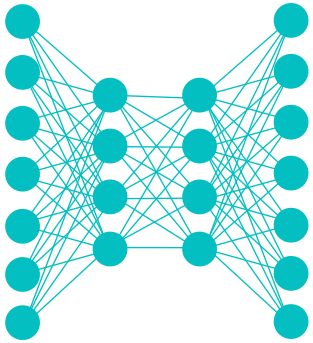
Lecture Notes for

Deep Learning II

De-biasing Strategies



Logistics and Agenda

- Logistics
  - Lab due soon!
- Last Time:
  - Ethical Guidelines
  - Case Studies
- Agenda
  - Finish Case Studies
  - Paper Presentation
  - Word Embedding Review
  - Implicit and Explicit de-biasing

Ethical Principles in ML

From Australian Government, Department of Science

- **Reliability**: does system operate in accordance with intended purpose?
- **Fairness**: will system be inclusive and accessible? Will it involve or result in unfair discrimination against individuals, communities, or groups?

Model Measurement and Objective Alignment

- **Beneficence**: does system benefit individuals, society, or environment?
- **Respect**: does system respect human rights and autonomy of individuals?

Forethought and Insight

- **Privacy**: will system respect and uphold privacy rights and data protection, and ensure the security of data?
- **Transparency**: will system ensure people know when they are engaging with an AI system? Or know if significantly impacted?
- **Contestable**: will there be a timely process to allow people to challenge the use or output of the AI system?

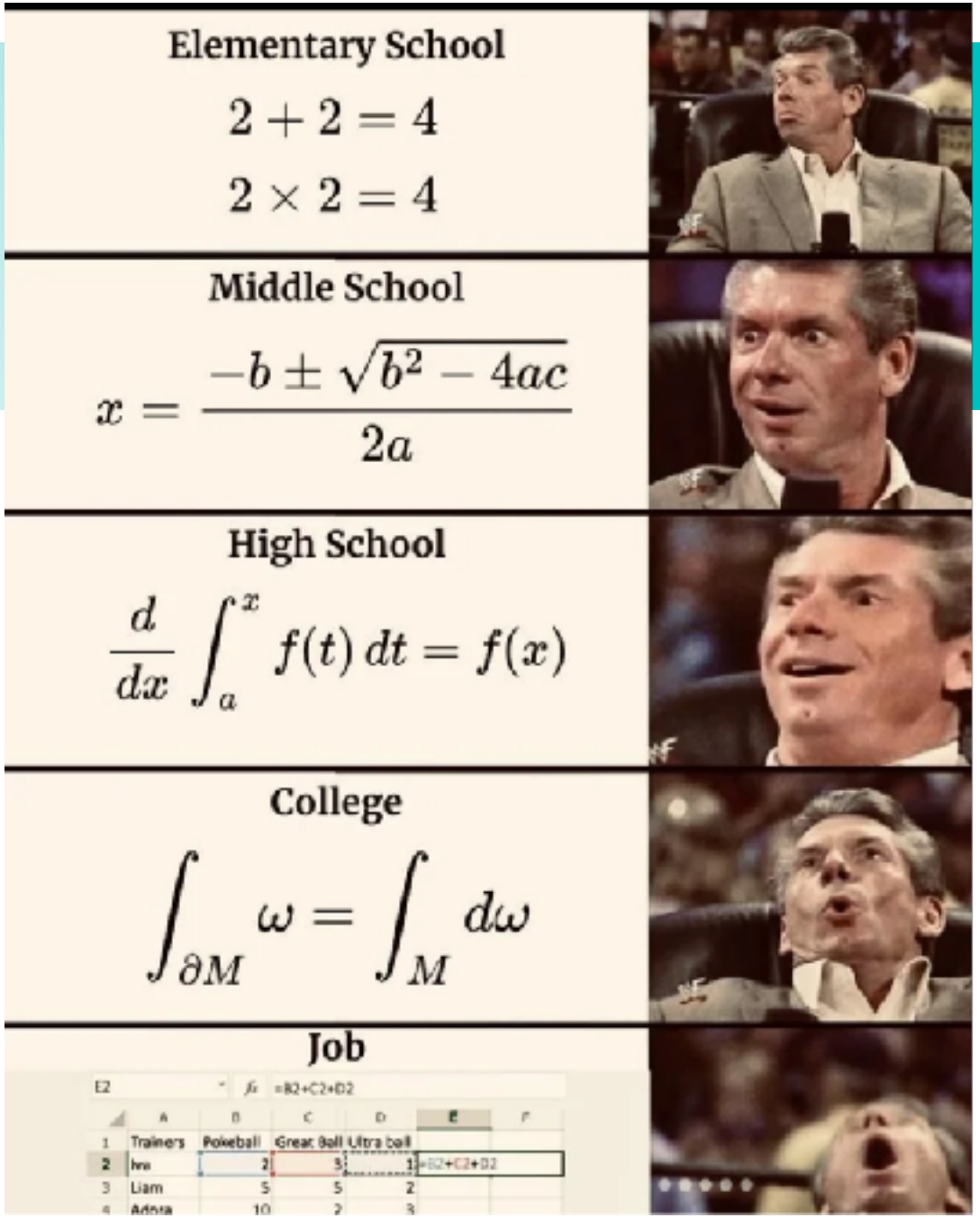
Deployment Design

- **Accountability**: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and *human oversight* of AI systems should be enabled.

Organizational Structure

https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

Case Studies Continued



Ethical Considerations in Military App.

- Ethical guidelines in combat
  - **One Interpretation**: Combat is not ethical because harm of individuals is incentivized
  - **Another**: Certain ethical guidelines can still be considered, accepting that the aim of combat is, in many instances, to create a weapon
- These are both common (maybe valid) interpretations and it is up to you to decide if combat should be included in ethical guidelines
  - **My Opinion**: combat should not be considered ethical, but is unavoidable in the presence of nefarious actors and therefore guidelines need to be in place
  - Many individuals will disagree with me on this and have excellent points, that I do not disagree with

AI Warfare

Class Discussion: Should this use of AI be allowed in Military application?

1. Model measure and objective: Reliability and fairness (does it work equally where needed?)
2. Forethought in design: beneficent and respect (who benefits, autonomy protected?)
3. Deployment: Privacy, transparency, contestability (if wrong, can it be detected and recover properly?)

The US and 30 Other Nations Agree to Set Guardrails for Military AI

The tech-centric war in Ukraine and the success of ChatGPT h
prompted new interest in figuring out how to prevent military A
awry.

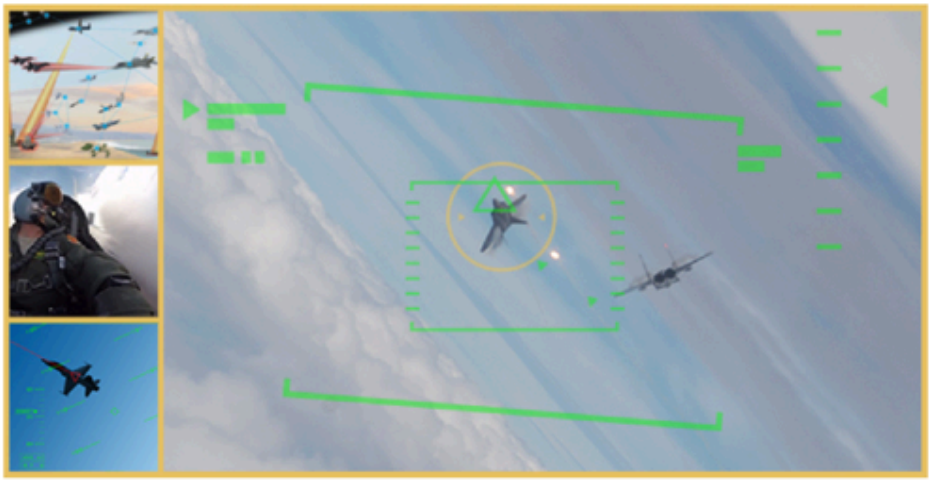
—Lauren Kahn, Senior Researcher at Georgetown University

Defense Advanced Research Projects Agency > News And Events

Training AI to Win a Dogfight

Trusted AI may handle close-range air combat, elevating pilots' role to cockpit-based mission commanders

OUTREACH@DARPA.MIL
5/8/2019



Most Common Use: Planning and Data Collection

Analyzing incoming data and surveillance sources to understand threats, then suggesting capabilities for those scenarios.

Defensive versus Offensive Use

Some autonomous weapons already exist, including defensive systems aboard battleships that can automatically shoot down incoming missiles. But there have only been a couple of reports of potential use of lethal systems that incorporate modern AI in warfare.

Paper Presentation

MITRA: Mixed Synthetic Priors for Enhancing Tabular Foundation Models

Xiyuan Zhang
Amazon

Danielle C. Maddix
Amazon

Junming Yin
Amazon

Nick Erickson
Amazon

Abdul Fatir Ansari
Amazon

Boran Han
Amazon

Shuai Zhang
Amazon

Leman Akoglu
Amazon and CMU

Christos Faloutsos
Amazon and CMU

Michael W. Mahoney
Amazon

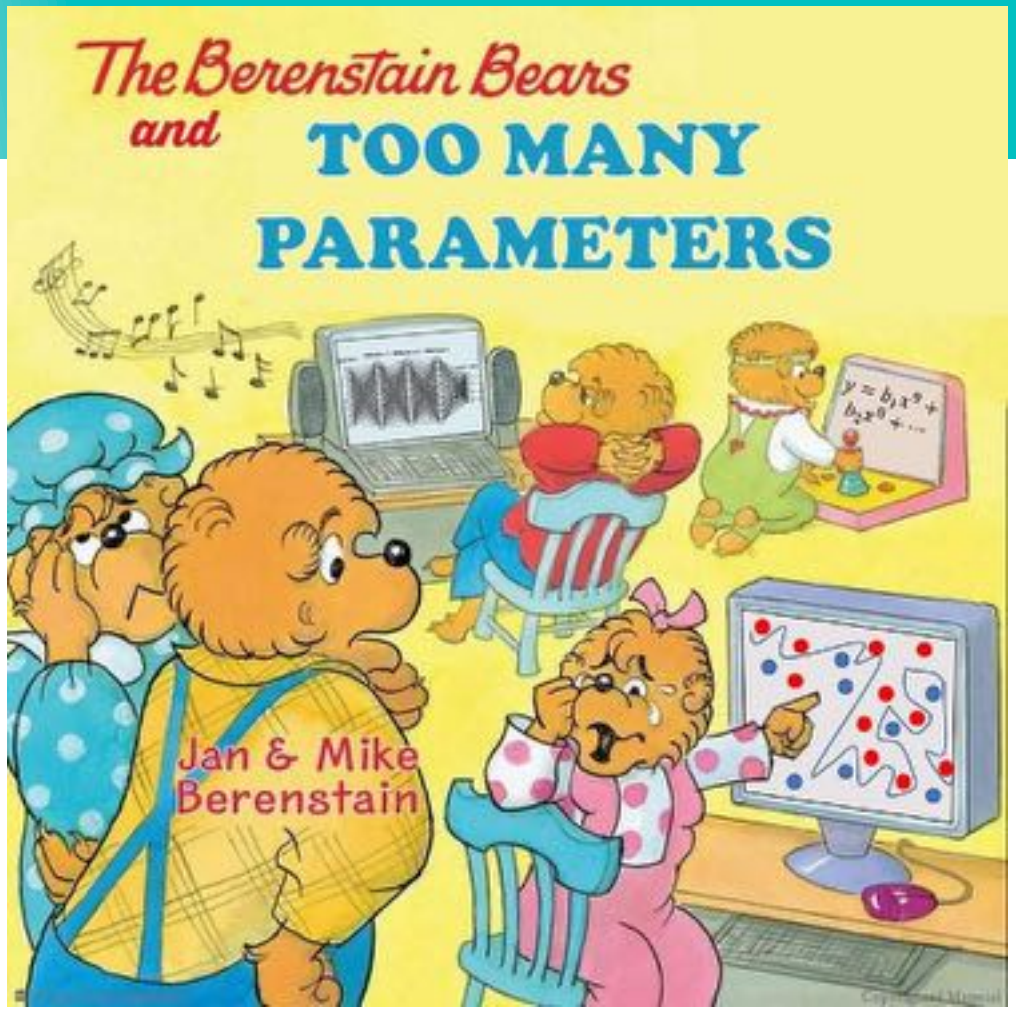
Cuixiong Hu
Amazon

Huzefa Rangwala
Amazon

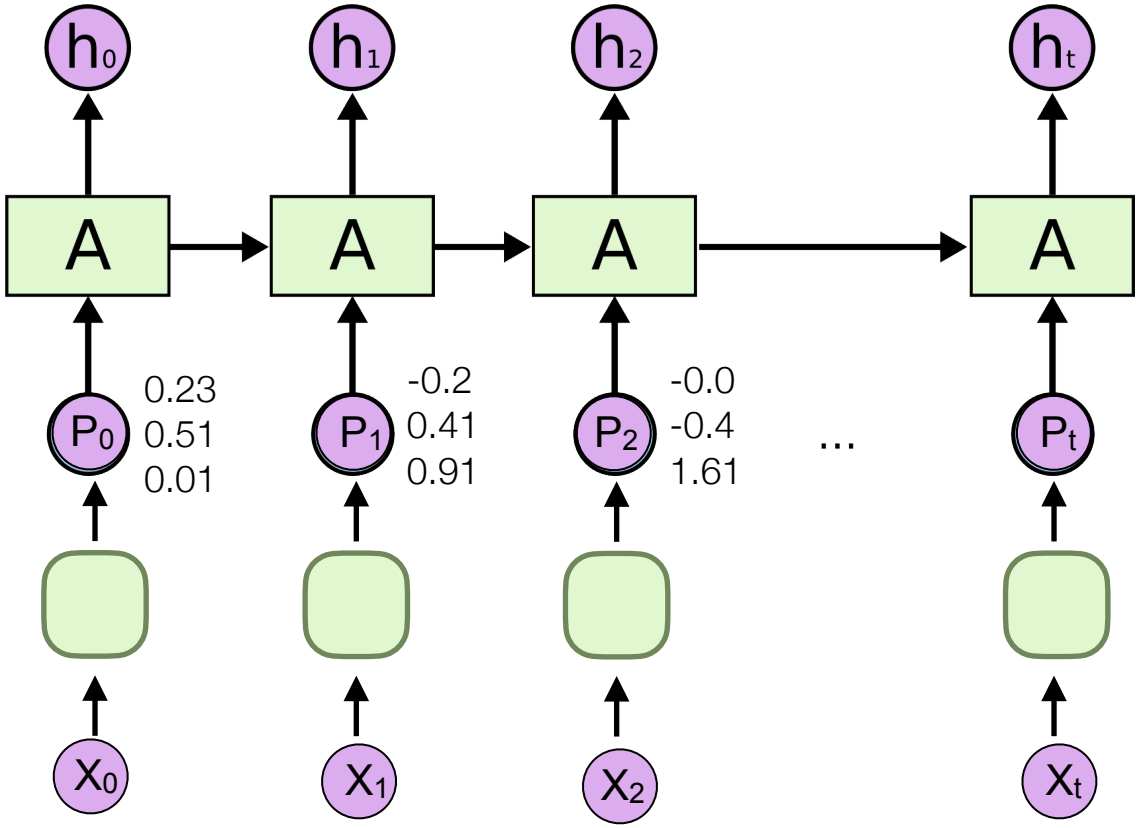
George Karypis
Amazon

Bernie Wang
Amazon

NLP Embeddings Review



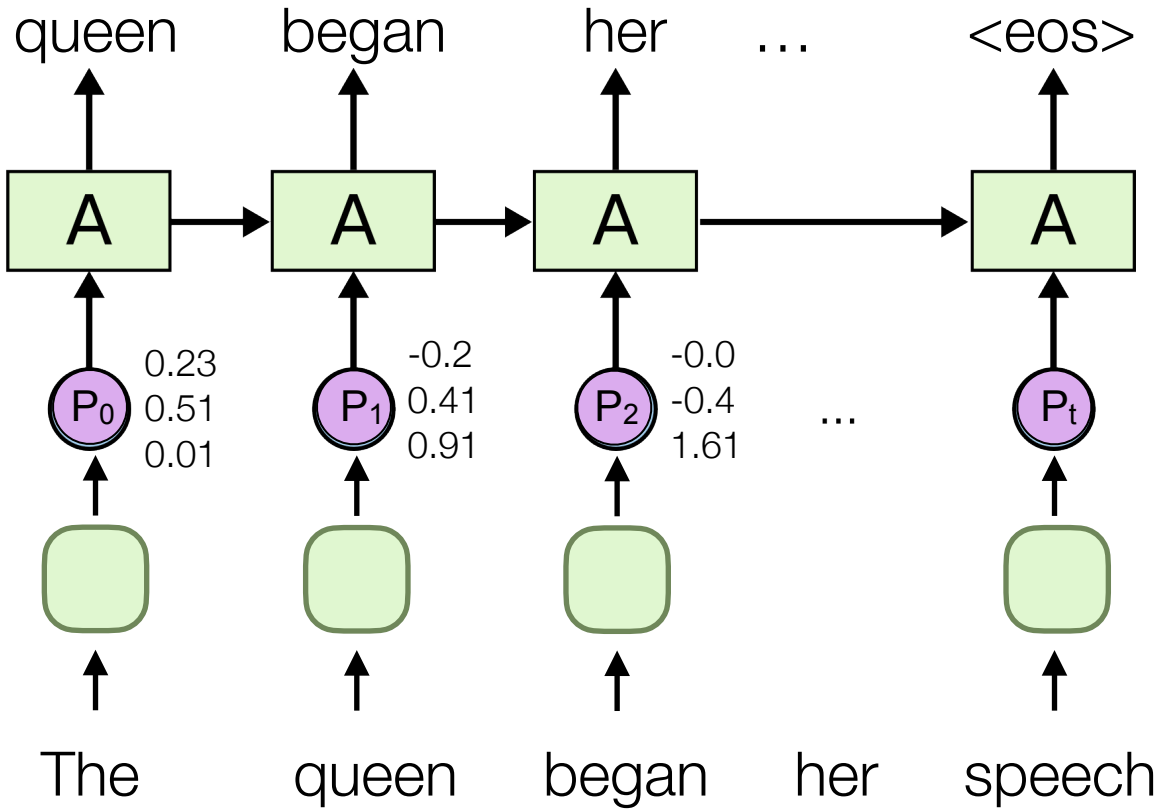
Word Embeddings Review



int:	3	1	17	
one hot:	0 0 1 0 0 0 0	1 0 0 0 0 0 0	0 0 0 0 0 0 0	

Word Embeddings: Training Review

- many training options exist
  - a popular option, next word prediction



GloVe Review

GloVe

Global Vectors for Word Representation

Highlights

1. Nearest neighbors

The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. For example, here are the closest words to the target word *frog*:

0. *frog*
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

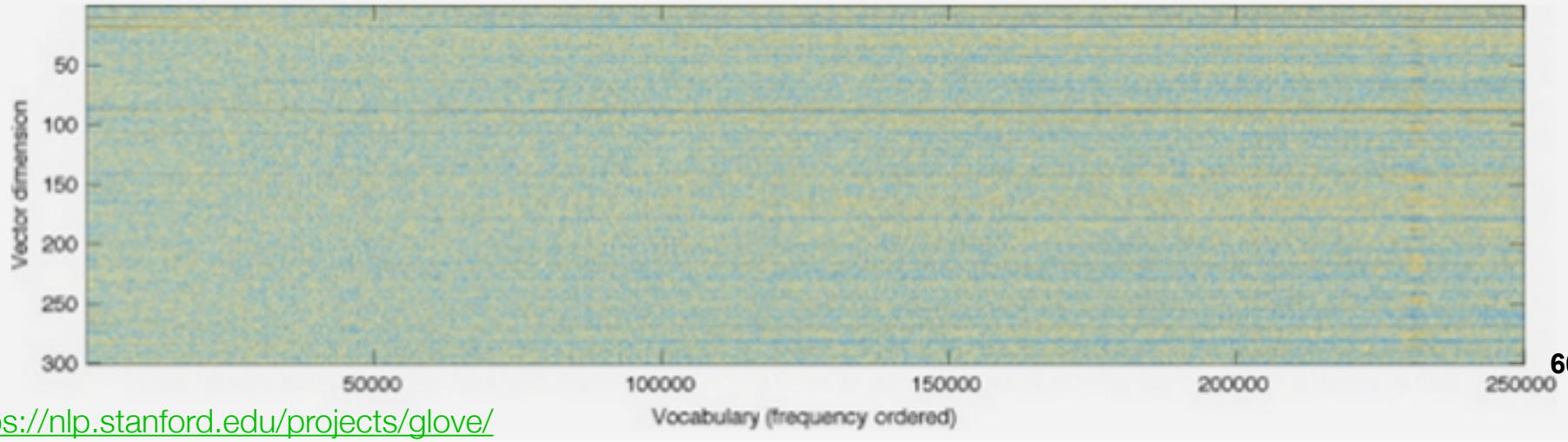
3. litoria

4. leptodactylidae

5. rana

7. eleutherodactylus

GloVe produces word vectors with a marked banded structure that is evident upon visualization:

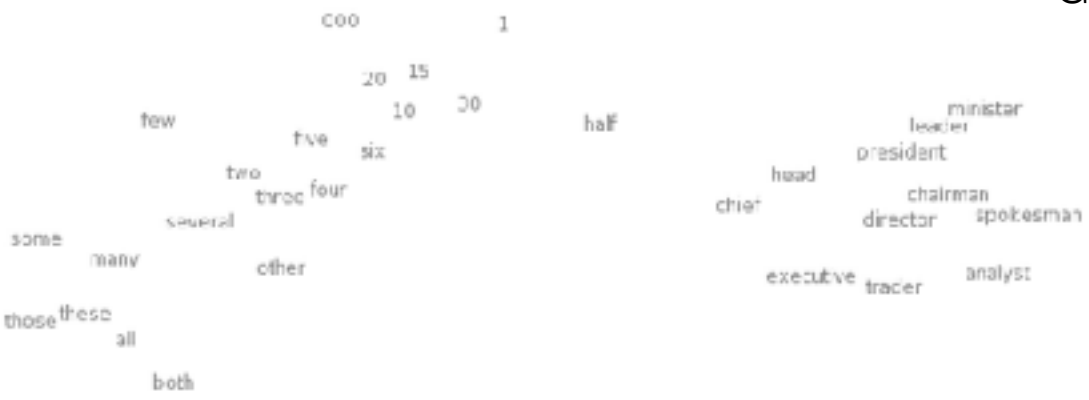


https://nlp.stanford.edu/projects/glove/

Word Embeddings: proximity

GloVe Review

Global Vectors for Word Representation



t-SNE visualizations of word embeddings. Left: Number Region; Right: Jobs Region. From Turian *et al.* (2010), see complete image.

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

What words have embeddings closest to a given word? From Collobert *et al.* (2011)

The **chairman** called the **meeting** to order.

The **director** called the **conference** to order.

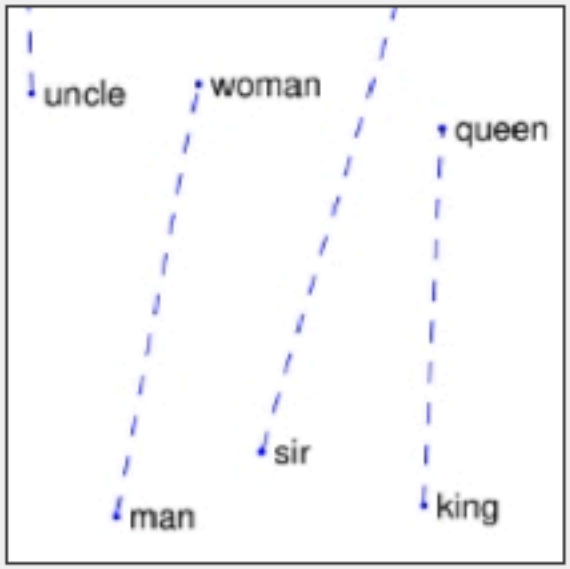
The **chief** called the **council** to order.

http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

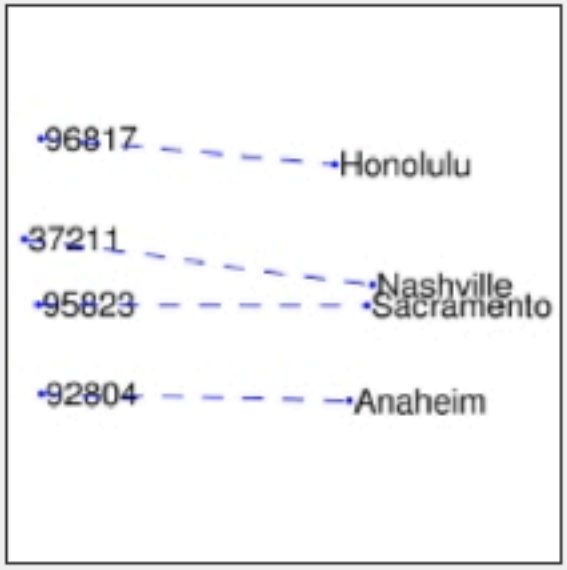
Word Embeddings: Analogy

GloVe Review

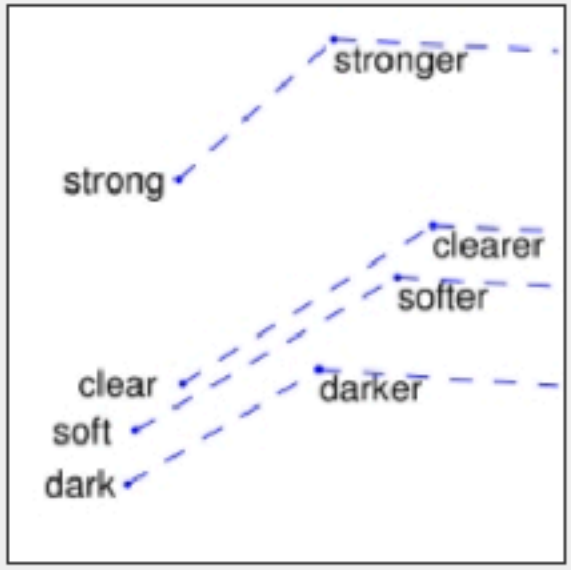
Global Vectors for Word Representation



man - woman



city - zip code



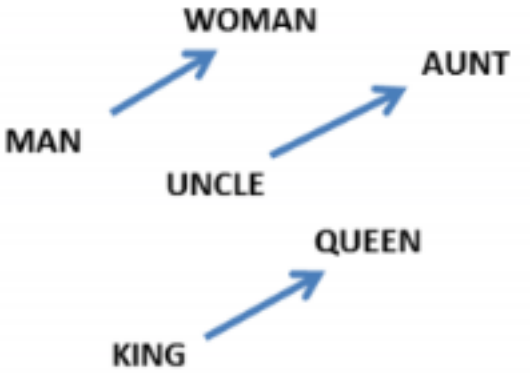
comparative - superlative

each vector difference **might** encode analogy

https://nlp.stanford.edu/projects/glove/ http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

Word Embeddings: Analogy?

GloVe Review



From Mikolov *et al.* (2013a)

$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“aunt”}) - W(\text{“uncle”})$$

$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“queen”}) - W(\text{“king”})$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. fighter pilot	12. boss

Trained on New York Times

Bolukbasi et al., NeurIPs 2016
https://arxiv.org/pdf/1607.06520.pdf

https://nlp.stanford.edu/projects/glove/