Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 5: Clustering overview

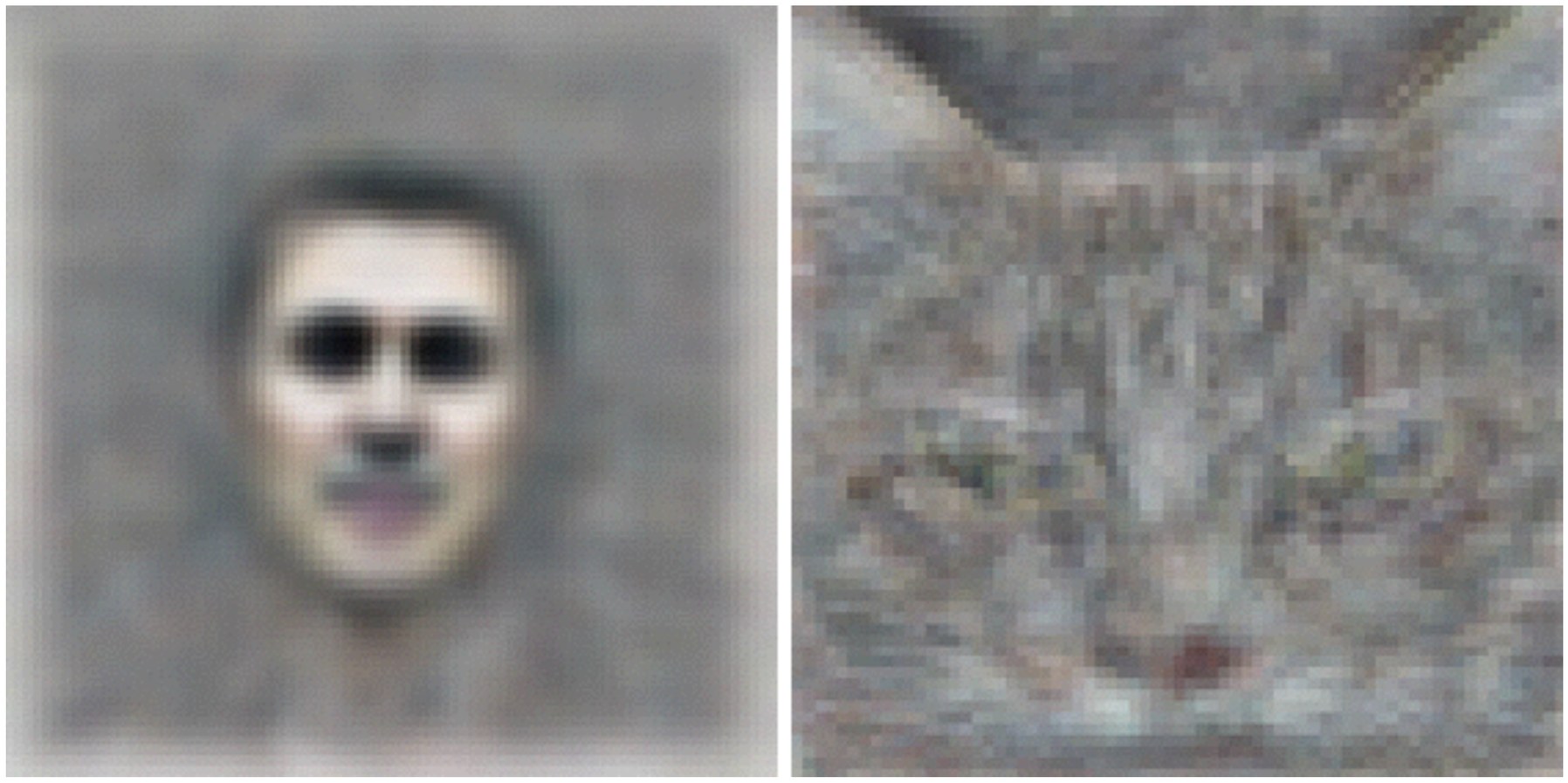
Jan 31, 2016

Clustering

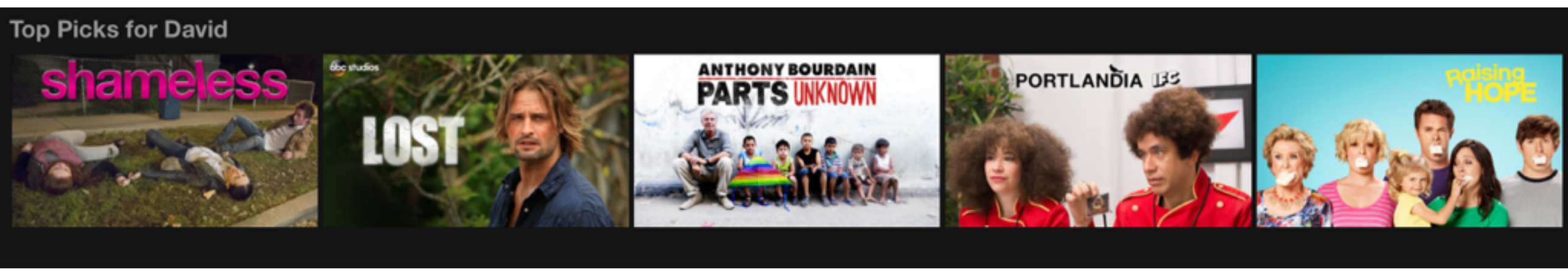
- Clustering (and unsupervised learning more generally) finds *structure* in data, using just X

X = a set of skyscrapers

Unsupervised Learning



Le et al. (2012), “Building High-level Features Using Large Scale Unsupervised Learning” (ICML)



Netflix

Your Amazon.com › Recommended for You
(If you're not David Bamman, click here.)

Recommendations

- Amazon Video
- Appliances
- Appstore for Android
- Arts, Crafts & Sewing
- Automotive
- Baby
- Books
- Books on Kindle
- Camera & Photo
- CDs & Vinyl

These recommendations are based on items you own and more.

view: **All** | New Releases | Coming Soon

1.

Eagle America 415-9307 Dovetail Marker
by Eagle America (October 22, 2013)
Average Customer Review: (133)
In Stock

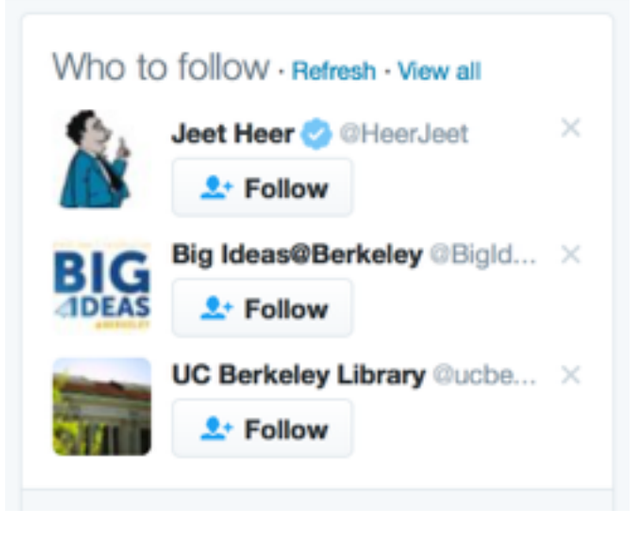
Price: $24.86
3 used & new from $18.00

I own it Not interested Rate this item

Recommended because you purchased **Stanley 15-106A Coping Saw** and more (Fix this)

Amazon

Twitter



MOST EMAILED MOST VIEWED **RECOMMENDED FOR YOU**

1. The Two Americas of 2016

2. Donald Trump's Son-in-Law, Jared Kushner, Tests Legal Path to White House Job

3. Donald Trump Selects Senator Jeff Sessions for Attorney General

4. NICHOLAS KRISTOF
A 12-Step Program for Responding to President-Elect Trump

5. Hillary Clinton, in Emotional Speech, Implores Supporters to Keep Believing in America

New York Times

Unsupervised Learning

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

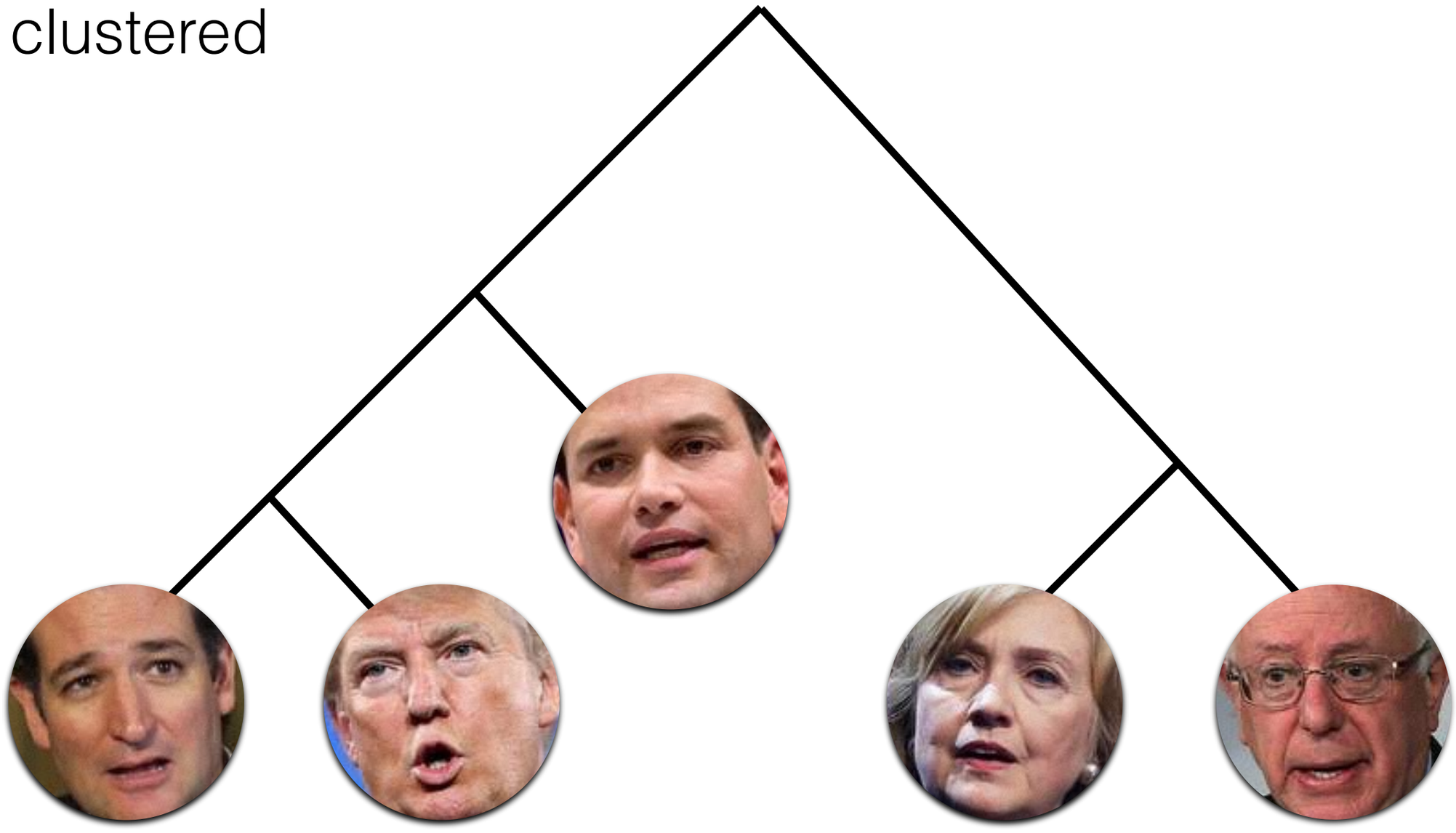
	Ann	Bob	Chris	David	Erik
Star Wars	5	5	4	5	3
Bridget Jones		4		4	1
Rocky	3		5		
Rambo		?		2	5

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

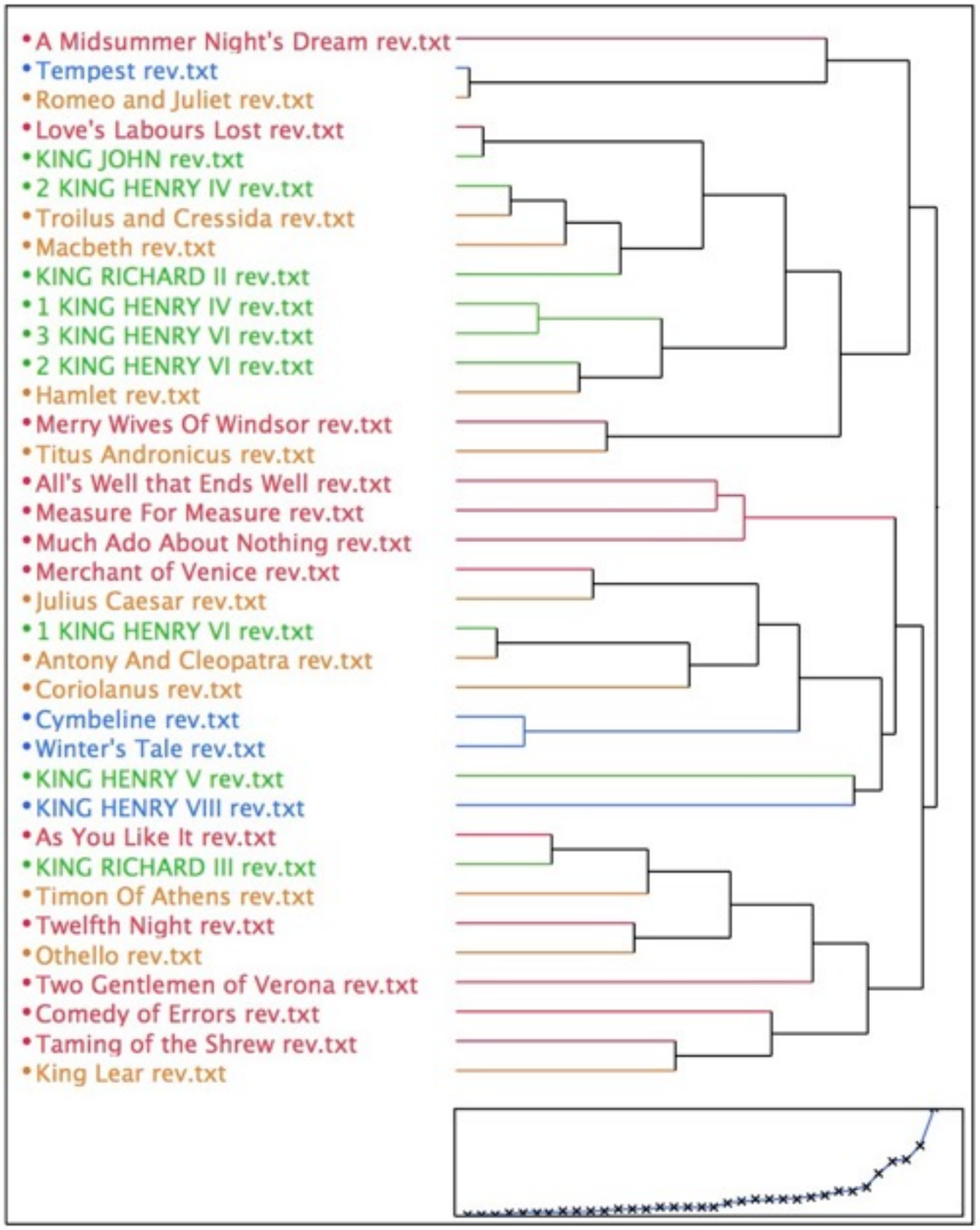
- *Hierarchical* order among the elements being clustered



Dendrogram

Shakespeare's plays

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



Bottom-up clustering

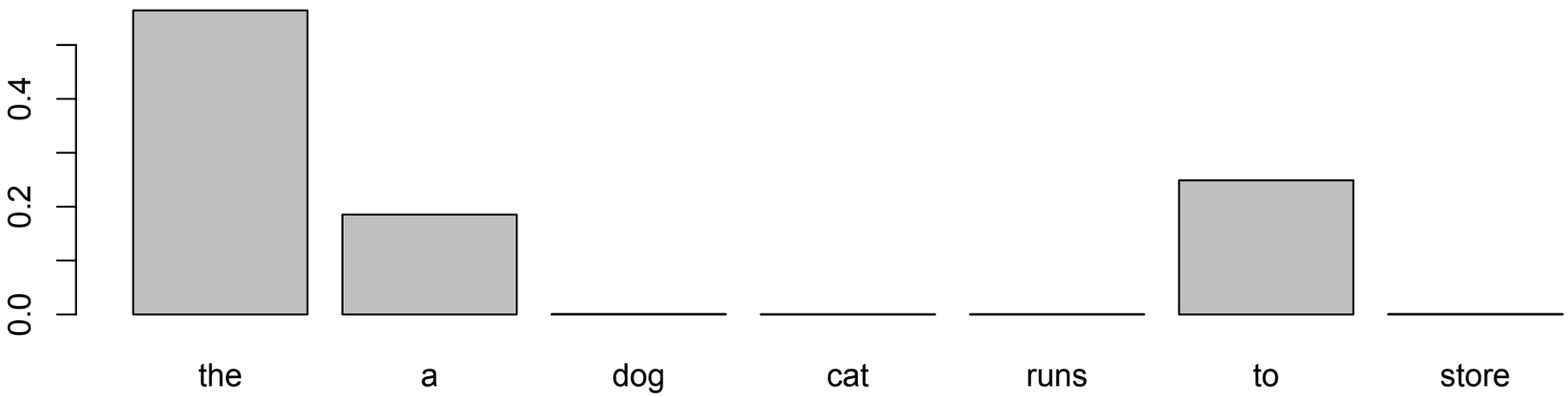
1 Given: a set $\mathcal{X} = \{x_1, \ldots x_n\}$ of objects
2 a function sim: $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R}$
3 **for** $i := 1$ **to** n **do**
4 $c_i := \{x_i\}$ **end**
5 $C := \{c_1, \ldots, c_n\}$
6 $j := n + 1$
7 **while** $C > 1$
8 $(c_{n_1}, c_{n_2}) := \arg\max_{(c_u, c_v) \in C \times C} \mathrm{sim}(c_u, c_v)$
9 $c_j = c_{n_1} \cup c_{n_2}$
10 $C := C \backslash \{c_{n_1}, c_{n_2}\} \cup \{c_j\}$
11 $j := j + 1$

Similarity

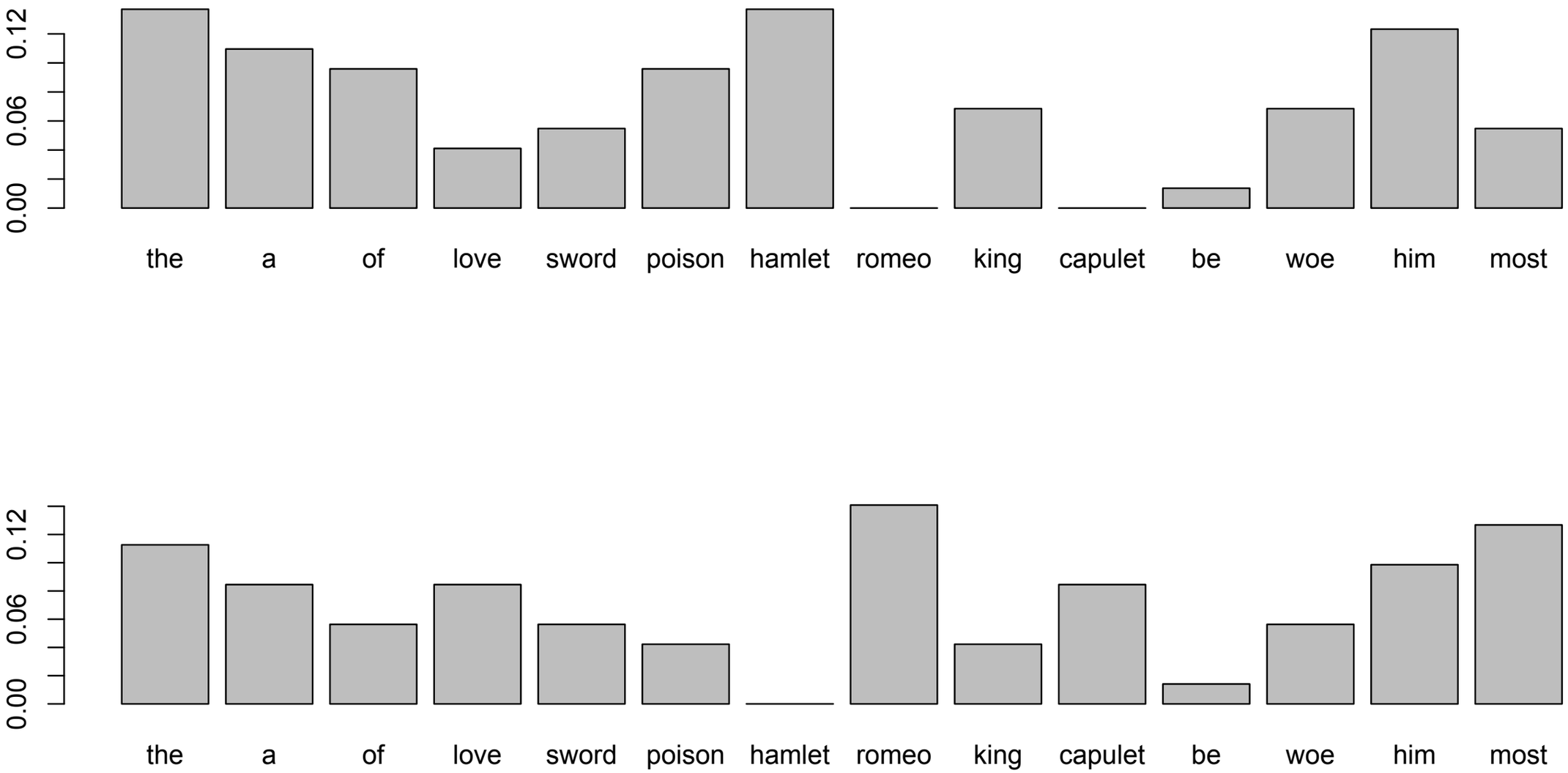
$$\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$$

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

Probability



Unigram probability



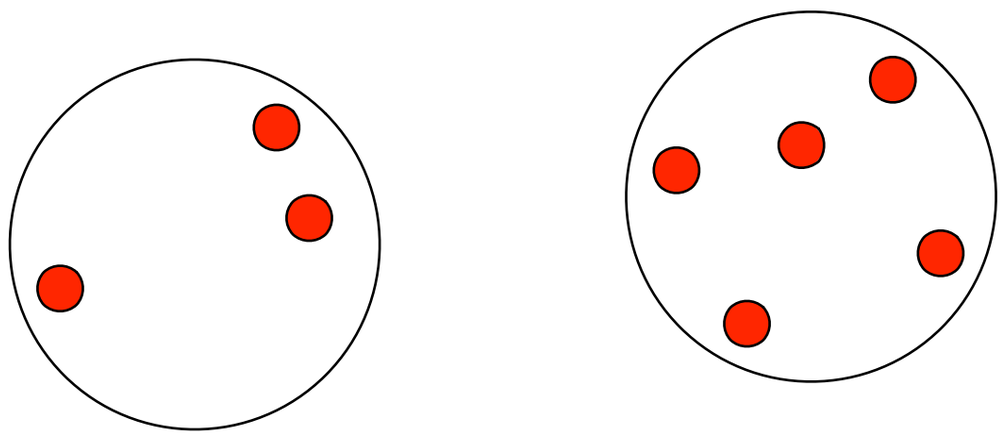
Similarity

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

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

B

A

C

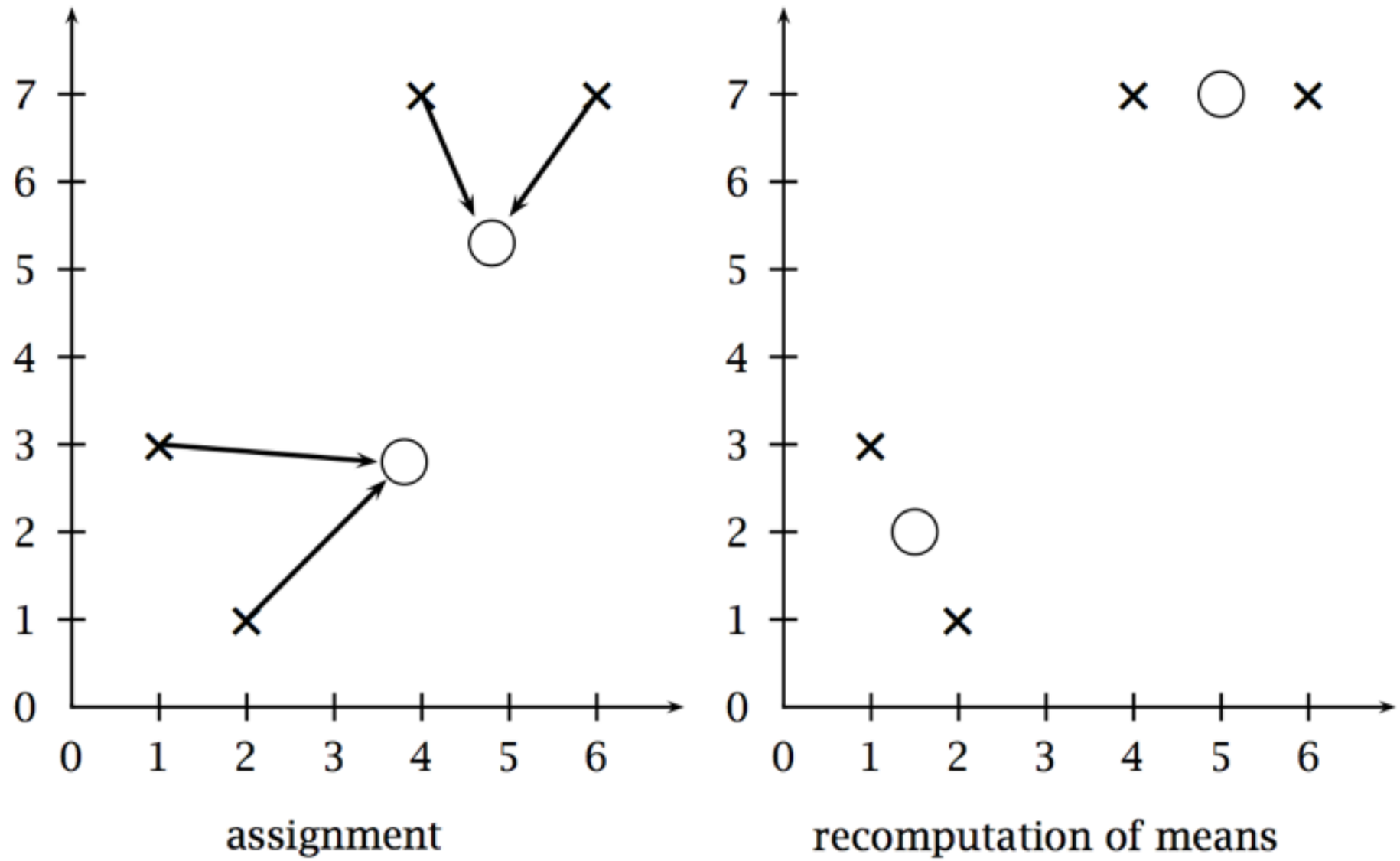
Flat Clustering

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

K-means

1 Given: a set $\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_n\} \subseteq \mathbb{R}^m$
2 a distance measure $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$
3 a function for computing the mean $\mu : \mathcal{P}(\mathbb{R}) \to \mathbb{R}^m$
4 Select k initial centers $\vec{f}_1, \ldots, \vec{f}_k$
5 **while** stopping criterion is not true **do**
6 **for** all clusters c_j **do**
7 $c_j = \{\vec{x}_i \mid \forall \vec{f}_l \, d(\vec{x}_i, \vec{f}_j) \leq d(\vec{x}_i, \vec{f}_l)\}$
8 **end**
9 **for** all means $\vec{f}_j$ **do**
10 $\vec{f}_j = \mu(c_j)$
11 **end**
12 **end**

K-means

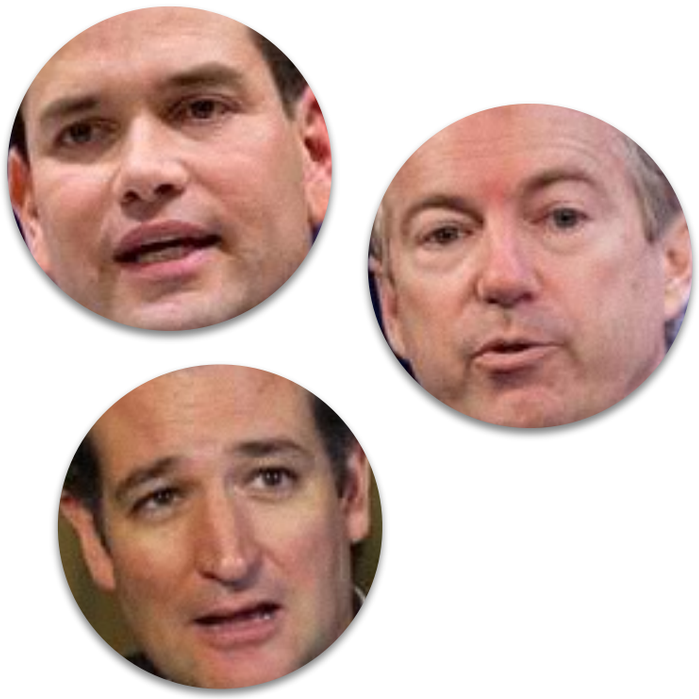

7
6
5
4
3
2
1
0
0 1 2 3 4 5 6
assignment
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6
recomputation of means

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



Yes on abortion access	1
Yes on expanding gun rights	0
Yes on tax breaks	0

Voting behavior

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

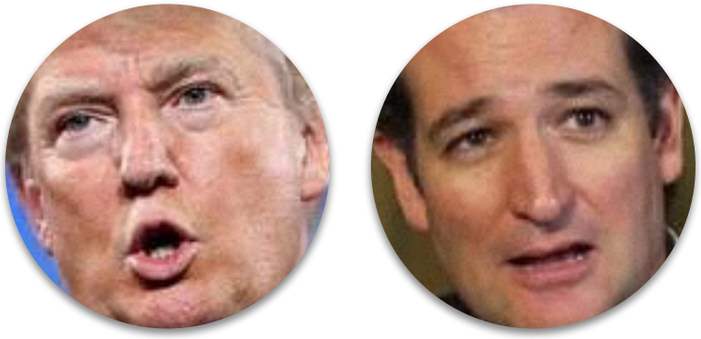
- Elements within clusters should be more similar to each other

- 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)

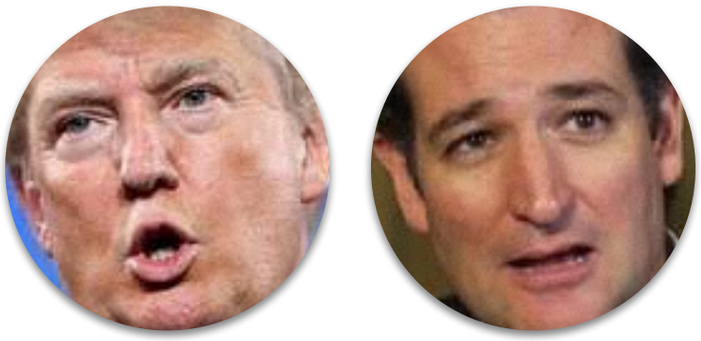
$$\mathcal{G} = \{g_1 \ldots g_k\}$$

- 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

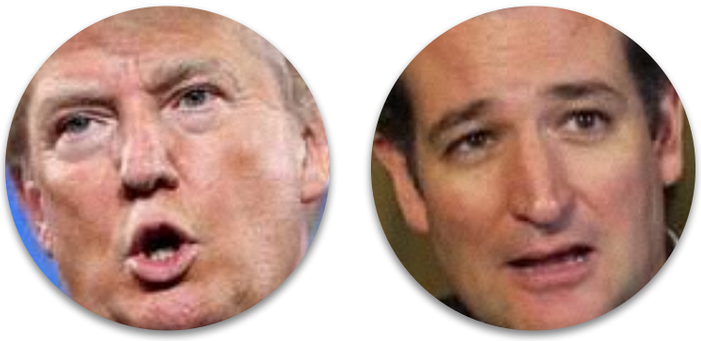
B

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)

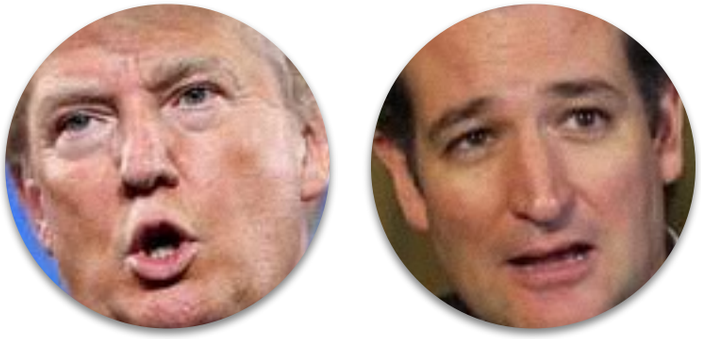
Learned (G)

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)

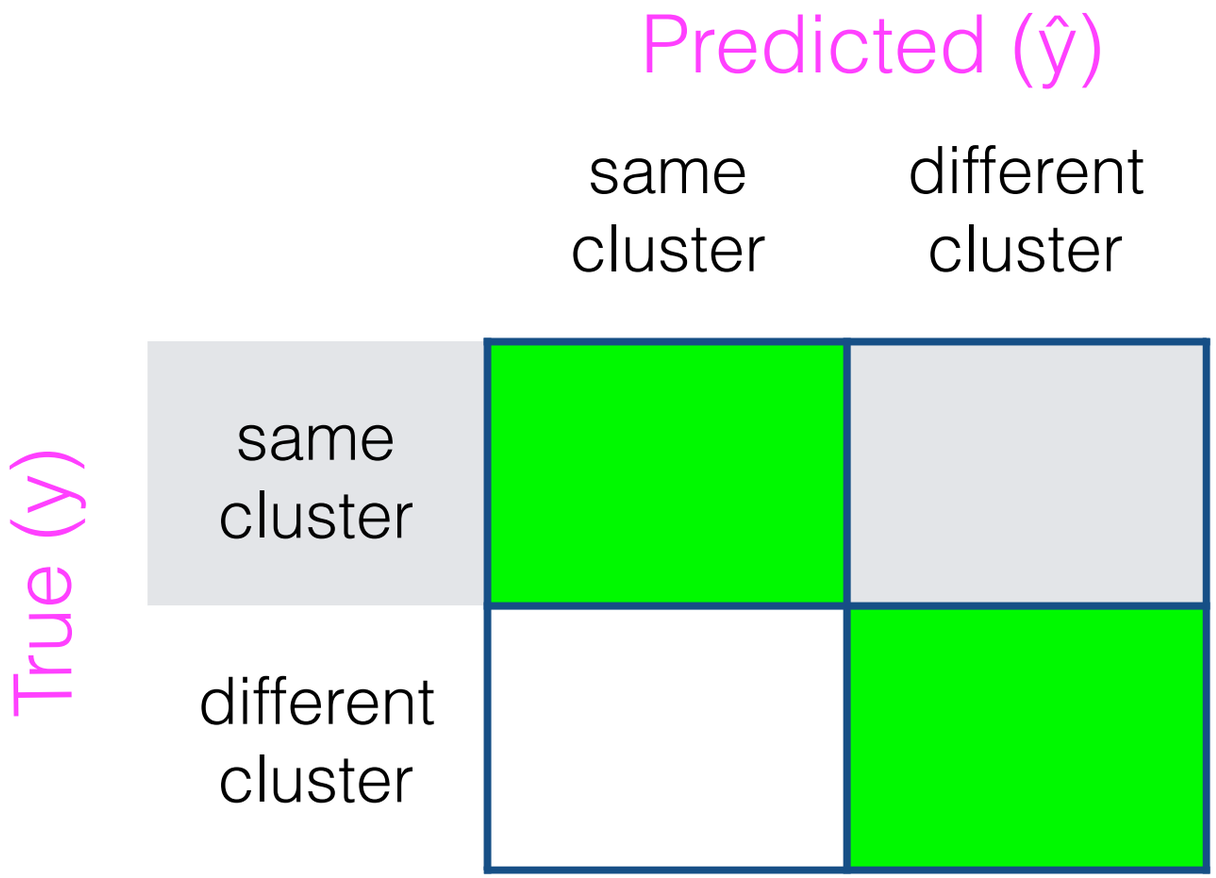
Evaluation: Rand Index

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



21 decisions

$N(N-1)/2$

Learned







Predicted (ŷ)

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 (ŷ)	
		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

Personas

attribute	dark major henchman warrior sergeant
agent	shoot aim overpower interrogate kill

Highest weighted features:

- Male
- Action
- War film

Jason Bourne, Bourne Supremacy

Personas

patient	capture corner transport imprison trap
agent	infiltrate deduce leap evade obtain
agent	flee escape swim hide manage

Highest weighted features:

- Female
- Action
- Adventure

Ginormica (Monsters vs. Aliens)

Evaluation I: Names

- Gold clusters: characters with the same name (sequels, remakes)

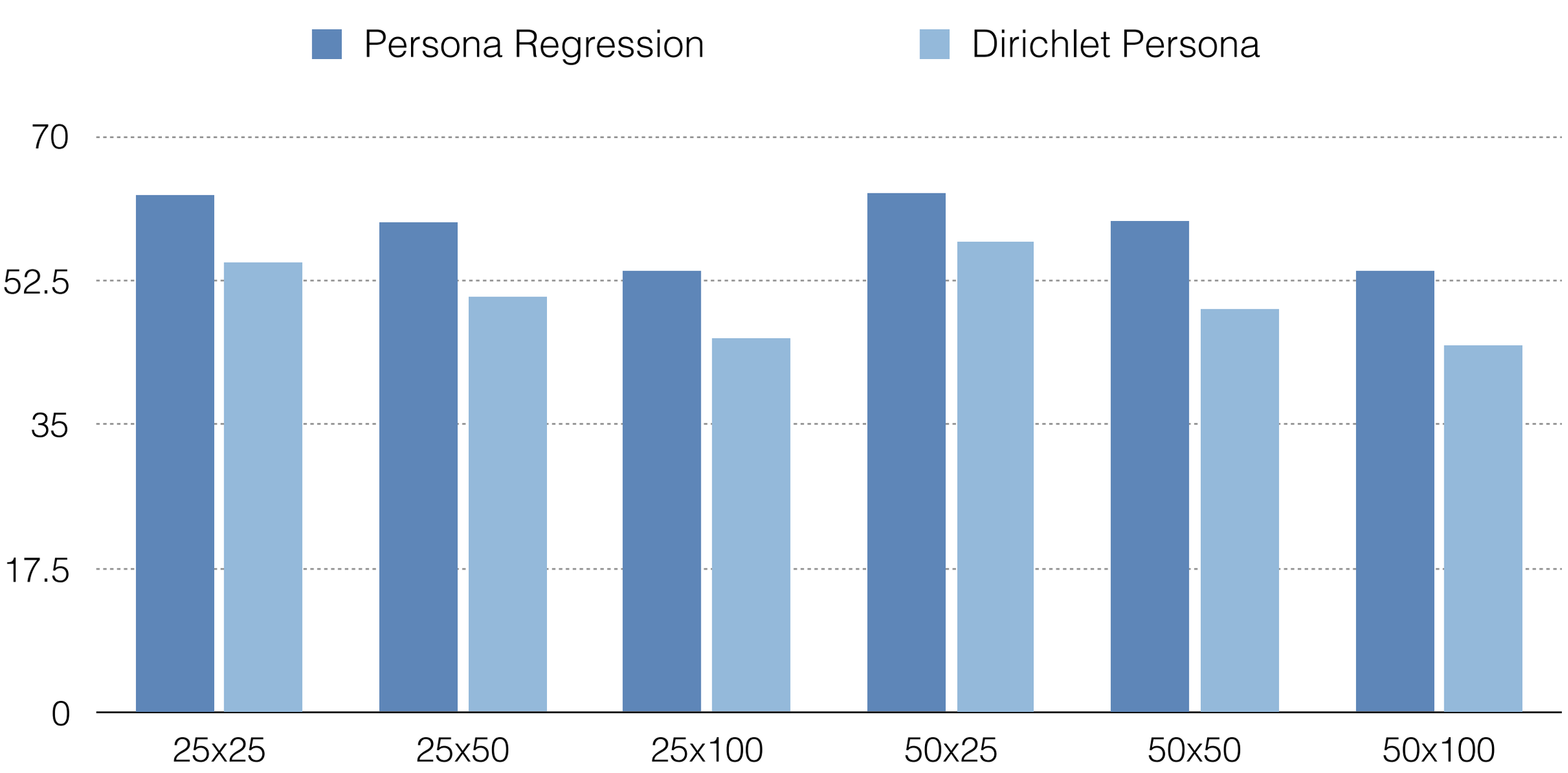
- Noise: “street thug”

- 970 unique character names used twice in the data; n=2,666

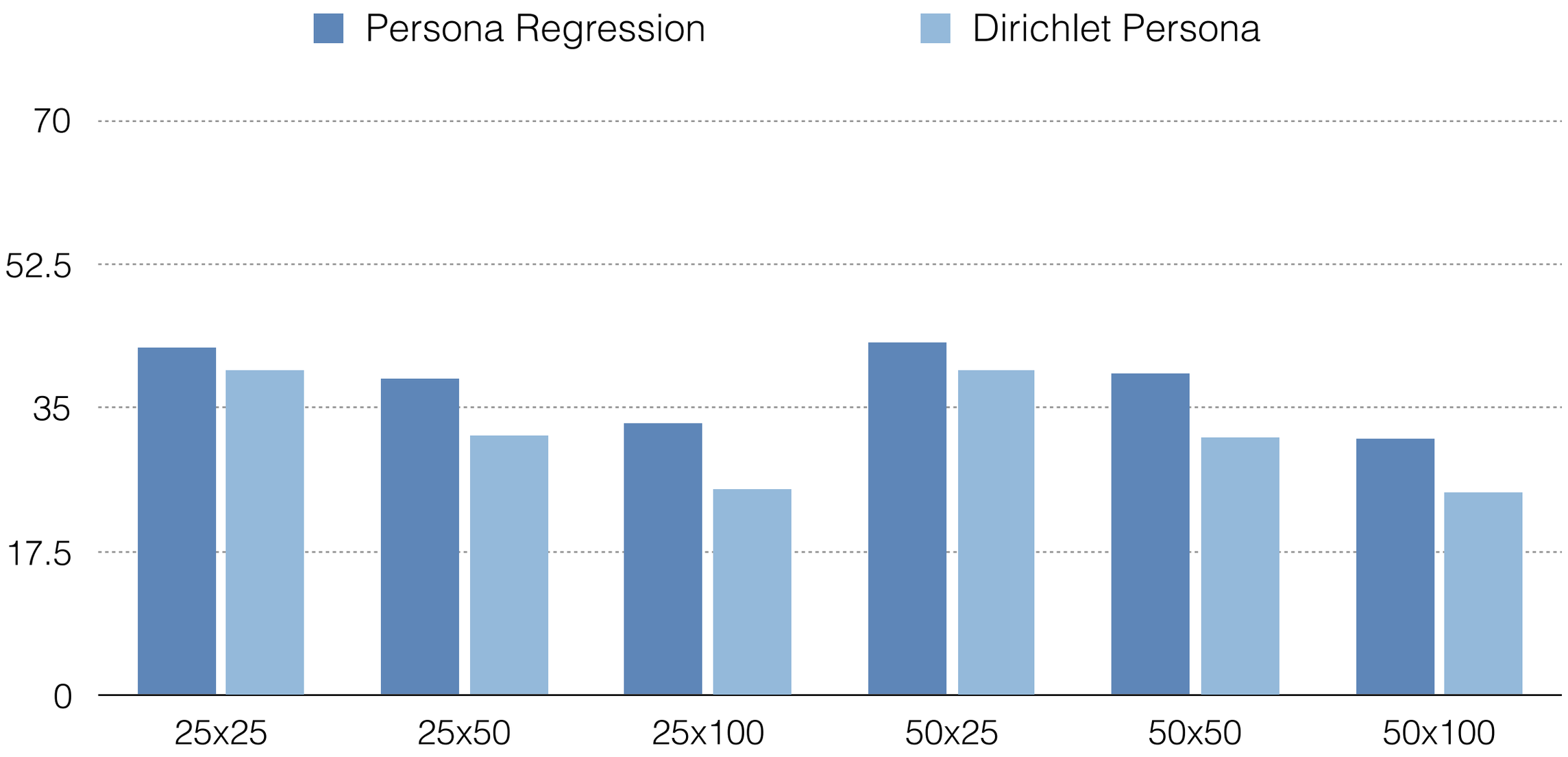
Evaluation II: TV Tropes

- 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

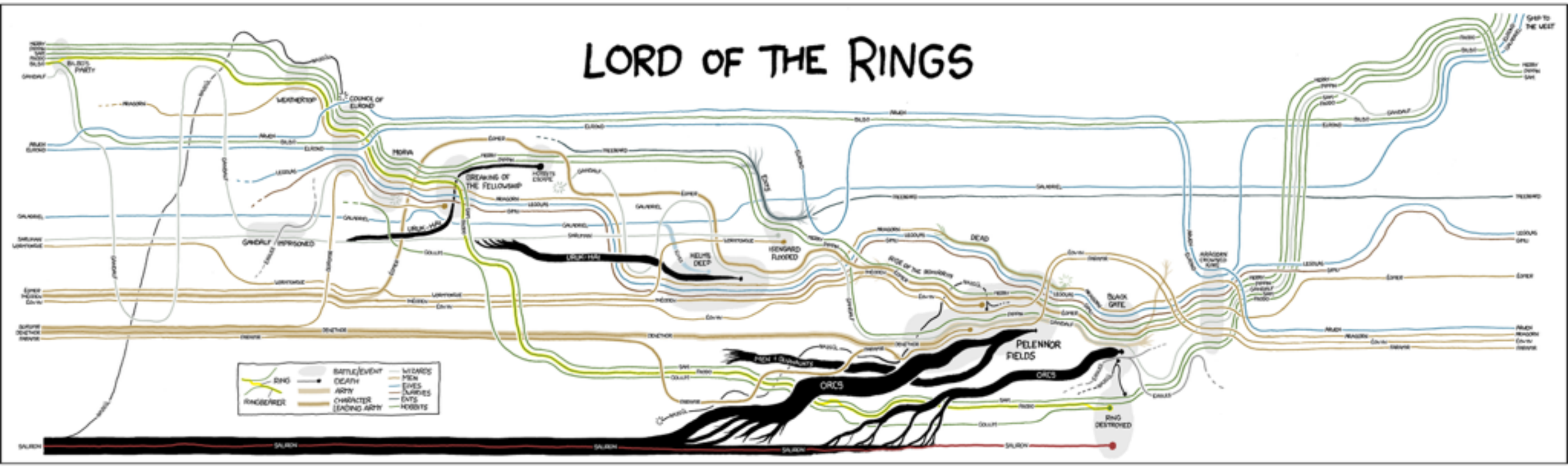
task	x
learn patterns that define architectural styles	set of skyscrapers
learn patterns that define genre	set of books
learn patterns that suggest “types” of customer behavior	customer data

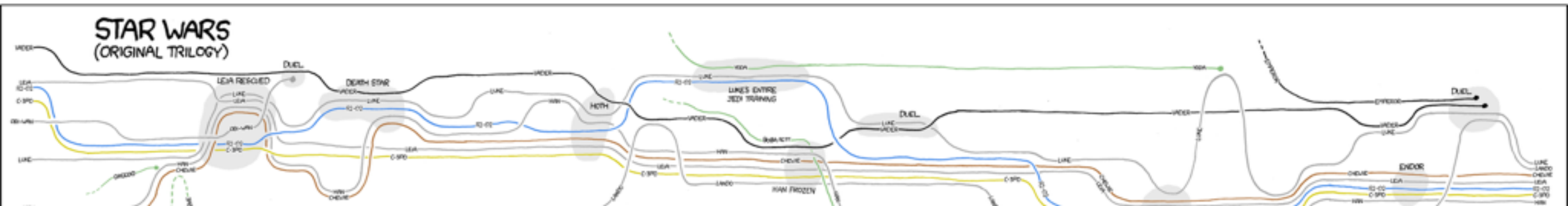
Digital Humanities

- Marche (2012), Literature Is not Data: Against Digital Humanities

- Underwood (2015), Seven ways humanists are using computers to understand text.

Text visualization





Characteristic vocabulary

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

Finding and organizing texts

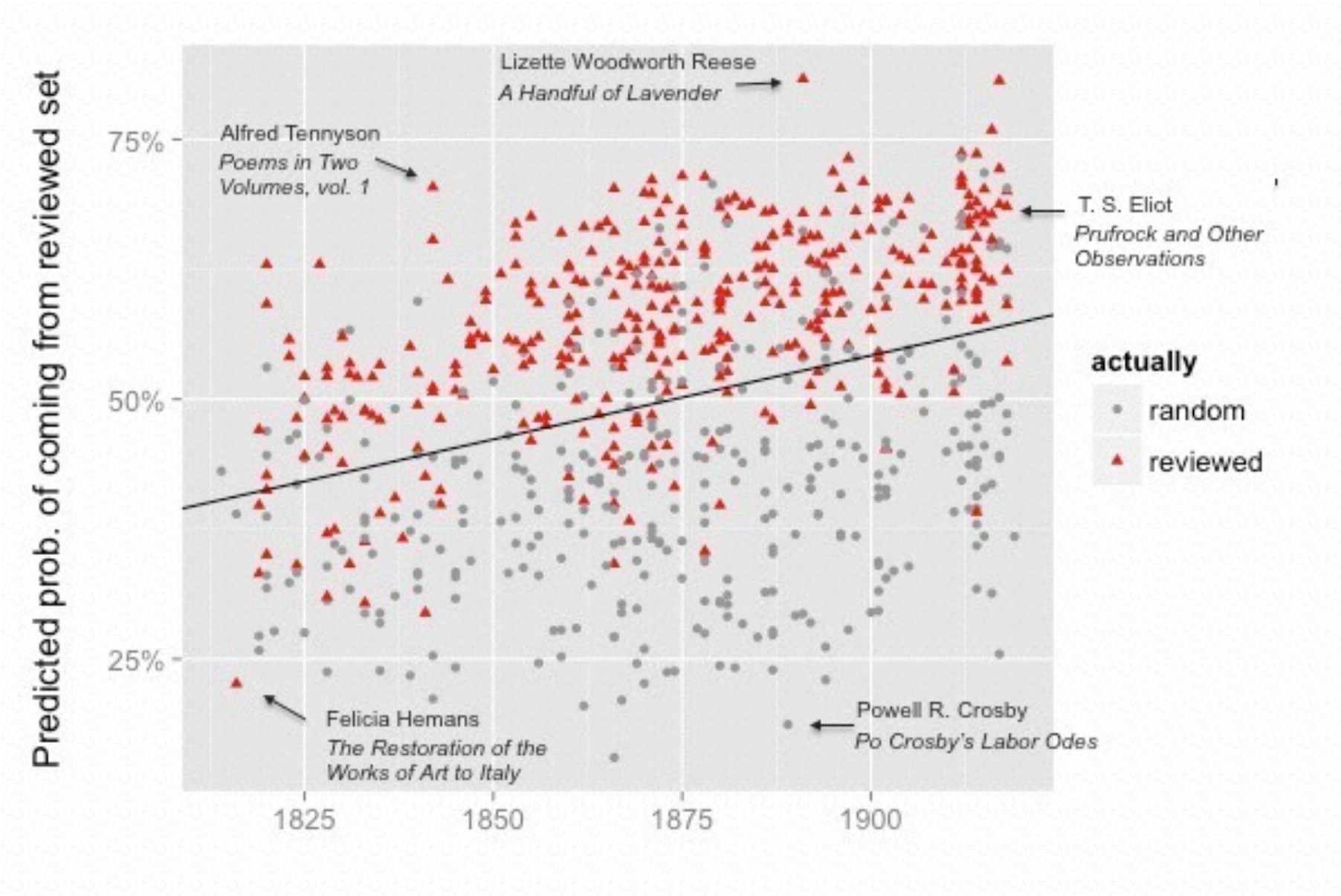
- e.g., finding all examples of a complex literary form (Haiku).

- Supplement traditional searches: book catalogues, search engines.

Modeling literary forms

- What features of a text are predictive of Haiku?

Modeling social boundaries



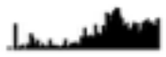
Predicting reviewed texts [Underwood and Sellers (2015)]

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%