Resource Description

David Bamman
Info 202: Information Organization and Retrieval

September 28, 2016

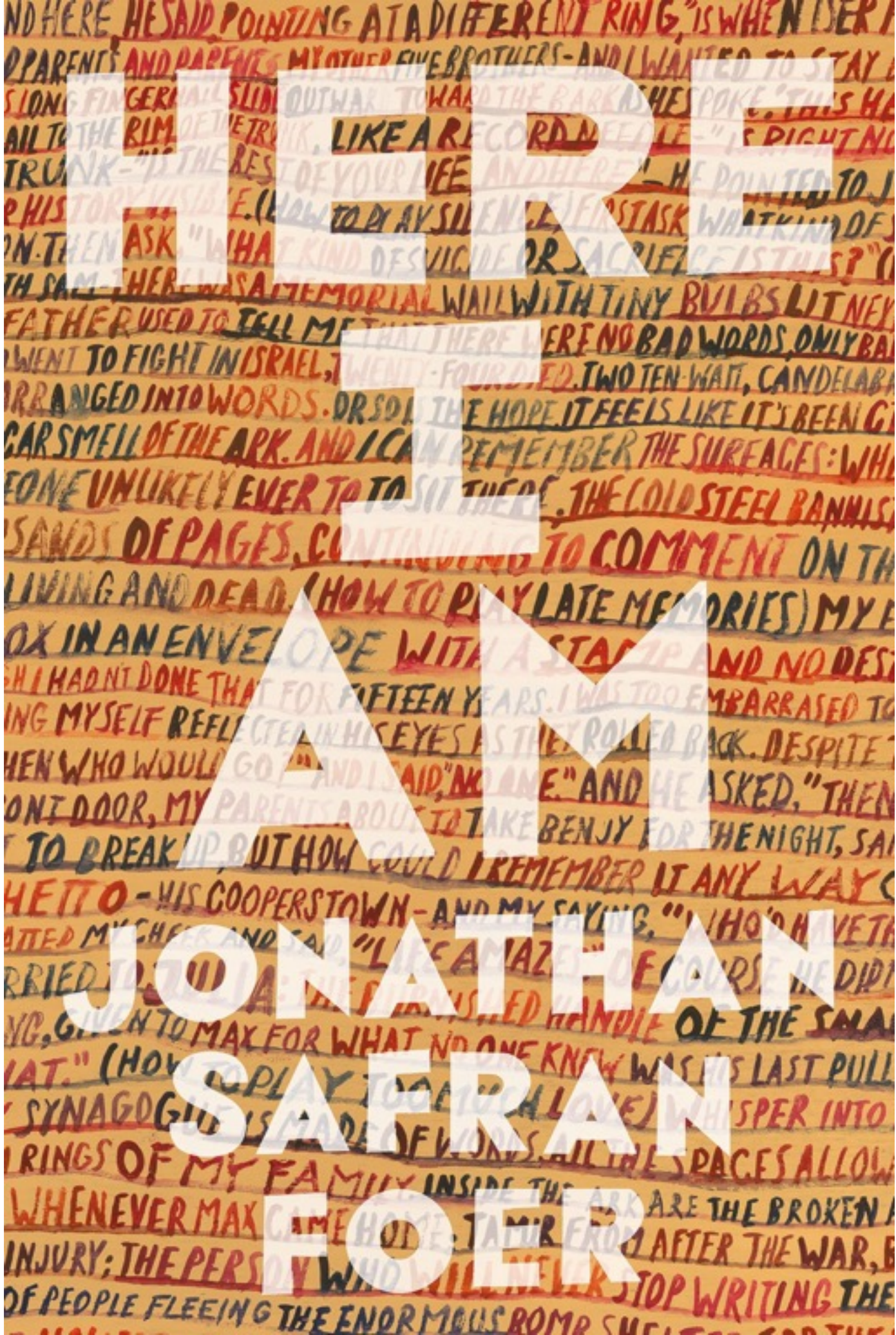
Data science lifecycle

Cross Industry Standard Process for Data Mining (CRISP-DM)

Feature engineering

How do we represent a given data point in a computational model?

How do we describe a given resource in a computational model?



Resource description

- We describe resources so we can refer to them, select them, organize them, interact with them (especially to compare them), and maintain them

- But different types of resources must also have differentiating properties, or there would be no basis or reason to distinguish them

Resource description

Interactions:

- presentation to public in museum
- shipping to Metropolitan Museum
- restoring cracked surface

Resource description

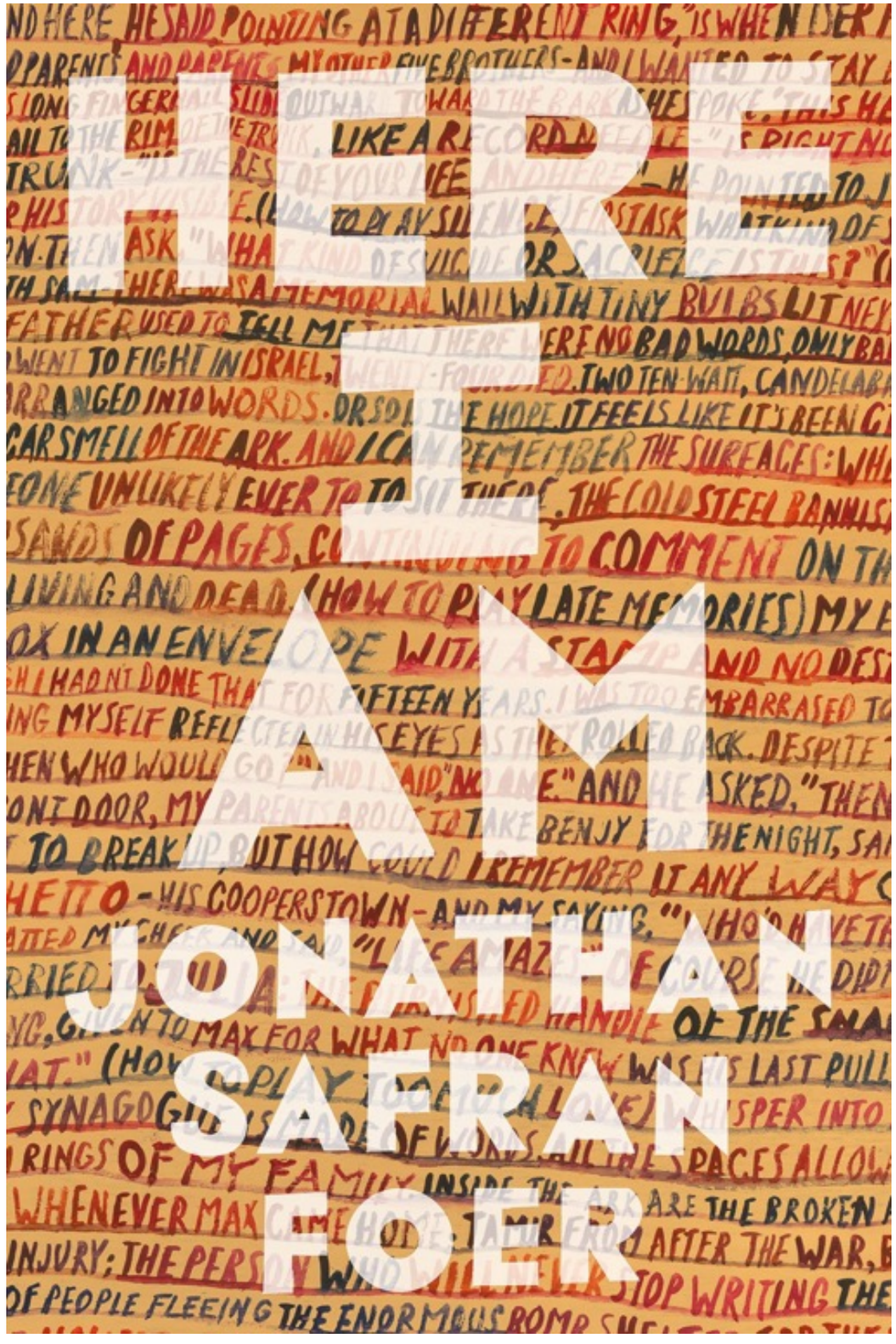
- Descriptions are tied to the interactions they are meant to support (e.g., informational description of book vs. physical properties)

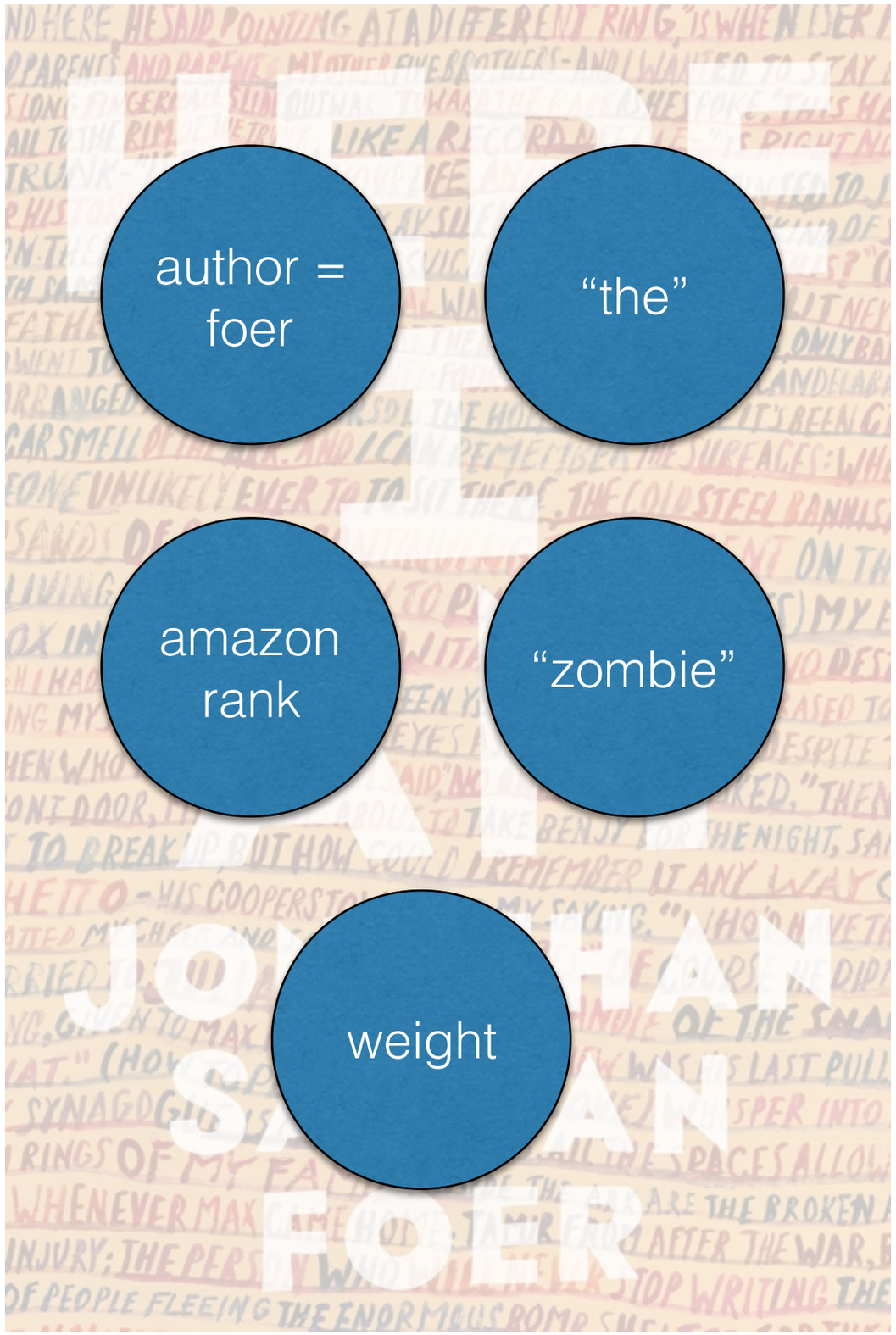
Resource description in data science

- Resource description = “feature”

- Description holds both to the description of the object (= predictor) and to any label (= response)

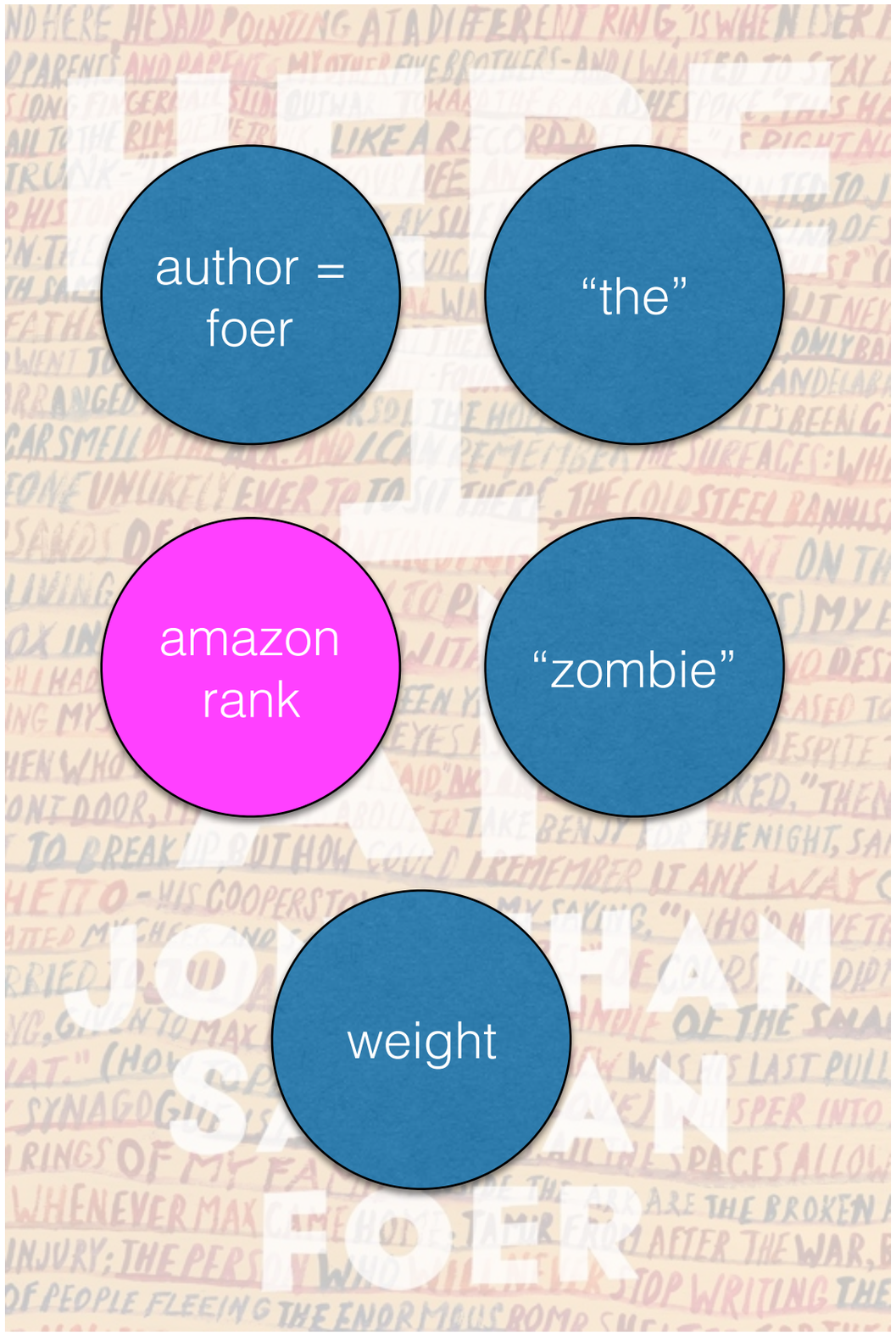
author: foer	TRUE
author: austen	FALSE
pub year	2016
height (inches)	9.2
weight (pounds)	2
contain: the	TRUE
contains: zombies	FALSE
amazon rank @ 1 month	159

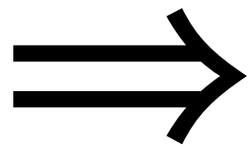


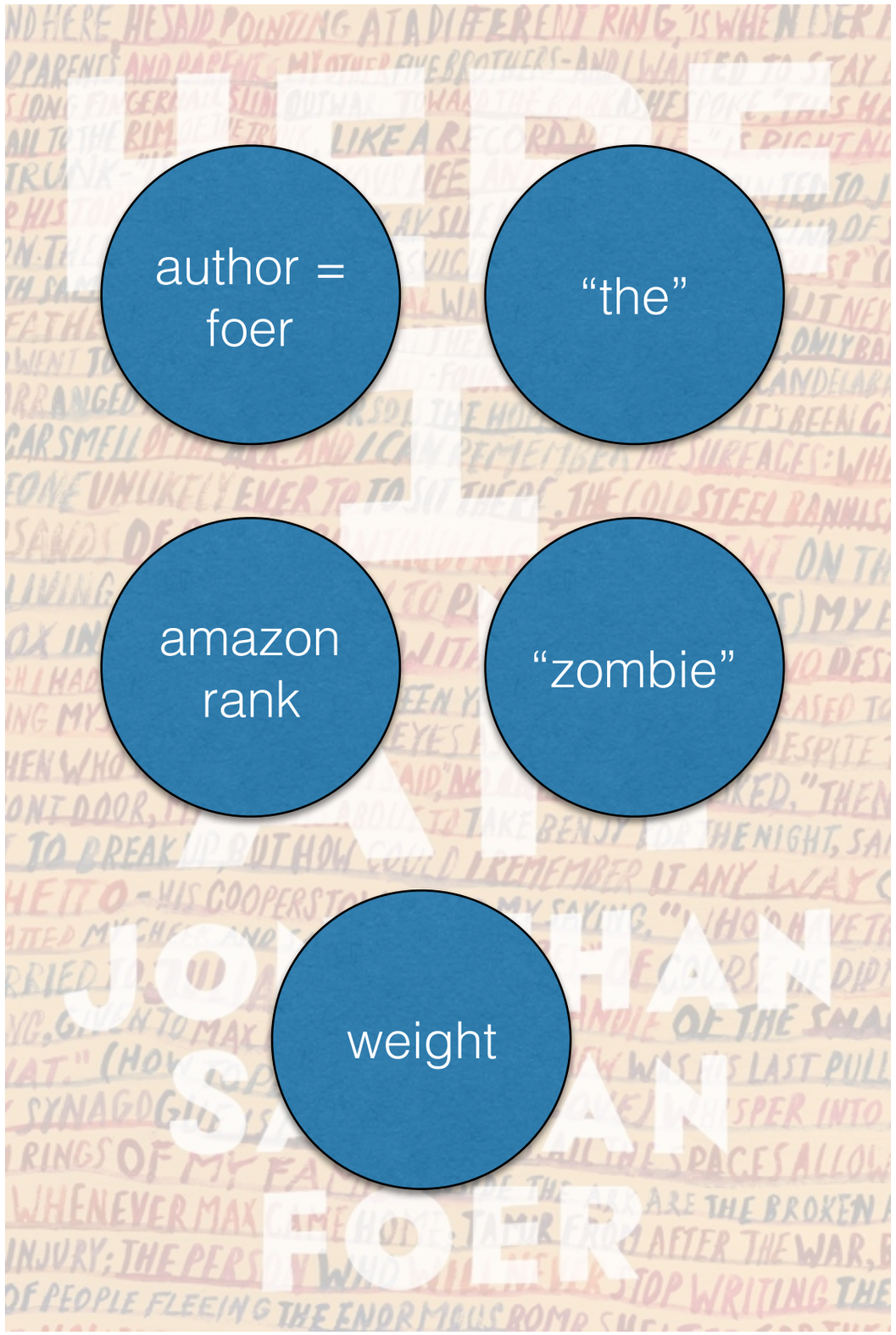

author =
foer
“the”
amazon
rank
“zombie”
weight

predictor

response

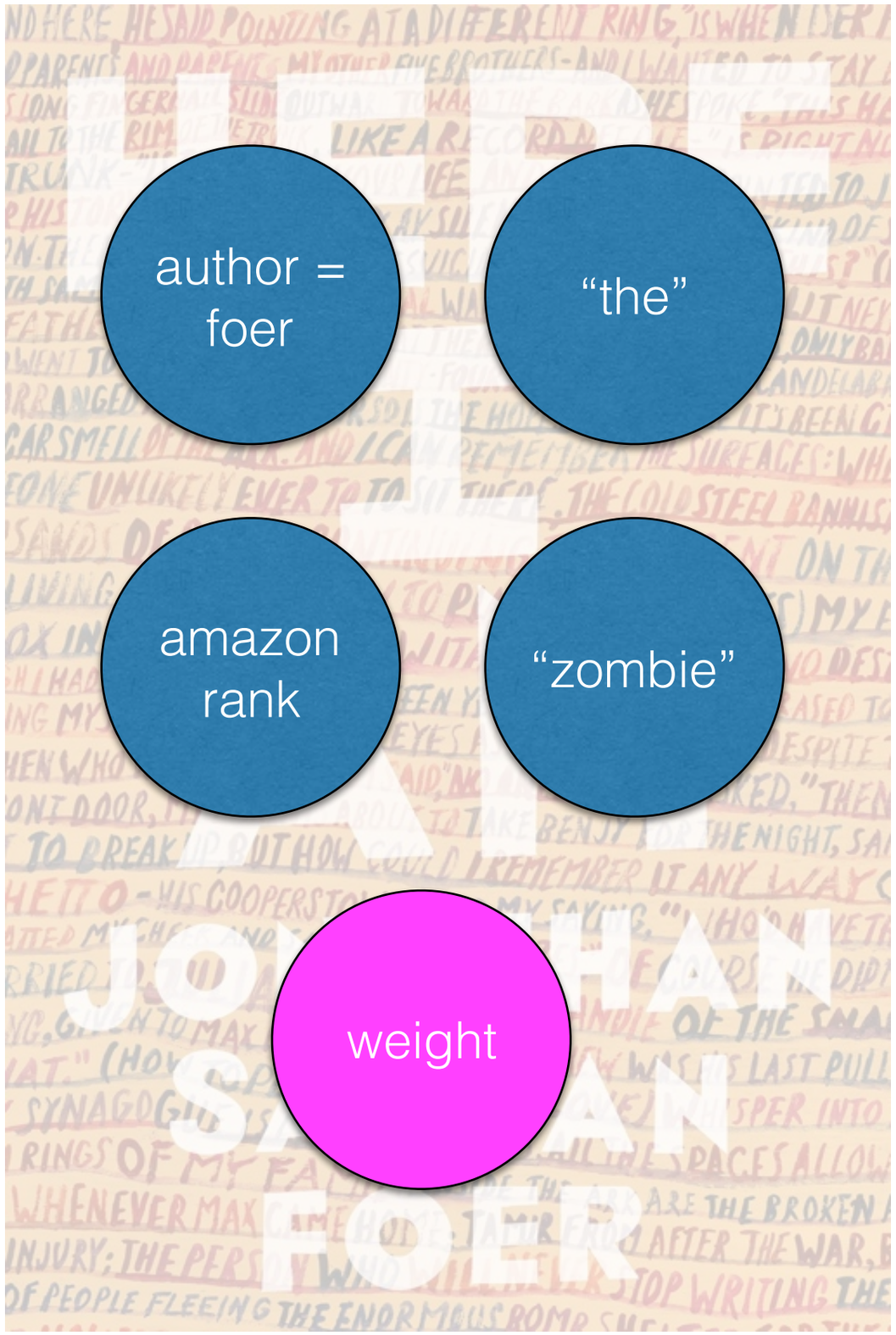





author =
foer
“the”
amazon
rank
“zombie”
weight

predictor

response



⇒

Levels of measurement

- Binary indicators
- Counts
- Frequencies
- Ordinal

Feature design

- What features to include?
- How do we operationalize them? What values are we encoding in that operationalization?
- How do we assign their levels?

Design choices

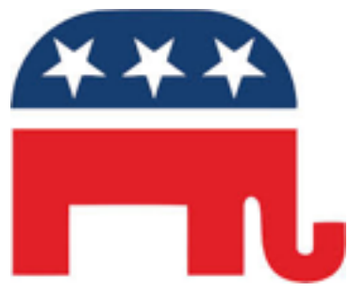
- Gender
- Intrinsic/extrinsic?
- Static/dynamic?
- Binary/n-ary?

Facebook gender options

- Agender
- Androgyne
- Androgynous
- Bigender
- Cis
- Cisgender
- Cis Female
- Cis Male
- Cis Man
- Cis Woman
- Cisgender Female
- Cisgender Male
- Cisgender Man
- Cisgender Woman
- Female to Male
- FTM
- Gender Fluid
- Gender Nonconforming
- Gender Questioning
- Gender Variant
- Genderqueer
- Intersex
- Male to Female
- MTF
- Neither
- Neutrois

Design choices

- Political preference
  - Intrinsic/extrinsic?
  - Static/dynamic?
  - Binary/n-ary?
  - Categorical/real valued?
  - One dimension or several dimensions?)



Scope

- Properties that obtain only of the resource

- Contextual properties (relate to the situation in which a thing exists)

Resource description

David Bamman
@dbamman

Assistant Professor, School of Information, UC Berkeley. Natural language processing, machine learning, computational social science, digital humanities.

Berkeley, CA

people.ischool.berkeley.edu/~dbamman/

Joined October 2009

TWEETS 542 | FOLLOWING 455 | FOLLOWERS 990 | LIKES 162 | LISTS 2

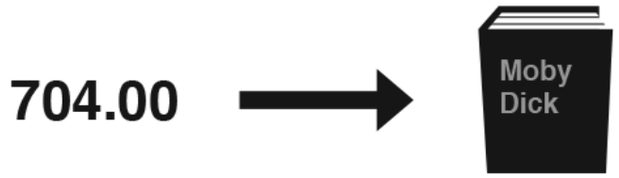
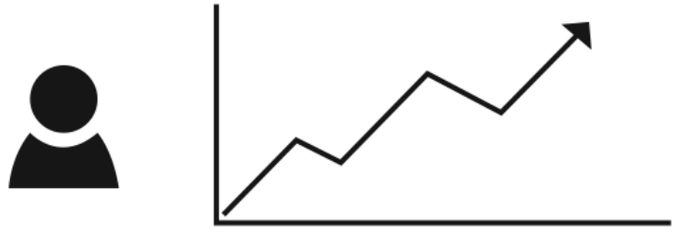
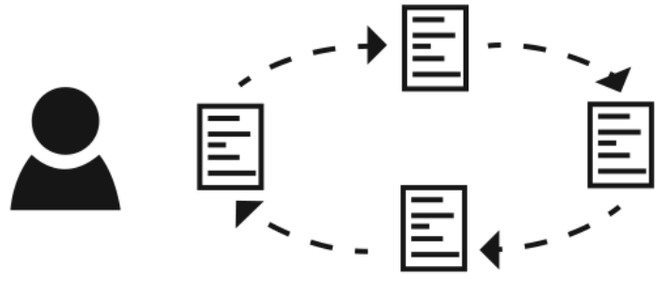
Tweets | Tweets & replies | Media

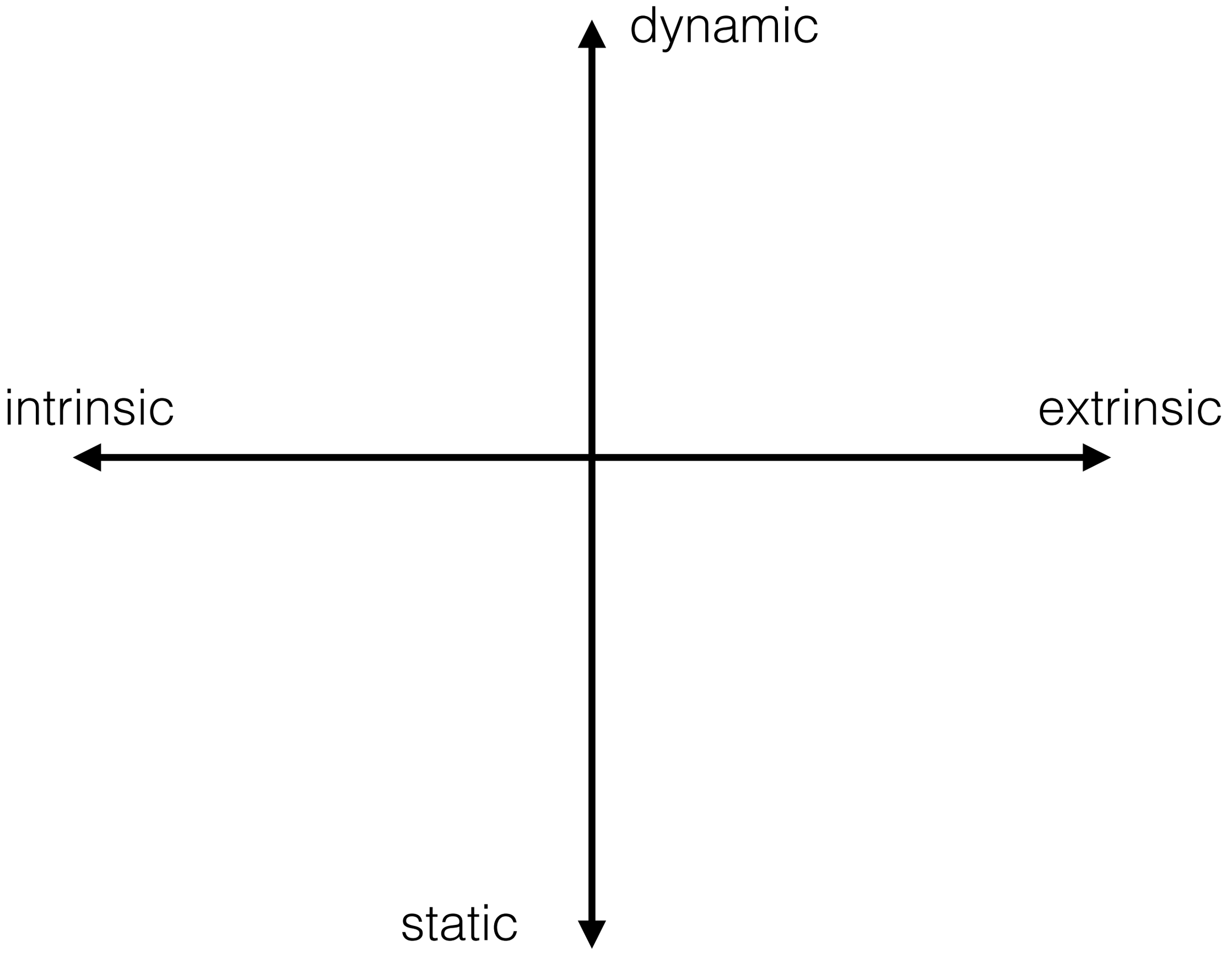
David Bamman @dbamman · Sep 23

Rounding out a quick NY trip for @NYUDataScience with a talk here today

Resource description

Property Essence

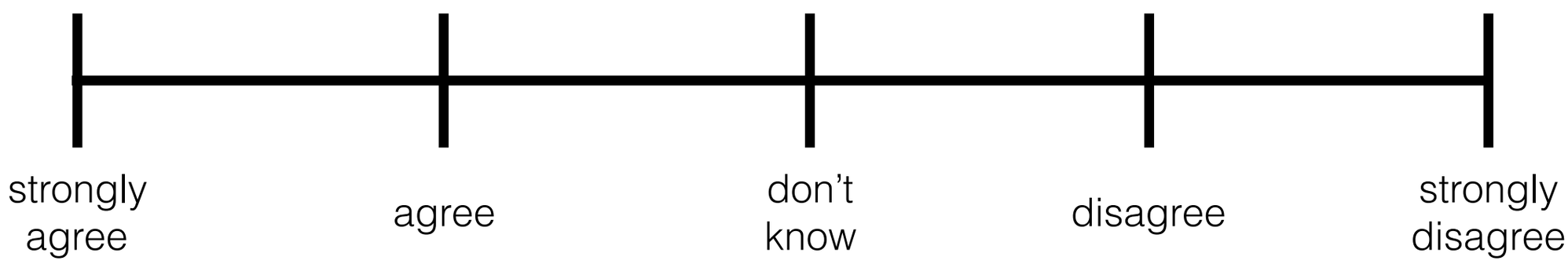
Property Persistence	Intrinsic	Extrinsic
Static	**Intrinsic Static** **Definition**: Directly experienced, subject matter, implicit, inherent properties. **Examples**: Size, color, shape, author, date of creation. 	**Extrinsic Static** **Definition**: Assigned to resource, name, identifier. **Examples**: Dewey decimal 
Dynamic	**Intrinsic Dynamic** **Definition**: Inherent properties; change over time. **Examples**: Skills, experience 	**Extrinsic Dynamic** **Definition**: Behavioral and contextual properties **Examples**: Current owner, location, best seller lists. 


dynamic
intrinsic
extrinsic
static

Extrinsic measurement

- Human judgments
- Likert scale
  - Fixed-choice
  - Measure attitudes
  - Usually 5 or 7

"Global warming is one of the most important issues today"



Stability

- The degree to which some feature value is stable over repeated measurements of the same thing

- Physical instruments (e.g., thermometers) should give the same readings under the same conditions

- Subjective measures (e.g., survey responses) should have similar responses by the same individual

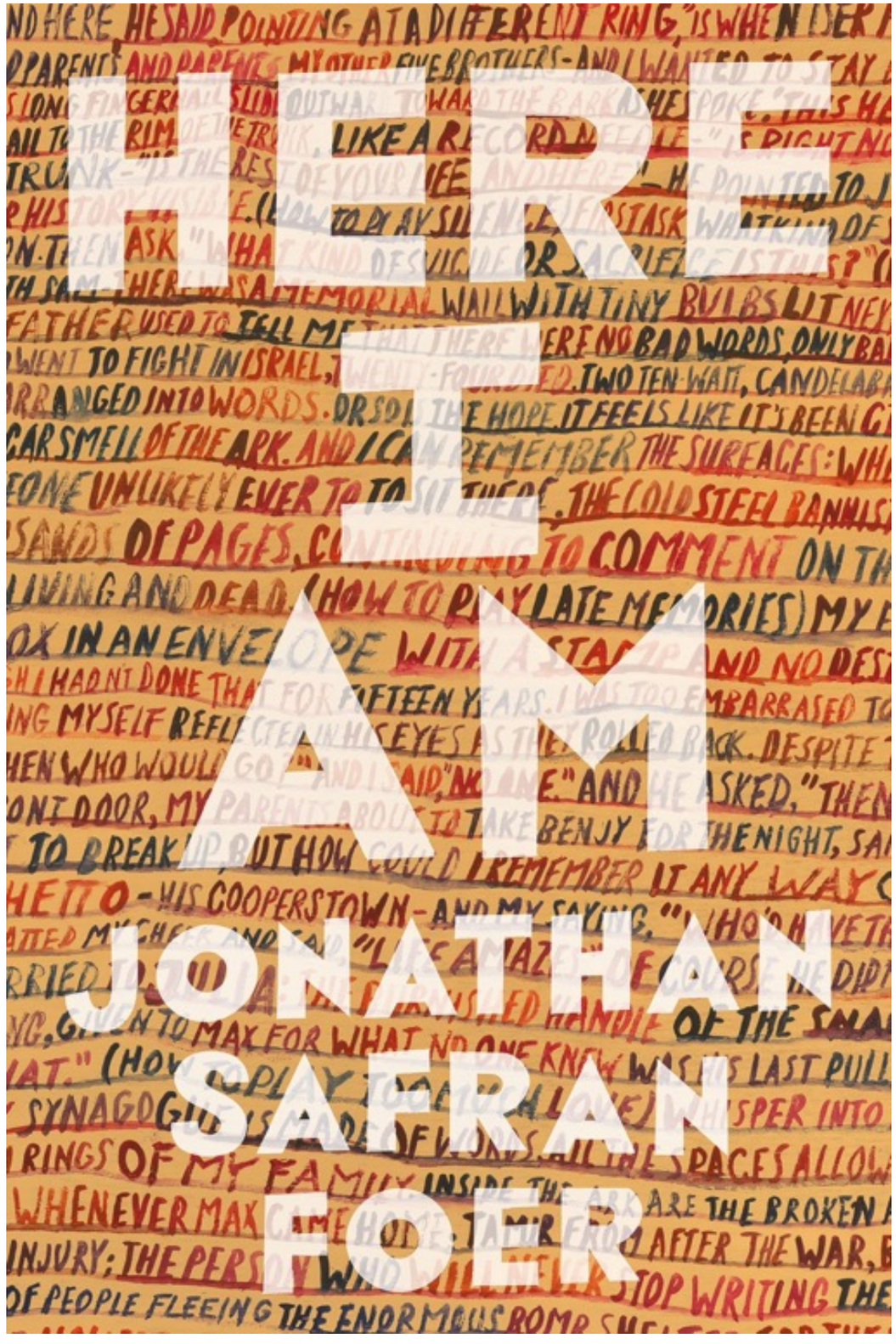
Consistency

- The degree to which some feature value is consistent over repeated measurements of different things.

- If asking different people for judgments, how often do they give the same response?

- Inter-annotator agreement, inter-coder reliability, etc.

author: foer	TRUE
author: austen	FALSE
pub year	2016
height (inches)	9.2
weight (pounds)	2
contain: the	TRUE
contains: zombies	FALSE
amazon rank @ 1 month	159



genre: fiction

genre: world literature

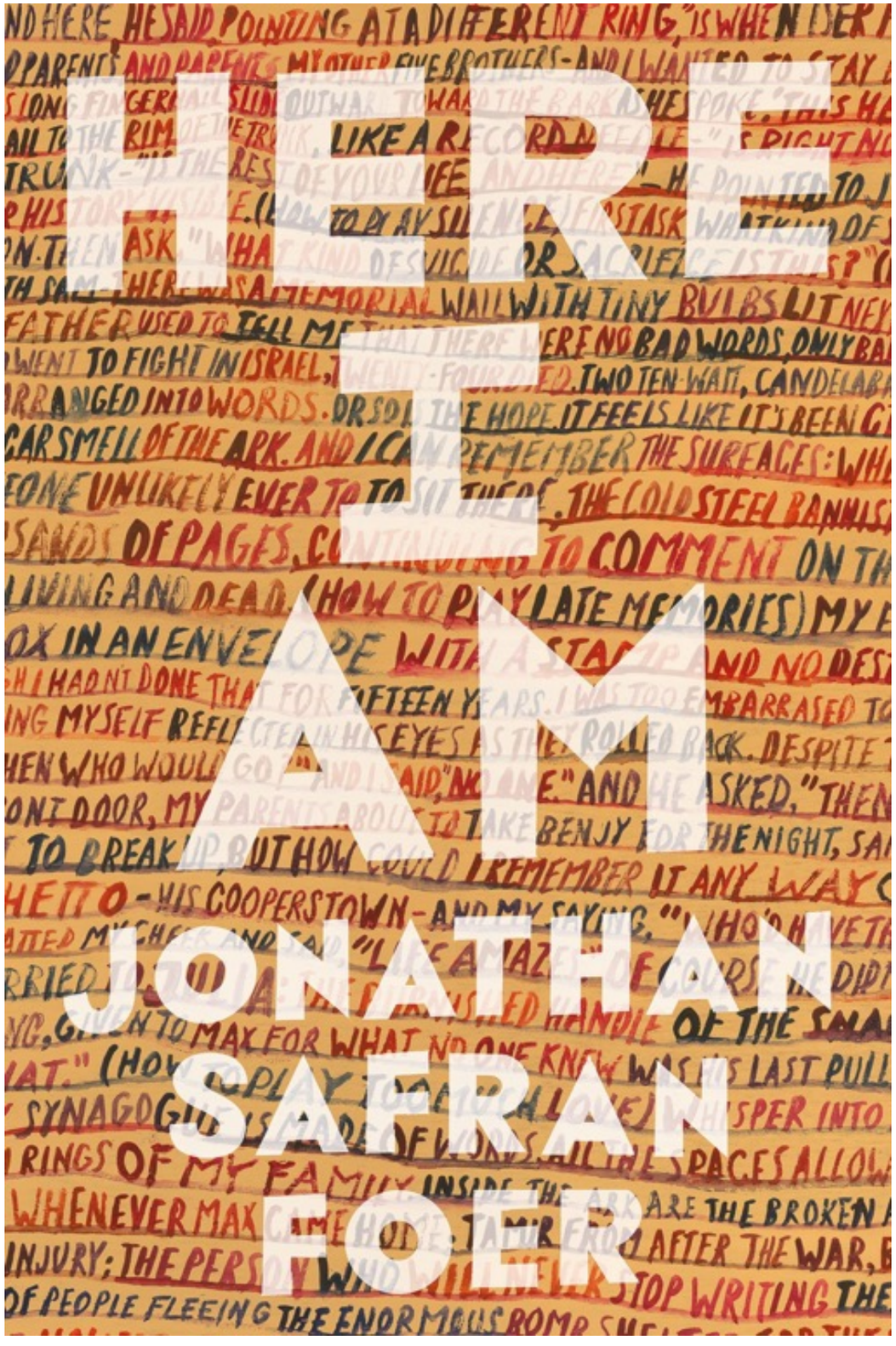
genre: religion and spirituality

strong female lead

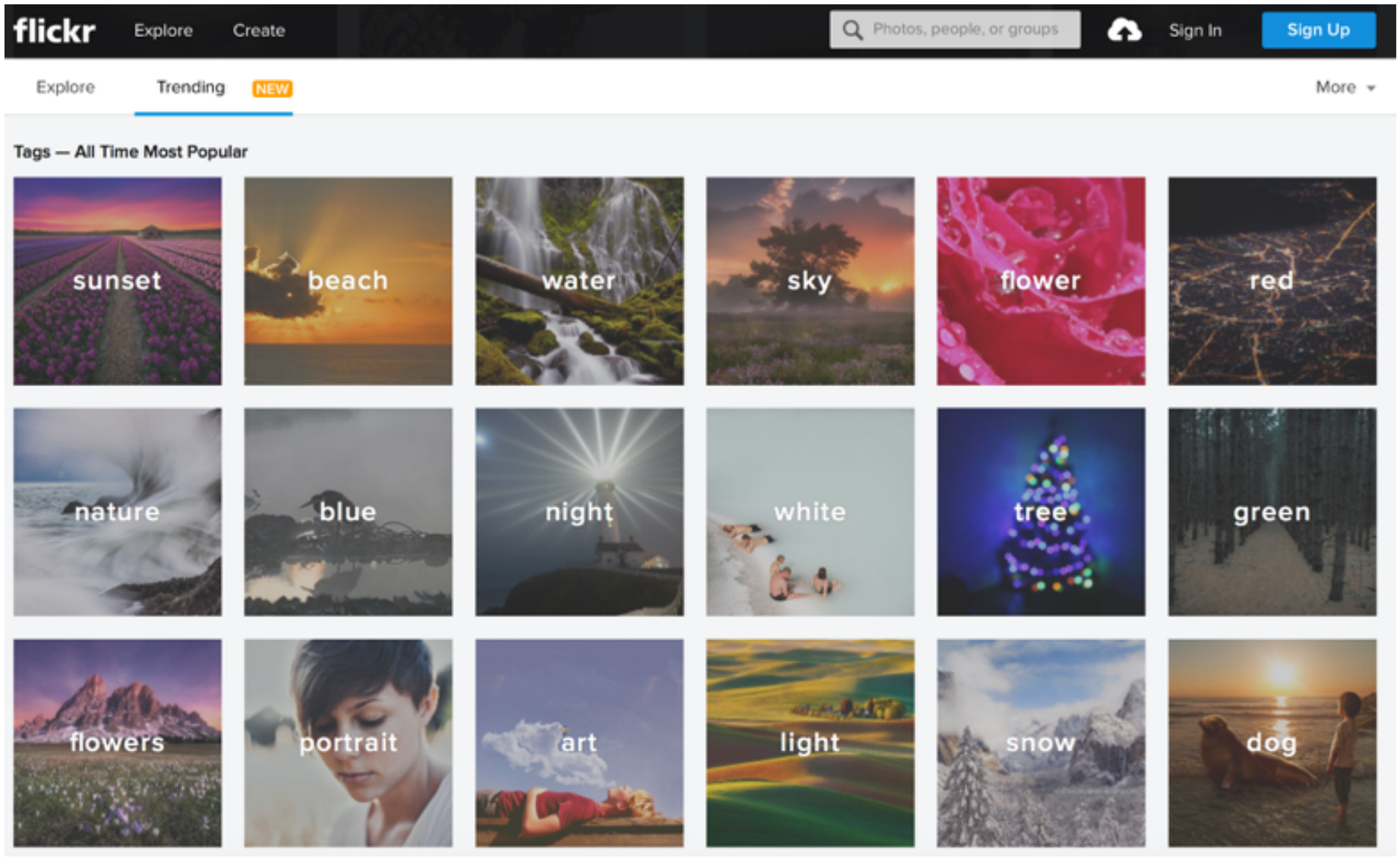
strong male lead

happy ending

sad ending



Open vocabularies



Interannotator agreement

annotator A

annotator B

	puppy	fried chicken
puppy	6	3
fried chicken	2	5

observed agreement = 11/16 = 68.75%

https://twitter.com/teenybiscuit/status/705232709220769792/photo/1

Interannotator agreement

But we need to correct for the agreement that would just happen by chance

Two annotators: Cohen's κ
Multiple annotators: Fleiss' κ

$$\text{Cohen's } \kappa = \frac{p_o - p_e}{1 - p_e}$$

annotator A

annotator B	puppy	fried chicken
puppy	93	3
fried chicken	3	1

observed agreement = 94/100 = 94%
expected agreement = 92.32%
κ = .218

Interannotator agreement

Word sense disambiguation: I'm going to the bank

- $bank_1$ = "financial institution"
- $bank_2$ = "sloping mound"
- $bank_3$ = "biological repository"
- $bank_4$ = "building where a $bank_1$ does its business"

1. a. The shop, office, or place of business of a money changer or moneylender.
b. The table or counter of a money changer or moneylender. Chiefly hist.
c. A pawnbroking establishment set up to provide loans to the poor at low interest;

2. a. An institution that invests money deposited by customers or subscribers, typically pays interest on deposits, and usually offers a range of other financial services
b. With the and capital initial: (in England and Wales) = Bank of England

3. a. A sum of money, an amount. Now rare.
b. spec. A sum of money upon which to draw, esp. a fund for disbursement for a particular purpose. Now somewhat rare.
c. U.S. colloq. Without article: large amounts of money; a fortune. Freq. to make bank.

4. A fund created for commercial purposes from the contributions of many; a joint stock or capital.

5. In games of chance and some board games: an amount or pile of money held centrally, or by a player who plays against all the others, e.g. the proprietor of the gaming table; (also) the person holding the bank in some gambling or board games; the banker.

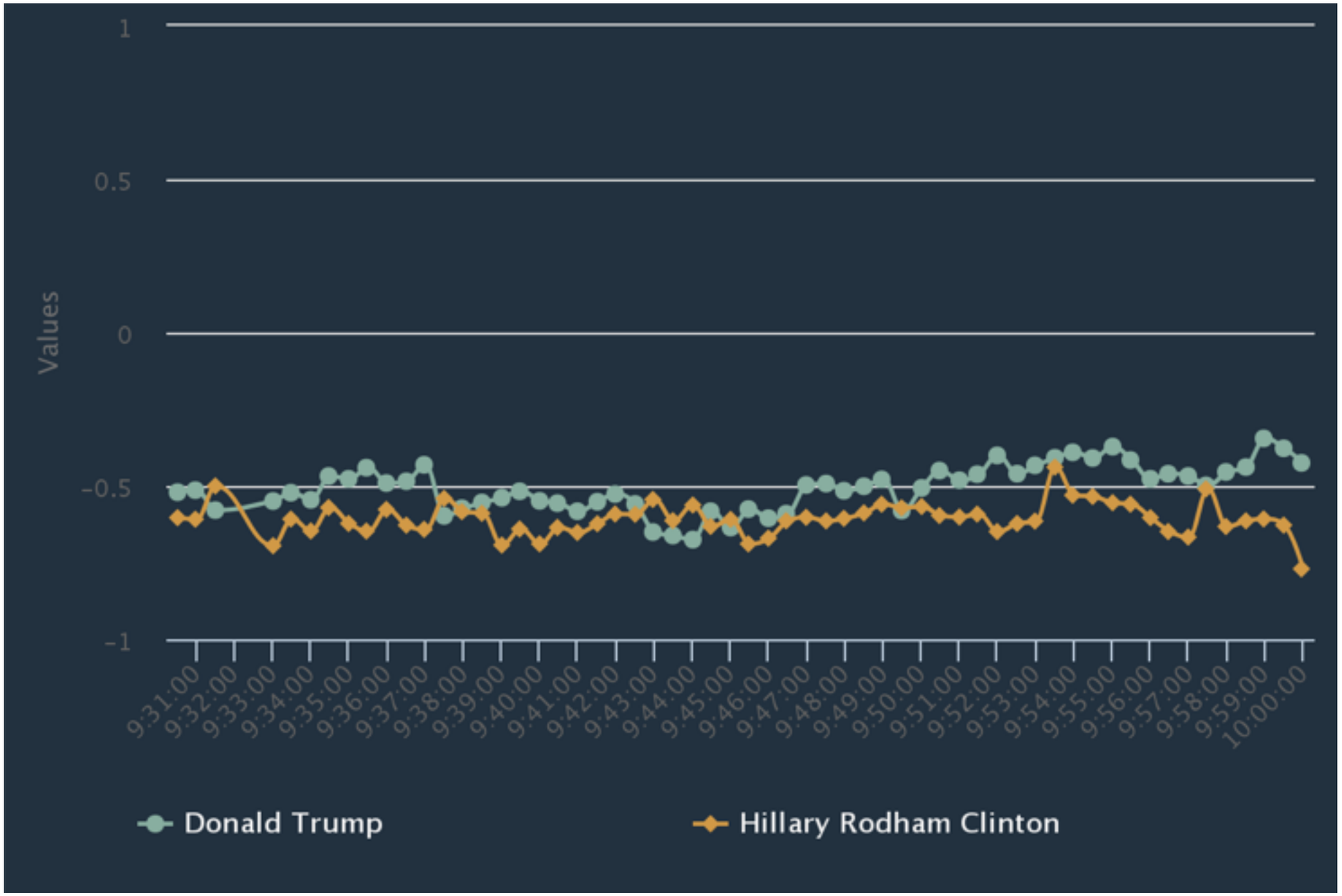
6. a. A stock or repository of something immaterial
b. A pool of people whose skills, services, etc., may be drawn upon when required
c. A stock of something held for use in an emergency or shortage
d. U.S. A laboratory storing blood, cells, tissues, or organs for transfusion, transplantation

7. Computing. A group of units of memory that a computer has access to

8. Brit. A site or receptacle where certain used items may be deposited for recycling.

OED, “bank”

Sentiment analysis



https://techcrunch.com/2016/09/26/cognovi-labs-will-study-the-presidential-debates-tonight-with-their-emotional-sentiment-tool/

Sentiment analysis

Christopher Hayes @chrislhayes · Sep 26

Did the audience just laugh at **Trump** saying "I have a better temperament" ?

1.4K 4.2K

Concurrent validity

- Does a measure correlate with another trusted variable?

Discriminant validity

- Does a measure not correlate with measures of *different* phenomena?

Tradeoffs

- Size of feature space (parameters) vs. interpretability of results

- Overfitting: memorizing the nuances (and noise) of the training data that prevents generalizing to unseen data

Tradeoffs

A	B
the	democrat
he	republican
can	
business	
taxes	
dont	
pig	
dog	
country	
money	
nuclear	
…	

Tradeoffs

- Size of feature space (parameters) vs. interpretability of results

- Overfitting: memorizing the nuances (and noise) of the training data that prevents generalizing to unseen data

Tradeoffs

- Many methods for feature selection (determining which features are important for prediction etc.)

- Resource description importantly specifies which information as algorithm has access to.