CS224C: NLP for CSS

Computational Basics

Diyi Yang

Stanford CS

Regression

Regression

A mapping from input data x (drawn from instance space X) to a point y in R

R: the set of real numbers

x = the empire state building

y = 17444.5625″

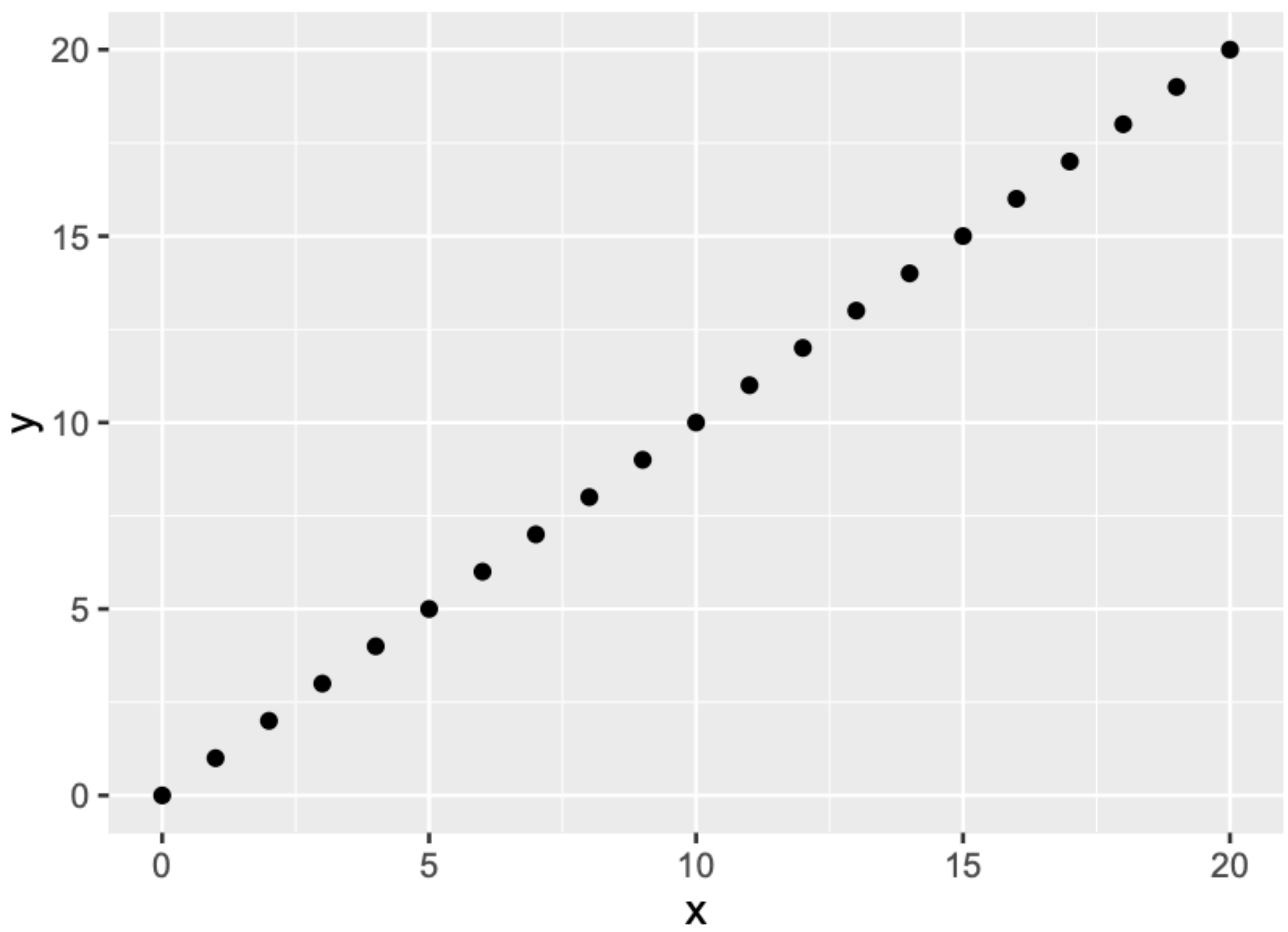
Slide content credit to David Bamman

Linear Regression

Suppose we have n data points. For each data point i, we observe

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

Linear regression states that $\hat{y}_i = \sum_{i=1}^{F} x_i \beta_i$



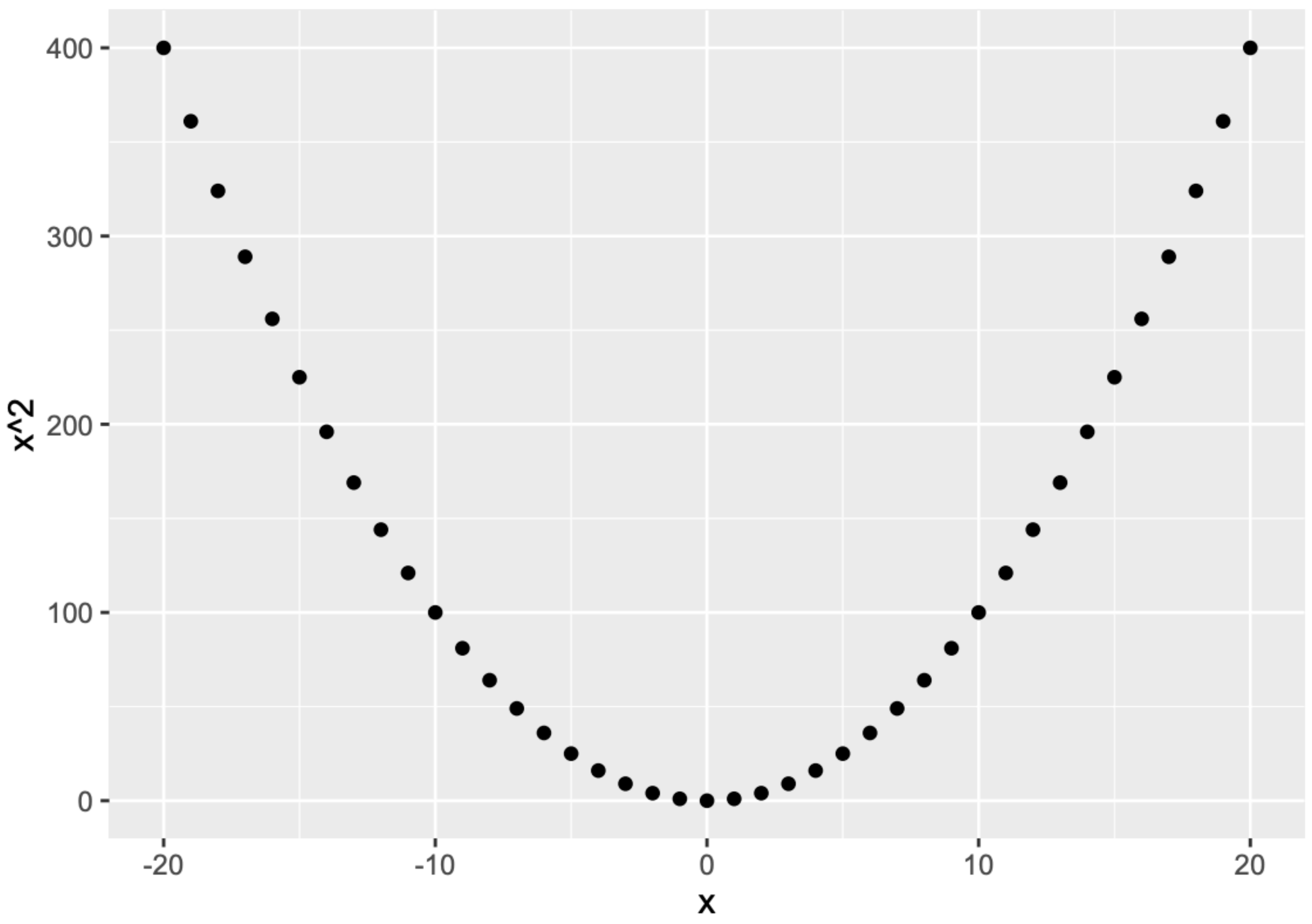
Slide content credit to David Bamman

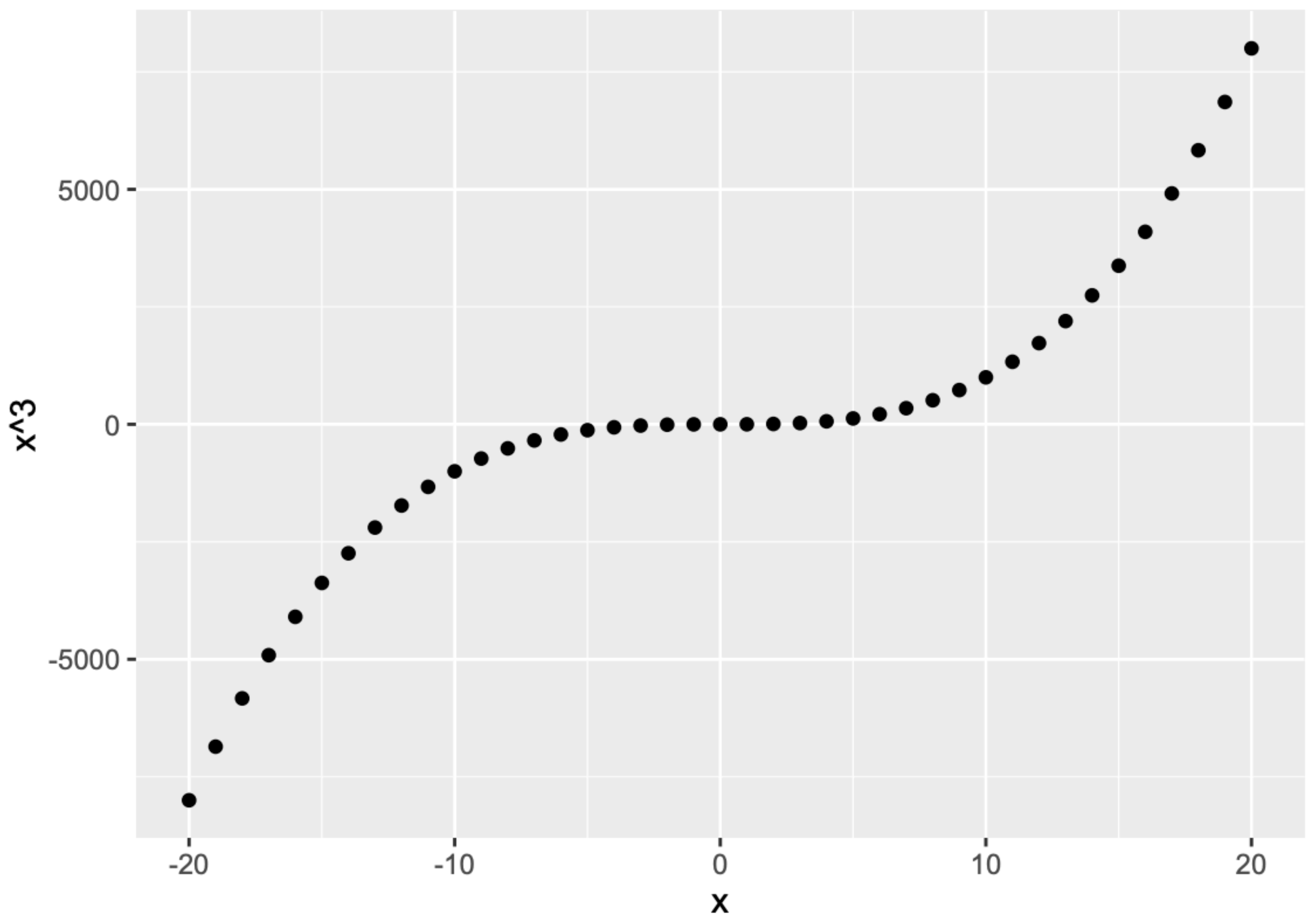
Regression for Social Sciences

Polynomial Regression

$$\hat{y}_i = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i}$$

$$\hat{y}_i = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i} + \sum_{i=1}^{F} x_i^3 \beta_{c,i}$$





Slide content credit to David Bamman

Nonlinear Regression

Support vector machines (regression)

Neural Networks

…

Slide content credit to David Bamman

Number of Parameters

$$\hat{y}_i = \sum_{i=1}^{F} x_i \beta_{a,i}$$

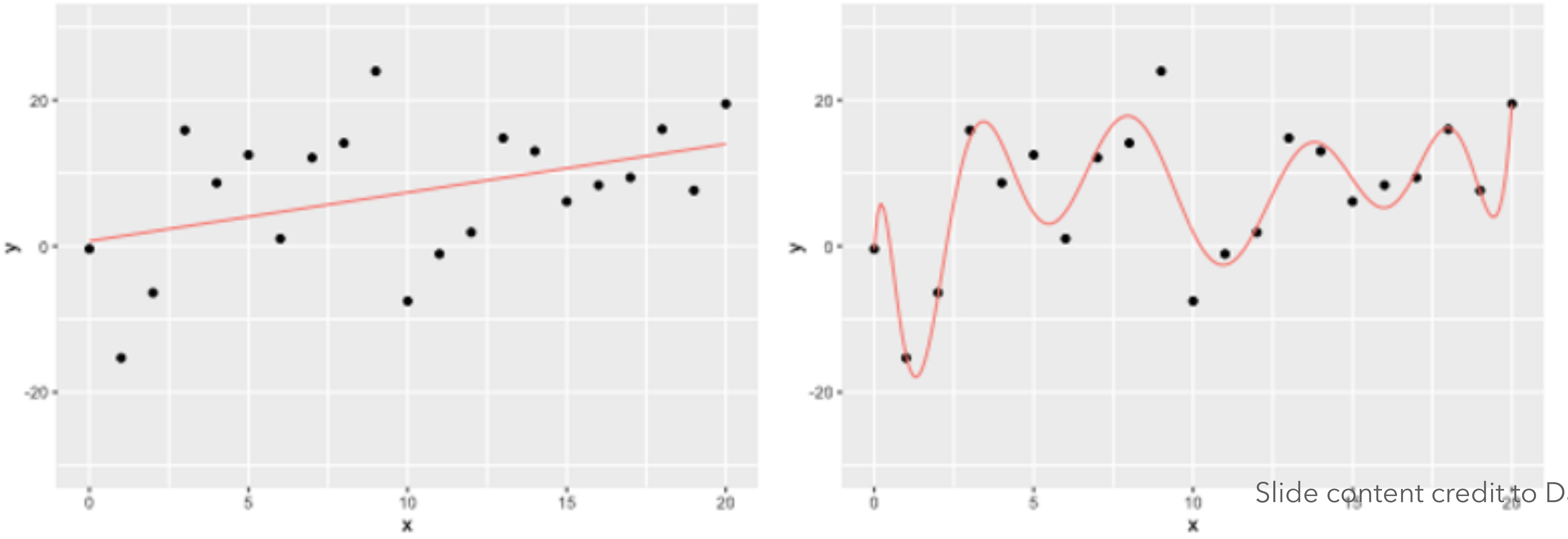
$$\hat{y}_i = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i}$$

$$\hat{y}_i = \sum_{i=1}^{F} x_i \beta_{a,i} + \sum_{i=1}^{F} x_i^2 \beta_{b,i} + \sum_{i=1}^{F} x_i^3 \beta_{c,i}$$

Slide content credit to David Bamman

Overfitting

Memorizing the nuances (and noise) of the training data that prevents generalizing to unseen data



Slide content credit to David Bamman

Sources of Error

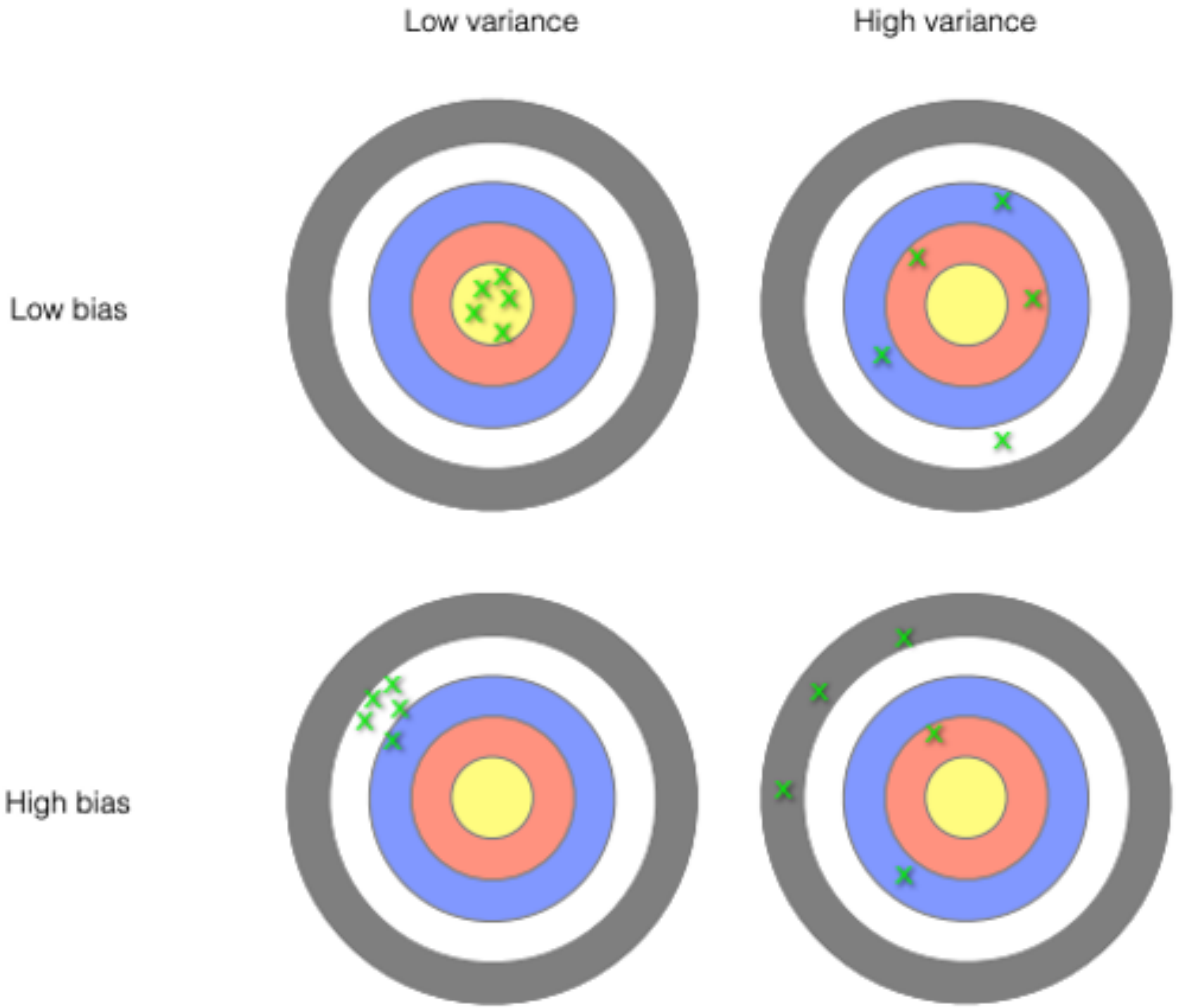
Bias: Error due to mis-specifying the relationship between input and output

Too few parameters, or the wrong kinds

Variance: Error due to sensitivity to random fluctuations in the training data. If you train on different data, do you get radically different predictions?

Too many parameters

Slide content credit to David Bamman



Slide content credit to David Bamman

Image from Flach 2012

Regression for Social Sciences

Regression analysis is a very useful tool for social sciences

- ✦ Understand the relationship between variables, adjusting for other potential confounders
- ✦ Predict the value of one variable based on others

In Other Terminology

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Intersect

Dependent Variable = Independent Variable + Independent Variable

How good is the Fit?

Mean squared error (MSE) $\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$

Mean absolute error (MAE) $\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$

How good is the "fit"?

Sum of the squares total (SST): total variability about the mean

$$\sum (Y - \bar{Y})^2$$

Sum of the squared error (SSE): variability about the regression line

$$\sum (Y - \hat{Y})^2$$

Sum of the squares due to regression (SSR): total variability that is explained by the model

$$\sum (\hat{Y} - \bar{Y})^2$$

Coefficient of Determination r^2

The proportion of the variability explained by regression model

$$r^2 = \frac{\text{SSR}}{\text{SST}}$$

Recommendations for Building Regression Models

A high r^2 is desired with a reasonable set of variables

When more variables get added to the model, r^2 usually increases.

Thus, adjusted r^2 is often used to account for the number of variables

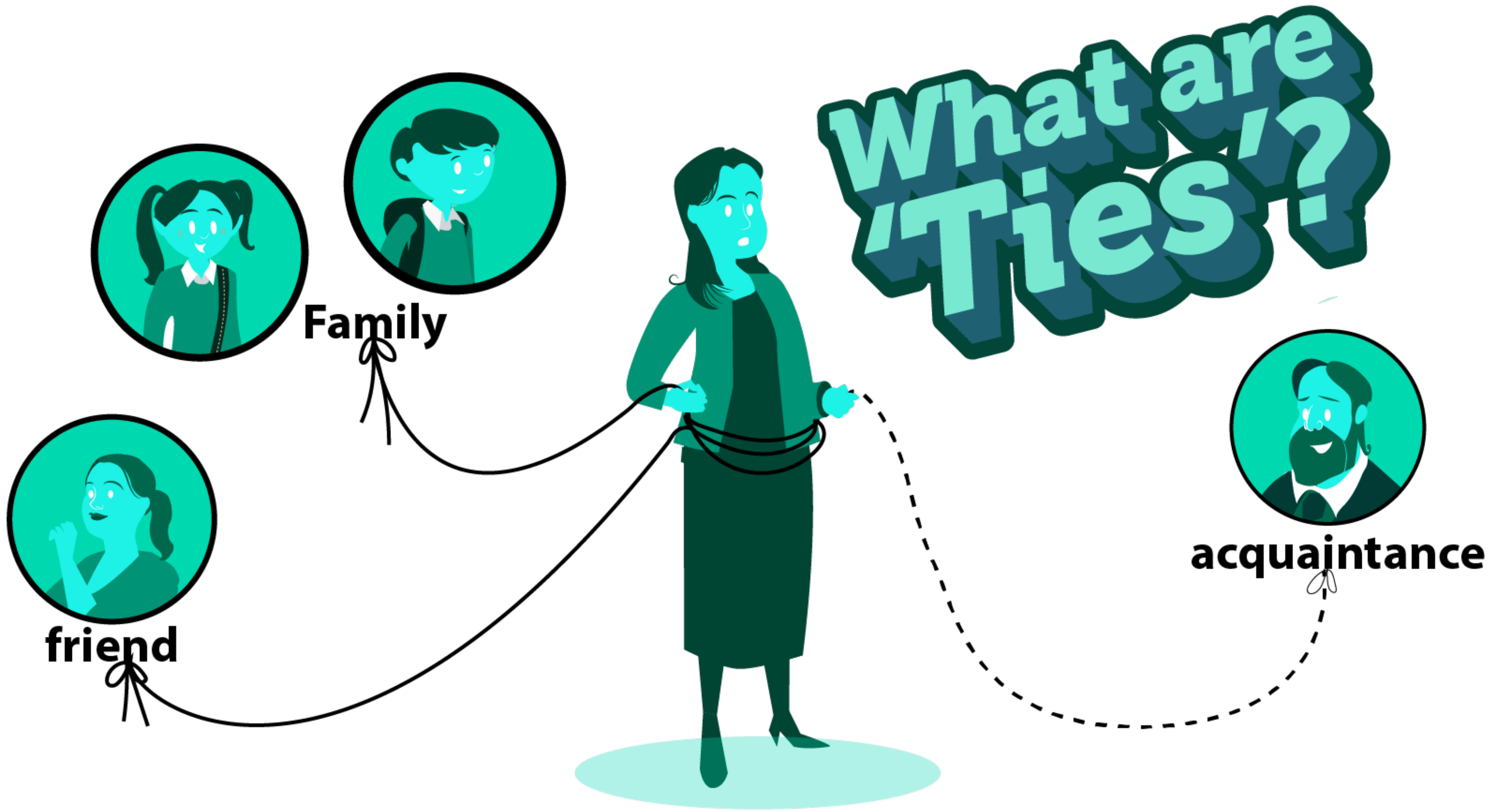
Independent variables might contain **duplicated** information

Colinear if two variables are correlated

Multicolinearity if more than two variables are correlated - this will make the interpretation of regression coefficient problematic

Let’s predict tie strength on Facebook

1. **Why** is this a regression task?
2. **What** is tie strength?
3. How can we get the **ground truth**?
4. How to get **data**?
5. How can we **evaluate** it?
6. Does the system really **work**?



https://murraydare.co.uk/marketing-theory/strong-weak-ties

Let's predict tie strength on Facebook

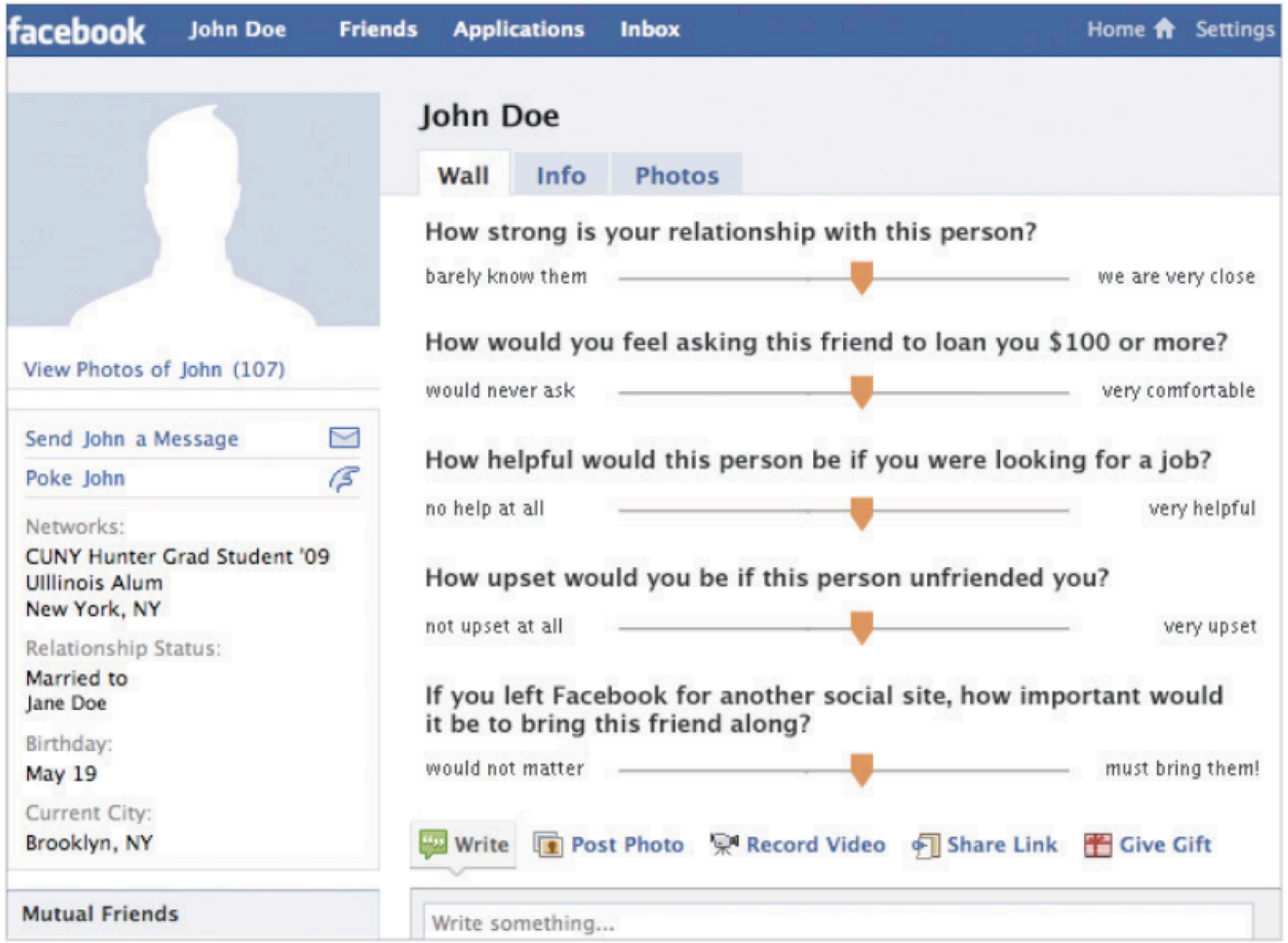
Mark Granovetter introduced the concept of **tie strength** in1973

"**The Strength of Weak Ties**"

The strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie

Gilbert, Eric, and Karrie Karahalios. "Predicting tie strength with social media." In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 211-220. 2009.

Let's predict tie strength on Facebook



How strong is your relationship with this person?

barely know them ——— we are very close

How would you feel asking this friend to loan you $100 or more?

would never ask ——— very comfortable

How helpful would this person be if you were looking for a job?

no help at all ——— very helpful

How upset would you be if this person unfriended you?

not upset at all ——— very upset

If you left Facebook for another social site, how important would it be to bring this friend along?

would not matter ——— must bring them

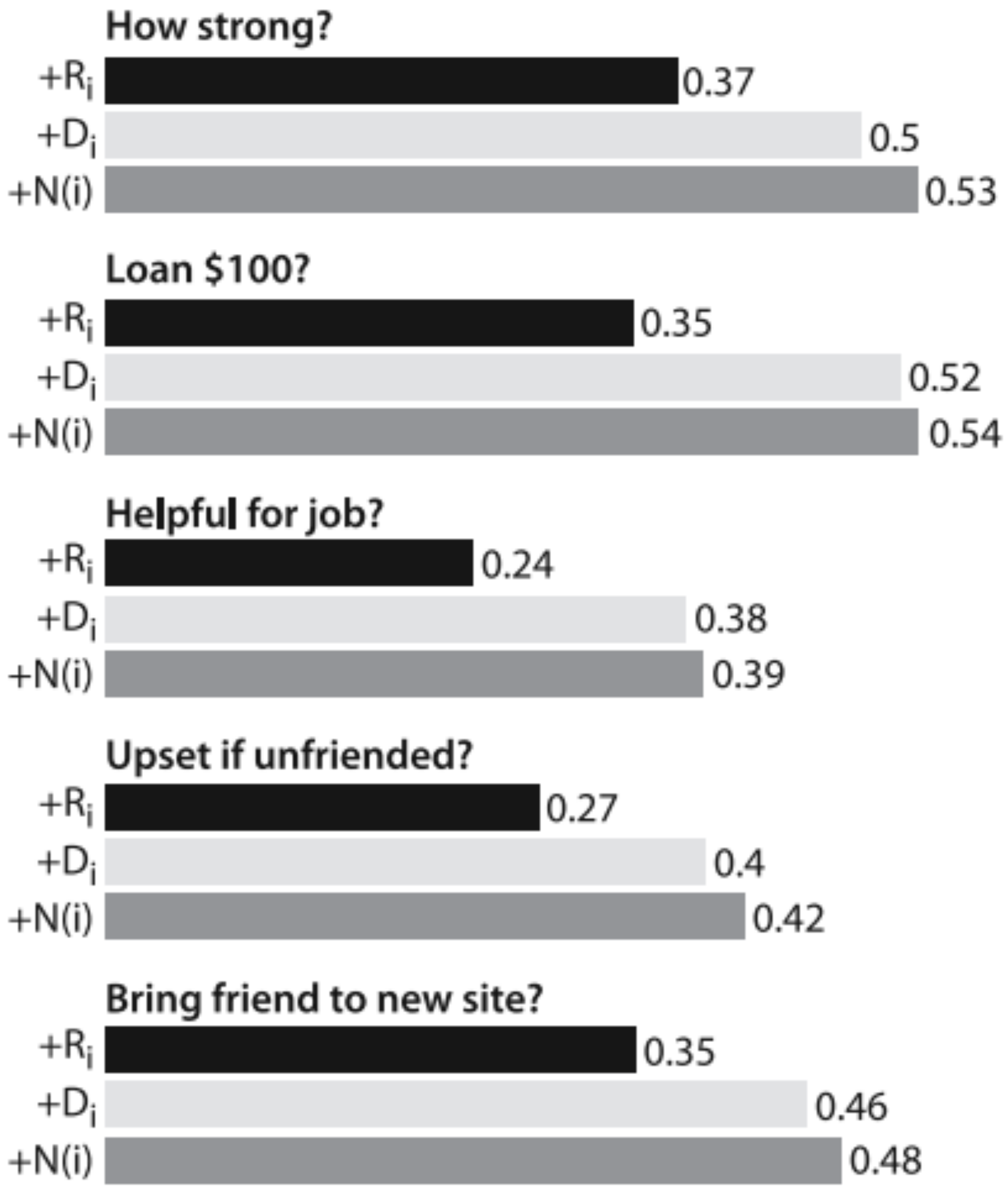
Let’s predict tie strength on Facebook

What features we could use to predict self-reported tie strength?

Predictive Intensity Variables	**Distribution**	**Max**
Wall words exchanged		9549
Participant-initiated wall posts		55
Friend-initiated wall posts		47
Inbox messages exchanged		9
Inbox thread depth		31
Participant’s status updates		80
Friend’s status updates		200
Friend’s photo comments		1352
Intimacy Variables		
Participant’s number of friends		729
Friend’s number of friends		2050
Days since last communication		1115
Wall intimacy words		148
Inbox intimacy words		137
Appearances together in photo		73
Participant’s appearances in photo		897
Distance between hometowns (mi)		8182
Friend’s relationship status	6% engaged, 32% married, 30% single, 30% in relationship	

Duration Variable		
Days since first communication		1328
Reciprocal Services Variables		
Links exchanged by wall post		688
Applications in common		18
Structural Variables		
Number of mutual friends		206
Groups in common		12
Norm. TF-IDF of *interests* and *about*		73
Emotional Support Variables		
Wall & inbox positive emotion words		197
Wall & inbox negative emotion words		51
Social Distance Variables		
Age difference (days)		5995
Number of occupations difference		8
Educational difference (degrees)		3
Overlapping words in *religion*		2
Political difference (scale)		4

Let’s predict tie strength on Facebook



The model’s Adjusted R2 values for all five dependent variables, broken down by the model’s three main terms.

Modeling interactions between tie strength dimensions results in a substantial performance boost.

The model performs best on Loan $100? and How strong?, the most general question

Let's predict tie strength on Facebook

Top 15 Predictive Variables	β	F	p-value
Days since last communication	-0.76	453	< 0.001
Days since first communication	0.755	7.55	< 0.001
Intimacy × Structural	0.4	12.37	< 0.001
Wall words exchanged	0.299	11.51	< 0.001
Mean strength of mutual friends	0.257	188.2	< 0.001
Educational difference	-0.22	29.72	< 0.001
Structural × Structural	0.195	12.41	< 0.001
Reciprocal Serv. × Reciprocal Serv.	-0.19	14.4	< 0.001
Participant-initiated wall posts	0.146	119.7	< 0.001
Inbox thread depth	-0.14	1.09	0.29
Participant's number of friends	-0.14	30.34	< 0.001
Inbox positive emotion words	0.135	3.64	0.05
Social Distance × Structural	0.13	34	< 0.001
Participant's number of apps	-0.12	2.32	0.12
Wall intimacy words	0.111	18.15	< 0.001

The fifteen predictive variables with highest standardized beta coefficients.

The two Days since variables have large coefficients because of the difference between never communicating and communicating once.

The utility distribution of the predictive variables forms a power-law distribution: **with only these fifteen variables, the model has over half of the information it needs to predict tie strength.**

Let’s predict tie strength on Facebook

Don’t forget error analysis

rating: 0.96; prediction: 0.47

This friend is very special. He and I attended the same high school, we interacted a lot over 3 years and we are very very close. We trust each other. My friend are I are still interacting in ways other than Facebook such as IM, emails, phones. Unfortunately, that friend and I rarely interact through Facebook so I guess your predictor doesn't have enough information to be accurate.

rating: 0; prediction: 0.44

I don't know why he friended me. But I'm easy on Facebook, because I feel like I'm somehow building (at least a miniscule amount of) social capital, even when I don't know the person. We went to the same high school and have a few dozen common friends. We've never interacted with each other on Facebook aside from the friending.

rating: 0.6; prediction: 0.11

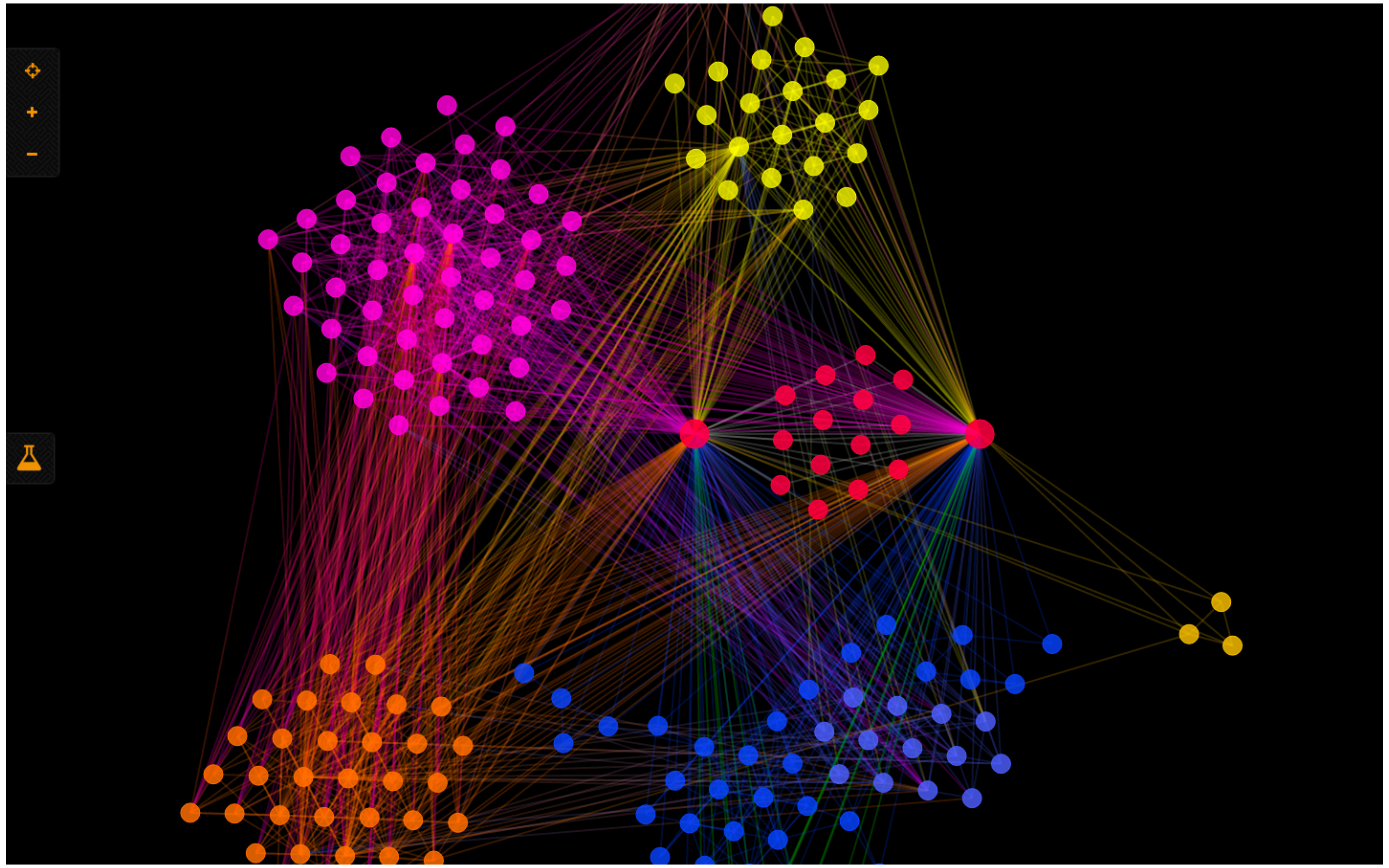
Ah yes. This friend is an old ex. We haven't really spoken to each other in about 6 years, but we ended up friending each other on Facebook when I first joined. But he's still important to me. We were best friends for seven years before we dated. So I rated it where I did (I was actually even thinking of rating it higher) because I am optimistically hoping we’ll recover some of our “best friend”-ness after a while. Hasn't happened yet, though.

Clustering

Clustering

Group a set of data points into a number of clusters, so that

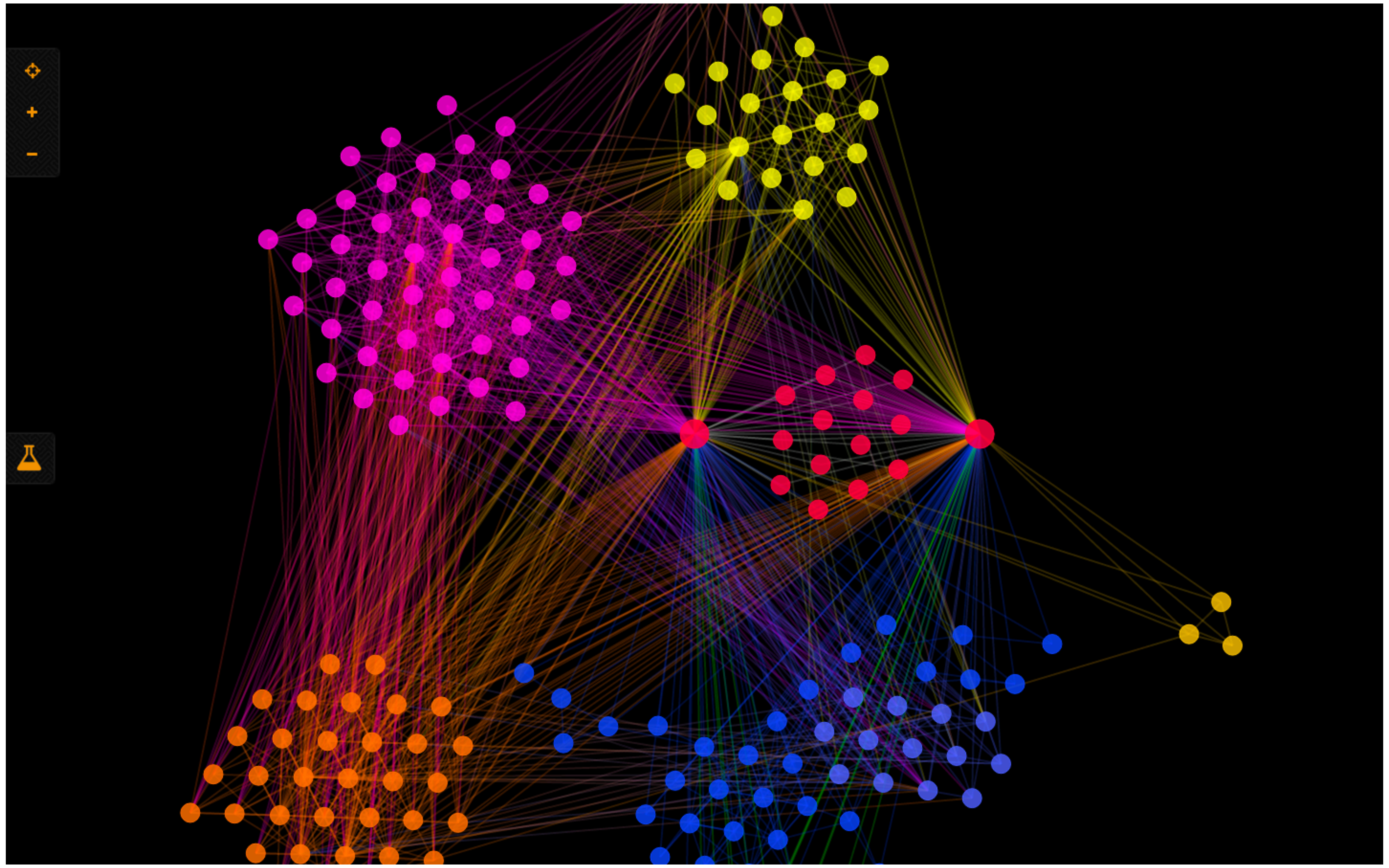
- Data points in the same cluster are similar to each other
- Data points in different clusters are dissimilar



https://graphalchemist.github.io/Alchemy/images/features/cluster_team.png

Clustering

Finding structures in data, using just X



https://graphalchemist.github.io/Alchemy/images/features/cluster_team.png

What are Structures?

Partitioning a group of data point into K disjoint sets (K-means clustering)

Assigning X to hierarchical structures (Hierarchical clustering)

Assigning X to partial membership in K different sets (Graphic models, GMM)

Learning a representation of x that puts similar data points closer to each other (Deep learning)

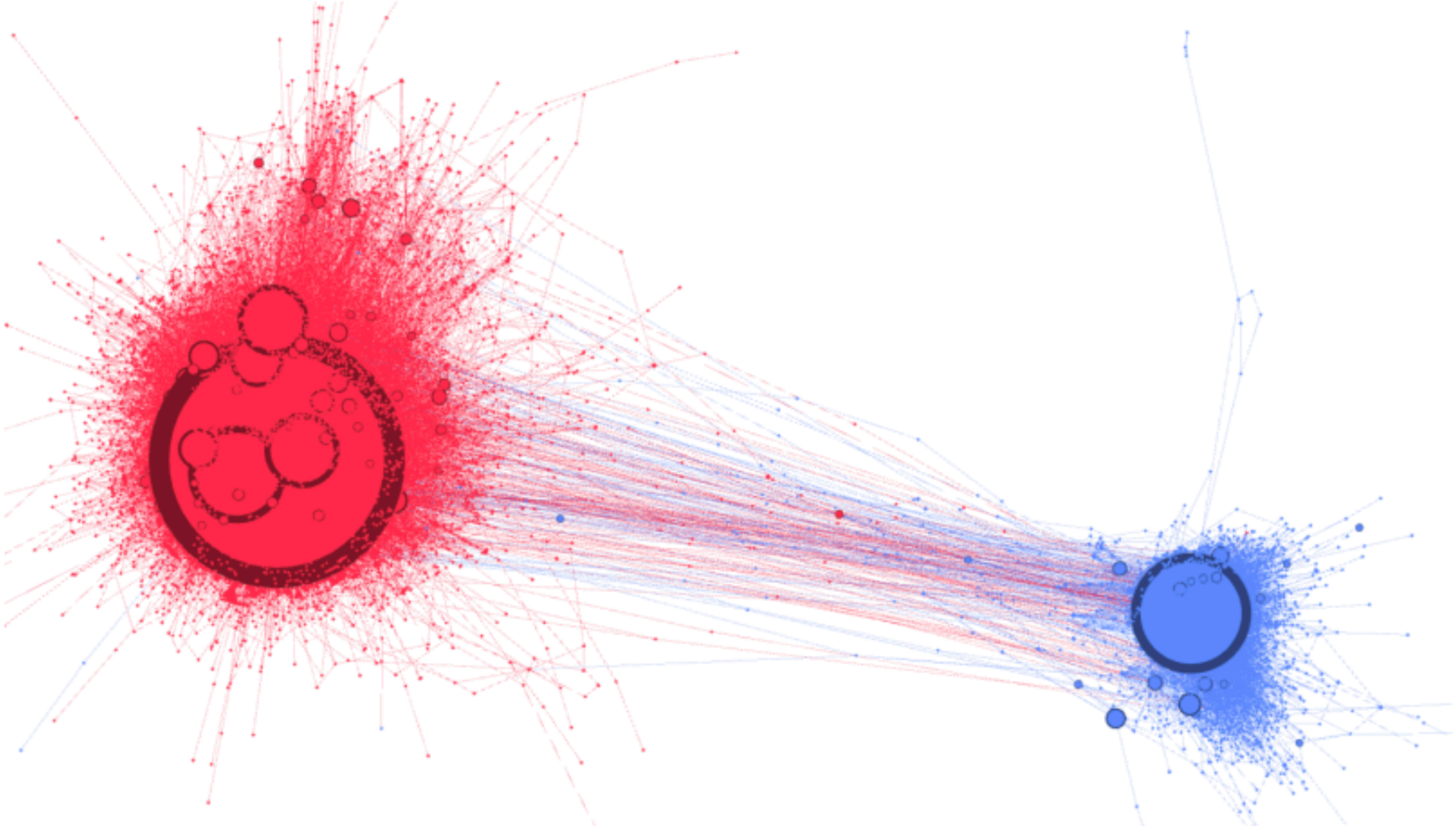
Slide content credit to David Bamman

Why and when do I need clustering?

Discovering interesting or unexpected structures can be useful for hypothesis generation

Unsupervised learning generates alternative representation **as features** for some subsequent supervised models

Slide content credit to David Bamman



The structure of the White Helmets discourse has two clear clusters of accounts—a pro-White Helmets cluster that supports the organization and an anti-White Helmets cluster that criticizes them, using Twitter conversations.

Wilson, Tom, and Kate Starbird. "Cross-platform disinformation campaigns: lessons learned and next steps." Harvard Kennedy School Misinformation Review 1, no. 1 (2020).

Key Design Choices for Clustering

How to **represent** each data point?

How to calculate the **similarity** between data points?

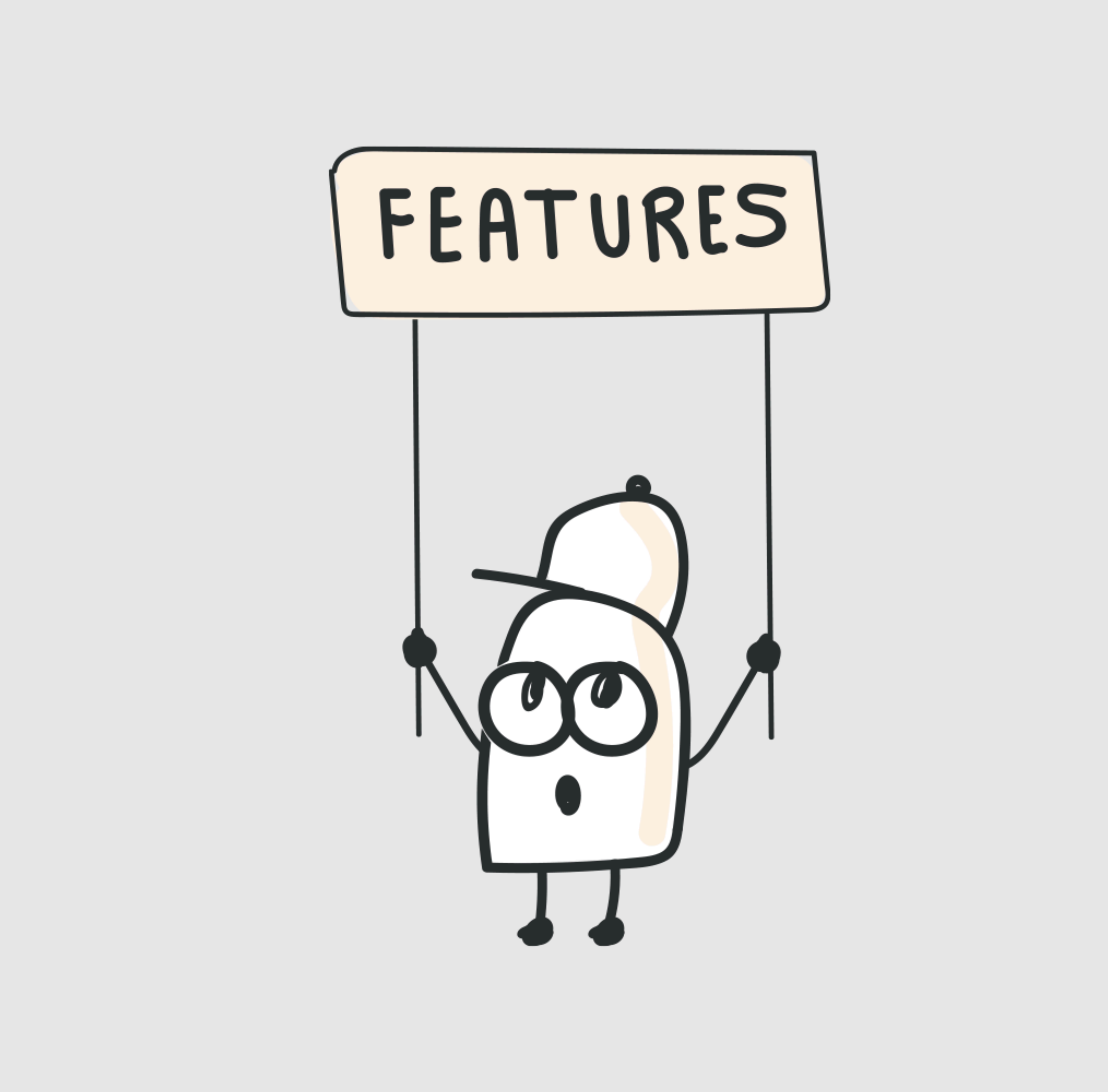
What is the **number of clusters** to use?

How can we **evaluate** the resulting clusters?

Representation

Unigrams, bigrams
Word embeddings,
metadata …

This is a huge decision that impacts what you can learn



https://www.upvoty.com/how-to-avoid-building-features-that-nobody-will-use/

Similarity

Cosine similarity for vectors $\dfrac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$

Jaccard similarity for sets $\dfrac{|X \cap Y|}{|X \cup Y|}$

Euclidean distance for points $\sqrt{\sum |x_i - y_i|^2}$

Number of Clusters

When our desired number of clusters is obtained

Assume we know the best number of clusters

Or when stopping criterion is met

E.g., stop if similarity exceeds threshold

Evaluation

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

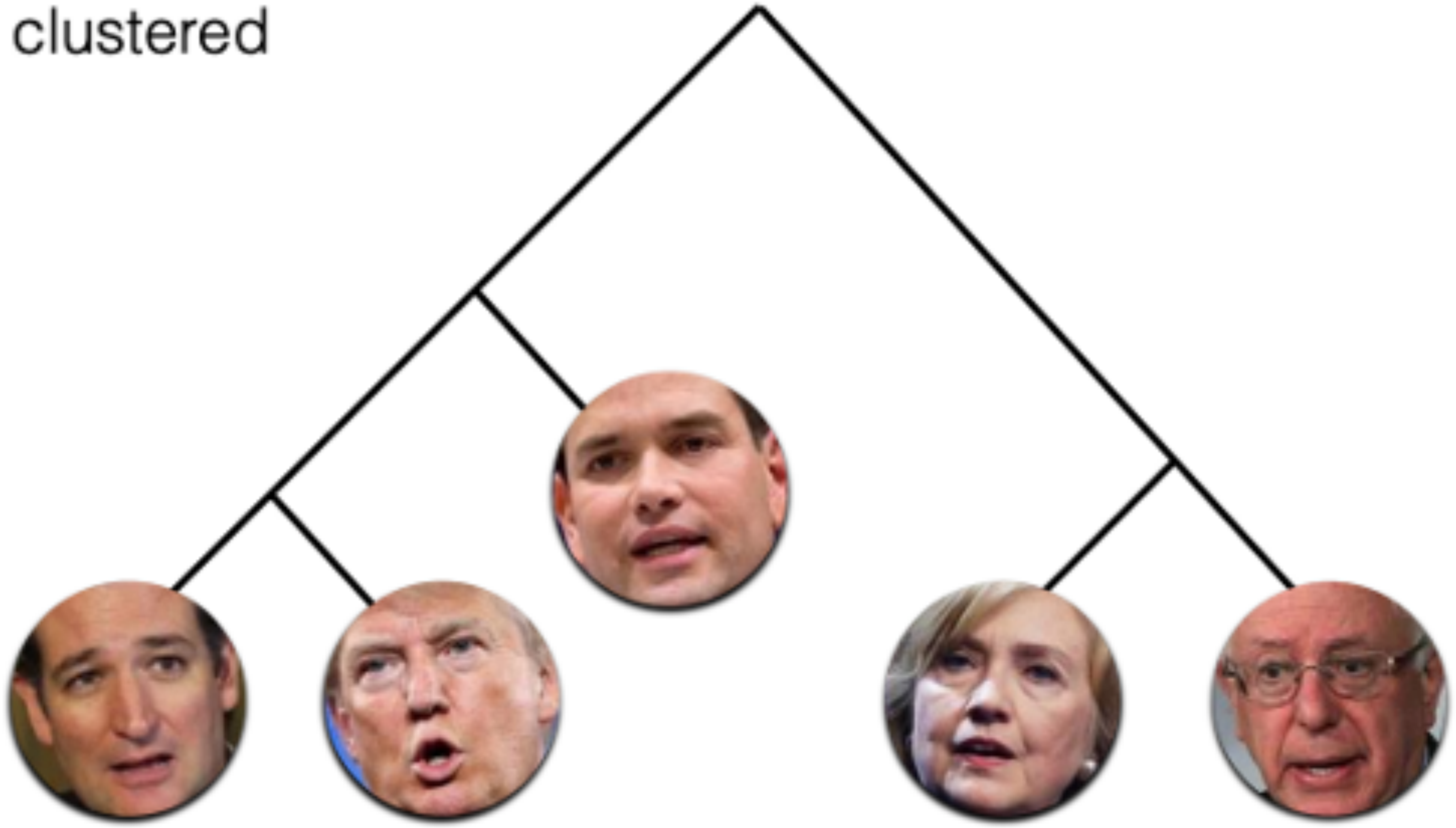
Elements in different clusters should be less similar to each other

External criteria

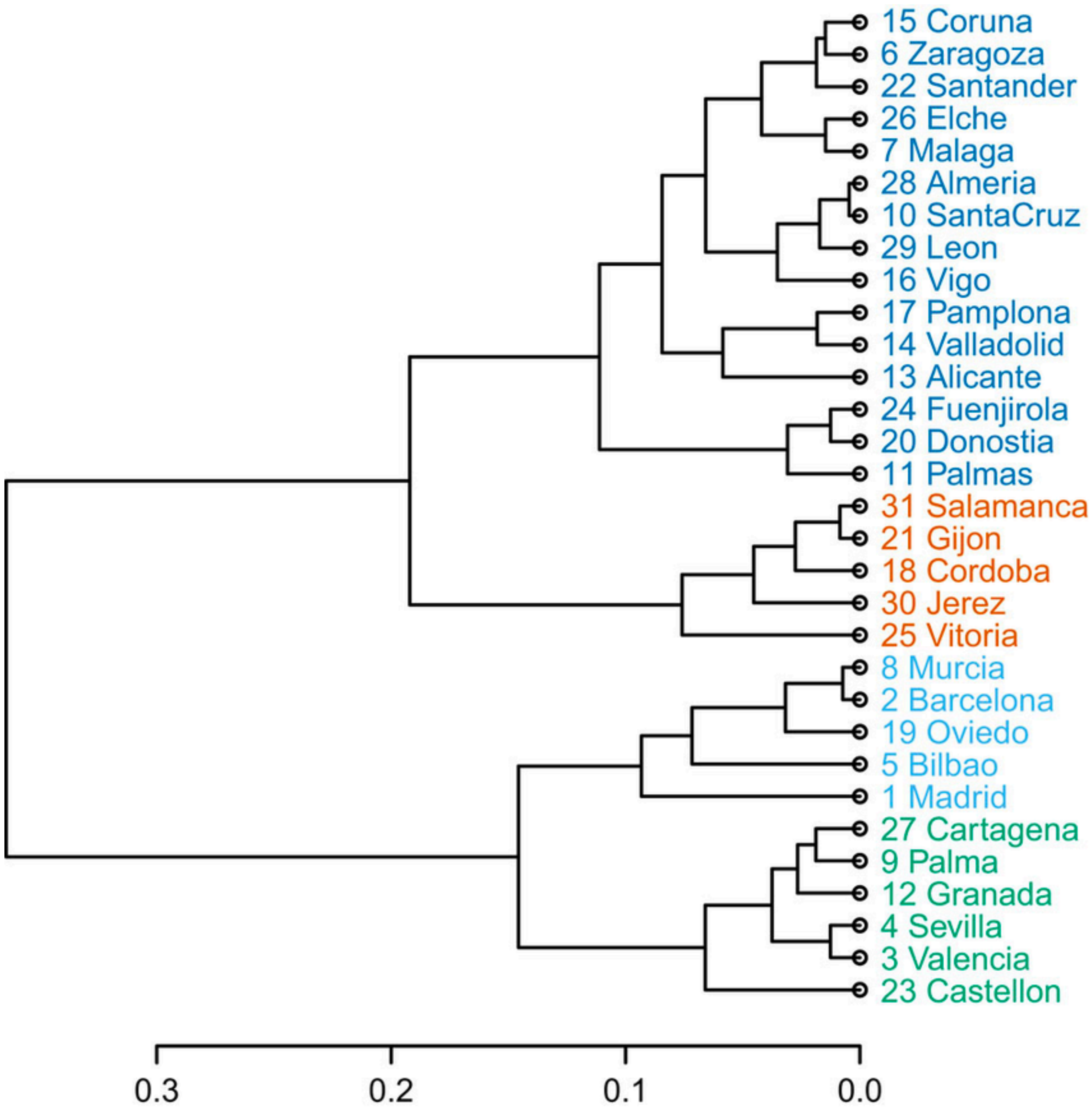
How closely does your clustering reproduce gold standard clustering?

Some highlight: Hierarchical Clustering

Hierarchical order
among the elements
being clustered



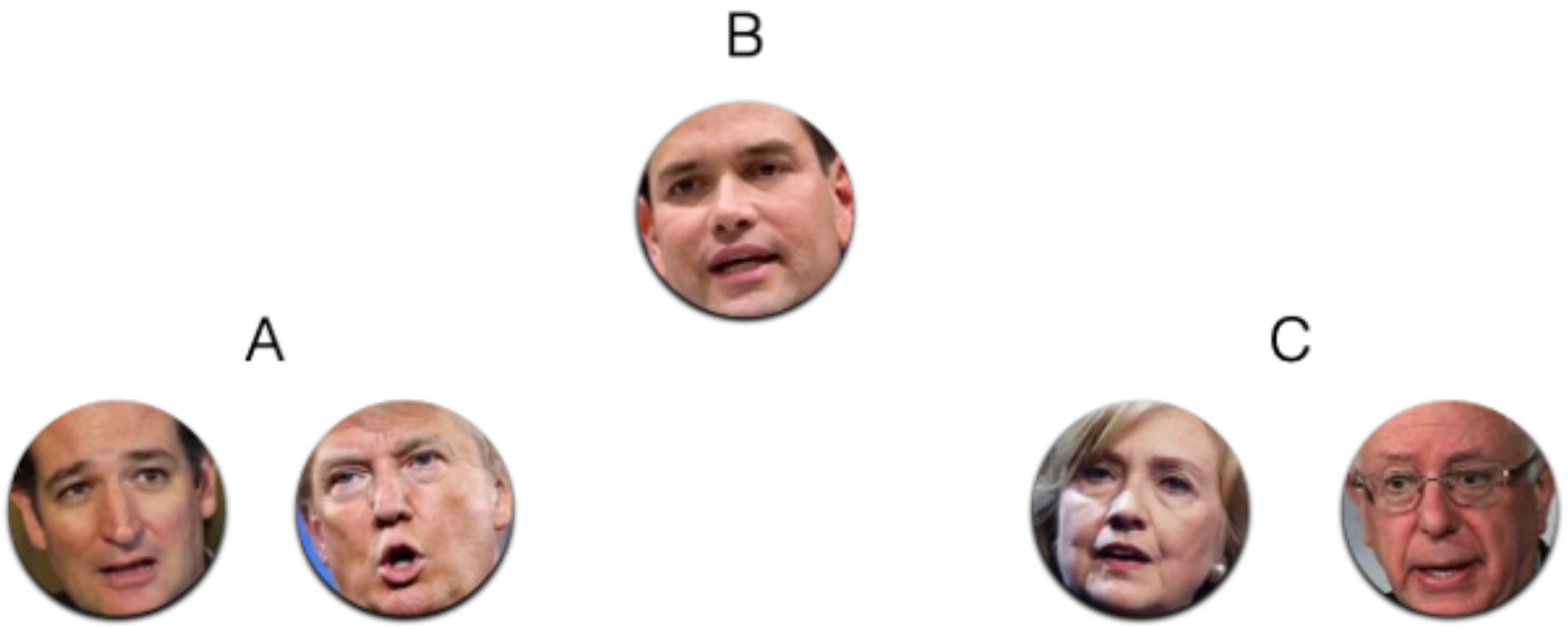
Slide content credit to David Bamman



Slide content credit to David Bamman

Louail, Thomas, Maxime Lenormand, Miguel Picornell, Oliva Garcia Cantu, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, and Marc Barthelemy. "Uncovering the spatial structure of mobility networks." Nature communications 6, no. 1 (2015): 1-

Some highlight: K-means Clustering



Slide content credit to David Bamman

Some highlight: K-means Clustering

Given a set of data points $\{x_1, x_2, x3, ...x_m\}$

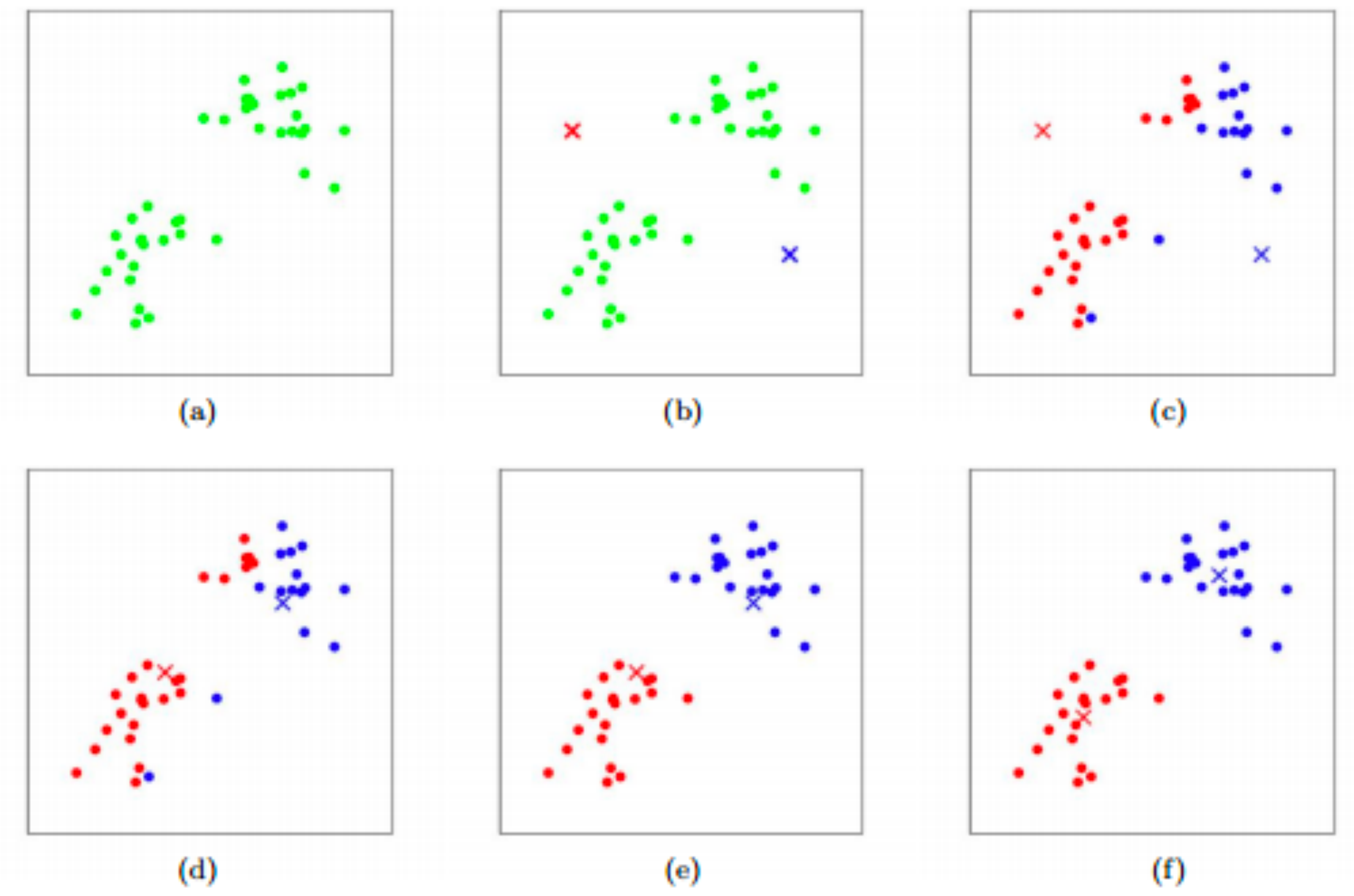
First initialize cluster centroid $\{\mu_1, \mu_2, \ldots, \mu_k\}$ randomly

Repeat until convergence:

Assign labels $c_i := \arg\min_j \| x_i - \mu_j \|^2$

Update centroids $\mu_j := \dfrac{\sum_{i=1}^{m} \mathbf{1}\{c_i = j\} x_i}{\sum_{i=1}^{m} \mathbf{1}\{c_i = j\}}$

Some highlight: K-means Clustering



K-means algorithm. Training examples are shown as dots, and cluster centroids are shown as crosses.

(a) Original dataset.
(b) Random initial cluster centroids.
(c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid (shown by "painting" the training examples the same color as the cluster centroid to which is assigned); then we move each cluster centroid to the mean of the points assigned to it.

Images courtesy of Michael Jordan.

Checkout: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

Let’s find different groups of people in support groups

Imagine this is on an online social support community …

1. **Why** is this a clustering task?
2. **What** is “group” of people?
3. How can we get the **ground truth**?
4. How **many** groups?
5. What **features** should we use?
6. How can we **evaluate** it?

Let's find different groups of people in support groups

Imagine this is on an online social support community …

We need to come up with **a lot of features**

Agent: `members on CSN …`

Interaction: `medical/treatment topics, emotions …`

Expectation: `report to moderators …`

Context: `private vs. public discussion …`

Goal: `social support …`

Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities

Diyi Yang
Language Technologies Institute
Carnegie Mellon University
diyiy@andrew.cmu.edu

Robert Kraut
Human-Computer Interaction Institute
Carnegie Mellon University
robert.kraut@cmu.edu

Tenbroeck Smith
Behavioral Research
American Cancer Society
tenbroeck.smith@cancer.org

Elijah Mayfield
Language Technologies Institute
Carnegie Mellon University
elijah@cmu.edu

Dan Jurafsky
Department of Linguistics
Stanford University
jurafsky@stanford.edu

Let's find different groups of people in support groups

Agent: members on CSN …
Interaction: medical/treatment topics, emotions …
Expectation: report to moderators …
Context: private vs. public discussion …
Goal: social support

The Facet of Goal: Social Support

Since you are a triple positive they can put you on hormones and the chance of recurrence is low. Listen to your chemo nurse ...

Informational Support

It gives me faith that you can have cancer and live a full life. Sorry to hear that. God bless you. Please stay strong!

Emotional Support

Let's find different groups of people in support groups

Intuition: *a user is a mixture of different social roles*

Let's find different groups of people in support groups

Emotional Support Provider	Private Support Provider
Newcomer Welcom	All-round Expert
Informational Support	
Story Sharer	
Informational Support	
Private Communic	

Role Name	Prevalence (%)	Typical Behaviors Listed in Importance
Emotional Support Provider	33.3	Provide emotional support Provide empathy Participate in a large number of cancer-specific forums
Welcomer	15.9	Frequently talking to newcomers Provide encouragement Higher number of replies
Informational Support Provider	13.3	Provide informational support Higher usage of words related to symptoms and treatment
Story Sharer	10.2	Higher level of self-disclose Seek emotional support Initialize higher number of threads

Let's find different groups of people in support groups

Work with 6 moderators on CSN to assess the derived roles

*" It seems very **comprehensive** and there are so many different examples, so I feel like it is **covered very well** with your different roles and labels. "*

The identified roles were comprehensive

Is it a classification/regression/clustering problem?

I want to predict a star value {1,2,3,4,5} for a product review

I want to find all of the texts that have allusions to Paradise Lost

I want to predict the stock price

I want to tell which team will win

I want to associate photographs of cats with animals in a taxonomic hierarchy

I want to reconstruct an evolutionary tree for languages

Computational Social Science in the Age of Big Data

danah boyd and Kate Crawford (2012), "Critical Questions for Big Data," Information, Communication and Society

1 “Big data” changes the definition of knowledge

How do computational methods/quantitative analysis pragmatically affect epistemology?

Restricted to what data is available (twitter, data that’s digitized, google books, etc.). How do we counter this in experimental designs?

Establishes alternative norms for what “research” looks like

2 Claims to objectivity and accuracy are misleading

Data collection, selection process is subjective, reflecting belief in what matters.

Model design is likewise subjective

- model choice (classification vs. clustering etc.)
- representation of data
- feature selection

Claims need to match the sampling bias of the data

3 Bigger data is not always better data

Uncertainty about its source or selection mechanism [Twitter, Google books]

Appropriateness for question under examination

How did the data you have get there?

Are there other ways to solicit the data you need?

Remember **the value of small data**: individual examples and case studies

4 Taken out of context, big data loses its meaning

A representation (through features) is a necessary approximation; what are the consequences of that approximation?

Example: quantitative measures of “tie strength” and its interpretation

5 Just because it is accessible does not make it ethical

Anonymization practices for sensitive data (even if born public)

Accountability both to research practice and to subjects of analysis

6 Limited access to big data creates new digital divides

Inequalities in access to data and the production of knowledge

Privileging of skills required to produce knowledge