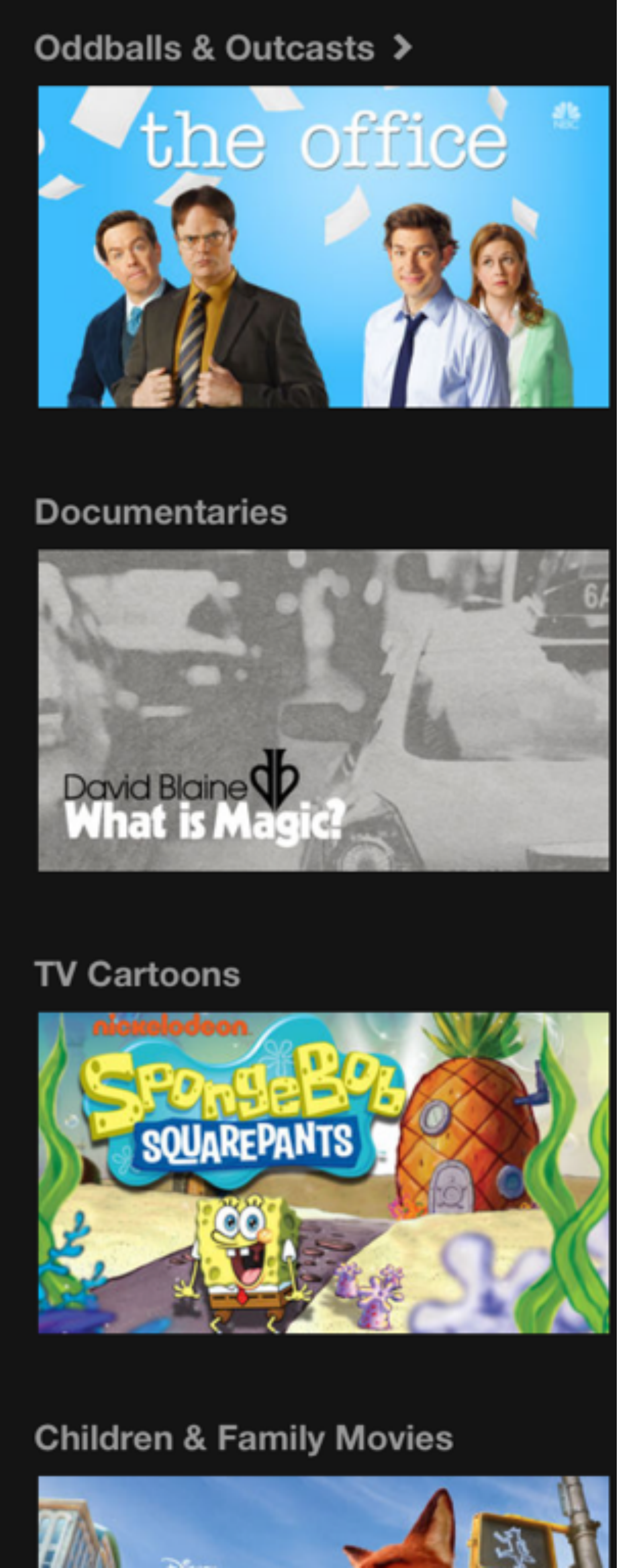
Computational Categories

David Bamman
Info 202: Information Organization and Retrieval

October 24, 2016



Categories

- Categories provide the framework for organizing resources

- Classification assigns individual resources to categories.

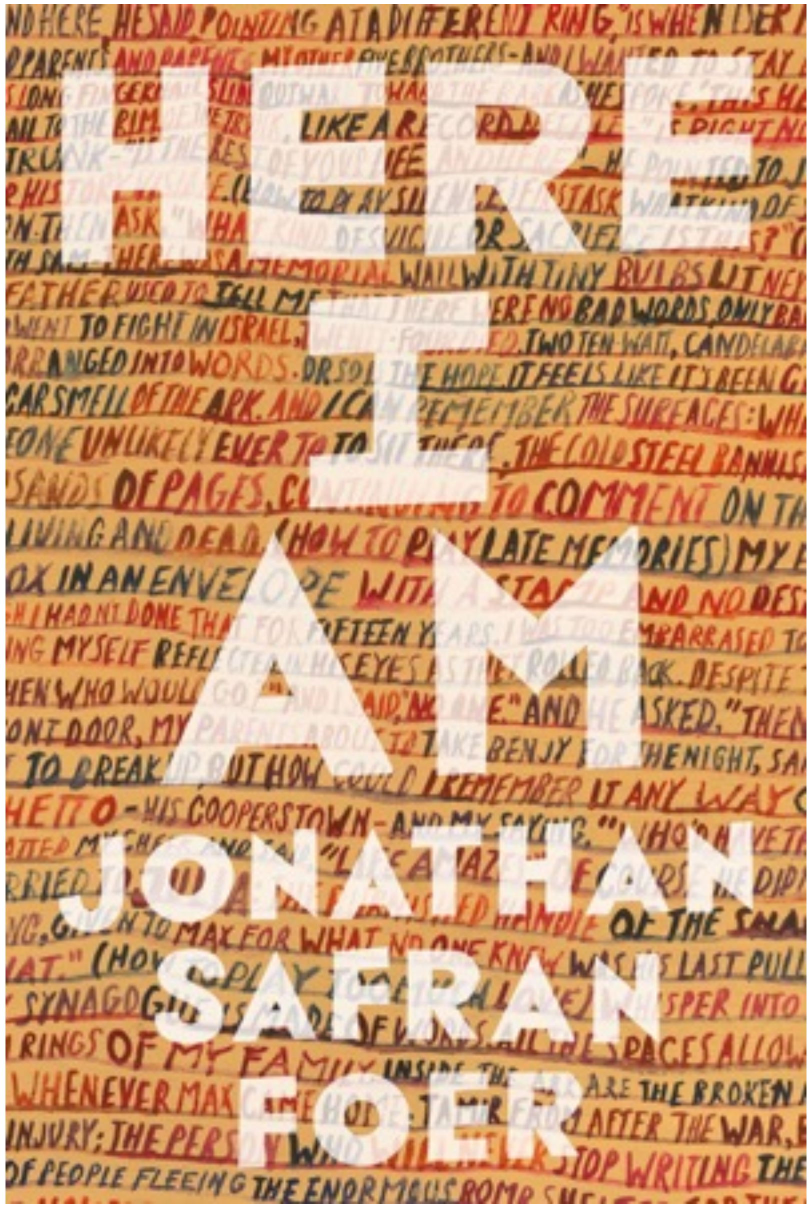
Cerebral Foreign Movies from the 1970s (669)
Cerebral Foreign Political Dramas (2742)
Cerebral Foreign War Movies (157)
Cerebral French-Language Crime Dramas (2935)
Cerebral French-Language Dramas (4521)
Cerebral French-Language Dramas from the 1960s (102)
Cerebral French-Language Movies (3623)
Cerebral French-Language Movies from the 1950s (2642)
Cerebral French-Language Movies from the 1960s (2672)
Cerebral French-Language Movies from the 1970s (2703)
Cerebral Independent Biographical Movies (2518)
Cerebral Independent Comedies (1474)
Cerebral Independent Crime Movies (3717)
Cerebral Independent Movies (551)
Cerebral Independent Movies from the 1980s (1451)
Cerebral Independent Political Movies (368)
Cerebral Italian Dramas (1553)
Cerebral Japanese Dramas (3720)
Cerebral Military Movies (3156)
Cerebral Movies (1813)
Cerebral Movies based on Books (3555)
Cerebral Movies directed by Akira Kurosawa (4359)
Cerebral Political Dramas (814)
Cerebral Political Movies (3152)
Cerebral Scandinavian Movies (995)

Cult B-Horror Movies (2622)
Cult Crime Comedies (1571)
Cult Crime Movies from the 1960s (475)
Cult Crime Movies from the 1970s (510)
Cult Crime Movies from the 1980s (538)
Cult Movies based on Books (4201)
Cult Movies on Blu-ray (4310)
Cult Psychological Horror Movies (186)
Cult Satanic Stories (3527)
Cult Sci-Fi & Fantasy (4734)
Cult Sci-Fi & Fantasy from the 1950s (117)
Cult Sci-Fi & Fantasy from the 1970s (168)
Cult Sci-Fi & Fantasy from the 1980s (193)
Cult Sci-Fi Thrillers (2521)
Czech Movies (1697)
Dance Workouts (1498)
Dark Action & Adventure based on Books (858)
Dark Action Sci-Fi & Fantasy (1452)
Dark Alien Sci-Fi (3166)
Dark British Dramas (494)
Dark British Dramas based on Books (4382)
Dark British Independent Dramas (831)
Dark British Independent Movies (666)
Dark British Movies from the 1980s (482)
Dark British Political Movies (2414)

http://ogres-crypt.com/public/NetFlix-Streaming-Genres2.html

Categories

- There are many ways we can carve up the world, and different categorizations accomplish different ends and have different caveats attached.
- Many choices to make when creating or adopting a categorization system



Categories in DS

- Categories define the classification task
- Before jumping in to the technical details of classifying, we need to make sure that the classes we are trying to discriminate correspond to what we want to learn.
- Sentiment (positive vs. negative)
- Political preference (democrat vs. republican)
- A “good” employee

Categories in DS

- If a predictive model fares poorly at discriminating between categories (given sufficient training data), maybe we should look at the categories again.

- Classification: use an existing set of categories to predict the category for a new data point

- Clustering: infer a set of new categories from structure in the data.

Why?

Document clustering

Donald Trump launches rare attack at Michelle Obama
Telegraph.co.uk - 9 hours ago
Donald **Trump** launched a rare attack at the First Lady Michelle Obama, referencing an attack line she used in 2007. He also said "all she wants ...

How Donald **Trump** Broke the Al Smith Dinner
International - The Atlantic - Oct 21, 2016

View all

'Brexit times five': could **Trump** really win despite polls favoring ...
The Guardian - 3 hours ago
Outside **Trump** Tower, the war looks to be over. As smoke clears from weeks of political bombardment, White House watchers are convinced ...

Trump Says US Election Result Will Be Like 'Brexit Times Five'
NBCNews.com - 16 hours ago

View all

Google News

Behavioral clustering

- $260 average order
- duration: 5 years
- frequently bought categories: furniture, kitchen appliances

- $13 average order
- duration: 21 days
- frequently bought categories: books

Topic models

A Topic Model of Literary Studies Journals

Overview | Topic | Article | Word | Bibliography | Word index | Settings | About

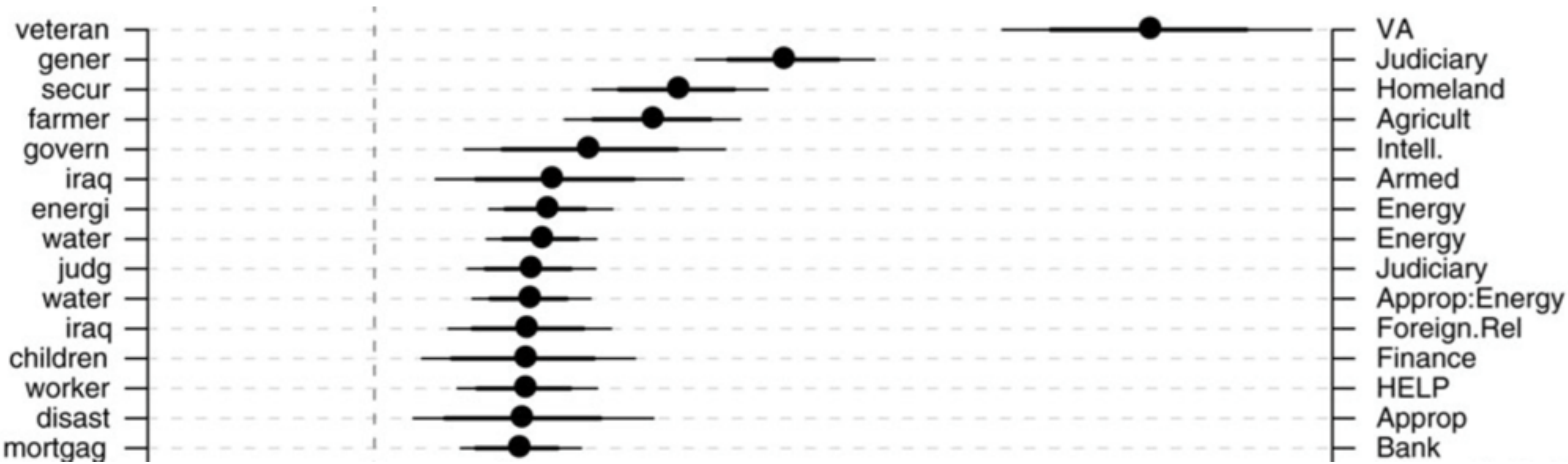
List | Grid | Years

click a column label to sort; click a row for more about a topic

topic	1889—2013	top words	proportion of corpus
1		see both own view role university further account critical particular	2.5%
2		other both two form same even each part experience process	2.6%
3		old beowulf english ic mid swa pe poet ond grendel	0.3%

Goldstone and Underwood (2014), The Quiet Transformations of Literary Studies

Topic models

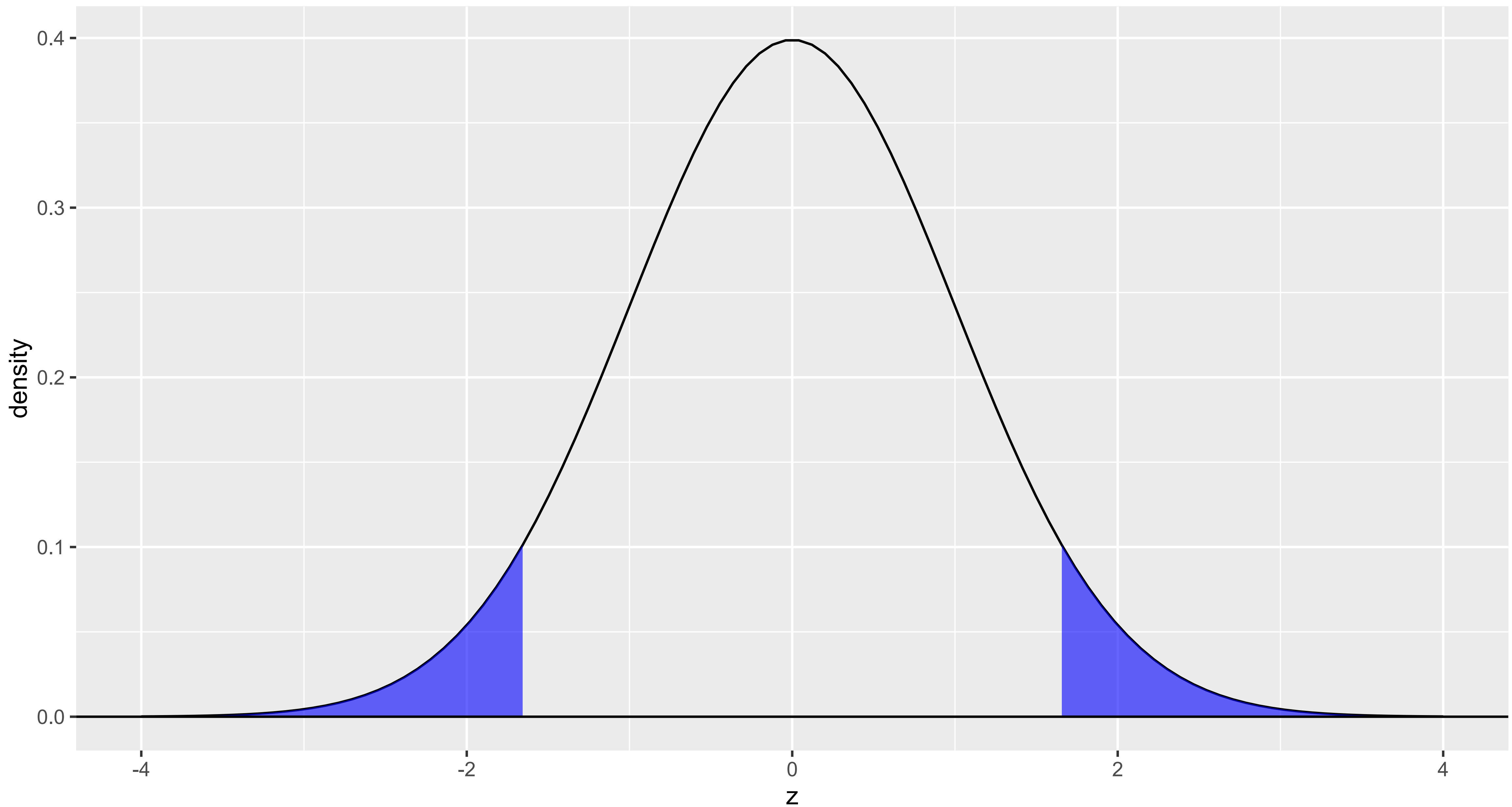


Grimmer (2010), A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases

Descriptive statistics

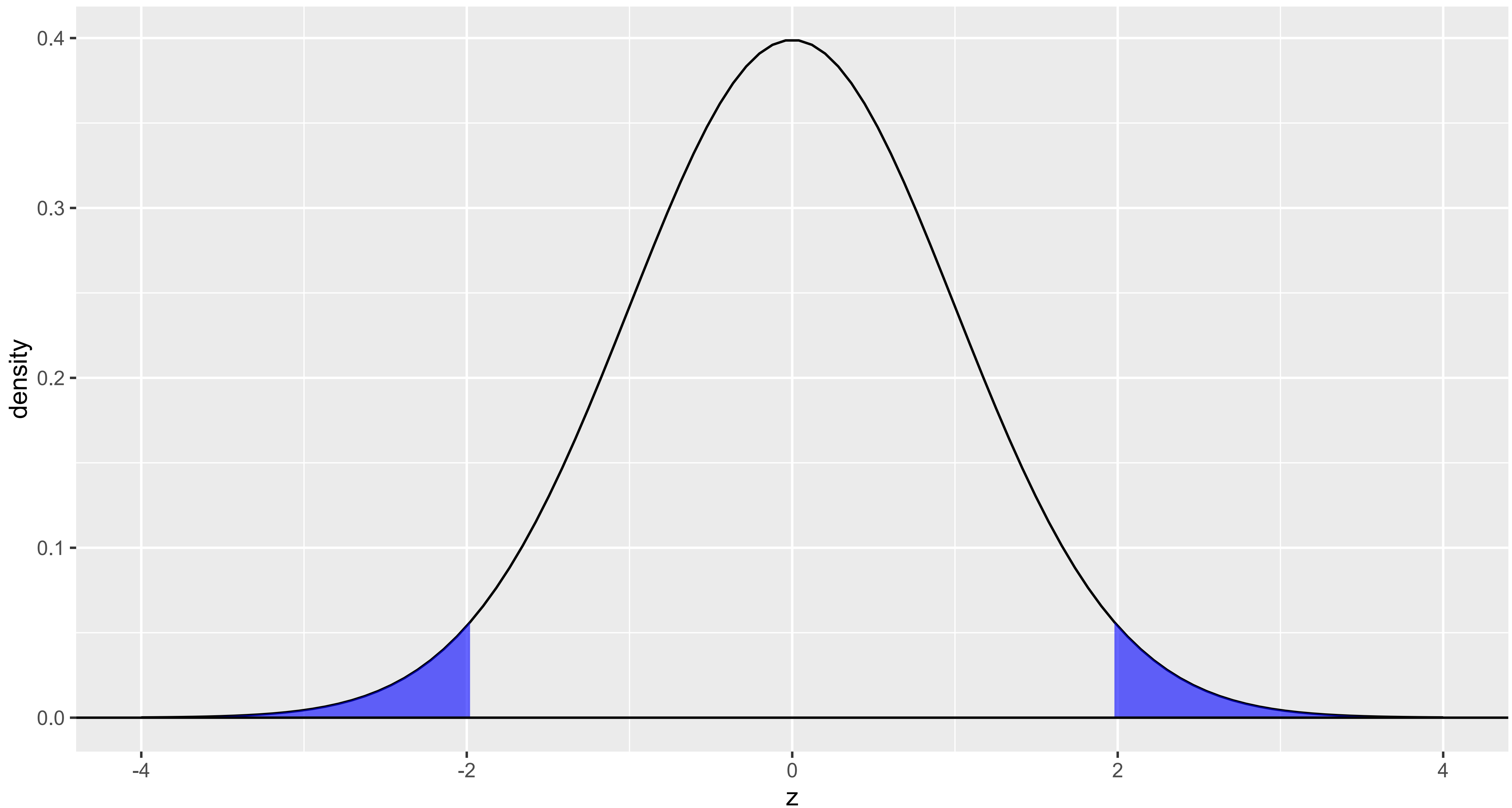
- The simplest computational categories are those derived from descriptive statistics
- Not based on features of the data, but rather on where how typical a data point is respect to the rest of the collection

Outlier vs. non-outlier



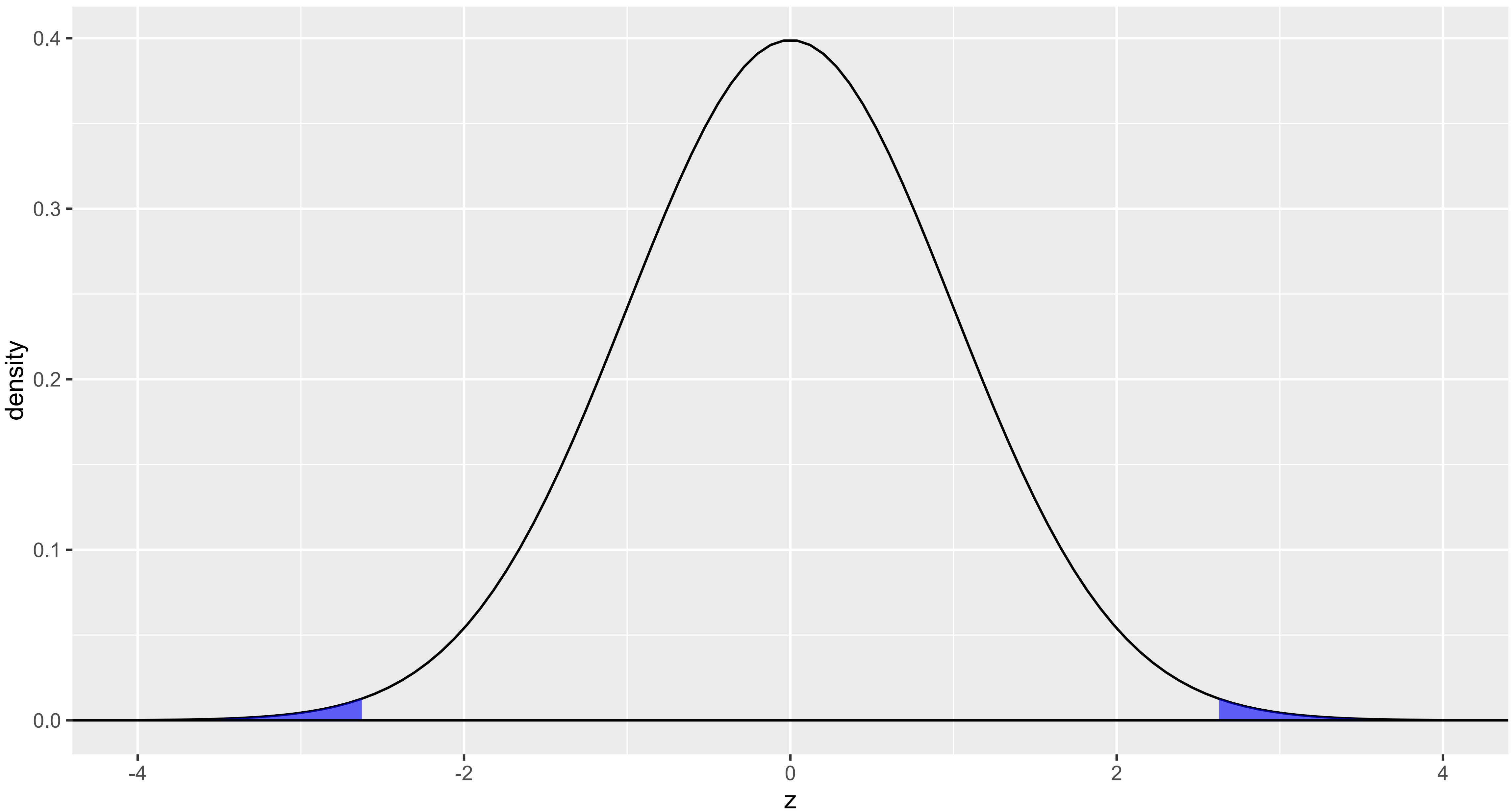
“least likely” 10%

Outlier vs. non-outlier



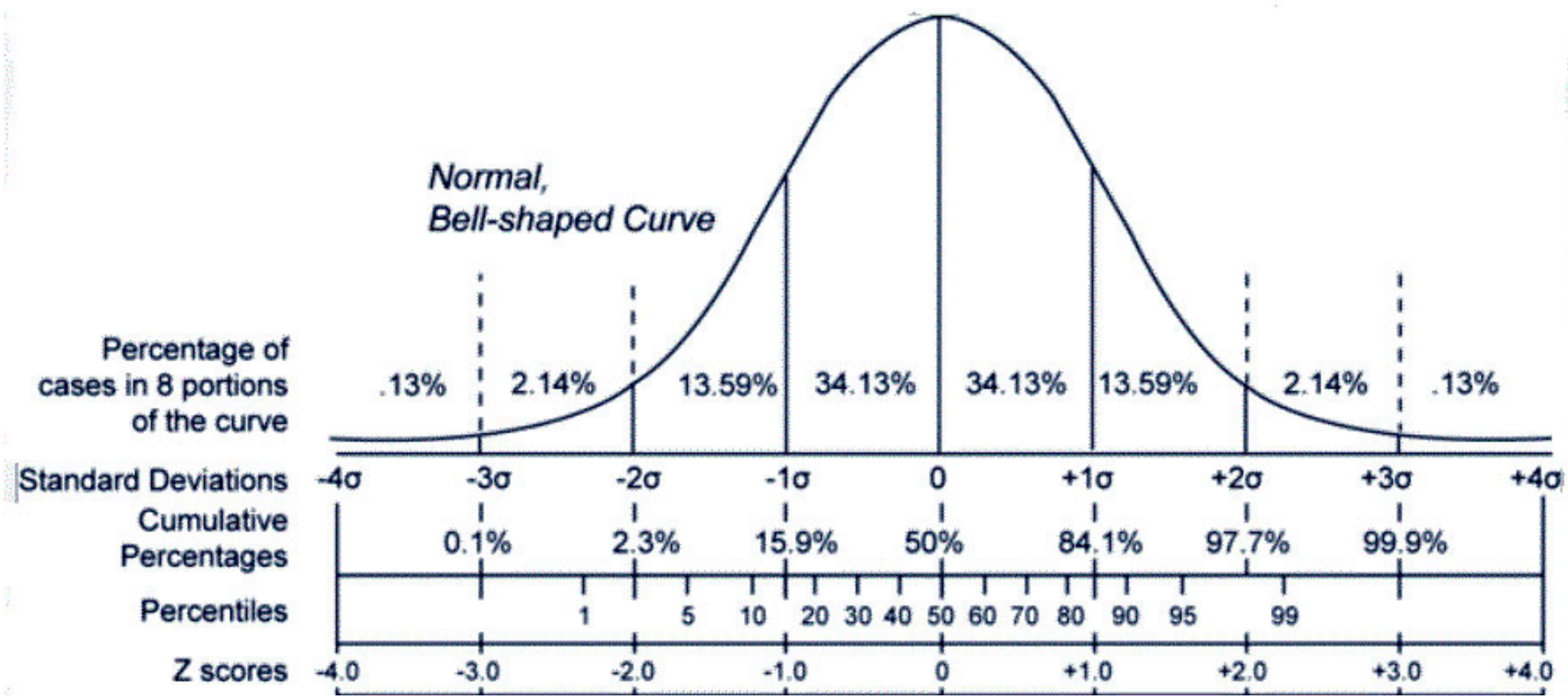
“least likely” 5%

Outlier vs. non-outlier



“least likely” 1%

Quantiles



Normal,
Bell-shaped Curve
Percentage of cases in 8 portions of the curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Percentiles
1
5
10
20
30
40
50
60
70
80
90
95
99
Z scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0

Unsupervised learning

- Classification is an example of supervised learning (where supervision is provided by examples of data points paired with their known categories)

- Unsupervised learning finds *interesting structure* in data.

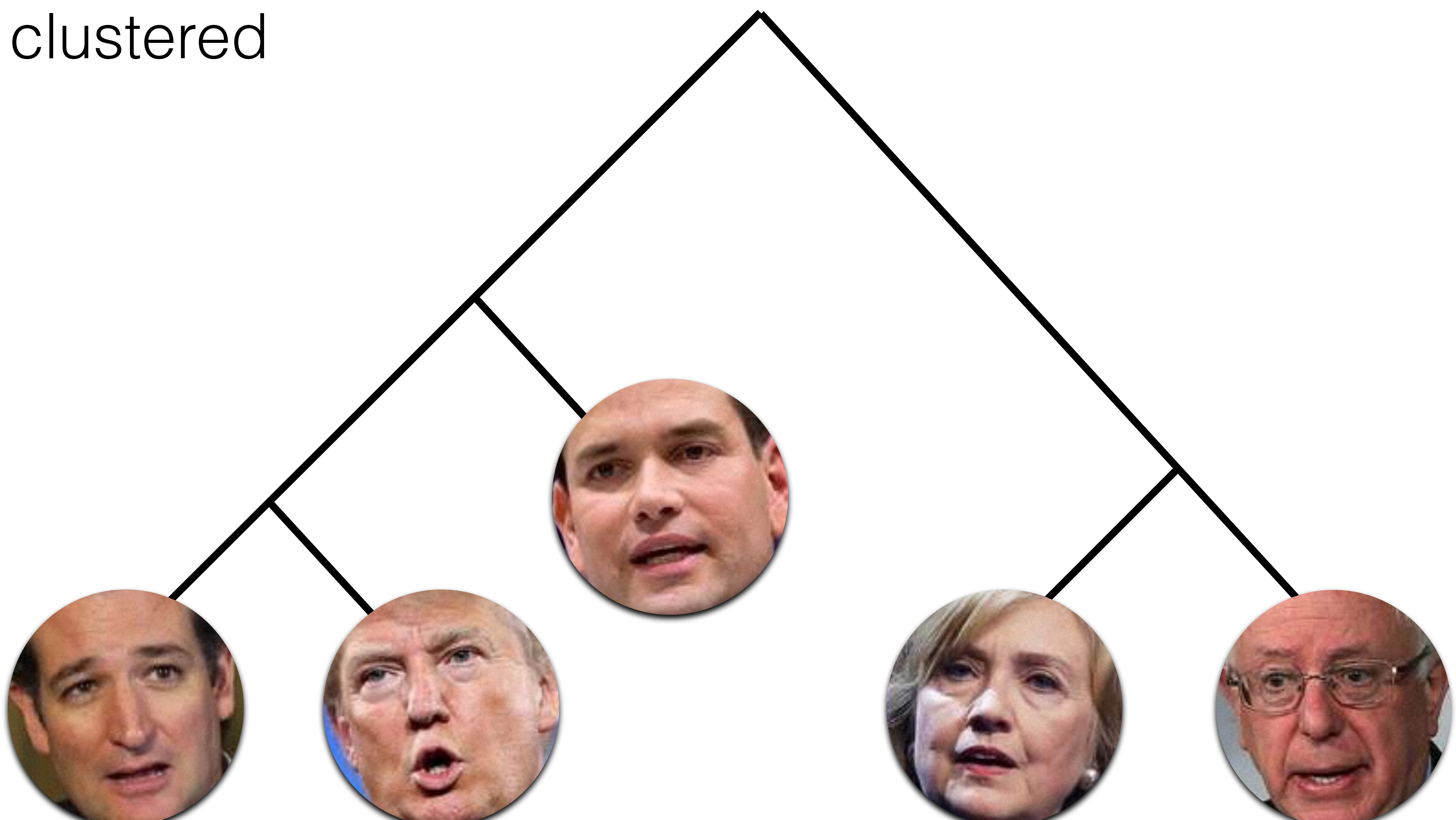
  - clustering data into groups
  - discovering “factors”
  - discovering graph structure

Types of clusters

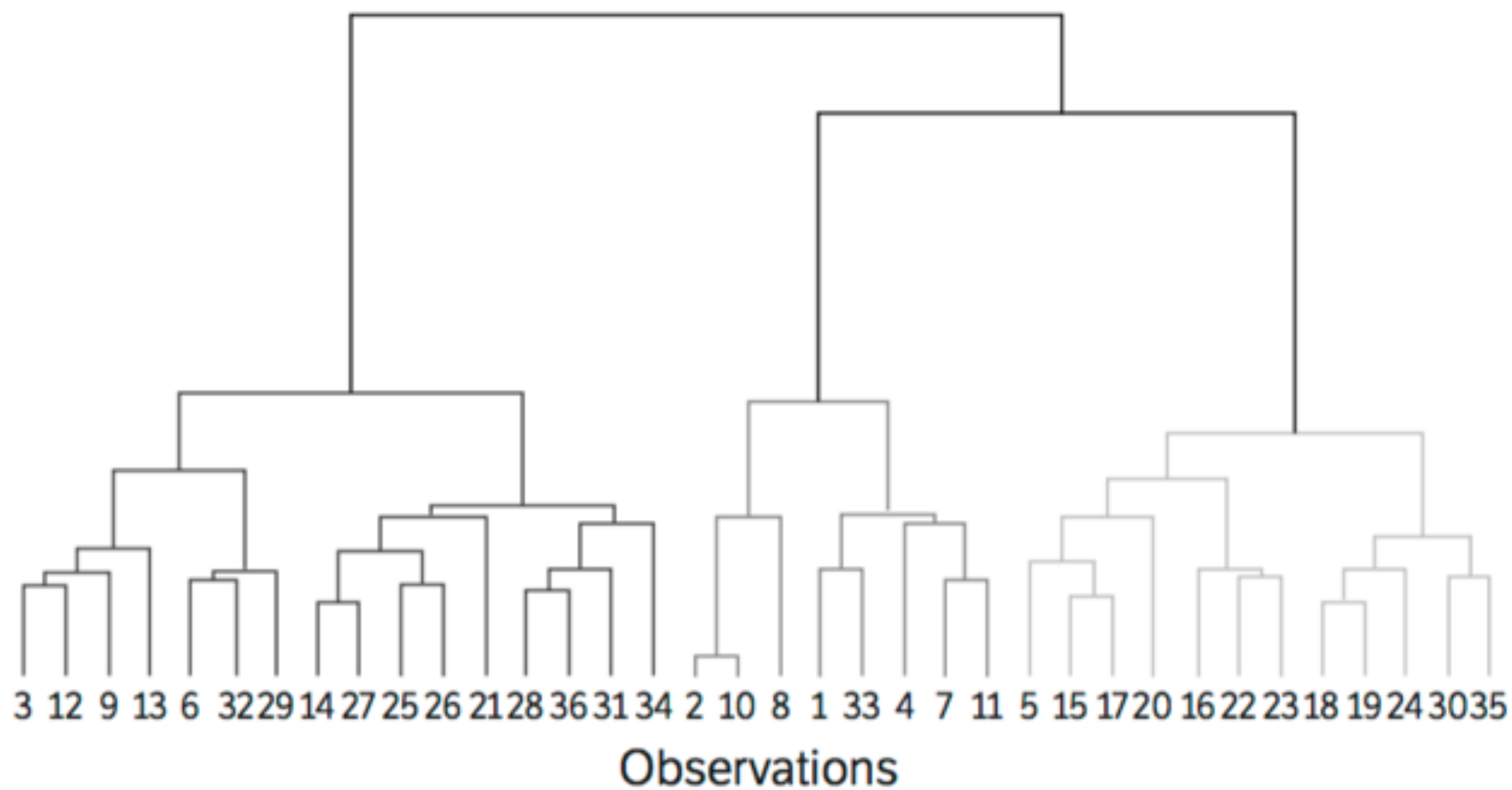
- Many different ways of learning computational categories; the two most common differ in terms of the structure between the categories
    - hierarchical clusters
    - flat clusters

Hierarchical Clustering

- *Hierarchical* order among the elements being clustered



Hierarchical clustering



A Midsummer Night's Dream (3)
Twelfth Night (12)
Much Ado About Nothing (9)
Two Gentlemen (13)
Measure for Measure (6)
Othello (32)
Julius Caesar (29)

The Winter's Tale (14)
Cymbeline (27)
Antony and Cleopatra (25)
Coriolanus (26)
Henry VIII (21)
Hamlet (28)
Troilus and Cressida (36)
Macbeth (31)
Timon of Athens (34)

All's Well That Ends Well (2)
Taming of the Shrew (10)
Merry Wives of Windsor (8)
A Midsummer Night's Dream (1)
Romeo and Juliet (33)
Comedy of Errors (4)
Merchant of Venice (7)
The Tempest (11)

Love's Labours' Lost (5)
1 Henry IV (15)
2 Henry IV (17)
Henry V (20)
1 Henry VI (16)
King John (22)
Richard II (23)

2 Henry VI (18)
2 Henry VI (19)
Richard III (24)
King Lear (30)
Titus Andronicus (35)

Allison et al. 2009

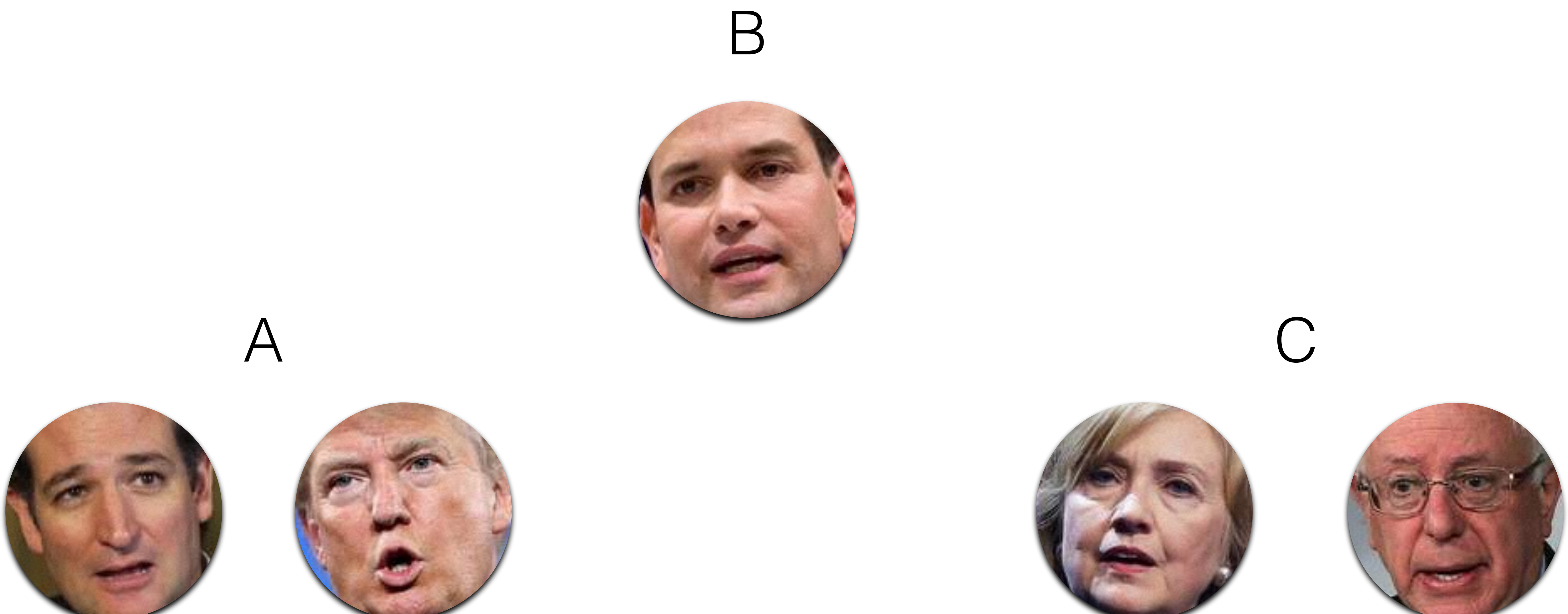
Bottom-up clustering

Algorithm 1 Hierarchical agglomerative clustering

```
Data: N training data points x ∈ ℝ^F
Let X denote a set of objects x
Given some linkage function d(X, X′) → ℝ
Initialize clusters 𝒞 = {C_1, ..., C_N} to singleton data points
while data points not in one cluster do
    Identify X, Y as clusters with smallest linkage function among clusters in 𝒞
    Create new cluster Z = X ∪ Y
    remove X, Y from 𝒞
    add Z to 𝒞
end while
```

Flat Clustering

- Partitions the data into a set of *K* clusters



Flat Clustering

- Partitions the data into a set of *K* clusters

K-means

Algorithm 1 K-means

Data: training data $x \in \mathbb{R}^F$
Given some distance function $d(x, x') \to \mathbb{R}$
Select k initial centers $\{\mu_1, \ldots, \mu_k\}$
while not converged **do**
    for $i = 1$ to N **do**
        Assign x_i to $\arg\min_c d(x_i, \mu_c)$
    end for
    for $i = 1$ to K **do**
        $\mu_i = \frac{1}{D_i} \sum_{j=1}^{D_i} x_i$
    end for
end while

Similarity

- Both hierarchical clustering and flat clustering rely on measuring the similarity between data points
- How you choose to represent a data point (i.e., in terms of which features to describe) will influence the clusters you learn.

Thomas Jefferson's Library
Library of Congress

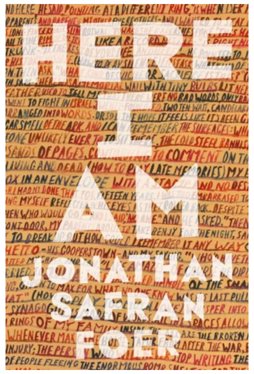
Latent variables

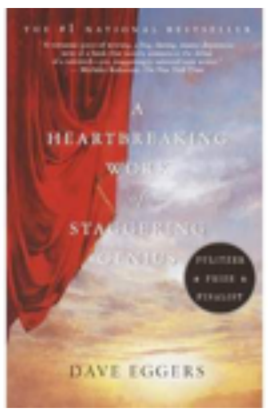
- A latent variable is one that's unobserved, either because:
    - we are predicting it (but have observed that variable for other data points)
    - it is unobservable

Latent variables

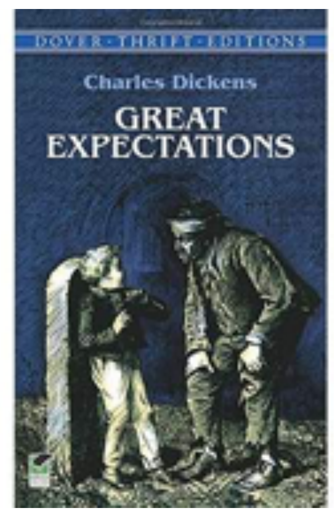
	observed variables	latent variables
email	text, date, sender	
novels		
social network		
fitbit data		
legislators		
netflix users		

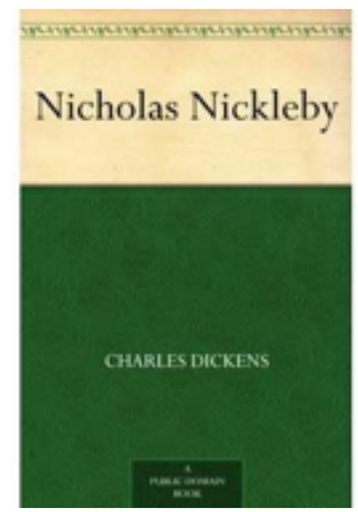
Example: clustering





cluster A

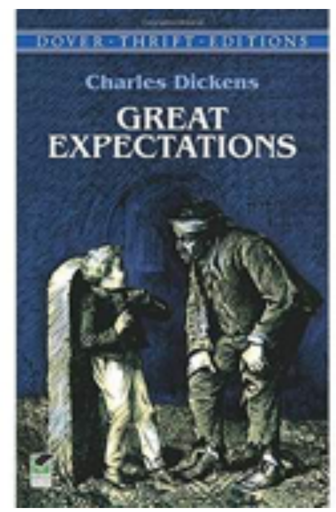


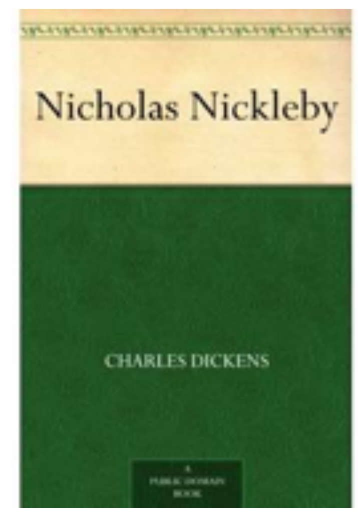


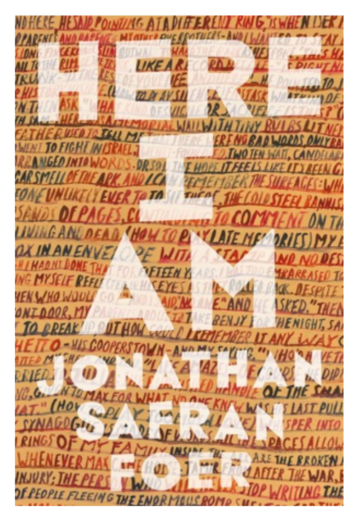
cluster B

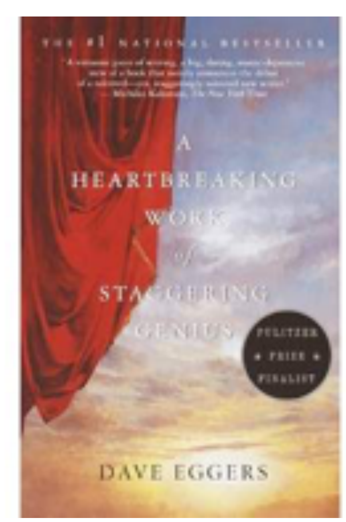
Example: clustering

Resource description matters for inferring computational descriptions too





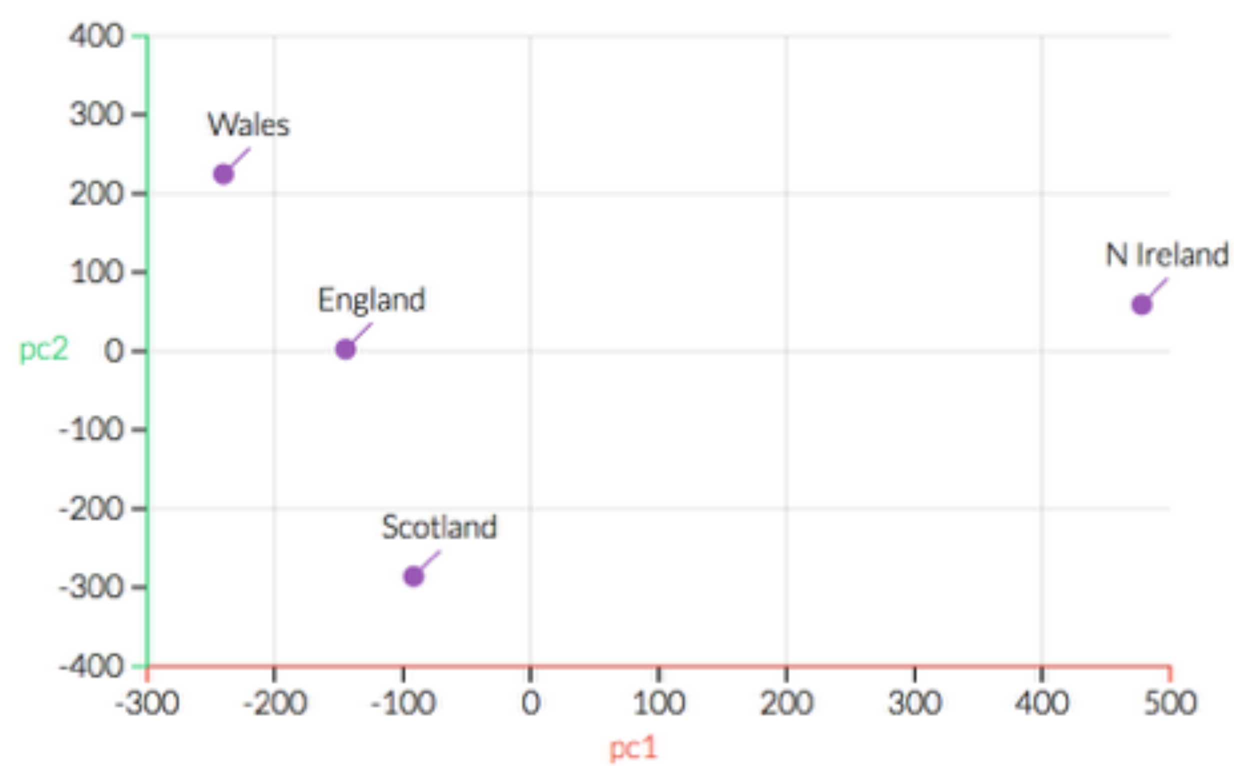




Principle Component Analysis

Method for transforming a set of original (possible correlated) observations into new (uncorrelated) values.

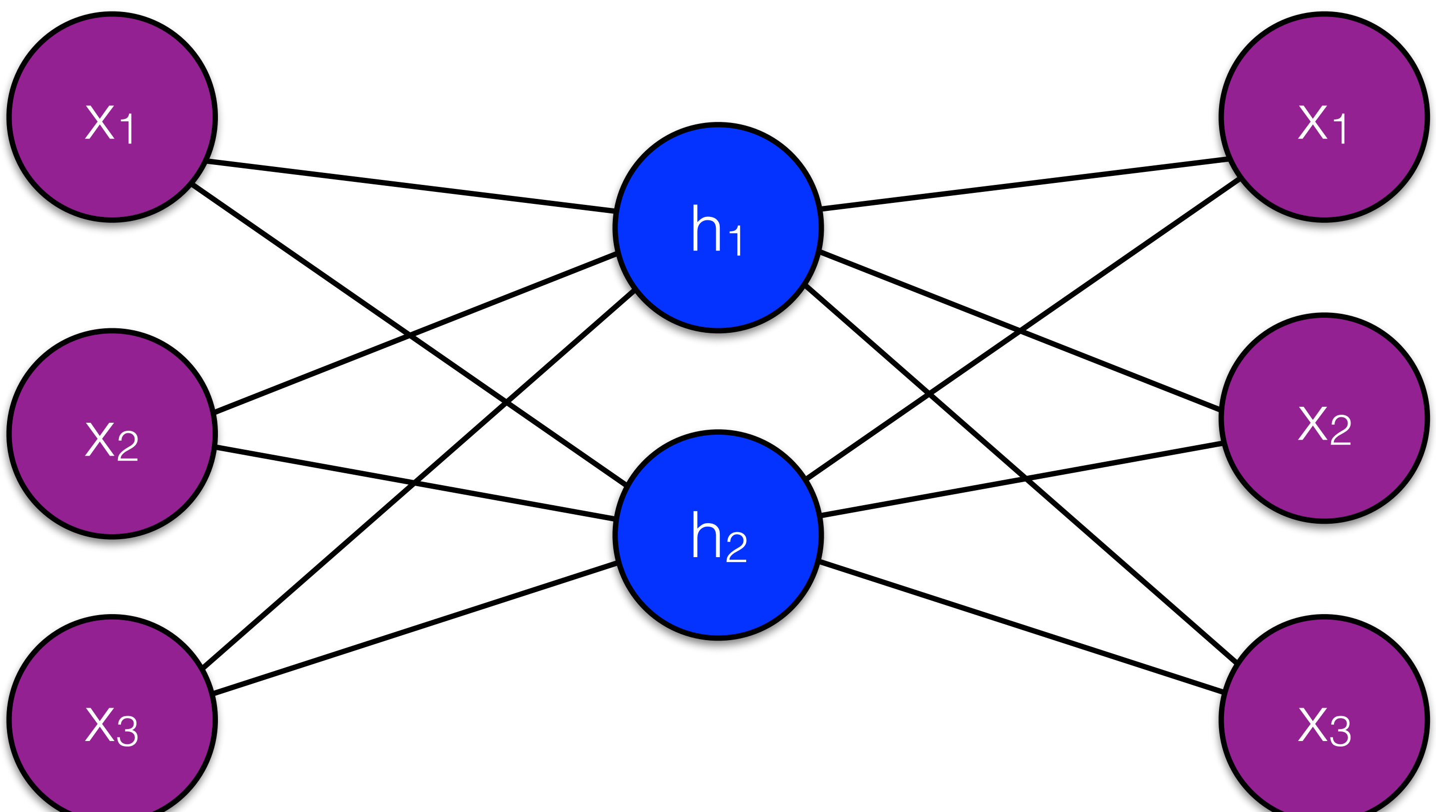
	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

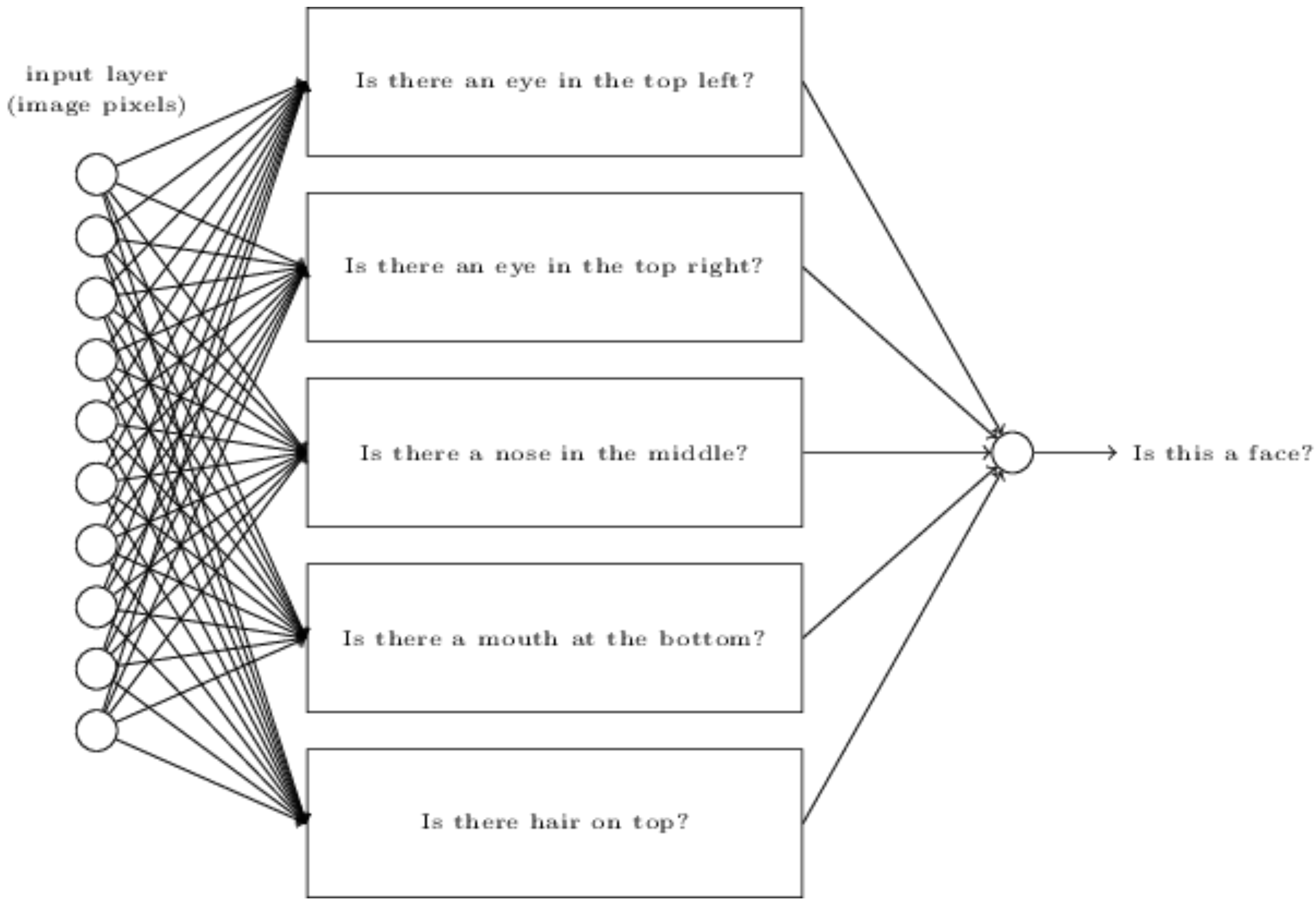


http://setosa.io/ev/principal-component-analysis/

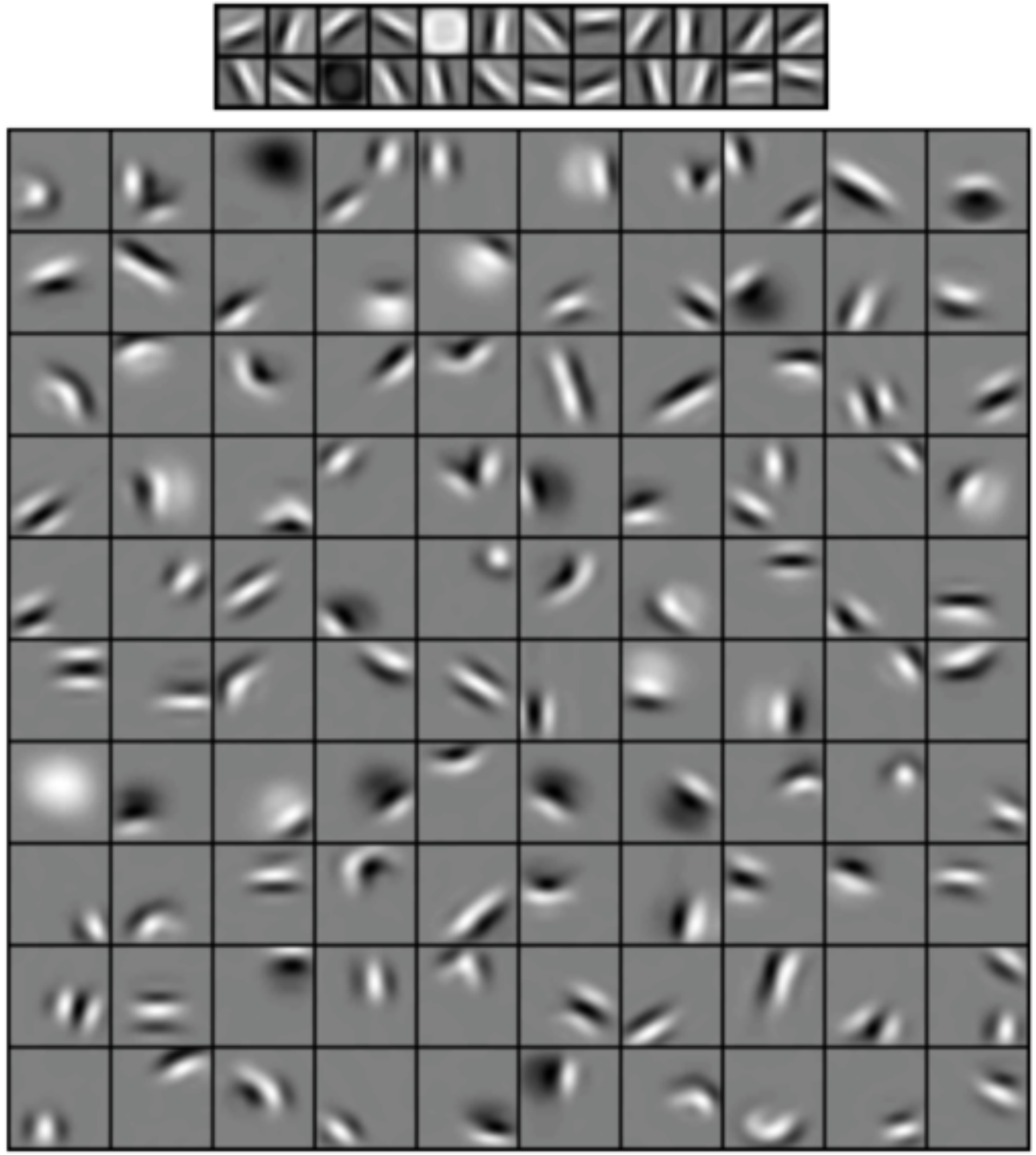
Unsupervised neural networks

- Learns a low-dimensional representation of x by predicting itself





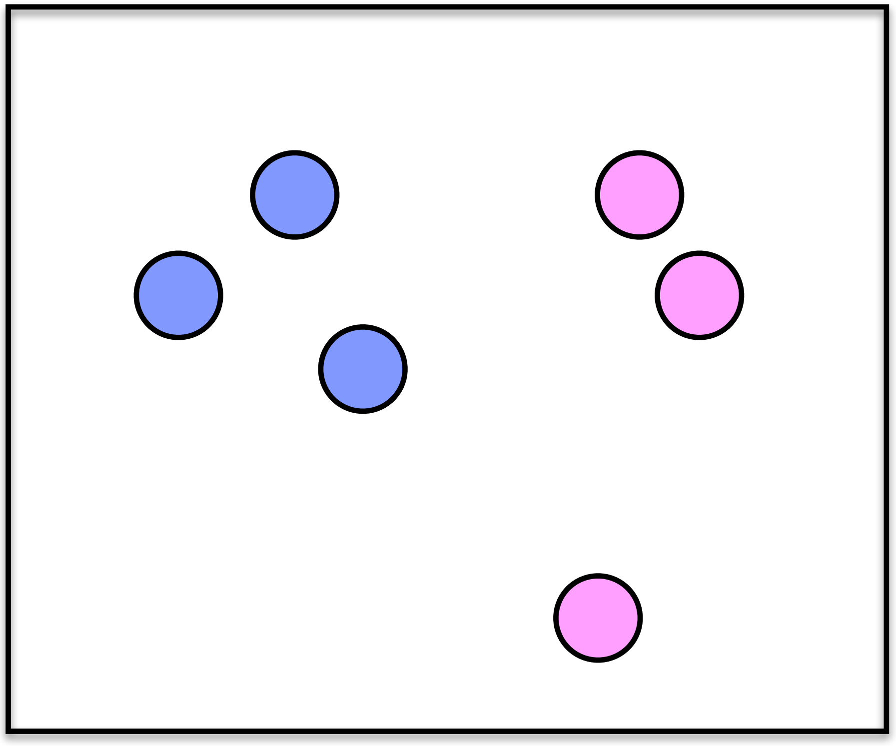
http://neuralnetworksanddeeplearning.com/chap1.html

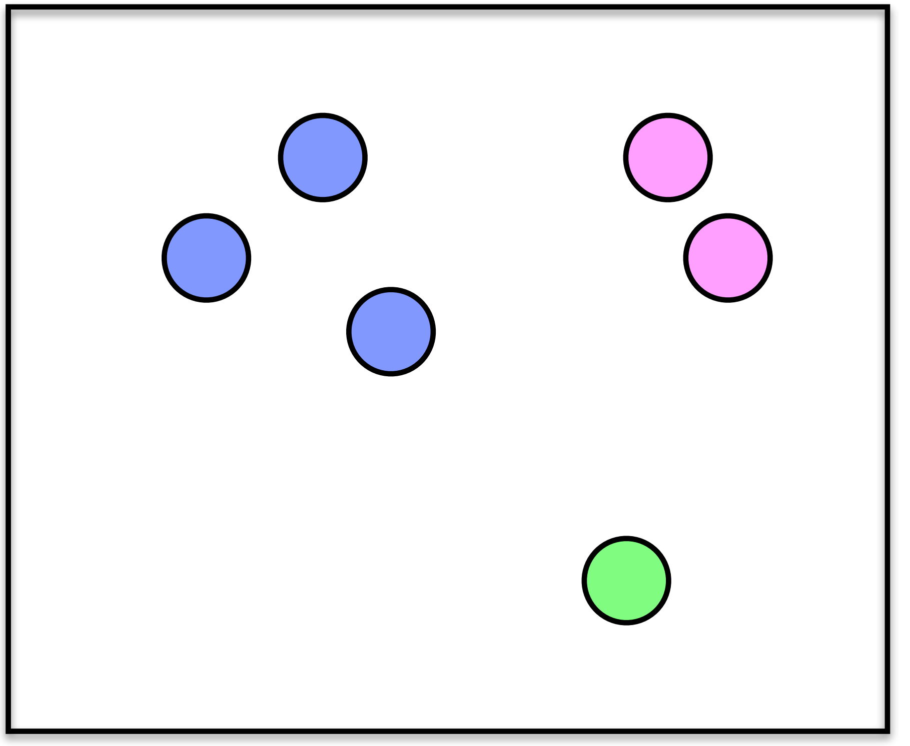


Higher order features learned for image recognition
Lee et al. 2009 (ICML)

Evaluation

- How do you know when a clustering is valid?
- Much more complex than supervised classification since there's often no notion of "truth"





Internal criteria

- Elements within clusters should be more similar to each other
- Elements in different clusters should be less similar to each other

External criteria

- How closely does your clustering reproduce another (“gold standard”) clustering?

Interpretability

- Good human-created categories generally have interpretable semantics, and high agreement rates between annotators

- When inferring categories through clustering, it’s often difficult to interpret what commonalities it’s learning between data points.

Clustering → classification

- Clustering can interact with classification in several ways
- Automatically inferred clusters can become the raw material for manual refinement
- Assignment of data points to clusters can act as features for downstream classification

id	Age mean	Age s.d.	% Fem.	words
14	95.29	42.65	12.1	statue, unveiled, memorial, plaque, anniversary, erected, monument, death, bronze, memory
472	92.39	46.06	13.0	national, historic, park, state, house, named, memorial, home, honor, museum
369	83.62	35.79	15.0	stamp, named, issued, team, century, australian, series, anniversary, service, rugby league
208	82.66	41.43	21.8	film, portrayed, played, based, movie, actor, novel, character, life, starring
179	81.98	40.39	23.8	film, music, museum, book, released, work, published, history, american, part
250	81.39	45.11	11.4	wrote, book, death, years, time, said, made, work, john, history
262	80.84	9.95	13.8	died, age, death, home, california, aged, new york, hospital, cancer, heart attack
446	77.3	36.30	16.9	published, biography, book, wrote, life, press, john, edited, written, work

- $260 average order
- duration: 5 years
- frequently bought categories: furniture, kitchen appliances

- $13 average order
- duration: 21 days
- frequently bought categories: books

Predict whether an individual will make a purchase next week; inferred clusters as features allow you to back off to others with similar behavior