MACHINE LEARNING IN CLINICAL MEDICINE

Frank Rudzicz

UHN Toronto Rehabilitation Institute

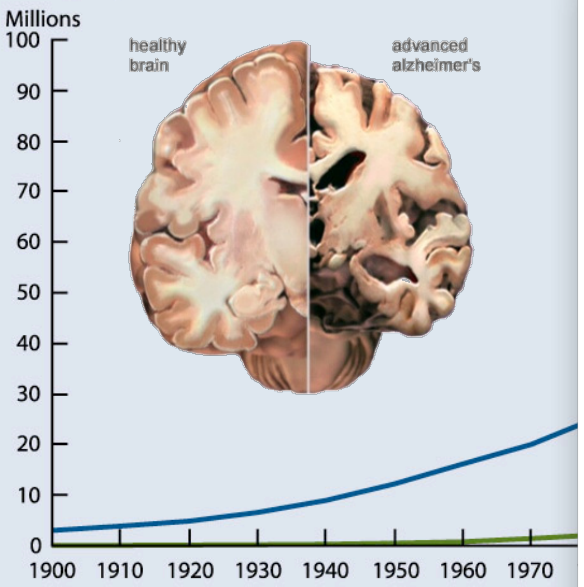
UNIVERSITY OF TORONTO

TODAY

- I'm going to tell you:
  - How talking about *cookies* can reveal *dementia*.
  - How *counting* words can tell you what they *mean*.
  - How you can *ignore errors* in speech recognition.

THE RISING TIDE OF DEMENTIA

Number of people age 65 and over, by age group, selected years 1900–2006 and projected 2010–2050



Note: Data for 2010–2050 are projections of the population.
Reference population: These data refer to the resident population.
Source: U.S. Census Bureau, Decennial Census, Population Estimates and Projections.

Mini-Mental State Examination (MMSE)

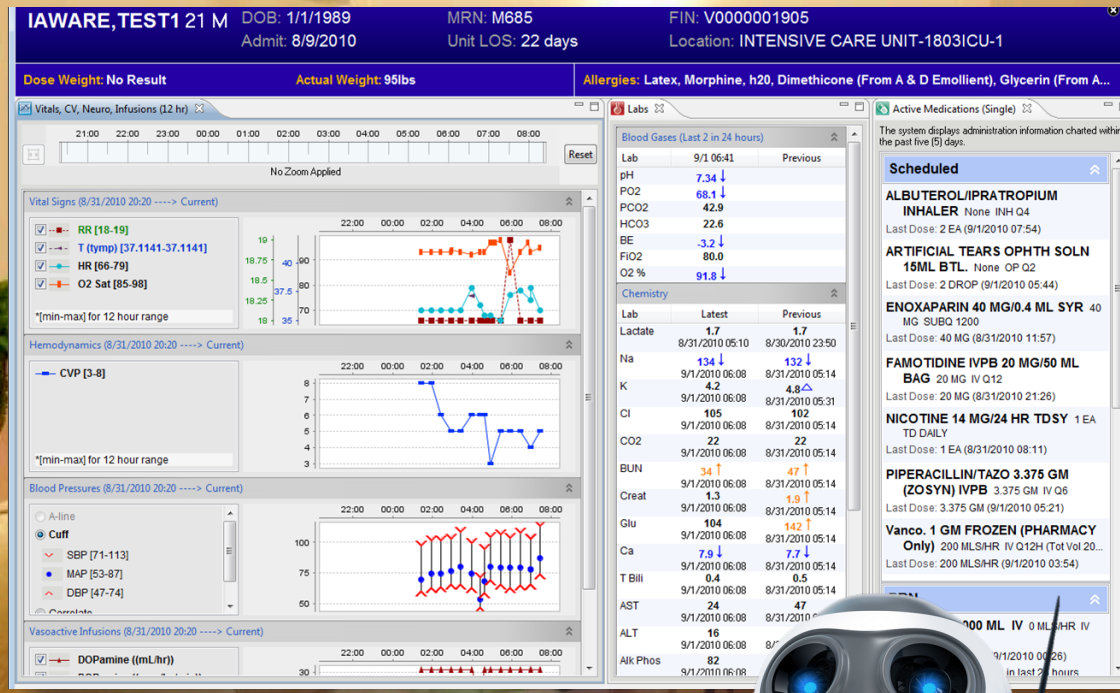
Patient's Name: _____ Date: _____

Instructions: Score one point for each correct response within each question or activity.

Maximum Score	Patient's Score	Questions
5		"What is the year? Season? Date? Day? Month?"
5		"Where are we now? State? County? Town/city? Hospital? Floor?"
3		The examiner names three unrelated objects clearly and slowly, then the instructor asks the patient to name all three of them. The patient's response is used for scoring. The examiner repeats them until patient learns all of them, if possible.
5		"I would like you to count backward from 100 by sevens." (93, 86, 79, 72, 65, …) Alternative: "Spell WORLD backwards." (D-L-R-O-W)

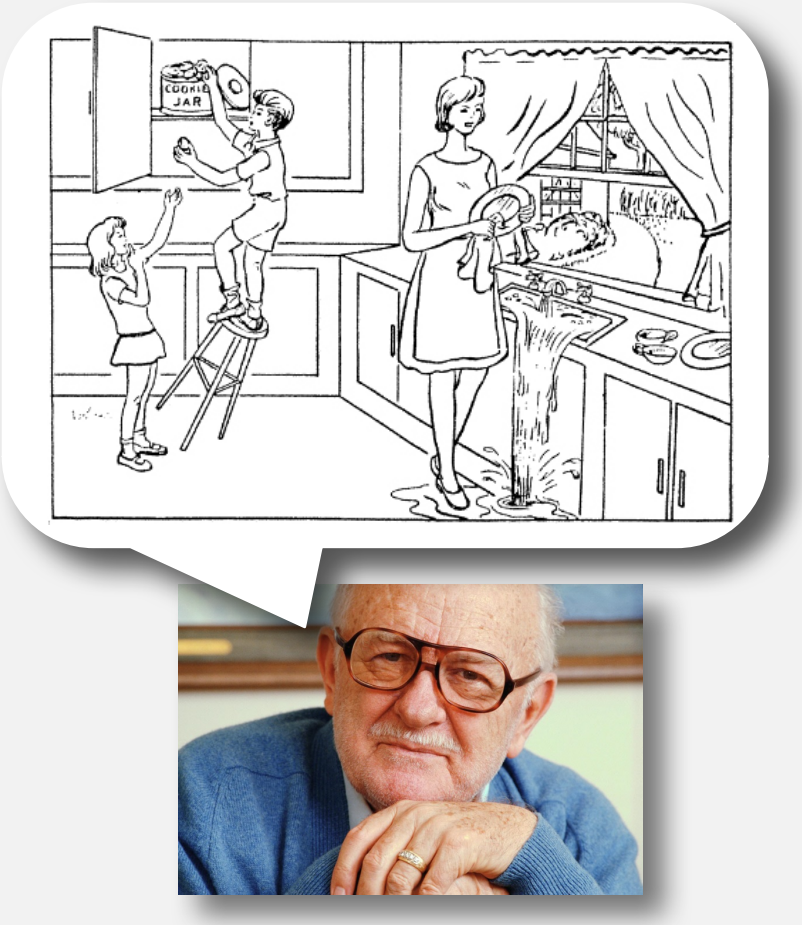
CLINICAL DECISION SUPPORT

"Clinical decision support systems link health observations with health knowledge to influence health choices by clinicians for improved health care"



What kind of data is useful?

ASSESSING ALZHEIMER'S AUTOMATICALLY



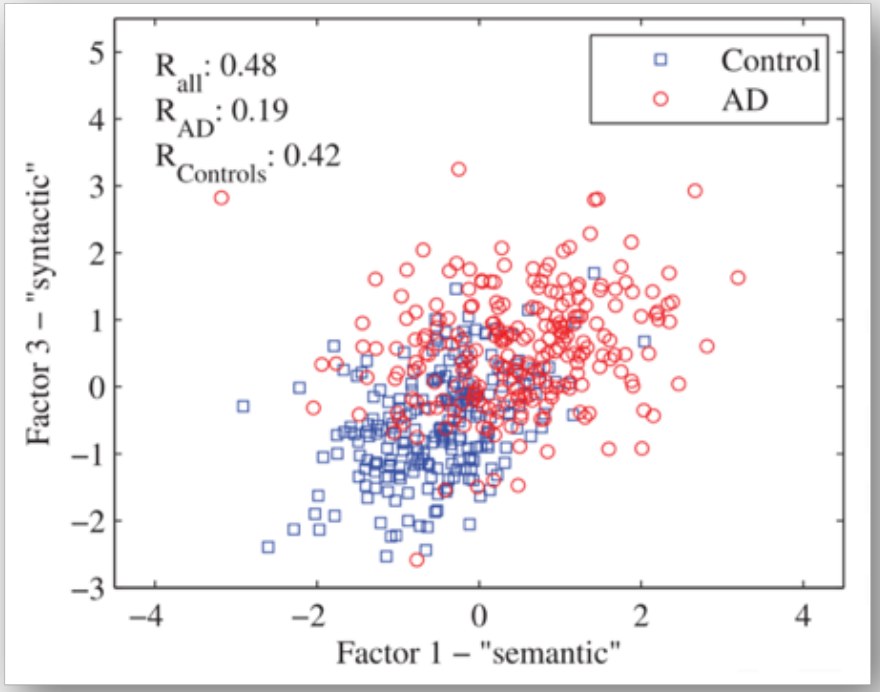
- A task that can be done in less than a minute, on the couch.

- **DementiaBank:**
  240 samples from167 people with AD,
  233 samples from 97 controls.
  - Free-form descriptions of "Cookie Theft" (incl. audio)
  - Transcribed and annotated, e.g., with filled pauses, paraphasias, and unintelligible words.
  - **Mini-mental state exam (MMSE)**

ASSESSING ALZHEIMER'S AUTOMATICALLY

Extract many, many, many, many features related to:

- Words ('lexical')
- Grammar ('syntax')
- Meaning ('semantics')
- Context ('pragmatics')
- Voice ('acoustics')

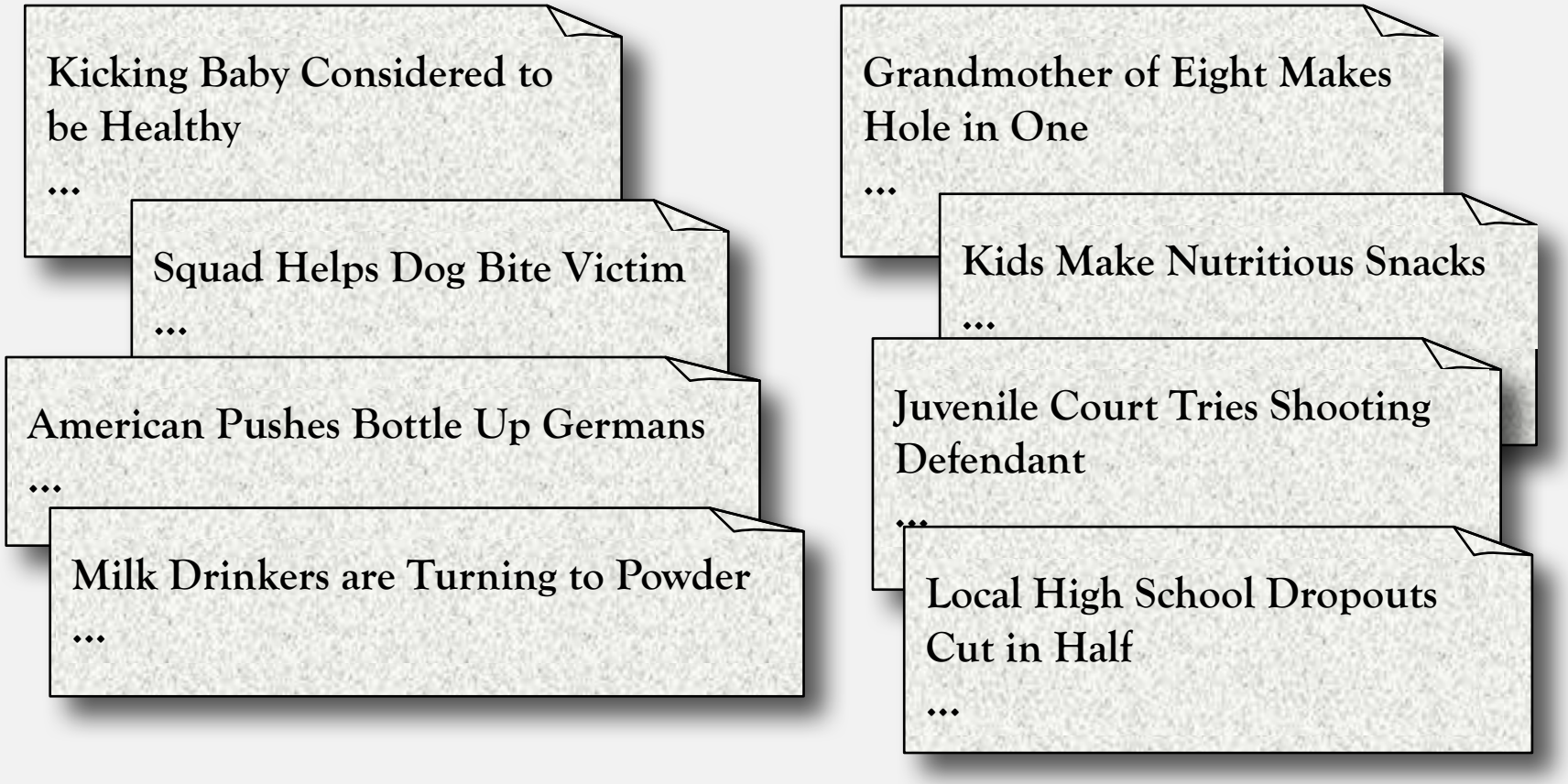


State-of-the-art accuracy: 85% - 92%

*Is this **easy**?*

NO

- **Ambiguity** is everywhere. E.g., newspaper headlines:



AI TO THE RESCUE!

- AI involves **resolving ambiguity** at all levels.
  - **Reasoning** with **world knowledge.**
    - In the early days, knowledge was **explicitly encoded** in artificial **symbolic** systems (e.g., context-free grammars) by **experts.**
  - Now, algorithms **learn** using **probabilities** to distinguish between subtly different competing hypotheses.
    - E.g., does a clinical note indicate diabetes or not?
    - Examine many *examples* of both, and then compute something like:

$$P(diabetes) > P(no\ diabetes) > 0$$

*How do you **learn semantics**?*

LATENT SEMANTIC INDEXING

- Consider the following:

	Term 1	Term 2	Term 3	Term 4
Query	ignoring	sink		
Record 1			water	overflowing
Record 2	ignoring	sink	water	overflowing

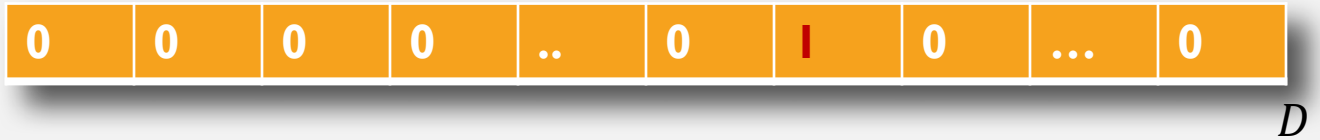
  - Record 1 appears to be **related** to the query although it contains **none** of the query terms.
    - The query and Record 1 are **semantically related**.

*How do you learn semantics **of words**?*

BAG OF WORDS

- Words are often treated as if they're **marbles in a bag**.
- Imagine each of D available words is a 0-vector with a unique 1.

sink =

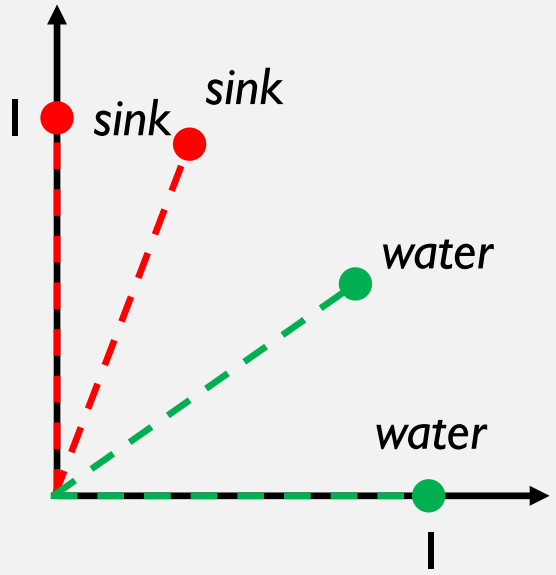


In this approach, words **do not overlap**:

sink = $[0,0,0,\ldots,0,\mathbf{1},0,\ldots,0]$, &

water = $[0,0,0,\ldots,0,0,\mathbf{1},\ldots,0]$

There is no shared information



CO-OCCURRENCE MATRIX

Co-occurrence: when two or more terms occur in similar contexts more often than by chance.

a										31	18	…
boy										32	3	…
on										56	33	…
stool										62	1	…
the										12	12	…
girl										16	34	…
wants										88	23	…
to										32	12	…
give											7	…
…	…	…	…	…	…	…	…	…	…	…	…	…

Corpus

I see a boy on a stool

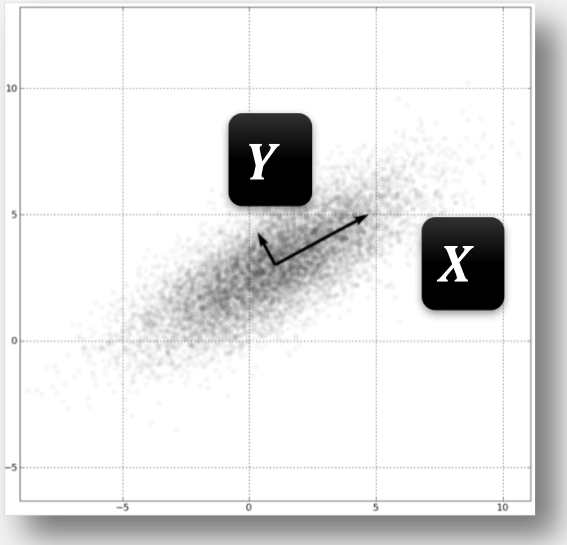
How do you pull out ***hidden information****?*

I don't … I don't know what the… what it is

…

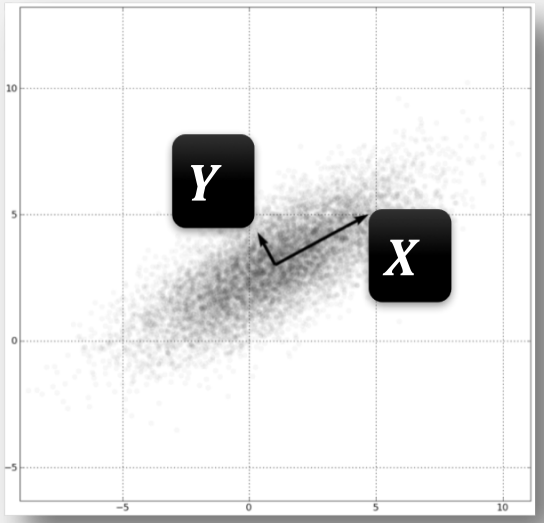
Co-occurrence

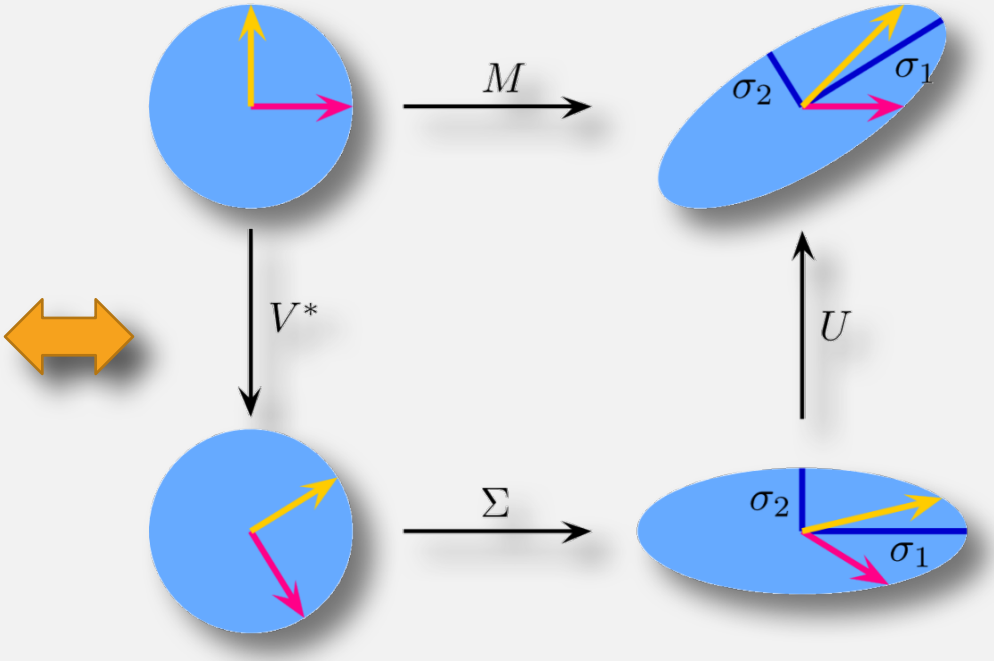
LATENT DIMENSIONS



- Principal components analysis (PCA) finds **latent dimensions** of maximum variance within a dataset.
- Imagine each grey dot is a row of our **co-occurrence matrix** – one dot per word.
- We can rotate and project words *down* onto fewer latent dimensions.

SINGULAR VALUE DECOMPOSITION I





$$M = U \cdot \Sigma \cdot V^*$$

PCA

SVD

SINGULAR VALUE DECOMPOSITION 2

$M =$

	i	see	a	boy	on	stool	the	girl	wants	to	give	…
i	81	47	72	47	23	41	38	62	23	48	50	…
see		26	43	21	33	45	44	21	24	12	9	…
a			80	41	40	22	12	34	56	31	18	…
boy				43	38	45	11	51	20	32	3	…
on					56	41	67	68	31	56	33	…
stool						42	51	13	51	62	1	…
the							86	51	42	12	12	…
girl								23	33	16	34	…
wants									18	88	23	…
to										32	12	…
give											7	…
…	…	…	…	…	…	…	…	…	…	…	…	…

$= U \cdot \Sigma \cdot V^*$

“embedding”

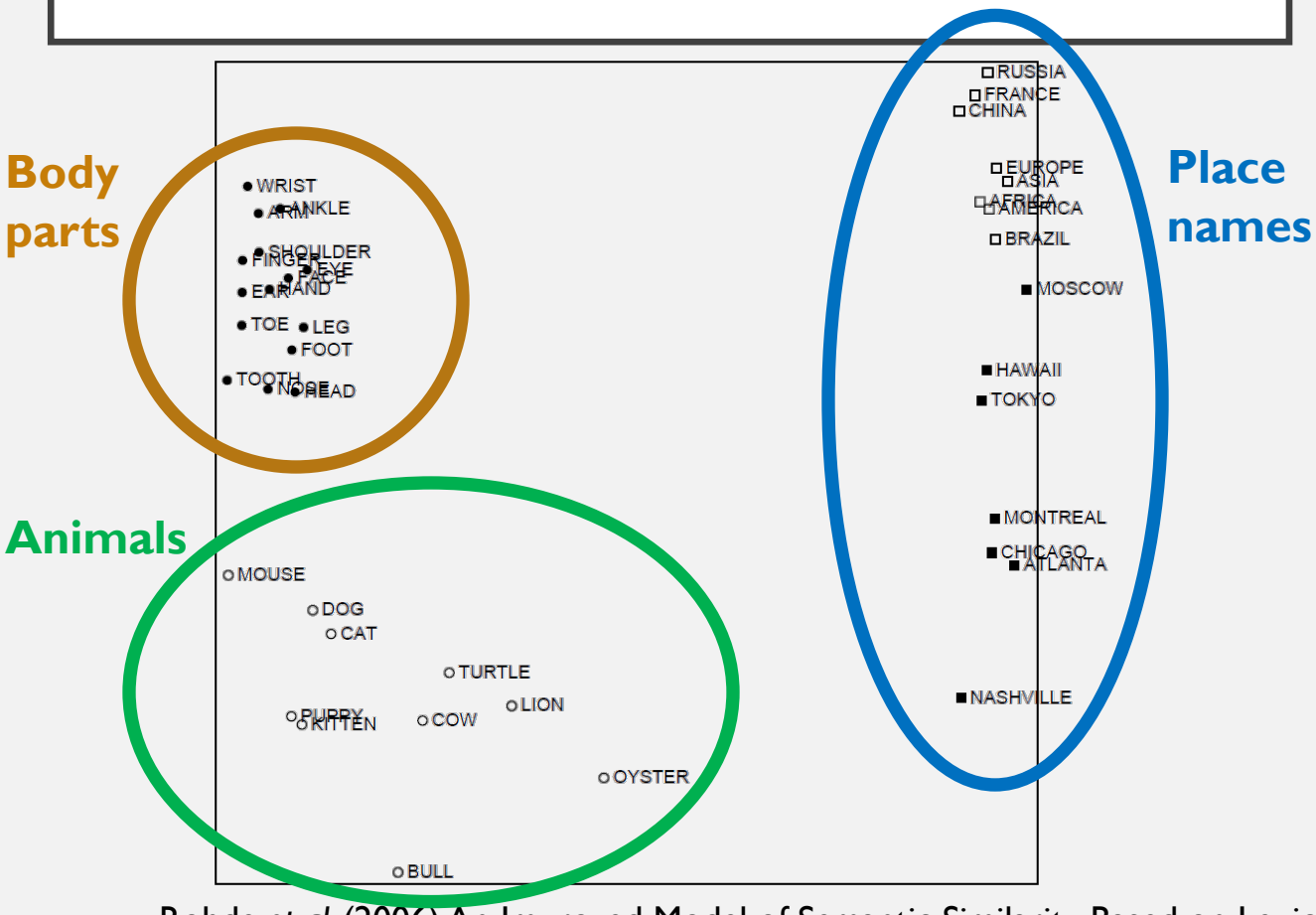
$\Sigma =$

2.16	0	0	0	…
0	1.59	0	0	…
0	0	1.28	0	…
0	0	0	1	…
…	…	…	…	…

$U =$

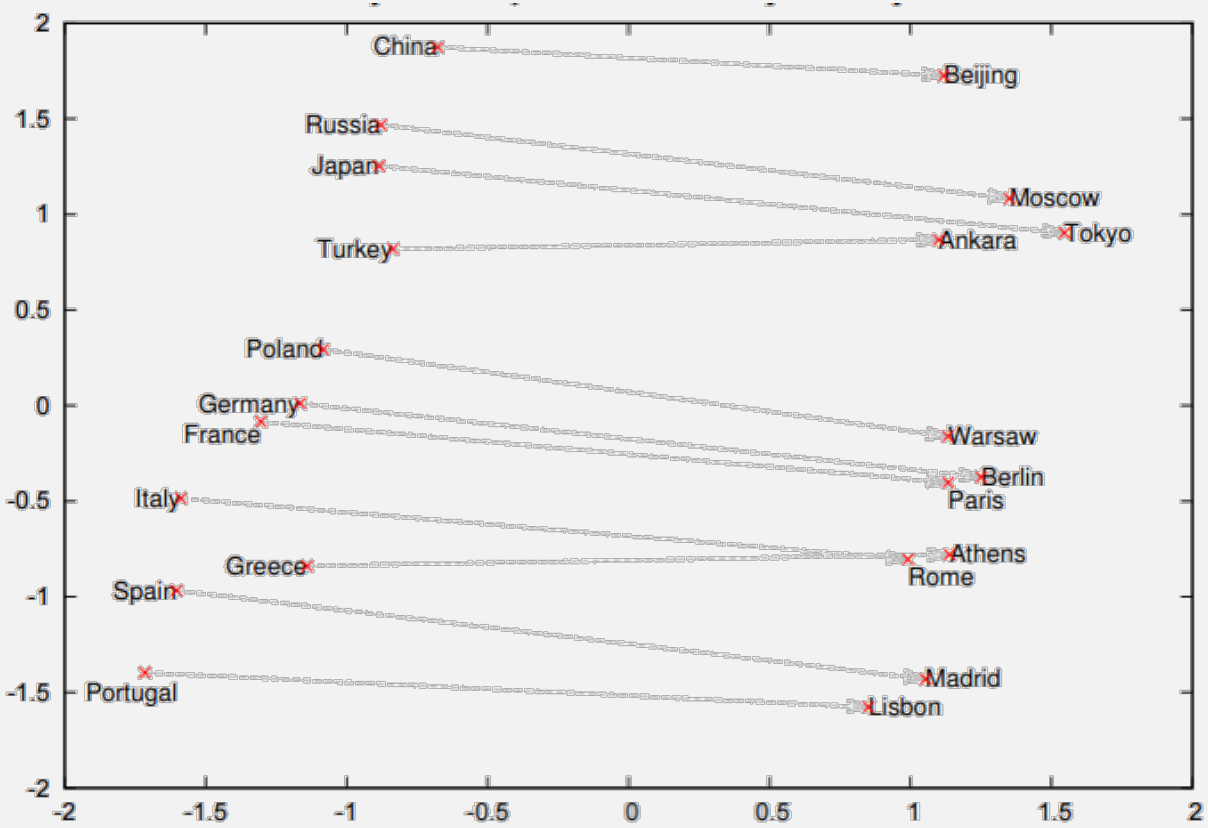
ignore	-0.44	-0.30	0.57	0.58	…
sink	-0.13	-0.33	-0.59	0	…
water	-0.48	-0.51	-0.37	0	…
overfl.	-0.70	0.35	0.15	-0.58	…
…	…	…	…	…	…

SINGULAR VALUE DECOMPOSITION 3



Rohde *et al.* (2006) An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. *Communications of the ACM* **8**:627-633.

REGULARITIES IN WORD-VECTOR SPACE



Trained on the Google news corpus with over 300 billion words.

REGULARITIES IN WORD-VECTOR SPACE

Expression	Nearest token
Paris – France + Italy	Rome
Bigger – big + cold	Colder
Sushi – Japan + Germany	bratwurst
Cu – copper + gold	Au
Windows – Microsoft + Google	Android

Analogies: apple:apples :: octopus:octopodes
Hypernymy: shirt:clothing :: chair:furniture

Similar relations will be discoverable in genetics texts.

FROM SVD TO NEURAL NETWORKS

- **SVD**: Computational costs grow quickly with M.
  'Hard' to incorporate new words.

- **Neural networks**: Don't capture co-occurrence directly
  Just try to model *surrounding* words.

Build a model that *can* do accurate predictions *in order* to learn relations.

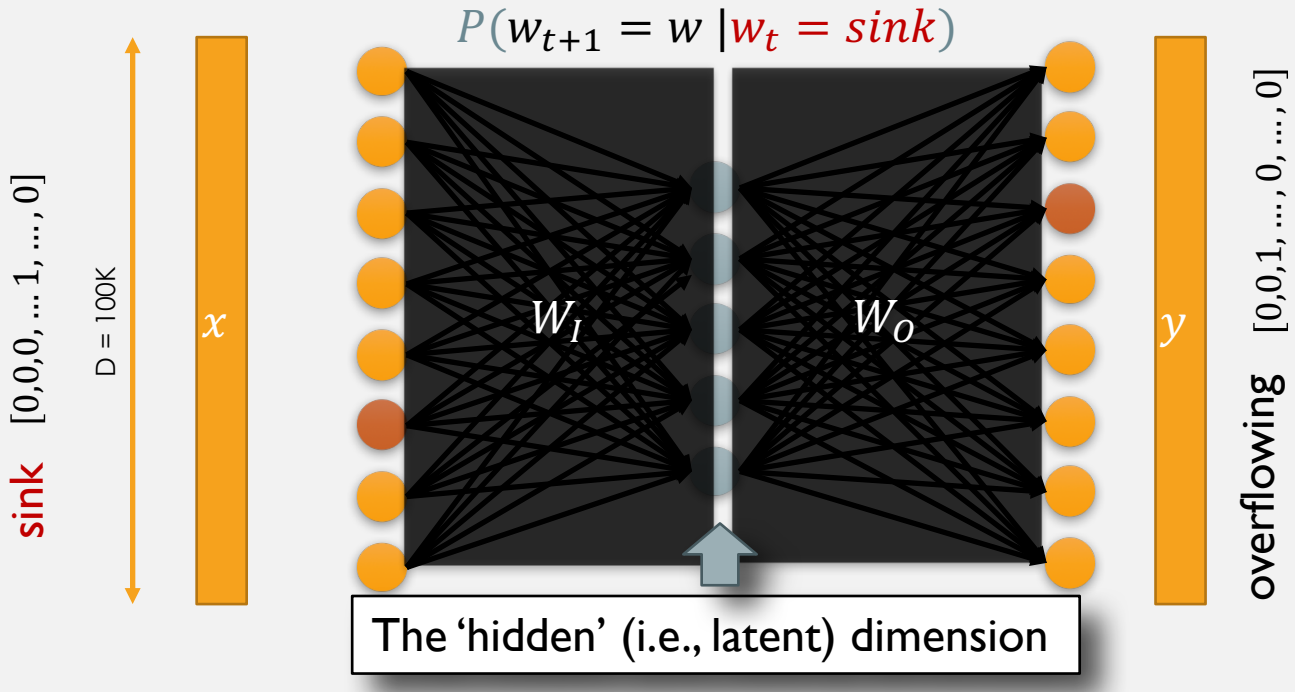
$$P(w_{t+2} = overflowing | w_t = sink)$$

and the sink was overflowing

and the water was overflowing

…

NEURAL NETWORKS



We i) '**plug in**' *each* word *in sequence*, ii) **perform** matrix multiplication, iii) **compare** the result to the next word, and iv) **propagate** the error back through the weights.

THE LEARNING BIT

- Our model is $\theta = [W_I, W_O]$
- The model is used in: $P(w_{t+1}|w_t) = \frac{\exp((yW_O)^\intercal xW_I)}{\sum_{w=1}^{W} \exp((wW_O)^\intercal xW_I)}$ "softmax"
- To see how well our network is adjusted, we want to maximize an '**objective function**' that characterizes our prediction:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \log P(w_{t+1}|w_t)$$

- Adjust model θ using a 'learning rate' η and a derivative of $J(\theta)$:

$$\theta^{(new)} \leftarrow \theta^{(old)} - \eta \nabla_\theta J(\theta)$$

What a wonderful question.

USING WORD REPRESENTATIONS

Originally, words didn't overlap:

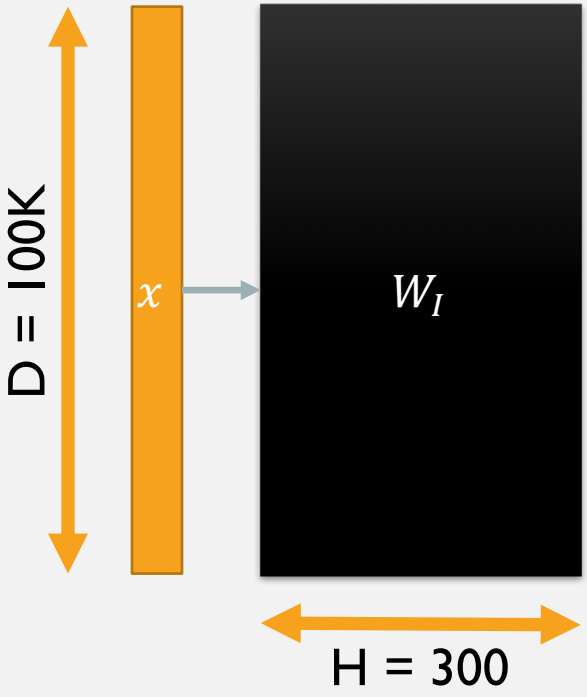
sink = [0,0,0, … , 0,1,0, … , 0], &

water = [0,0,0, … , 0,0,1, … , 0] so

Similarity = 0.0

Transform

$v_w = x^T W_1$



Through learning, they now do:

sink = [0.67,0.9, … , 0.1, 0.32], &

water = [0.74,0.8, … , 0.2, 0.44] so

Similarity = 0.67

https://code.google.com/p/word2vec/

SUMMARY: LATENT SPACES

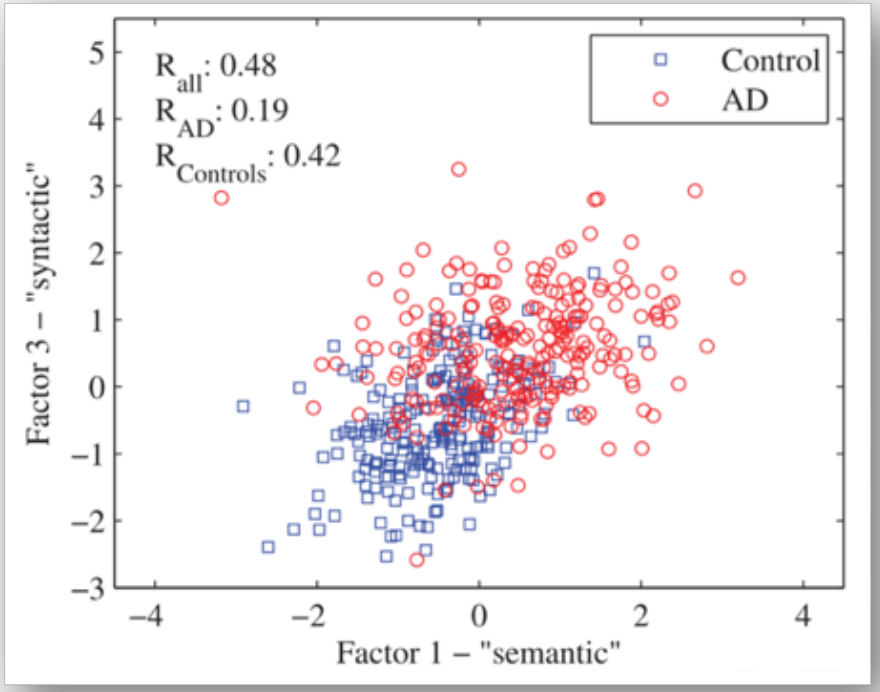
- SVD and neural networks transform words into lower-dimensional '**latent spaces**' that encode **information**.
- This is (part of) what Google does when it ranks the **relatedness** of *web pages* given *search terms*.
  - It is applicable to (almost) any information retrieval task in biology.
- We don't need a formal representation of **meaning**, we can just use some statistics of how words co-occur.

"words of a feather flock together."
"you shall know a word by the company it keeps."
- J.R. Firth (1957)

ASSESSING ALZHEIMER'S AUTOMATICALLY

Extract many, many, many, many features related to:

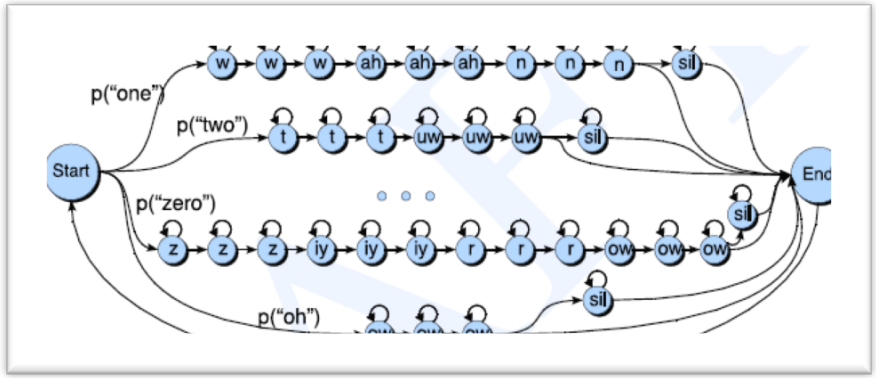
- Words ('lexical')
- Grammar ('syntax')
- Meaning ('semantics')
- Context ('pragmatics')
- Voice ('acoustics')

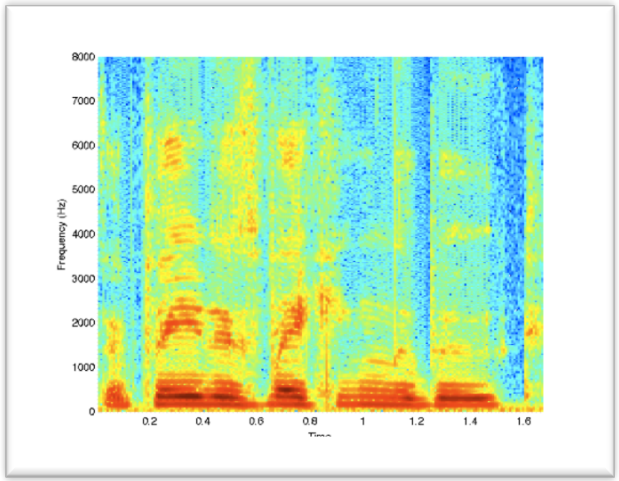


State-of-the-art accuracy: 85% - 92%

*What if words are **misheard**?*

SPEECH RECOGNITION (ASR)





Language model

Acoustic model

TYPES OF ERROR

- We can compute **word-error rate (WER)**, to count different **kinds** of errors:
- **Substitution error**: A word being mistook for another
  e.g., '*think*' given '*sink*'
- **Deletion error**: An input word that is 'skipped'
  e.g. '*She ignoring*' given '*She is ignoring*'
- **Insertion error**: A 'hallucinated' word not said.
  e.g., '*He wants the delicious cookies*'
  given '*He wants the cookies*'
- *How do these errors affect subsequent features?*

FEATURE SELECTION I

- **Different errors** in the text data will have **different effects** on the features we extract.
- We want to only use features that are **robust** against error.
- So we choose a subset S^* of k features f_i that are best at differentiating category c (e.g., Alzheimer's disease), using Spearman correlation ρ:

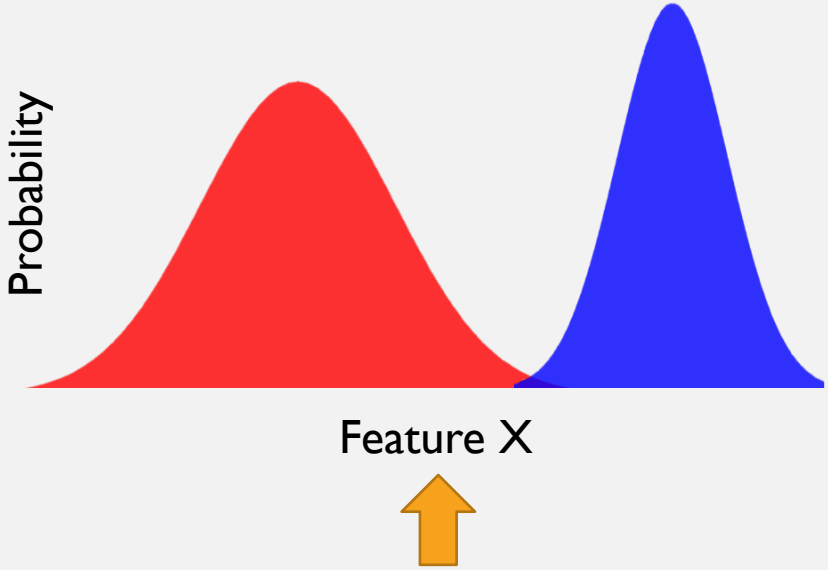
$$S^* = \underset{S}{\operatorname{argmax}} \frac{\sum_{f_i \in S} \rho_{c f_i}}{\sqrt{k + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \rho_{f_i f_j}}}$$

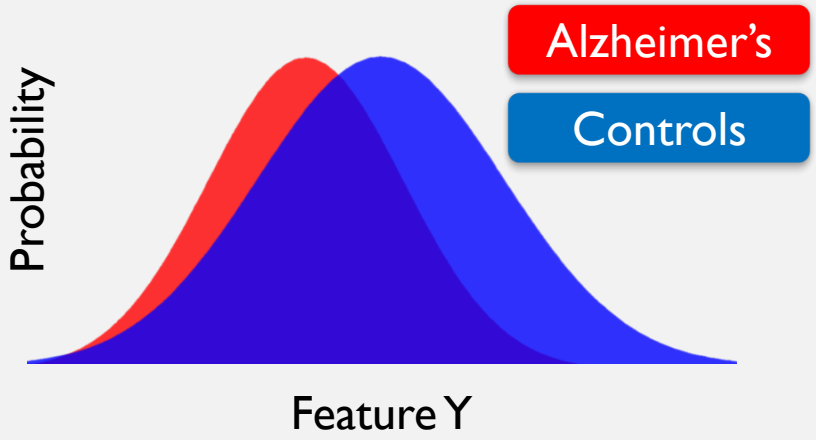
Maximize relevance

Minimize redundancy

FEATURE SELECTION 2

- Another approach is to use **statistical hypothesis testing**.

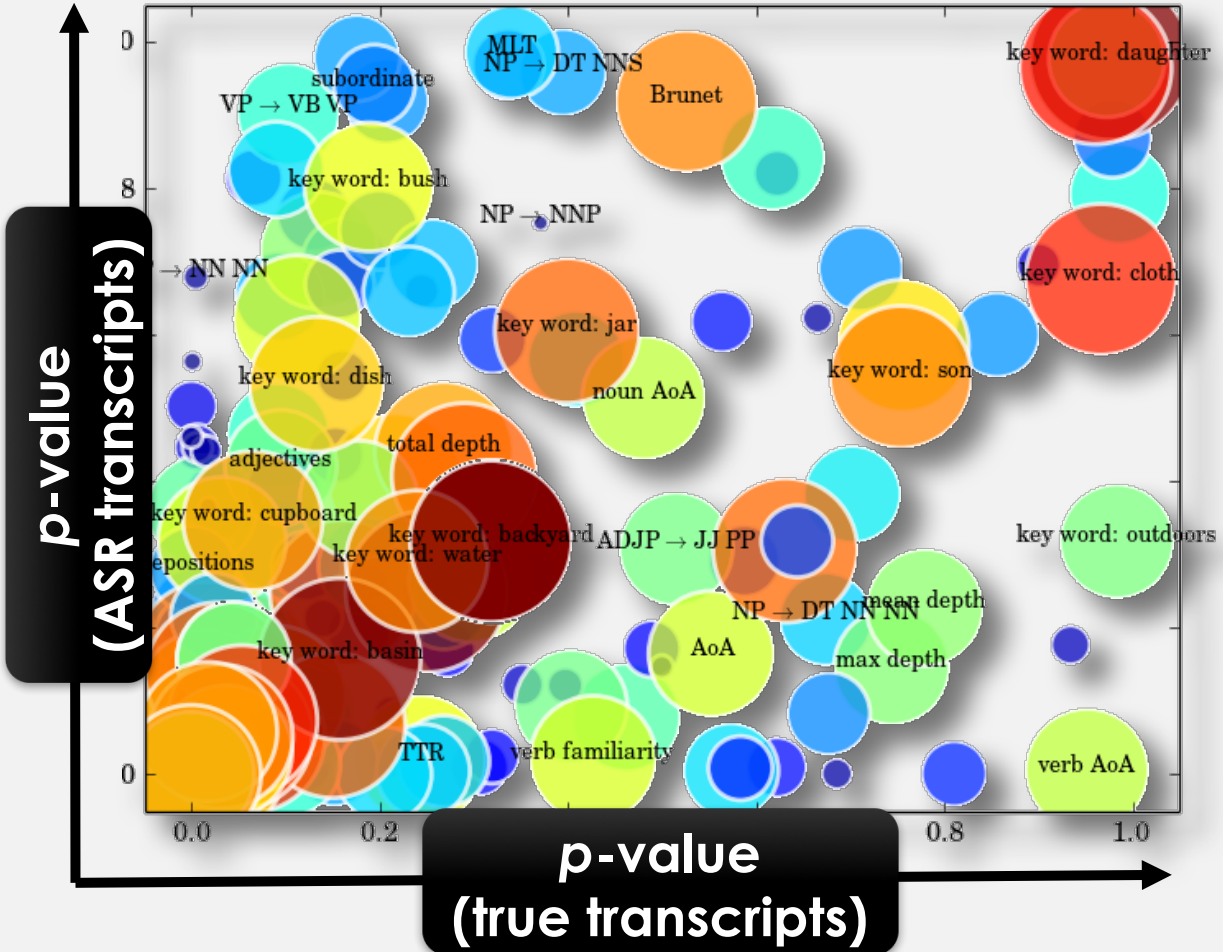




Feature X is better at telling groups apart.

(i.e., feature X gives a smaller *p*-value, given a *t*-test of significance)

ROBUSTNESS FROM ERROR



SUMMARY

- Useful clinical solutions are possible given Big Data and i) natural behavioural tasks, ii) *many* extracted features, and iii) modern machine learning.
  - However, relevant features can be *hidden* below the surface
- We can infer hidden information by using latent-space models, including modern neural networks.
  - However, these can be affected by errors or 'noise'.
- We can overcome errors and noise by selecting features that are appropriate.

Thank you!