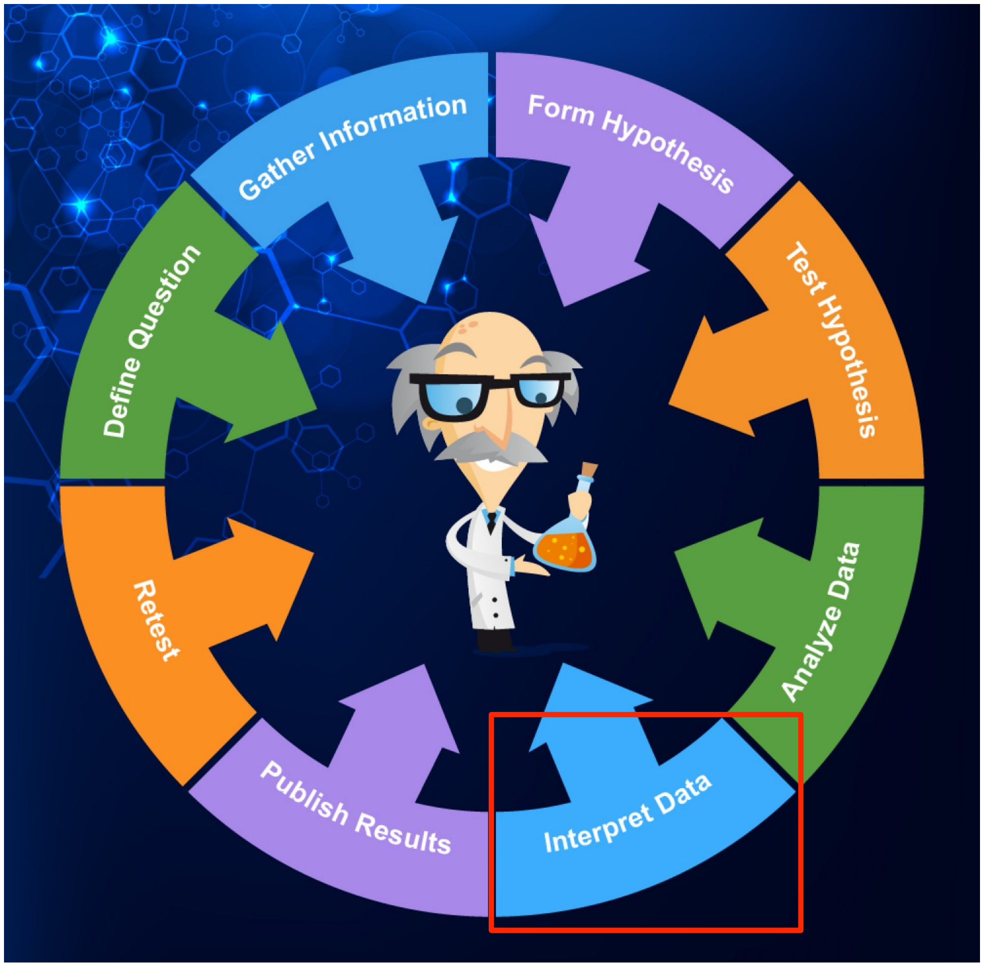
By picking the data of your interest.

That your data can answer.

Via various data-science tools.

Regression, Distributions
Stochastic Gradient Descent,
Clustering, Reduction

David's mindmap



Based on your observations.

By picking the data of your interest.

That your data can answer.

Via various data-science tools.

Regression, Distributions
Stochastic Gradient Descent,
Clustering, Reduction

Based on your visualization!

David's mindmap

Based on your observations.

By picking the data of your interest.

That your data can answer.

Via various data-science tools.

Regression, Distributions
Stochastic Gradient Descent,
Clustering, Reduction

Based on your visualization!

You learned Javascript and D3!

David's mindmap

Based on your observations.

By picking the data of your interest.

That your data can answer.

Via various data-science tools.

Regression, Distributions
Stochastic Gradient Descent,
Clustering, Reduction

Based on your visualization!

You learned Javascript and D3!

Based on comments and feedback.

He is ambitious!

1. Critical ingredient of readability?

2. Can visualization inspire new/better models?

1. Must be human interpretable

(even if the data is complex & high-dimension)

1. Must be human interpretable
(even if the data is complex & high-dimension)

2. Yes!

Cognitive limit

Only 2D (or 3D at most) figures are readable.

Cognitive limit

Only 2D (or 3D at most) figures are readable.

What if with more than 2 feature variables?

(Estimate housing price by {size, distance, area, ...})

Cognitive limit

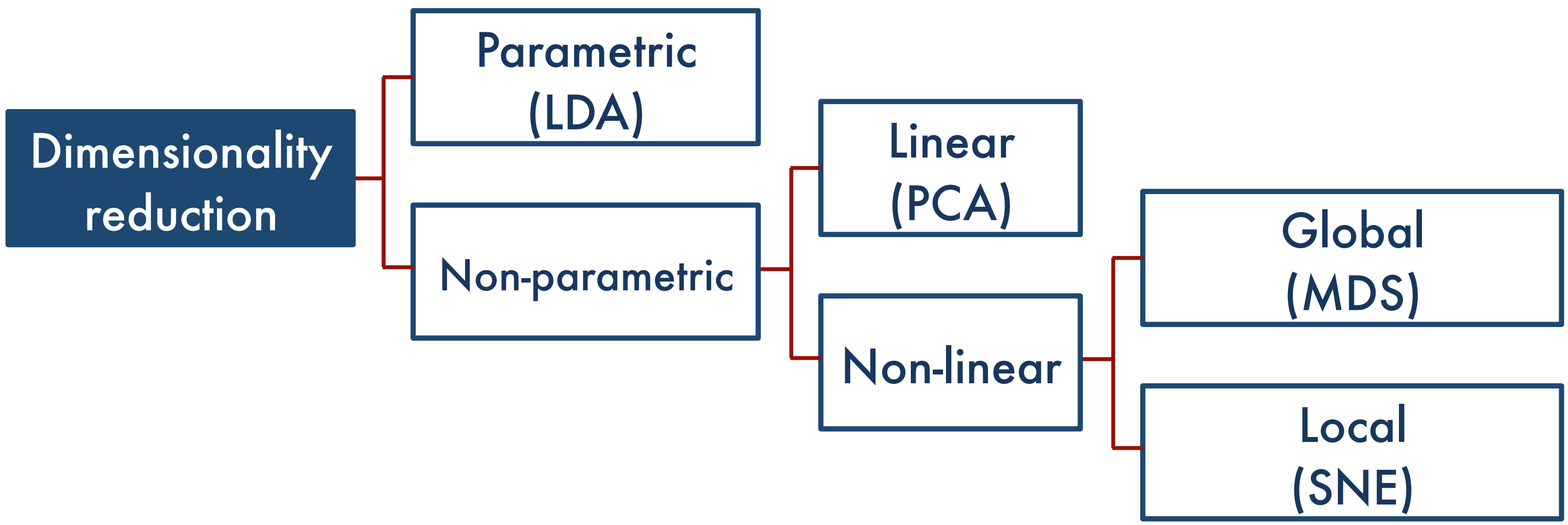
Only 2D (or 3D at most) figures are readable.

What if with more than 2 feature variables?

(Estimate housing price by {size, distance, area, ...})

Dimensionality reduction!

Map of the tools



Common assumptions

NOT that complex (in reality)!

NOT that complicated (in details)!

Common assumptions

NOT that complex (in reality)!

- Looking complex, but lies in much simple space.
- That space gives better representations.

NOT that complicated (in details)!

Common assumptions

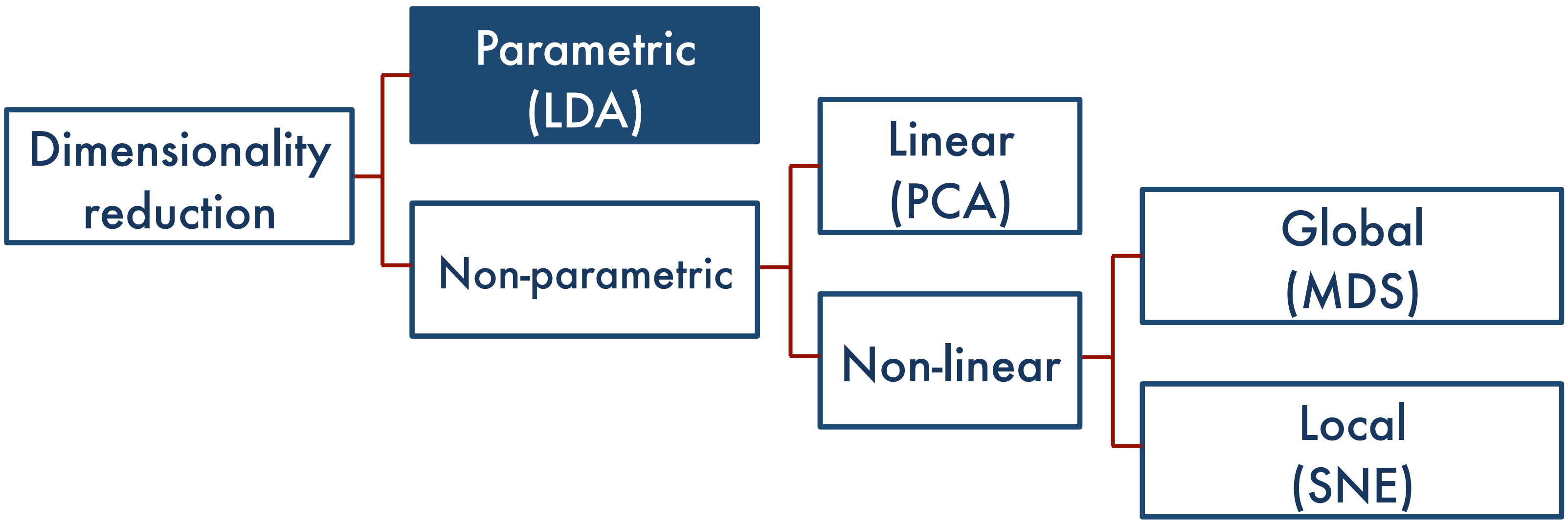
NOT that complex (in reality)!

- Looking complex, but lies in much simple space.
- That space gives better representations.

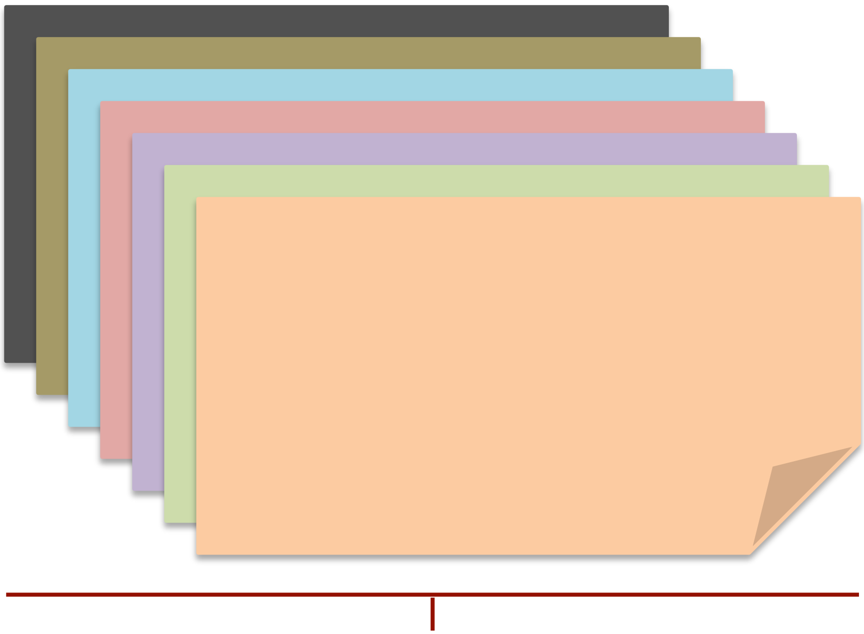
NOT that complicated (in details)!

- Most information is pairwise between data points.
- In particular, by pairwise distances.

Map of the tools

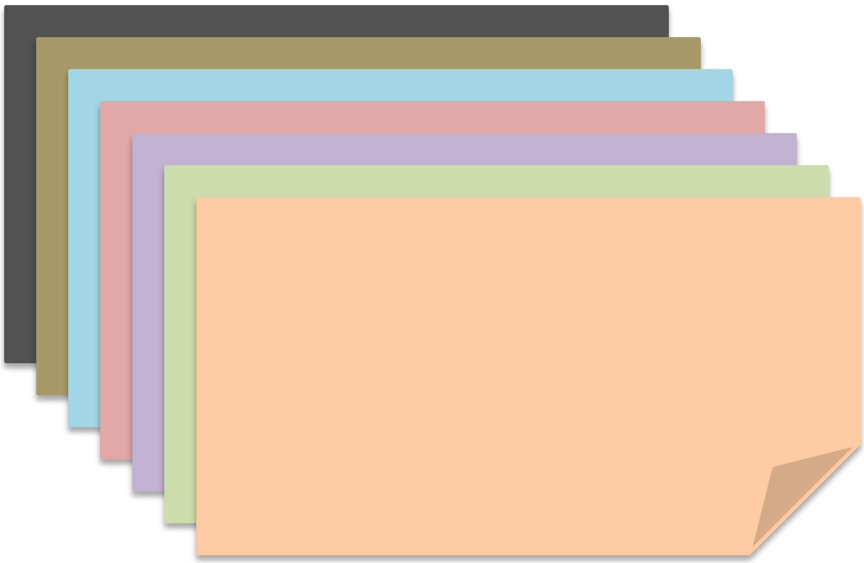


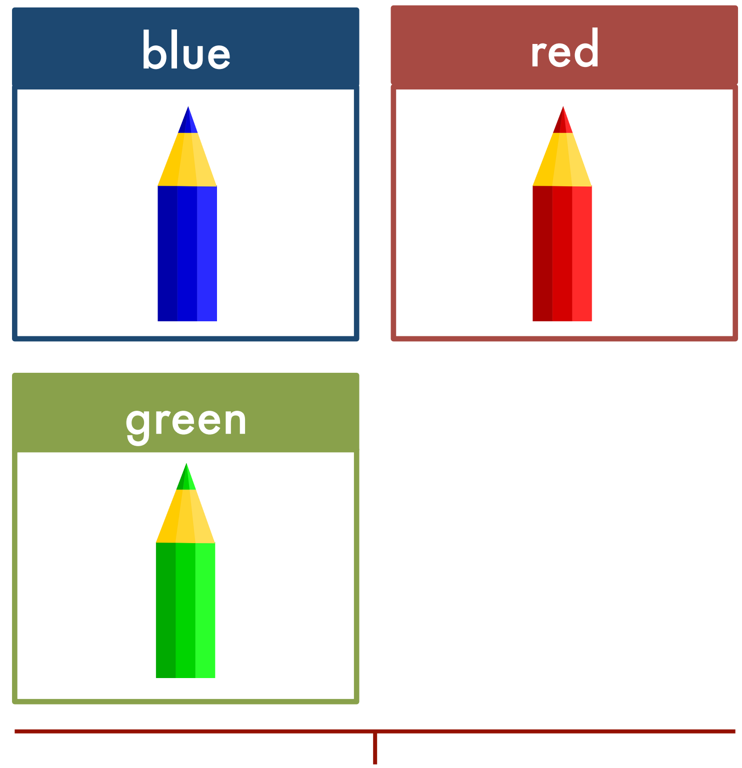
Motivation: coloring



2^{24} sheets with different colors

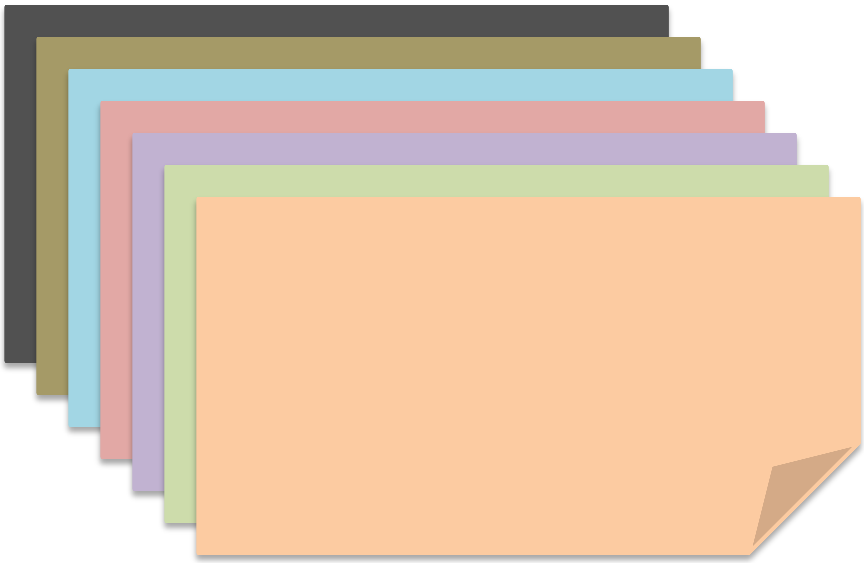
Motivation: coloring



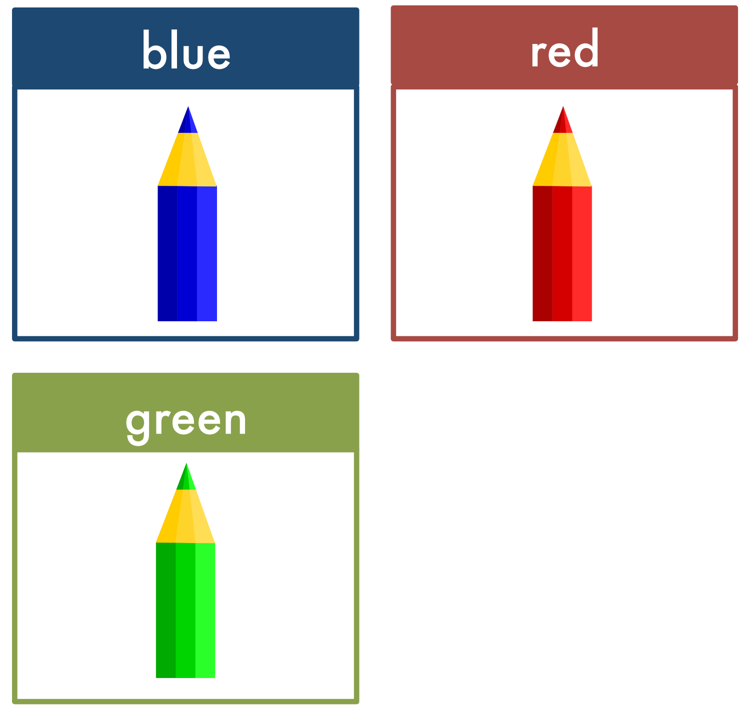


Bring basis pencils

Motivation: coloring

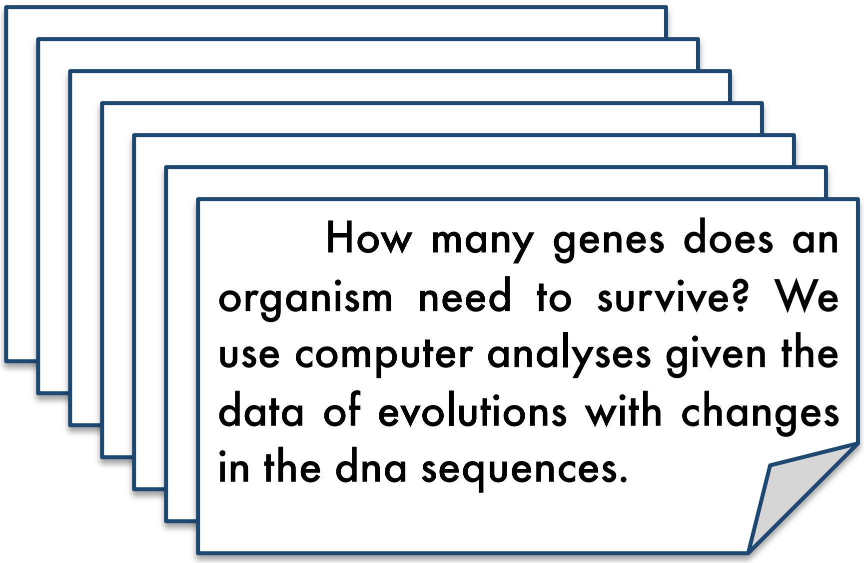


(R=255, G=199, B=143)



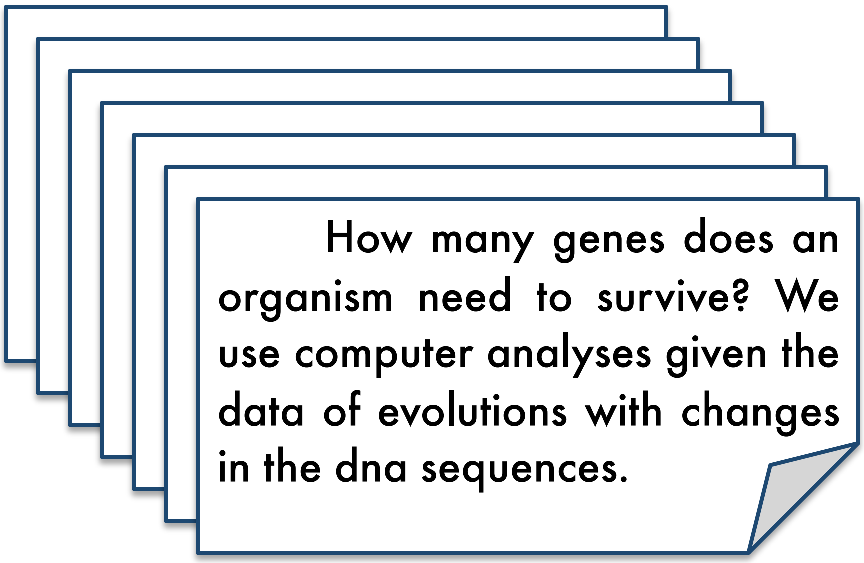
Bring basis pencils

Topic modeling



17,000 Articles from the journal "Science"

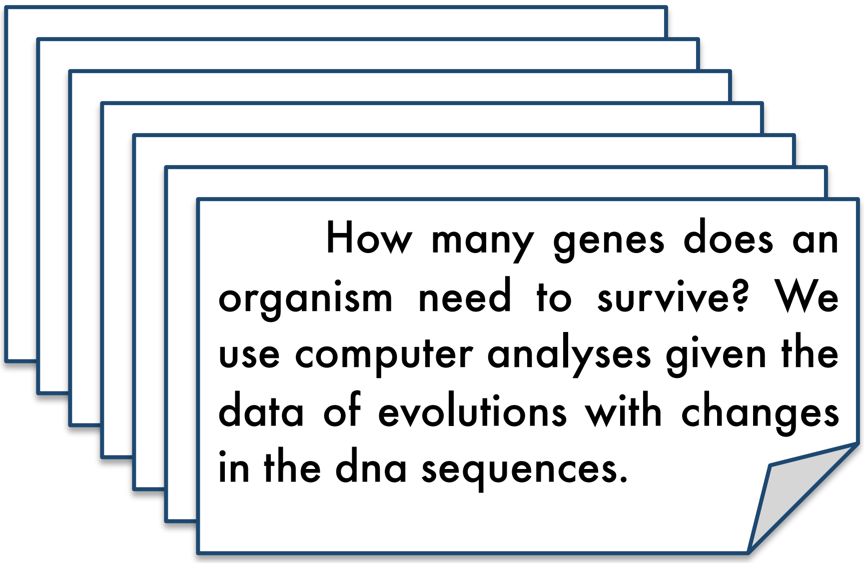
Topic modeling



17,000 Articles from the journal "Science"

user input: 4 topics

Topic modeling



b_1: biology	
dna	0.04
genes	0.02
data	0.01
...	...

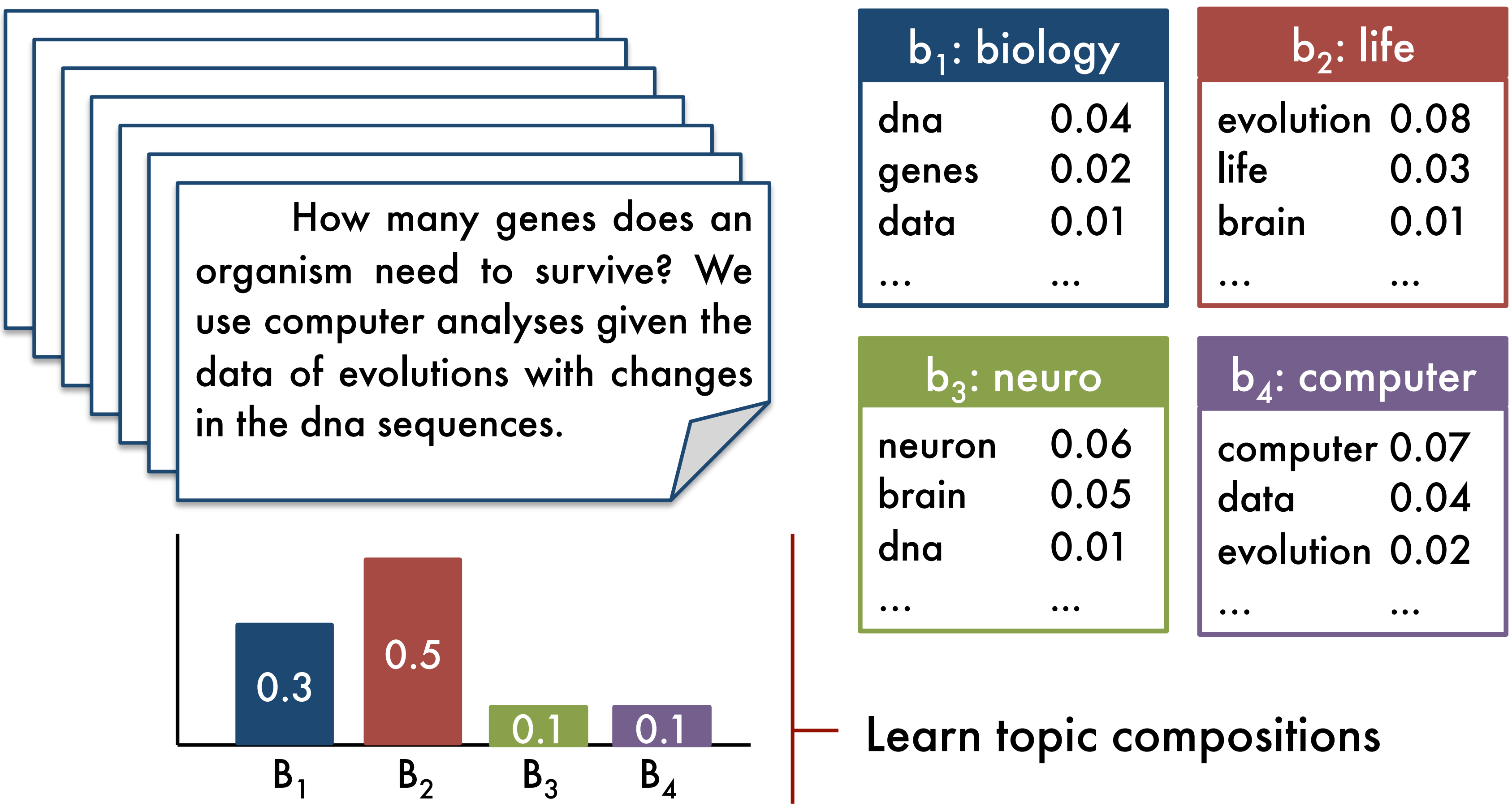
b_2: life	
evolution	0.08
life	0.03
brain	0.01
...	...

b_3: neuro	
neuron	0.06
brain	0.05
dna	0.01
...	...

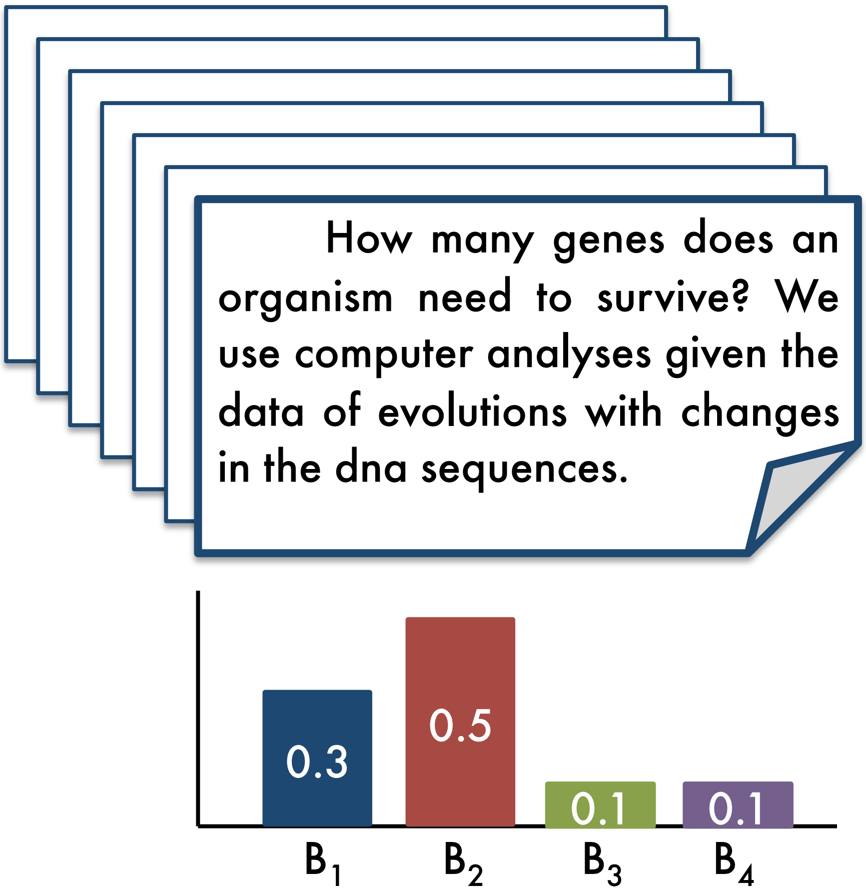
b_4: computer	
computer	0.07
data	0.04
evolution	0.02
...	...

Learn topics

Topic modeling



Topic modeling



b_1: biology	
dna	0.04
genes	0.02
data	0.01
…	…

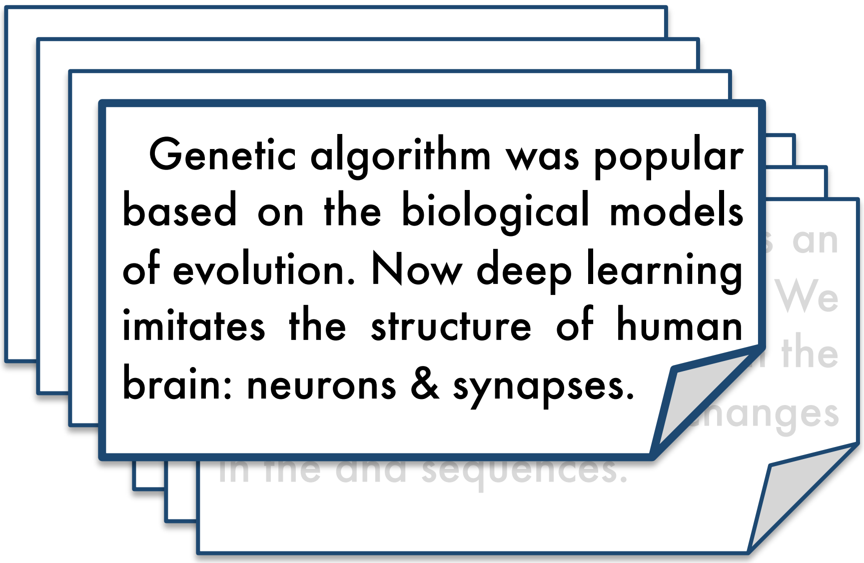
b_2: life	
evolution	0.08
life	0.03
brain	0.01
…	…

b_3: neuro	
neuron	0.06
brain	0.05
dna	0.01
…	…

b_4: computer	
computer	0.07
data	0.04
evolution	0.02
…	…

$$\text{document} = (0.3, 0.5, 0.1, 0.1) \in \mathbb{R}^4$$

Topic modeling



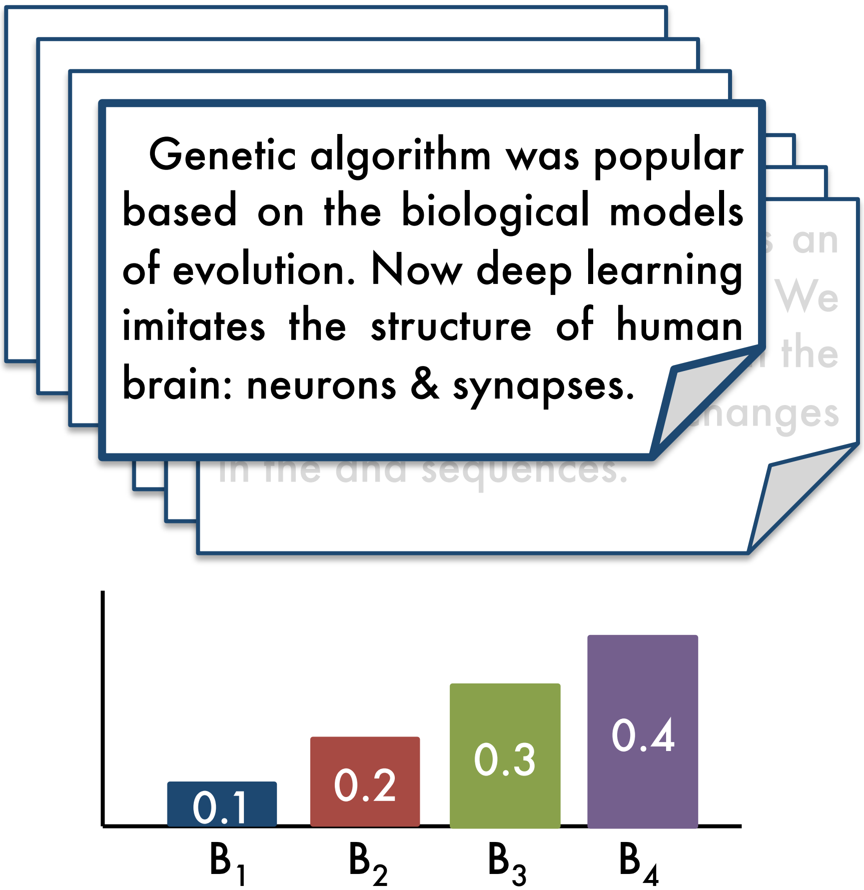
b$_1$: biology	
dna	0.04
genes	0.02
data	0.01
…	…

b$_2$: life	
evolution	0.08
life	0.03
brain	0.01
…	…

b$_3$: neuro	
neuron	0.06
brain	0.05
dna	0.01
…	…

b$_4$: computer	
computer	0.07
data	0.04
evolution	0.02
…	…

Topic modeling



b_1: biology	
dna	0.04
genes	0.02
data	0.01
...	...

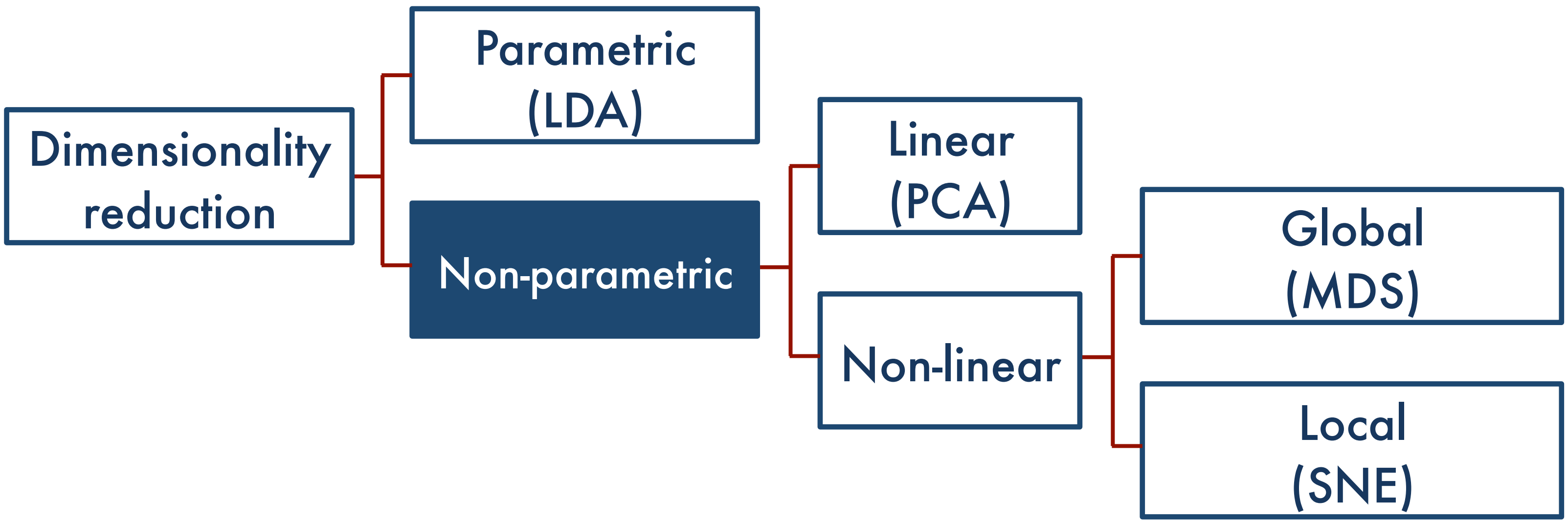
b_2: life	
evolution	0.08
life	0.03
brain	0.01
...	...

b_3: neuro	
neuron	0.06
brain	0.05
dna	0.01
...	...

b_4: computer	
computer	0.07
data	0.04
evolution	0.02
...	...

$$\text{document} = (0.1, 0.2, 0.3, 0.4) \in \mathbb{R}^4$$

Map of the tools



Non-parametric reduction

Goals:

- Represent high-dim data in a low-dim space.
- Try to preserve certain properties in the original data.

Non-parametric reduction

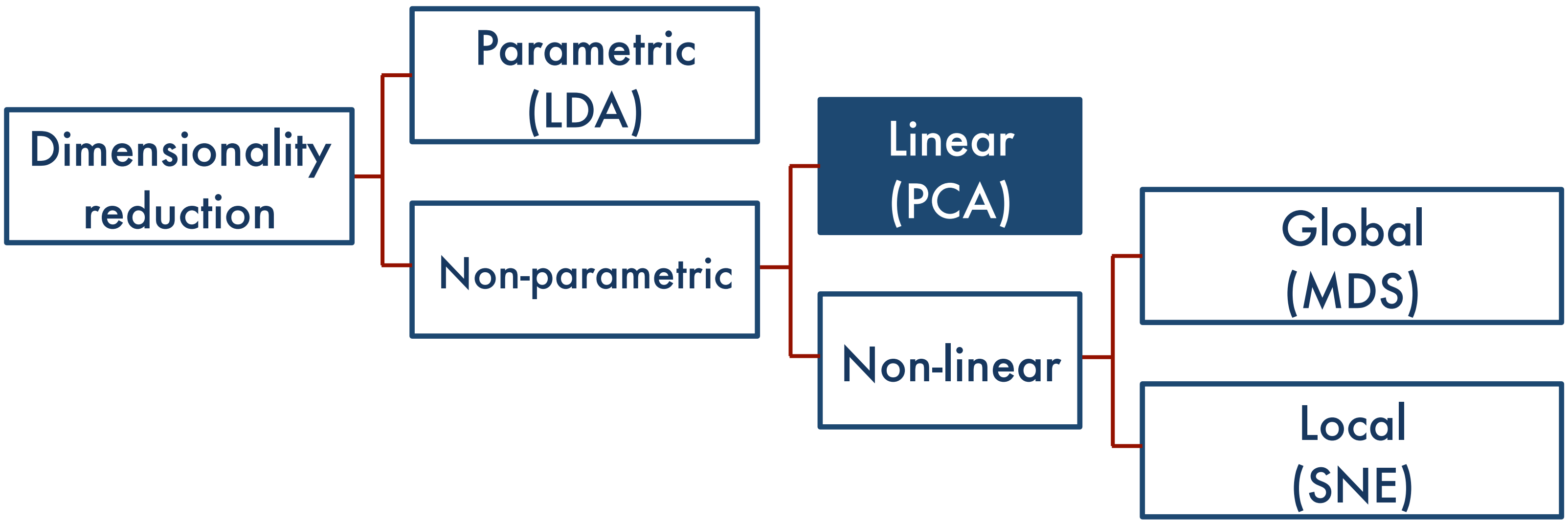
Goals:

- Represent high-dim data in a low-dim space.
- Try to preserve certain properties in the original data.

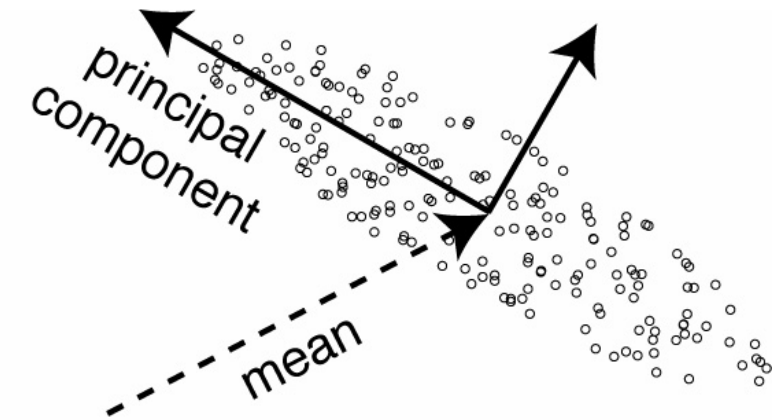
NOT interested in

- Parametric encoding of high-to-low mapping.
- Parametric decoding of low-to-high reconstruction.

Map of the tools

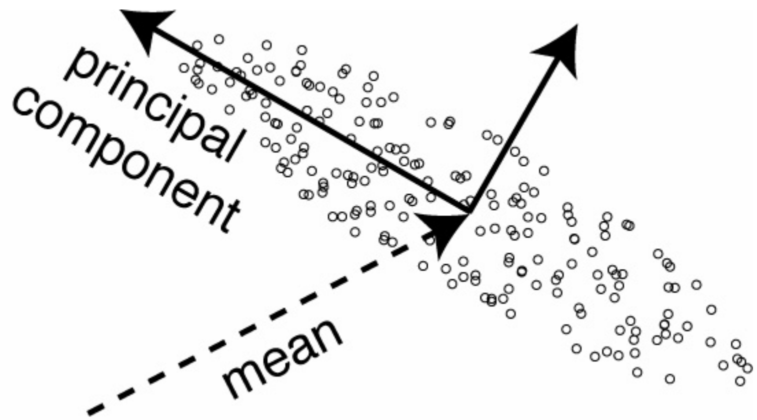


Principal Component Analysis



- For each x, y = f(x) = $V^T x$ (V: n x k linear transform)
- 1st PC: the direction with the maximum variance
- 2nd PC: the orthogonal direction to the 1st PC
  with the maximum residual variance.

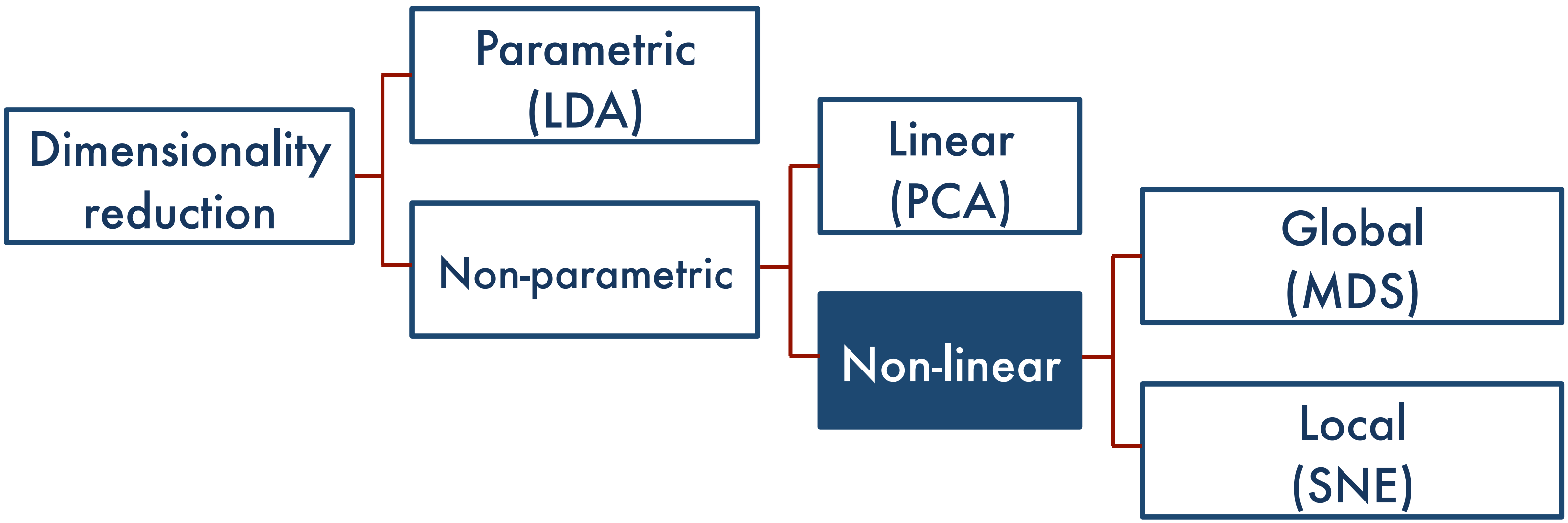
Principal Component Analysis



- For each x, y = f(x) = $V^T x$ (V: n x k linear transform)
- 1^{st} PC: the direction with the maximum variance
- 2^{nd} PC: the orthogonal direction to the 1^{st} PC
  with the maximum residual variance.

Original axes information will be lost!

Map of the tools



Non-linear reduction

Global methods:

Local methods:

Non-linear reduction

Global methods:

- All pairwise distances are equally important.
- Choose a low-dim space considering all pairs.

Local methods:

Non-linear reduction

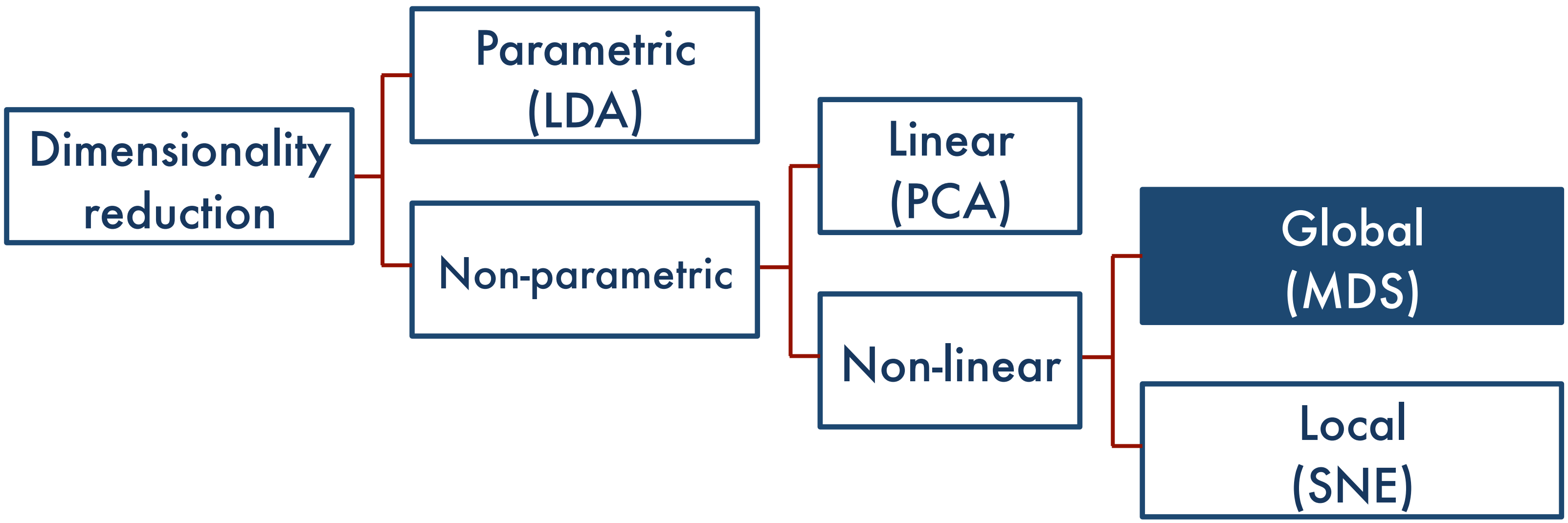
Global methods:

- All pairwise distances are equally important.
- Choose a low-dim space considering all pairs.

Local methods:

- Only local pairwise distances are reliable.
- Put more weights on preserving local pairs.

Map of the tools



Multi-Dimensional Scaling

$$\min Cost = \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2$$

$$d_{ij} = \| x_i - x_j \|^2$$

$$\hat{d}_{ij} = \| y_i - y_j \|^2$$

- d_{ij}: the distance between i and j in the original space.
- đ$_{ij}$: the distance between I and j in the new space.
- By minimizing cost, we preserve all pairwise distances. (penalized if two distances are different)

Multi-Dimensional Scaling

$$\min Cost = \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2$$

$$d_{ij} = \| x_i - x_j \|^2$$

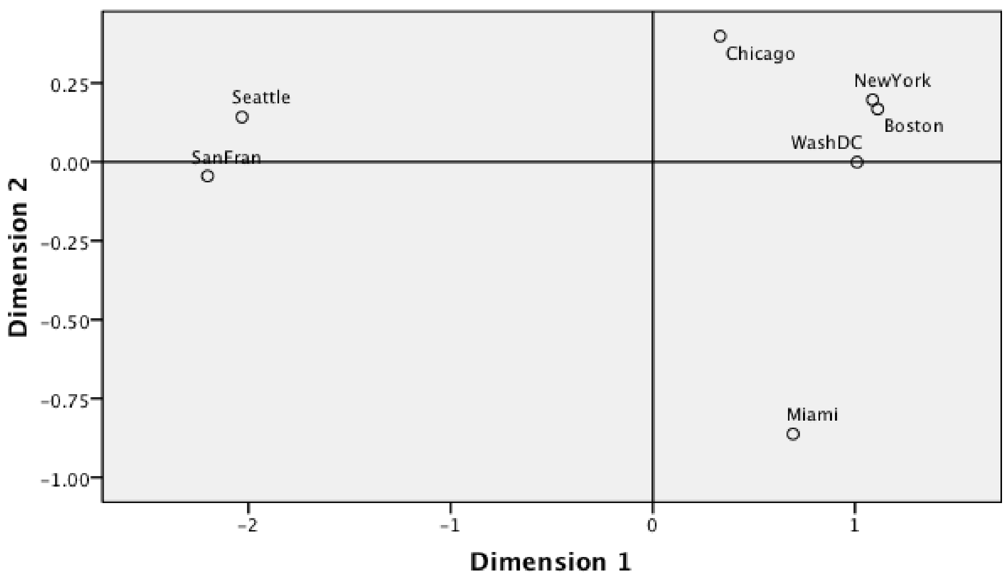
$$\hat{d}_{ij} = \| y_i - y_j \|^2$$

- d_{ij}: the distance between i and j in the original space.
- $\hat{d}_{ij}$: the distance between I and j in the new space.
- By minimizing cost, we preserve all pairwise distances. (penalized if two distances are different)

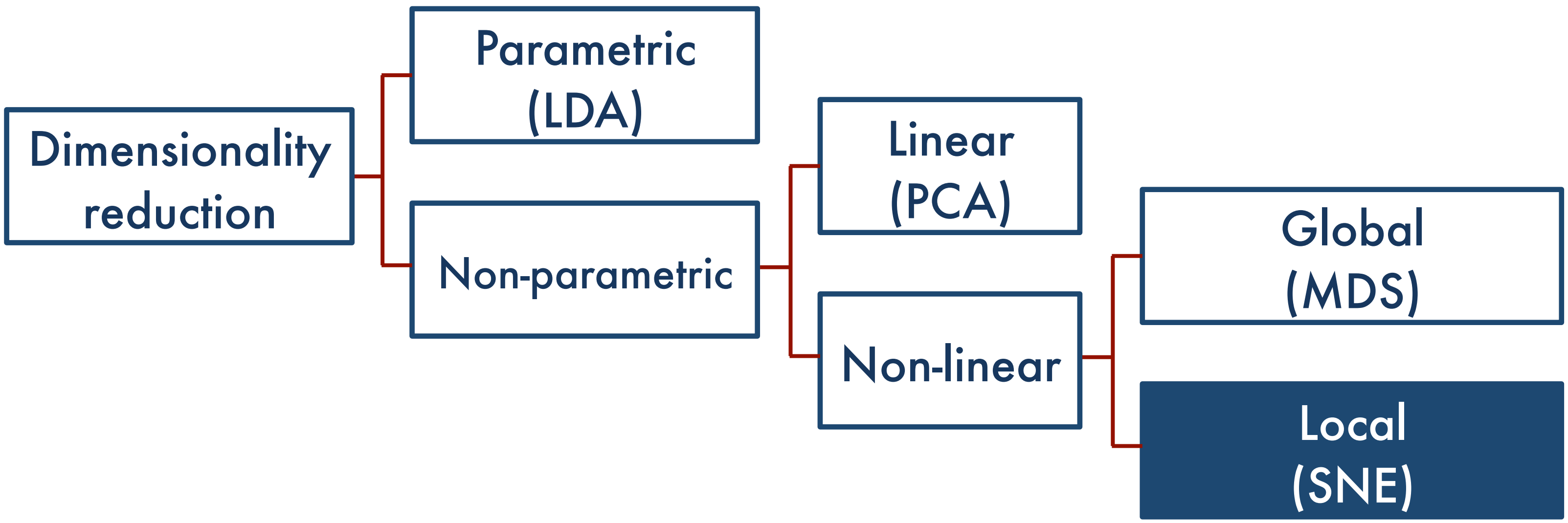
Too careful in far distant points!

What are the dimensions?

		1	2	3	4	5	6	7	8	9
		BOST	NY	DC	MIAM	CHIC	SEAT	SF	LA	DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0



Map of the tools



Stochastic Neighbor Embedding

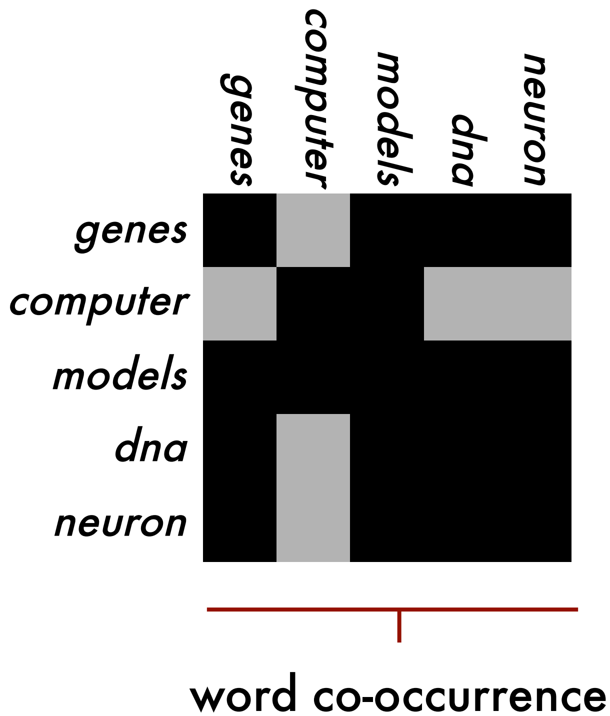
$$\min Cost = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

$$p_{j|i} = \frac{e^{-\|x_i - x_j\|^2 / 2\sigma_i^2}}{\sum_k e^{-\|x_i - x_k\|^2 / 2\sigma_i^2}} \qquad q_{j|i} = \frac{e^{-\|y_i - y_j\|^2}}{\sum_k e^{-\|y_i - y_k\|^2}}$$

- Preserves local distances correctly > infinitesimal ones.
- Probabilistically decide the locality of pairwise distance.
- Define the similarity by probabilistic distance.
  (so called KL-divergence)

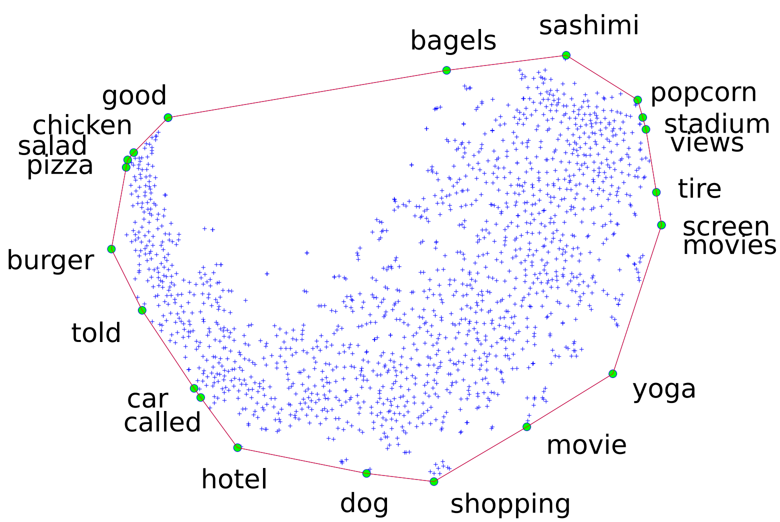
Using t-SNE for topic modeling

Co-occurrence space & anchor words on Yelp [Moontae and David]



Using t-SNE for topic modeling

Co-occurrence space & anchor words on Yelp [Moontae and David]



Better topics | Better interpretability

THANKS
&
ASK ME!

If we still have a time…


Moderato e maestoso. (♩ = 112)
p
cresc.
f
dimin.
cresc.
f
dimin.


Moderato e maestoso. (♩ = 112)
p
cresc.
f
dimin.
cresc.
f
dimin.


Moderato e maestoso. (♩ = 112)
p
cresc.
f
dimin.
cresc.
f
dimin.


Moderato e maestoso. (♩ = 112)
p
cresc.
f
dimin.
cresc.
f
dimin.


Moderato e maestoso. (♩ = 112)
p
cresc.
f
dimin.
cresc.
f
dimin.


dimin.
cresc.


dimin.
cresc.


f
dimin.
p
cresc.
f