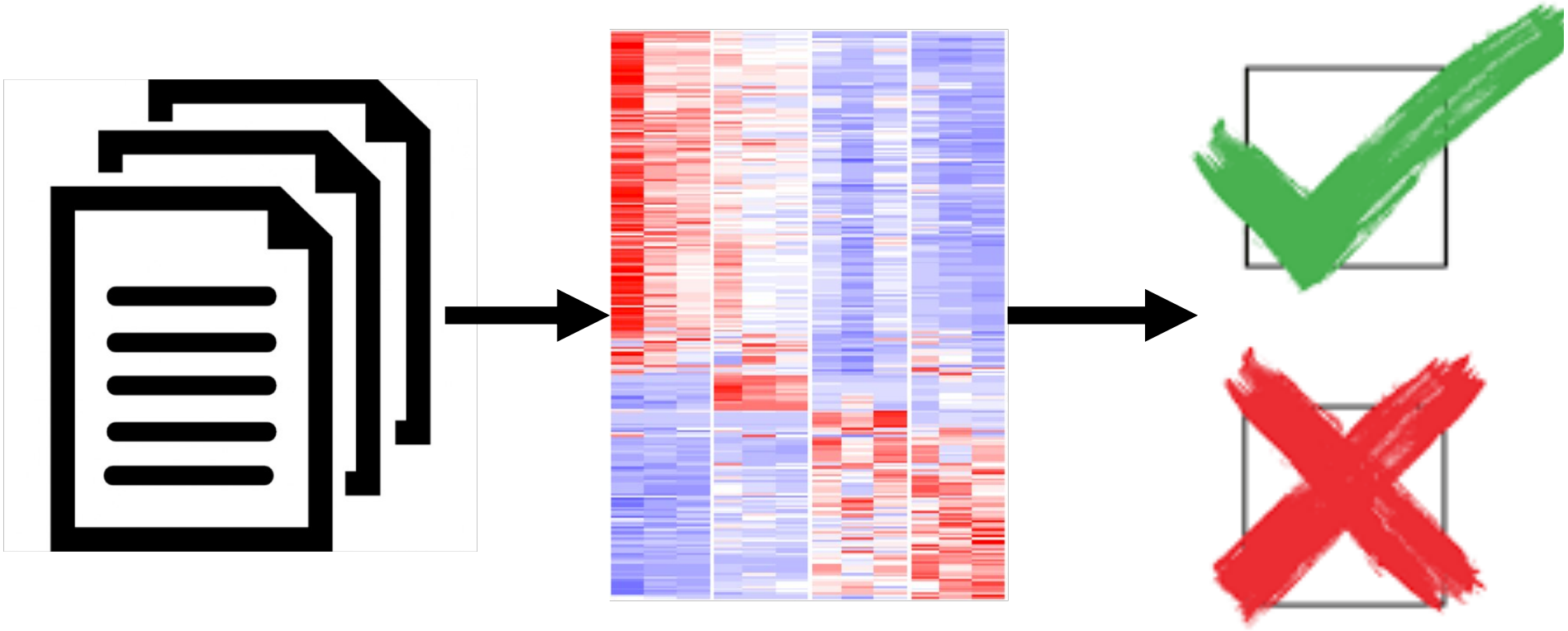
Text mining: an overview and a motivating example

Drew Hendrickson

Department of Cognitive Science & Artificial Intelligence

Extracting value from text



Extracting value from text

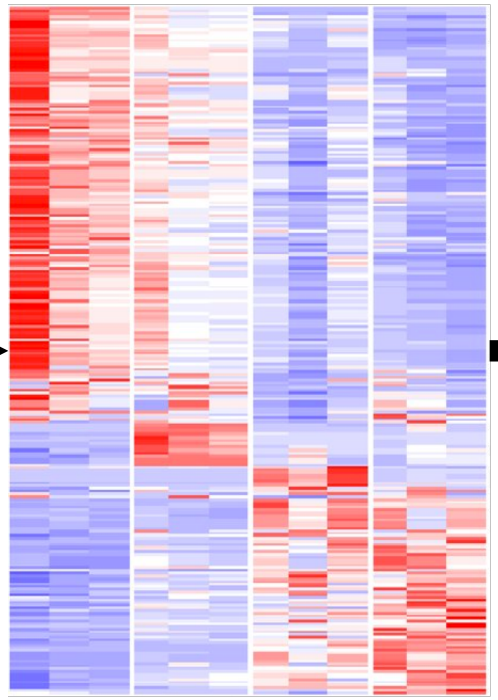
- Contracts or legal documents
- News coverage and opinion pieces
- Transcripts of phone calls or interviews
- Reviews and user comments
- Descriptions

Extracting value from text

- Predict writer's emotional state
- Predict writer's attitude toward topic
- Answer writer's questions*
- Engage in productive dialogue*

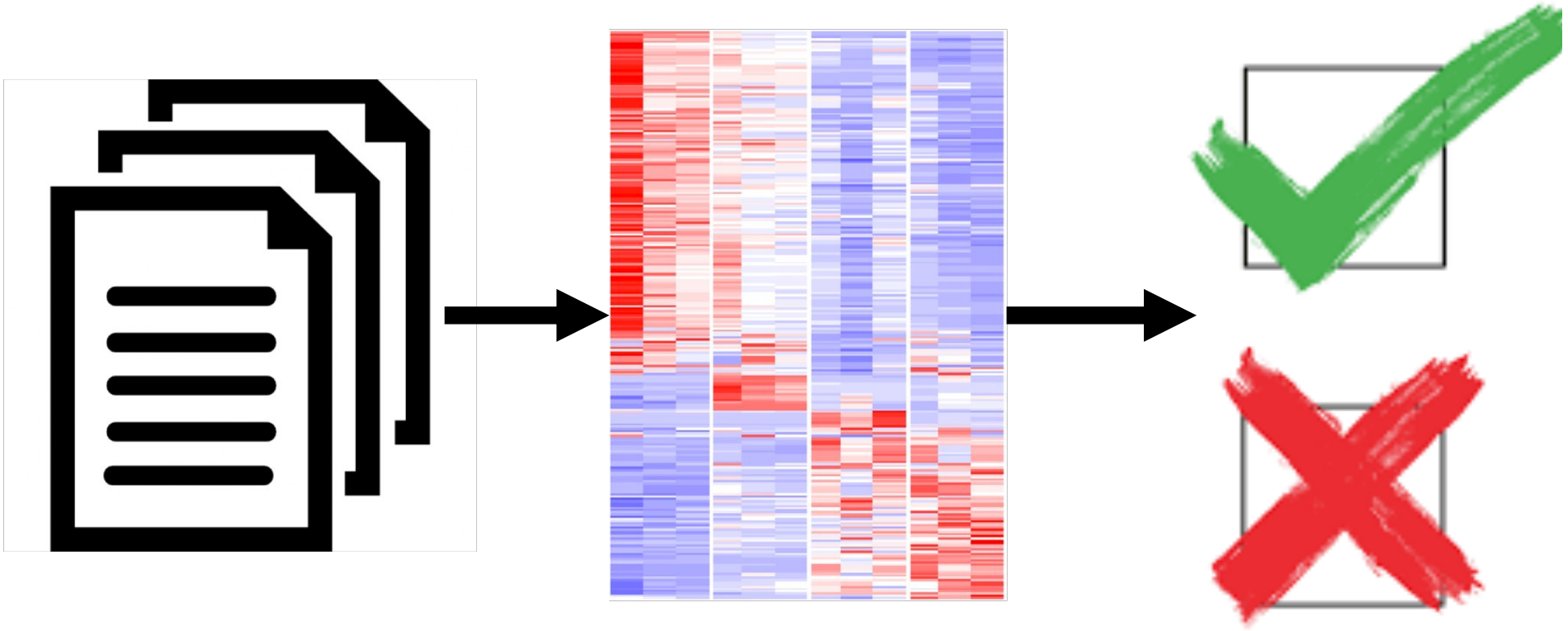
Extracting value from text

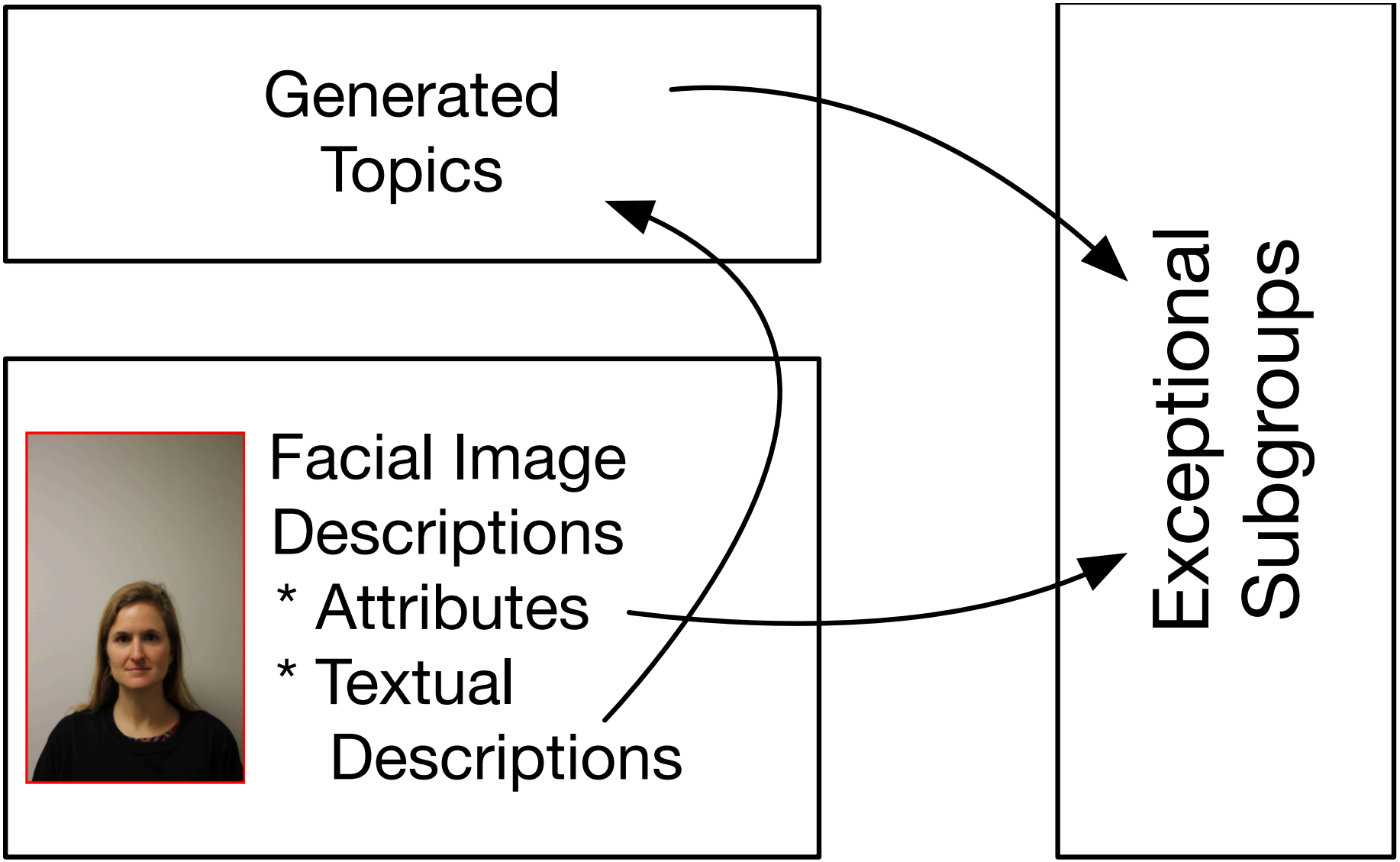
- Training deep neural networks
  - RNNs and LSTMs
- Pre-trained word embeddings
  - Word2Vec
  - FastText
  - GloVe
  - ELMo



- Syntax parsing
  - Part-of-speech tagging
  - Context free grammars
- Collection of words
  - LSA
  - LDA
  - TF-IDF

An example: finding exceptional subgroups




Generated
Topics
Facial Image
Descriptions
* Attributes
* Textual
Descriptions
Exceptional
Subgroups

Description of the Dataset

- 2,491 unique descriptions and ratings of 193 faces by 500 raters
- Self-reported information about each rater:
  - Age, gender, ethnicity, country of origin
- Judgments about the attributes of the image:
  - Age, hair color, eye color, ethnicity, gender, attractiveness, typicality, occupation
- Additionally two written descriptions were provided:
  - A description of the physical attributes (the only description used in this work)
  - A description of non-physical attributes (‘what you can guess about the person‘)

Please answer a few questions about the person in this image:

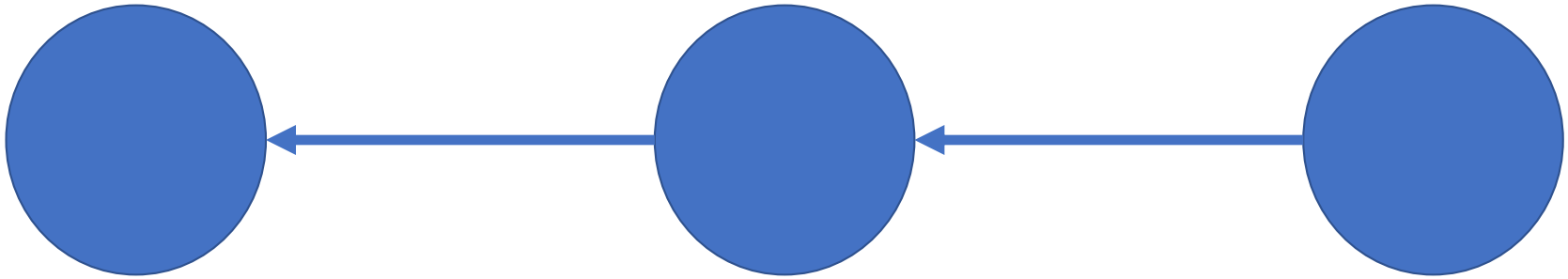
How old do you think they are?
What do you think their hair color is?
What do you think eye color is?
Their ethnicity (select one or more options): African Asian Latino White
Ethnicity details (optional):
What do you think is their gender?
Male Female Unsure
This person is attractive:
Strongly Agree Agree Neither agree or disagree Disagree Strongly Disgree
This person is typical:
Strongly Agree Agree Neither agree or disagree Disagree Strongly Disgree
What do you think their occupation is?
Please write a physical description of this person (excluding their clothes):

Please write a short list of things you can guess about this person:

Submit

Figure 2: Example depicting the description generation task: The figure shows the image on the left, the questions relating to the (nominal) attributes on the top right, and the two types of free text descriptions on the bottom right.

Latent Dirichlet Allocation Model

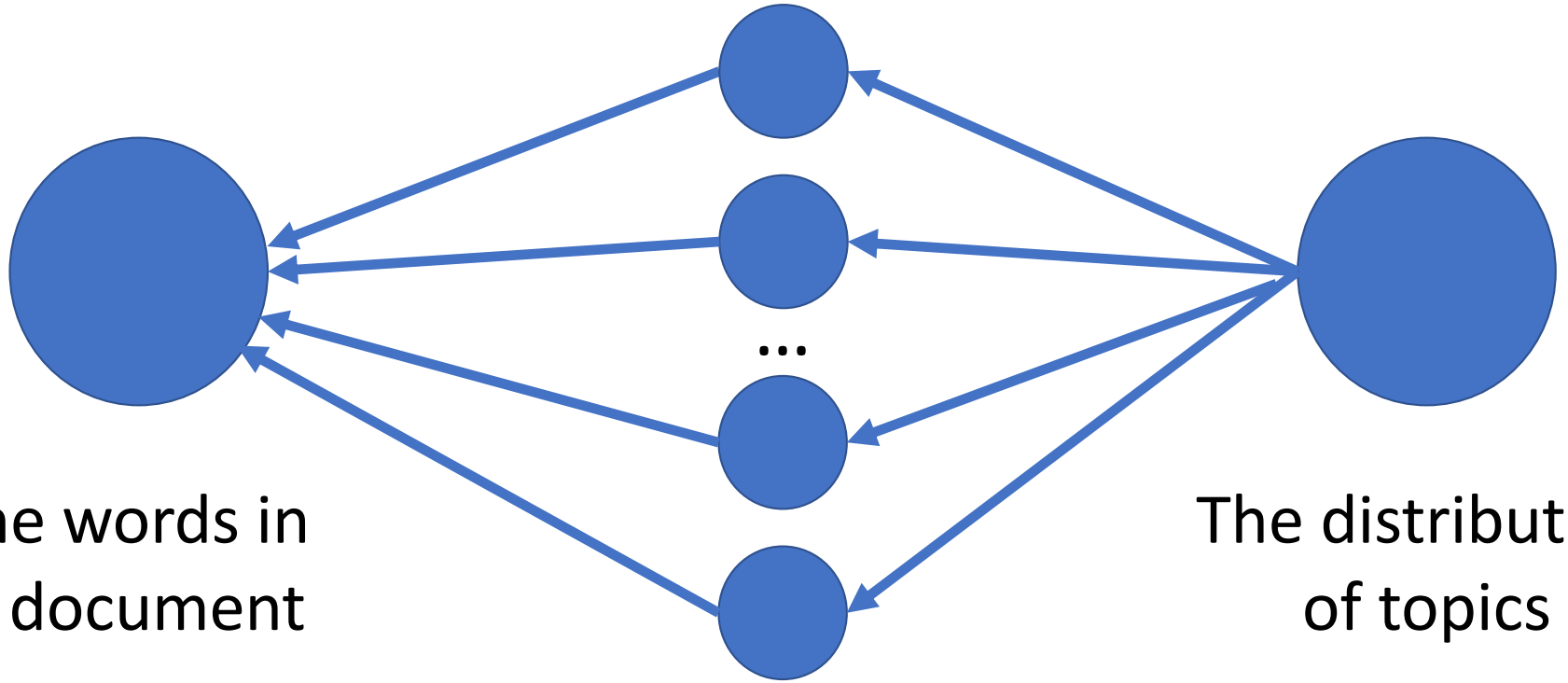


The words in a document

The topic of the document

The distribution of topics

Latent Dirichlet Allocation Model



Topic Modeling

- Modeling topics
  - Latent Dirichlet Allocation (LDA) modeling across the 193 documents consisting of the 193 images of faces
- Grid search on LDA parameters
  - Maximizing difference between documents
    - Sum of the cosine similarity all pairs of documents of a topic
  - Minimize number of topics per document
    - conditional entropy across topic distributions for all documents
- Result: 9 topics
- For 2491 individual descriptions: Probability distribution on topics

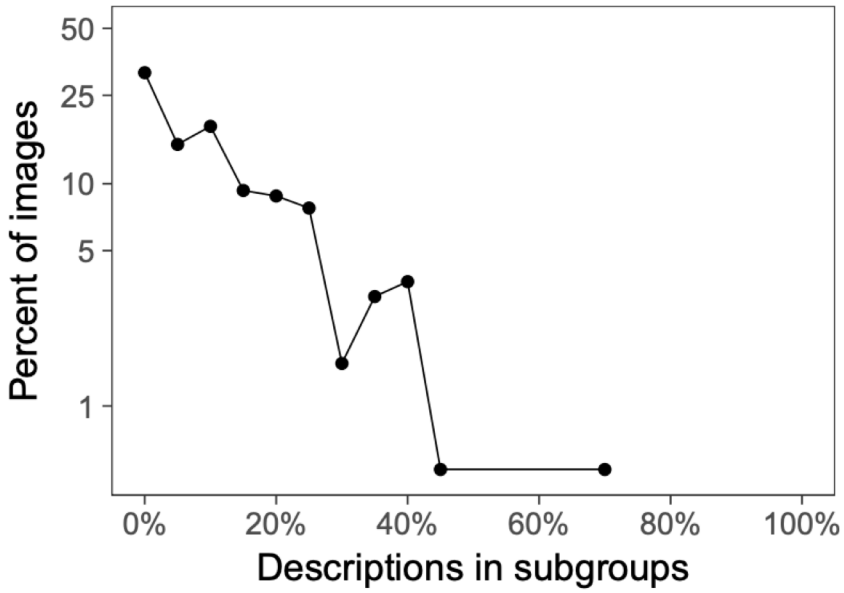
Topic 1	Topic 2	Topic 3
enjoys	good	wise
games	student	tshirt
video	hard	black
play	friends	girl
friends	college	wearing
Topic 4	**Topic 5**	**Topic 6**
sales	authority	hair
figure	dude	eyes
punctual	challenge	long
son	legal	style
advantage	spirited	black
Topic 7	**Topic 8**	**Topic 9**
activities	energetic	warm
outdoor	moments	metal
laugh	notice	heavy
pets	pick	pays
awkward	predictable	hipster

Figure 4: The five words with the highest posterior probability for each of the nine topics.

Exceptional Subgroup Discovery

- Find *descriptions* of subsets in the data, that *differ* significantly for the total population with respect to a *target concept.*
- Example:
  - “50% of all images rated by females aged between 30-40 show a strongly deviating typicality rating compared to the total population.”
- Identifying exceptional subgroups is necessary for plans to differentially process subgroups
- In this application, performed at the level of individual descriptions
- Evaluate the 20 most exceptional subgroups of descriptions

How exceptional are people's descriptions?



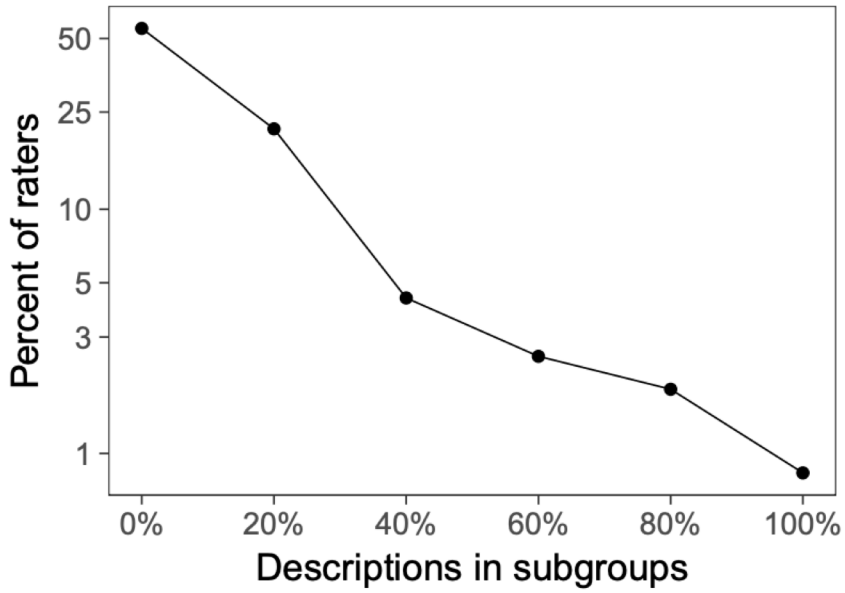


Fig. 1. The proportion of descriptions of specific images (left) and raters (right) that occur in at least one exceptional subgroup. The y-axis of both plots is in log units.

Which features are often exceptional?

R. Country	**R. Gender**	**I. Gender**	**I. Eye Color**	**I. Hair Color**	**I. Ratings**
USA (3) India (2)	Female (3) Male (7)	Female (8) Male (3)	Black (12) Brown (1) Green (1)	Black (6) Blond (1)	Typicality (4) Attract. (5)

Discussion

- Overall: Descriptions that significantly deviate from the population of descriptions are relatively frequent
- Topics and words people use when describing other people may vary widely
- Some findings for this particular application:
  - Male raters and female images are attributes that are likely to define deviating subgroups
  - Black hair and black eyes are the attributes of images most likely to identify exceptional subgroups

Going forward, new questions:

- To what degree does the choice of a representation impact the exceptional subgroups we find?
- Can enriching the word representations with linguistic information improve the process?
- To what degree does knowing exceptional subgroups improve prediction performance in other tasks?

Going forward, new questions:

Please enter your question:

Submit Question | Ready to Finish

Thanks!

Drew Hendrickson
Cognitive Science & Artificial Intelligence
Tilburg University

More information: drewhendrickson.github.io
Accessible via: csai.nl