Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 5: Clustering overview

Feb 3, 2016

Clustering

- Clustering (and unsupervised learning more generally) finds *structure* in data, using just X

X = a set of skyscrapers

Unsupervised Learning

- Matrix completion (e.g., user recommendations on Netflix, Amazon)

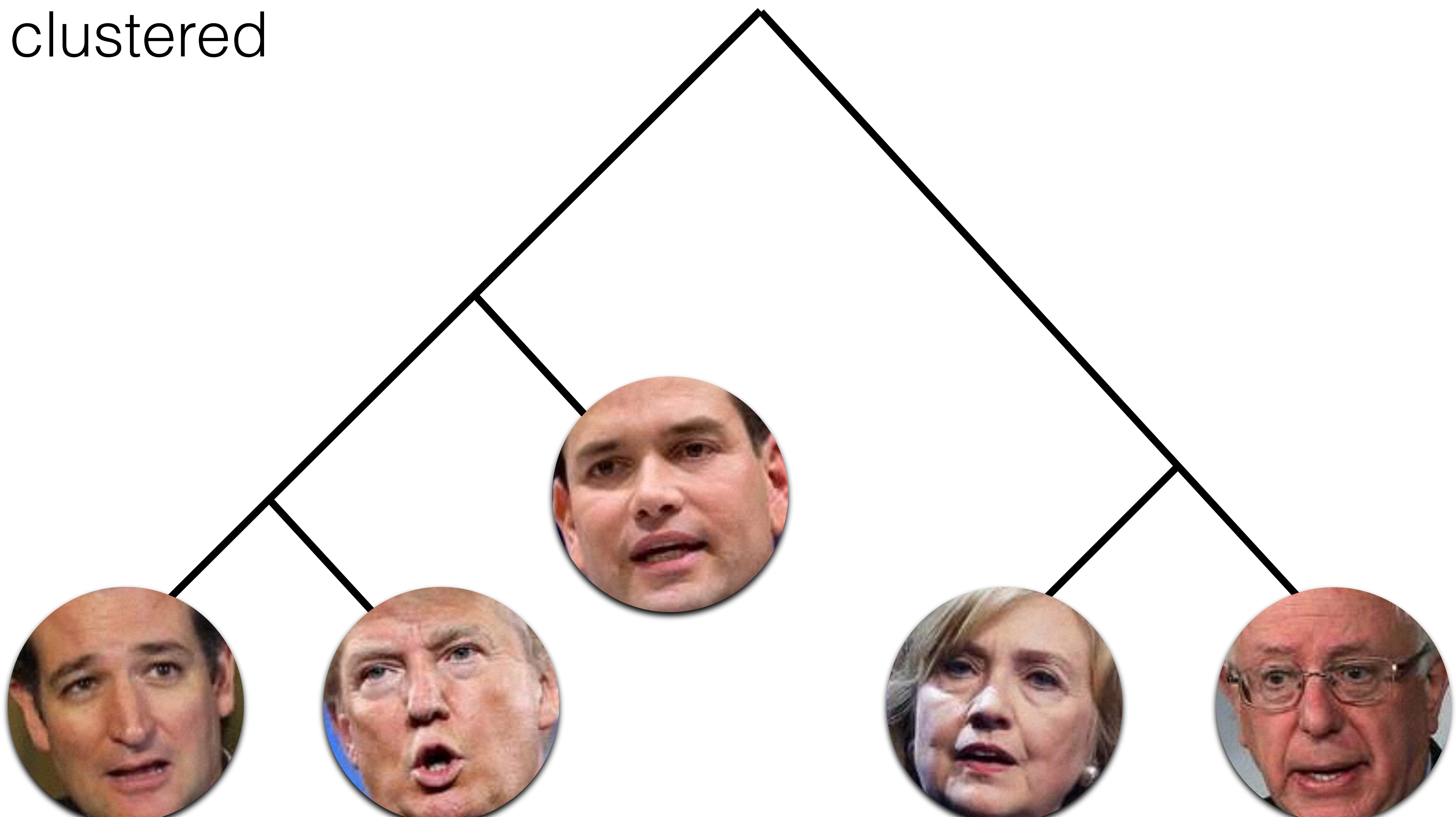
	Ann	Bob	Chris	David	Erik
Star Wars	5	5	4	5	3
Bridget Jones		4		4	1
Rocky	3		5		
Rambo		?		2	5

task	x
learn patterns that define architectural styles	set of skyscrapers
learn patterns that define genre	set of books
learn patterns that suggest “types” of customer behavior	customer data

Methods differ in the kind of structure learned

Hierarchical Clustering

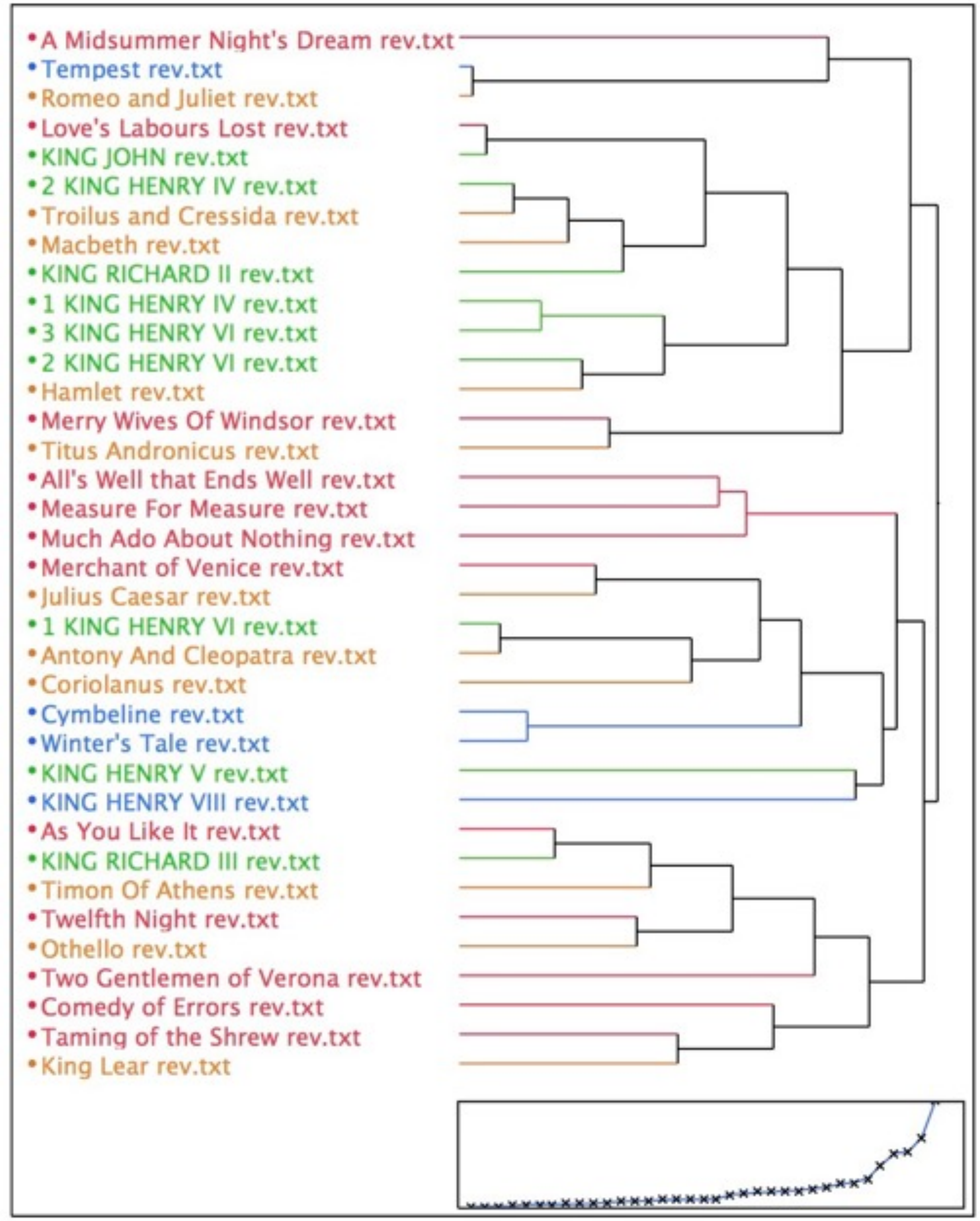
- *Hierarchical* order among the elements being clustered



Dendrogram

Shakespeare's plays

Witmore (2009) http://winedarksea.org/?p=519



Bottom-up clustering

Given: a set $\mathcal{X} = \{x_1, \ldots x_n\}$ of objects

a function sim: $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$

for $i := 1$ **to** n **do**

$\quad c_i := \{x_i\}$ **end**

$C := \{c_1, \ldots, c_n\}$

$j := n + 1$

while $C > 1$

$\quad (c_{n_1}, c_{n_2}) := \arg\max_{(c_u, c_v) \in C \times C} \mathrm{sim}(c_u, c_v)$

$\quad c_j = c_{n_1} \cup c_{n_2}$

$\quad C := C \backslash \{c_{n_1}, c_{n_2}\} \cup \{c_j\}$

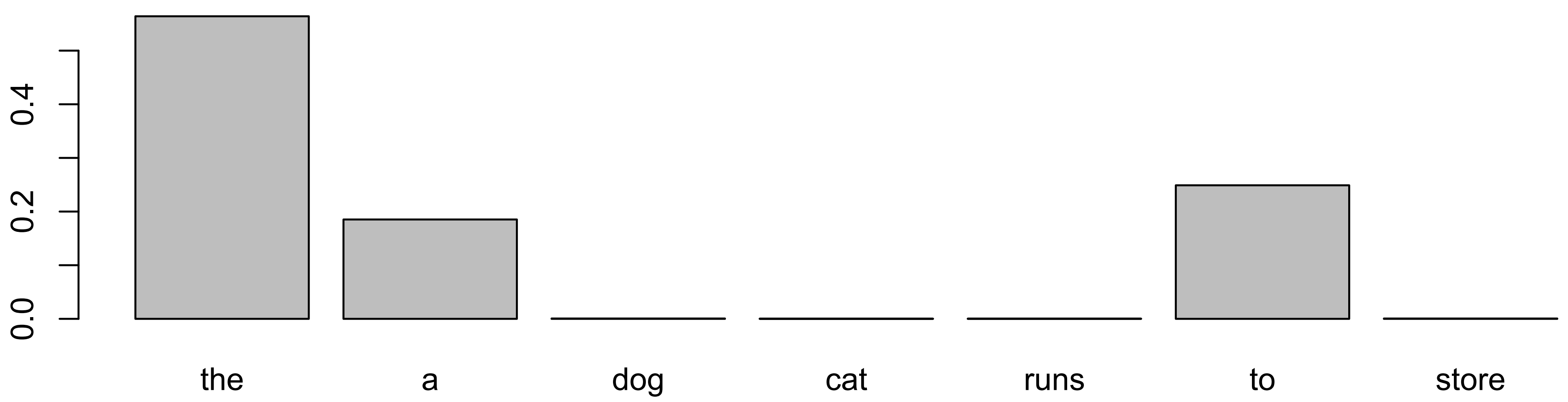
$\quad j := j + 1$

Similarity

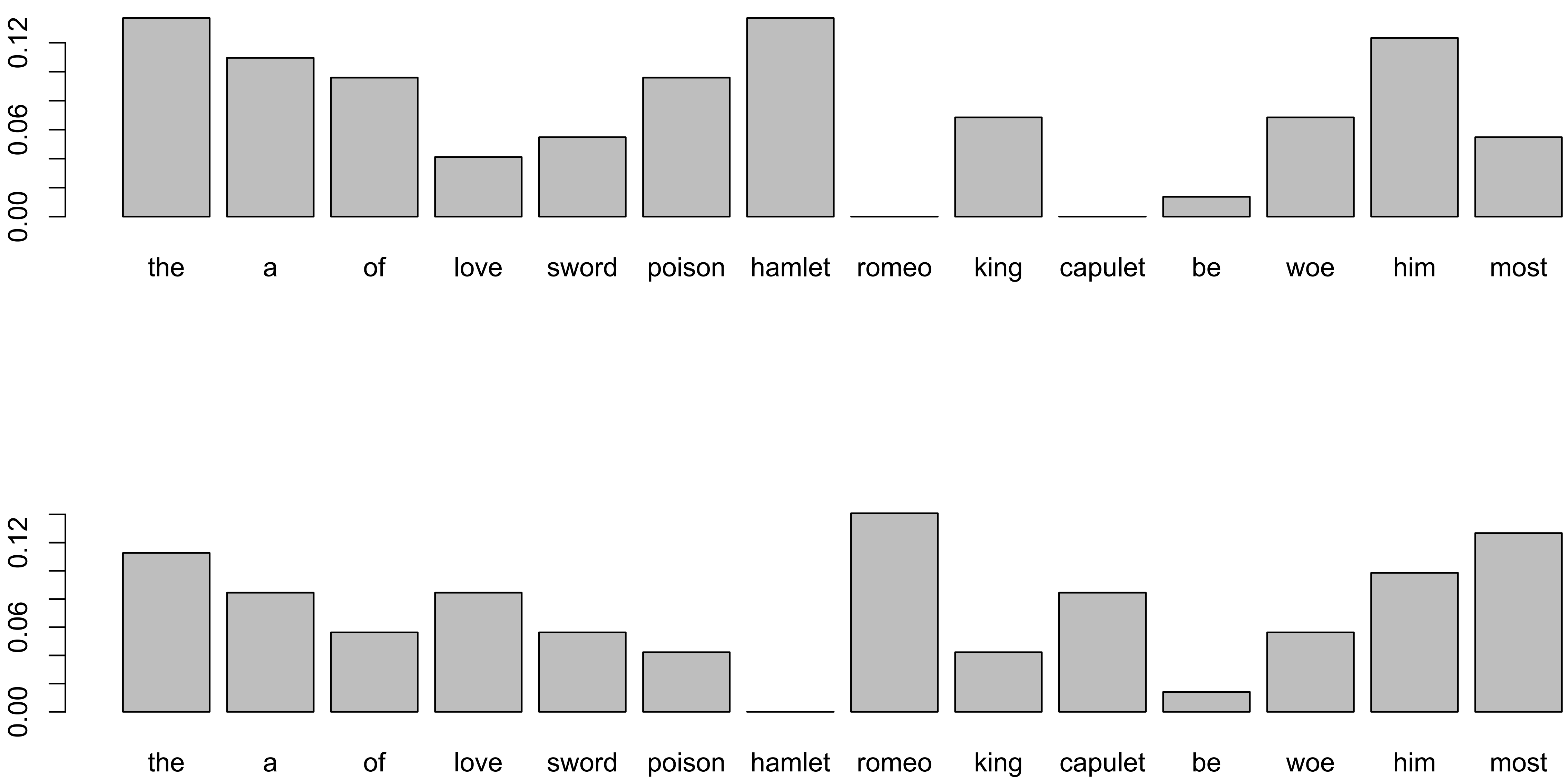
$$\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$$

- What are you comparing?
- How do you quantify the similarity/difference of those things?

Probability



Unigram probability



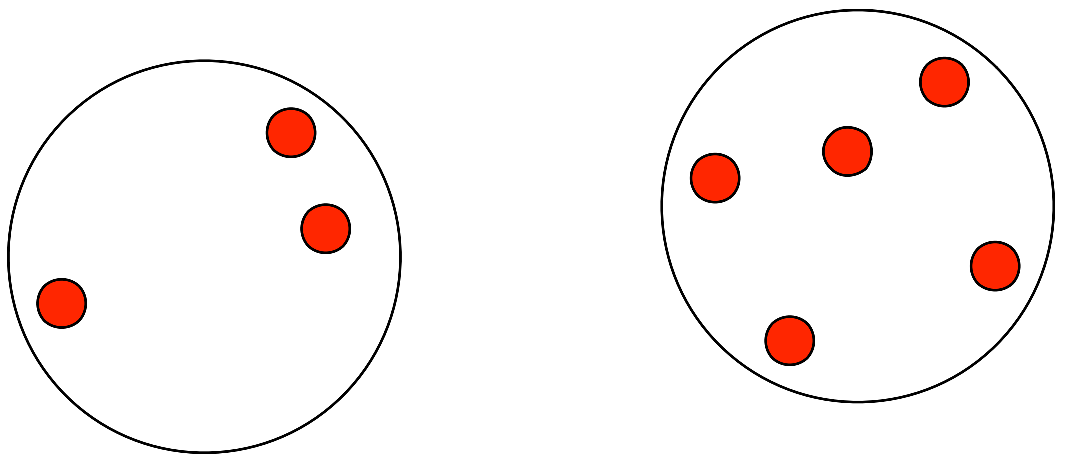
Similarity

$$\text{Euclidean} = \sqrt{\sum_{i}^{vocab} \left(P_i^{\text{Hamlet}} - P_i^{\text{Romeo}}\right)^2}$$

Cosine similarity, Jensen-Shannon divergence…

Cluster similarity

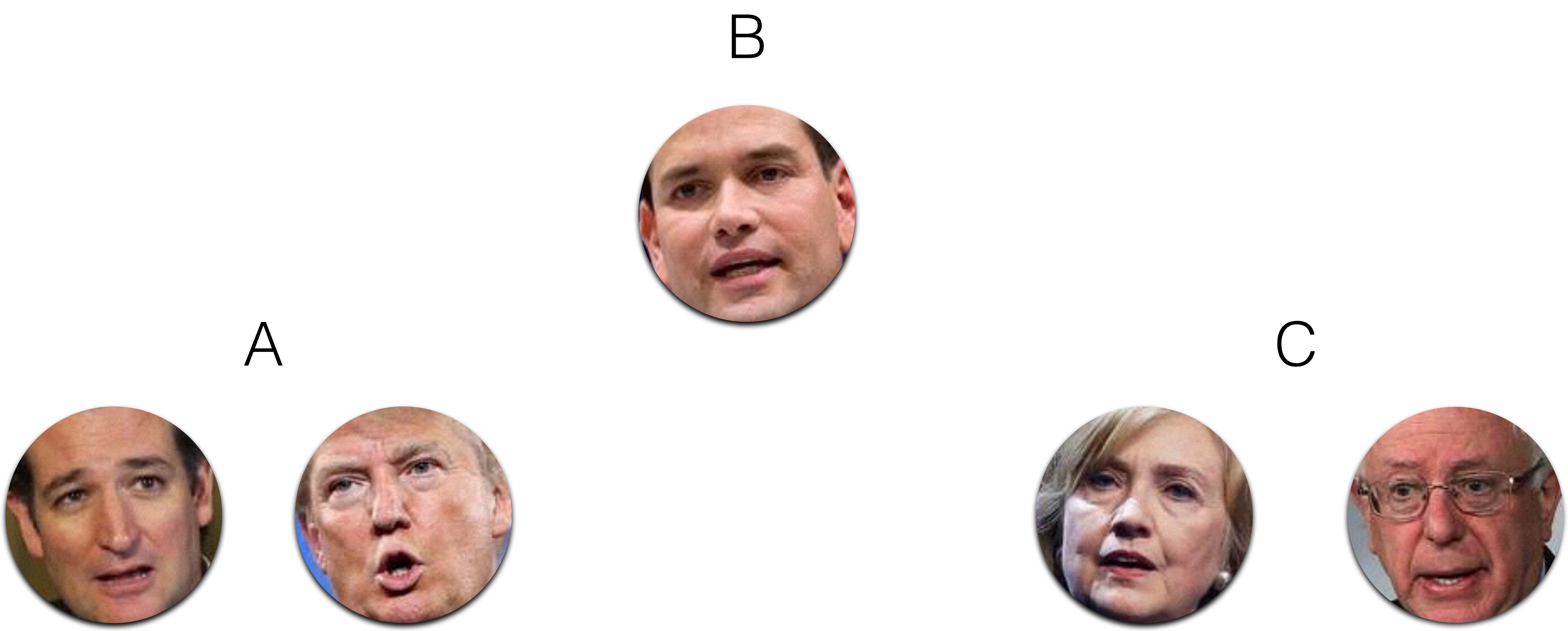
Cluster similarity



- Single link: two **most** similar elements
- Complete link: two **least** similar elements
- Group average: average of all members

Flat Clustering

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



Flat Clustering

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

K-means

Given: a set $\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_n\} \subseteq \mathbb{R}^m$

  a distance measure $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$

  a function for computing the mean $\mu : \mathcal{P}(\mathbb{R}) \to \mathbb{R}^m$

Select k initial centers $\vec{f}_1, \ldots, \vec{f}_k$

while stopping criterion is not true **do**

  for all clusters c_j **do**

    $c_j = \{\vec{x}_i \mid \forall \vec{f}_l \; d(\vec{x}_i, \vec{f}_j) \leq d(\vec{x}_i, \vec{f}_l)\}$

  end

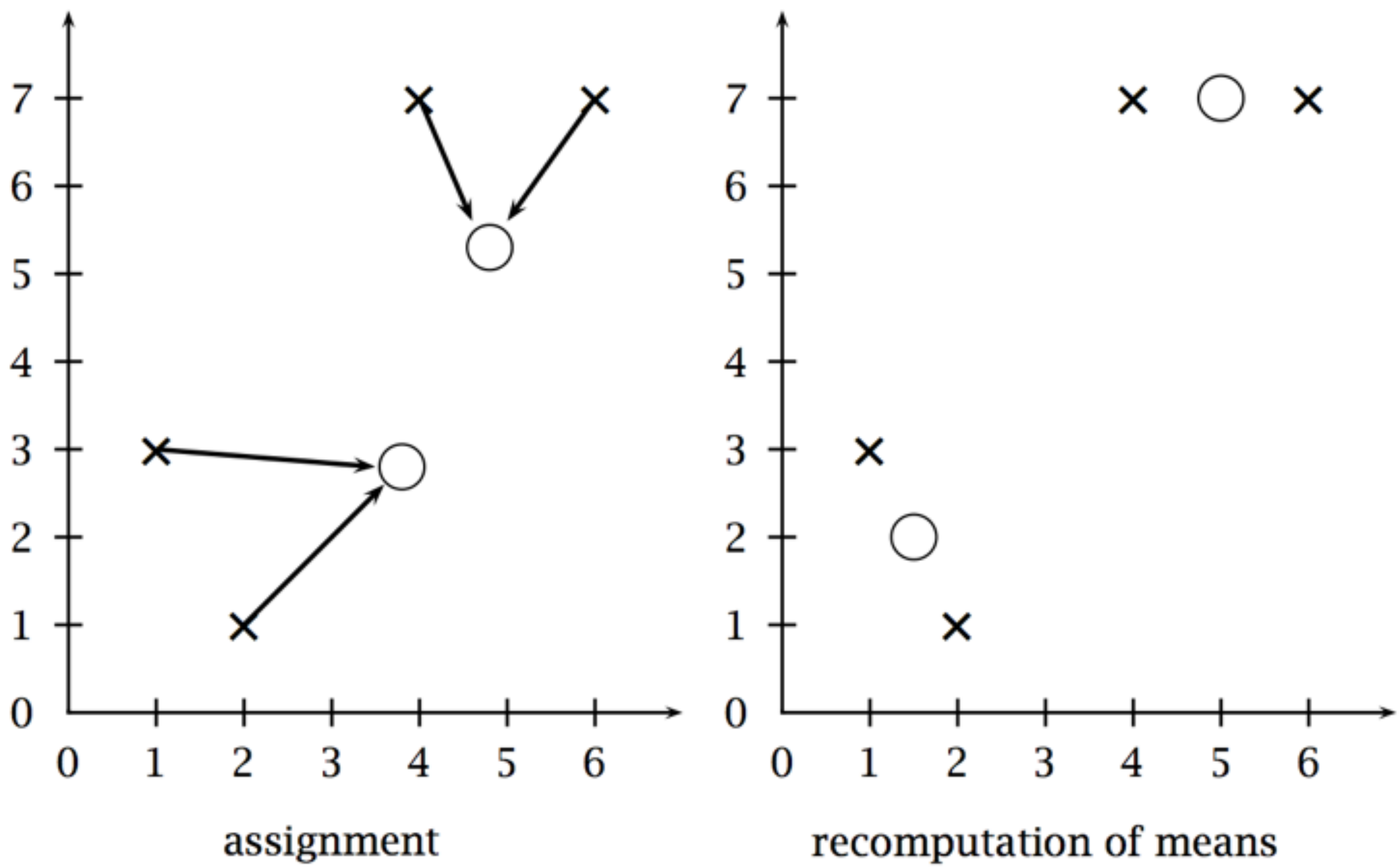
  for all means $\vec{f}_j$ **do**

    $\vec{f}_j = \mu(c_j)$

  end

end

K-means



Representation

$$x \in \mathbb{R}^F$$

[x is a data point characterized by F real numbers, one for each feature]

- This is a huge decision that impacts what you can learn





Voting behavior


Yes on abortion access	1
Yes on expanding gun rights	0
Yes on tax breaks	0
Yes on ACA	1
Yes on abolishing IRS	0

$$x \in \mathbb{R}^5$$

First letter of last name

Last name starts with < “A”	0
Last name starts with < “B”	0
Last name starts with < “C”	1
Last name starts with < “D”	1
…	1
Last name starts with < “Z”	1

$$x \in \mathbb{R}^{26}$$

Representation

task	x
learn patterns that define architectural styles	set of skyscrapers
learn patterns that define genre	set of books
learn patterns that suggest “types” of customer behavior	customer data

Evaluation

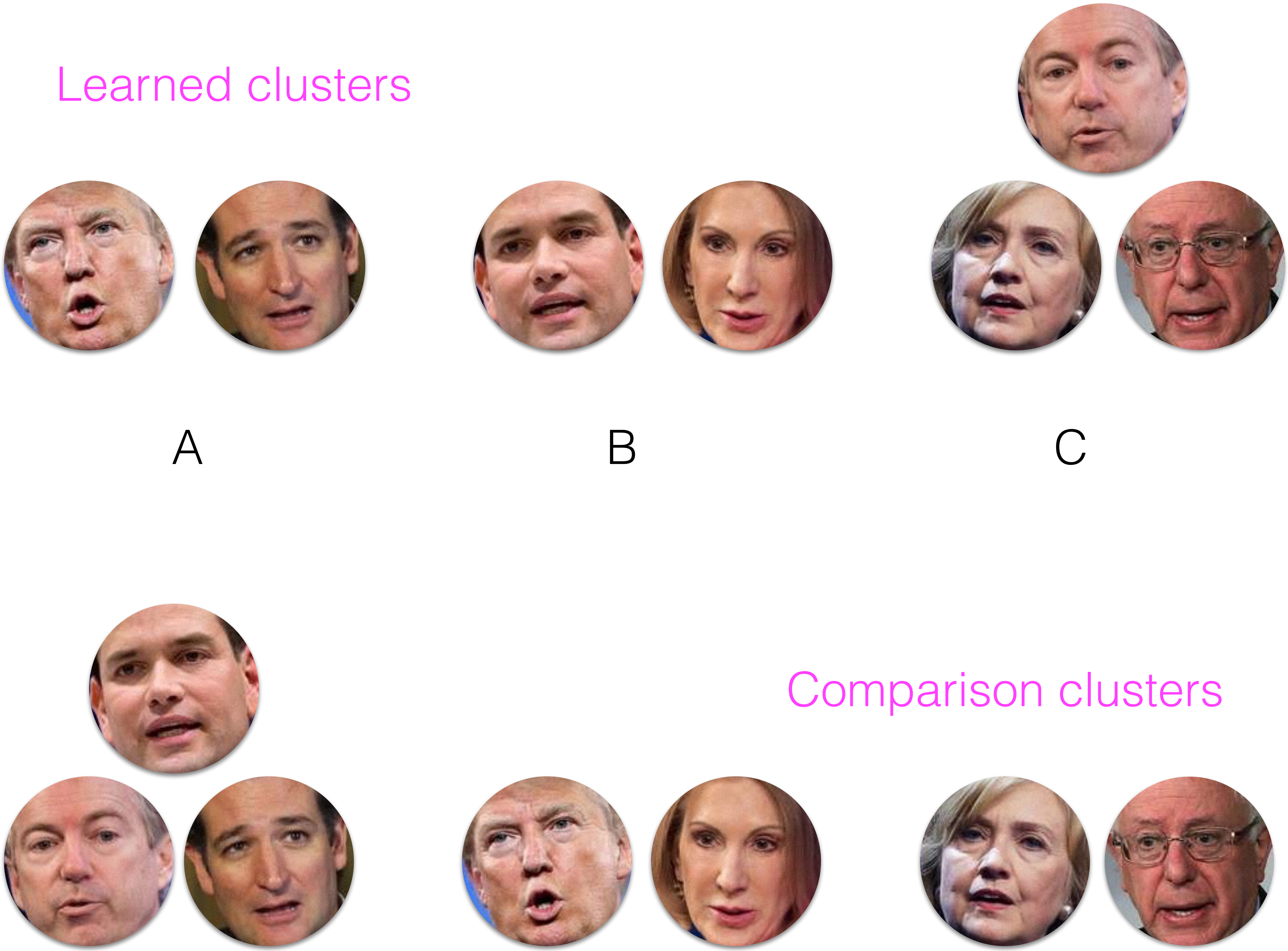
- Much more complex than supervised learning since there’s often no notion of “truth”

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?


Learned clusters
A
B
C
Comparison clusters

Evaluation: Purity

- Learned clusters
  (as learned by our algorithm)

$$\mathcal{G} = \{g_1 \ldots g_k\}$$

- External clusters
  (from some external source)

$$\mathcal{C} = \{c_1 \ldots c_j\}$$

$$\text{Purity} = \frac{1}{N} \sum_k \max_j |g_k \cap c_j|$$

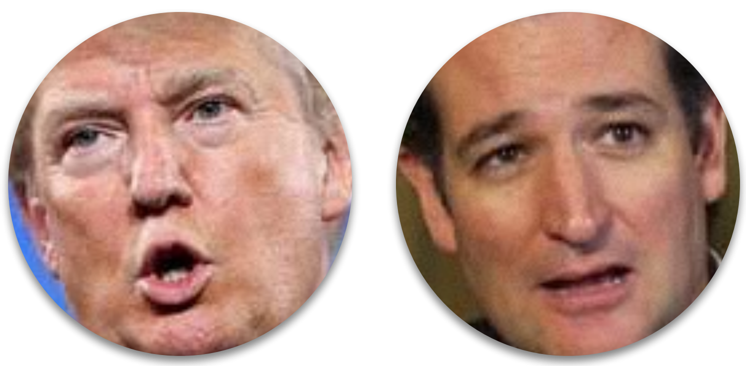
Learned (G)

A B C

$$= \frac{1}{N} \sum_{k} \max_{j} |g_k \cap c_j|$$

External (C)

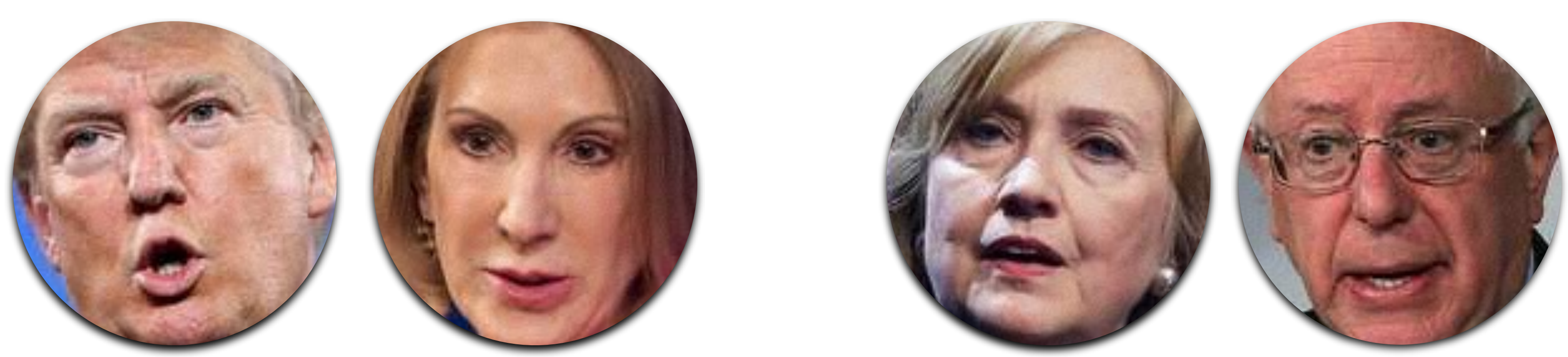
Learned (G)



A B C

$$= \frac{1}{N} \sum_k \max_j |g_k \cap c_j|$$

External (C)



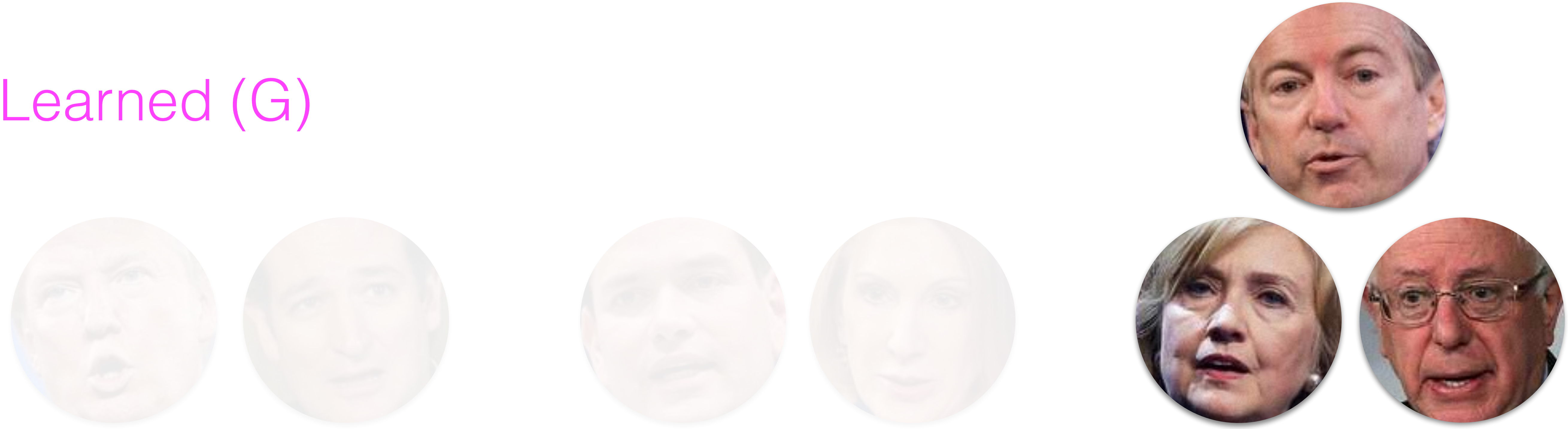
Learned (G)

A B C

$$= \frac{1}{N} \sum_{k} \max_{j} |g_k \cap c_j|$$

External (C)

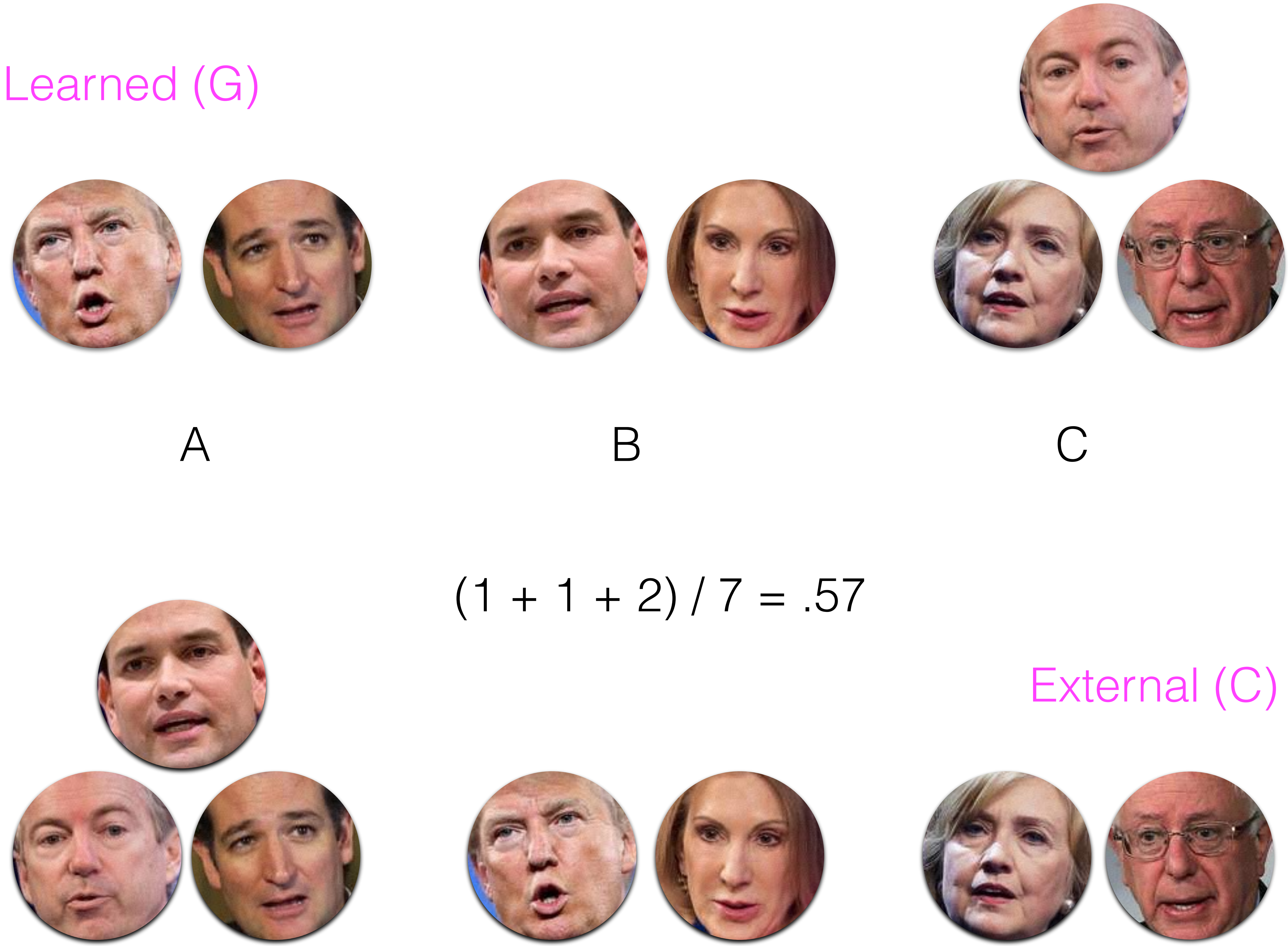
Learned (G)



A B C

$$= \frac{1}{N} \sum_k \max_j |g_k \cap c_j|$$

External (C)

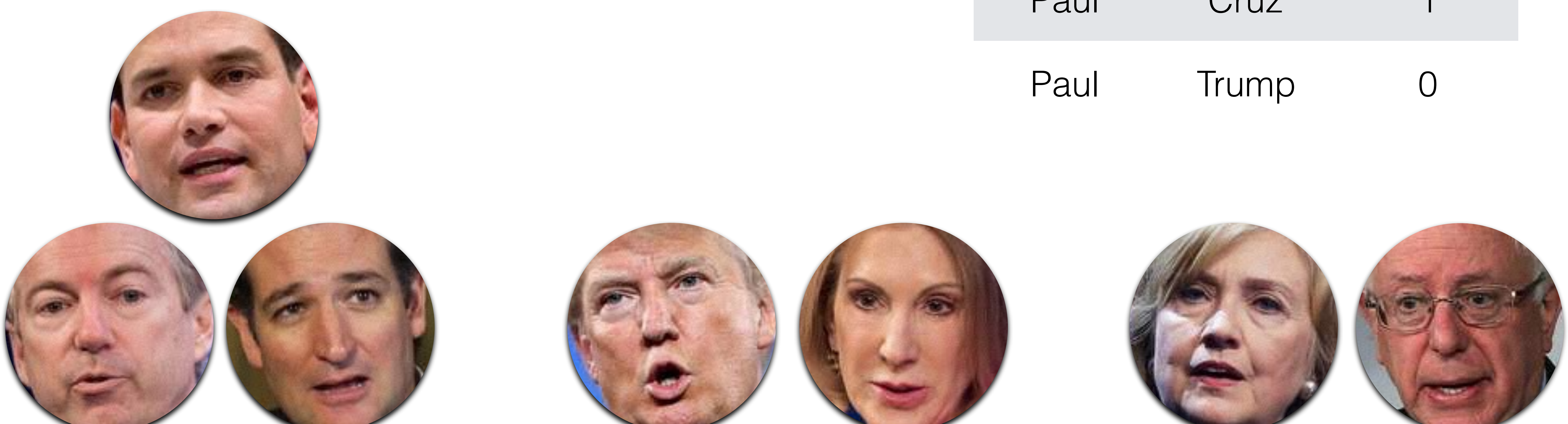

Learned (G)
A
B
C
(1 + 1 + 2) / 7 = .57
External (C)

Evaluation: Rand Index

Every pair of data points is either in the same external cluster, or it's not. = binary classification

Rand Index

		same cluster?
Rubio	Paul	1
Rubio	Cruz	1
Rubio	Trump	0
Rubio	Fiorina	0
Rubio	Clinton	0
Rubio	Sanders	0
Paul	Cruz	1
Paul	Trump	0



Rand Index

Predicted (ŷ)

True (y)	same cluster	different cluster
same cluster		
different cluster		

21 decisions

$$N(N-1)/2$$

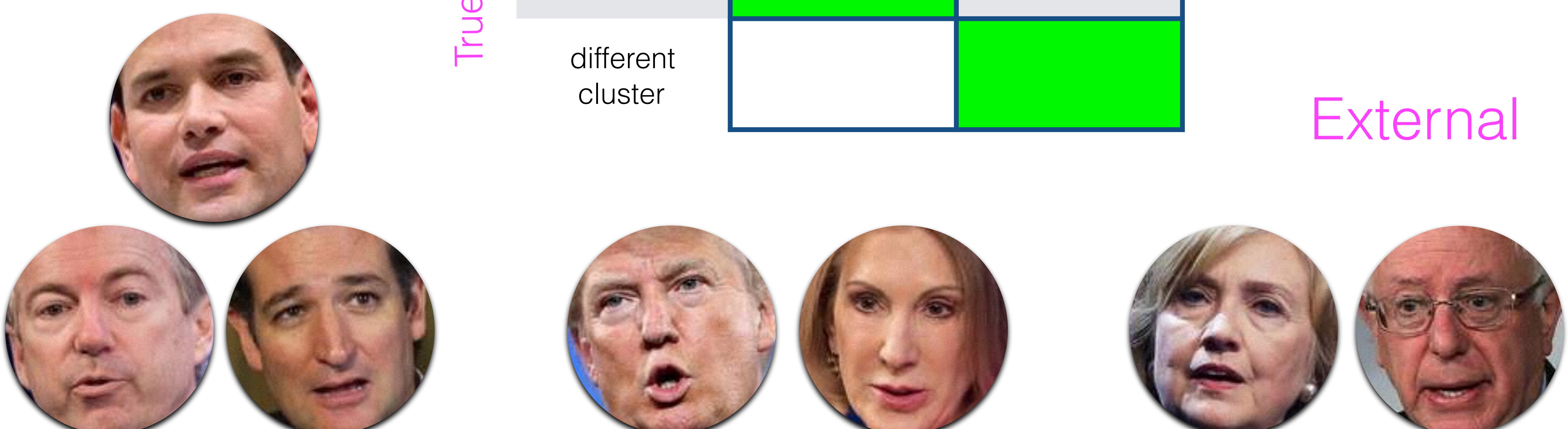
Learned

Predicted ($\hat{y}$)

True (y)

	same cluster	different cluster
same cluster		
different cluster		

External



Rand Index

From the confusion matrix, we can calculate standard measures from binary classification

The Rand Index = accuracy

(1 + 12) / 21 = .619

		Predicted ($\hat{y}$)	
		same cluster	different cluster
True (y)	same cluster	1	4
	different cluster	4	12

Example

Clustering characters
into distinct types

The Villain

- Does (agent): kill, hunt, severs, chokes
- Has done to them (patient): fights, defeats, refuses
- Is described as (attribute): evil, frustrated, lord

The Villain

- Is character in the movie "Star Wars"
    - Science Fiction, Adventure, Space Opera, Fantasy, Family Film, Action
- Is played by David Prowse
    - Male
    - 42 years old in 1977

Task

Learning character types from textual descriptions of characters.

Data	Source
42,306 movie plot summaries	Wikipedia
15,099 English novels (1700-1899)	HathiTrust

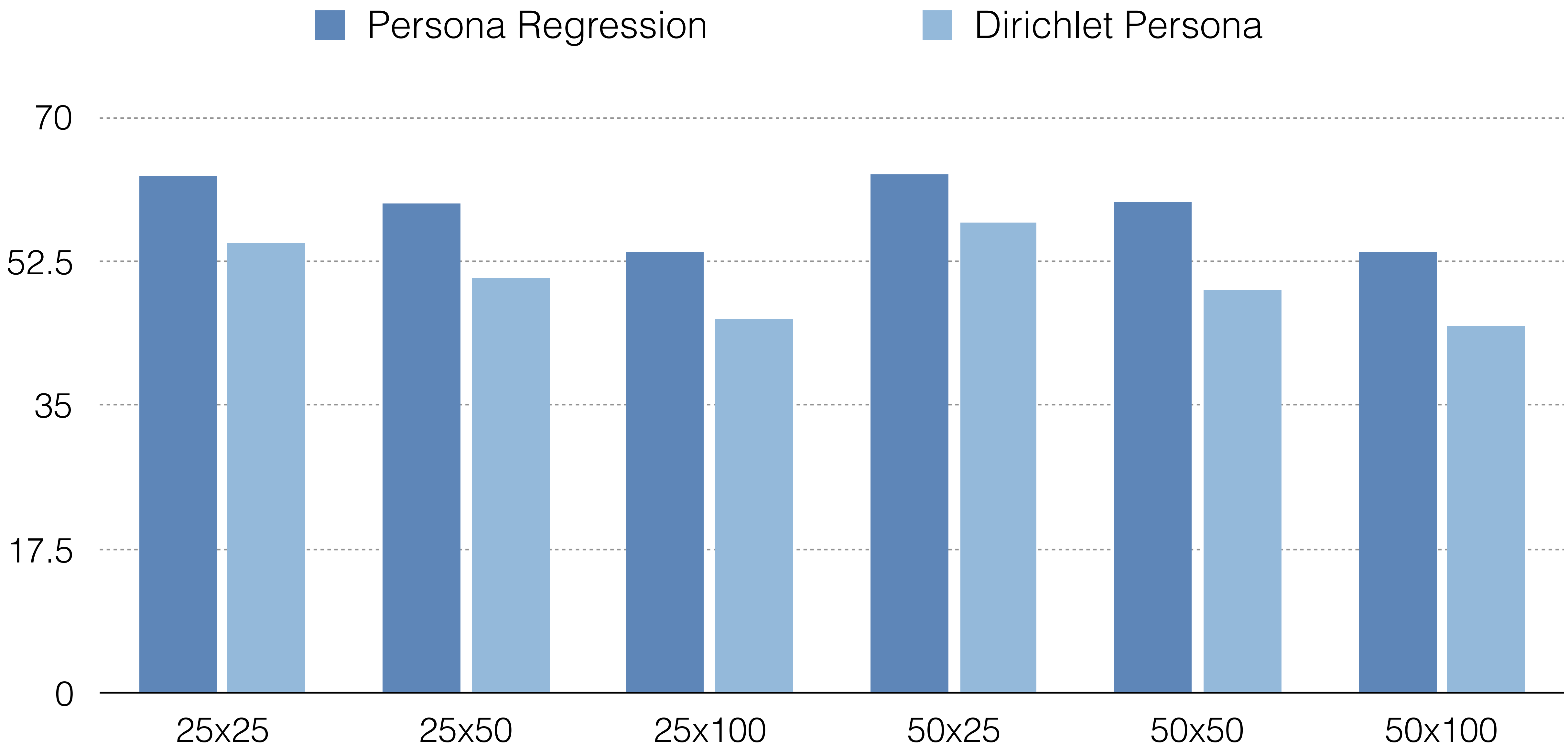
Evaluation I: Names

- Gold clusters: characters with the same name (sequels, remakes)
- Noise: “street thug”
- 970 unique character names used twice in the data; n=2,666

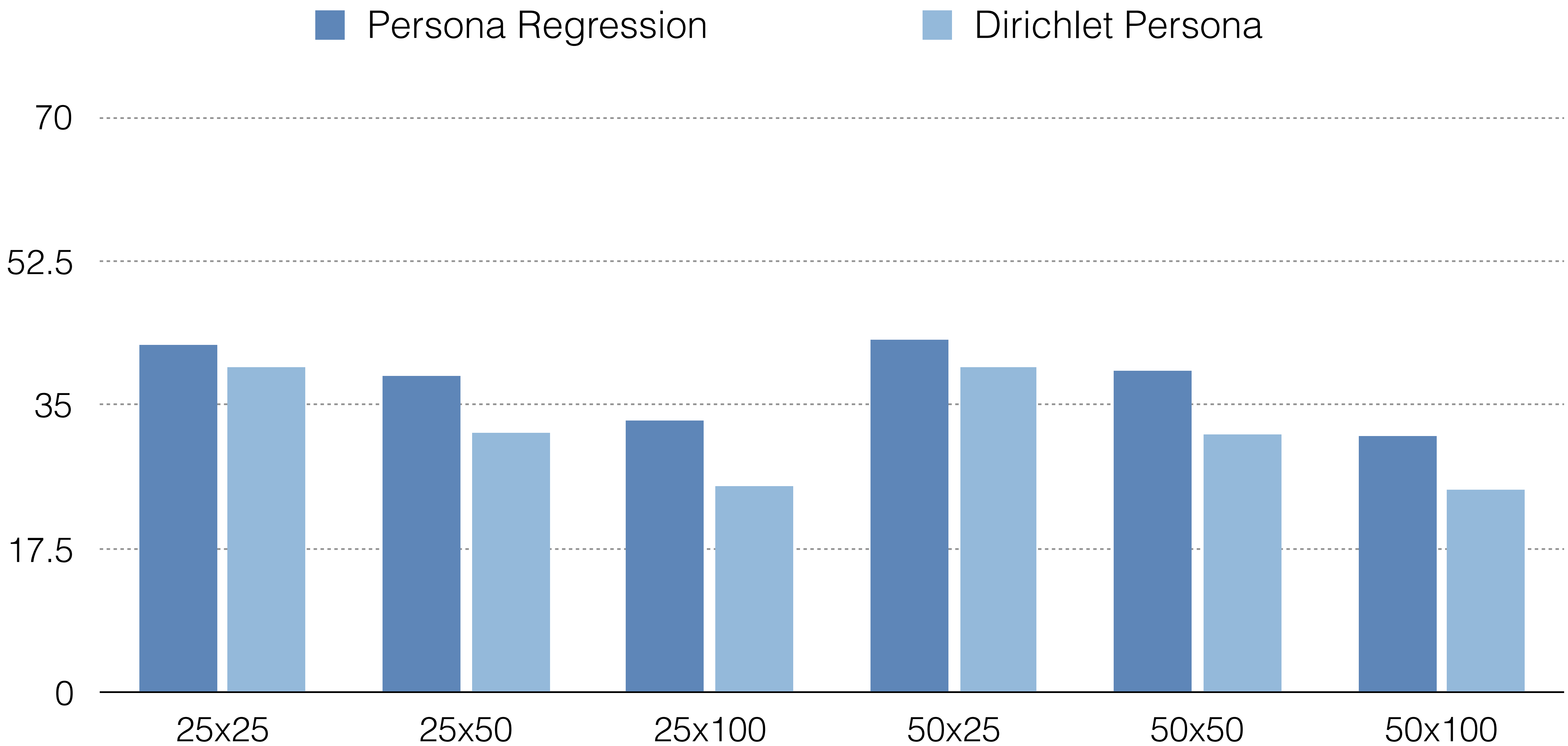
Evaluation II: TV Tropes

- Gold clusters: manually clustered characters from www.tvtropes.com
    - “The Surfer Dude”
    - “Arrogant Kung-Fu Guy”
    - “Hardboiled Detective”
    - “The Klutz”
    - “The Valley GIrl”
- 72 character tropes containing 501 characters

Purity: Names



Purity: TV Tropes



Evaluation

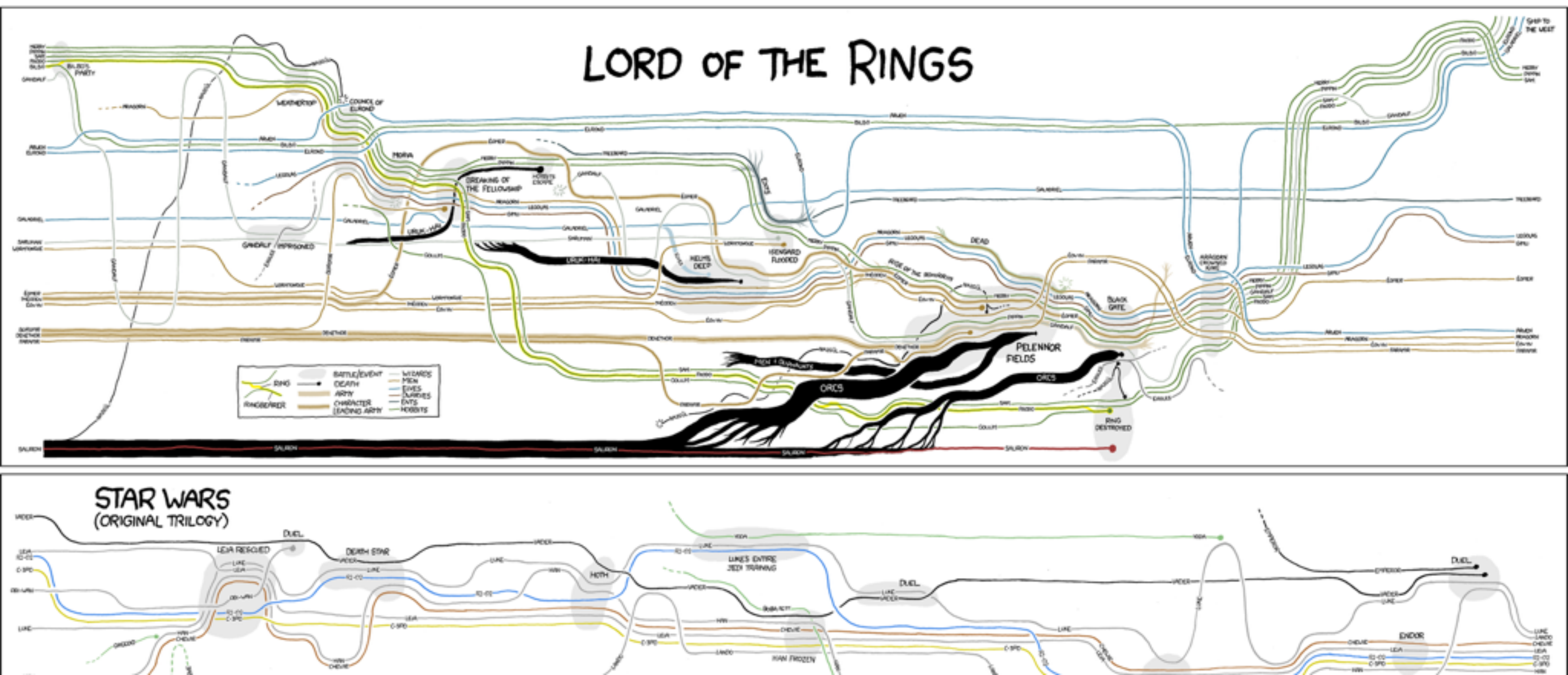
task	x
learn patterns that define architectural styles	set of skyscrapers
learn patterns that define genre	set of books
learn patterns that suggest “types” of customer behavior	customer data

Digital Humanities

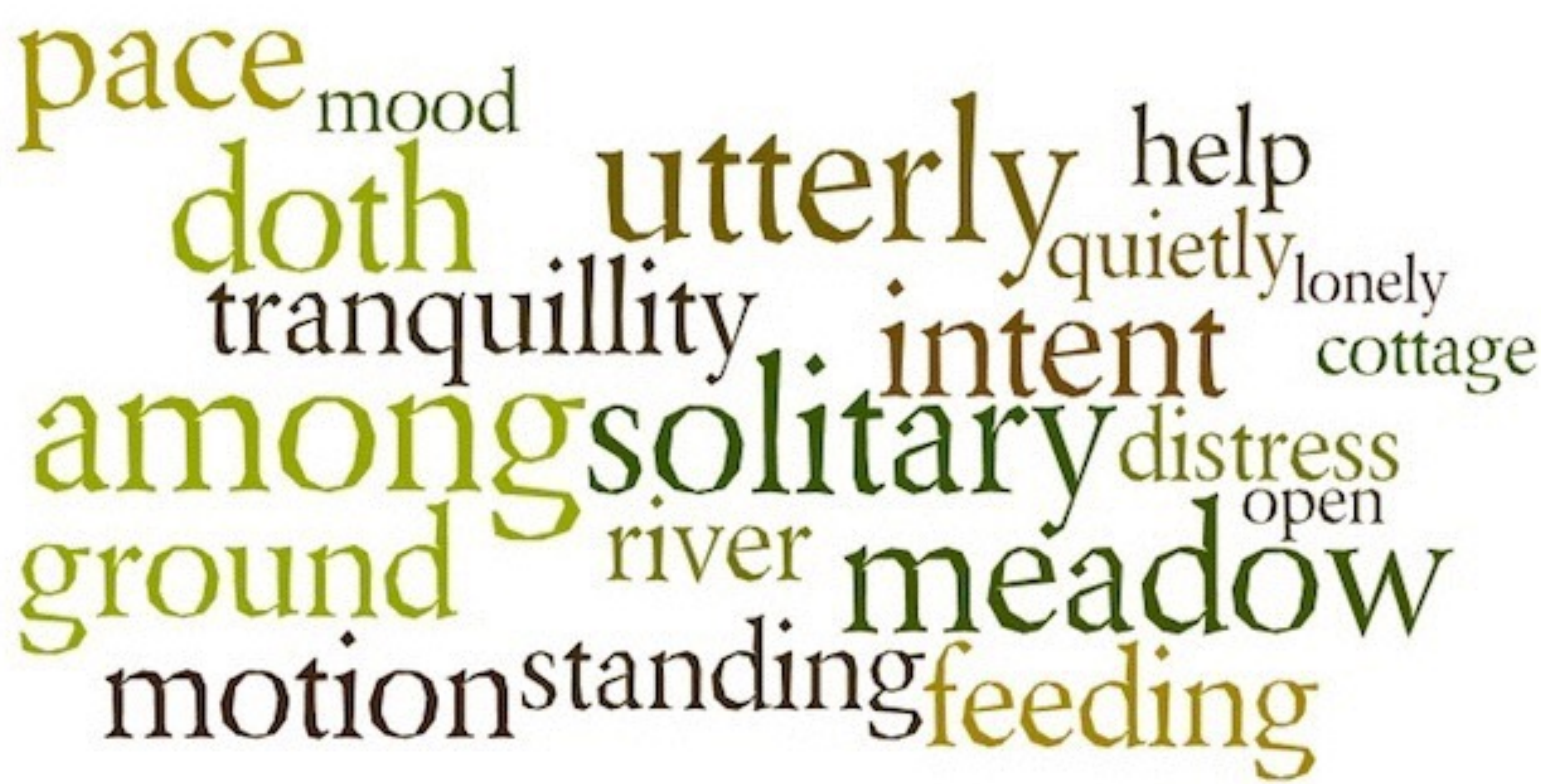
- Marche (2012), Literature Is not Data: Against Digital Humanities
- Underwood (2015), Seven ways humanists are using computers to understand text.

Text visualization

THESE CHARTS SHOW MOVIE CHARACTER INTERACTIONS.
THE HORIZONTAL AXIS IS TIME. THE VERTICAL GROUPING OF THE LINES INDICATES WHICH CHARACTERS ARE TOGETHER AT A GIVEN TIME.



Characteristic vocabulary



Characteristic words by William Wordsworth (in comparison to other contemporary poets) [Underwood 2015]

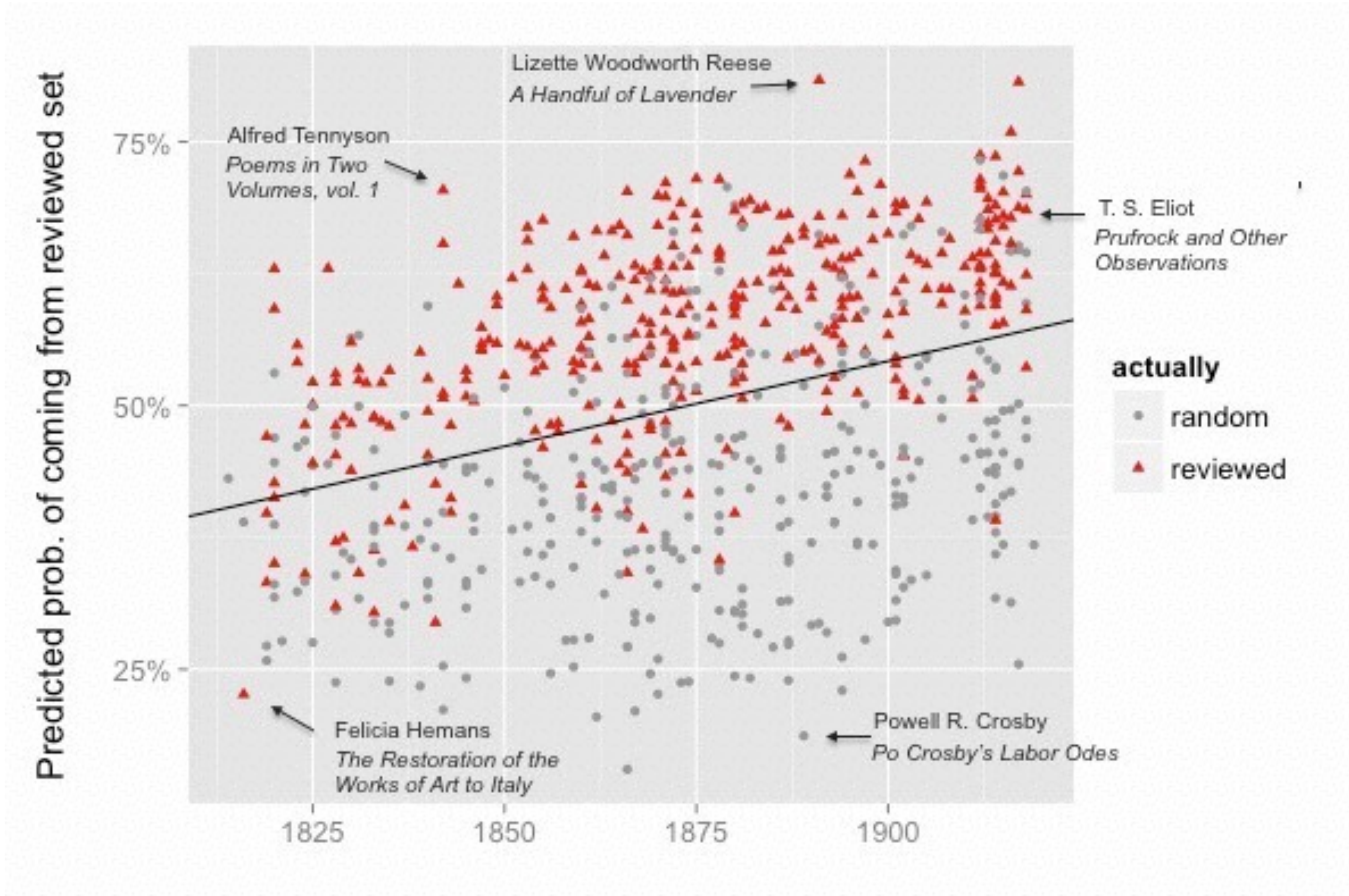
Finding and organizing texts

- e.g., finding all examples of a complex literary form (Haiku).
- Supplement traditional searches: book catalogues, search engines.

Modeling literary forms

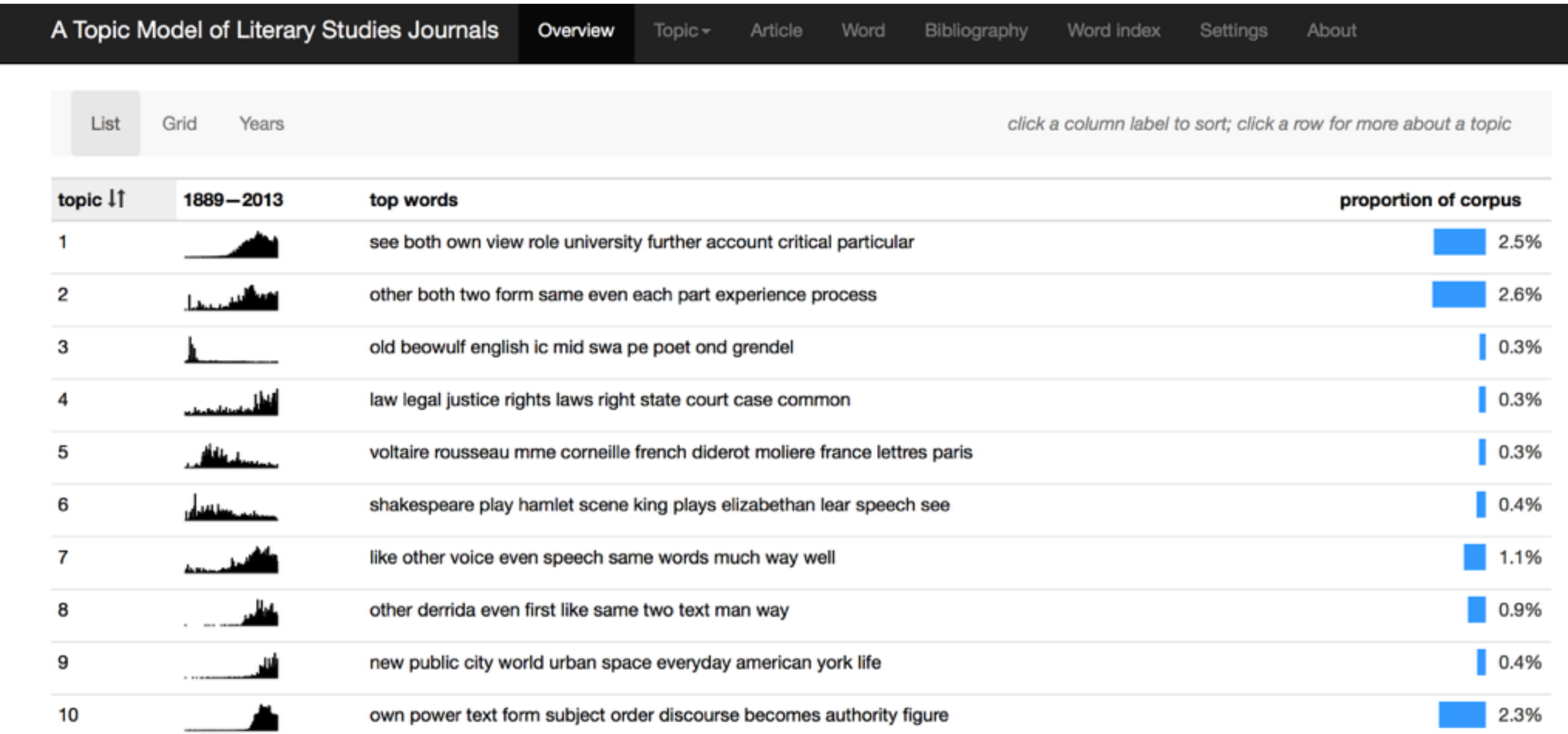
- What features of a text are predictive of Haiku?

Modeling social boundaries



Predicting reviewed texts [Underwood and Sellers (2015)]

Unsupervised modeling



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%
4		law legal justice rights laws right state court case common	0.3%
5		voltaire rousseau mme corneille french diderot moliere france lettres paris	0.3%
6		shakespeare play hamlet scene king plays elizabethan lear speech see	0.4%
7		like other voice even speech same words much way well	1.1%
8		other derrida even first like same two text man way	0.9%
9		new public city world urban space everyday american york life	0.4%
10		own power text form subject order discourse becomes authority figure	2.3%

Homework 1

Representation

- Part one (everyone): Design an *ideal* representation of Oscar nominees to enable good prediction/analysis.

Representation

- Part IIa. Implementation option. Instantiate a subset of those features for all nominees from 1960-2015. Deliverable: 6 feature files we will use to make predictions from.

feature name	feature value	nominee canonical id
boxoffice	60700000	/wiki/127_Hours
boxoffice	1000000	/wiki/12_Angry_Men_(1957_film)
boxoffice	168800000	/wiki/12_Monkeys
boxoffice	187700000	/wiki/12_Years_a_Slave_(film)
boxoffice	190000000	/wiki/2001:_A_Space_Odyssey_(film)
boxoffice	60400000	/wiki/21_Grams
boxoffice	2250000	/wiki/42nd_Street_(film)
boxoffice	9300000	/wiki/45_Years
boxoffice	5000000	/wiki/49th_Parallel_(film)

Representation

- Part IIb. Critical option. The prediction process here is conditioned on being the nominee. Lots of public critique of the Academy this year for nominating no minority actors.

- First, how would you model the Academy's (human) nomination process? How might this result in the underrepresentation of minorities?

- Second, consider an algorithmic approach to nominee prediction. What are the ways in which a similar underrepresentation can occur? What are the risks of training a supervised model?

- How does *representation* of data influence these processes?

- Deliverable: 3 page essay (single-spaced)