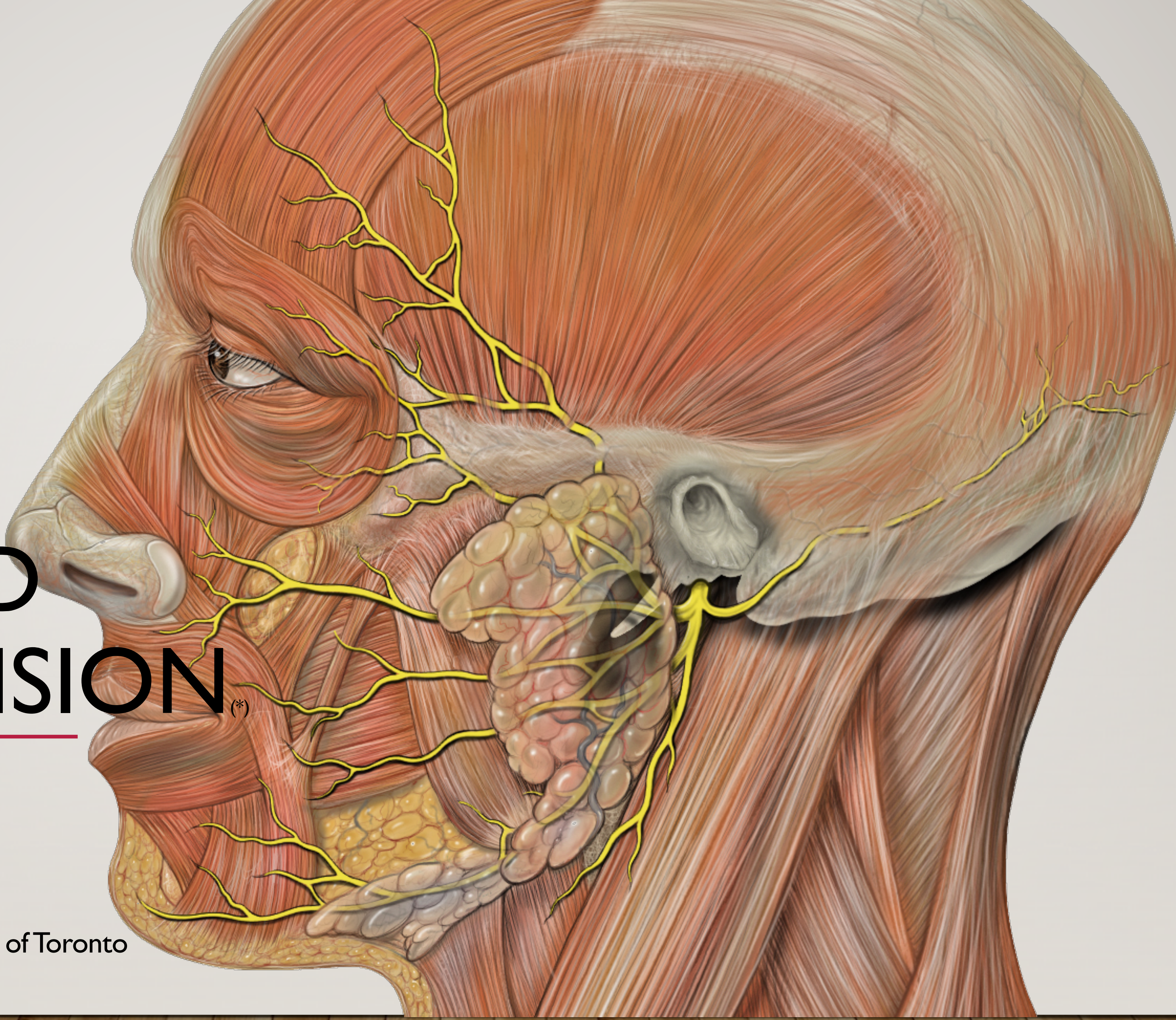
SOUND AND VISION(*)



AND LANGUAGE

CSC490/2600 Fall 2016
Frank Rudzicz, University of Toronto
Lecture 2

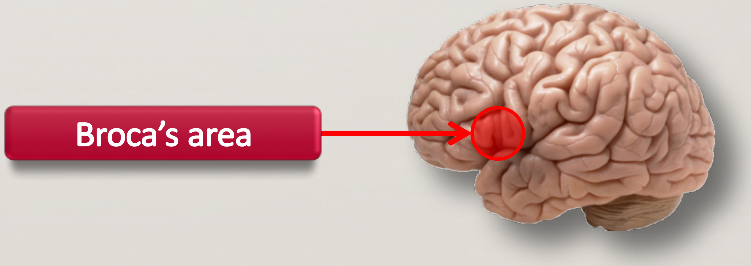
(*) David Bowie

SPEECH AND LANGUAGE DISORDERS

STUDYING HOW SYSTEMS BREAK DOWN

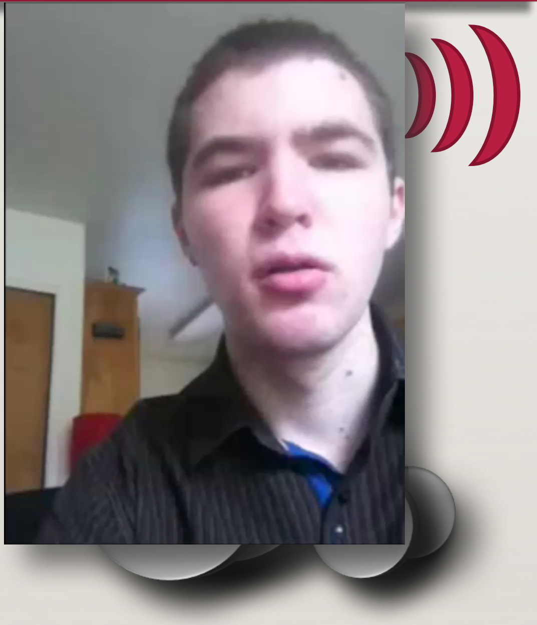
- Observing how **closed systems** ***fail*** can be a **valuable method** in discovering how those systems **work**.

- **Paul Broca** (left) discovered, in 1861, that a **lesion** in the **left** ventro-posterior **frontal lobe** caused **expressive aphasia.**
- This was the first **direct** evidence that **language function** was **localized**.
  - It hinted at a **mechanistic** view of **speech production**.



DYSARTHRIA

Neuro-motor articulatory disorders resulting in **unintelligible** speech.



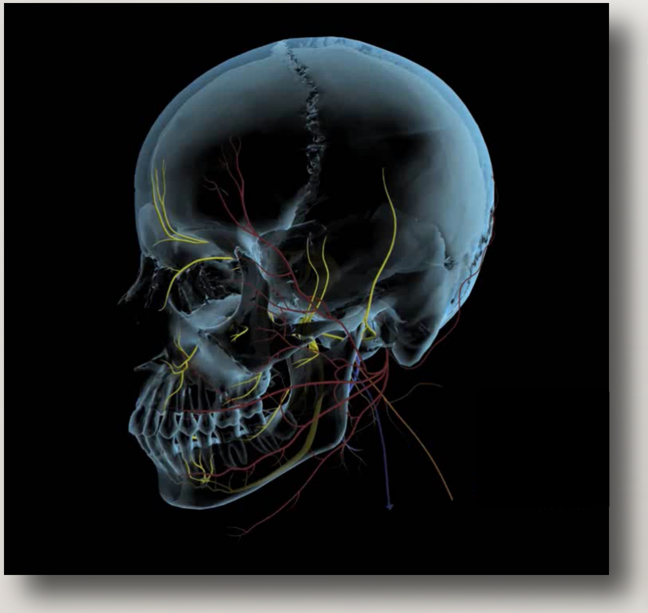
7.5 million Americans have **dysarthria**

- Cerebral palsy,
- Parkinson's,
- Amyotrophic lateral sclerosis)

(National Institute of Health)

NEURAL ORIGINS

- **Types** of dysarthria are related to **specific sites** in the subcortical nervous system.

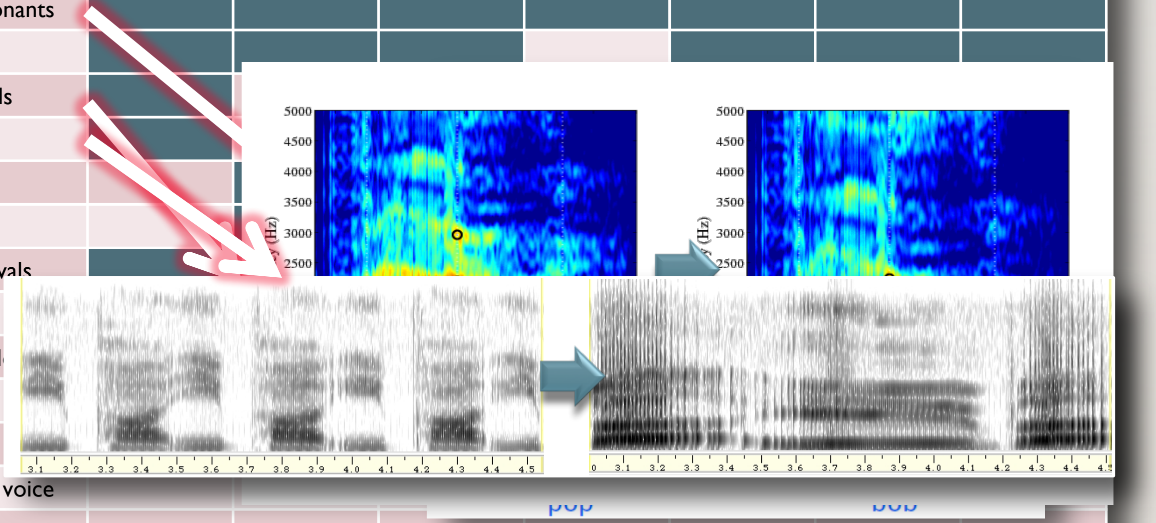


Type	Primary lesion site
Ataxic	Cerebellum or its outflow pathways
Flaccid	Lower motor neuron (≥1 cranial nerves)
Hypo-kinetic	Basal ganglia (esp. substantia nigra)
Hyper-kinetic	Basal ganglia (esp. putamen or caudate)
Spastic	Upper motor neuron
Spastic-flaccid	Both upper and lower motor neurons

(After Darley *et al.*, 1969)

CHARACTERISTICS OF DYSARTHRIA

	Ataxic	Flaccid	Hypo-kinetic	Hyper-kinetic, chorea	Hyper-kinetic, dystonia	Spastic	Spastic-flaccid (ALS)
Monopitch	■	■	■	■	■	■	■
Harshness	■	■	■	■	■	■	■
Imprecise consonants	■	■	■	■	■	■	■
Mono-loud	■	■	■		■	■	■
Distorted vowels	■	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Slow rate	■	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Short phrases		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Hypernasal		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Prolonged intervals	■	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Low pitch	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Inappropriate sil[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Variable rate	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Breathy voice	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Strain-strangled voice	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
…							



(After Darley *et al.*, 1969)

SPEECH RECOGNITION

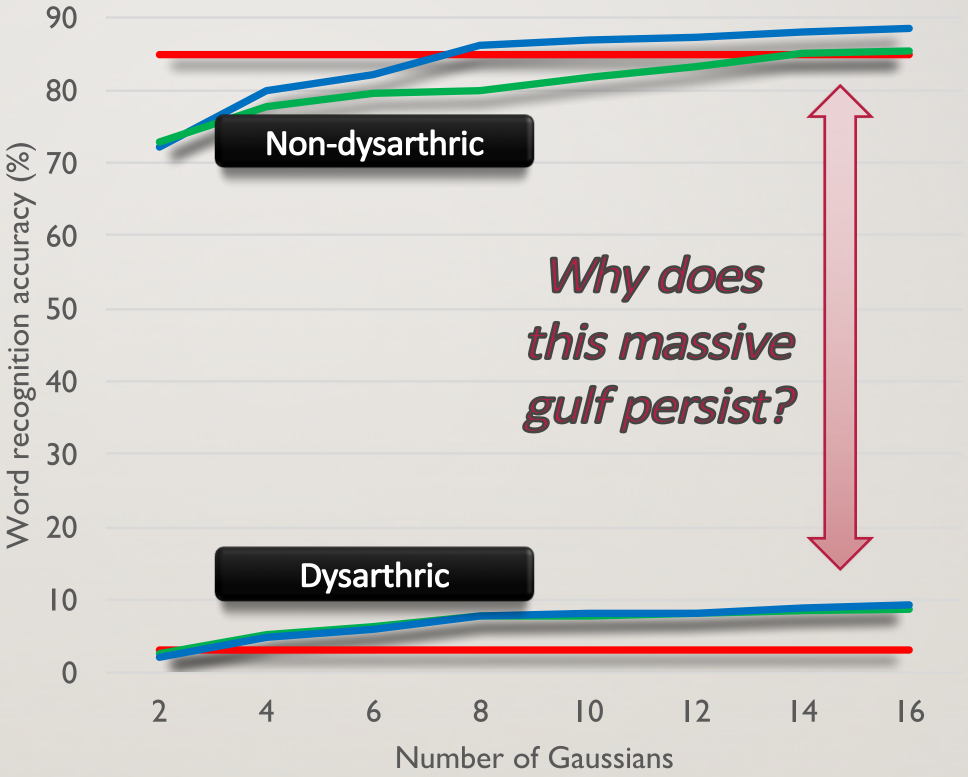
84.9%

Traditional ASR

Speaker-dependent

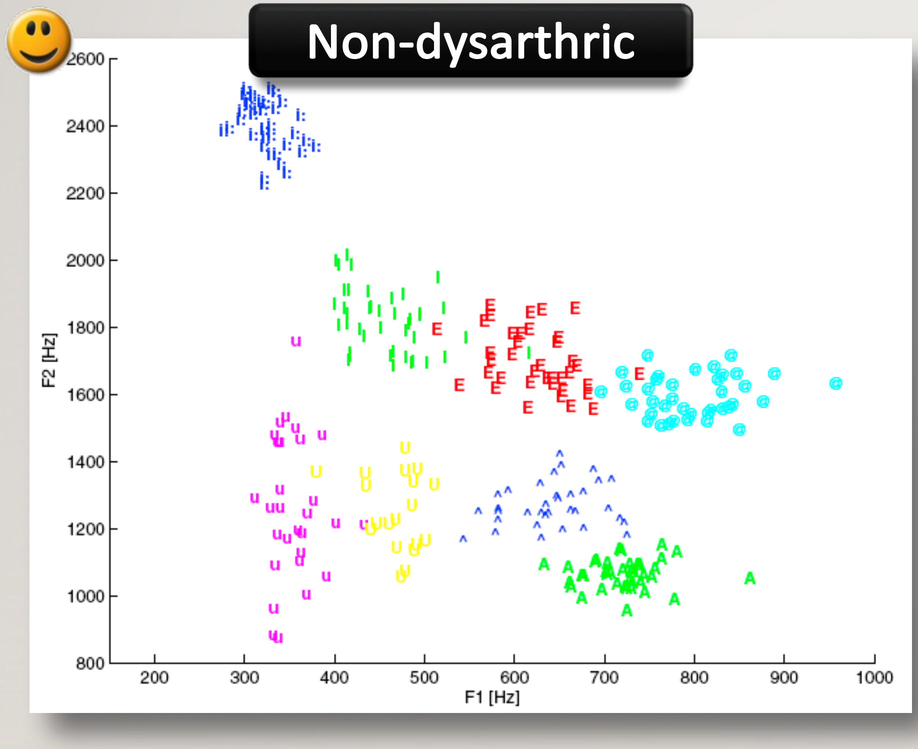
Speaker-retrained

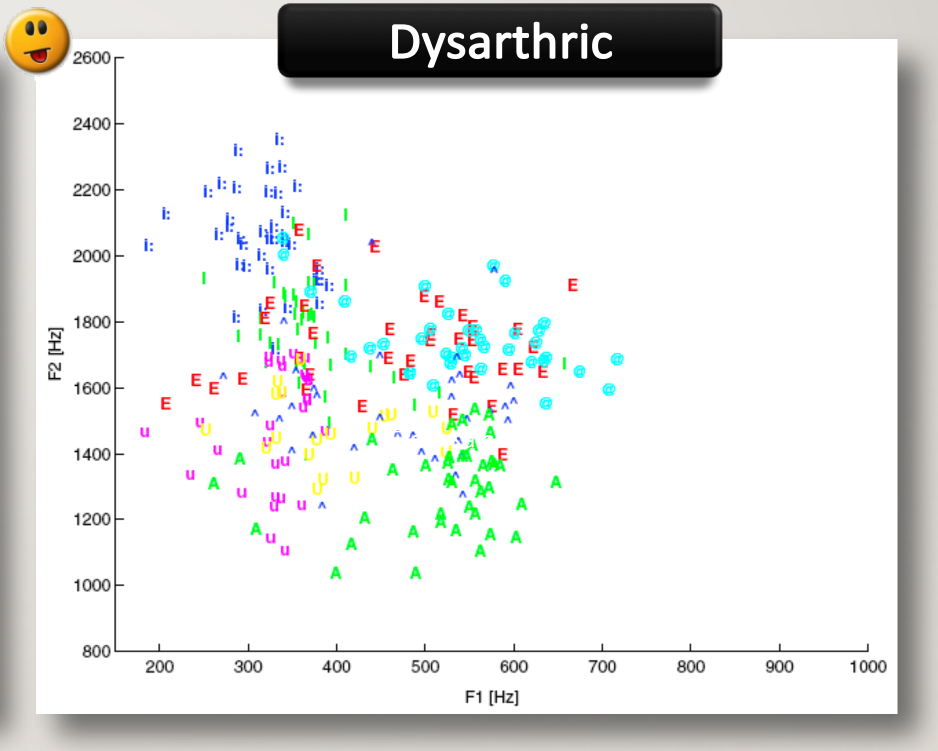
3.1%



Why does this massive gulf persist?

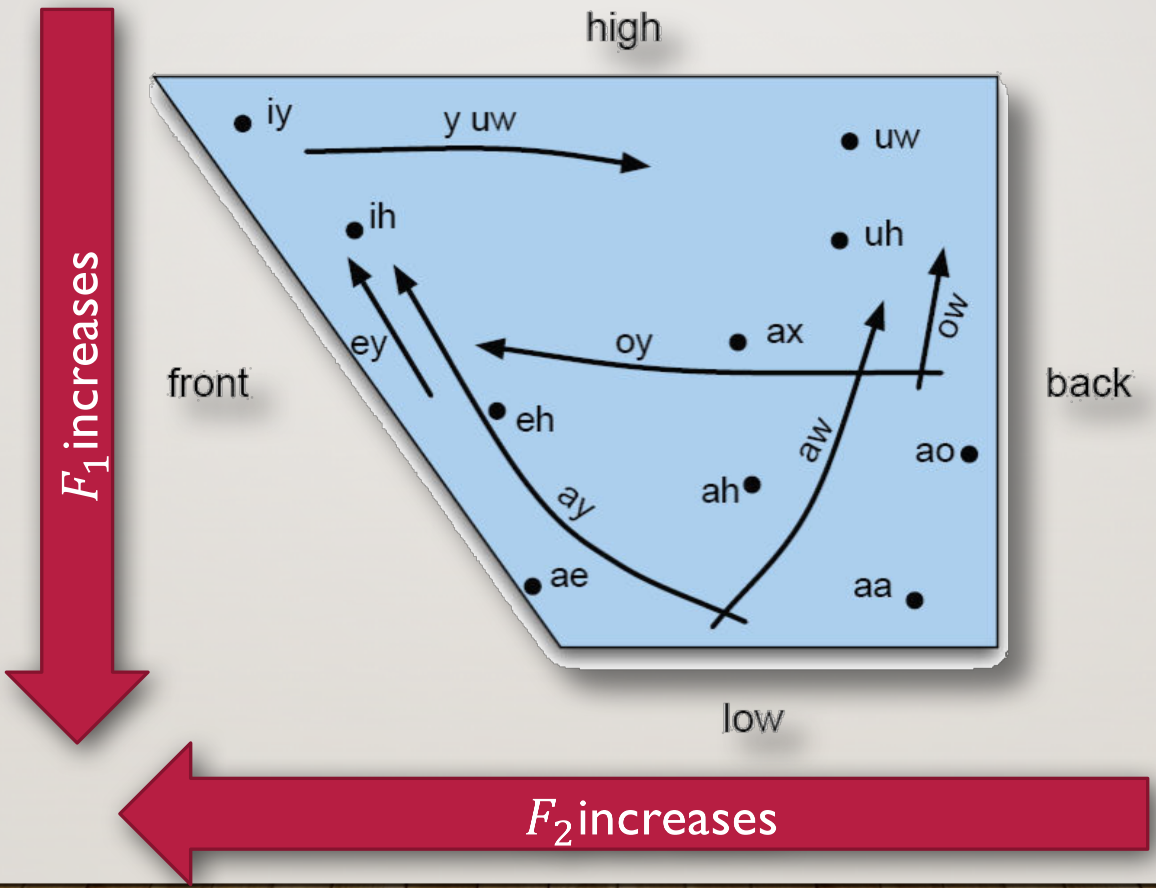
ACOUSTIC AMBIGUITY


Non-dysarthric


Dysarthric

This **acoustic** behaviour is indicative of underlying **articulatory** behaviour.

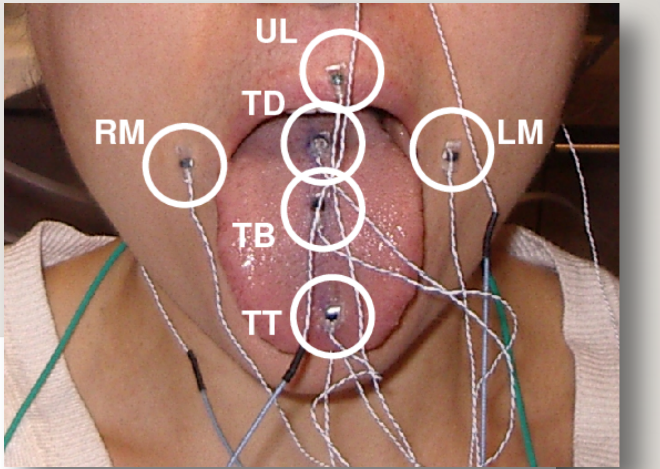
THE VOWEL TRAPEZOID

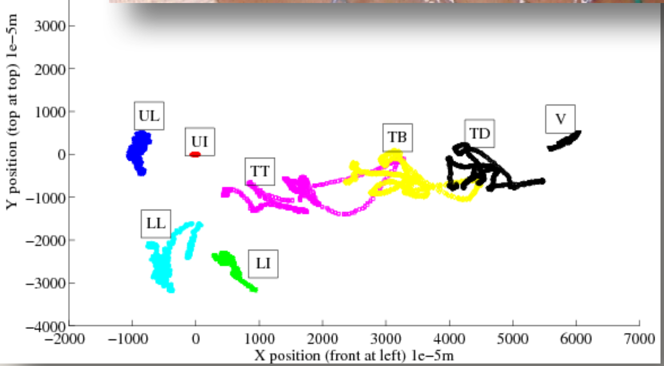


F_1 increases

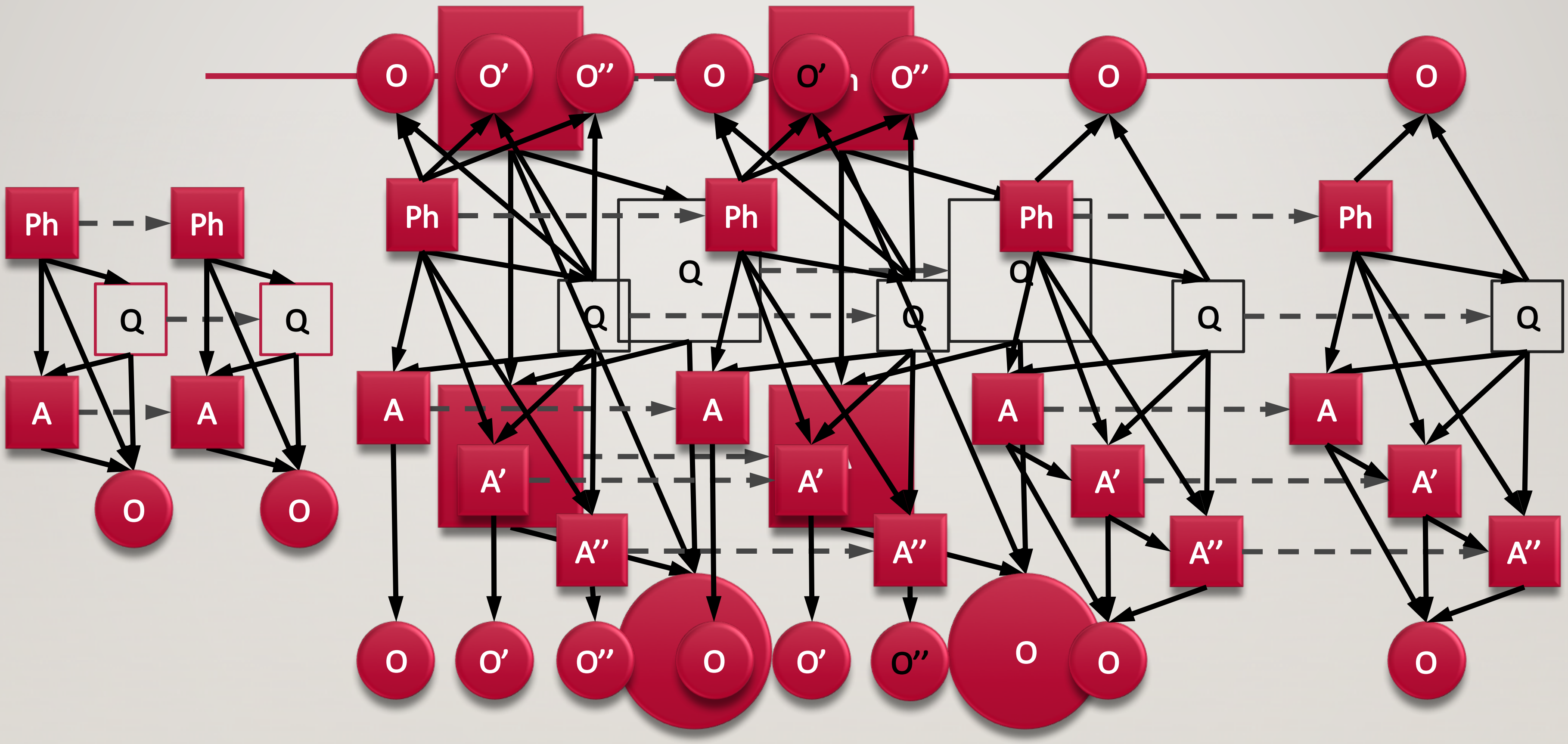
F_2 increases

ARTICULATORY DATA





AUDIO-VISUAL MODELS



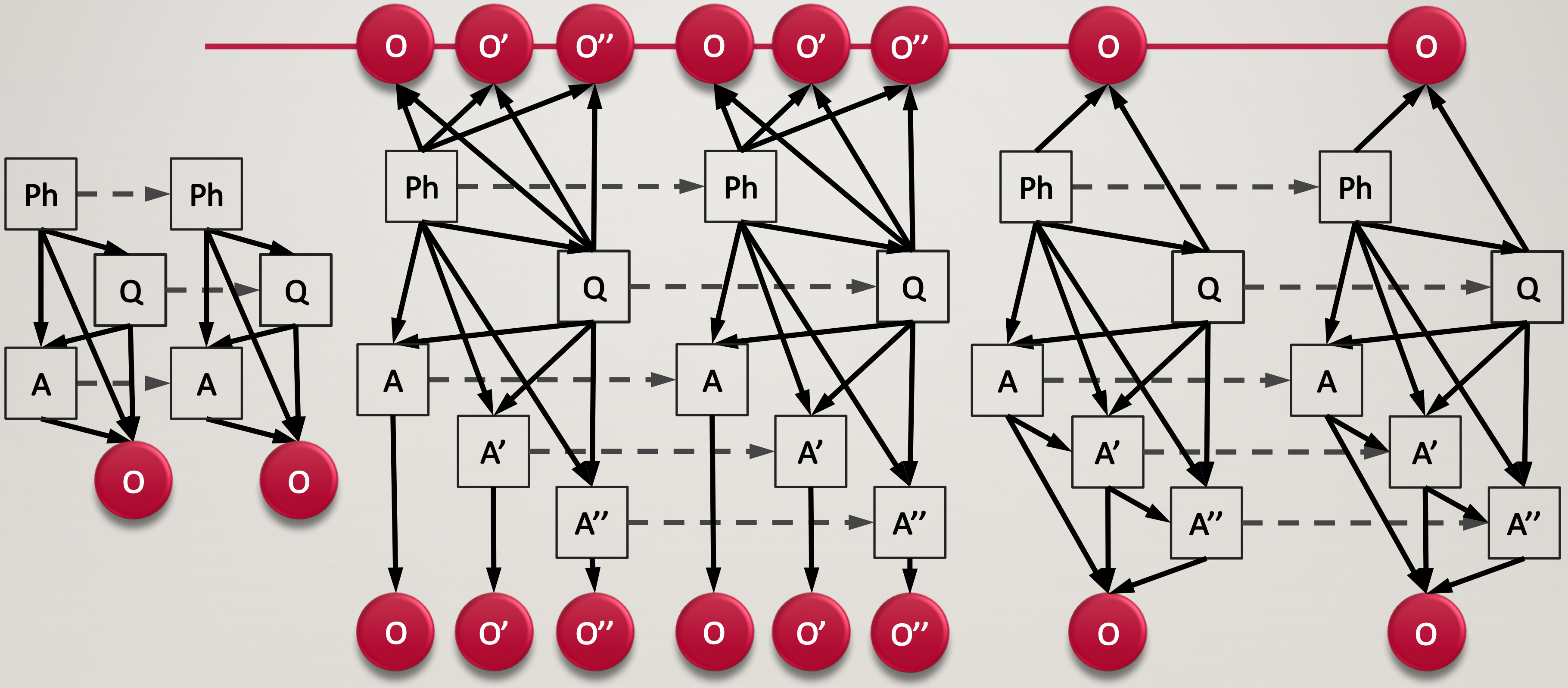
DBN-A

DBN-A2

DBN-A3

DBN = dynamic Bayes network

AUDIO-VISUAL MODELS



DBN-A

DBN-A2

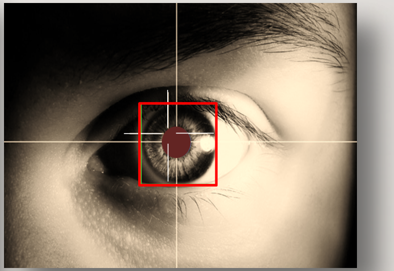
DBN-A3

AUGMENTATIVE/ALTERNATIVE COMMUNICATION (AAC)

- There are several 'physical' means to enter text.

Switches

Touch

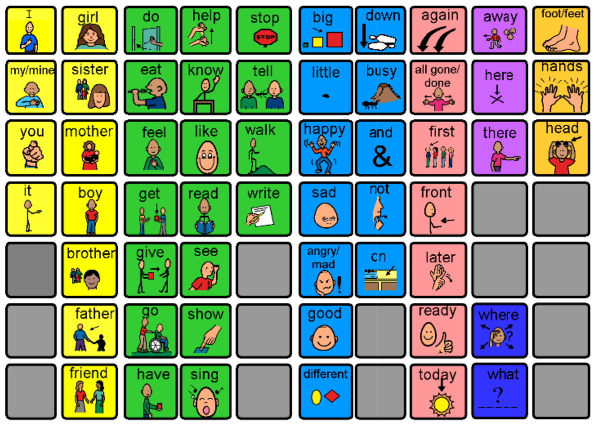


Eye

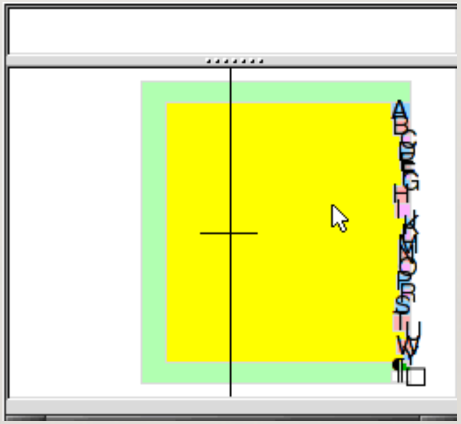
- Each can depend on the physical limits of the user.

SPEECH OUTPUT DEVICES

- There are several ‘soft’ means to enter text.
  - **Scanning** involves a **cursor** moving at a constant rate through an **array of symbols** until one is selected.



I	girl	do	help	stop	big	down	again	away	foot/feet
my/mine	sister	eat	know	tell	little	busy	all gone/ done	here	hands
you	mother	feel	like	walk	happy	and	first	there	head
it	boy	get	read	write	sad	not	front		
	brother	give	see		angry/ mad	cn	later		
	father	go	show		good		ready	where	
	friend	have	sing		different		today	what	



- **Word prediction** (with *N*-grams) can be invaluable.

SPEECH OUTPUT DEVICES

- **Rate enhancement** remains a challenge.
  - In addition to **word prediction**, **semantic compaction** and **lemmatization** can increase output to ~12 words/minute.

- AAC can **improve independent speech** in children with autism or developmental delays in 89% cases (Millar *et al.*, 2006).

- Use of AAC devices **significantly improves** quality of life, including social interaction and employment.
  - >90% unemployment rate for severely disabled individuals.

Physical perception

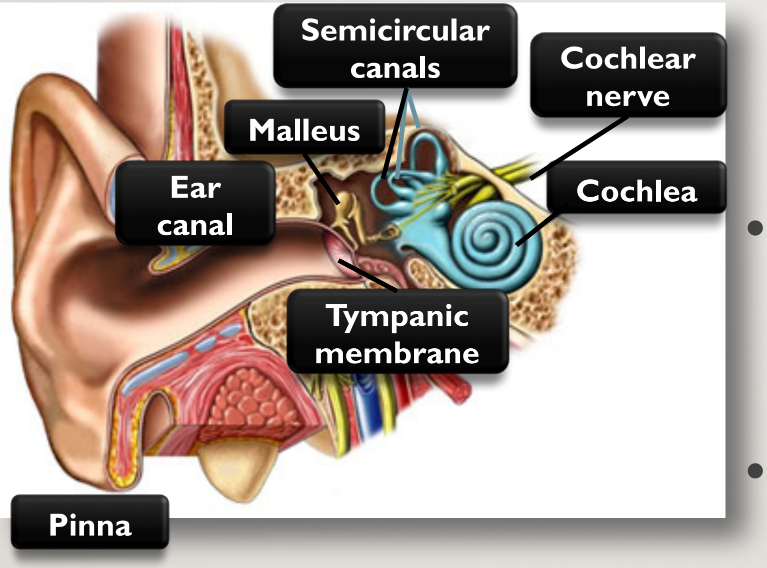
PROBLEMS OF PERCEPTION

- 0.1% of children are born with **pathological hearing loss**, including auditory nerve damage.
- ~33% of adults over 60 have **acquired hearing loss**.

- **Conductive** deafness interferes with sound to the inner ear.
- **Sensorineural** deafness involves the auditory nerve itself.

- **Tinnitus** involves noise (e.g., pulsing, hissing, ringing) that can be acute and debilitating.

THE INNER EAR

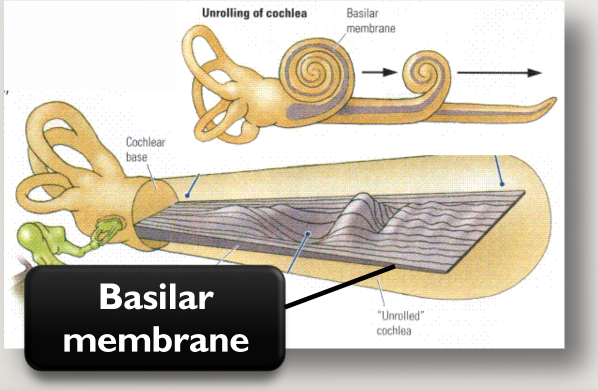


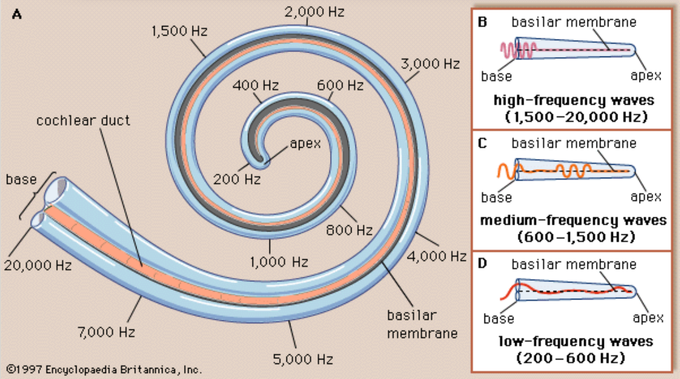
- Time-variant waves enter the ear, vibrating the **tympanic membrane**.

- This membrane causes tiny bones (incl. **malleus**) to vibrate.

- These bones in turn vibrate a structure within a shell-shaped bony structure called the **cochlea**.

THE COCHLEA AND BASILAR MEMBRANE





©1997 Encyclopaedia Britannica, Inc.

- The **basilar membrane** is covered with tiny hair-like nerves – some near the **base**, some near the **apex**.

- **High** frequencies are picked up near the base, **low** frequencies near the apex.

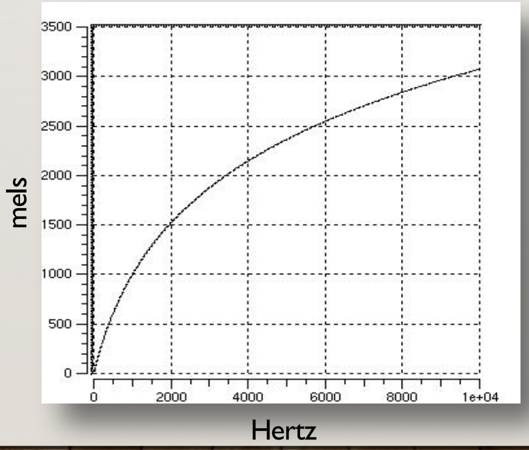
- These nerves fire when activated, and communicate to the brain.

THE MEL SCALE

- Human hearing is **not** equally sensitive to **all** frequencies.
  - We are **less** sensitive to frequencies > 1 kHz.

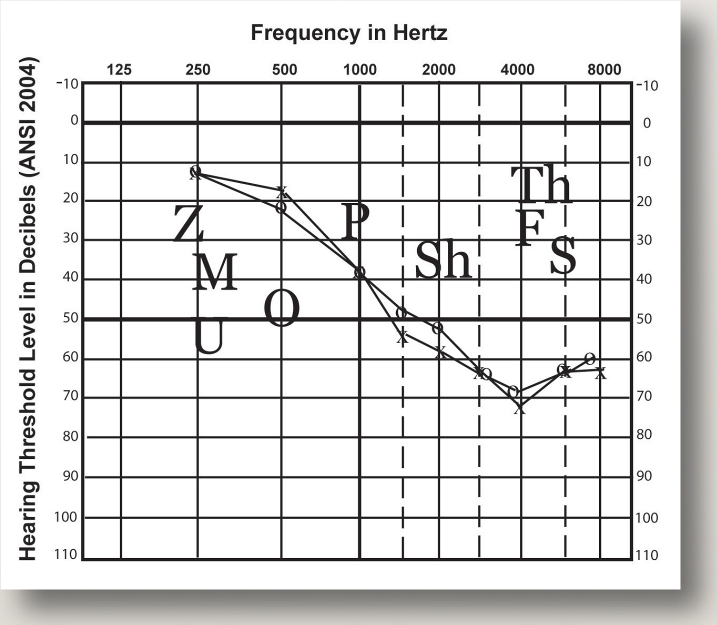
- A **mel** is a unit of pitch. Pairs of sounds which are **perceptually** equidistant in pitch are separated by an equal number of **mels**.

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$



ASSESSING PERCEPTION

- **Otologists** and **audiologists** administer audiograms, which measures hearing loss across tones (and words) at various frequencies and amplitudes.



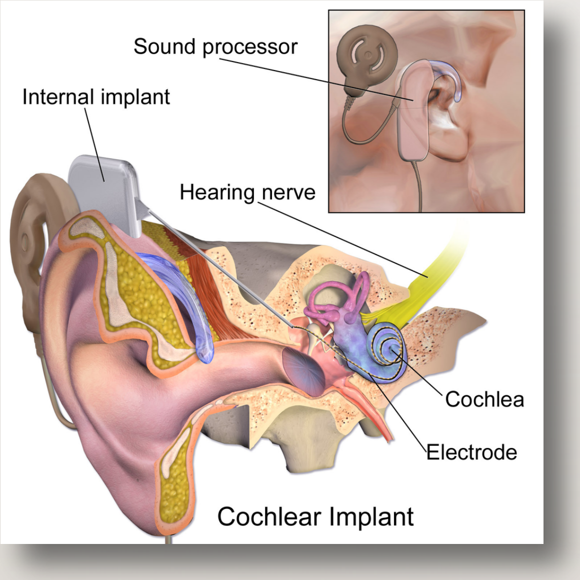
OVERCOMING PROBLEMS OF PERCEPTION

- **Hearing aids** usually **amplify** sound in certain frequencies.

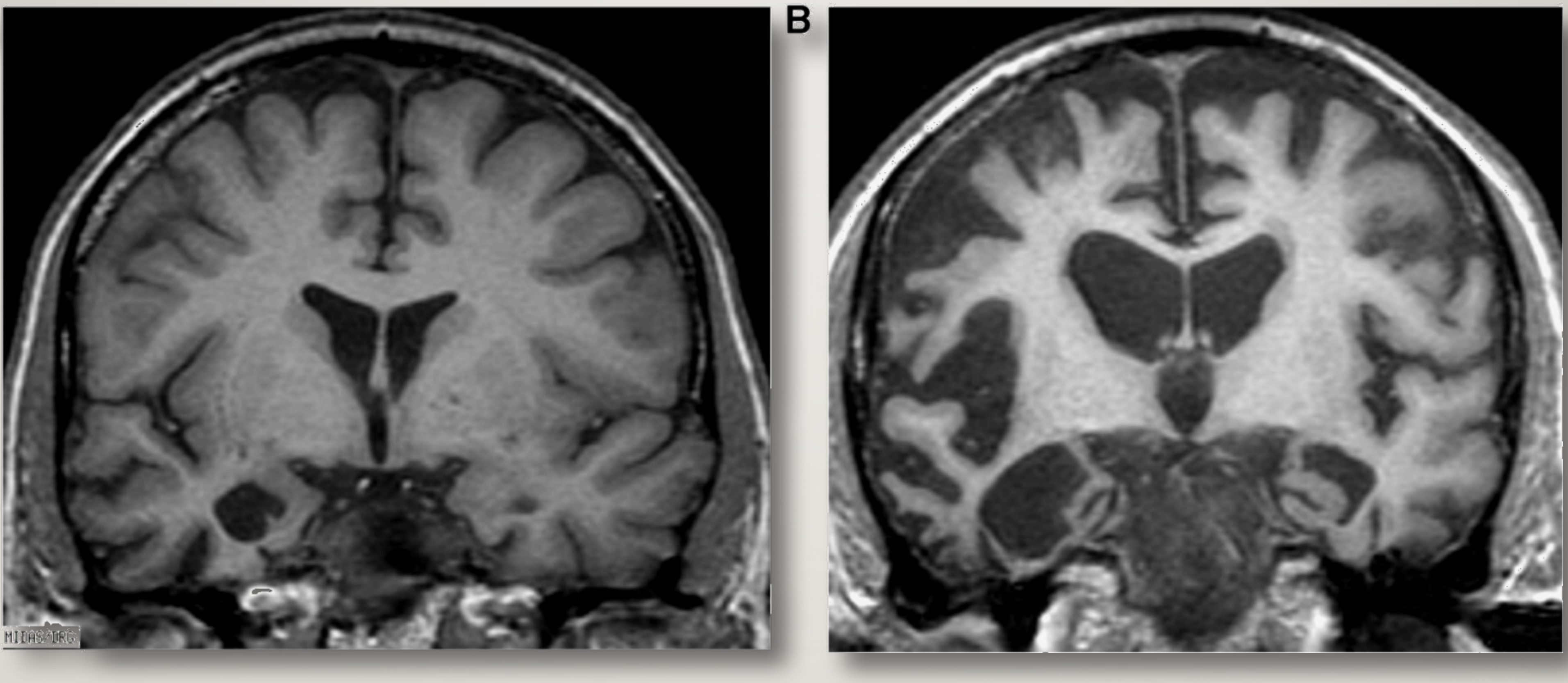
- Issues include:
  - **Occlusion effect** where person perceives "hollow" or "booming" echo-like sounds of their own voice caused by reverberations that normally pass *out* of the open air canal.
  - **Lombard effect** where people modify their own voice to compensate.
  - **Compression effect** where louder sounds need to be ‘capped’ to avoid further hearing damage.

OVERCOMING PROBLEMS OF PERCEPTION

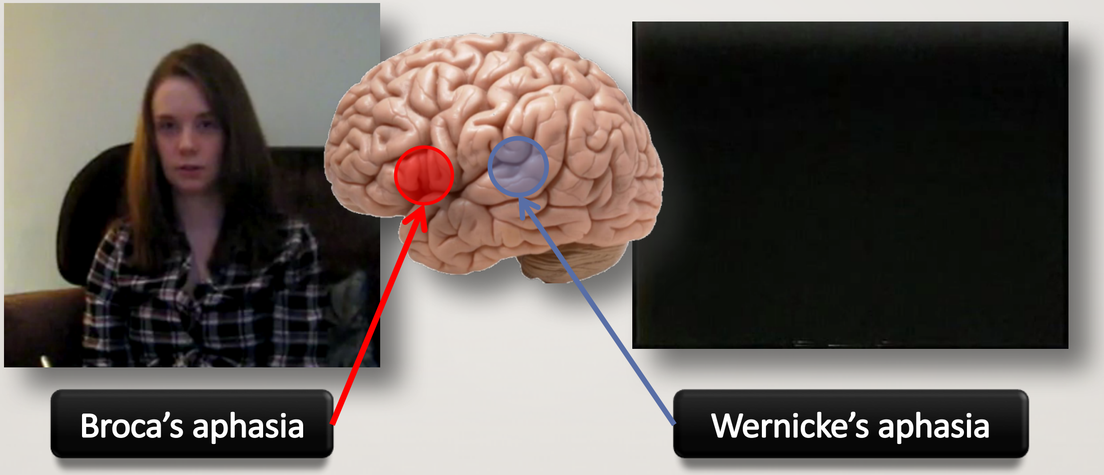
- **Cochlear implants** replace the basilar membrane and stimulate the auditory nerve directly.



Cortical atrophy and cognition



APHASIA



Broca's aphasia

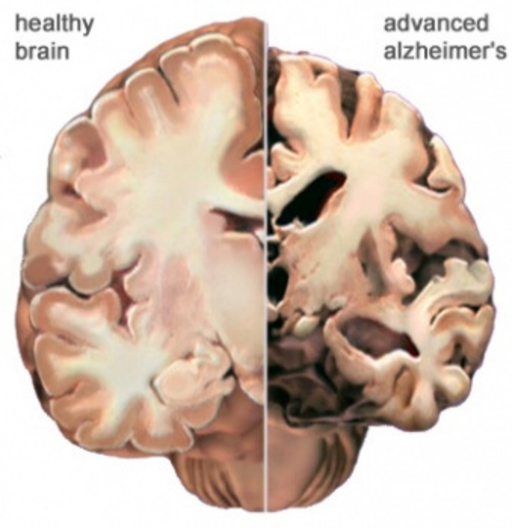
Wernicke's aphasia

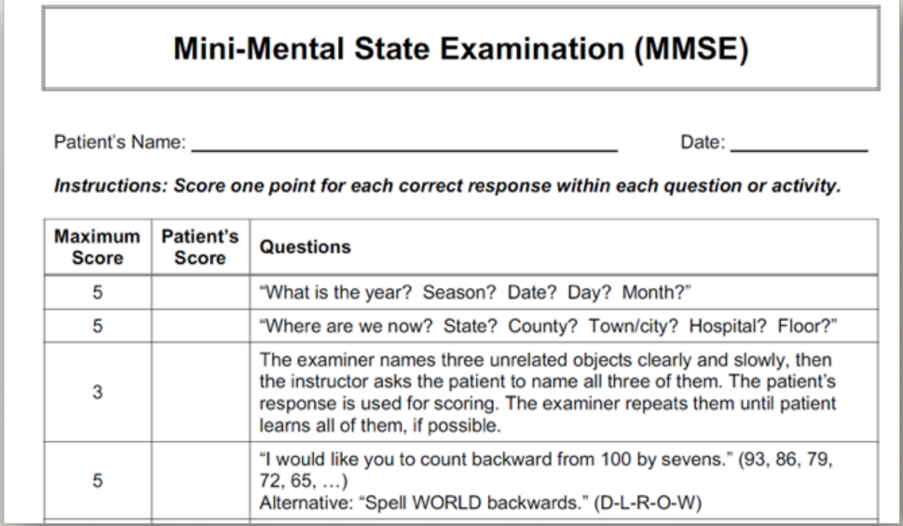
- **Reduced** hierarchical **syntax**.
- **Anomia**.
- **Reduced** "mirroring" between **observation** and **execution** of **gestures** (Rizzolatti & Arbib, 1998)**.**

- **Normal** intonation/rhythm.
- **Meaningless** words.
- '**Jumbled**' syntax.
- **Reduced** comprehension.

ALZHEIMER’S DISEASE

- **Alzheimer’s disease** (AD) is a progressive neuro-degenerative dementia characterized by **declines** in:
  - Cognitive ability (e.g., memory, reasoning),
  - Functional capacity (e.g., executive power), and
  - Social ability (e.g., linguistic abilities).





Mini-Mental State Examination (MMSE)

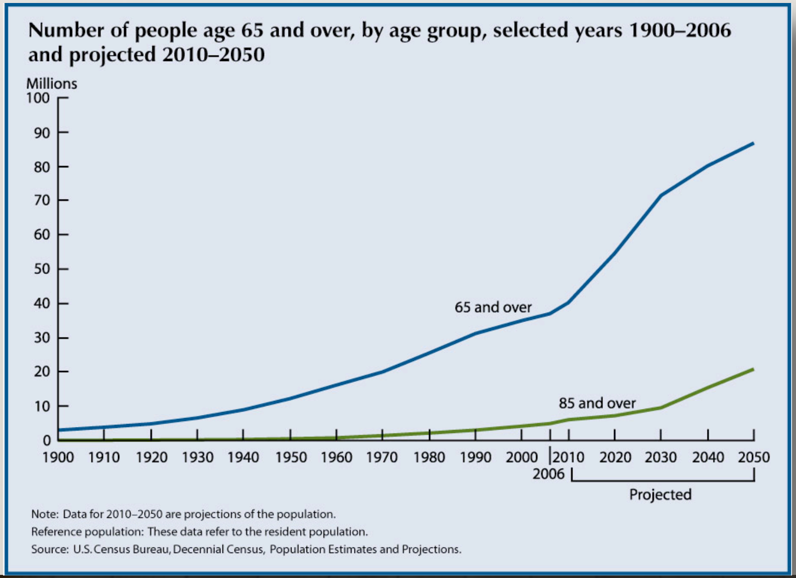
Patient’s Name: ______________________ Date: __________

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)

DEMOGRAPHIC CRISIS

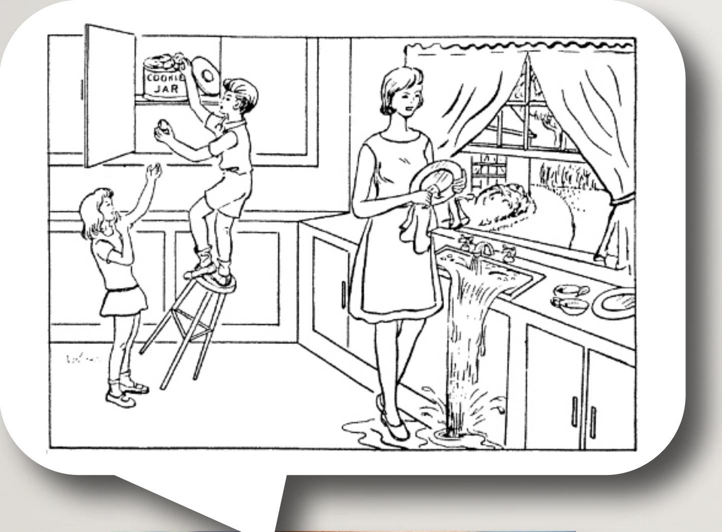
- Alzheimer's disease is pervasive (>48M people).
  - 1 in 9 adults aged ≥ 65; 1 in 3 aged ≥ 85
  - ($200B/year in care).
- As the population ages, the incidence of AD may double or triple in the next decade (Bharucha *et al.*, 2009).

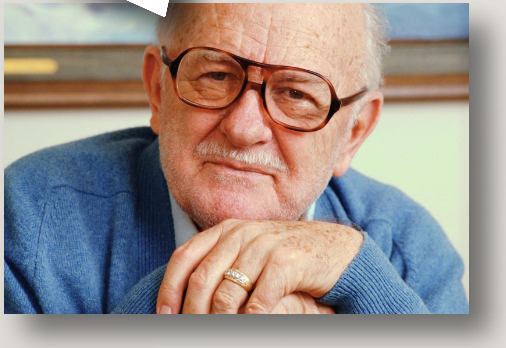


Number of people age 65 and over, by age group, selected years 1900–2006 and projected 2010–2050

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.

ASSESSING FOR ALZHEIMER'S AUTOMATICALLY





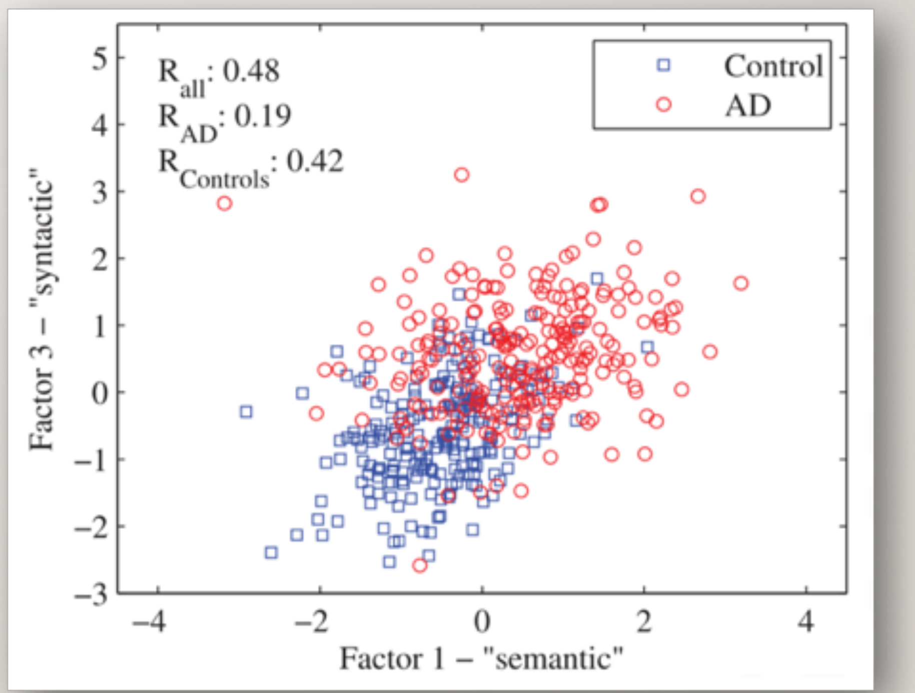
- **DementiaBank:**
  240 samples from167 participants with AD,
  233 samples from 97 controls.
  - Free-form descriptions of "Cookie Theft" (incl. audio)
  - Transcribed and annotated, e.g., with filled pauses, paraphasias, and unintelligible words.
  - **Mini-mental state exam (MMSE)**

ASSESSING FOR ALZHEIMER'S AUTOMATICALLY

Lexical	Frequency;
	Avg. word length;
	# demonstratives;
	Familiarity
	Honoré statistic

Syntactic	Parse tree depth;
	VP → VPG;
	VP → AUX VP;
	Coordinate conjunctions;
	Mean clause length

Acoustic	Phonation rate;
	Mean F2;
	Mean RPDE;
	Mean power;
	Pause::word ratio



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

NEUROPSYCHIATRIC MEASURES

- Very similar approaches can be taken for neuropsychiatric disorders such as depression, anxiety.
  - **Hamilton Depression Rating** scale: 21 questions with between 3 and 5 possible responses which increase in severity.
  - The **The Neuropsychiatric Inventory–Questionnaire (NPI-Q)** is self-administered or completed by informants about patients for whom they care.
    - Each of the 12 NPI-Q domains contains a survey question that reflects cardinal symptoms of that domain (e.g., delusions, aggression, depression, anxiety, aberrant motor, …)

DESCRIPTIVE TEXT IN EMRS AND OTHERWISE

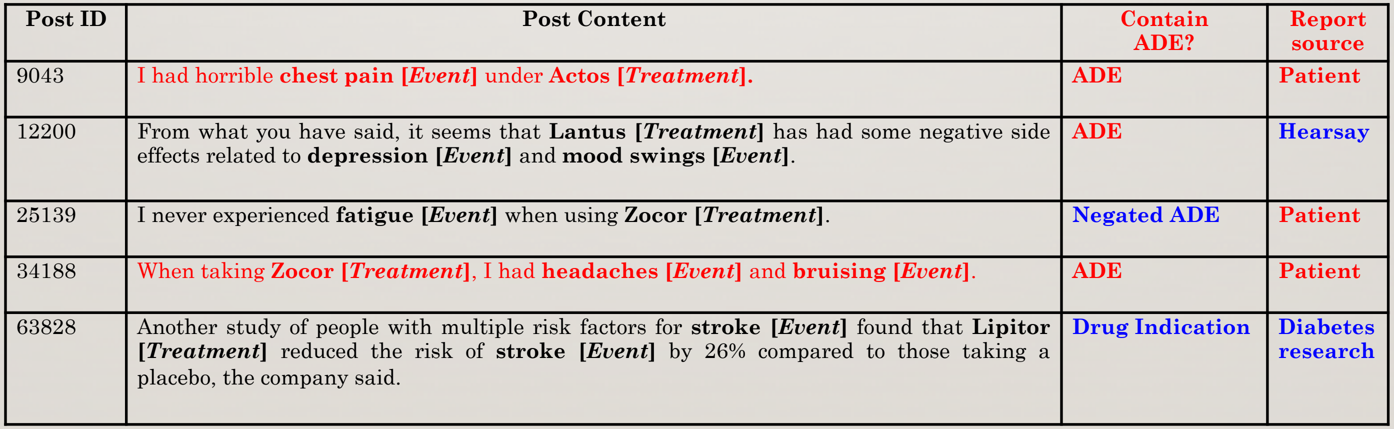
TEXT MINING IN HEALTH DATA

- Text mining
  - Information extraction
    - Named entity recognition
  - Information retrieval
- Clinical text vs. biomedical text vs. patient-centric text
  - Biomedical text: medical literature (well-formed, precise)
  - Clinical text: EMR notes (noisy, brief)
  - Patient-centric: websites for online discussion
    - E.g., /r/depression, PatientsLikeMe, DailyStrength
    - Disease, symptoms, treatments, lifestyle, emotional support

Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel Data Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology and Therapeutics*, ***91***(6), 1010–1021.

CASE STUDY:
ADVERSE DRUG EVENTS

- Extracting patient-reported **adverse drug events** (ADE) faces several challenges.
  - Topics in social media cover various **sources**, including *news, research*, *hearsay,* and *experience*. Redundant and noisy information often masks salient data.
  - Currently, extracting ADEs from comments gives in low precision due to confounding with **drug indications** (legitimate medical conditions a drug is used for) and negated ADE (contradiction or denial of experiencing ADEs).



Post ID	Post Content	Contain ADE?	Report source
9043	I had horrible **chest pain [*Event*]** under **Actos [*Treatment*].**	**ADE**	**Patient**
12200	From what you have said, it seems that **Lantus [*Treatment*]** has had some negative side effects related to **depression [*Event*]** and **mood swings [*Event*]**.	**ADE**	**Hearsay**
25139	I never experienced **fatigue [*Event*]** when using **Zocor [*Treatment*]**.	**Negated ADE**	**Patient**
34188	When taking **Zocor [*Treatment*]**, I had **headaches [*Event*]** and **bruising [*Event*]**.	**ADE**	**Patient**
63828	Another study of people with multiple risk factors for **stroke [*Event*]** found that **Lipitor [*Treatment*]** reduced the risk of **stroke [*Event*]** by 26% compared to those taking a placebo, the company said.	**Drug Indication**	**Diabetes research**

Material from H. Chen and X. Liu, University of Arizona
Liu, X., & Chen, H. (2015). Identifying adverse drug events from patient social media: A case study for diabetes. IEEE Intelligent Systems, **30**(3):44–51.

PRIOR PHARMACOVIGILANCE RESEARCH IN HEALTH SOCIAL MEDIA

Previous Studies	Test Bed	Focus	Methods			Results
			Classification	Medical Entity Recognition	Adverse Drug Event Extraction	
Leaman et al. 2010	DailyStrength.com	Adverse Drug Events	Not Applied	Lexicon based: UMLS, MedEffect, SIDER	Co-occurrence based	Precision: 78.3%; Recall: 69.9%; F-measure: 73.9%
Nikfarjam et al. 2011	DailyStrength.com	Adverse Drug Events	Not Applied	Association rule mining	Co-occurrence based	Precision: 70% recall:66.32% F-measure:67.96%
Chee et al. 2011	Health Forums from Yahoo! Groups	Drug- patient opinions	Ensemble Classifier with SVM and Naïve Bayes	Lexicon based: UMLS, MedEffect, SIDER	Not Applied	The ensemble classifier is able to identify risky drugs for FDA's scrutiny.
Benton et al. 2011	Breastcancer.org, komen.org, csn.cancer.org	Adverse Drug Events	Not Applied	Lexicon based: CHV; AERS	Co-occurrence based	Precision 35.1% Recall:77% F-measure: 52.8%
Yang et al. 2012	MedHelp	Adverse Drug Events	Not Applied	Lexicon based: CHV	Co-occurrence based	Promising to detect ADR reported by FDA.
Bian et al. 2012	Twitter	Adverse Drug Events	Machine Learning: SVM	Lexicon based: AERS	Not Applied	Accuracy: 74%; AUC value: 0.82
Mao et al. 2013	Breast cancer forums	Adverse Drug Events, Drug switching	Not Applied	Lexicon based: CHV; AERS	Co-occurrence based	Online discussions of breast cancer drugs can help to understand drug switching and discontinuation behaviors

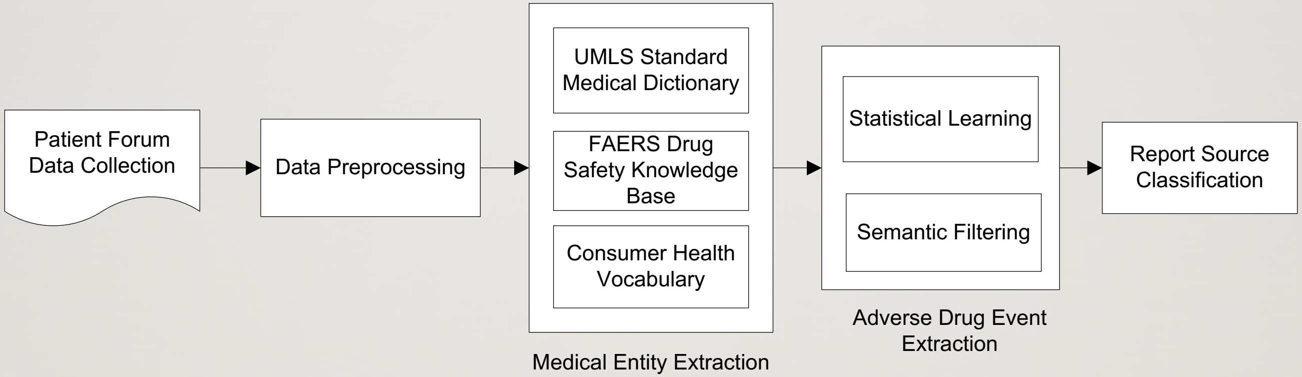
BIOMEDICAL RELATION EXTRACTION

Author	Test Bed	Focus	Approach	Method	Result
Fundel et al. 2007	Medline Abstracts	Gene protein relations	Rule-based	Rules based on dependency parse trees	F-measure of 80%
Li et al. 2008	Medline Abstracts	Gene-disease relations	Statistical Learning	Composite kernel with word, sequence kernel and tree kernel	F-measure of 70.75%
Miwa et al. 2009	Biomedical literature	Protein-protein interaction	Statistical learning	Composite kernel with BOW, Sub tree, Shortest dependency path and Graph kernel	F-measure of 60.9%
Yang et al. 2010	Biomedical literature from DIP database	protein-protein interaction	Statistical learning	Feature based: word features, keyword features, entity distance, link path features	F-measure of 57.85
Thomas et al. 2011	Medical literature	drug-drug interaction	Statistical learning	ensemble learning based on all-paths graph kernel, shortest dependency path kernel and shallow linguistic kernel	F-measure of 65.7%
Segura-Bedmar et al 2011	Biomedical text from DrugBank	drug-drug interaction	Statistical learning	shallow linguistic kernel	F-measure of 60.01%
Bui et al, 2011	Biomedical literature	protein-protein interaction	Hybrid	syntactic rules for relation detection; SVM based relation classification with lexical, distance and POS tag features	F-measure of 83.0%
Yang et al. 2012	health social forums(MedHelp)	adverse drug events	co-occurrence analysis	assumes a relation exists when two entities co-occur within 10 tokens	NA
Mao et al. 2013	Breast Cancer Patient forums	adverse drug events	co-occurrence analysis	assumes a relation exists when two entities co-occur within 20 tokens	NA

RESEARCH QUESTIONS

- How to develop an integrated & scalable framework for mining patient-reported ADEs from patient forums?
- How to augment statistical learning with health-relevant semantic filtering?
- How to identify true patient reported ADEs among noisy forum discussions?

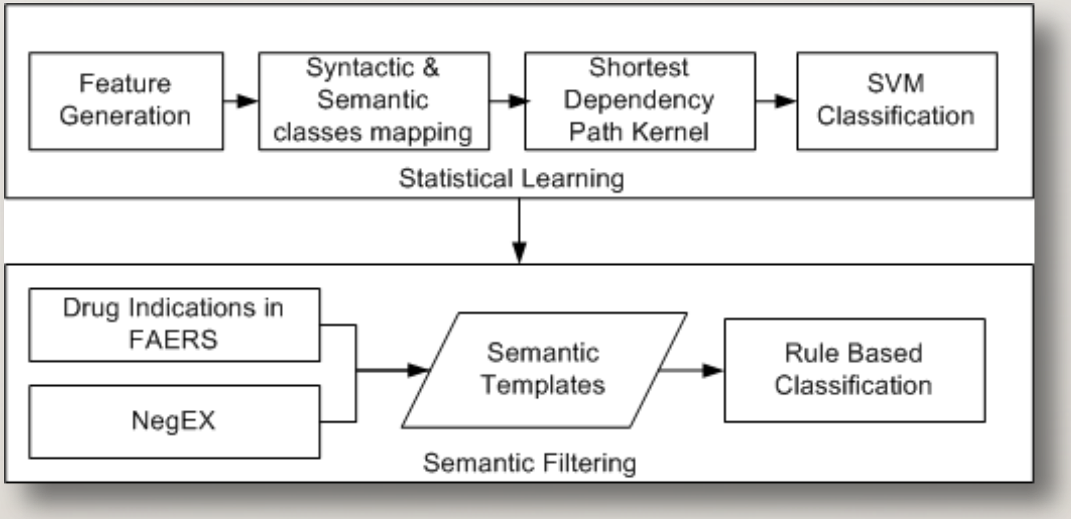
RESEARCH FRAMEWORK


Patient Forum Data Collection
Data Preprocessing
UMLS Standard Medical Dictionary
FAERS Drug Safety Knowledge Base
Consumer Health Vocabulary
Medical Entity Extraction
Statistical Learning
Semantic Filtering
Adverse Drug Event Extraction
Report Source Classification

- **Patient Forum Data Collection:** collect patient forum data through a web crawler
- **Data Preprocessing**: remove noisy text including URL, duplicated punctuation, etc.
- **Medical entity extraction**: identify treatments and adverse events discussed in forum
- **ADE extraction**: identify drug-event pairs indicating an adverse drug event based on results of medical entity extraction
- **Report source classification**: classify the source of reported events either from patient experience or hearsay

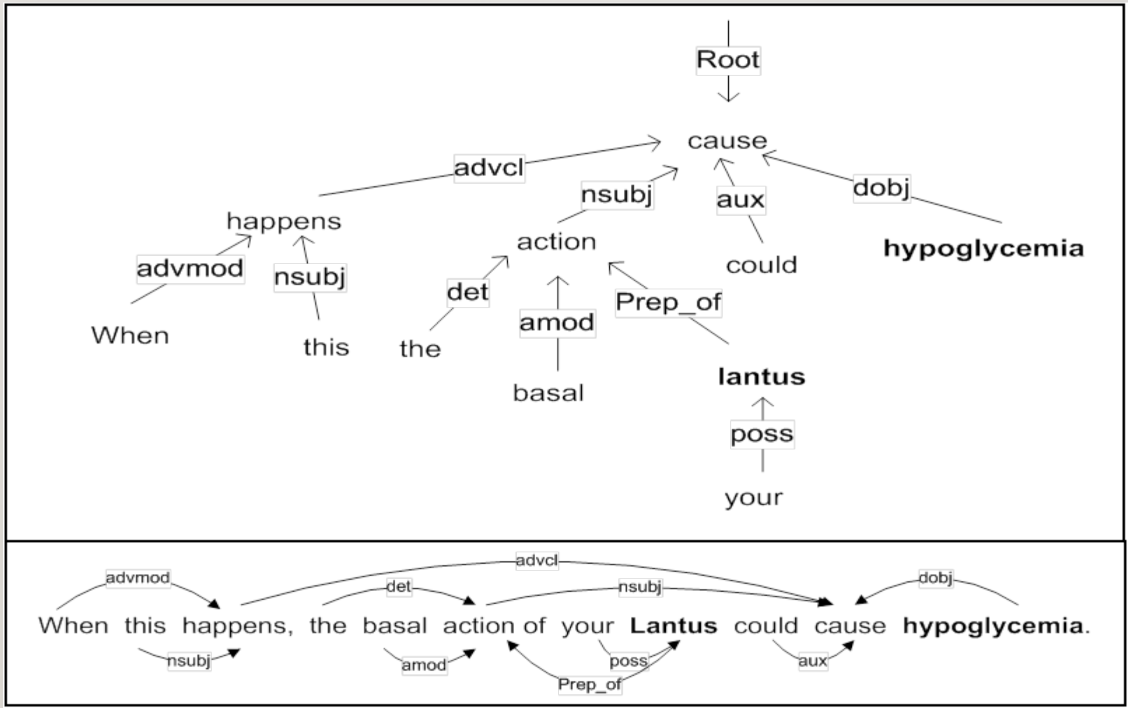
ADE EXTRACTION

- Chen and Liu incorporate kernel-based learning and semantic filtering with explicit medical and linguistic knowledge bases to identify adverse drug events in social media discussions.


Feature Generation
Syntactic & Semantic classes mapping
Shortest Dependency Path Kernel
SVM Classification
Statistical Learning
Drug Indications in FAERS
NegEX
Semantic Templates
Rule Based Classification
Semantic Filtering

ADE EXTRACTION: STATISTICAL LEARNING

Stanford Parser for dependency parsing.



http://nlp.stanford.edu/software/stanford-dependencies.shtml

ADE EXTRACTION: STATISTICAL LEARNING

Syntactic and Semantic Classes Mapping

- Word classes include part-of-speech (POS) extracted with Stanford CoreNLP packages.
- Semantic types (Event and Treatments) are used for the two ends of the shortest path.

$$\begin{bmatrix} \mathit{Hypoglycemia} \\ \mathit{NN} \\ \mathit{Noun} \\ \mathit{Event} \end{bmatrix} \times [->] \times \begin{bmatrix} \mathit{cause} \\ \mathit{VB} \\ \mathit{Verb} \end{bmatrix} \times [<-] \times \begin{bmatrix} \mathit{action} \\ \mathit{NN} \\ \mathit{Noun} \end{bmatrix} \times [<-] \times \begin{bmatrix} \mathit{Lantus} \\ \mathit{NN} \\ \mathit{Noun} \\ \mathit{Treatment} \end{bmatrix}$$

Syntactic and Semantic Classes Mapping from dependency graph

StanfordCoreNLP:http://nlp.stanford.edu/software/corenlp.shtml
Penn Tree Bank Guideline: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports

ADE EXTRACTION: SEMANTIC FILTERING

```
ALGORITHM SEMANTIC FILTERING
Input:      a relation instance i with a pair of related drug and medical events,
            R(drug, event).
Output:     The relation type.
If drug exists in FAERS:
    Get indication list for drug;
    For indication in indication list:
            If event = indication:
                    Return R(drug, event) = 'Drug Indication';
    For rule in NegEX:
            If relation instance i matches rule:
                    Return R(drug, event) = 'Negated Adverse Drug Event';
    Return R(drug, event) = 'Adverse Drug Event';
```

"indication" for a drug refers to the use of that drug for treating a particular disease. E.g., diabetes is an indication for insulin.

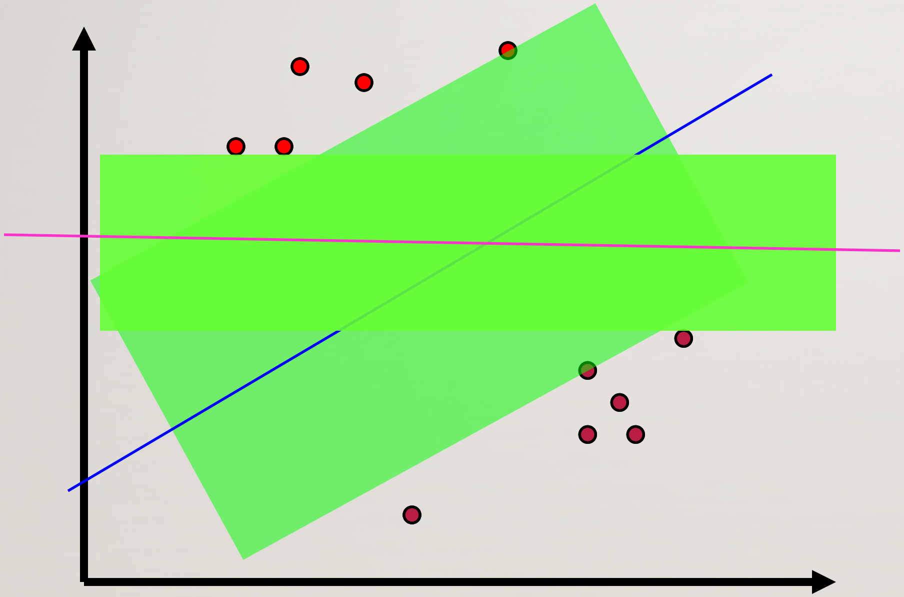
FAERS: FDA Adverse Event Reporting System

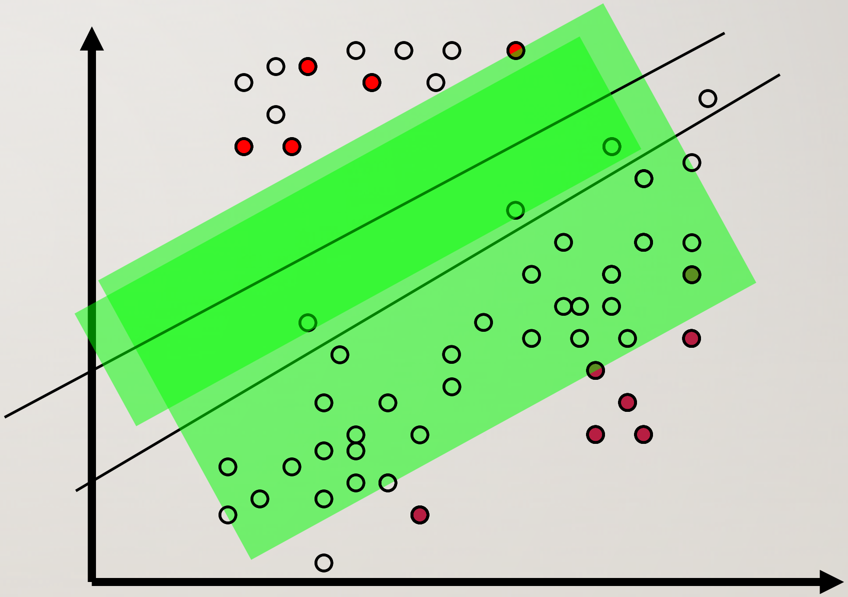
NegEx: University of Pittsburgh tool to detect negated terms from clinical text.

REPORT SOURCE CLASSIFICATION

- Chen and Liu adopted BOW features and transductive support vector machines for classification.
  - Semi-supervised classification methods such as transductive SVMs, which leverage labeled and unlabeled data, can build the model with a small set of annotated data and conduct transductive inference in unlabeled data (Joachims 1999).
  - This is more scalable than traditional supervised methods because of the large amount of unlabeled data available in social media.

TRANSDUCTIVE SVMS



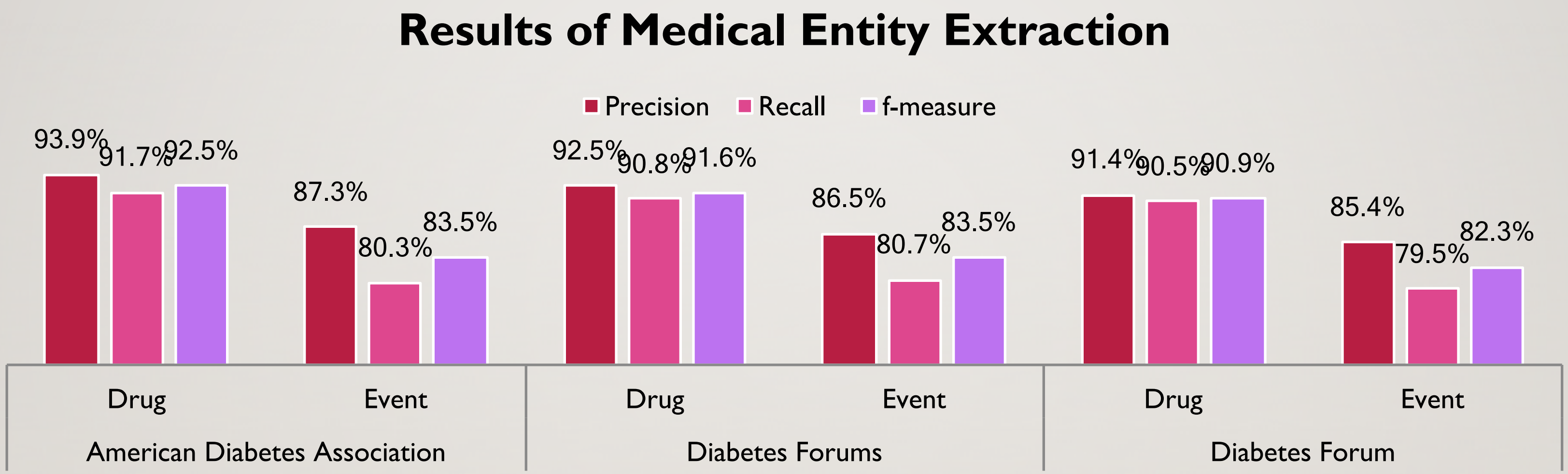


EVALUATION

- The test bed was developed from three major diabetes patient forums in the United States, i.e., the **American Diabetes Association** online community, **Diabetes Forums**, and **Diabetes Forum**.
  - Diabetes affects **25.8 million** people. A large number of treatments exist to help control glucose and prevent organ damage from hyperglycemia. However, many treatments have a number of adverse events that range from minor to serious.

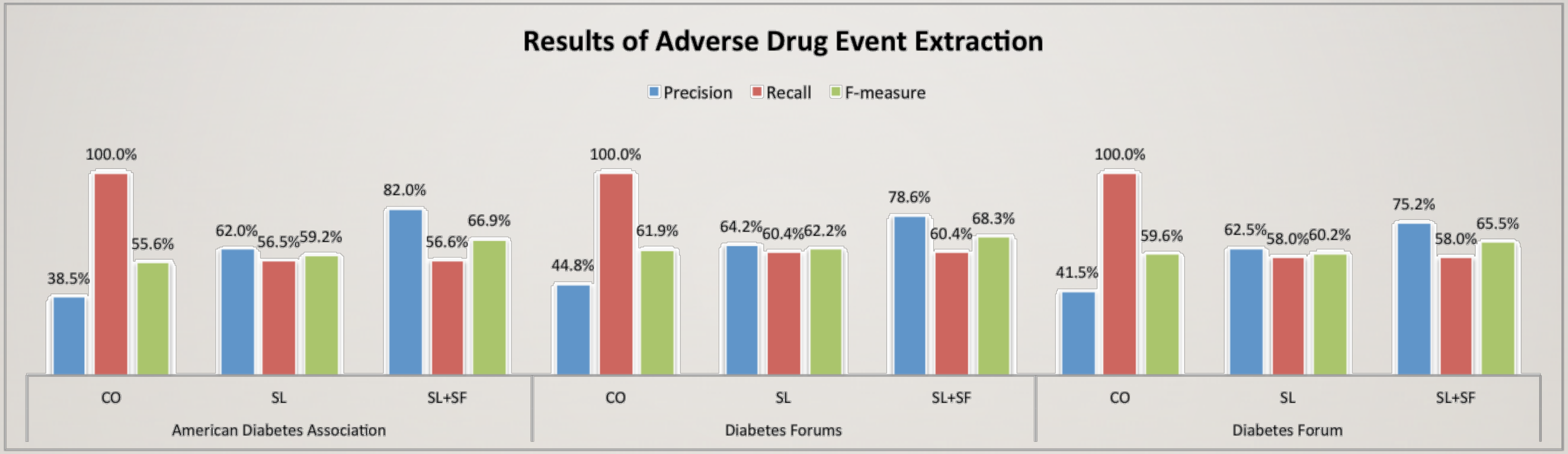
Forum Name	Number of Posts	Number of Topics	Number of Member Profiles	Time Span	Total Number of Sentences
American Diabetes Association	184,874	26,084	6,544	2009.2-2012.11	1,348,364
Diabetes Forums	568,684	45,830	12,075	2002.2-2012.11	3,303,804
Diabetes Forum	67,444	6,474	3,007	2007.2-2012.11	422,355

EVALUATION ON MEDICAL ENTITY EXTRACTION



- The performance of their system (F-measure, 82%-92%) beat prior studies (F-measure 73.9%), which had applied UMLS and MedEffect to extract adverse events from DailyStrength (Leaman et al., 2010).

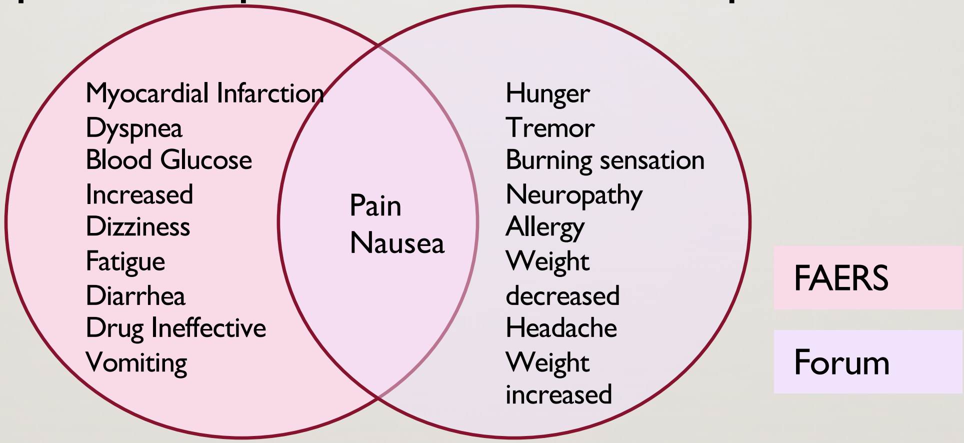
EVALUATION ON ADVERSE DRUG EVENT EXTRACTION



- Compared to co-occurrence based approach (CO), statistical learning (SL) increased precision from around 40% to above 60% while recall dropped from 100% to around 60%. **F-measure of SL is better than CO by 0.3-3.6% ($p = 0.029$).**
- Semantic filtering (SF) further improved precision from 60% to about 80%. **F-measure of SF-SL is better than CO by 6-12% ($p = 0.022$).**

ANALYSIS OF DOCUMENTED VS. FOUND ADVERSE EVENTS

- Differences between Top 10 adverse events from FDA's AERS reports and patient social forum reports



Myocardial Infarction
Dyspnea
Blood Glucose Increased
Dizziness
Fatigue
Diarrhea
Drug Ineffective
Vomiting

Pain
Nausea

Hunger
Tremor
Burning sensation
Neuropathy
Allergy
Weight decreased
Headache
Weight increased

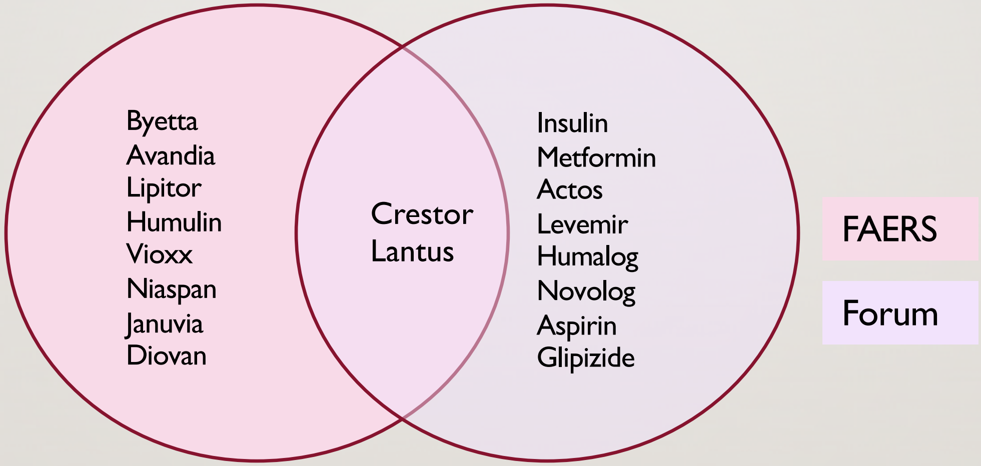
FAERS

Forum

- Top reported adverse events from FAERS contain more severe events such as myocardial infarction
- Forum reports have more minor events but closely related to diabetes daily management such as weight changes and hunger.

ANALYSIS OF DOCUMENTED VS. FOUND TOP REPORTED DRUGS

- Differences between Top 10 reported drugs from FDA's AERS reports and patient social forum reports



- Top reported medications from FAERS contain more drugs known to cause severe adverse events such as **Byetta, Avandia** and **Vioxx**.
- Top reported medications from forums have more common diabetes treatments such as **insulin** and **Metformin**, reflecting the popularity of the treatments among patients.

NLP TOOLS I

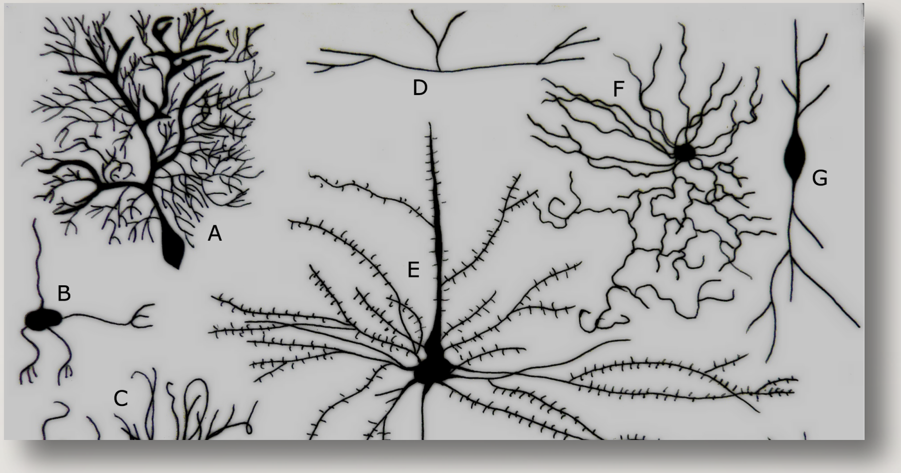
- clinical Text and Knowledge Extraction System (cTAKES): cTAKES is built on top of Apache UIMA, and is composed of sets of UIMA processors that are assembled together into pipelines. Some of the processors are wrappers for Apache OpenNLP components, and some are custom built. cTAKES was developed at the Mayo Clinic, and is distributed by the Open Health NLP Consortium.
- Health Information Text Extraction (HITEX): HITEx was developed as part of the i2b2 project. It is a rule-based NLP pipeline based on the GATE framework.
- Computational Language and Education Research toolkit (cleartk): cleartk has been developed at the University of Colorado at Boulder, and provides a framework for developing statistical NLP components in Java. It is built on top of Apache UIMA.

NLP TOOLS 2

- NegEx (NegEx): NegEx is a tool developed at the University of Pittsburgh to detect negated terms from clinical text. The system utilizes trigger terms as a method to determine likely negation scenarios within a sentence.
- ConText (ConText): ConText is an extension to NegEx, and is also developed by the University of Pittsburgh. ConText extends NegEx to not only detect negated concepts, but to also find temporality (recent, historical or hypothetical scenarios) and who the experiencer is (patient or other) of the concept.
- National Library of Medicine's MetaMap (MetaMap): MetaMap is a comprehensive concept tagging system which is built on top of the Unified Medical Language System (UMLS). It requires an active UMLS Metathesaurus License Agreement for use. The program may execute by itself, although there has been done some work to create a UIMA Wrapper to allow MetaMap to act as a UIMA component.

NLP TOOLS 3

- (MedEx): MedEx processes free-text clinical records to recognize medication names and signature information, such as drug dose, frequency, route, and duration. Use is free with a UMLS license. It is a standalone application for Linux and Windows.
- SecTag – section tagging hierarchy (SecTag): SecTag recognizes note section headers using NLP, Bayesian, spelling correction, and scoring techniques. The link here includes the SQL and CSV files for the section terminologies. Use is free with either a UMLS or LOINC license.
- Stanford Named Entity Recognizer (NER): Stanford's NER is a Conditional Random Field sequence model, together with well-engineered features for Named Entity Recognition in English and German.
- Stanford CoreNLP (CoreNLP): Stanford CoreNLP is an integrated suite of natural language processing tools for English in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference.



NEURAL MODELS OF WORD REPRESENTATION

With material from Yoshua Bengio, Fréderic Godin, Richard Socher, and others (where indicated).

WORDS

- Given a corpus with D (e.g., $= 100K$) unique words, the **classical binary approach** is to uniquely assign **each word** with an index in D-dimensional vectors ('one-hot' representation).

0	0	0	0	..	0	1	0	…	0

D

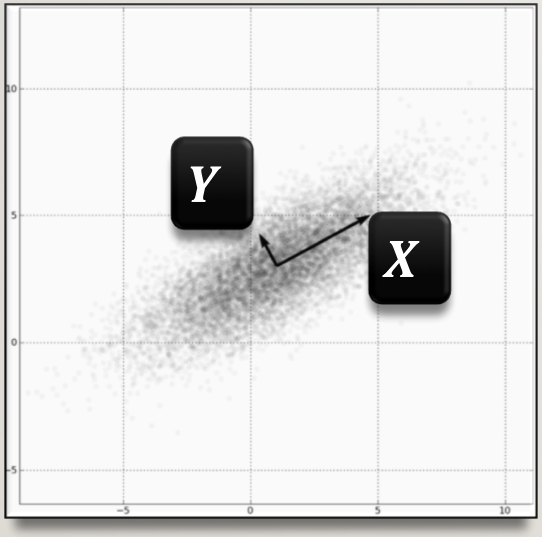
- Classic **word-feature representation** assigns **features** to each index.
  - E.g., 'VBG', 'positive', 'age-of-acquisition'.

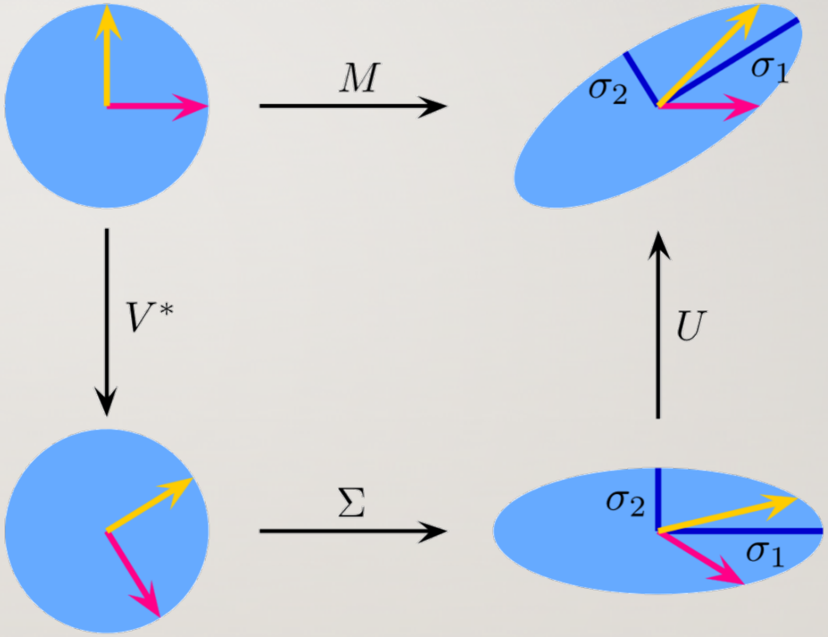
1	0.8	4.5	0.81	…	99

$d \ll D$

- Is there a way to *learn* something *like* the latter?

SINGULAR VALUE DECOMPOSITION



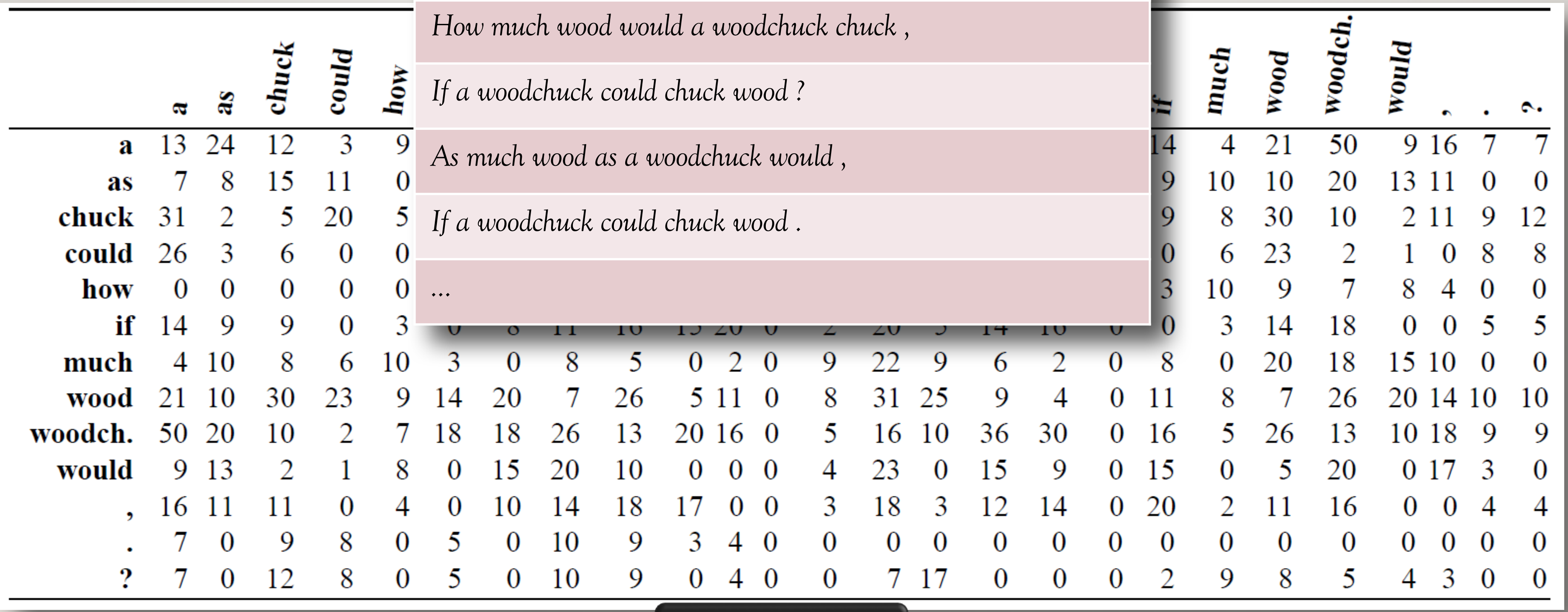


$$M = U \cdot \Sigma \cdot V^*$$

PCA

SVD

SINGULAR VALUE DECOMPOSITION



	a	as	chuck	could	how														if	much	wood	woodch.	would	,	.	?
a	13	24	12	3	9														14	4	21	50	9	16	7	7
as	7	8	15	11	0														9	10	10	20	13	11	0	0
chuck	31	2	5	20	5														9	8	30	10	2	11	9	12
could	26	3	6	0	0														0	6	23	2	1	0	8	8
how	0	0	0	0	0														3	10	9	7	8	4	0	0
if	14	9	9	0	3	0	8	11	16	15	20	0	2	20	5	14	16	0	0	3	14	18	0	0	5	5
much	4	10	8	6	10	3	0	8	5	0	2	0	9	22	9	6	2	0	8	0	20	18	15	10	0	0
wood	21	10	30	23	9	14	20	7	26	5	11	0	8	31	25	9	4	0	11	8	7	26	20	14	10	10
woodch.	50	20	10	2	7	18	18	26	13	20	16	0	5	16	10	36	30	0	16	5	26	13	10	18	9	9
would	9	13	2	1	8	0	15	20	10	0	0	0	4	23	0	15	9	0	15	0	5	20	0	17	3	0
,	16	11	11	0	4	0	10	14	18	17	0	0	3	18	3	12	14	0	20	2	11	16	0	0	4	4
.	7	0	9	8	0	5	0	10	9	3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
?	7	0	12	8	0	5	0	10	9	0	4	0	0	7	17	0	0	0	2	9	8	5	4	3	0	0

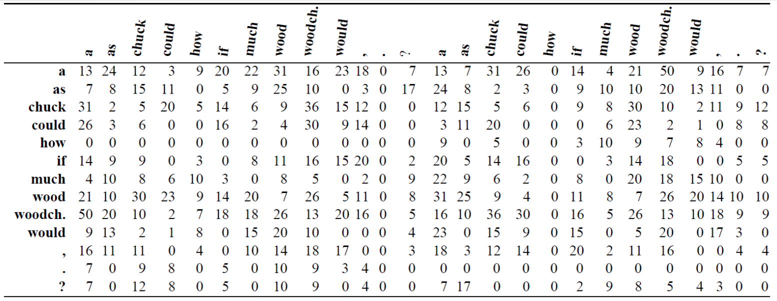
Corpus
How much wood would a woodchuck chuck ,
If a woodchuck could chuck wood ?
As much wood as a woodchuck would ,
If a woodchuck could chuck wood .
…

Co-occurrence

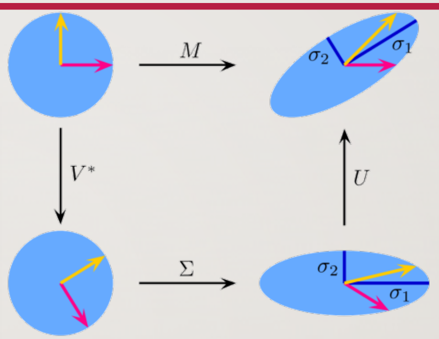
Rohde *et al.* (2006) An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. *Communications of the ACM* **8**:627-633.

SINGULAR VALUE DECOMPOSITION

$M =$



	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	13	24	12	3	9	20	22	31	16	23	18	0	7	13	7	31	26	0	14	4	21	50	9	16	7	7
as	7	8	15	11	0	5	9	25	10	0	3	0	17	24	8	2	3	0	9	10	10	20	13	11	0	0
chuck	31	2	5	20	5	14	6	9	36	15	12	0	0	12	15	5	6	0	9	8	30	10	2	11	9	12
could	26	3	6	0	0	16	2	4	30	9	14	0	0	3	11	20	0	0	0	6	23	2	1	0	8	8
how	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	5	0	0	3	10	9	7	8	4	0	0
if	14	9	9	0	3	0	8	11	16	15	20	0	2	20	5	14	16	0	0	3	14	18	0	0	5	5
much	4	10	8	6	10	3	0	8	5	0	2	0	9	22	9	6	2	0	8	0	20	18	15	10	0	0
wood	21	10	30	23	9	14	20	7	26	5	11	0	8	31	25	9	4	0	11	8	7	26	20	14	10	10
woodch.	50	20	10	2	7	18	18	26	13	20	16	0	5	16	10	36	30	0	16	5	26	13	10	18	9	9
would	9	13	2	1	8	0	15	20	10	0	0	0	4	23	0	15	9	0	15	0	5	20	0	17	3	0
,	16	11	11	0	4	0	10	14	18	17	0	0	3	18	3	12	14	0	20	2	11	16	0	0	4	4
.	7	0	9	8	0	5	0	10	9	3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
?	7	0	12	8	0	5	0	10	9	0	4	0	0	7	17	0	0	0	2	9	8	5	4	3	0	0



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

$$A = U_{[:,1:2]}\Sigma_{[1:2,1:2]}$$

$U =$

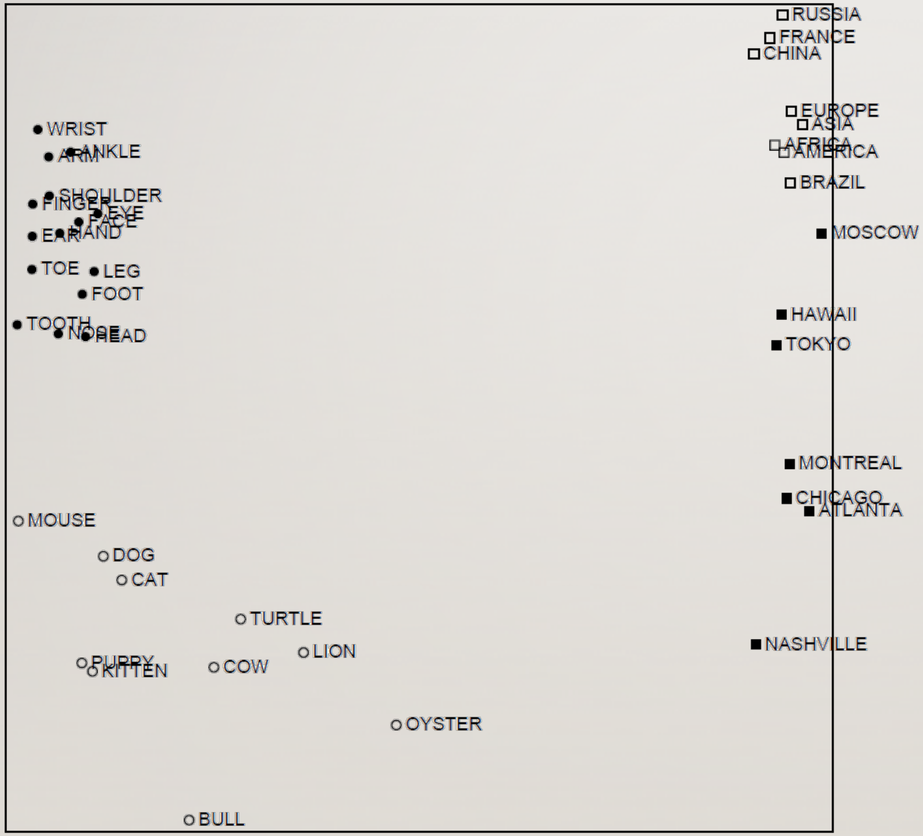
a	-0.44	-0.30	0.57	0.58	…
as	-0.13	-0.33	-0.59	0	…
chuck	-0.48	-0.51	-0.37	0	…
could	-0.70	0.35	0.15	-0.58	…
…	…	…	…	…	…

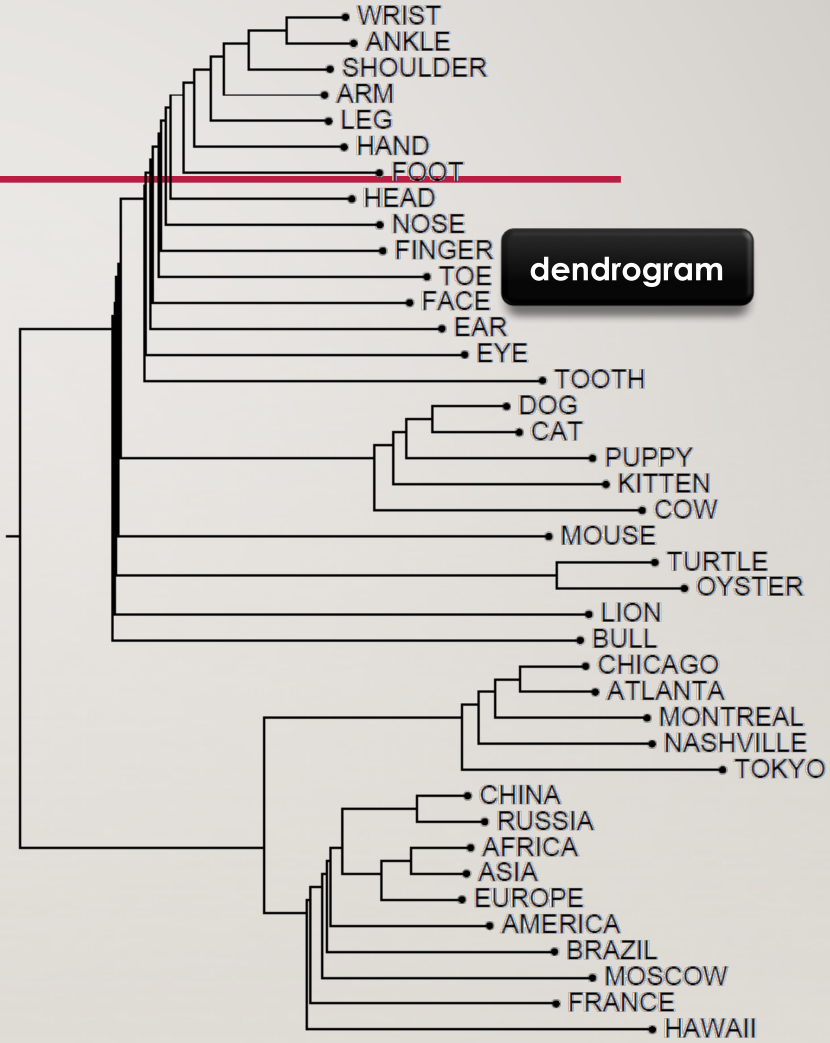
$\Sigma =$

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

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

SINGULAR VALUE DECOMPOSITION

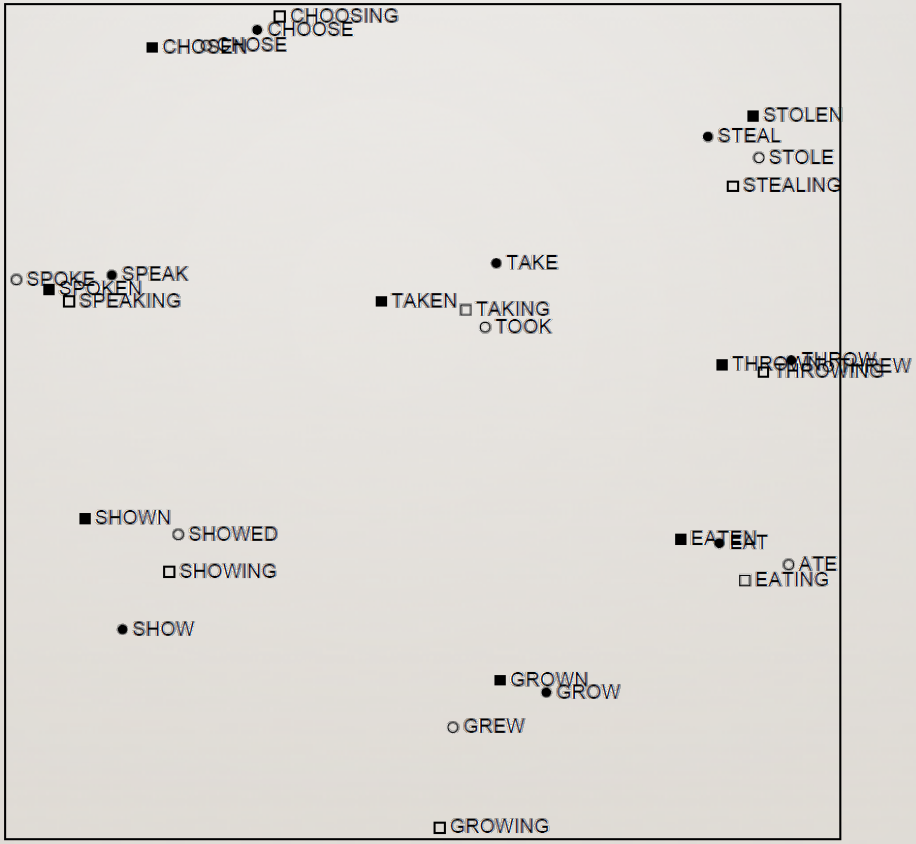




dendrogram

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

SINGULAR VALUE DECOMPOSITION



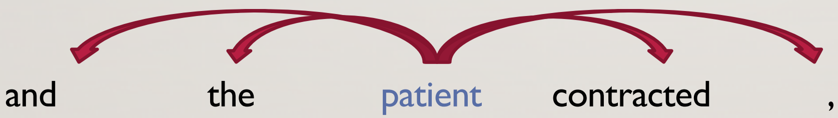
Rohde *et al.* (2006) An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. *Communications of the ACM* **8**:627-633.

PROBLEMS WITH SVD; INTRO TO WORD2VEC

- SVD: Computational costs scale quadratically with M.
  ‘Hard’ to incorporate new words.

- Word2vec: Don’t capture co-occurrence directly
  Just try to predict surrounding words.

$$P(w_{t+1} = contracted | w_t = patient)$$



and the patient contracted ,

and the wife contracted ,

…

Here, we predict the center word given the context.

‘softmax’

$$P(w_o | w_i) = \frac{\exp(V_{w_o}^{\intercal} v_{w_i})}{\sum_{w=1}^{W} \exp(V_w^{\intercal} v_{w_i})}$$

Where v_w is the ‘input’ vector for word w,
and V_w is the ‘output’ vector for word w.

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

LEARNING WORD REPRESENTATIONS

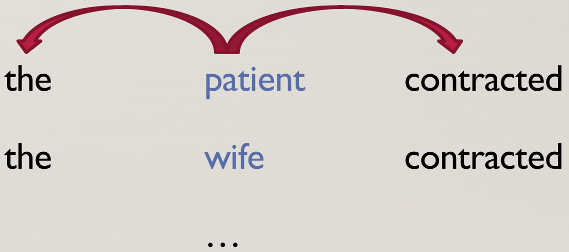
- Word representations can be learned using the following **objective function:**

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c<j<c,j\neq 0} \log P(w_{t+j}|w_t)$$

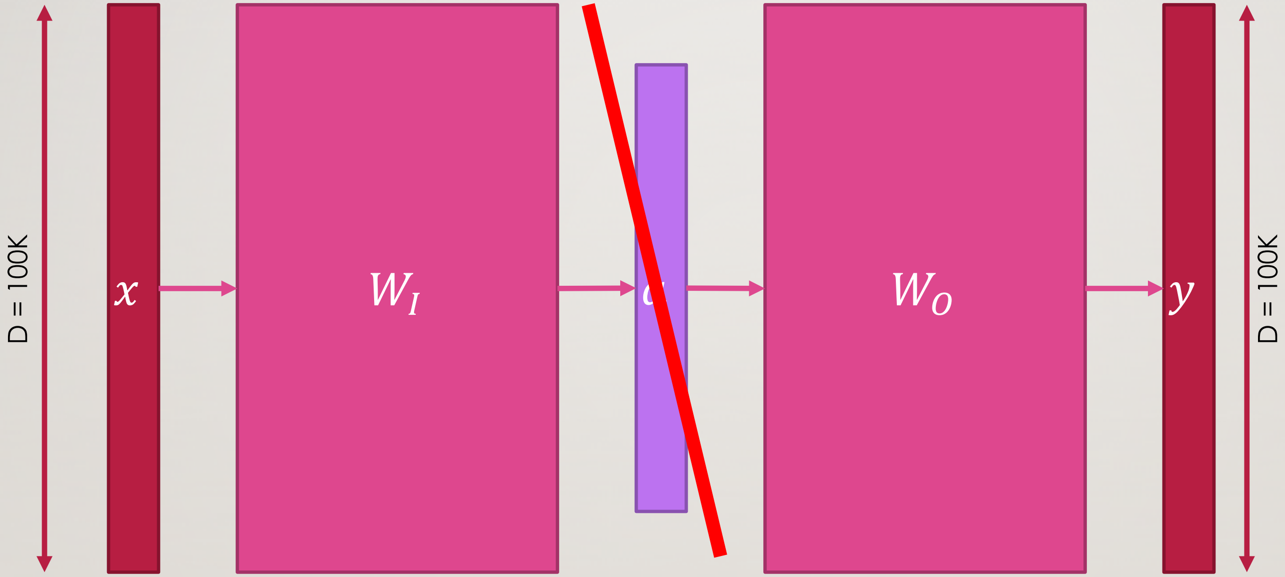
where w_t is the t^{th} word in a sequence of T words.

- This is closely related to **word prediction.**
  - *"words of a feather flock together."*
  - *"you shall know a word by the company it keeps."*

- J.R. Firth (1957)



LEARNING WORD REPRESENTATIONS



Note: we now have **two** representations of each word:
v_w comes from the rows of W_I
V_w comes from the cols of W_O

$[0,0,0, \dots 1, \dots, 0]$
patient

the	patient	contracted
the	wife	contracted
“outside”	**“inside”**	**“outside”**
	…	

$[0,1,0, \dots, 0, \dots, 0]$ the

$[0,0,1, \dots, 0, \dots, 0]$ contracted

Continuous bag of words (CBOW)

USING WORD REPRESENTATIONS

Without a latent space,

patient = $[0,0,0, \ldots, 0,1,0, \ldots, 0]$, &

wife = $[0,0,0, \ldots, 0,0,1, \ldots, 0]$ so

Similarity = $\cos(x, y) = 0.0$

Transform

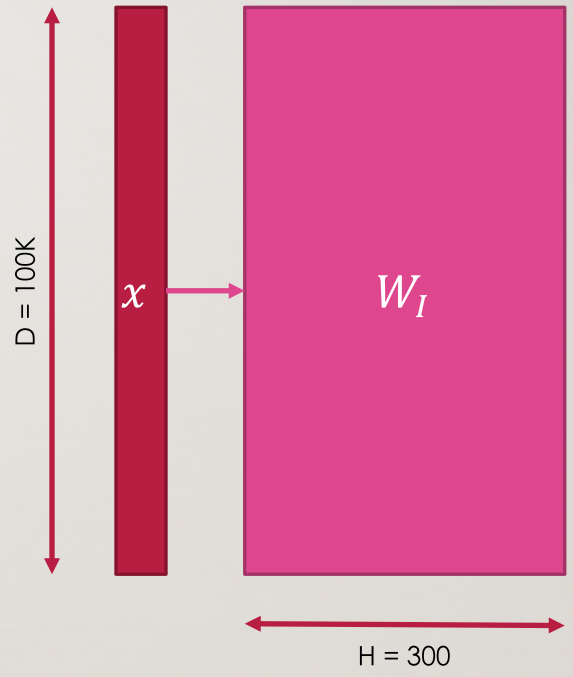
$v_w = xW_1$

In latent space,

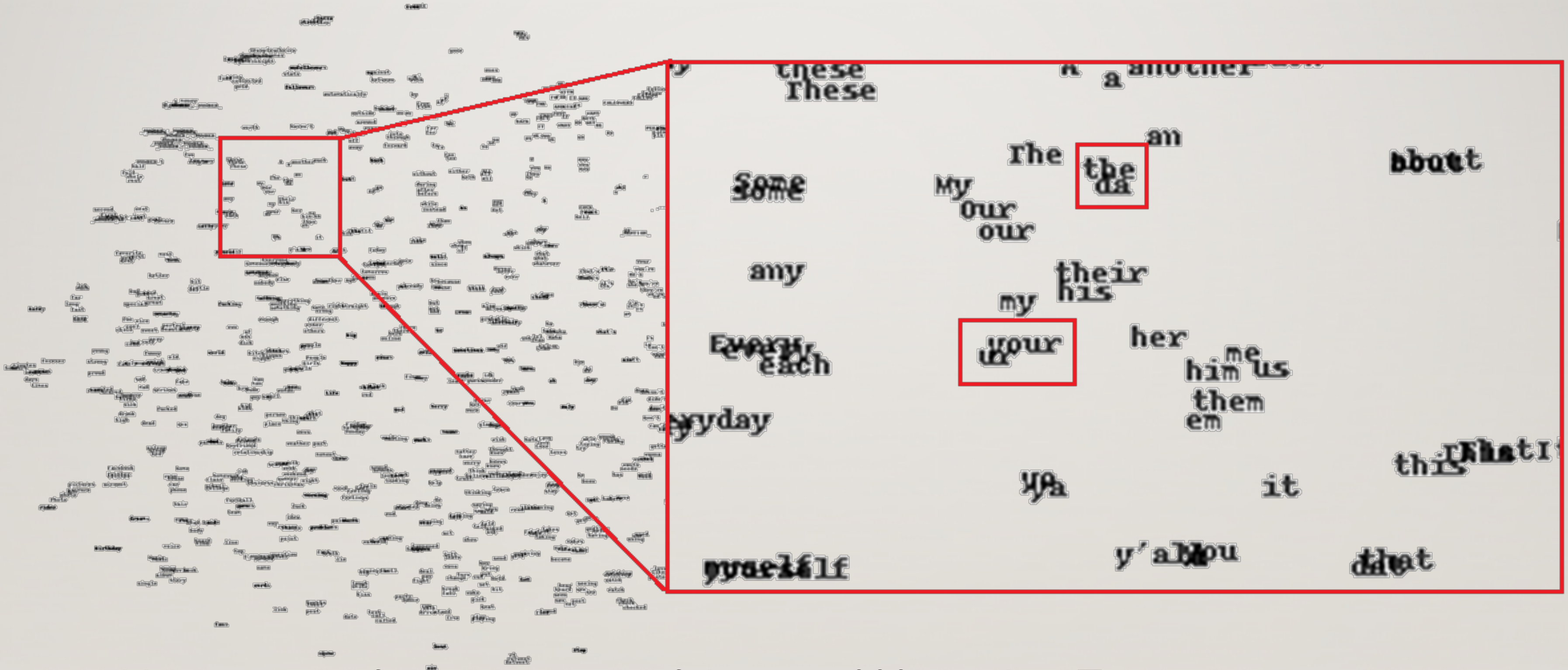
patient = $[0.8,0.69,0.4, \ldots, 0.05]_H$, &

wife = $[0.9,0.7,0.43, \ldots, 0.05]_H$ so

Similarity = $\cos(x, y) = 0.67$



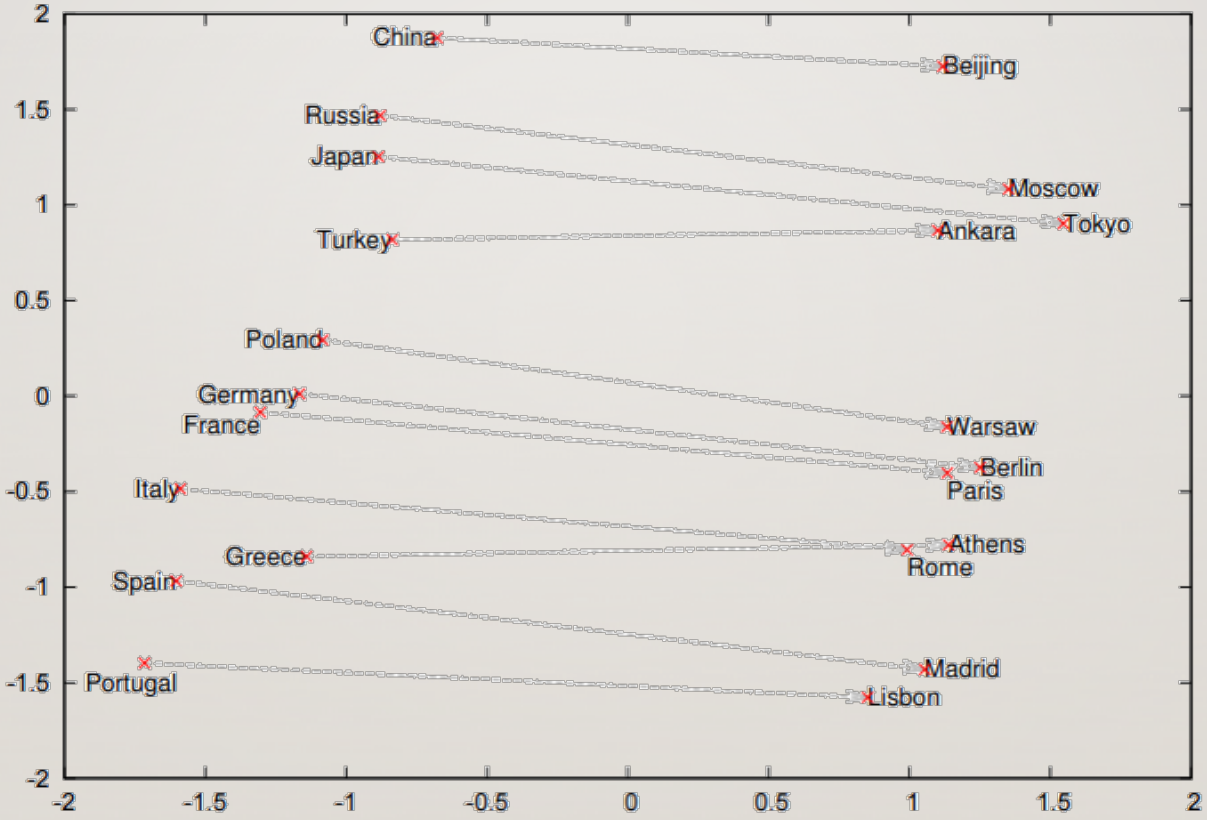
LINGUISTIC REGULARITIES IN WORD-VECTOR SPACE



Visualization of a vector space of the top 1000 words in Twitter

Trained on 400 million tweets having 5 billion words

LINGUISTIC REGULARITIES IN WORD-VECTOR SPACE



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

LINGUISTIC 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

Ha ha – isn’t that nice? But it’s easy to cherry-pick…

ACTUALLY LEARNING

First, let's define what our parameters are. Given H-dimensional vectors, and V words:

$$\theta = \begin{bmatrix} v_a \\ v_{aardvark} \\ \vdots \\ v_{zymurgy} \\ V_a \\ V_{aardvark} \\ \vdots \\ V_{zymurgy} \end{bmatrix} \in \mathbb{R}^{2VH}$$

ACTUALLY LEARNING

Many options. Gradient descent is popular.

We want to optimize

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c<j<c, j\neq 0} \log P(\overset{\text{“outside”}}{w_{t+j}} | \overset{\text{“inside”}}{w_t})$$

And we want to update vectors $V_{w_{t+j}}$ then v_{w_t} within θ

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

so we'll need to take the derivative of the (log of the) softmax function:

$$P(w_{t+j}|w_t) = \frac{\exp(V_{w_{t+j}}^\intercal v_{w_t})}{\sum_{w=1}^{W} \exp(V_w^\intercal v_{w_t})}$$

ACTUALLY LEARNING

We need to take the derivative of the (log of the) softmax function:

$$\frac{\delta}{\delta v_{w_t}} \log P\left(w_{t+j} \middle| w_t\right) = \frac{\delta}{\delta v_{w_t}} \log \frac{\exp(V_{w_{t+j}}^\intercal v_{w_t})}{\sum_{w=1}^{W} \exp(V_w^\intercal v_{w_t})}$$

$$= \frac{\delta}{\delta v_{w_t}} \log \exp\left(V_{w_{t+j}}^\intercal v_{w_t}\right) - \log \sum_{w=1}^{W} \exp(V_w^\intercal v_{w_t})$$

$$= V_{w_{t+j}} - \frac{\delta}{\delta v_{w_t}} \log \sum_{w=1}^{W} \exp(V_w^\intercal v_{w_t})$$

[apply the chain rule $\frac{\delta f}{\delta v_{w_t}} = \frac{\delta f}{\delta z} \frac{\delta z}{\delta v_{w_t}}$]

$$= V_{w_{t+j}} - \sum_{w=1}^{W} p(w|w_t) V_w$$

More details: http://arxiv.org/pdf/1411.2738.pdf

SMELL THE GLOVE

GloBal VeCtors for Word representations is a popular alternative to word2vec.

Trained on the non-zero entries of a global word-word co-occurrence matrix.

$$J(\theta) = \frac{1}{2} \sum_{ij} f\left(P_{ij}\right)\left(w_i \cdot \widetilde{w_j} - \log P_{ij}\right)^2$$

Fast and scalable.

Same kinds of benefits

Words close to *frog*

3. litoria

4. leptodactylidae

5. rana

7. eleutherodactylus

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

RESULTS – NOTE THEY'RE ALL EXTRINSIC

- Bengio *et al* 2001, 2003: beating N-grams on small datasets (Brown & APNews), but much slower.
- Schwenk *et al* 2002,2004,2006: beating state-of-the-art large-vocabulary speech recognizer using deep & distributed NLP model, with real-time speech recognition.
- Morin & Bengio 2005, Blitzer *et al* 2005, Mnih & Hinton 2007,2009: better & faster models through hierarchical representations.
- Collobert & Weston 2008: reaching or beating state-of-the-art in multiple NLP tasks (**SRL**, POS, NER, chunking) thanks to unsupervised pre-training and multi-task learning.
- Bai et al 2009: ranking & semantic indexing (info retrieval).

SENTIMENT ANALYSIS

Traditional bag-of-words approach used dictionaries of **happy** and **sad** words, simple counts, and regression or simple binary classification.

But consider these blog posts:

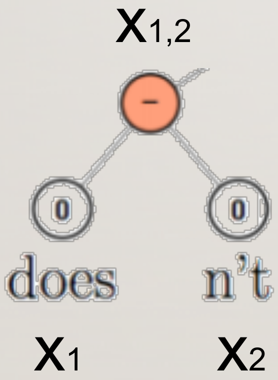
	Hamilton Rating for Depression
Best day of my life	0/50
Sunny and pleasant, despite some brief rain	8/50
I’m glad this stupid sunny day is over	19/50

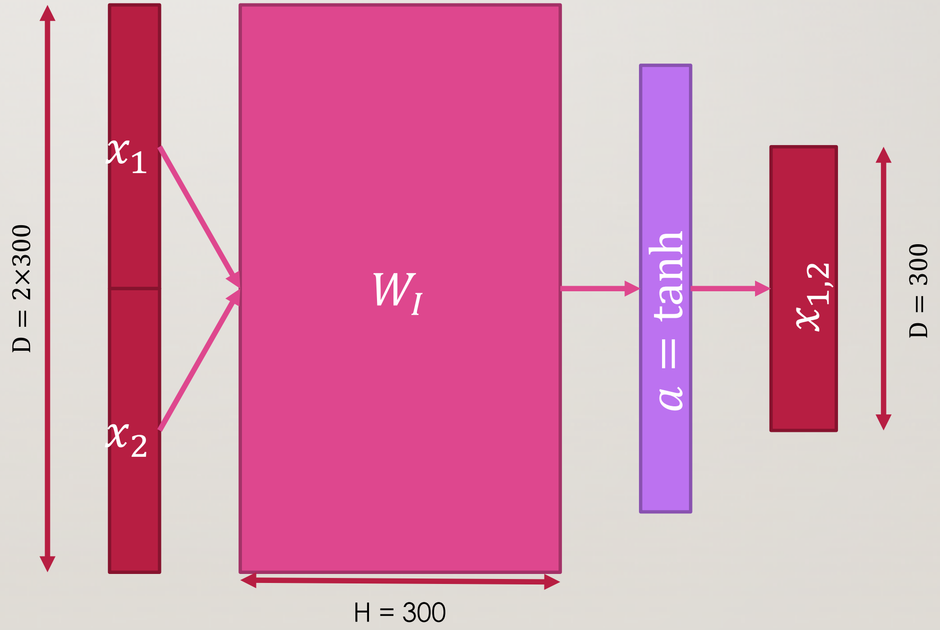
HAM-D:

0-7 = Normal
8-13 = Mild Depression
14-18 = Moderate Depression 19-22 = Severe Depression
≥ 23 = Very Severe Depression

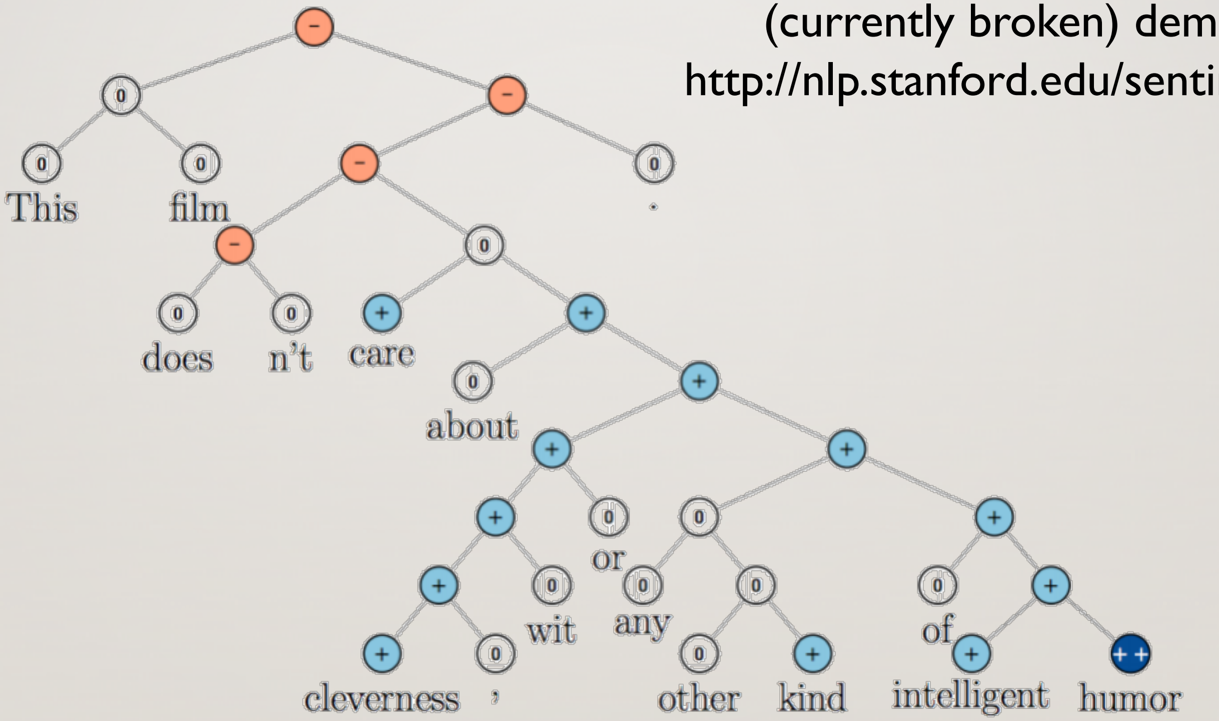
SENTIMENT ANALYSIS

We can combine **pairs** of words into **phrase** structures. Similarly, we can combine phrase and word structures hierarchically for classification.





TREE-BASED SENTIMENT ANALYSIS

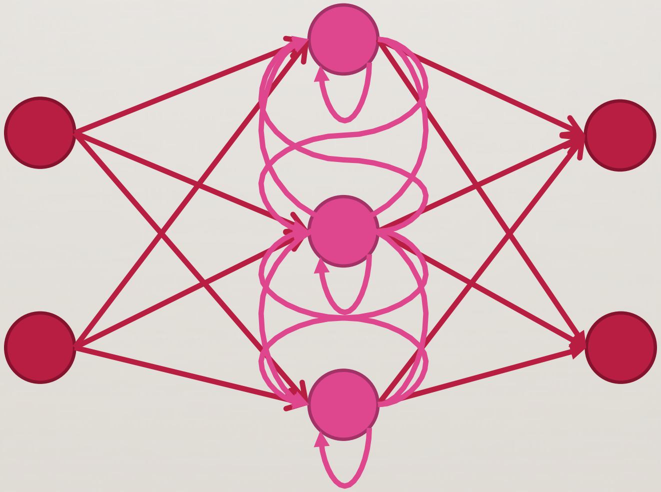


(currently broken) demo:
http://nlp.stanford.edu/sentiment/

RECURRENT NEURAL NETWORKS (RNNS)

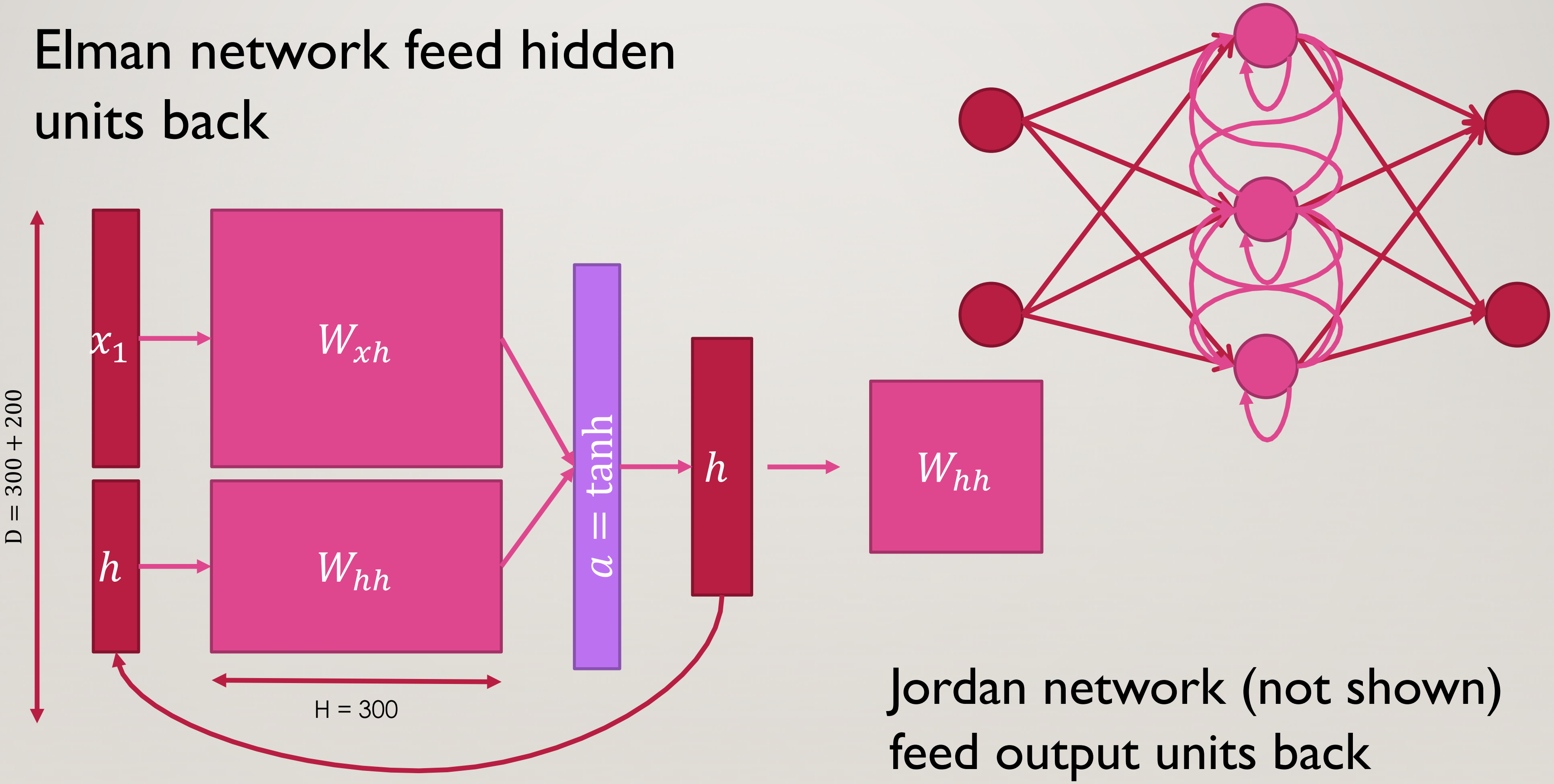
An RNN has feedback connections in its structure so that it ‘remembers’ n previous inputs, when reading in a sequence.

(e.g., can use current word input with hidden units from the previous word)



RECURRENT NEURAL NETWORKS (RNNS)

Elman network feed hidden units back



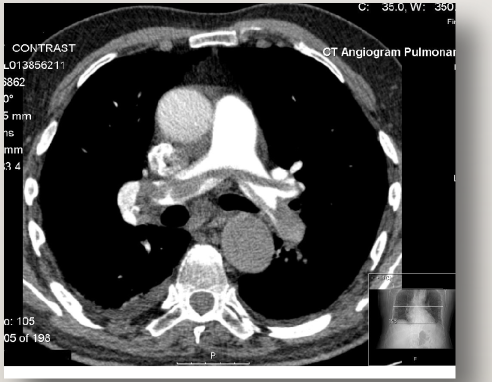
Jordan network (not shown) feed output units back

VISION

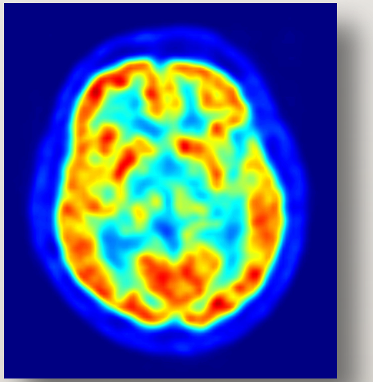
With material from Jimeng Sun (IBM TJ Watson Research),
Chandan K. Reddy (Wayne State University), an DS Raicu (de Paul University).

IMAGE DATA

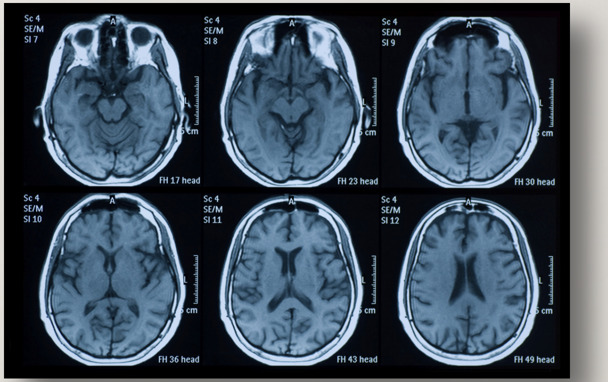
- In 2015, the average hospital had 0.7 petabytes (665 terabytes) of patient data, 80% of which was unstructured image data like CT scans and X-rays.
  - PACS (Picture Archival & Communication Systems) used for storage and retrieval.



Computed Tomography (CT)

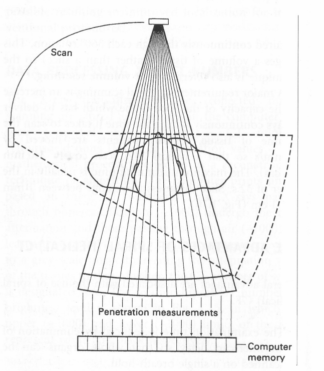


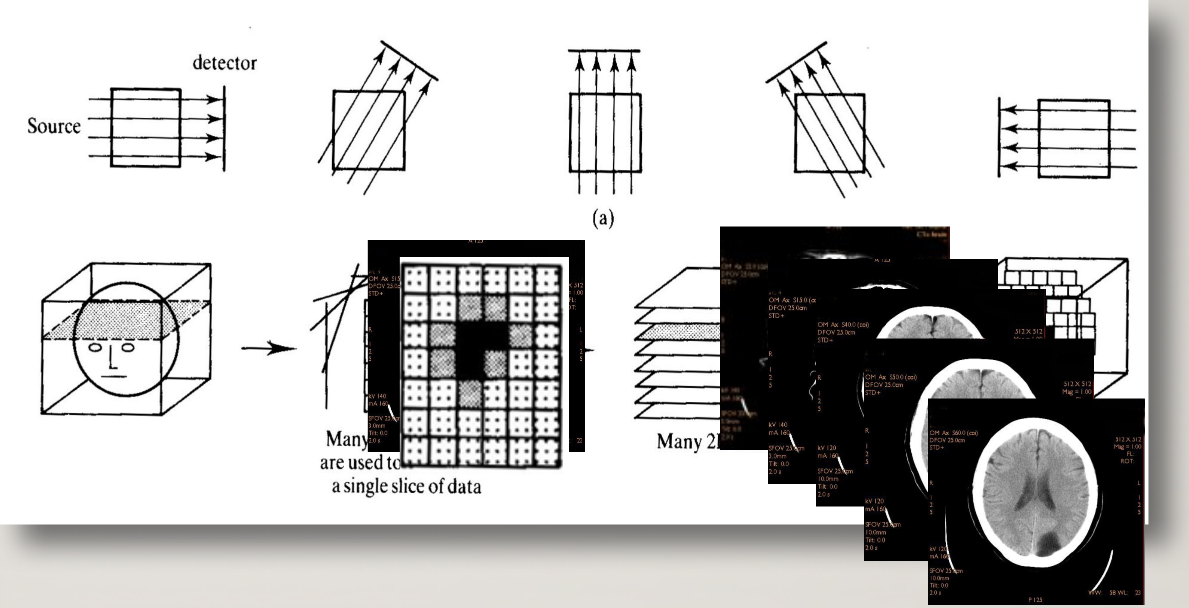
Positron Emission Tomography (PET)



Magnetic Resonance Imaging (MRI)

HOW ARE THESE DATA COMPUTED?

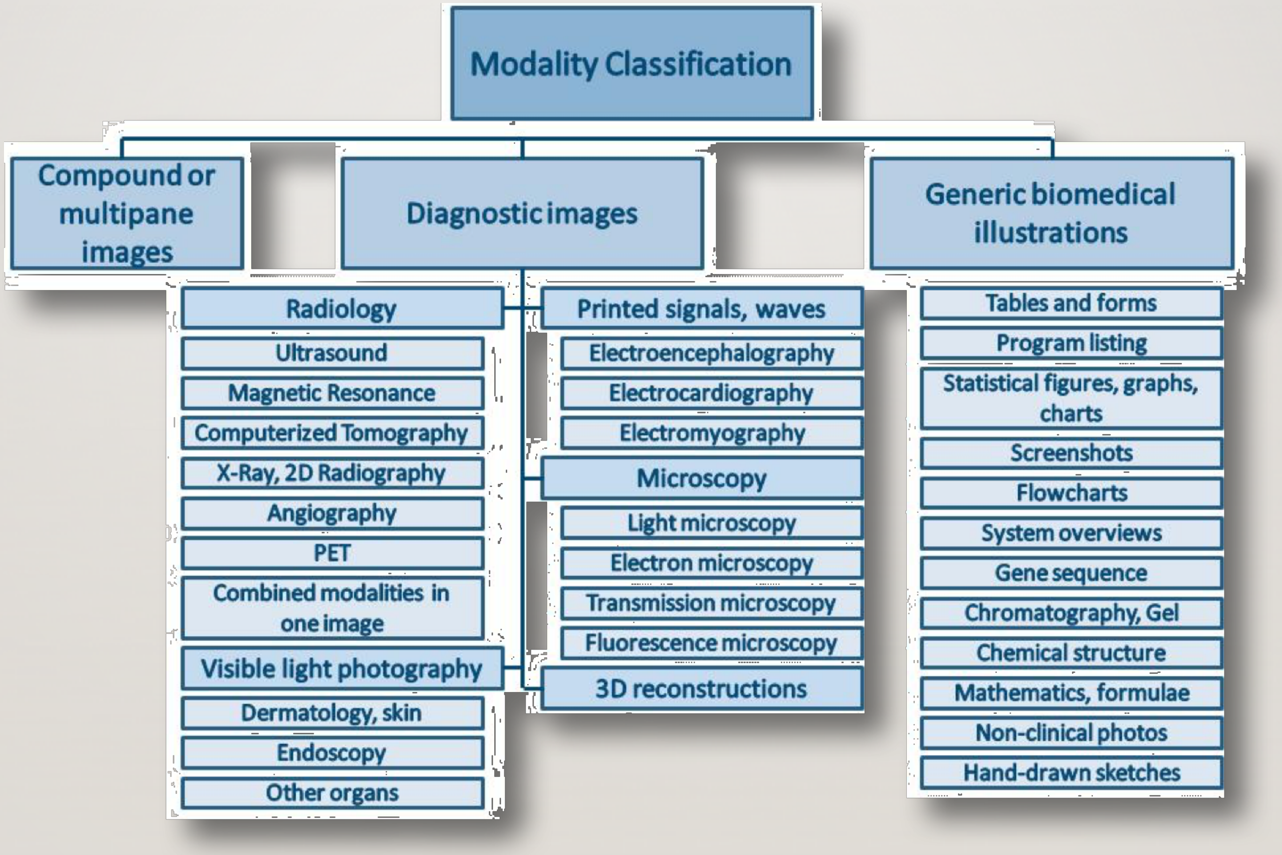




CONTENT-BASED IMAGE RETRIEVAL

- Image features/descriptors designed to encode color and/or texture properties of the image, the spatial layout of objects, and various geometric shape characteristics of coherent structures.
- Assessment of similarities between image features based on mathematical analyses.
  - E.g., vector affinity measures such as Euclidean distance, Mahalanobis distance, Kullback-Leibler divergence, and Earth Mover's distance.

MODALITY CLASSIFICATION



- Modality Classification
  - Compound or multipane images
  - Diagnostic images
    - Radiology
      - Ultrasound
      - Magnetic Resonance
      - Computerized Tomography
      - X-Ray, 2D Radiography
      - Angiography
      - PET
      - Combined modalities in one image
    - Visible light photography
      - Dermatology, skin
      - Endoscopy
      - Other organs
    - Printed signals, waves
      - Electroencephalography
      - Electrocardiography
      - Electromyography
    - Microscopy
      - Light microscopy
      - Electron microscopy
      - Transmission microscopy
      - Fluorescence microscopy
    - 3D reconstructions
  - Generic biomedical illustrations
    - Tables and forms
    - Program listing
    - Statistical figures, graphs, charts
    - Screenshots
    - Flowcharts
    - System overviews
    - Gene sequence
    - Chromatography, Gel
    - Chemical structure
    - Mathematics, formulae
    - Non-clinical photos
    - Hand-drawn sketches

IMAGE FEATURES

- Photo-metric features exploit color and texture cues and they are derived directly from raw pixel intensities.
- Geometric features: cues such as edges, contours, joints, polylines, and polygonal regions.
  - A suitable shape representation is extracted from pixel intensity information by region-of interest detection, segmentation, and grouping.

EXAMPLE FEATURES

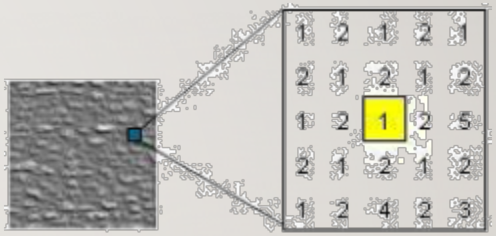
	Representation	Examples
Photometric	Grayscale/colour	Histograms; moments; block-based
	Texture	Texture co-occurrence; Fourier power spectrum; Gabor features; wavelets; Haralick's statistical features; multiresolution autoregression
Geometric	Contours/curves	Polygon approximation; edge histograms; Fourier; Curvature scale space
	Point sets	Shape spaces
	Surfaces	Level sets/distance transforms; Gaussian random fields
	Regions and parts	Statistical anatomical parts model; wavelet-based region descriptors; spatial distributions of regions of interest
	Other	Global shape (size, eccentricity, …); morphology; location and spatial relationships;

Akgül, Ceyhun Burak, et al. 2011 Content-based image retrieval in radiology: current status and future directions. *Journal of Digital Imaging* **24**(2):208-222.

Müller, Henning, et al. 2004 A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *International journal of medical informatics* **73**(1):1

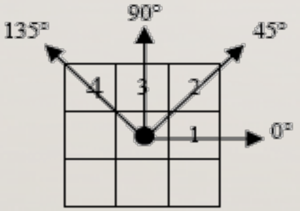
PIXEL-LEVEL EXTRACTION

- Consider texture around the pixel of interest.
- Capture texture characteristic based on estimation of joint conditional probability of pixel pair occurrences $P_{ij}(d, \theta)$.
    - P_{ij} denotes the normalized co-occurrence matrix specified by displacement vector (d) and angle (θ).



1	2	1	2	1
2	1	2	1	2
1	2	1	2	5
2	1	2	1	2
1	2	4	2	3

Neighborhood of a pixel



Direction from the centered pixel

0	0	1	3	1
3	1	1	3	1
2	1	3	0	3
3	2	1	0	3
3	3	2	1	2

Original Image

$$i \quad \begin{array}{c|cccc} & \multicolumn{4}{c}{j} \\ & 0 & 1 & 2 & 3 \\ \hline 0 & 1 & 1 & 0 & 2 \\ 1 & 1 & 1 & 1 & 3 \\ 2 & 0 & 3 & 0 & 0 \\ 3 & 1 & 3 & 2 & 1 \end{array}$$

Co-occurrence

Co-occurrence matrix for distance 1, direction 0°

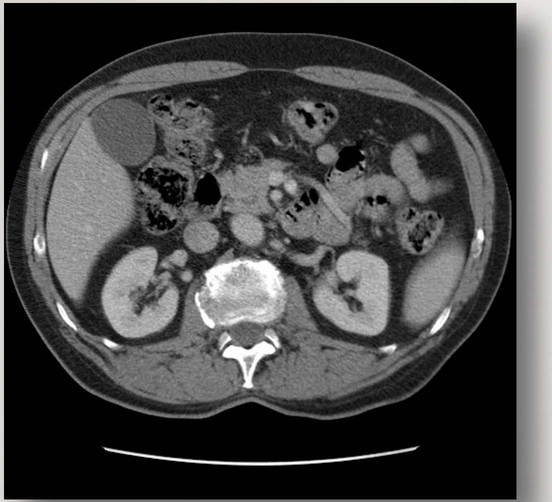
HARALICK TEXTURE FEATURES

- Entropy (randomness): $-\sum_i^M \sum_j^N P_{ij} \log P_{ij}$
- Energy (occurrence of repeated pairs): $\sum_i^M \sum_j^N P_{ij}^2$
- Contrast: $\sum_i^M \sum_j^N (i-j)^2 P_{ij}$
- Homogeneity: $\sum_i^M \sum_j^N \frac{P_{ij}}{|i-j|}$; where $i \neq j$
- Cluster tendency: $\sum_i^M \sum_j^N (i-\mu_r+j-\mu_c)^2 P_{ij}$

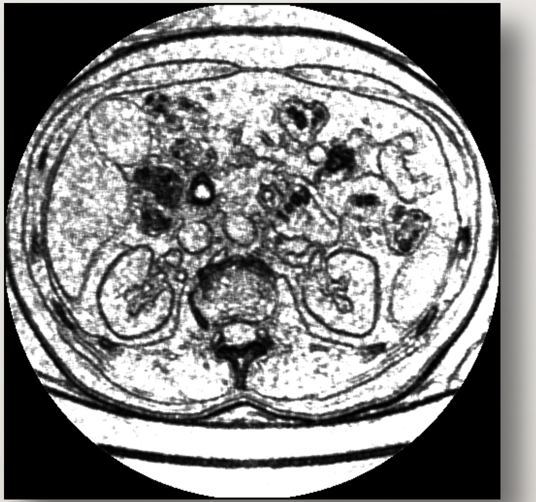
$$\mu_r = \sum_i^M \sum_j^N iP_{ij}$$

- Sum average; variance; correlation; inverse difference moment,…

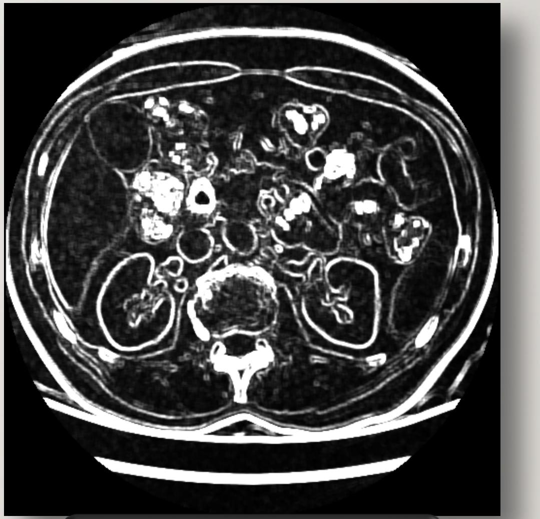
HARALICK TEXTURE FEATURES



Original

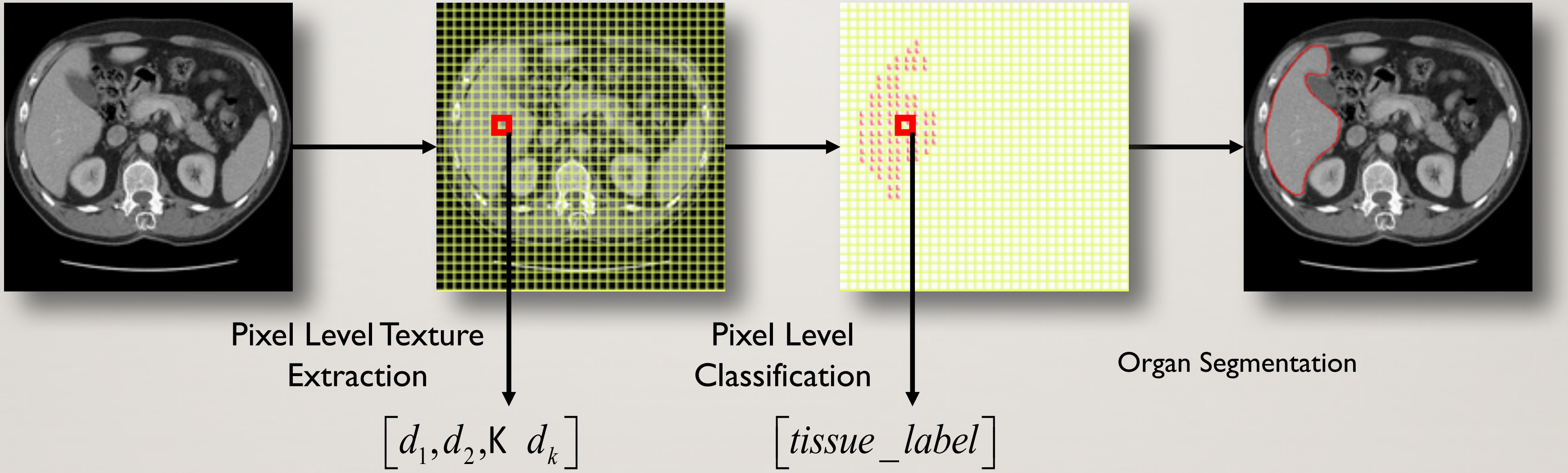


Energy



Cluster tendency

MALIGNANCY CLASSIFICATION



Pixel Level Texture Extraction

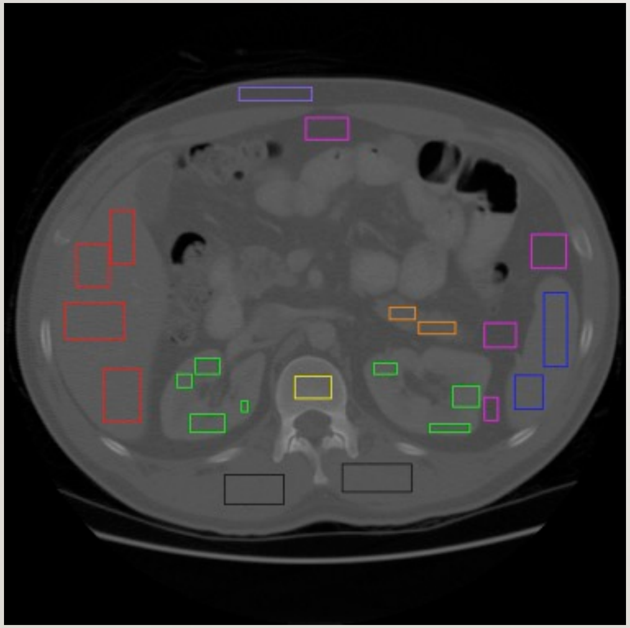
$\left[d_1, d_2, \mathrm{K}\ d_k\right]$

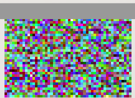
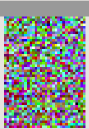
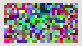
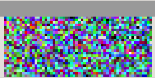
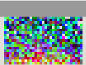
Pixel Level Classification

$\left[tissue\_label\right]$

Organ Segmentation

TISSUE CLASSIFICATION



Patch Samples	Ground truth: tissue names
	liver
	liver
	kidney
	fat
	muscle
	trabecular bone
…	…

TISSUE CLASSIFICATION

Organ-, patch-, and pixel- level classification using decision trees:

	Organ Level		Pure Patch Level		Pixel-level (9 x 9)		Pixel level (13 x 13)	
Organ	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Backbone (44)	100.0%	97.6%	97.7%	99.3%	100.0%	96.3%	100.0%	99.2%
Liver (259)	73.8%	95.9%	91.9%	97.9%	100.0%	99.0%	100.0%	98.4%
Heart (77)	73.6%	97.2%	79.2%	98.3%	81.1%	99.5%	66.7%	100.0%
Kidney (225)	86.2%	97.8%	91.6%	97.1%	78.9%	98.0%	96.6%	93.0%
Spleen (98)	70.5%	95.1%	65.3%	98.5%	94.4%	95.5%	100.0%	97.6%

- D. Xu, J. Lee, D.S. Raicu, J.D. Furst, D. Channin. (2005) Texture Classification of Normal Tissues in Computed Tomography, *The 2005 Annual Meeting of the Society for Computer Applications in Radiology*, Florida.
- D. Channin, D. S. Raicu, J. D. Furst, et al. (2004), Classification of Tissues in Computed Tomography using Decision Trees, Poster and Demo, *The 90th Scientific Assembly and Annual Meeting of Radiology Society of North America (RSNA04)***,** Chicago.

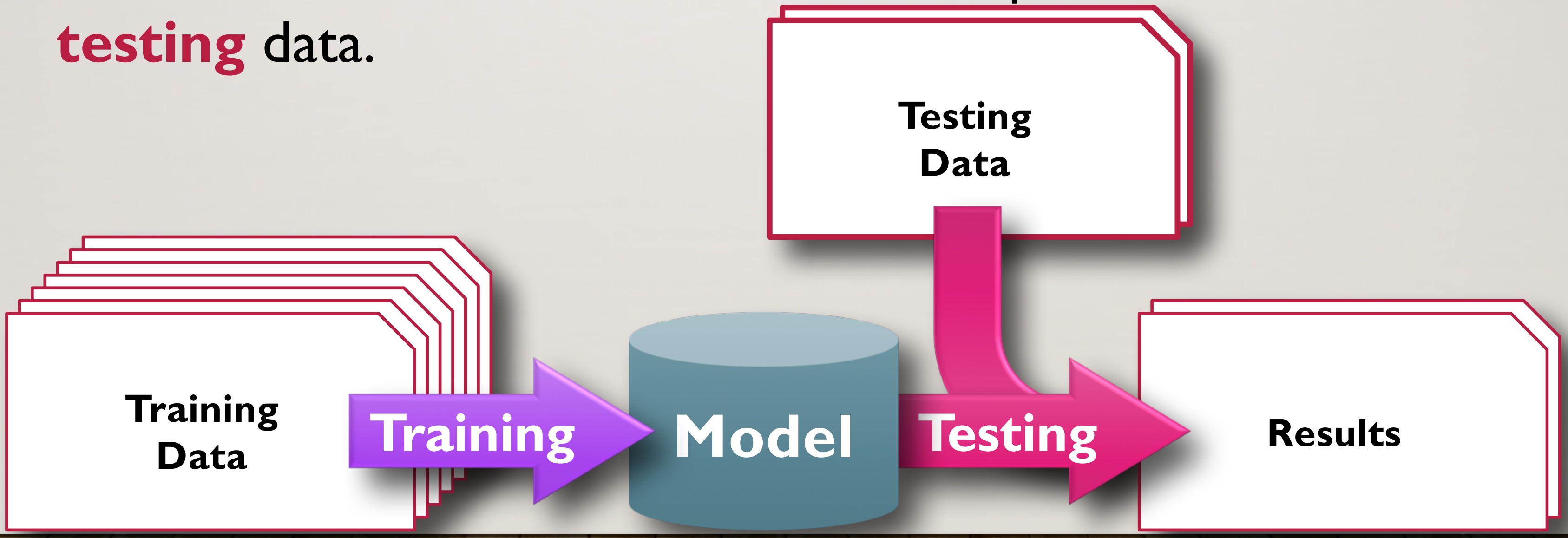
PUBLIC IMAGE DATABASES

	Modalities	# patients	# Images	Size	Notes/Applications	Link
Cancer Imaging Archive Database	CT DX CR	1010	244,527	241 GB	Lesion Detection and classification, Accelerated Diagnostic Image Decision, Quantitative image assessment of drug response	https://public.cancerimagingarchive.net/ncia/dataBasketDisplay.jsf
Digital Mammography database	DX	2620	9,428	211 GB	Research in Development of Computer Algorithm to aid in screening	http://marathon.csee.usf.edu/Mammography/Database.html
Public Lung Image Database	CT	119	28,227	28 GB	Identifying Lung Cancer by Screening Images	https://eddie.via.cornell.edu/crpf.html
Image CLEF Database	PET CT MRI US	unknown	306,549	316 GB	Modality Classification ,Visual ImageAnnotation, Scientific Multimedia Data Management	http://www.imageclef.org/2013/medical
MS Lesion Segment ation	MRI	41	145	36 GB	Develop and Compare 3D MS Lesion Segmentation Techniques	http://www.ia.unc.edu/MSseg/download .php
ADNI Database	MRI PET	2851	67,871	16GB	Define the progression of Alzheimer's disease	http://adni.loni.ucla.edu/data-samples/acscess-data/

SOME METHODOLOGICAL AND/OR EMPIRICAL CONSIDERATIONS

GENERAL PROCESS

1. We gather a big and relevant **training** dataset.
2. We learn our **parameters** (e.g., probabilities) from that dataset to build our **model**.
3. Once that model is fixed, we use those probabilities to evaluate **testing** data.



GENERAL PROCESS

- Often, **training data** consists of 80% to 90% of the available data.
  - Often, some subset of *this* is used as a **validation/development set**.

- **Testing data** is ***not*** used for training but comes from the same *source*.
  - It often consists of the remaining 10% to 20% of the available data.
  - Sometimes, it's important to **partition** patients so they **don't** appear in both training and testing.

BETTER PROCESS: *K*-FOLD CROSS-VALIDATION

- **K-fold cross validation**: *n.* splitting all data into *K* **partitions** and iteratively testing on each after training on the rest (report means and variances).

	Part 1	**Part 2**	**Part 3**	**Part 4**	**Part 5**	
Iteration 1	■	□	□	□	□	: Err1 %
Iteration 2	□	■	□	□	□	: Err2 %
Iteration 3	□	□	■	□	□	: Err3 %
Iteration 4	□	□	□	■	□	: Err4 %
Iteration 5	□	□	□	□	■	: Err5 %

5-fold cross-validation

■	**Testing Set**
□	**Training Set**

SYNTHETIC DATA: K-MEANS IMPUTATION

- If you have missing (NaN) variables in your data, guess (i.e., impute) them by looking at the k nearest vectors from the available data.
- E.g., given $k = 2$ and the table below, we would impute the hidden values as [2, 12, 0.75]

0.01	0.01	1	10	0.5	17
0.01	0.02	[redacted]	[redacted]	[redacted]	16
0.06	0.18	8	7	0.5	100
0.02	0.01	3	14	1.0	17.1

KNOWLEDGE

- **Anecdotes** are often useless except as proofs by contradiction.
  - E.g., *"My son has autism and took vaccines"* does **not** mean that *autism* is **always** (or even **likely** to be) correlated with vaccines.

- **Shallow statistics** are often not enough to be truly meaningful.
  - E.g.,*"My CDSS is 95% accurate on my test data. Yours is only 94.5% accurate, you horrible knuckle-dragging idiot."*
    - What if the test data was **biased** to favor my system?
    - What if we only used a **very small** amount of data?

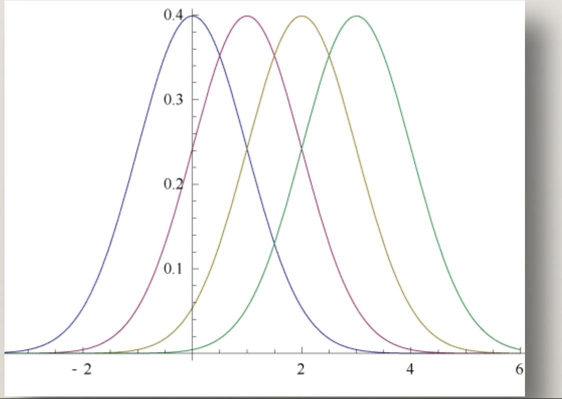
- We need a **test** to see if our statistics actually **mean** something.

DIFFERENCES DUE TO SAMPLING

- **Kullback-Leibler divergence** measures how **different** two distributions are from each other.

- But what if their difference is due to **randomness** in **sampling**?

- How can we tell that a distribution is ***really*** different from another?



HYPOTHESIS TESTING

- Often, we assume a **null hypothesis**, H_0, which states that the **two distributions are *the same*** (i.e., come from the same underlying model, population, or phenomenon).

- We **reject** the null hypothesis if the probability of it being true is too small.
  - This is often our goal – e.g., if my CDSS beats yours by 0.5%, I want to show that this difference is **not** a random accident.

  - As scientists, we have to be very **careful** to not reject H_0 too hastily.
    - How can we ensure our **diligence**?

CONFIDENCE

- We stated that we **reject** H_0 if it is **too improbable**.
  - How do we determine the value of 'too'?

- **Significance level α** $(0 \le \alpha \le 1)$ is the **maximum** probability that two distributions are **identical** allowing us to **disregard** H_0.
  - In practice, $\alpha \le 0.05$. Usually, it's much lower.
  - **Confidence level** is $\gamma = 1 - \alpha$
  - E.g., a confidence level of **95%** $(\alpha = 0.05)$ implies that we expect that our decision is correct 95% of the time, **regardless of the test data**.

THE *T*-TEST

- The **t-test** is a method to compute if distributions are significantly different from one another.

- It is based on the mean $(\overline{\boldsymbol{x}})$ and variance $(\boldsymbol{\sigma})$ of N samples.
- It compares $\bar{x}$ and σ to H_0 which states that the samples are drawn from a distribution with a **mean** $\boldsymbol{\mu}$.

- If $\boxed{t = \frac{\bar{x}-\mu}{\sqrt{\sigma^2/N}}}$ (the "t-statistic") is large enough, we can reject H_0.

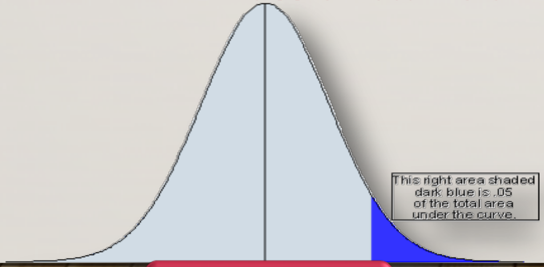
An example would be nice…

There are actually **several types** of *t*-tests for different situations…

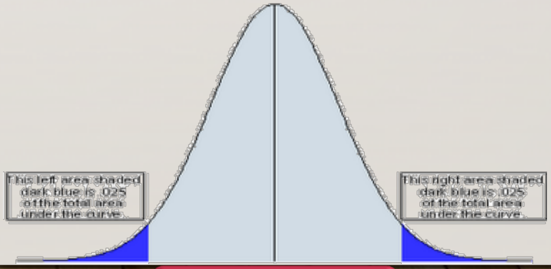
EXAMPLE OF THE T-TEST; TAILS

- Imagine that the average IQ of a UofT student is $\mu = 158$.
- We sample $N = 200$ UofT students from DCS and find that $\bar{x} = 169$ and $\sigma^2 = 2600$.
- Are DCS students significantly **smarter** than their peers?

- We use a '**one-tailed**' test because we want to see if DCS students measure significantly **higher**.
  - If we just wanted to see if DCS were significantly **different**, we'd use a **two-tailed** test.



one tail



two tails

EXAMPLE OF THE T-TEST: FREEDOM

- Imagine that the average IQ of a UofT student is $\mu = 158$.
- We sample $N = 200$ UofT students from DCS and find that $\bar{x} = 169$ and $\sigma^2 = 2600$.
- Are DCS students significantly **smarter** than their peers?

- **Degrees of freedom (d.f.):** *n.pl.* In *this t*-test, this is the sum of the number of observations in each group, minus 2 (because there are two groups).

- In our example, we have $N_{DCS} = 200$ for DCS students, but $N_{UofT} \approx \infty$ for the other group, so $d.f. = \infty$.

EXAMPLE OF THE T-TEST

- Imagine that the average IQ of a UofT student is $\mu = 158$.
- We sample $N = 200$ UofT students from DCS and find that $\bar{x} = 169$ and $\sigma^2 = 2600$.
- Are DCS students significantly **smarter** than their peers?

- So $t = \frac{\bar{x}-\mu}{\sqrt{\sigma^2/_N}} = \frac{169-158}{\sqrt{2600/_{200}}} \approx 3.05$

- In a ***t*-test table**, we look up the minimum value of t necessary to reject H_0 at $\alpha = 0.005$ (we want to be quite confident) for a 1-tailed test…

EXAMPLE OF THE T-TEST

- So $t = \frac{\bar{x}-\mu}{\sqrt{\sigma^2/_N}} = \frac{169-158}{\sqrt{2600/_{200}}} \approx 3.05$

- In a **t-test table**, we look up the minimum value of t necessary to reject H_0 at $\alpha = 0.005$, and find 2.576.
  - Since $3.05 > 2.576$, we can reject H_0 at the 99.5% level of confidence ($\gamma = 1 - \alpha = 0.995$) ; **DCS students are significantly smarter**.

	α **(one-tail)**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
d.f.	1	6.314	12.71	31.82	63.66	318.3	636.6
	10	1.812	2.228	2.764	3.169	4.144	4.587
	20	1.725	2.086	2.528	2.845	3.552	3.850
	∞	1.645	1.960	2.326	**2.576**	3.091	3.291

EXAMPLE OF THE T-TEST

- Some things to observe about the *t*-test table:
  - We need **more evidence**, *t*, if we want to be **more confident** (left-right dimension).
  - We need **more evidence**, *t*, if we have **fewer measurements** (top-down dimension).
- A common criticism of the *t*-test is that picking α is ad-hoc. There are ways to correct for the selection of α.

	α (one-tail)	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
d.f.	1	6.314	12.71	31.82	63.66	318.3	636.6
	10	1.812	2.228	2.764	3.169	4.144	4.587
	20	1.725	2.086	2.528	2.845	3.552	3.850
	∞	1.645	1.960	2.326	**2.576**	3.091	3.291

ANALYSIS OF VARIANCE

- **Analyses of variance (ANOVAs)** (there are several types) can be:
  - A way to **generalize *t*-tests** to more than two groups.
  - A way to **determine which** (if any) of several **variables** are **responsible** for the **variation** (and interaction) in observations.
- E.g., we measure the **accuracy** of CDSS for different settings of **empirical parameters** M and Q.

Accuracy (%)	$\boldsymbol{M = 2}$	$\boldsymbol{M = 4}$	$\boldsymbol{M = 16}$
$\boldsymbol{Q = 2}$	53.33	66.67	53.33
	26.67	53.33	40.00
	0.00	40.00	26.67
$\boldsymbol{Q = 5}$	93.33	26.67	100.00
	66.67	13.33	80.00
	40.00	0.00	60.00

H_0: no effect of source variables.

Source	$\boldsymbol{d.f.}$	$\boldsymbol{p}$ **value**	
Q	1	0.179	Accept H_0
M	2	0.106	Accept H_0
interaction	2	0.006	Reject H_0 at $\alpha = 0.01$

A completely fictional example

TESTING FOR NORMALITY

- Another problem with *t*-tests and ANOVAs are that they often assume that the underlying distribution of the data is Gaussian/Normal.
- The Lilliefors test, based on the Kolmogorov–Smirnov test, tests whether a distribution is Normal.
    1. Estimate the population mean and population variance based on the data.
    2. Find the maximum discrepancy between the empirical distribution function and the cumulative distribution function (CDF) of the normal distribution with the estimated mean and estimated variance.
    3. Assess whether the maximum discrepancy is large enough to be statistically significant, thus requiring rejection of the null hypothesis.
        1. Since the hypothesized CDF has been moved closer to the data by estimation based on those data, the maximum discrepancy has been made smaller than it would have been if the null hypothesis had singled out just one normal distribution. Tables for this distribution have been computed only by Monte Carlo methods.

REPEATED MEASURES

- People are weird – there are lots of individual differences.
- **Repeated measures design** uses the same subjects with every stage of the research, including the control.
  - E.g., repeated measurements of limb function are collected in a longitudinal study where change is measured over time
  - E.g., study the same individuals after taking medication ***and*** after taking a placebo.
- In a **between-subjects** ANOVA, $SS_{\text{Total}} = SS_{\text{Treatment}} + SS_{\text{Error}}$
- In a repeated measures design,

  $SS_{\text{Total}} = SS_{\text{Treatment (excluding individual difference)}} + SS_{\text{Subjects}} + SS_{\text{Error}}$

MULTIPLE COMPARISONS

- The multiple comparisons problem occurs when you consider a set of statistical inferences simultaneously.
- E.g., H_0 is that a coin is fair, so,

$$\mathrm{p}\left(\text{heads} \geq \frac{9}{10} times\right) = (10 + 1)\left(\frac{1}{2}\right)^{10} = 0.0107.$$

- You choose $\alpha = 0.05$, so if you see 9/10 tosses come up heads, you reject H_0.

MULTIPLE COMPARISONS

- The multiple comparisons problem occurs when you consider a set of statistical inferences simultaneously.
- E.g., H_0 is that a coin is fair, so,

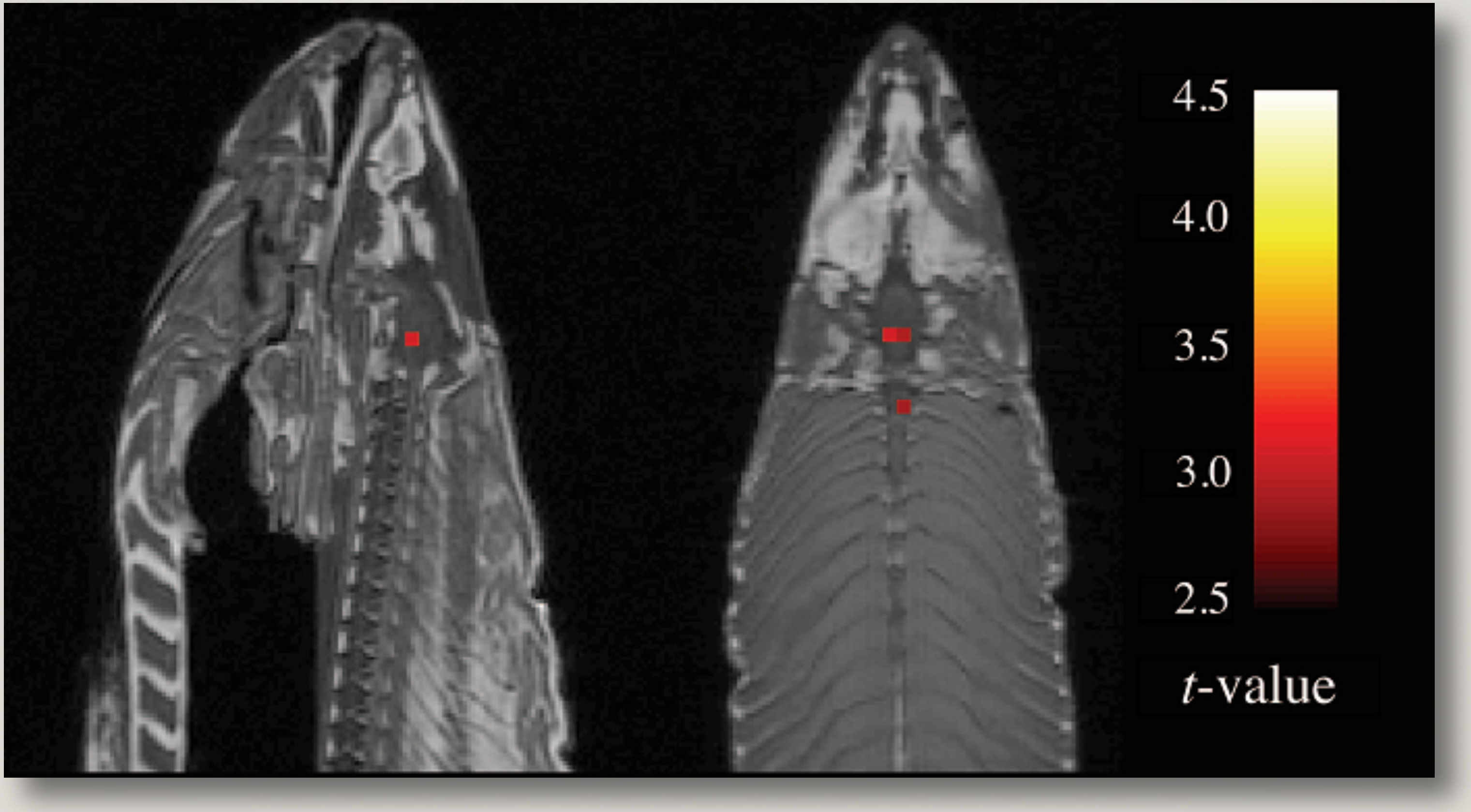
$$\mathrm{p}\left(\mathrm{heads} \geq \frac{9}{10} times\right) = (10+1)\left(\frac{1}{2}\right)^{10} = 0.0107.$$

  - You choose $\alpha = 0.05$, so if you see 9/10 tosses come up heads, you reject H_0.
  - No problem

MULTIPLE COMPARISONS

- Following the same example, what if you wanted to *simultaneously* test 100 coins?
  - The probability of a fair coin coming up 9 or 10 heads in 10 flips is 0.0107
- To see a ***particular*** *pre-selected* coin come up heads 9 or 10 times in 10 flips would still be unlikely.
- To see ***any*** coin behave that way, without concern for which one, would increase with the number of coins!
  - Precisely, the likelihood that *all* 100 fair coins are identified as fair by this criterion is $(1 - 0.0107)^{100} \approx 0.34$.
- Bonferroni correction sets the threshold for significance at α/m, given m hypotheses.

MULTIPLE COMPARISONS



Bennett et al. "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction" Journal of Serendipitous and Unexpected Results, 2010.