Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 7: Data and representation

Feb 7, 2016

“Data Science”

TensorFlow

knowledge

raw data

algorithm

Data

data category	example
behavioral traces	web logs, cell phone activity, tweets
sensor data	astronomical sky survey data
human judgments	sentiment, linguistic annotations
cultural data	books, paintings, music

“Raw” data

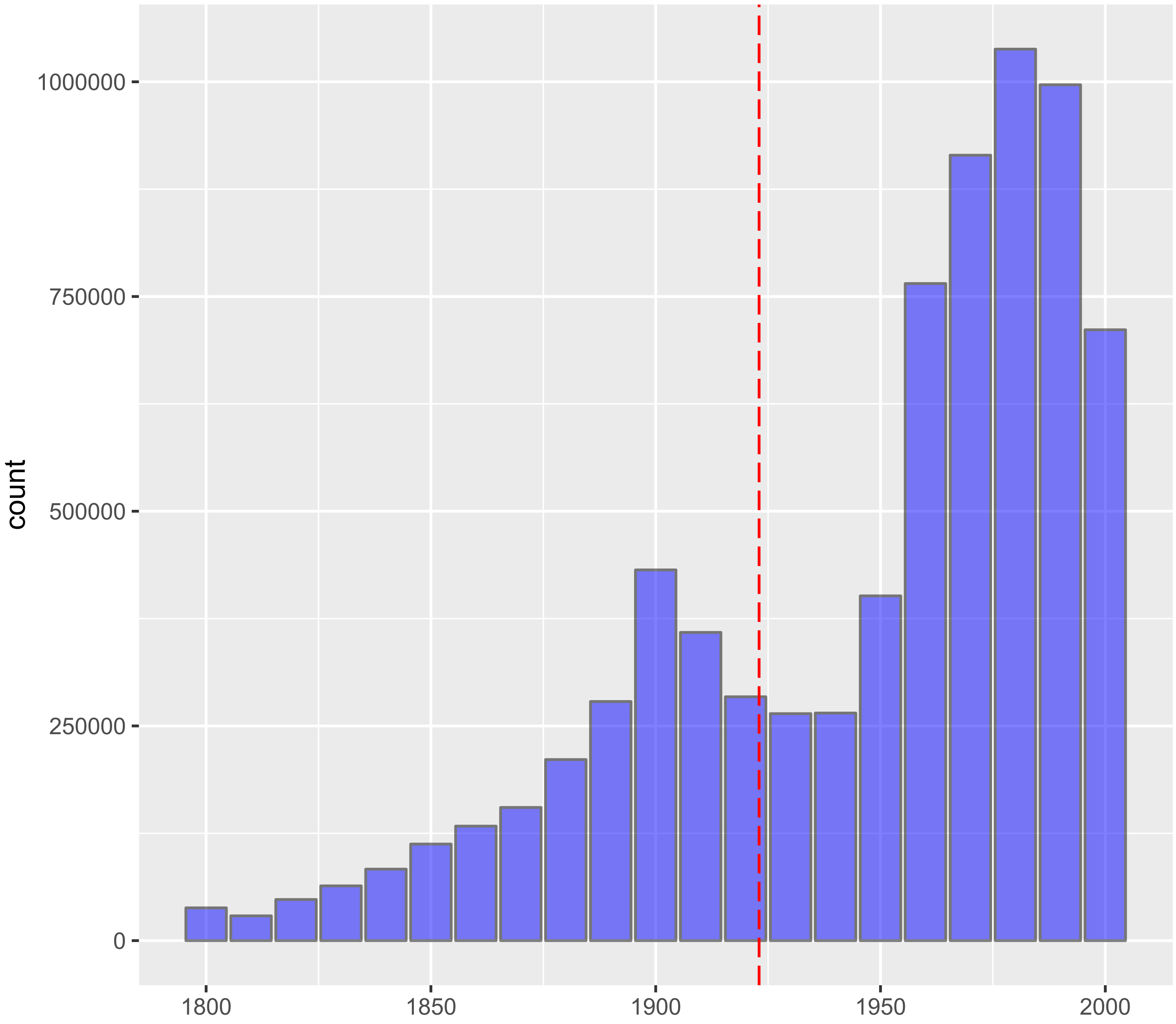
- Gitelman and Jackson (2013)
- Data is not self-evident, neutral or objective
- Data is collected, stored, processed, mined, interpreted; each stage requires our participation.

Provenance

- What is the process by which the data you have got to you?

Data

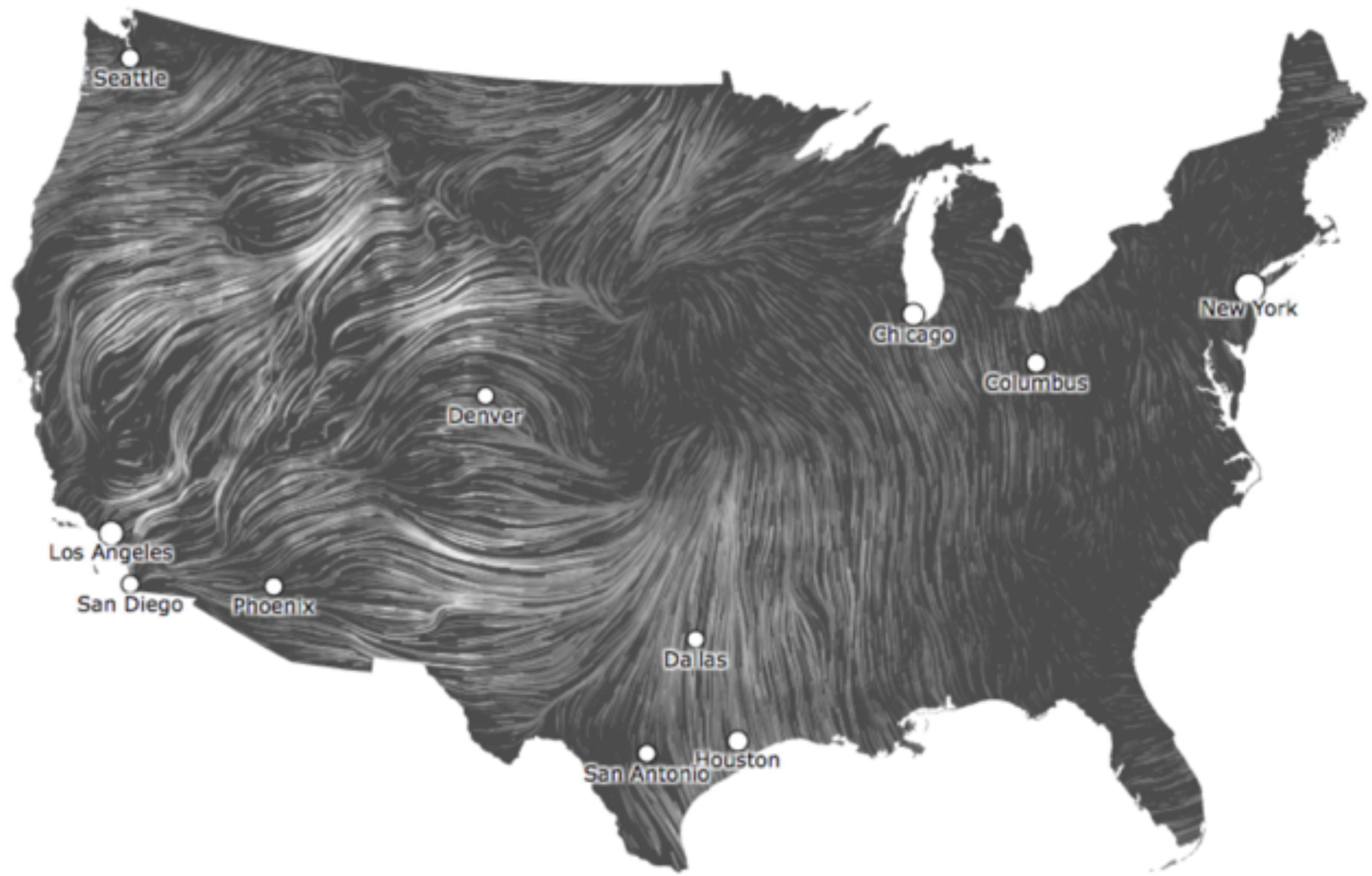
- Cultural analysis from printed books



Michel et al. (2010), "Quantitative Analysis of Culture Using Millions of Digitized Books," Science

Data

- Sensor data



Hill and Minsker (2010), "Anomaly detection in streaming environmental sensor data: A data-driven modeling approach," Environmental Modelling & Software

Edward Steichen, “The Flatiron” (1904)

Data Collection

- Data → Research Question
  - “Opportunistic data”
  - Research questions are shaped by what data you can find
- Research Question → Data
  - Research is driven by questions, find data to support answering it.

Audit trail (traceability)

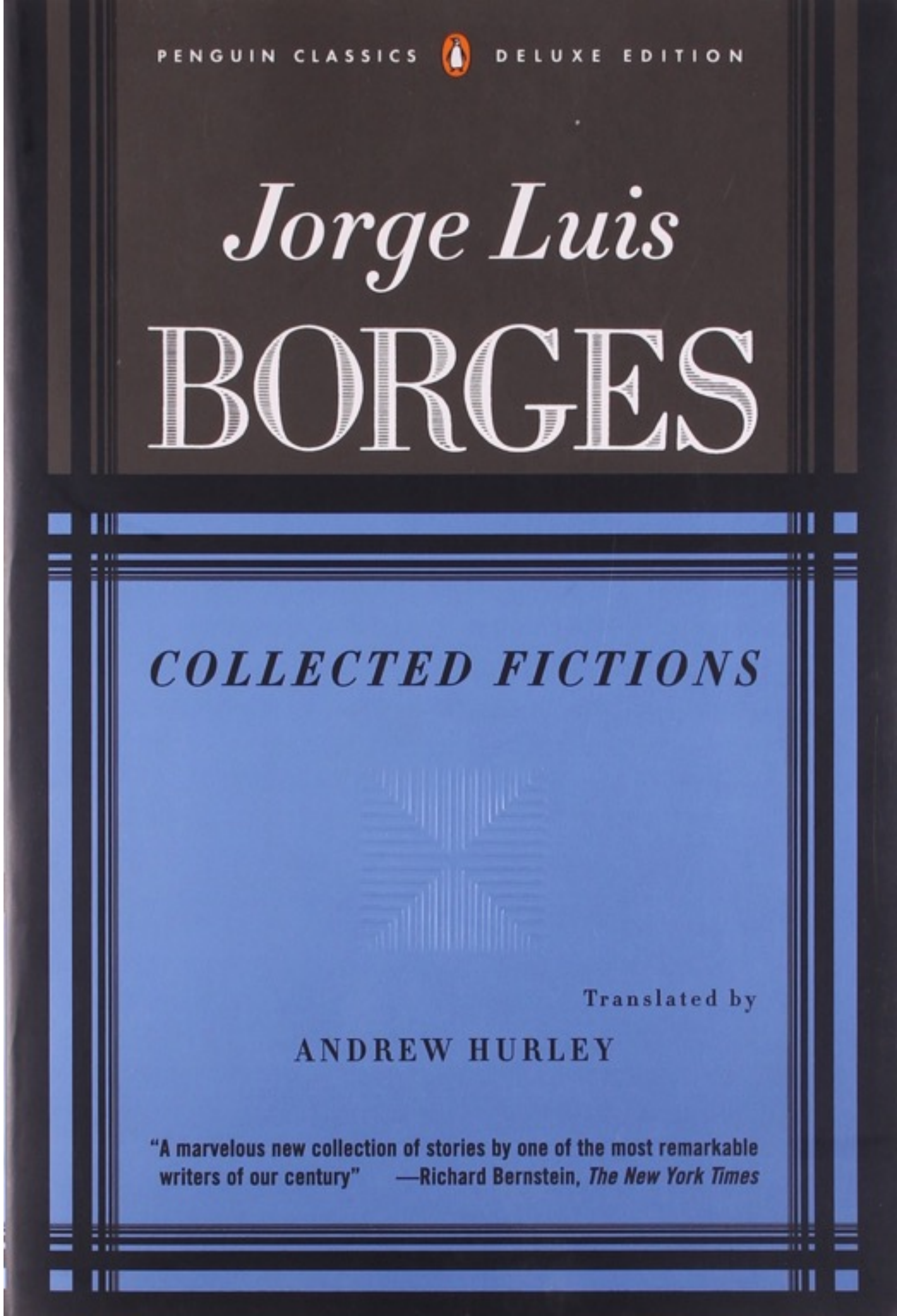
- Preserving the chain of decisions make can improve reproducibility and trust in an analysis.
- Trust extends to the interpretability of algorithms
- Practically: documentation of steps undertaken in an analysis

Data science lifecycle

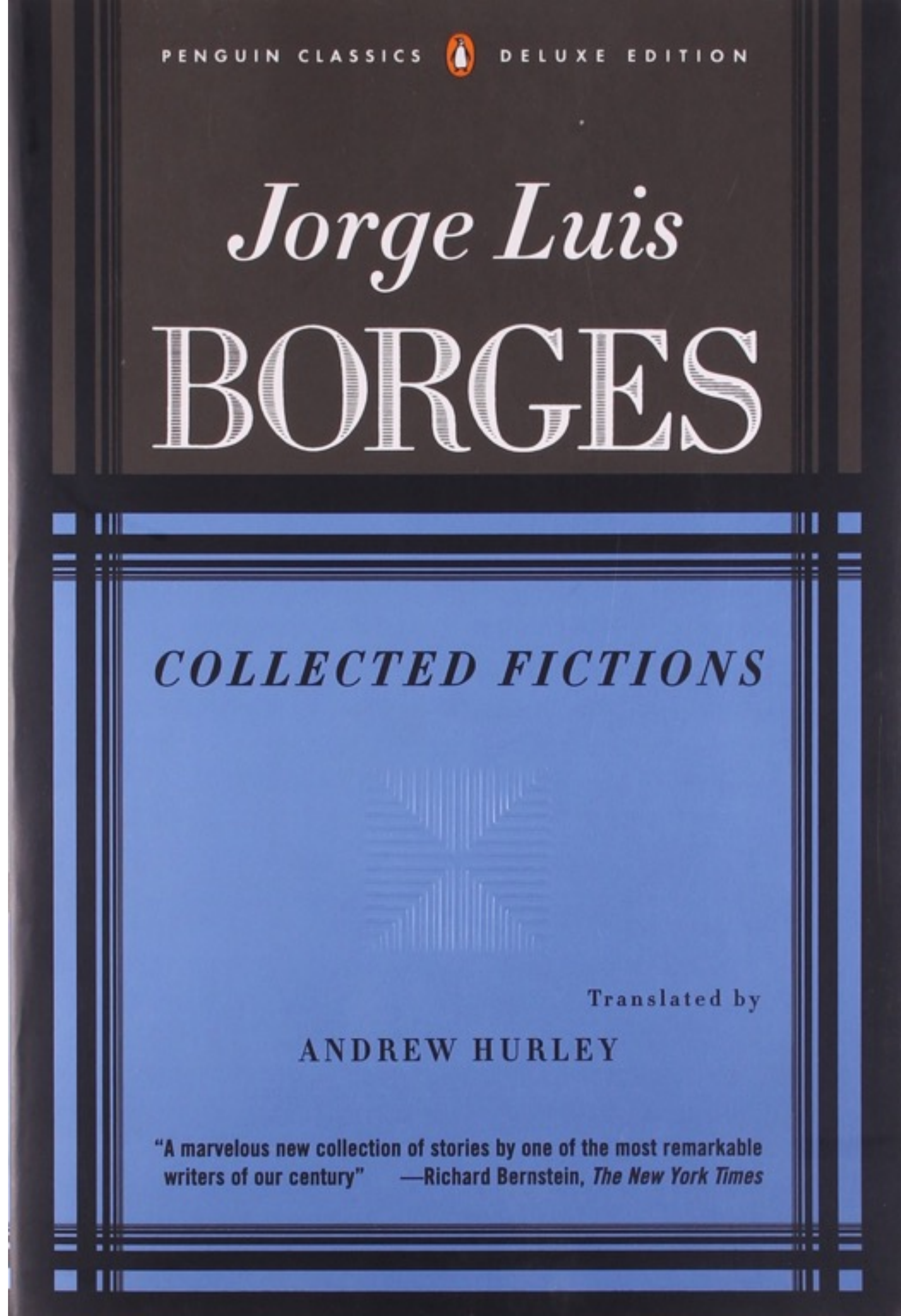
Cross Industry Standard Process for Data Mining (CRISP-DM)

Feature engineering

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



author: borges	TRUE
author: austen	FALSE
pub year	1998
height (inches)	9.2
weight (pounds)	2
contain: the	TRUE
contains: zombies	FALSE
amazon rank @ 1 month	159




PENGUIN CLASSICS DELUXE EDITION
author = borges
“the”
amazon rank
“zombie”
weight
Translated by

predictor

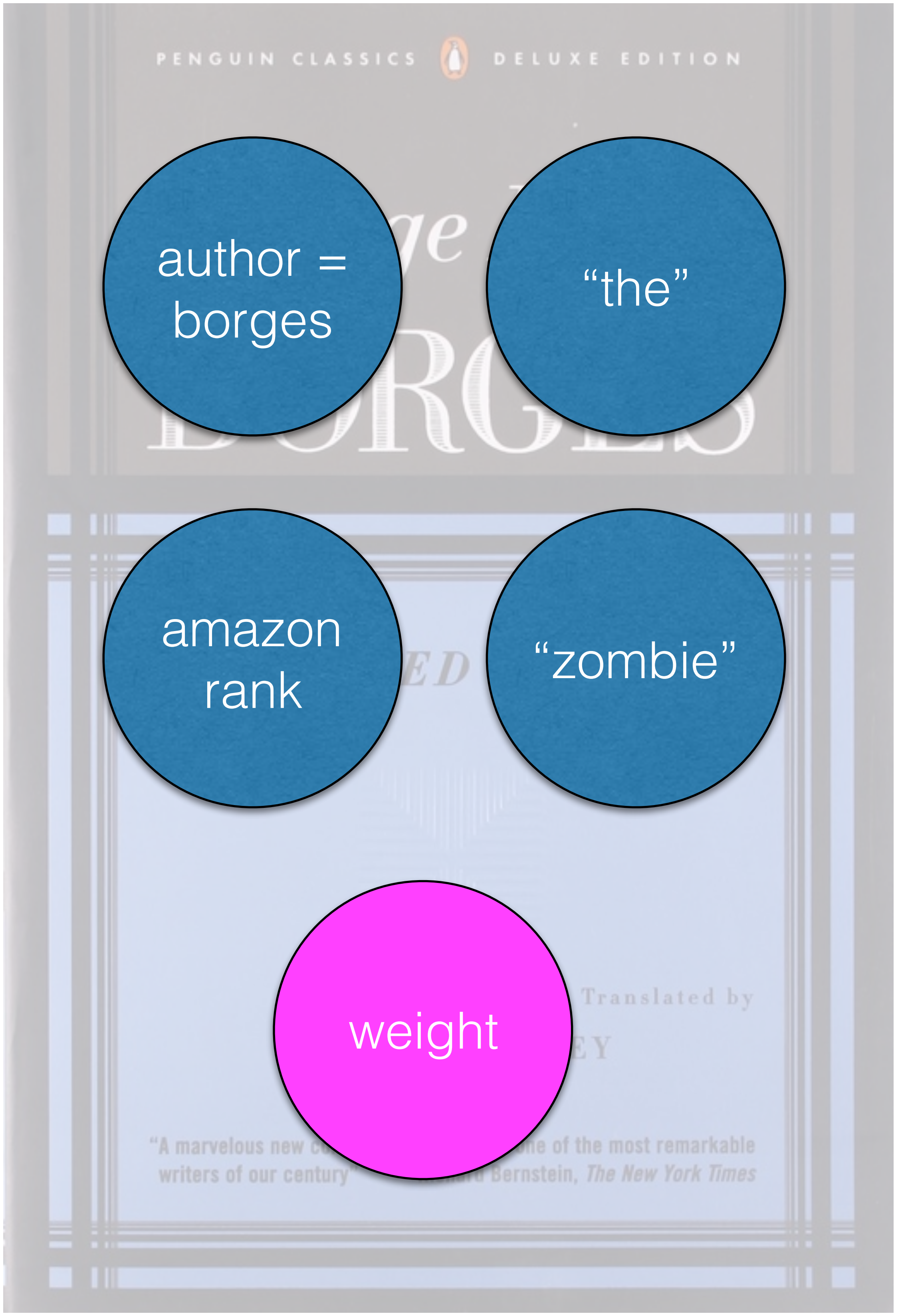
$\Rightarrow$

response


PENGUIN CLASSICS DELUXE EDITION
author = borges
“the”
amazon rank
“zombie”
weight
Translated by

predictor

response



$\Rightarrow$

genre: fiction

genre: world literature

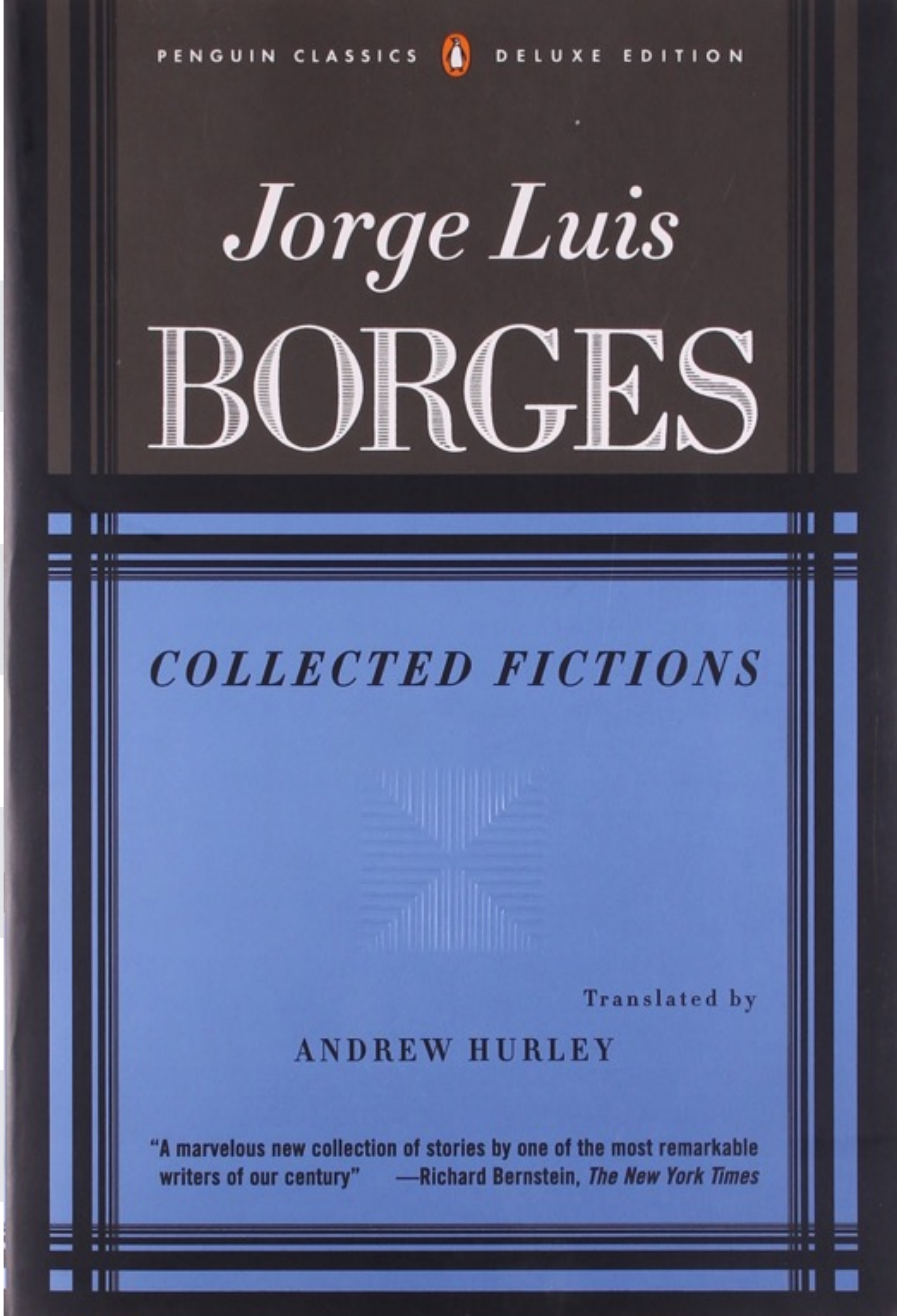
genre: religion and spirituality

strong female lead

strong male lead

happy ending

sad ending



Feature design

- What features to include? What's their scope?
- How do we operationalize them? What values are we encoding in that operationalization?
- What's their level of measurement?

Design choices

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

- 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

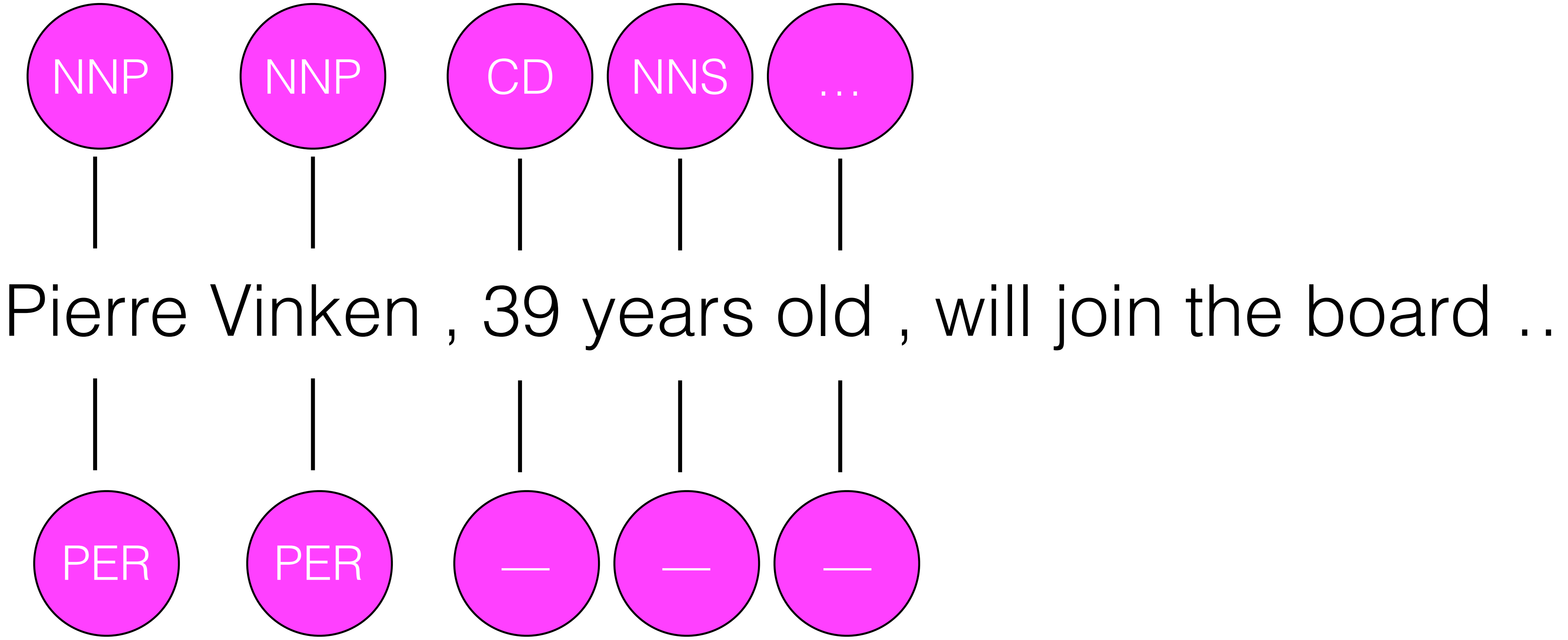
Facebook gender options

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


NNP
NNP
CD
NNS
...
Pierre Vinken , 39 years old , will join the board …
PER
PER
—
—
—

Scope


TWEETS
542
FOLLOWING
455
FOLLOWERS
990
LIKES
162
LISTS
2
Tweets
Tweets & replies
Media
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
David Bamman @dbamman · Sep 23
Rounding out a quick NY trip for @NYUDataScience with a talk here today
The New York Times

Scope

TWEETS
542
FOLLOWING
455
FOLLOWERS
990
LIKES
162
LISTS
2
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
14 Photos and videos
Tweets
Tweets & replies
Media
David Bamman @dbamman · Sep 23
Rounding out a quick NY trip for @NYUDataScience with a talk here today
The New York Times

Berkeley
SCHOOL OF INFORMATION
TWEETS
2,395
FOLLOWING
481
FOLLOWERS
3,729
LIKES
1,210
LISTS
6
UC Berkeley I School
@BerkeleyISchool FOLLOWS YOU
The UC Berkeley School of Information is a multi-disciplinary program devoted to enhancing the accessibility, usability, credibility & security of information.
Berkeley, California, USA
Tweets
Tweets & replies
Media
UC Berkeley I School Retweeted
Ljuba Miljkovic @ljuba · Sep 18
Looking back fondly at my grad school application, getting there…

Levels of measurement

- Binary indicators
- Counts
- Frequencies
- Ordinal

Binary

- $x \in \{0,1\}$

task	feature	value
text categorization	word	presence/absence

Continuous

- x is a real-valued number ($x \in \mathbb{R}$)

task	feature	value
text categorization	word	frequency
authorship attribution	date	year

Ordinal

- x is a categorical value, where members have ranked order ($x \in \{★, ★★, ★★★\}$), but the values are not inherently meaningful

- House numbers
- Likert scale responses

Categorical

- x takes one value out of several possibilities (e.g., $x \in$ {the, of, dog, cat})

task	feature	value
text categorization	token	word identity
political prediction	location	state identity

Features in models

- Not all models can accommodate features equally well.

	continuous	ordinal	categorical	binary
perceptron				
decision trees				
naive Bayes				

Transformations

Binarization

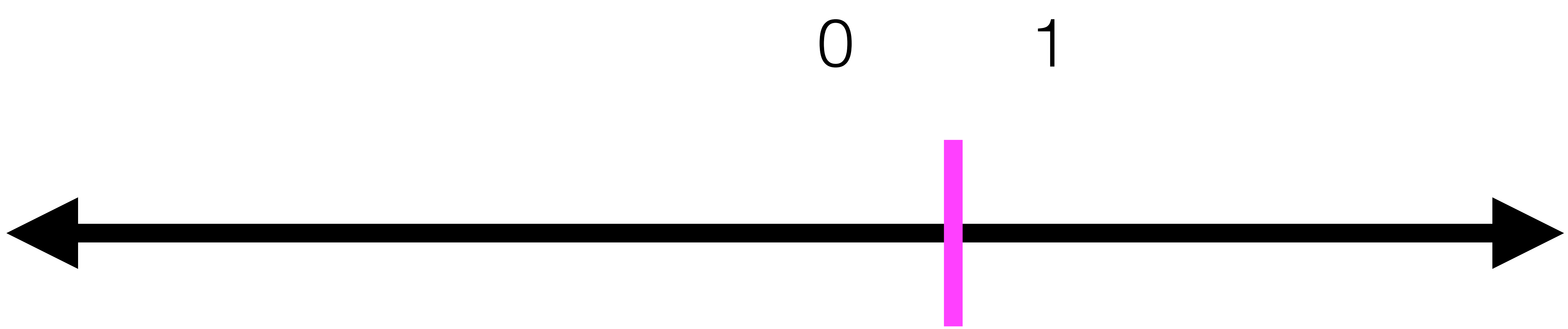
- Transforming a categorical variable of K categories into K separate binary features

Location: “Berkeley”

Berkeley	0
Oakland	1
San Francisco	0
Richmond	0
Albany	0

Thresholding

- Transforming a continuous variable into a single binary value



Decision trees

Algorithm 5.1: GrowTree(D, F) – grow a feature tree from training data.

Input : data D; set of features F.

Output : feature tree T with labelled leaves.

if Homogeneous(D) **then return** Label(D) ; // Homogeneous, Label: see text

$S \leftarrow$ BestSplit(D, F) ; // e.g., BestSplit-Class (Algorithm 5.2)

split D into subsets D_i according to the literals in S;

for each i **do**

  if $D_i \neq \emptyset$ **then** $T_i \leftarrow$ GrowTree(D_i, F) **else** T_i is a leaf labelled with Label(D);

end

return a tree whose root is labelled with S and whose children are T_i

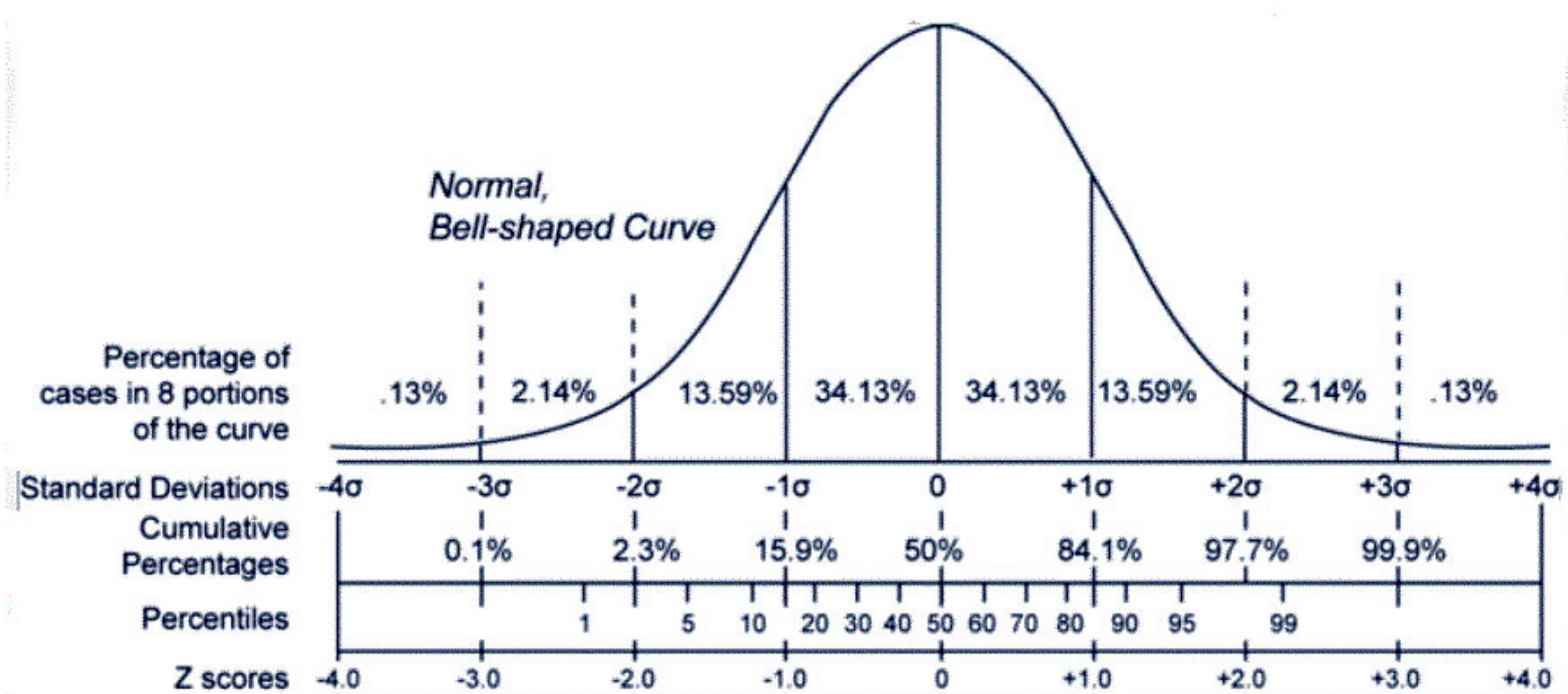
BestSplit identifies the feature with the highest information gain and partitions the data according to values for that feature

Decision trees

- Categorical/binary features: one child for each value
- Quantitative/ordinal features: binary split, with a single value as the midpoint.
    - Trees ignore the scale of a quantitative feature (monotonic transformations yield same ordering)

Discretizing/Bucketing

- Transforming a continuous variable into a set of buckets
- Equal-sized buckets = quantiles



Feature selection

- Many models have mechanisms built in for selecting which features to include in the model and which to eliminate (e.g., ℓ_1 regularization)
- Mutual information; Chi-squared test

Conditional entropy

- Measures your level of surprise about some phenomenon Y if you have information about another phenomenon X

- Y = word, X = preceding bigram ("the oakland ___")
- Y = label (democrat, republican), X = feature (lives in Berkeley)

Mutual information

- aka "Information gain": the reduction in entropy in Y as a result of knowing information about X

$$H(Y) - H(Y \mid X)$$

$$H(Y) = -\sum_{y \in \mathcal{Y}} p(y) \log p(y)$$

$$H(Y \mid X) = -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y \mid x) \log p(y \mid x)$$

	1	2	3	4	5	6
x_1	0	1	1	0	0	1
x_2	0	0	0	1	1	1
y	⊕	⊖	⊖	⊕	⊕	⊖

Which of these features gives you more information about y?

Feature	H(Y | X)
follow clinton	0.91
follow trump	0.77
"benghazi"	0.45
negative sentiment + "benghazi"	0.33
"illegal immigrants"	0
"republican" in profile	0.31
"democrat" in profile	0.67
self-reported location = Berkeley	0.80

$$MI = IG = H(Y) - H(Y \mid X)$$

H(Y) is the same for all features, so we can ignore it when deciding among them

χ^2

Tests the independence of two categorical events

x, the value of the feature
y, the value of the label

$$\chi^2 = \sum_x \sum_y \frac{(\text{observed}_{xy} - \text{expected}_{xy})^2}{\text{expected}_{xy}}$$

X^2

$$\chi^2 = \sum_x \sum_y \frac{(\text{observed}_{xy} - \text{expected}_{xy})^2}{\text{expected}_{xy}}$$

	A	B	Y
0	10	0	
1	0	5	

X

χ^2

	A	B	sum
0	10	0	10
1	0	5	5
sum	10	5	

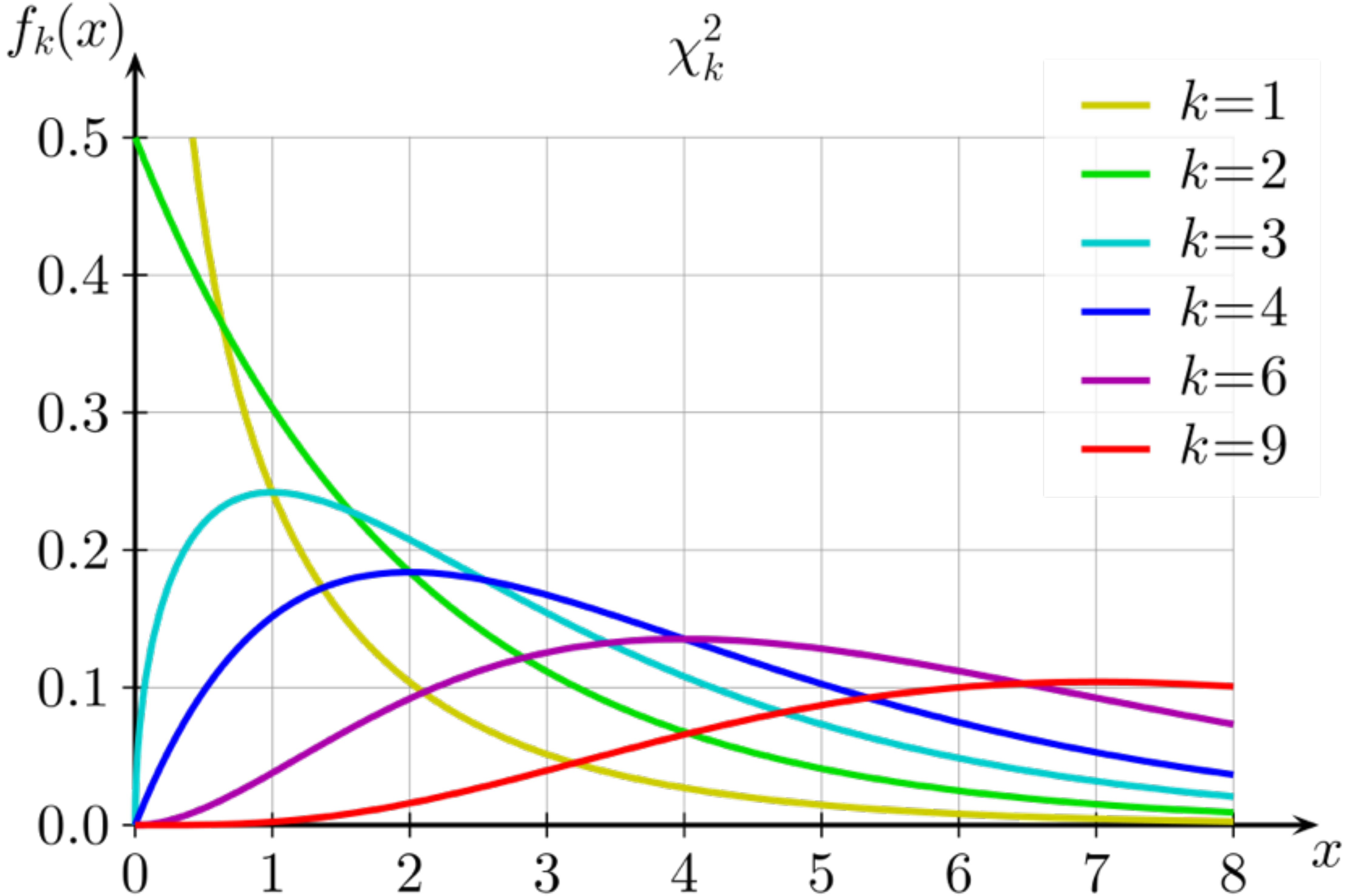
	A	B	marg. prob
0	10	0	0.66
1	0	5	0.33
marg prob	0.66	0.33	

X^2

	A	B	marg. prob
0	10	0	0.66
1	0	5	0.33
marg prob	0.66	0.33	

	A	B	sum
0	6.534	3.267	10
1	3.267	1.6335	5
sum	10	5	

Expected counts



$f_k(x)$
χ_k^2
$k=1$
$k=2$
$k=3$
$k=4$
$k=6$
$k=9$
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
7
8
x

Normalization

- For some models, problems can arise when different features have values on radically different scales

- Normalization converts them all to the same scale

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

Normalization

$$z = \frac{x - \mu}{\sigma}$$

- Normalization destroys sparsity (sparsity is usually desirable for computational efficiency)

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

TF-IDF

- Term frequency-inverse document frequency
- A scaling to represent a feature as function of how frequently it appears in a data point but accounting for its frequency in the overall collection
- IDF for a given term = the number of documents in collection / number of documents that contain term

TF-IDF

- Term frequency ($tf_{t,d}$) = the number of times term t occurs in document d

- Inverse document frequency = inverse fraction of number of documents containing (D_t) among total number of documents N

$$tfidf(t,d) = tf_{t,d} \times \log \frac{N}{D_t}$$

Latent features

- Explicitly articulated features provide the most control + interpretability, but we can also supplement them with *latent* features derived from the ones we observe

- Dimensionality reduction techniques (PCA/SVD) [Mar 9]

- Unsupervised latent variable models [Feb 23]

- Representation learning [Mar 14]

Brown clusters

Brown clusters trained from Twitter data: every word is mapped to a single (hierarchical) cluster

^001010110 (29)	never neva nvr gladly nevr #never neverr nver neverrr nevaa nevah nva neverrrr letchu letcha ne'er -never neveer glady #inever bever nevaaa neever nerver enver neeever nevet neeeever nevva
^001010111 (23)	ever eva evar evr everrr everr everrrr evah everrrrr everrrrrr evaa evaaa everrrrrrr nevar eveer evaaaa eveeer everrrrrrrr everrrrrrrrr evea eveeeer evaaaaa evur
^00101100 (16)	only onli onlyy ony onlii 0nly -only olny onlyyy onlt onlly onyl onlu onlee onle inly

http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html

Brown clusters

^001010110 (29)	never neva nvr gladly nevr #never neverr nver neverrr nevaa nevah nva neverrrr letchu letcha ne'er -never neveer glady #inever bever nevaaa neever nerver enver neeever nevet neeeever nevva
^001010111 (23)	ever eva evar evr everrr everr everrrr evah everrrrr everrrrrr evaa evaaa everrrrrrr nevar eveer evaaaa eveeer everrrrrrrr everrrrrrrrr evea eveeeer evaaaaa evur
^00101100 (16)	only onli onlyy ony onlii 0nly -only olny onlyyy onlt onlly onyl onlu onlee onle inly

author: foer	1
pub year	2016
contain: the	1
contains: zombies	0
contains: neva	1
contains: 001010110	1
contains: 001010111	0

Incomplete representations

- Missing at random
- Missing and depends on the missing value (e.g., drug use survey questions)

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

Incomplete representations

- Impute the mean
- Categorical values for being missing
- Predict the missing value from other features

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