Natural Language Processing

Lecture 11

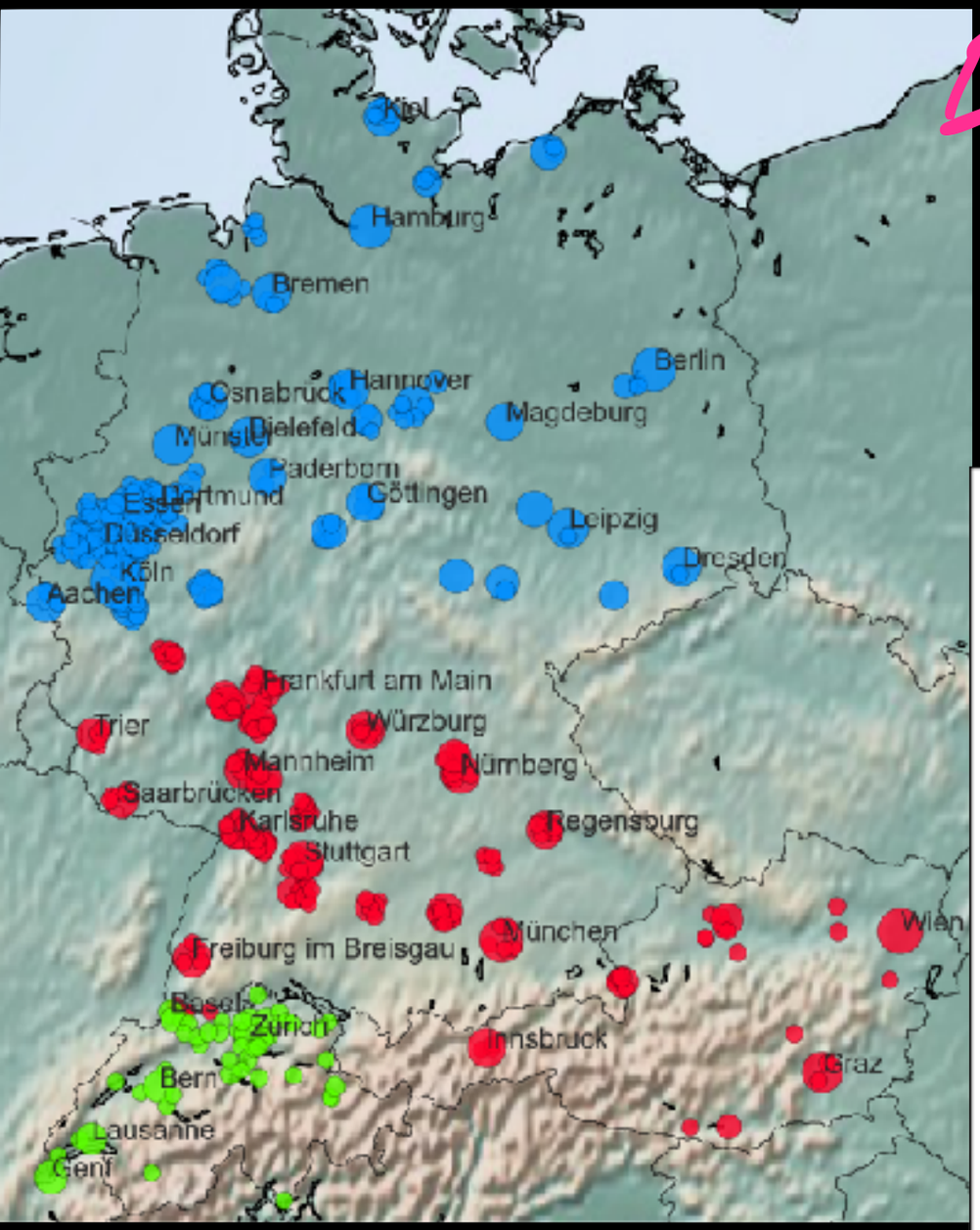
Dirk Hovy

dirk.hovy@unibocconi.it

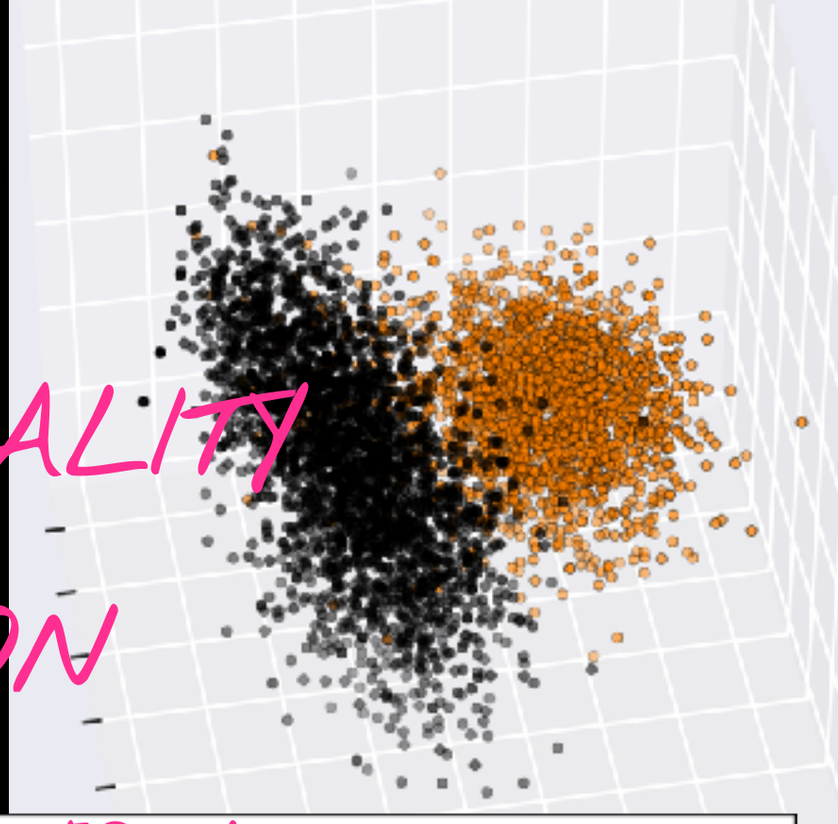
@dirk_hovy

Bocconi

Examples

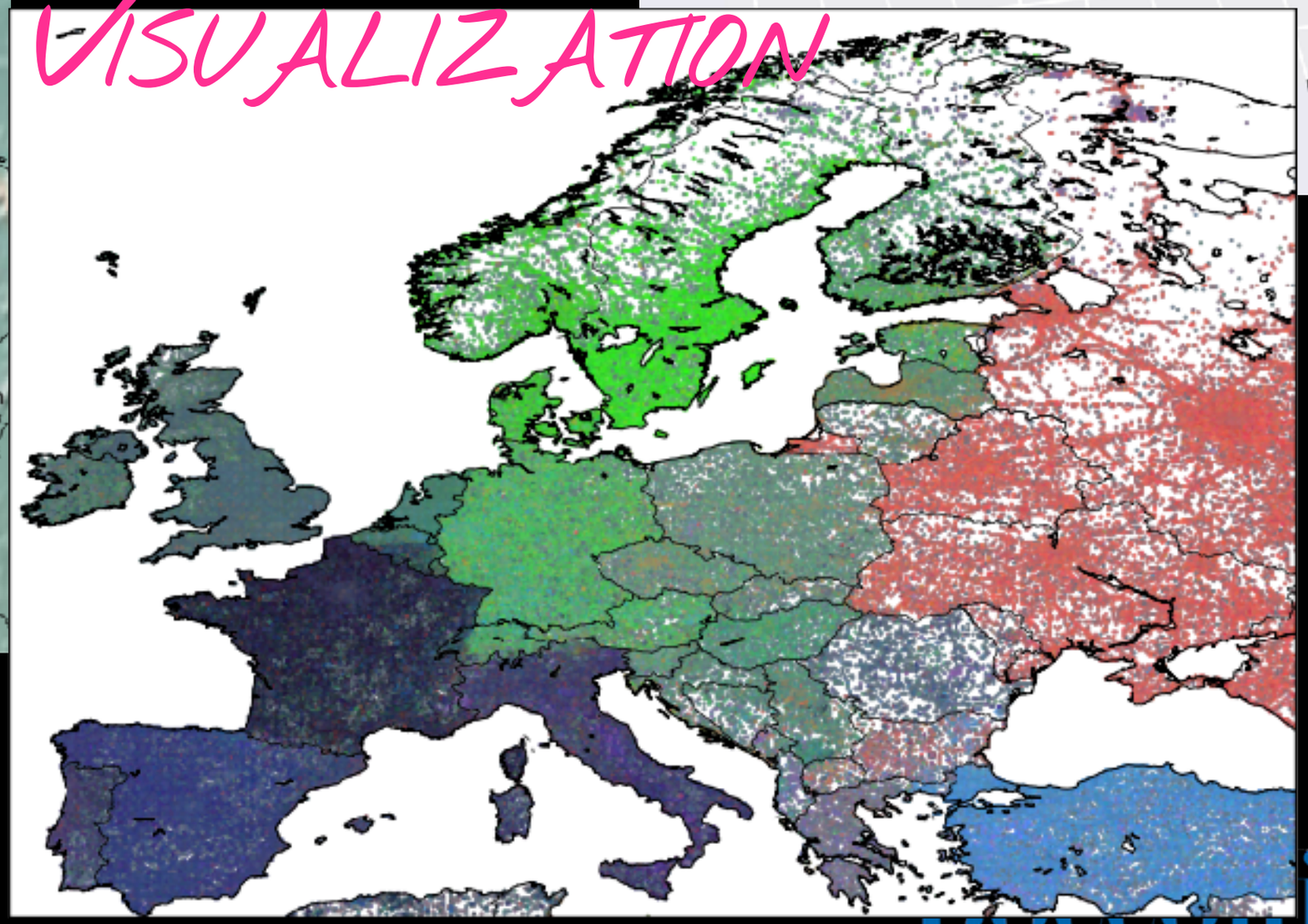


Clustering

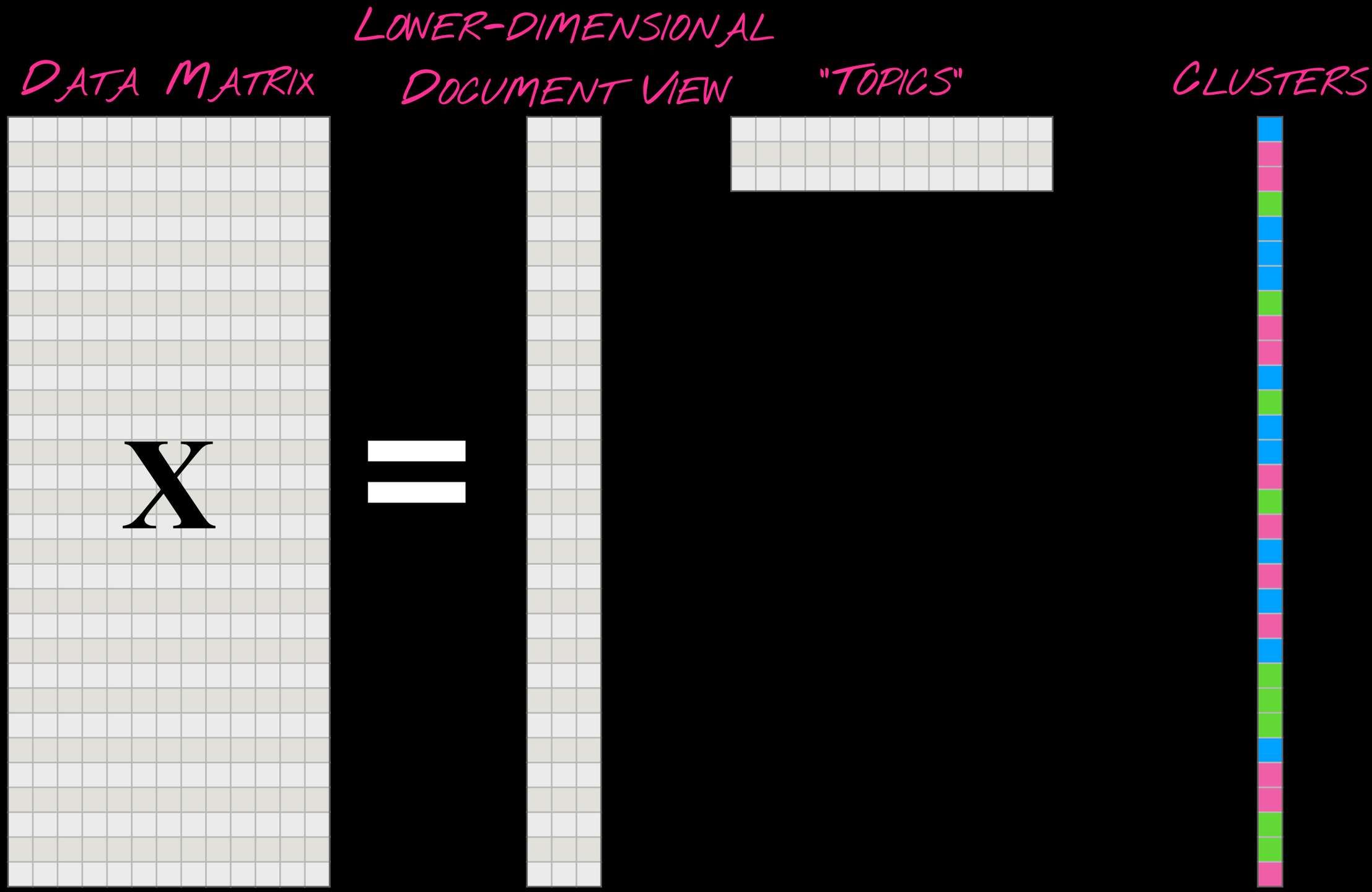


Dimensionality Reduction

Visualization



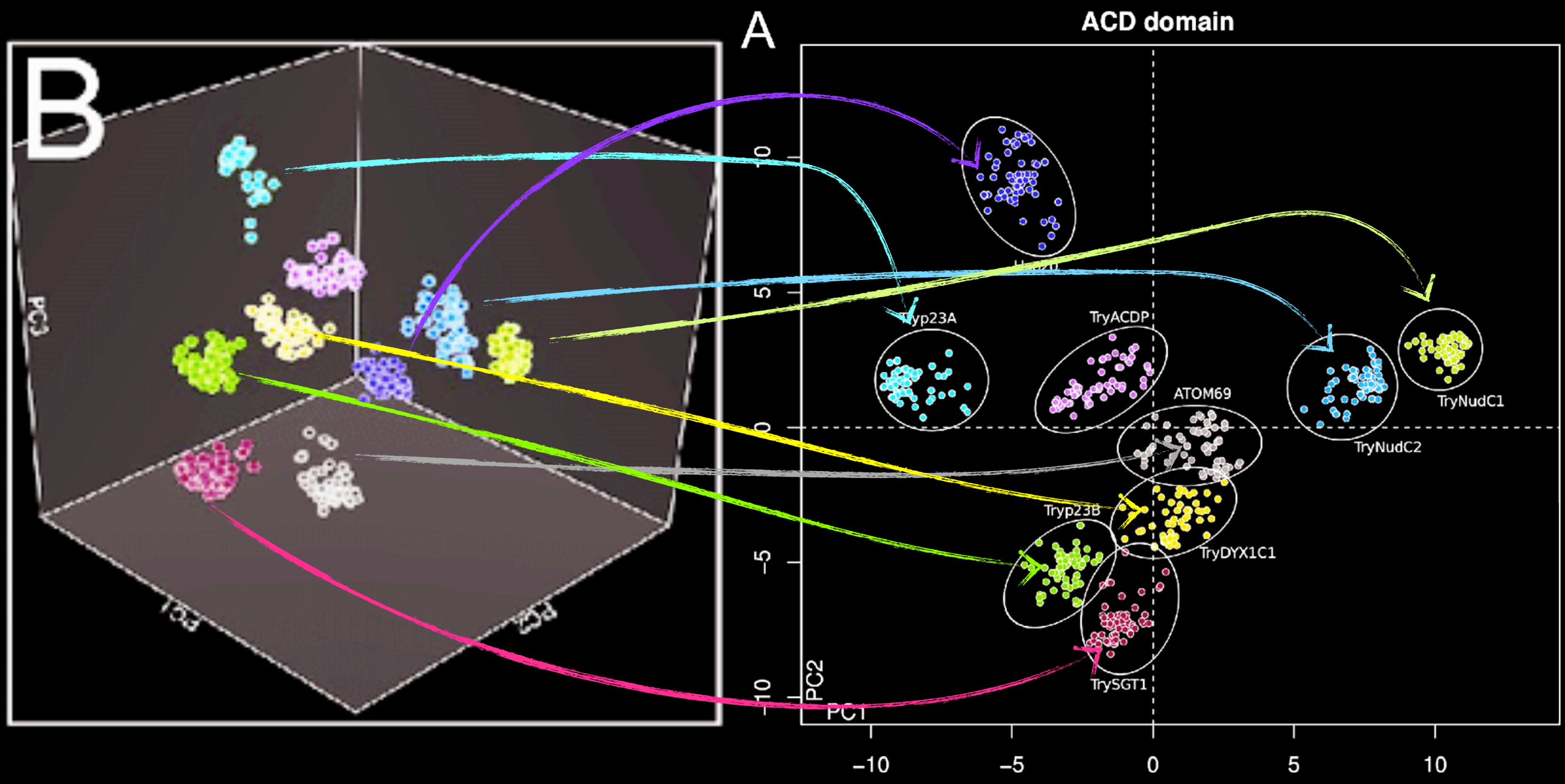
Latent Dimensions



Bocconi

Costa Martins et al. (2018)

Latent Dimensions



Bocconi

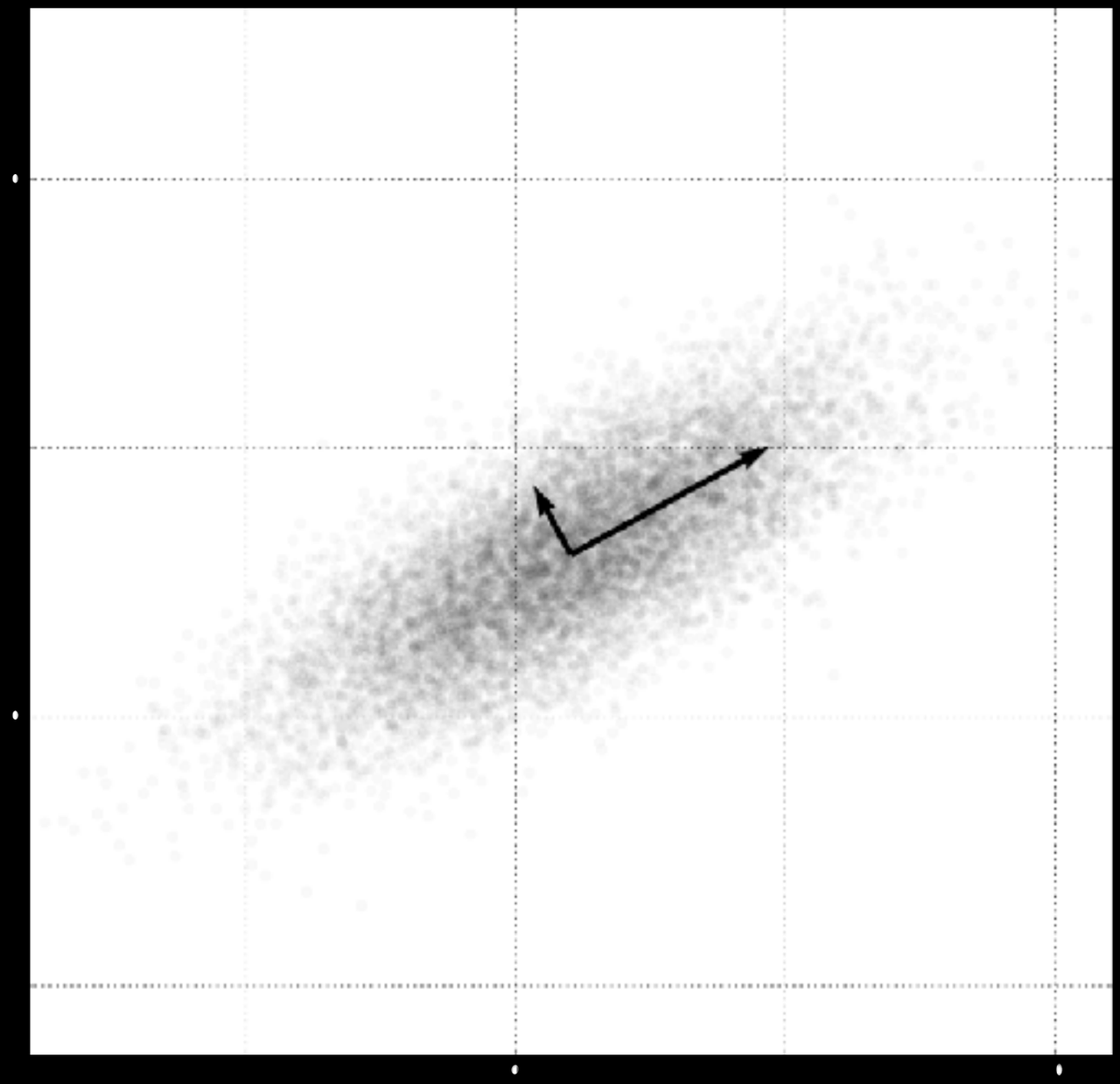
Goals for Today

- Learn about **matrix factorization** and its use for **semantic similarity** and **visualization**
- Learn about ***k*-means** and **agglomerative clustering**
- Learn about **evaluation** criteria

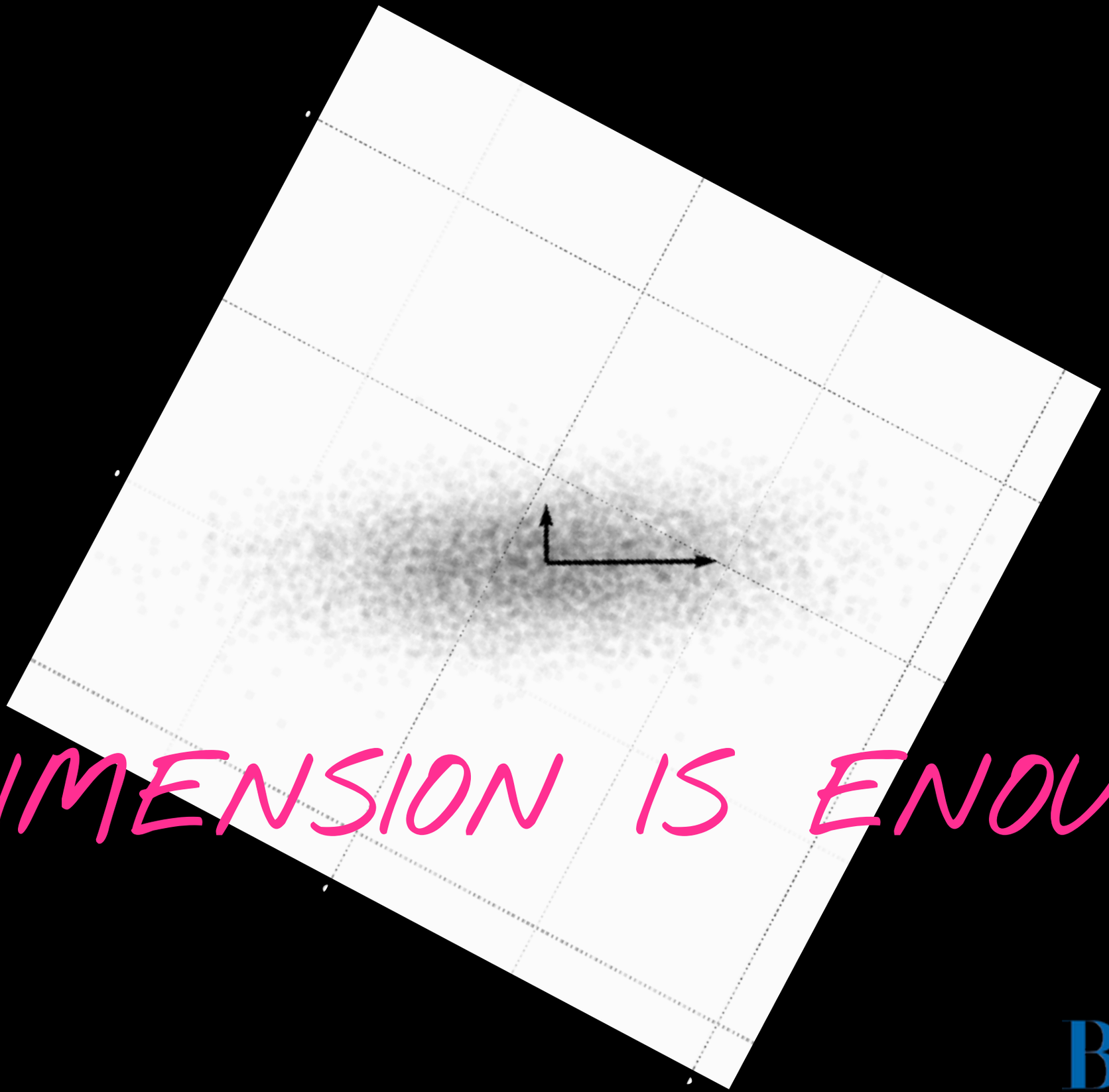
Matrix Factorization

Bocconi

Singular Value Decomposition



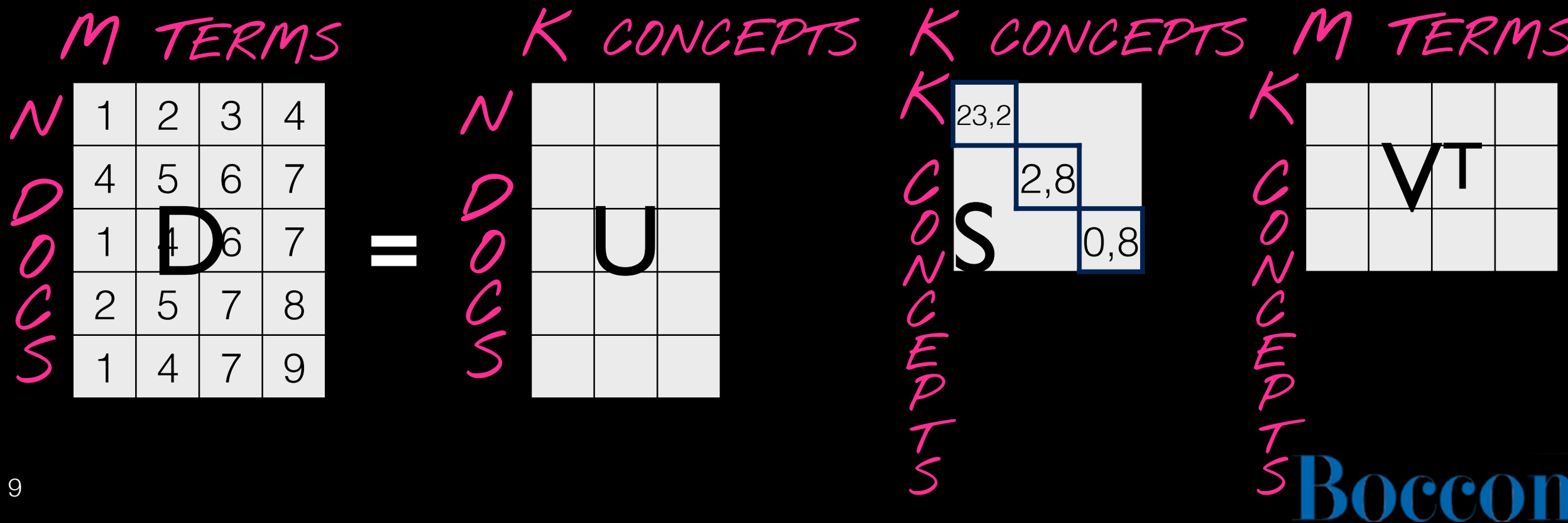
Singular Value Decomposition



I DIMENSION IS ENOUGH!

Singular Value Decomposition

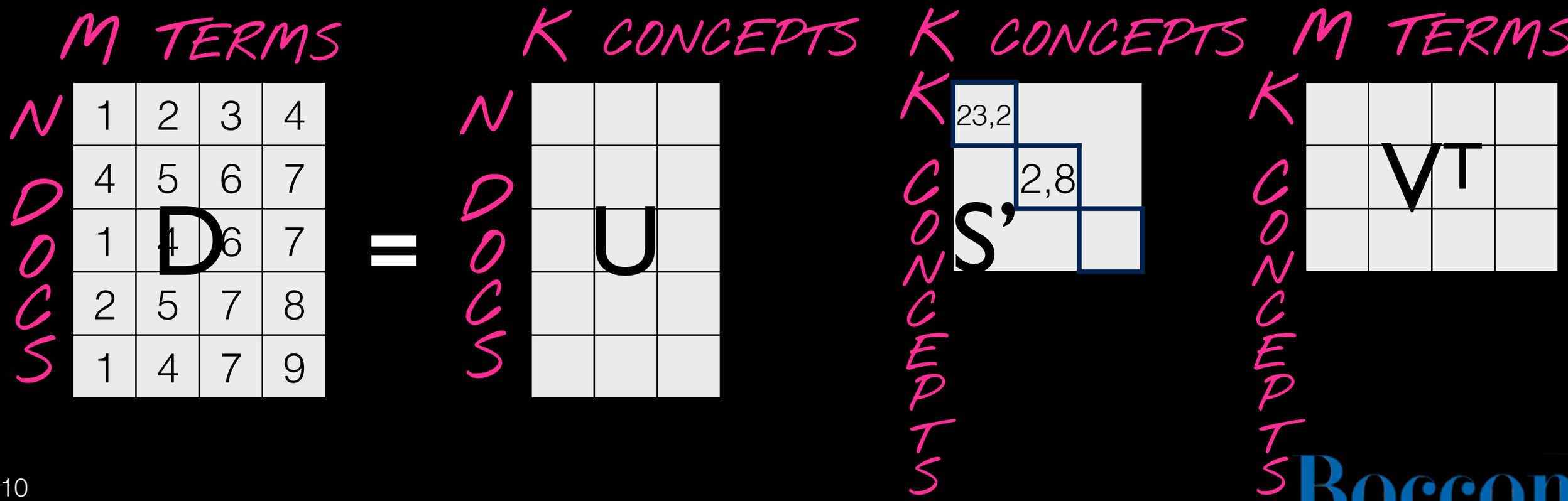
- “principal component analysis”: discover the dimensions that matter
- idea: matrix is made up of few hidden dimensions
- Dimensions correspond to **documents**, **terms**, and latent **concepts**



Bocconi

Singular Value Decomposition

- reduce principal components/concepts to smaller number



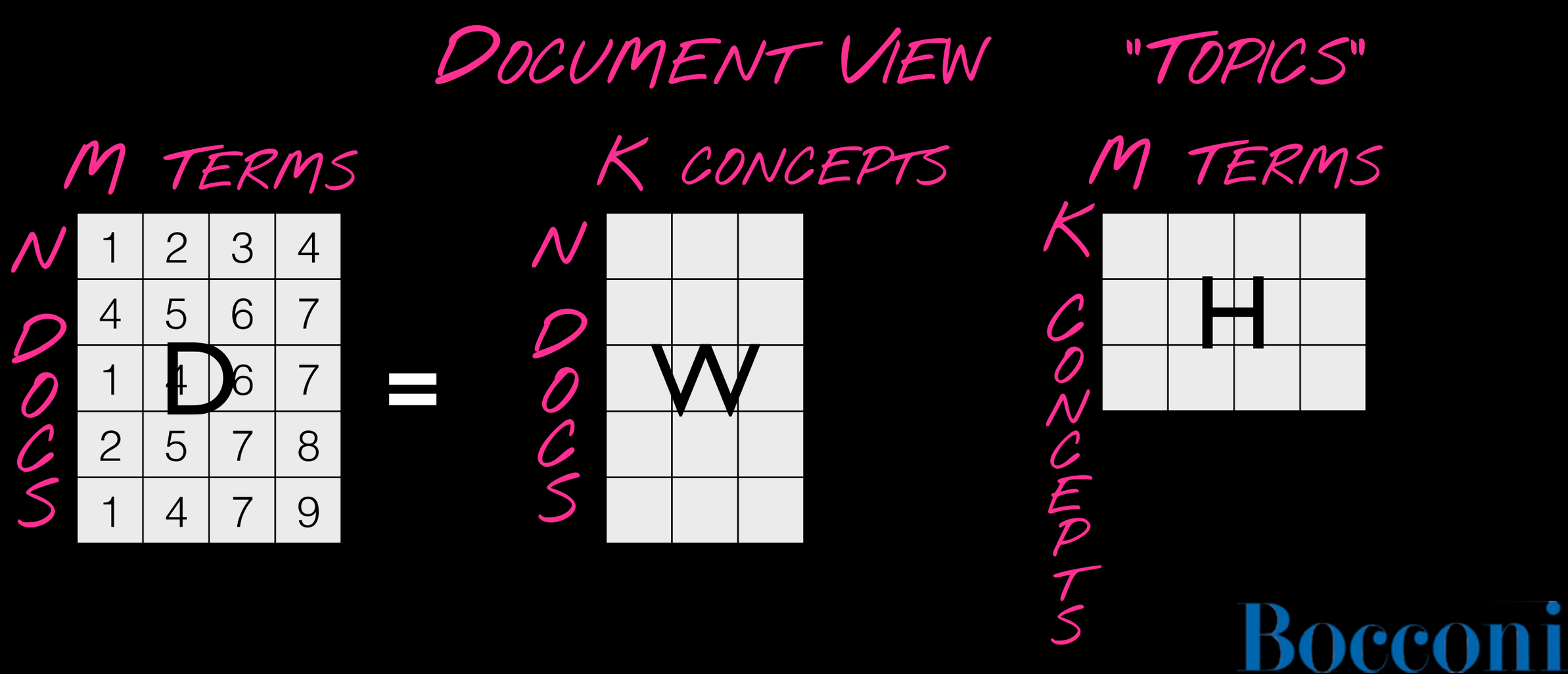
Singular Value Decomposition

- reconstruct original matrix in new concept space:
  Latent Semantic Analysis

Bocconi

Non-negative Matrix Factorization

- Use only positive values
- Find approximation of two components



Bocconi

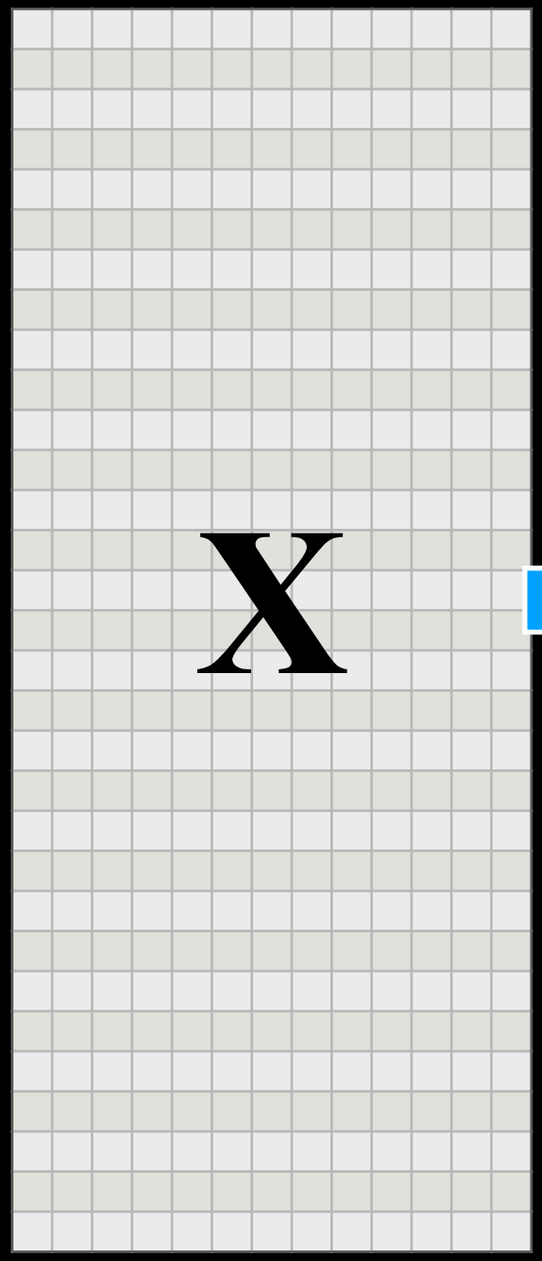
Comparison

	SVD	NMF
Negative values (embeddings) as input?	yes	no
#components	3: U, S, V	2: W, H
document view?	yes: U	yes: W
term view?	yes: V	yes: H
strength ranking?	yes: S	no
exact?	yes	no
"topic" quality	mixed	better
sparsity	low	medium

Yes, but: What is it Good for?

- Find latent **topic** dimensions (alternative: LDA)
- Find **word similarity** in latent space (alternative: Word2Vec)
- Find **document similarity** in latent space (alternative: Doc2Vec)
- Reduce dimensionality for **visualization**

Latent Word Dimension Topics



NMF(10) FOR MOBY DICK

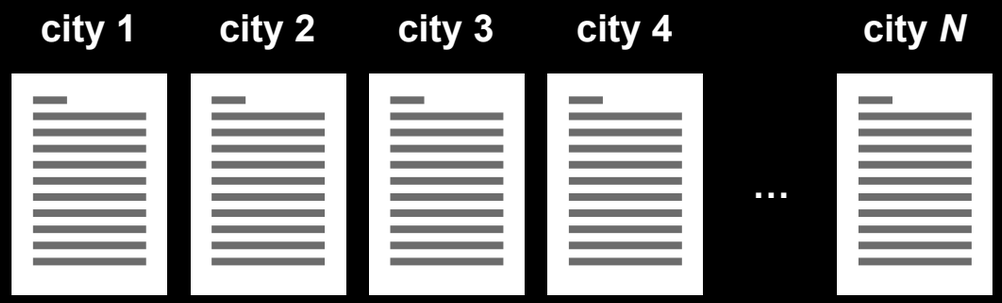
AHAB	ahab, captain, cried, captain ahab, cried ahab
STRUCTURE	chapter, folio, octavo, ii, iii
???	like, ship, sea, time, way
MEN	man, old, old man, look, young man
???	oh, life, starbuck, sweet, god
CHARACTERS	said, stubb, queequeg, don, starbuck
???	sir, aye, let, shall, think
OLD-TIMEY	thou, thee, thy, st, god
WHALES	whale, sperm, sperm whale, white, white whale
???	ye, look, say, ye ye, men

Dimensionality Reduction for Visualizations

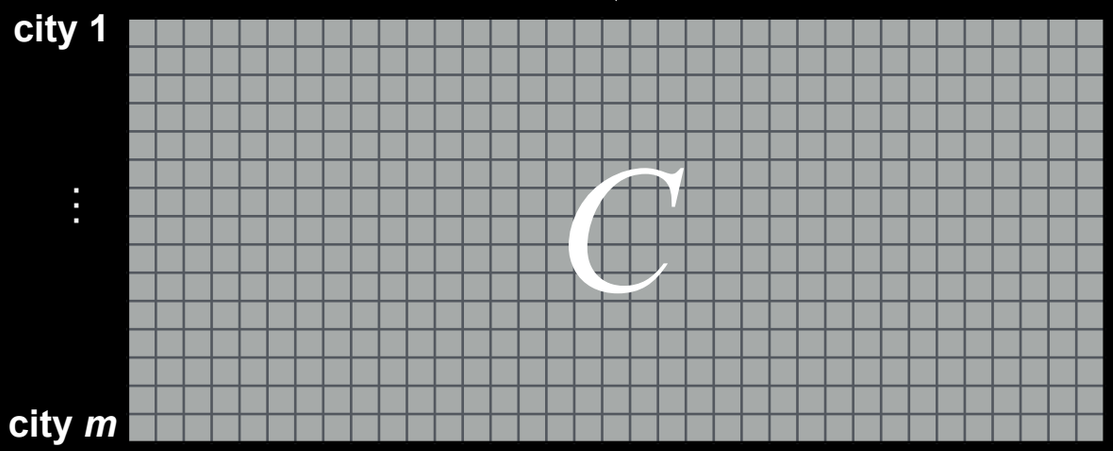
Bocconi

Hovy & Purschke, 2018

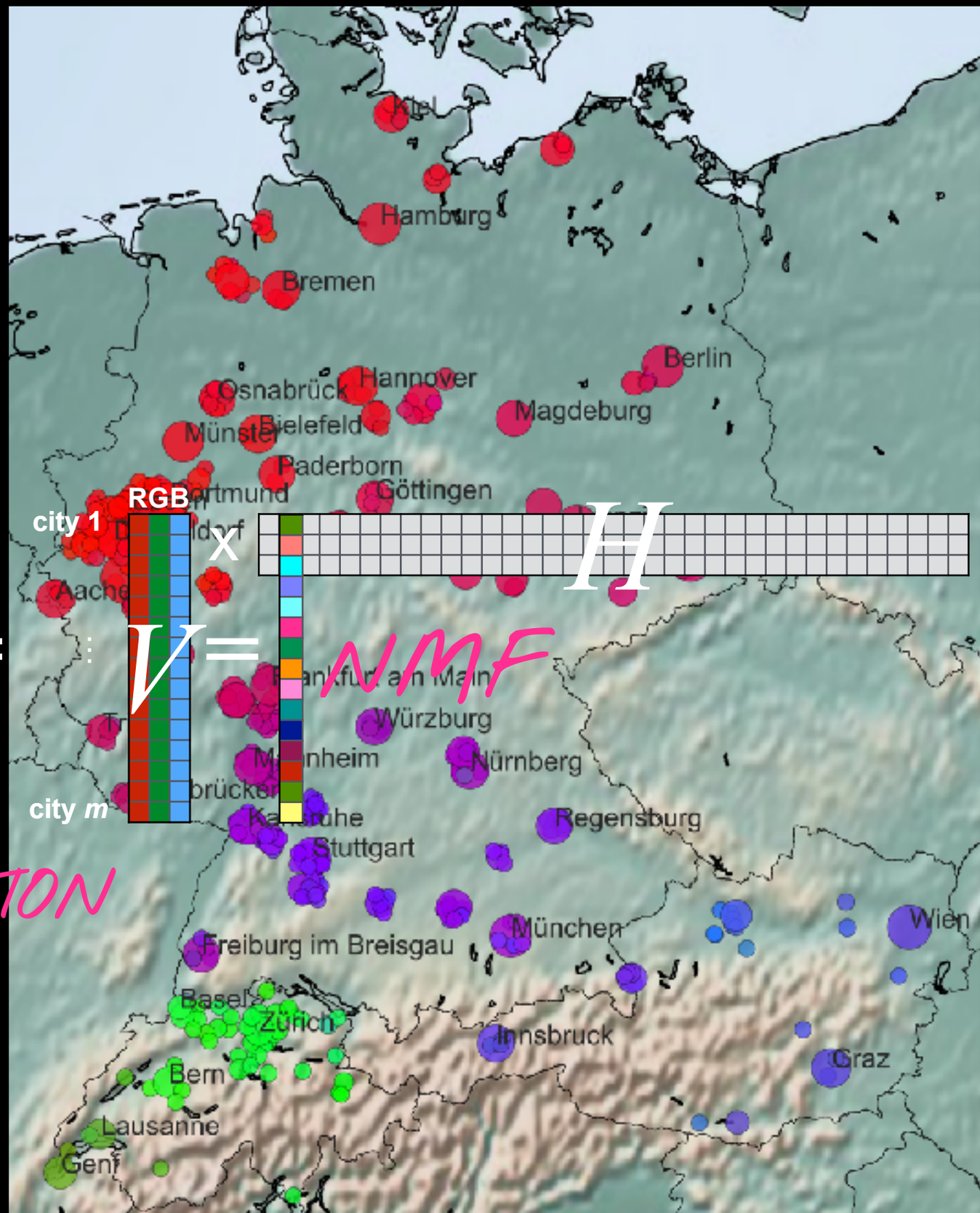
Dimensions as RGB



Doc2Vec



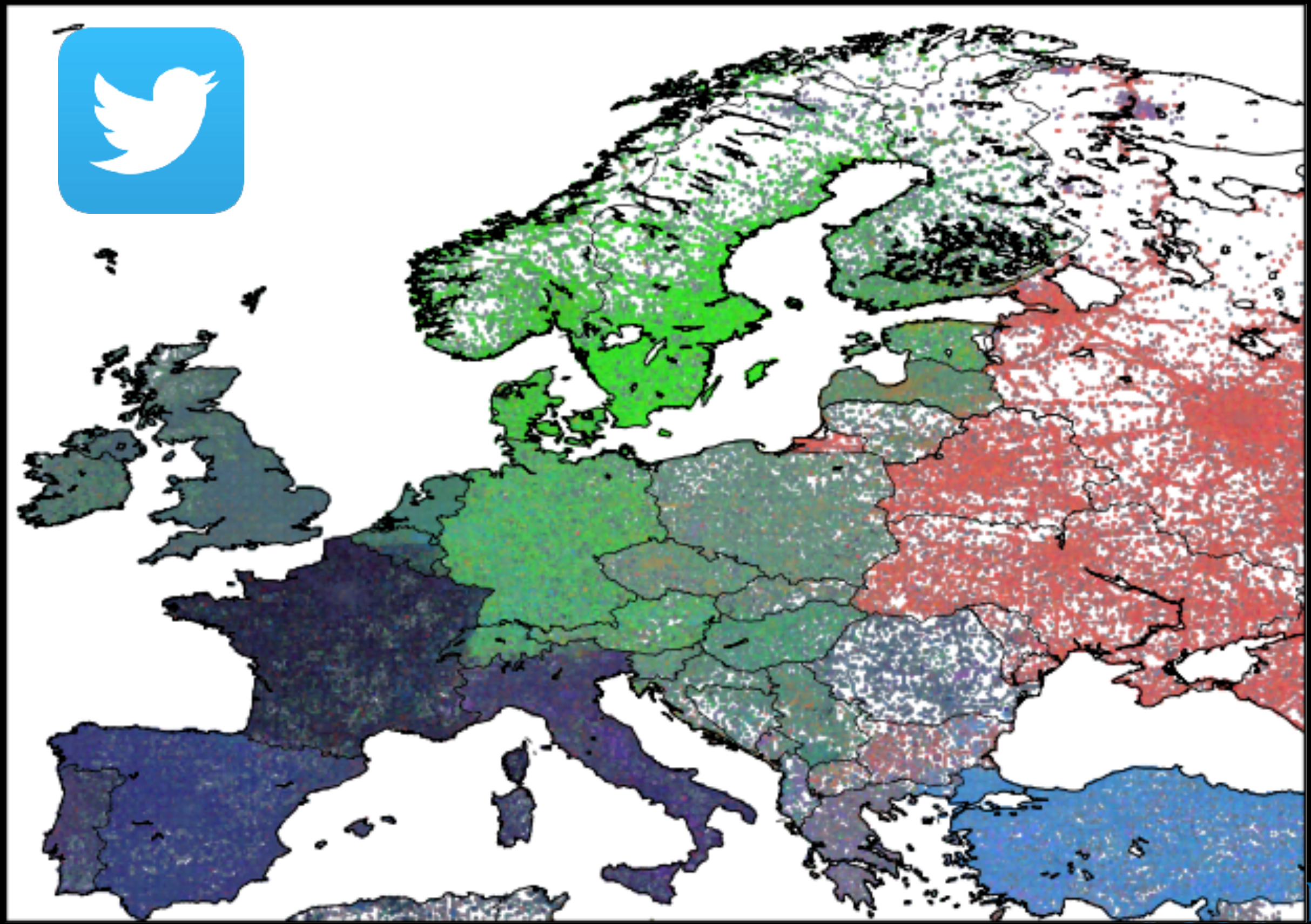
=



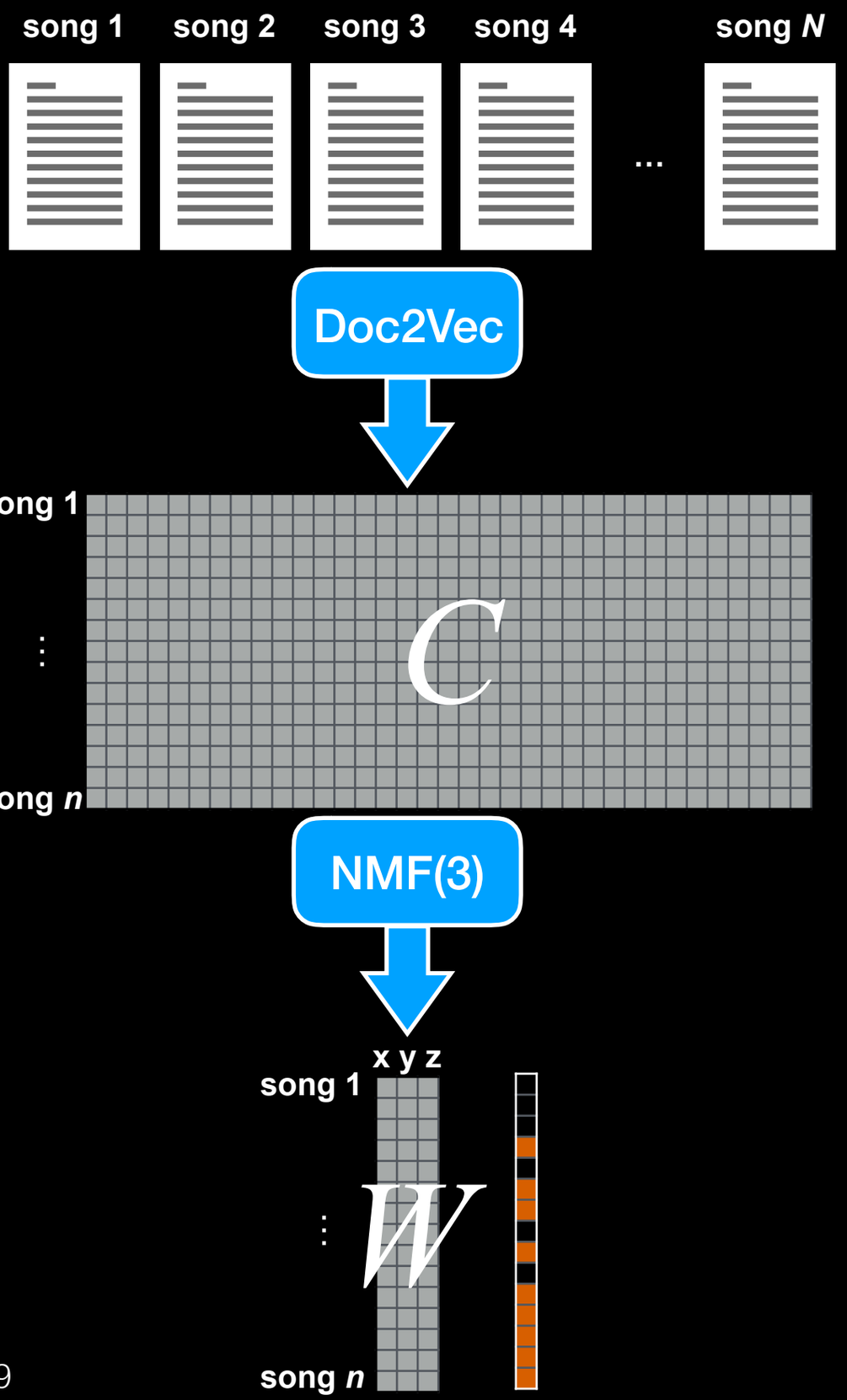
Dense Representation

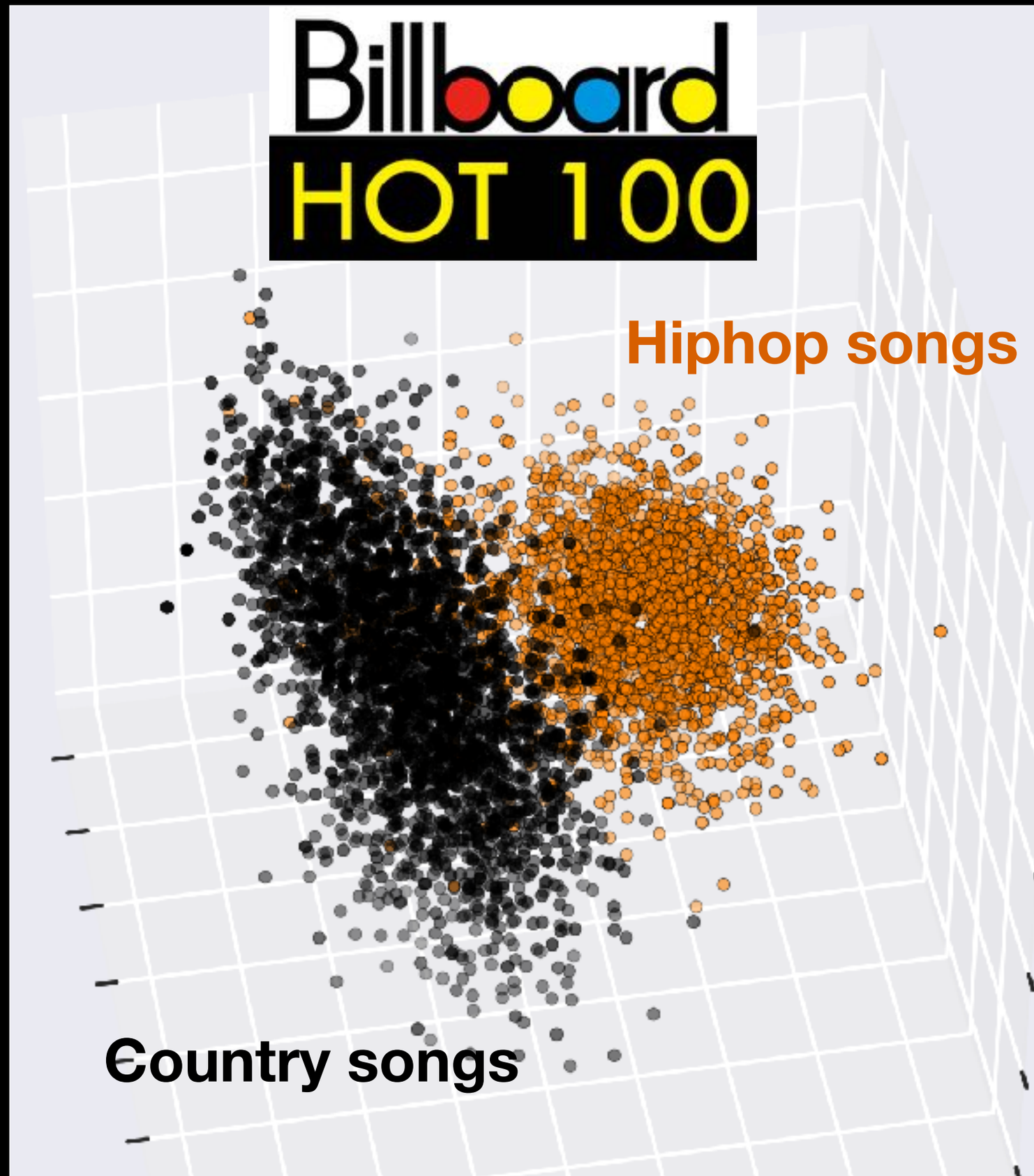
Hovy et al., 2019

Dimensions as RGB



Dimensions as Coordinates





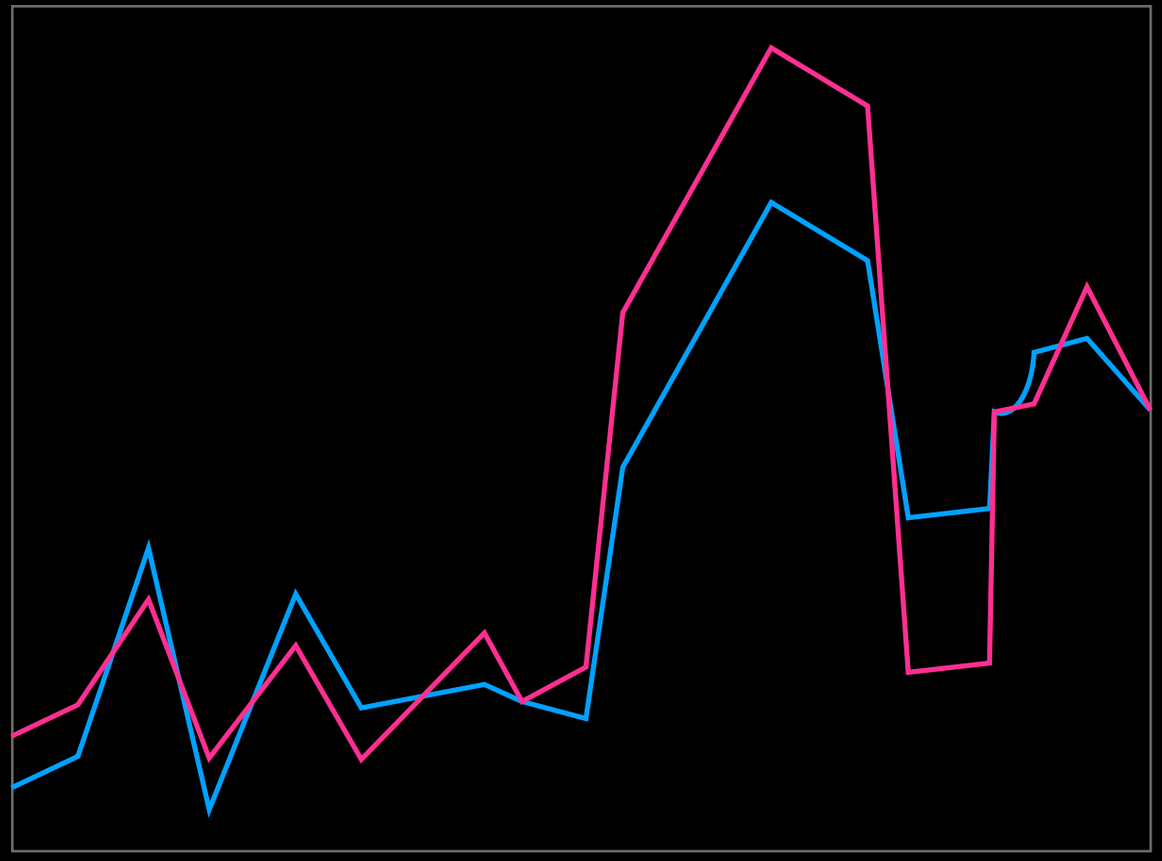
Van der Maaten & Hinton, 2008

t-SNE

- Map/preserve neighborhood structure of high-dimensional space in lower-dimensional space

- Minimize difference between probability distributions over neighbors in both dimensions

$P(n_j|n_i)$

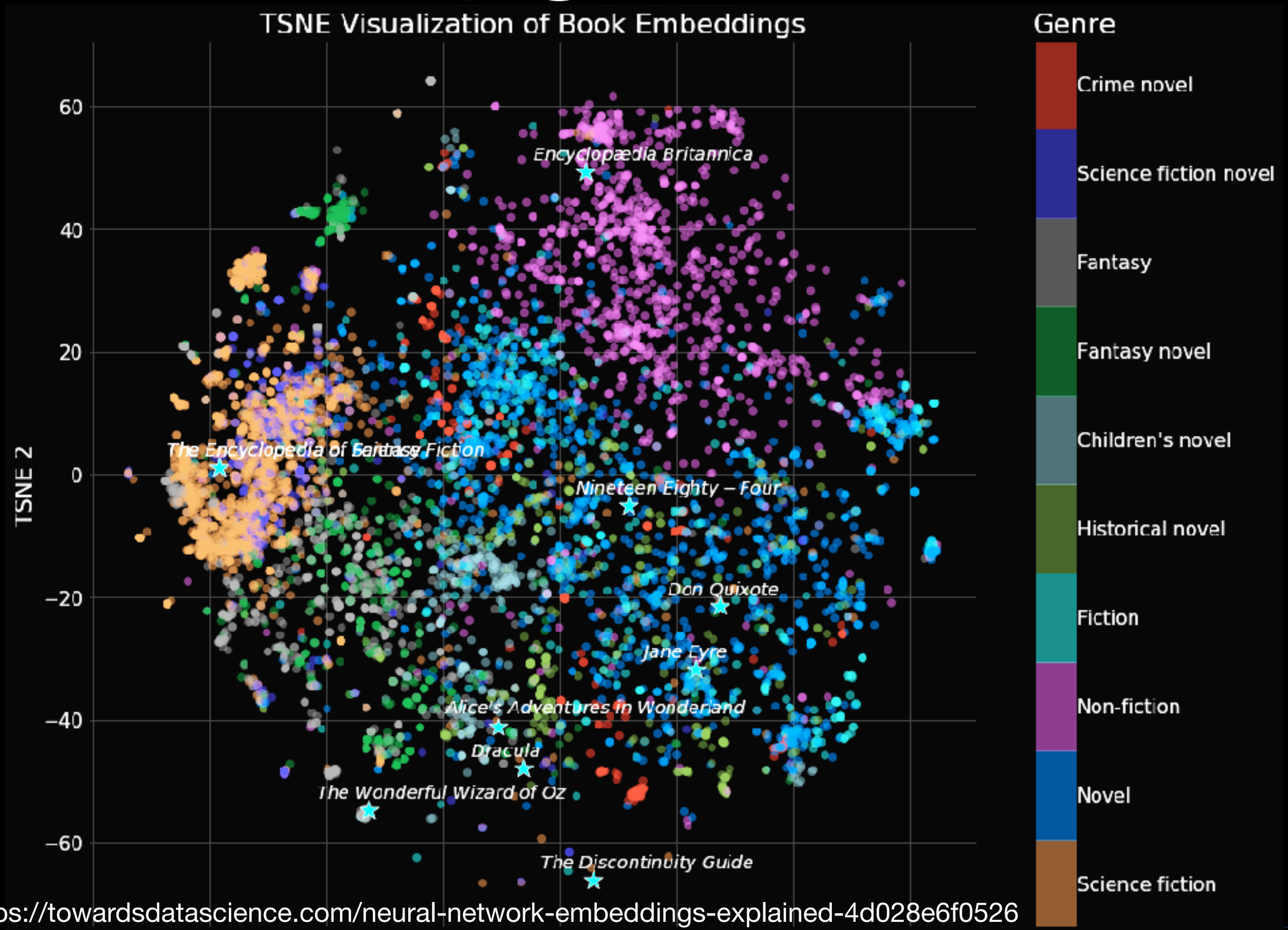


+ Nicer spatial images

– Stochastic (many runs)

– Many parameters

t-SNE



https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526

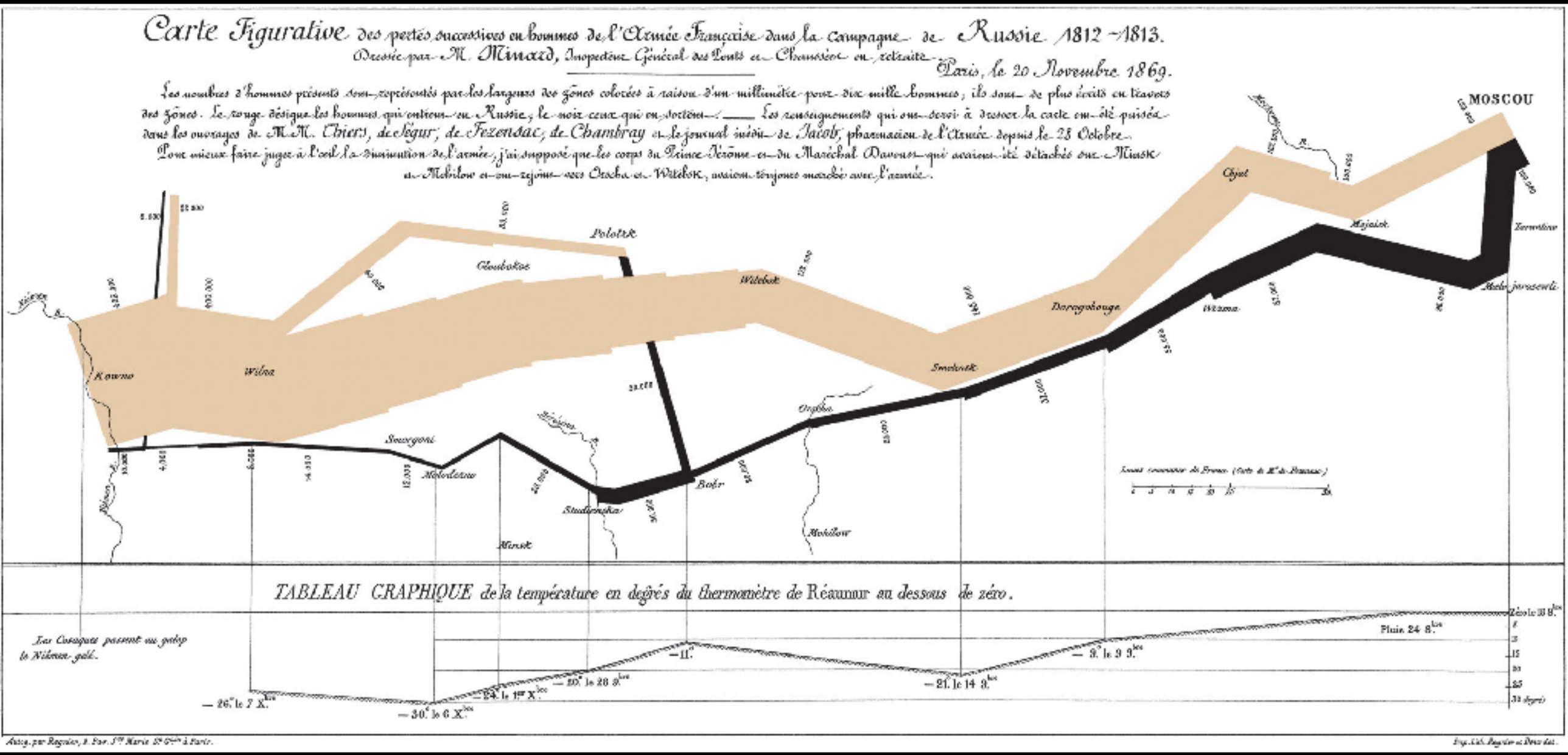
Data Visualization

Bocconi

How Vizs Help

- Goal: get a sense of the data
- understand the type of distribution
- spot outliers
- see (and then test) possible correlations
- ***always*** visualize your data!

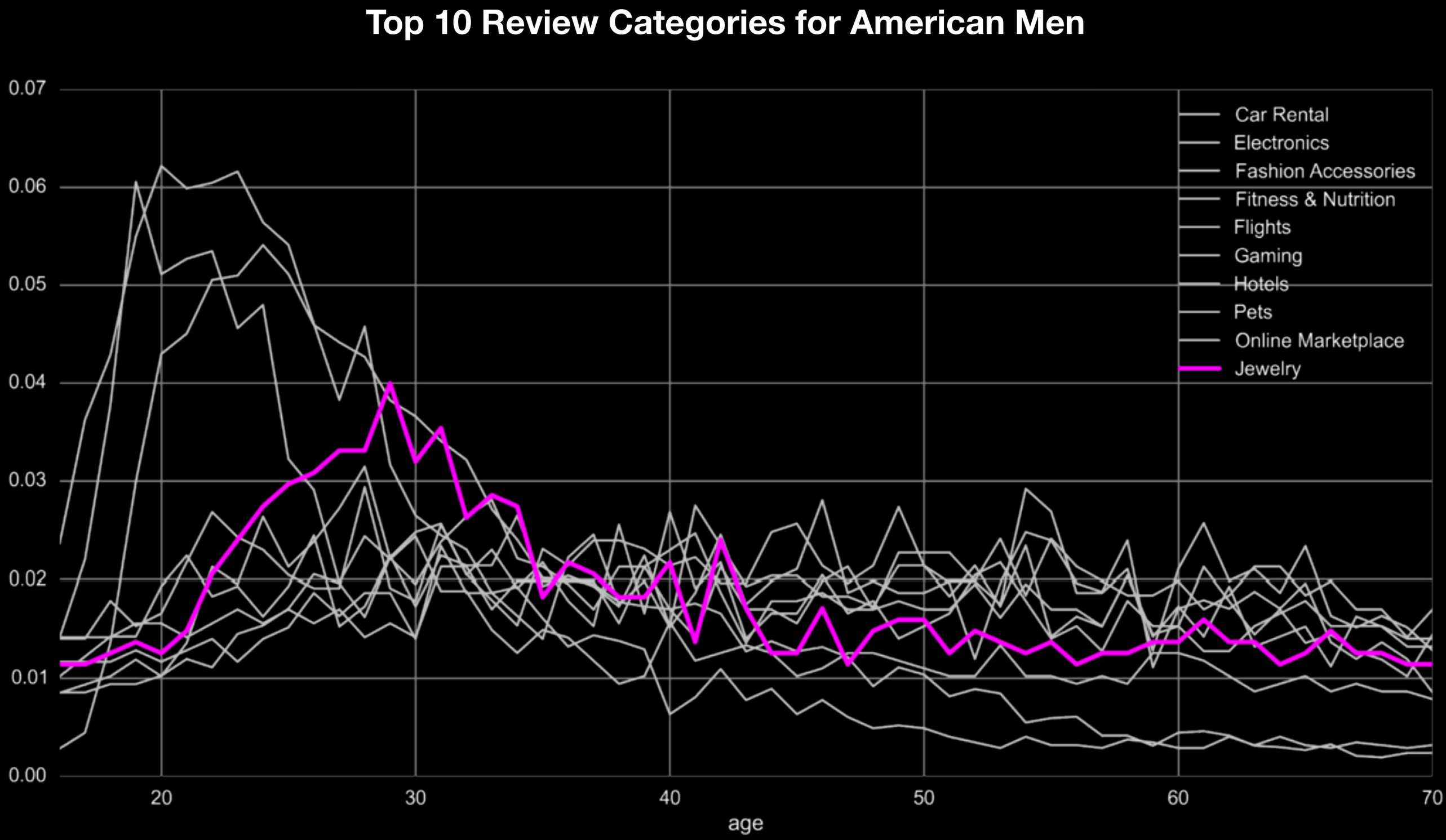
Communicate Insights & Patterns



Charles Joseph Minard: Napoleon losses 1812

Bocconi

Workhorses: Line Charts

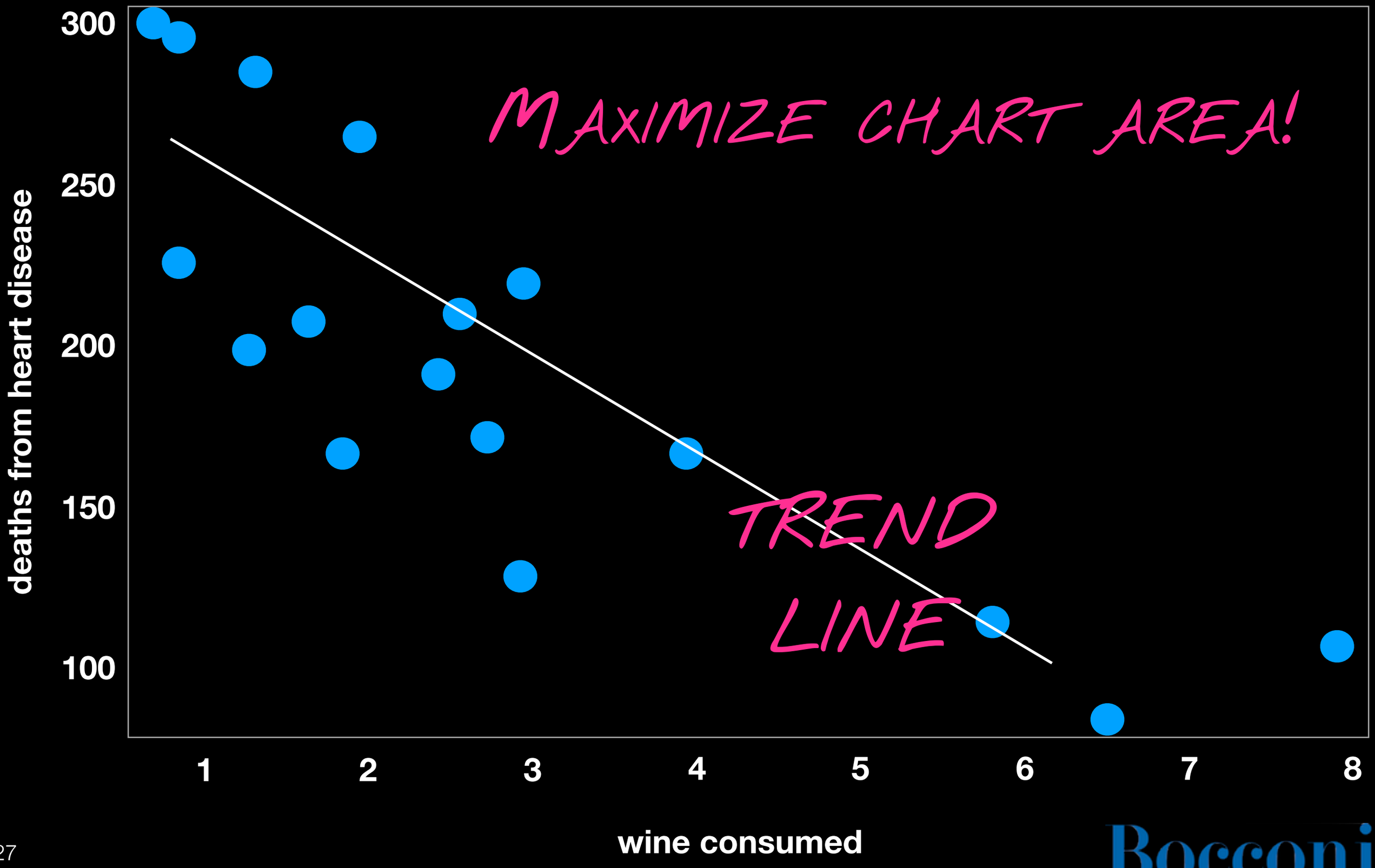


Bocconi

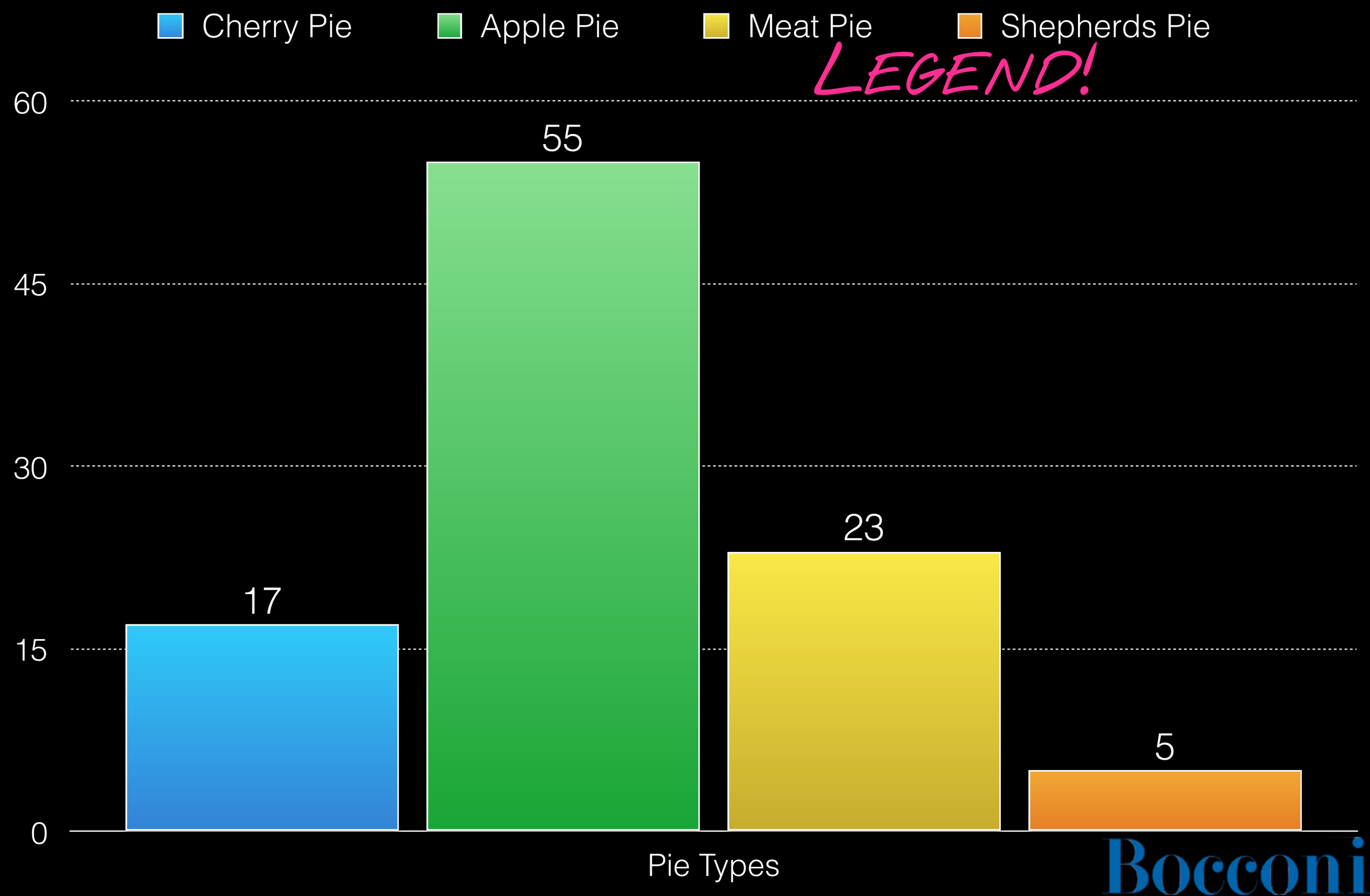
Caveat: False Leads!

Bocconi

Workhorses: Scatter Plot

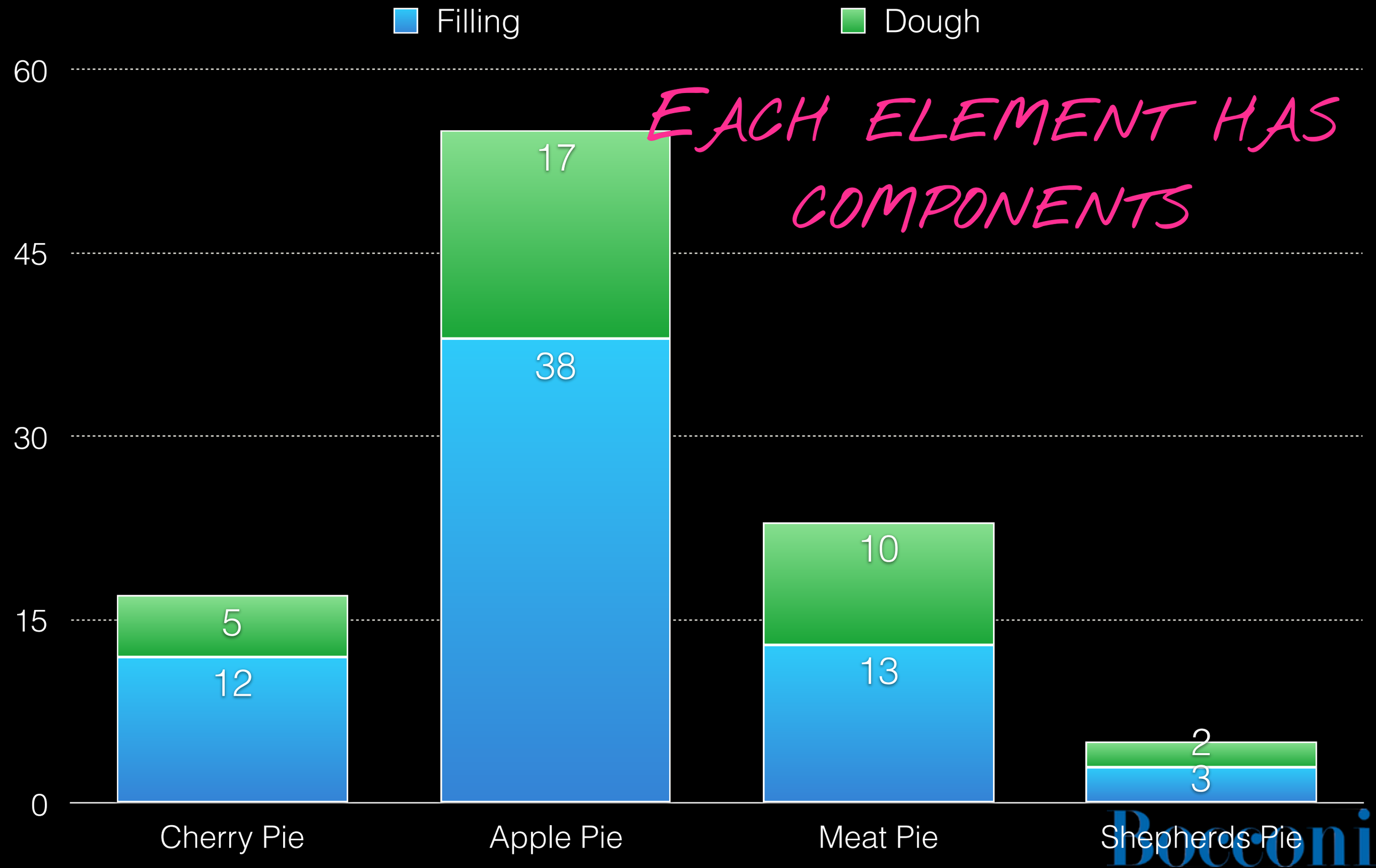


Workhorses: Bar Charts



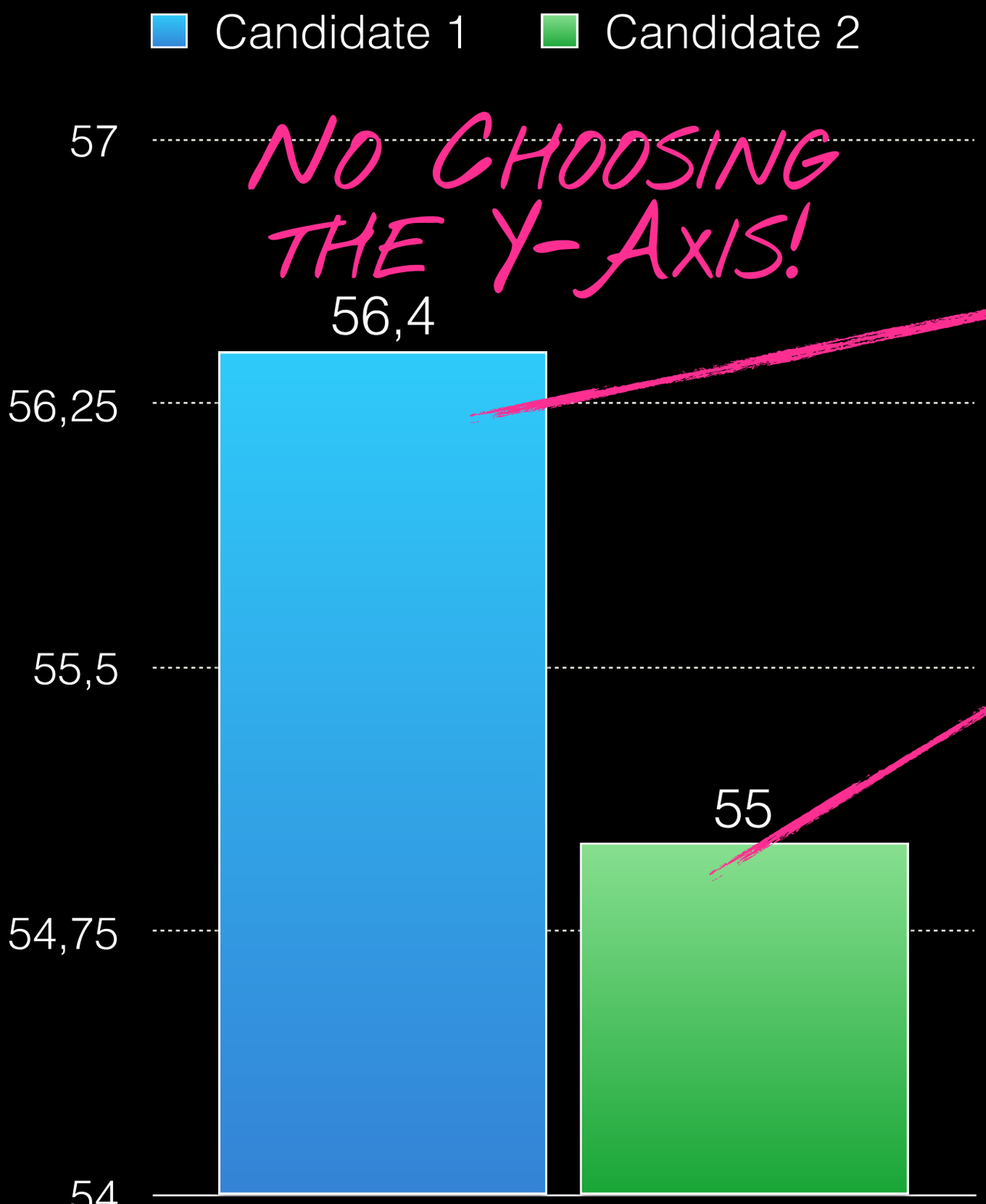
Bocconi

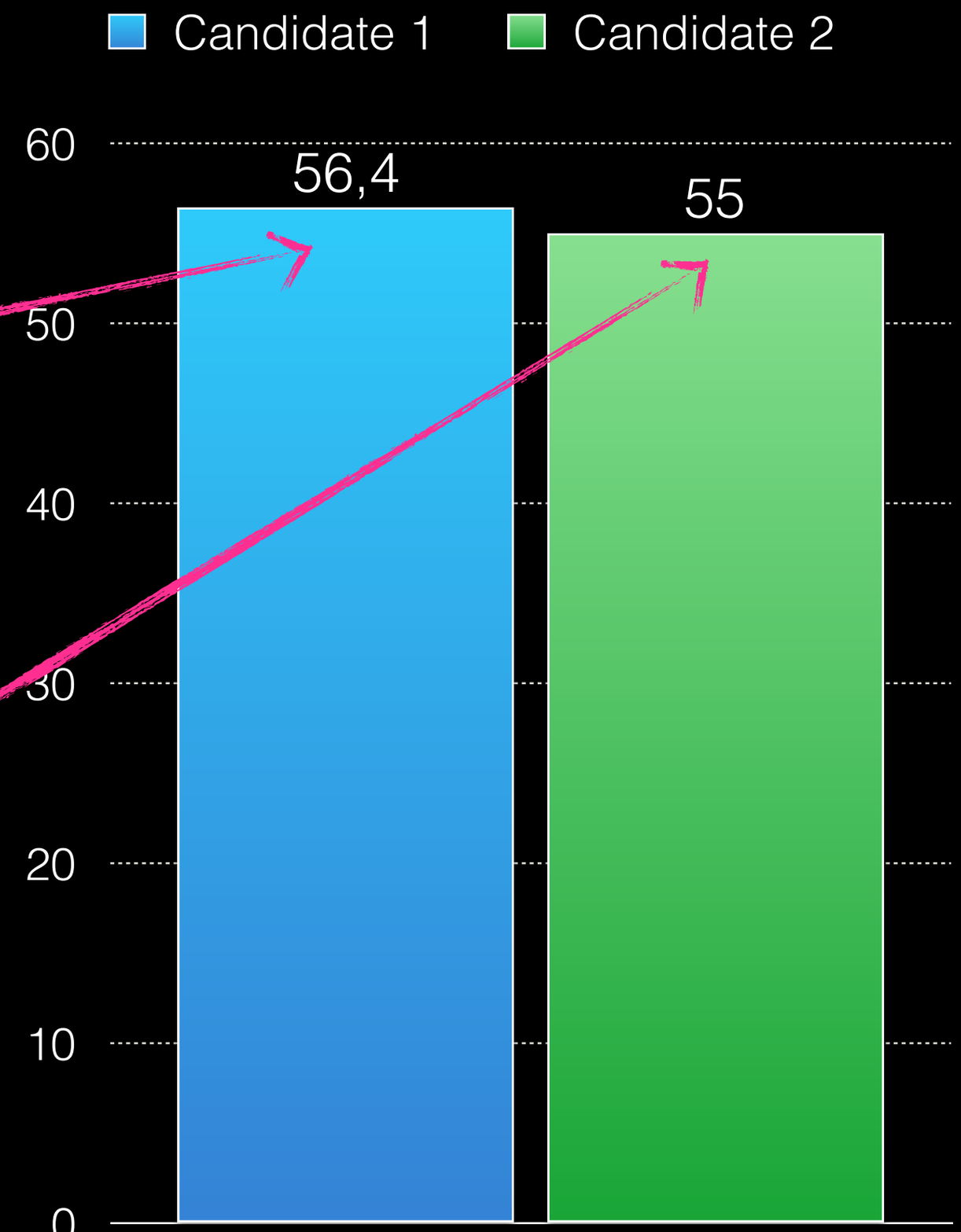
Workhorses: Bar Charts



EACH ELEMENT HAS COMPONENTS

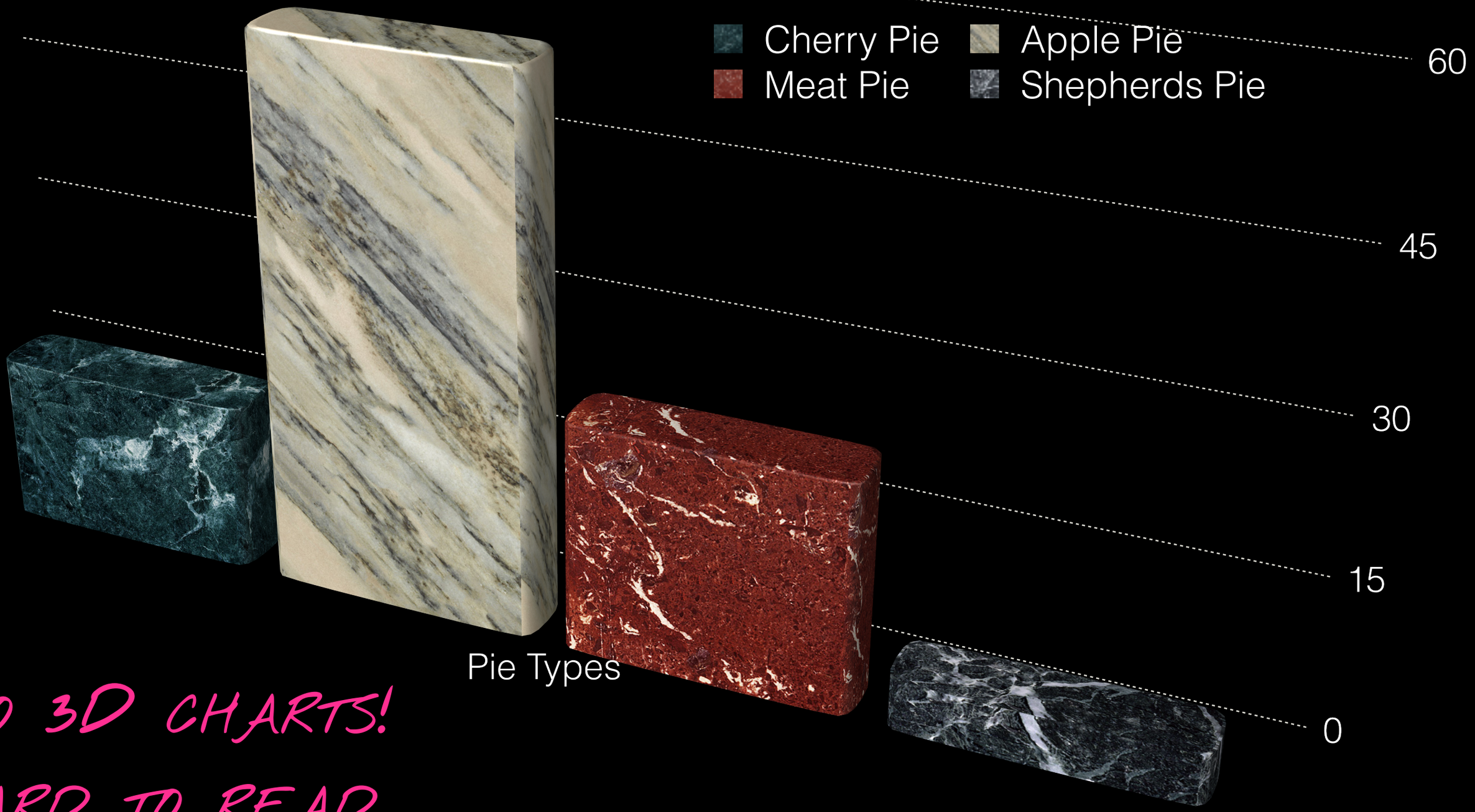
Chart No-nos!





Bocconi

Chart No-nos!



NO 3D CHARTS!

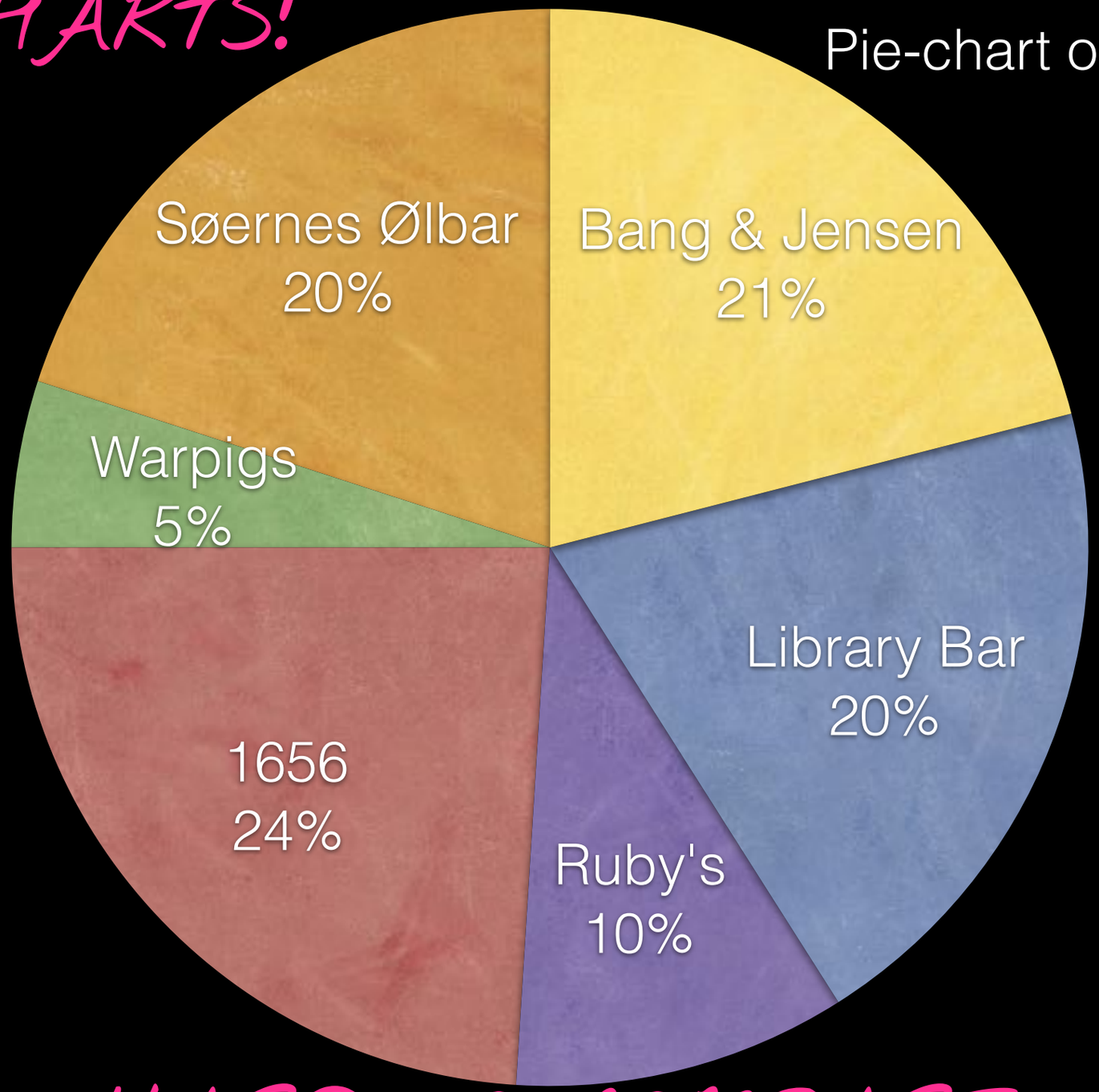
HARD TO READ

UNNECESSARY INFORMATION (DEPTH, TEXTURE)

Bocconi

More No-nos!

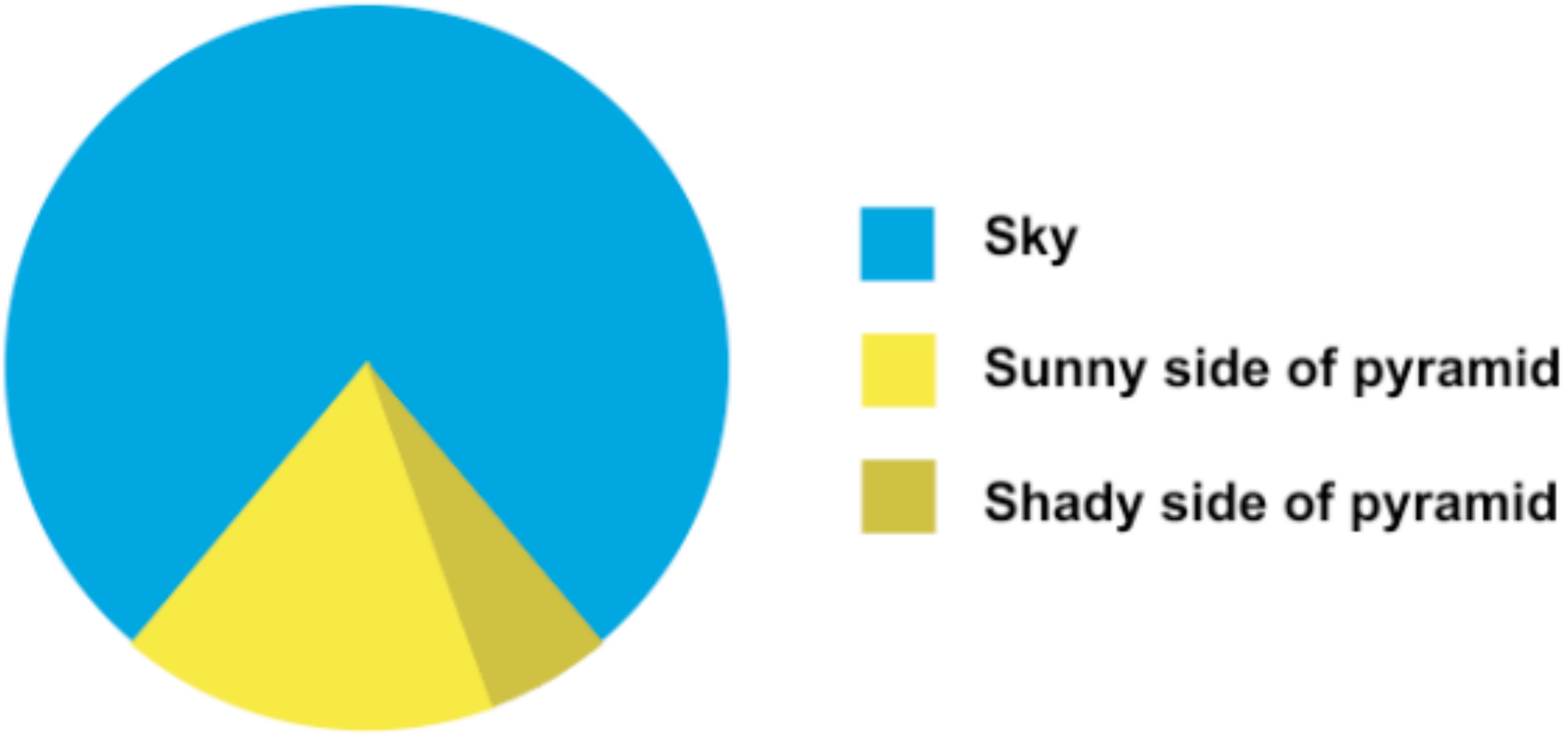
NO PIE CHARTS!



HARD TO COMPARE

AREA IS IRREGULAR SHAPE

The ONLY Exception!



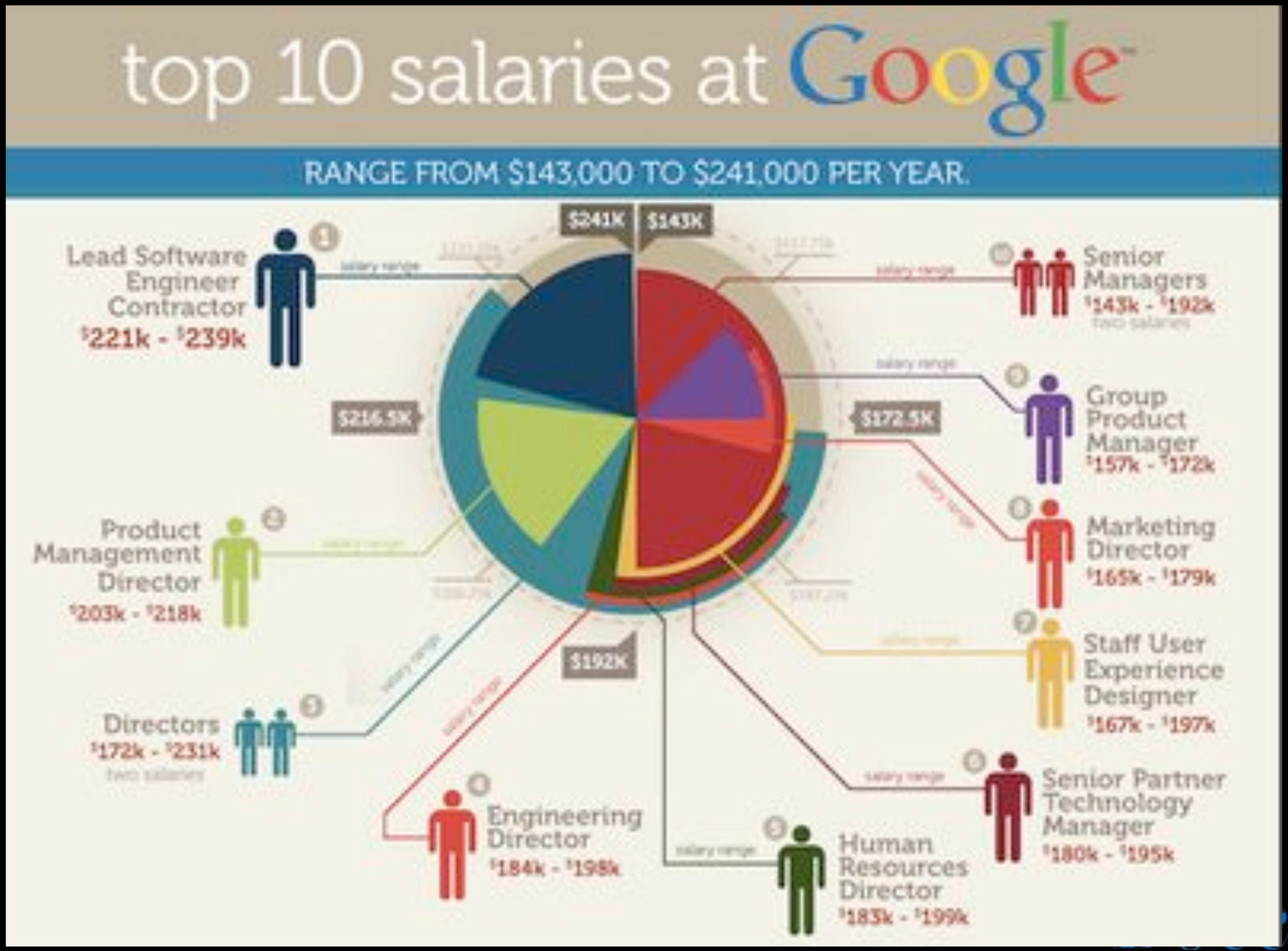
And… Just don’t…

NO DISCERNIBLE INFORMATION VALUE
LONG WORDS GET MORE SPACE
UNSCIENTIFIC

Bocconi

Improvements

http://junkcharts.typepad.com/junk_charts/2011/10/the-massive-burden-of-pie-charts.html



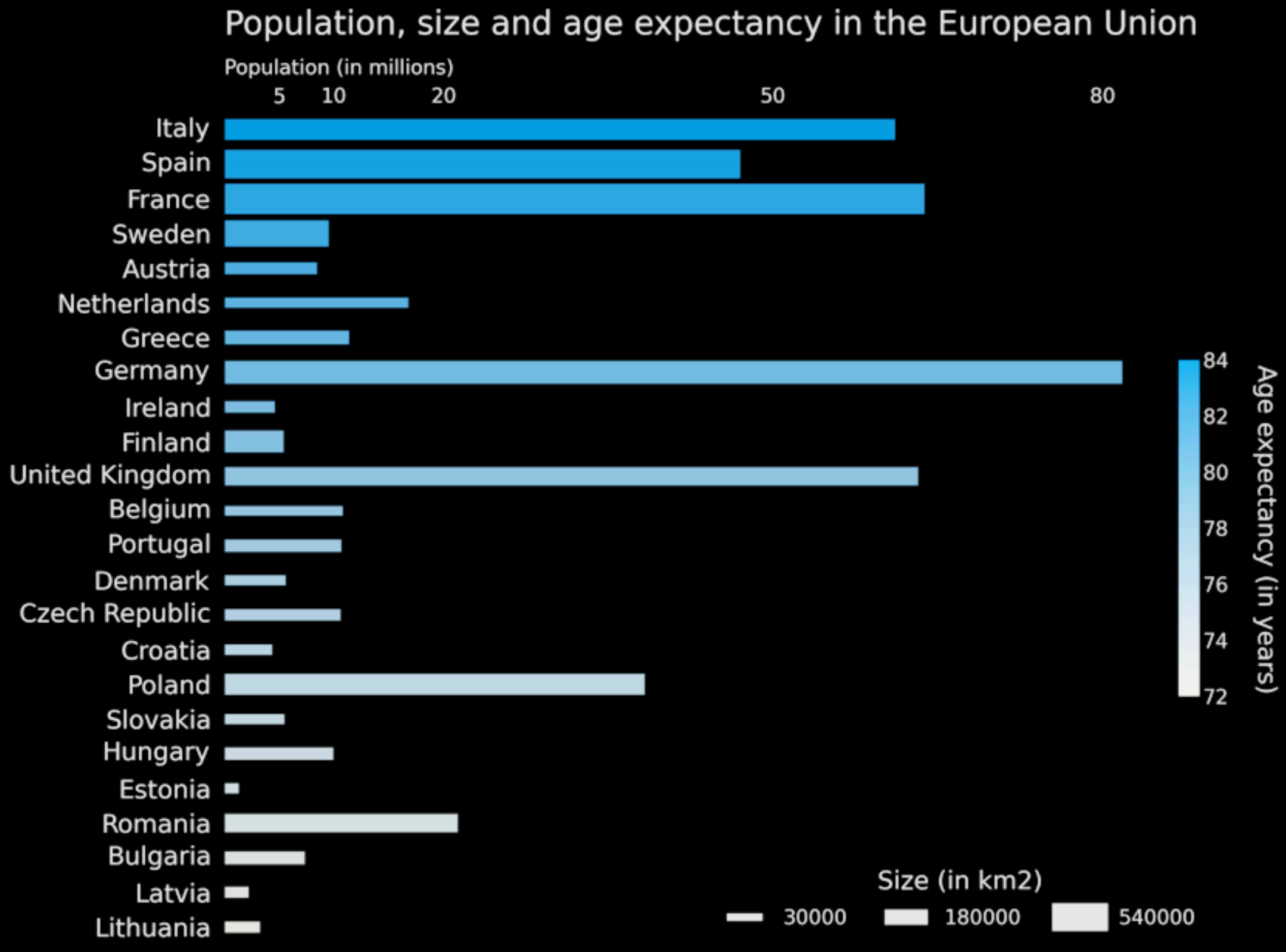
Improvements

http://junkcharts.typepad.com/junk_charts/2011/10/the-massive-burden-of-pie-charts.html

Bocconi

Improvements

https://datasciencelab.wordpress.com/2013/12/21/beautiful-plots-with-pandas-and-matplotlib



Dark Horses: Heatmaps

New COVID19 cases per 100k by week and age group

Neu gemeldete COVID19-Fälle pro 100.000 Personen in Deutschland nach Meldewoche und Altergruppe

	W20	W21	W22	W23	W24	W25	W26	W27	W28	W29	W30	W31	W32	W33	W34	W35	W36
0-4	3.0	3.1	2.2	2.7	2.8	4.0	3.4	3.5	3.7	3.5	4.0	4.0	4.7	6.9	7.1	5.6	5.2
5-9	3.0	2.3	3.0	2.7	3.3	4.7	3.7	3.3	3.1	3.7	4.1	5.3	7.6	10.9	10.1	9.3	7.4
10-14	3.2	3.0	3.2	2.8	3.2	5.2	4.0	2.8	2.7	4.1	5.2	6.8	9.8	14.7	15.5	12.9	9.9
15-19	4.7	4.0	4.2	4.2	2.9	6.3	4.3	3.6	3.9	5.0	6.5	9.5	14.2	19.2	22.8	21.5	17.9
20-24	9.0	6.8	6.4	5.0	5.4	8.1	5.5	5.6	5.1	6.5	8.0	11.3	15.9	25.3	33.3	28.6	23.9
25-29	7.9	7.1	6.4	4.5	4.9	8.0	6.5	6.4	4.6	6.2	8.3	10.2	12.3	16.8	24.0	24.4	20.9
30-34	7.0	5.7	4.8	3.6	4.0	7.2	5.9	4.6	4.6	5.3	6.7	7.6	9.1	12.4	16.8	15.1	15.0
35-39	6.7	5.3	4.6	2.9	3.6	7.0	5.6	4.2	3.8	4.8	6.7	7.2	9.7	12.1	14.2	13.3	10.8
40-44	7.1	4.8	4.2	3.4	3.1	8.8	5.4	4.2	3.8	5.0	6.3	8.5	9.6	12.5	14.1	12.9	11.1
45-49	6.5	5.1	4.3	3.2	2.9	7.4	5.5	4.4	3.6	4.6	6.6	7.7	9.0	11.4	13.5	11.9	10.6
50-54	5.8	3.9	3.7	2.3	2.3	4.8	4.1	2.8	2.6	3.2	4.5	5.4	6.4	7.2	8.6	8.2	7.4
55-59	5.2	3.9	3.5	2.1	2.0	2.7	3.0	2.3	1.7	2.5	3.6	4.2	4.5	4.8	6.3	5.1	5.1
60-64	4.1	3.1	2.9	1.9	1.5	2.1	1.9	1.9	1.6	1.9	2.6	2.6	3.5	3.3	4.0	4.2	3.8
65-69	3.0	2.5	2.2	1.2	1.1	1.4	1.3	1.3	1.3	1.4	1.7	1.7	2.5	2.2	2.4	2.7	2.2
70-74	3.2	2.3	2.0	1.2	1.0	1.4	1.0	1.0	1.4	1.5	1.6	2.3	2.3	2.0	2.4	2.3	2.4
75-79	5.1	3.0	2.4	1.7	1.3	1.9	1.6	1.2	1.2	1.2	1.6	2.0	1.6	2.0	1.8	1.9	2.0
80+	9.6	6.5	4.6	2.9	2.8	2.4	1.9	1.7	1.2	1.5	1.7	2.9	2.4	2.8	1.8	1.9	1.6

https://twitter.com/christoph_rothe/status/1303603360830980096

Bocconi

Chart Suggestions—A Thought-Starter

What would you like to show?

Comparison

Among Items

- One Variable per Item
  - Few Categories
    - Many Items → Bar Chart
    - Few Items → Column Chart
  - Many Categories → Table or Table with Embedded Charts
- Two Variables per Item → Variable Width Column Chart

Over Time

- Many Periods → Cyclical Data → Circular Area Chart
- Few Periods
  - Non-Cyclical Data → Line Chart
  - Single or Few Categories → Column Chart
  - Many Categories → Line Chart

Relationship

- Two Variables → Scatter Chart
- Three Variables → Bubble Chart

Distribution

- Single Variable
  - Few Data Points → Column Histogram
  - Many Data Points → Line Histogram
- Two Variables → Scatter Chart
- Three Variables → 3D Area Chart

Composition

Changing Over Time

- Few Periods
  - Only Relative Differences Matter → Stacked 100% Column Chart
  - Relative and Absolute Differences Matter → Stacked Column Chart
- Many Periods
  - Only Relative Differences Matter → Stacked 100% Area Chart
  - Relative and Absolute Differences Matter → Stacked Area Chart

Static

- Simple Share of Total → Pie Chart
- Accumulation or Subtraction to Total → Waterfall Chart
- Components of Components → Stacked 100% Column Chart with Subcomponents

Reading

- http://datavisualization.ch/
- http://www.princeton.edu/~ina/infographics/index.html
- http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches/
- http://abeautifulwww.com/2007/05/20/visualizing-the-power-struggle-in-wikipedia/
- http://www.designlabelblog.com/2008/03/data-visualization-and-infographics_30.html
- http://www.swiss-miss.com/data_visualization/page/2
- http://www.visualcomplexity.com/vc/
- http://flowingdata.com/2008/03/19/21-ways-to-visualize-and-explore-your-email-inbox/
- http://www.math.yorku.ca/SCS/Gallery/
- http://christopherbaker.net/projects/mymap/
- http://mashable.com/2007/05/15/16-awesome-data-visualization-tools/
- http://www.creativesynthesis.net/blog/2007/05/16/chi-review-qualities-of-perceived-aesthetic-in-data-visualization/

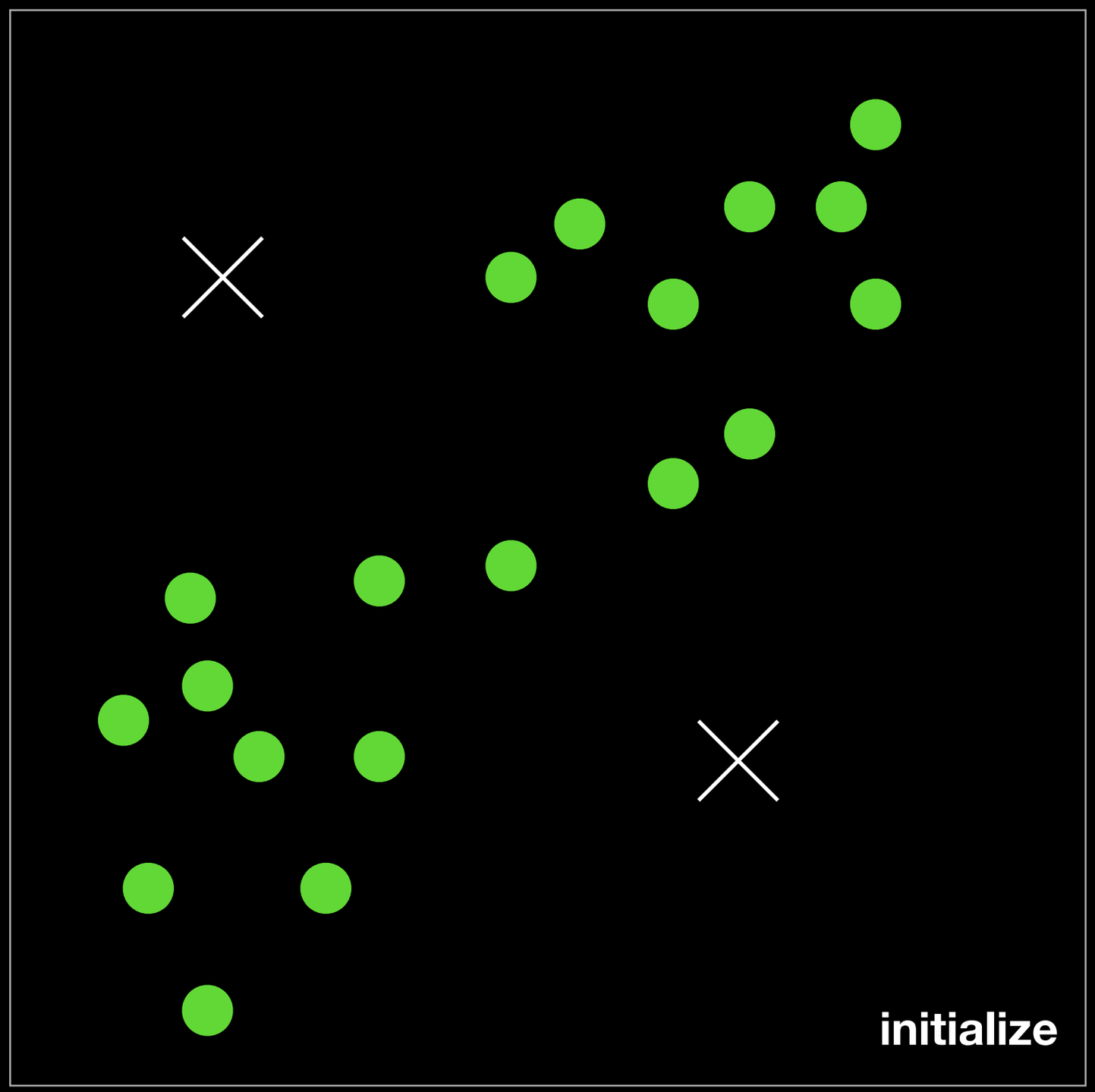
Clustering

Bocconi

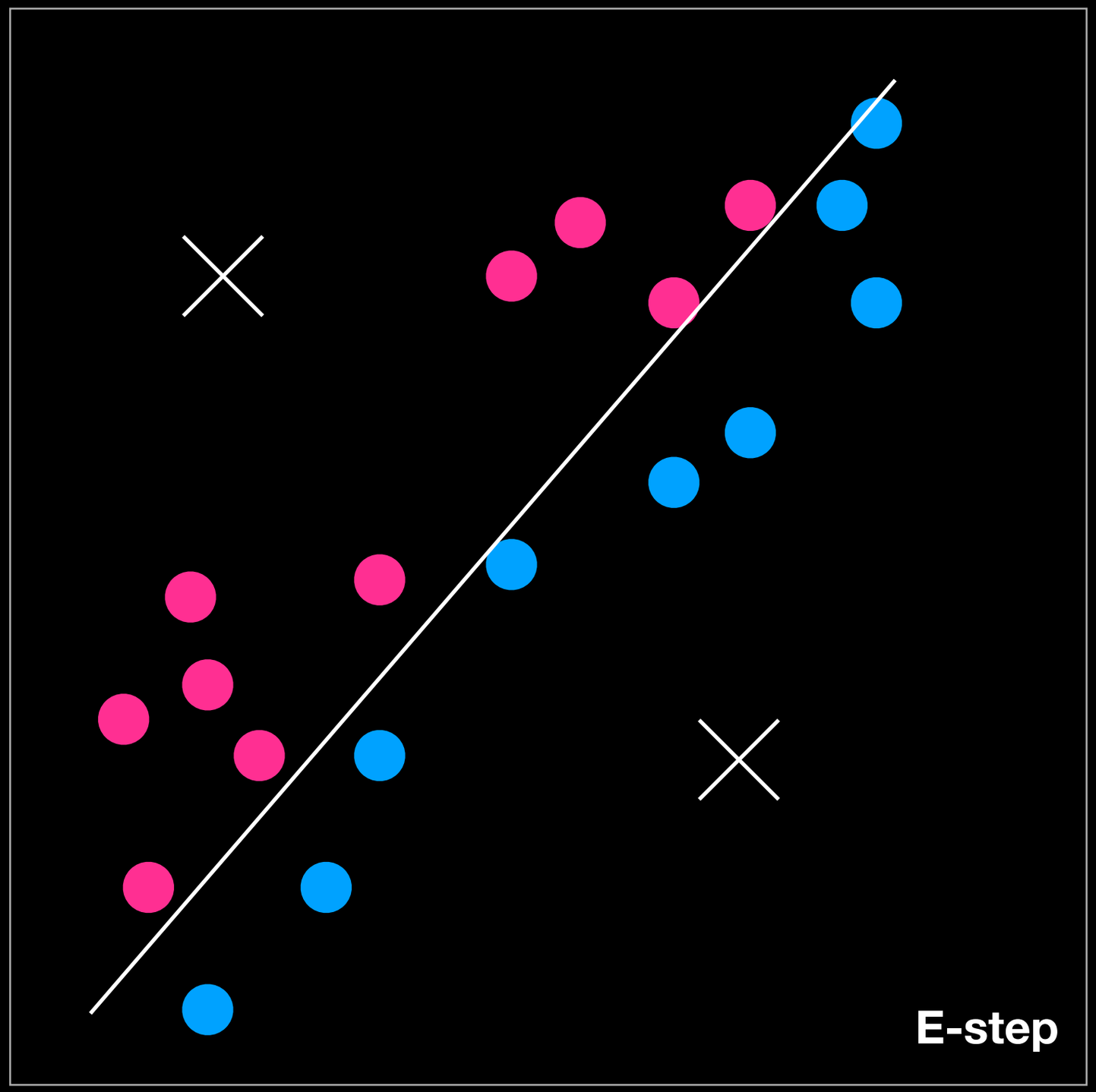
k-Means Clustering

Bocconi

k-Means

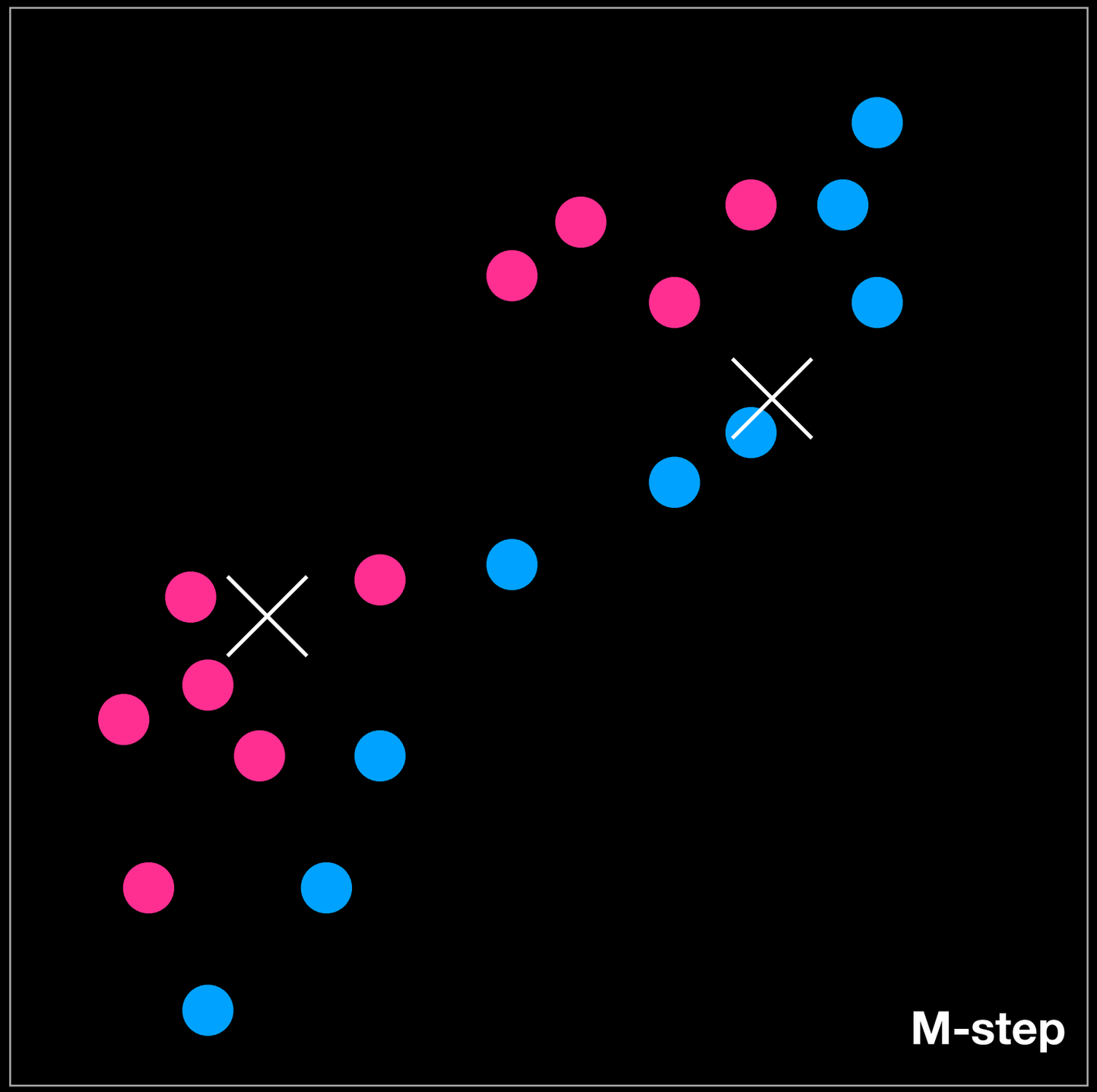


k-Means



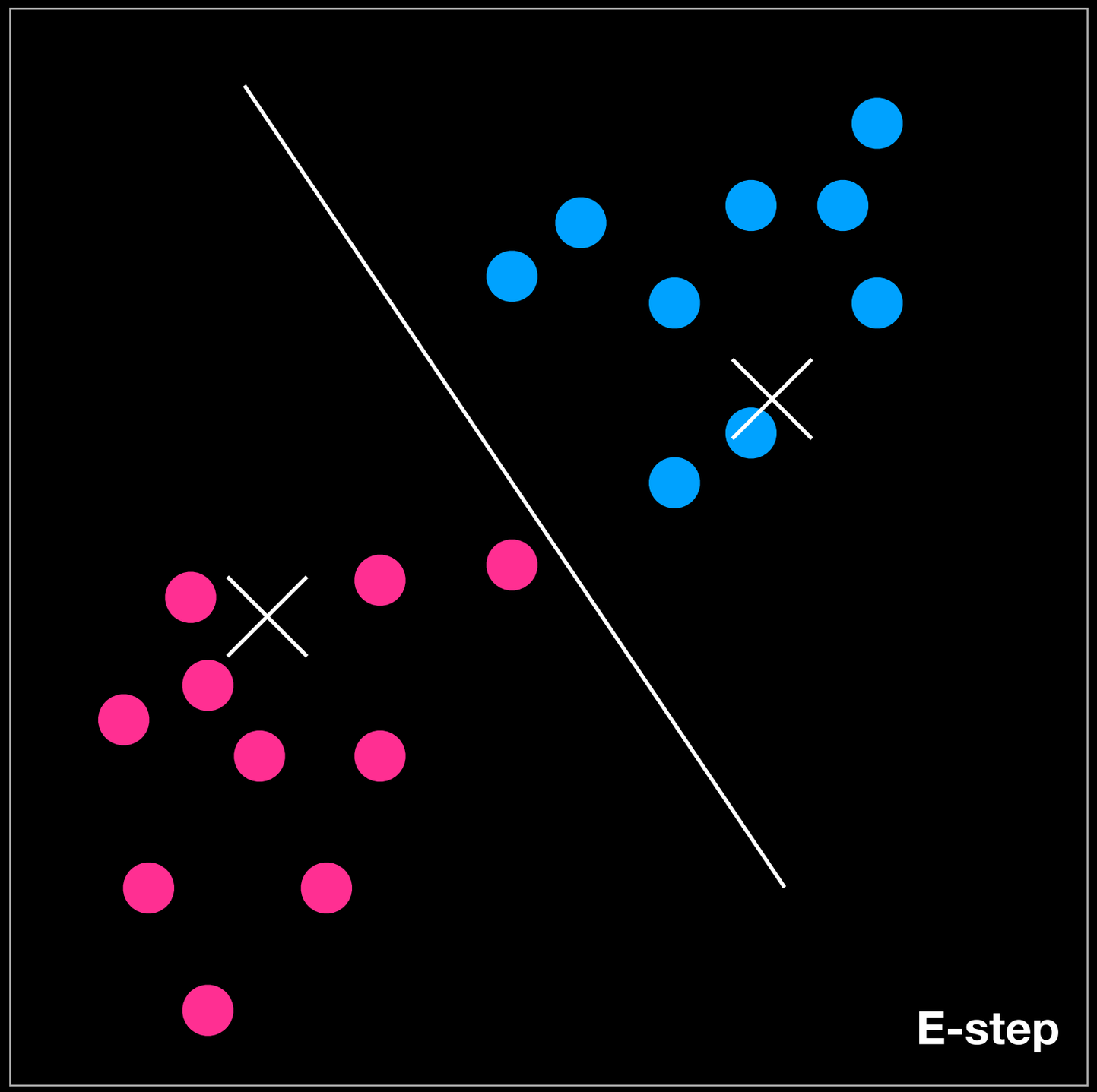
ASSIGN POINTS TO CENTROIDS

k-Means

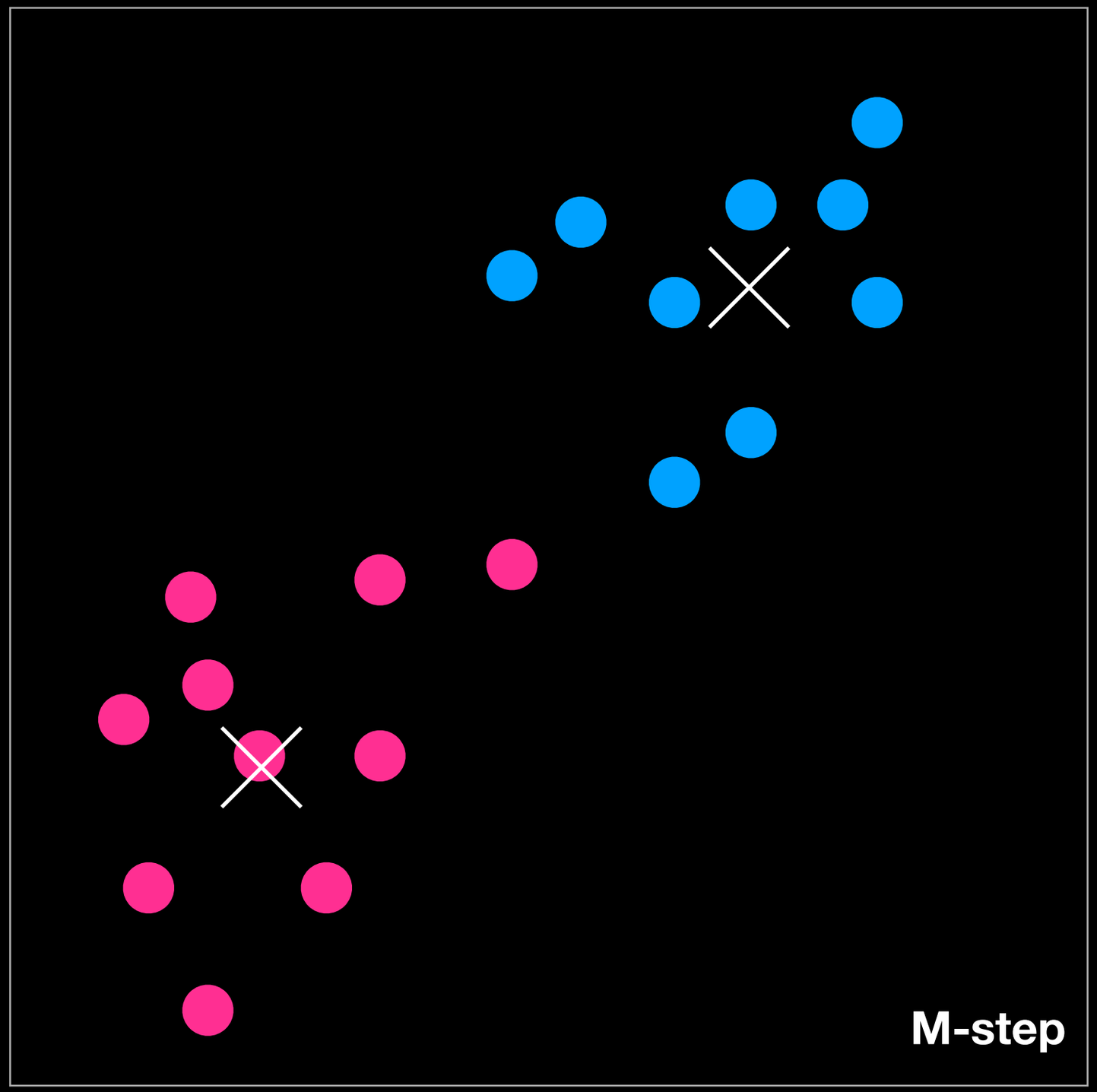


RECOMPUTE CENTROIDS

k-Means



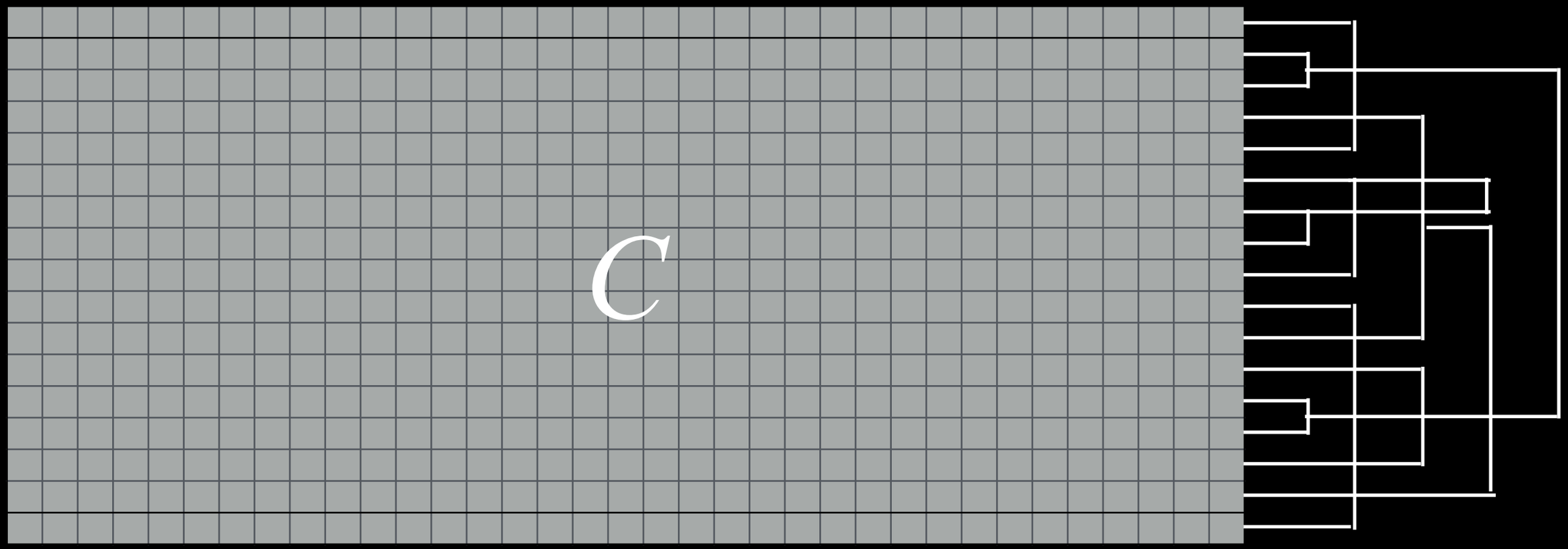
k-Means



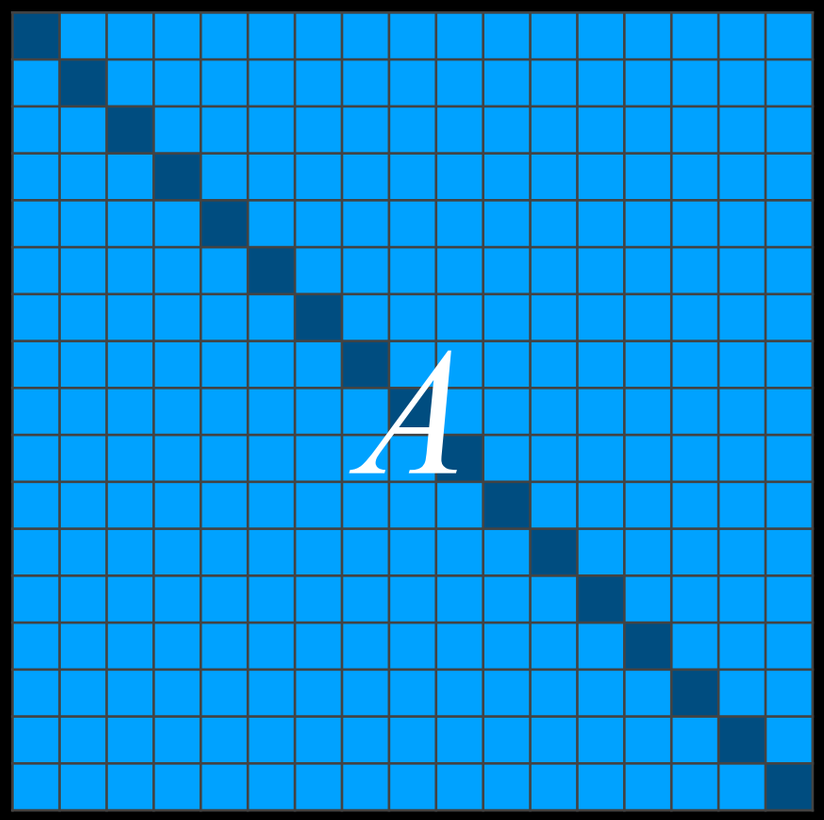
Agglomerative Clustering

Bocconi

Building Up

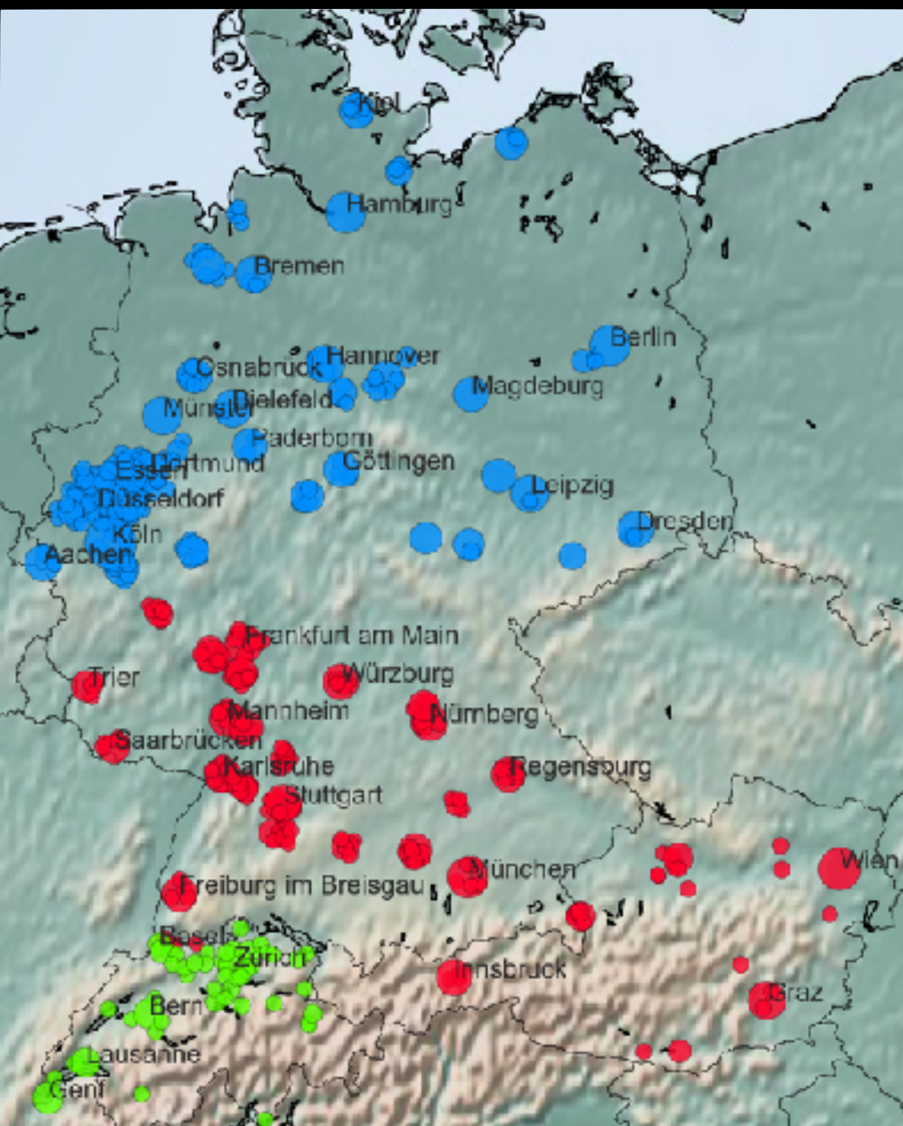


ADJACENCY MATRIX

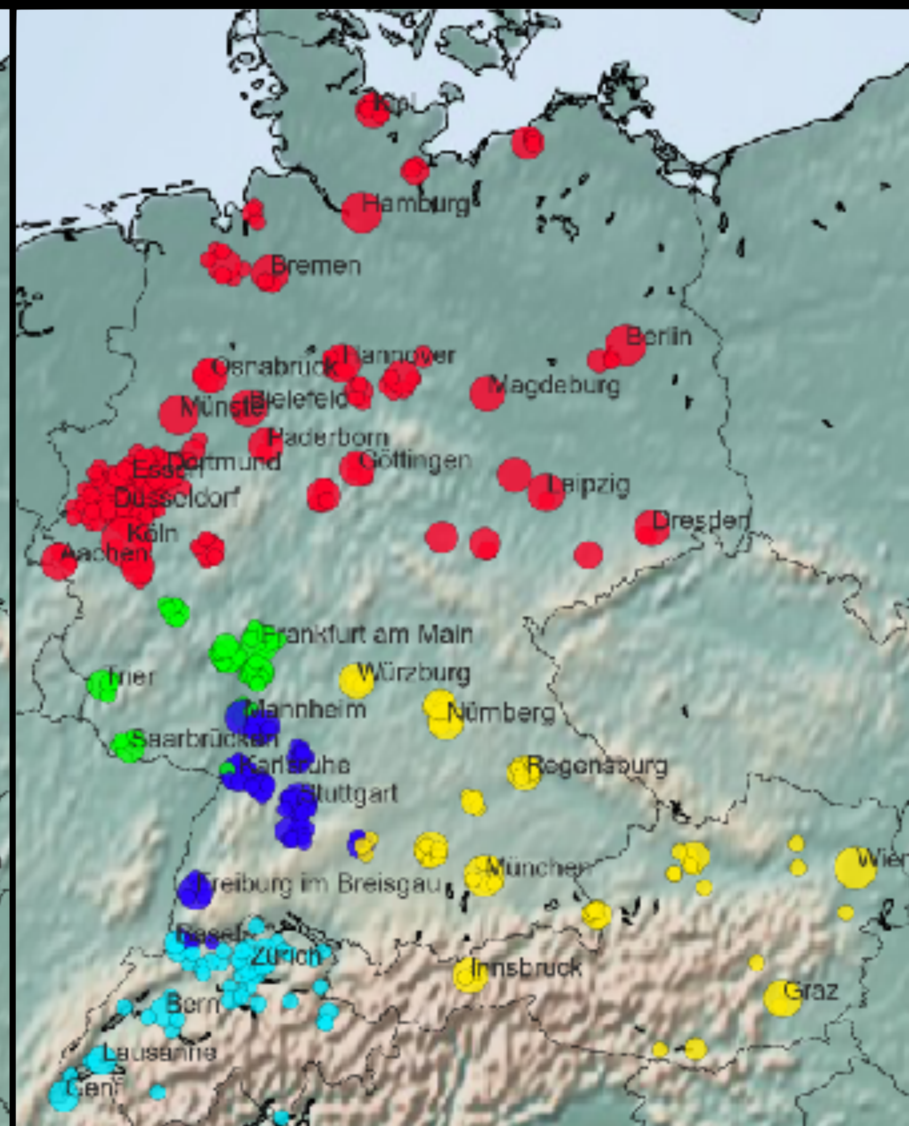


Dialect Clusters

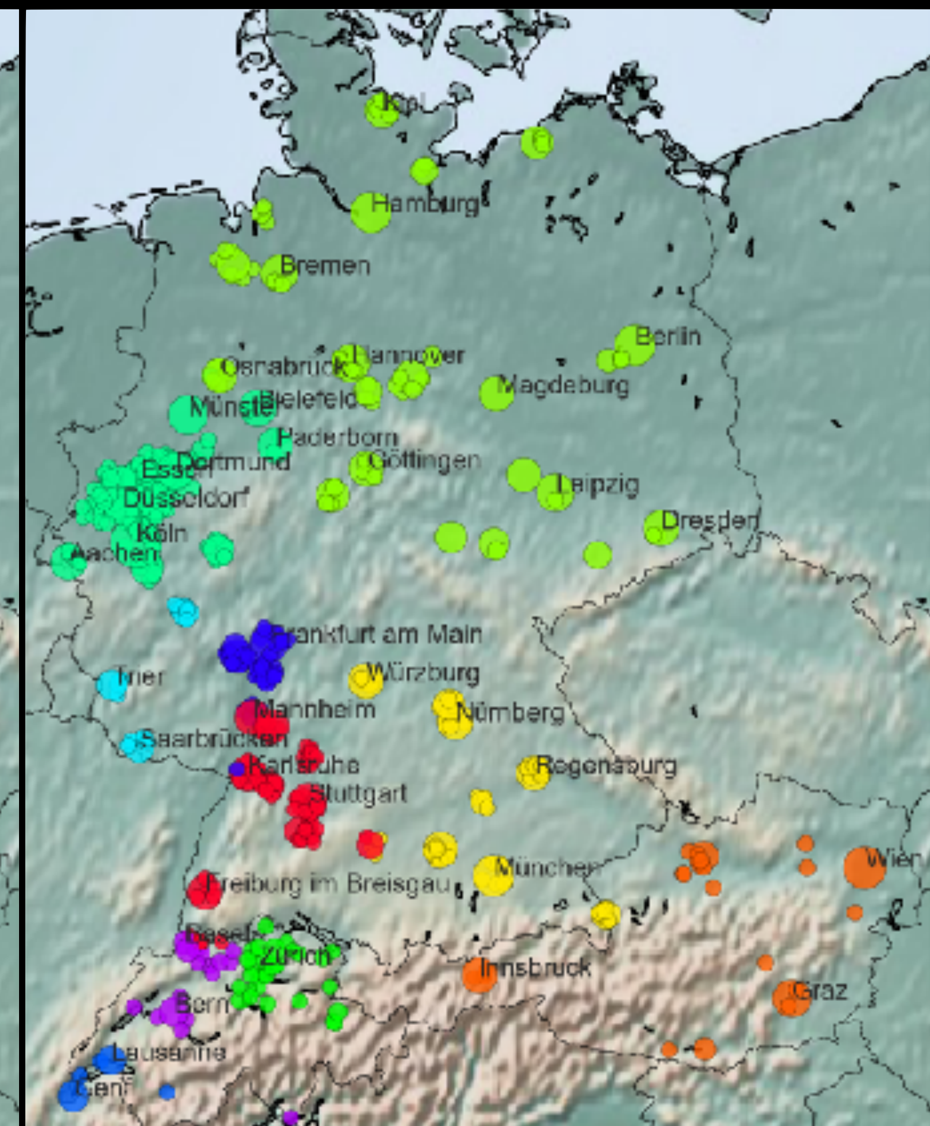
3



5



10



Evaluating Clusters

Bocconi

Making Sense of Clusters

NOUNS

NEUTRAL

VERBS

perspicacious

lacerate

president

kissed

fresh

induce

rotten

ROMANCE ROOT

GERMANIC ROOT

platypus

arsenic

mortify

mortgage

recognize

pie

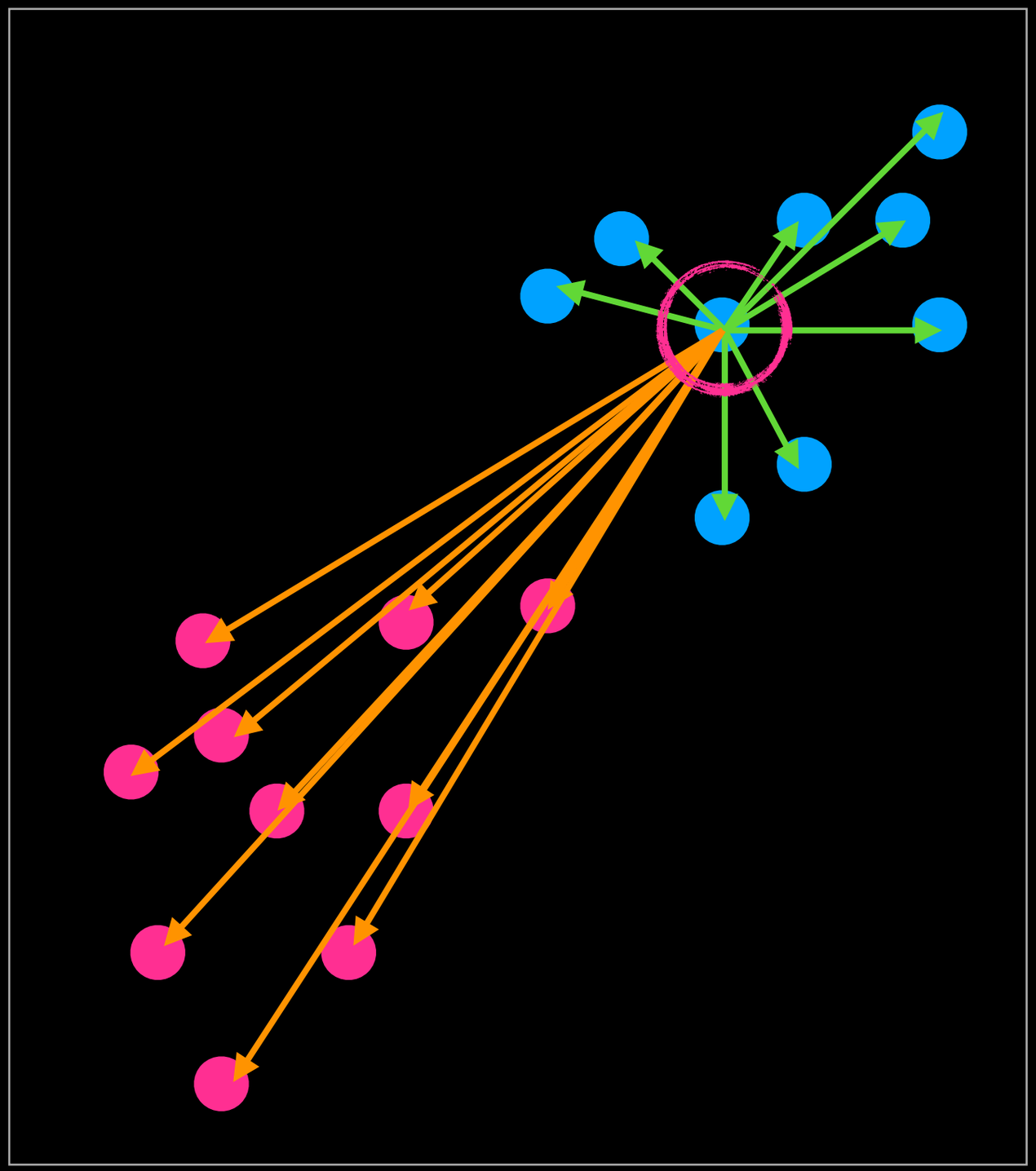
blue

POSITIVE

ADJECTIVES

NEGATIVE

How Many Clusters?



Silhouette Score

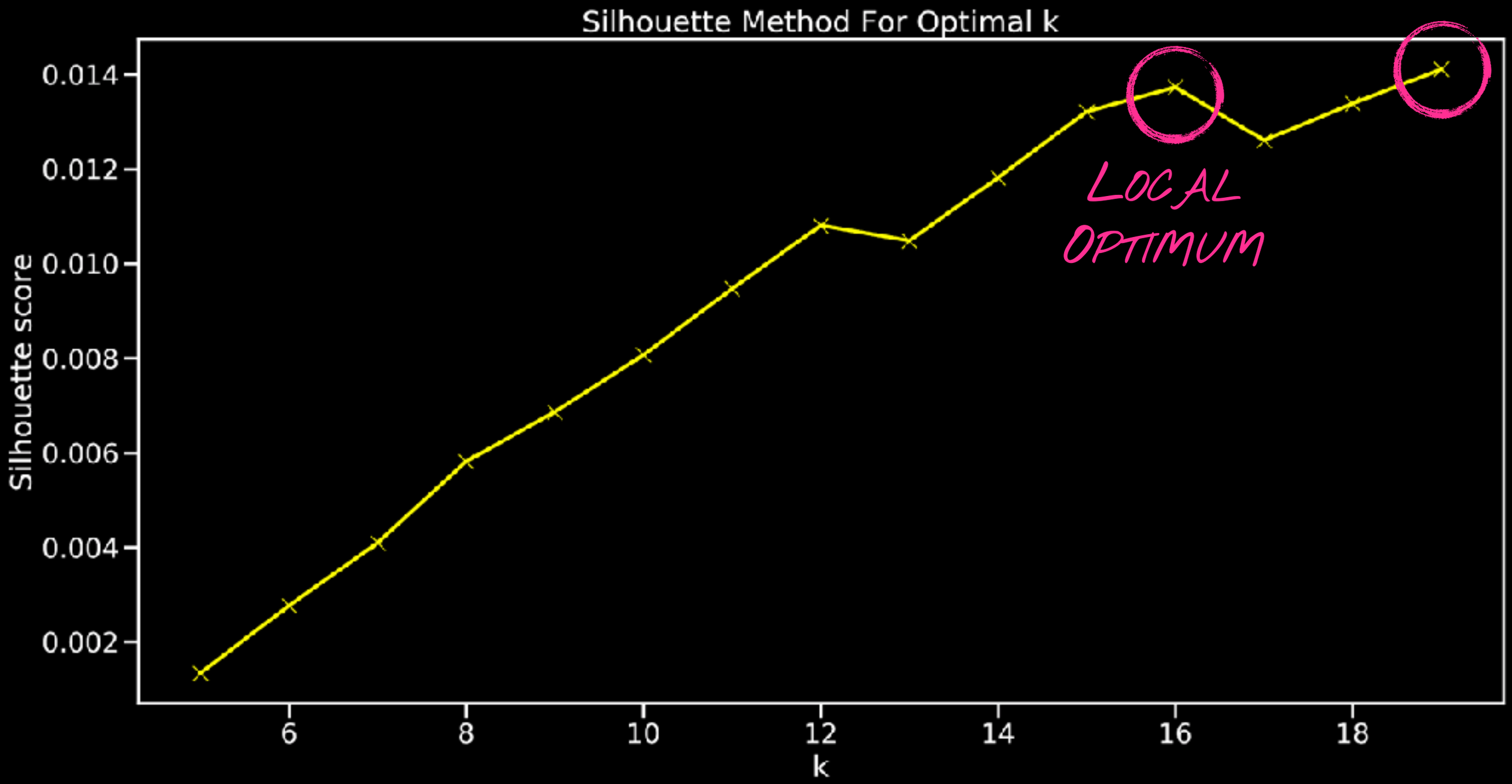
a = mean intra-cluster distance

$$S = \frac{(b - a)}{max(a, b)}$$

b = mean dist. nearest cluster

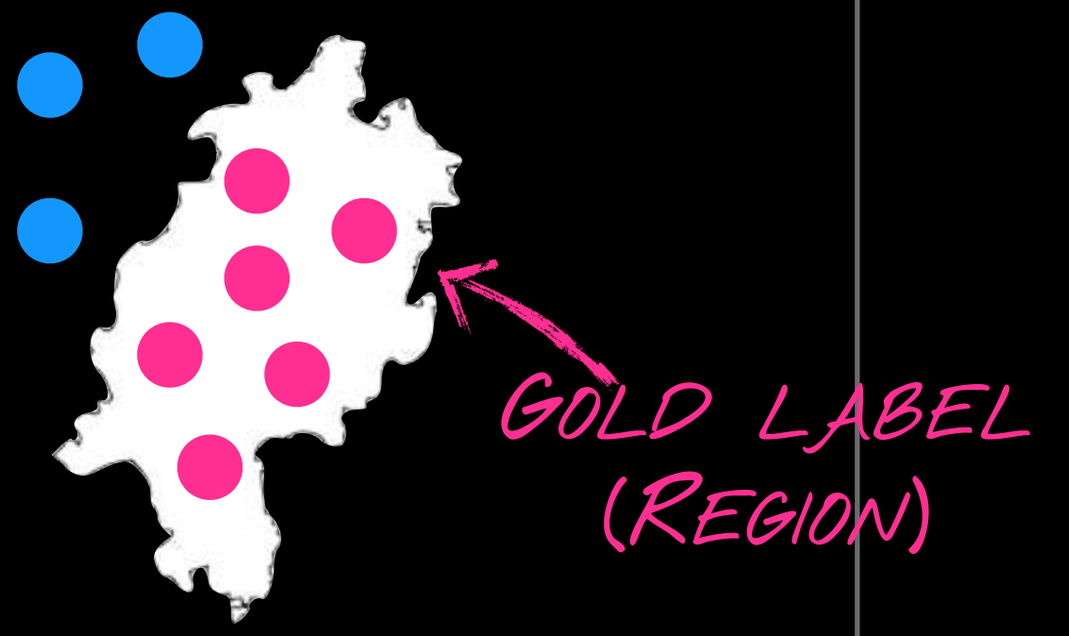
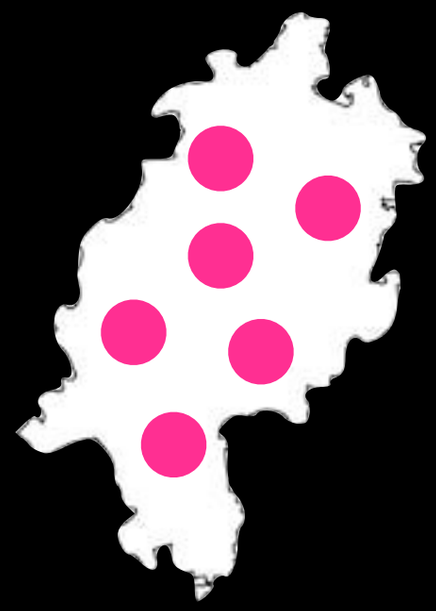
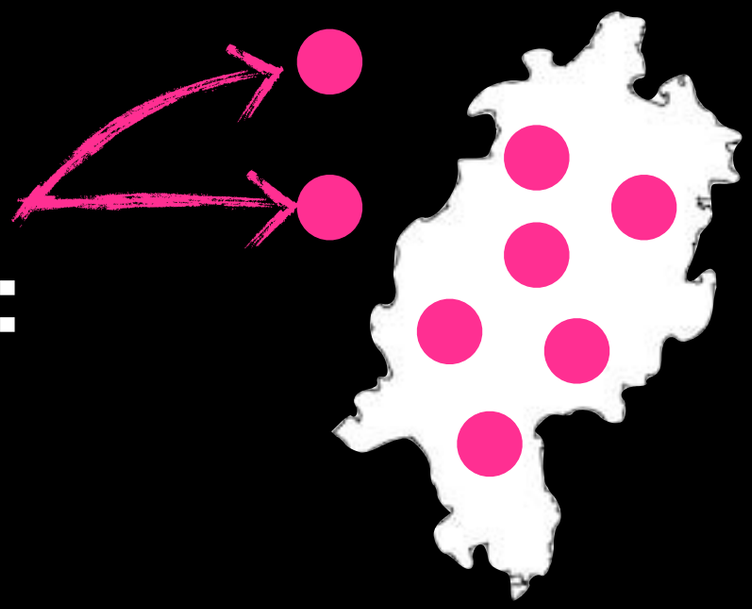
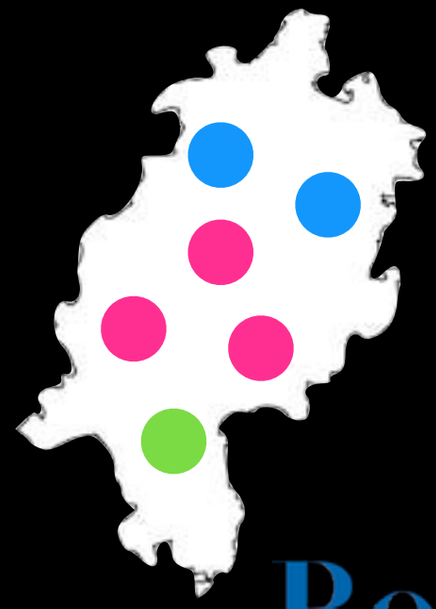
Silhouette Scores

DEPENDS ON PATIENCE|COMPUTE POWER



Bocconi

Supervised Evaluation Metrics

	Homogeneity *cluster* has only 1 gold label	**Completeness** *gold label* has only 1 cluster
Good:		
Bad:		

Comparison

	***k*-means**	**Agg**
scalable	yes	no (up to ~20k)
repeatable result	no	yes
include external info	no	yes
Good on dense clusters?	no	yes

Wrapping Up

Bocconi

When to Use What

	Discrete Features	Embeddings
Latent topics	NMF	*Not applicable*
RGB translation	NMF	SVD + scaling
Plotting	SVD	t-SNE
Clustering	*Reduce dimensions*	*Use as-is*

Take-Home Points

- **Matrix factorization** assumes latent concept dimensions
  - Can be used for semantic similarity (**LSA**)
  - Reduced components can be **visualized** in **graphs** or as **RGB** colors
- **Clusters** can group input in new ways
- Trade-off between speed and interpretability