cogs 105 this week

BIGDATA

today: latent semantic analysis

Types of Research

- Philosophical / theoretical
- Experimental
- Observational
- Computational
- Cognitive engineering

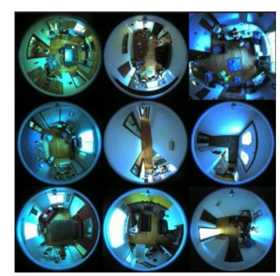
Types of Research

- Philosophical / theoretical
- **Experimental**
- **Observational**
- Computational
- Cognitive engineering

Experimental vs. Observational

involves direct intervention	intervention is avoided (or not possible)

E.g., setup experimental task in laboratory for babies



Deb Roy, MIT

Experimental vs. Observational

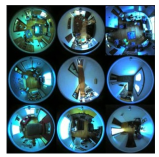
dependent variable
(you measure)

independent variable
(you control)

DV: Extent of play
IV: Depth of social familiarity

outcome variable
(variable of interest)

predictors and covariates
(to predict / explain outcome)



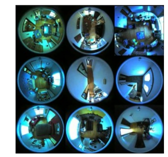
Outcome: Extent of play
Predictor: Depth of social familiarity
Covariates: Time of day, recent food, etc.

Experimental vs. Observational

causal
inferences often
acceptable

Enhanced social familiarity
causes increased play engagement

correlational
inferences are
preferred



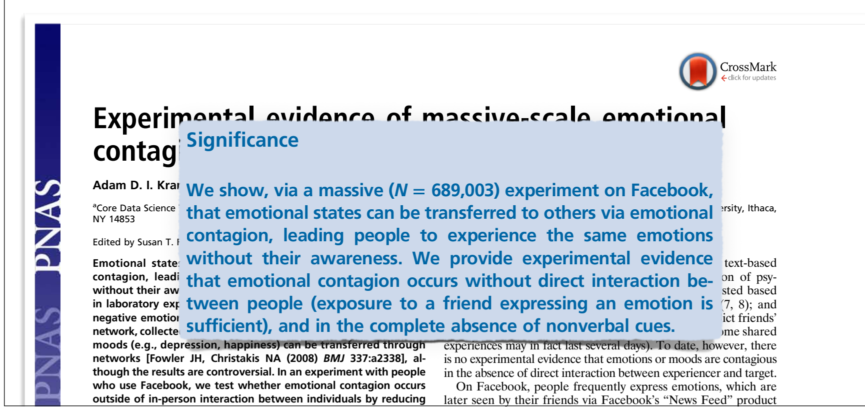
Enhanced social familiarity
is related to increased
play engagement.

Big Data

- Remember, “big data” is a general term that connotes a trend to utilize large and unseemly data sets to render new insights.
- Studies using big data are **primarily** observational in nature. (Correlational studies with lots of data.)
    - Big data studies can sometimes be experimental though. (Use of technology to setup experimental conditions and collect lots of data.)
    - Also big data can be **used to build tools** for experimental research.

Example

- Facebook’s controversial study.



PNAS

CrossMark

Experimental evidence of massive-scale emotional contag

Significance

We show, via a massive (N = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing

experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experiencer and target.

On Facebook, people frequently express emotions, which are later seen by their friends via Facebook’s “News Feed” product

Linguistic Tools

- Big data can also help us render new tools — for example, the development of semantic models.
- Latent semantic analysis (LSA).
  - Uses massive amounts of text to build a model that allows us to compare words to each other in terms of their "meaning."
- Thursday: LIWC

Starting Point

and rate.

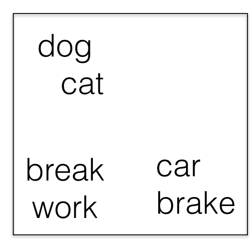
This leads us to ask the question: Suppose we have available a corpus of data approximating the mass of intrinsic and extrinsic language-relevant experience that a human encounters, a computer with power that could match that of the human brain, and a sufficiently clever learning algorithm and data storage method. Could it learn the meanings of all the words in any language it was given?

The keystone discovery for LSA was that using just a single simple con-

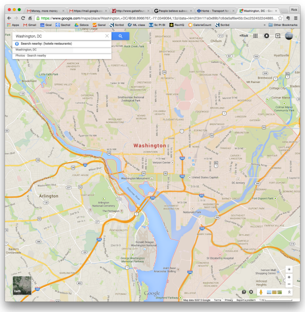
Philosophers, linguists, humanists, novelists, poets, and theologians have used the word "meaning" in a plethora of ways, ranging, for example, from the truth of matters to intrinsic properties of objects and happenings in the world, to mental constructions of the outside world, to physically irreducible mystical essences, as in Plato's ideas, to symbols in an internal communication and reasoning system, to potentially true but too vague no-

Mapping Meaning

- LSA goes from a huge amount of text data, to a distilled representation of word meaning in the form of a vector space or "map."
- In this space, words do not have "meaning" all on their own; their meanings are derived from their relationships to other words.



How LSA Works: Map Description



“massive text info”

LSA

“word meaning”

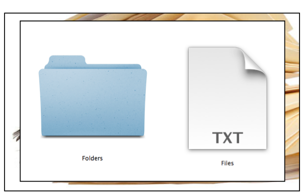
How LSA Works: Juicing Description

“massive text info”

LSA

“word meaning”

How LSA Works: Almost There



LSA

dog
cat

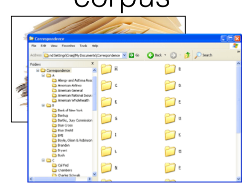
break
work

car
brake

How LSA Works: Almost There

Step 1: Word-by-Document Matrix

“corpus”

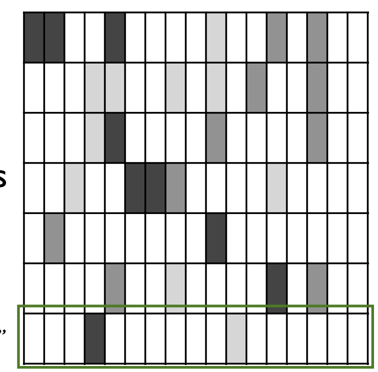


cells represent how often a word occurs in each file (represented by grayscale)

Files / documents

Words

“dog”



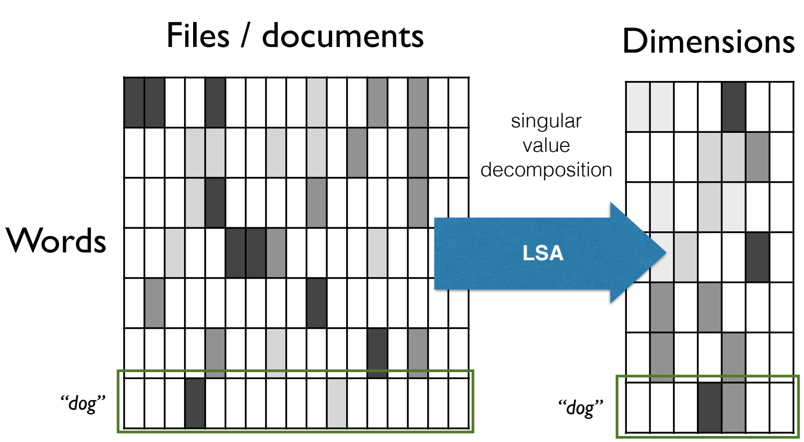
The Problem

- The cells in a word-by-document matrix are mostly empty; this creates great difficulties in relating word meaning.
  - Sometimes called “data sparsity” problem.
- LSA is a statistical techniques that acts like “squeezing the sponge” or “drawing the map” by extracting the **major trends/relationships among words in the matrix**.

A Simple Motivation…

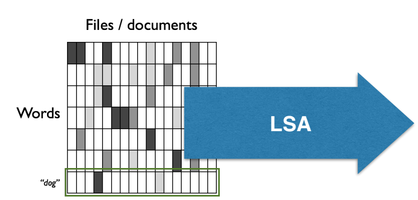
- “dog” may rarely or even never occur in the same document as either “parrot” or “pencil.”
- However, both “parrot” and “dog” may occur with similar words: “breathe, eat, drink, noise, interact, owner,” etc.
- LSA is able to extract these relationships — and so it would tell us, in our map of meaning, that “dog” and “parrot” are more similar than “dog” and “pencil.”

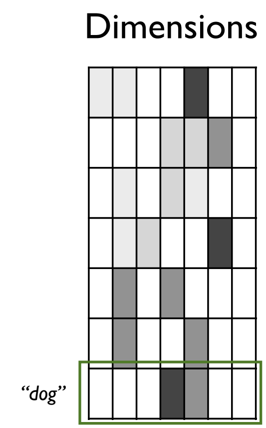
Finally…



How LSA Works: Almost There

Step 2: LSA space is a lower dimensional matrix



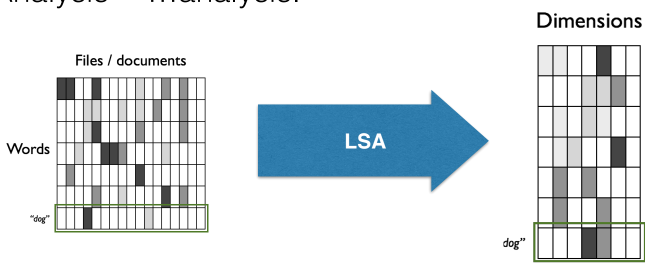


the dimensions
are now the space
in which words live
and can be related

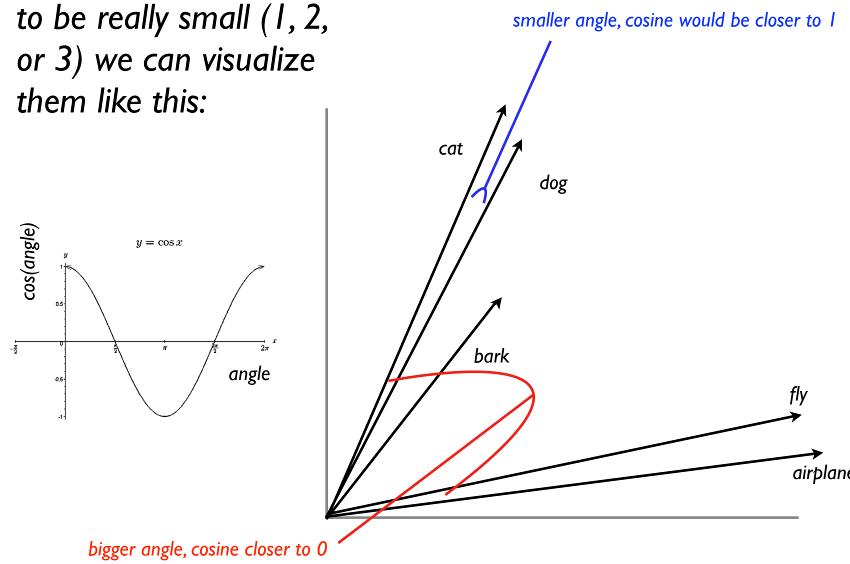
(…this is our
“map” or the
“juice”…)

Why “LSA”?

- Latent = “existing but not yet developed or manifest; hidden.”
- Semantic = “of or related to meaning.”
- Analysis = …analysis.



If dimensions happen to be really small (1, 2, or 3) we can visualize them like this:

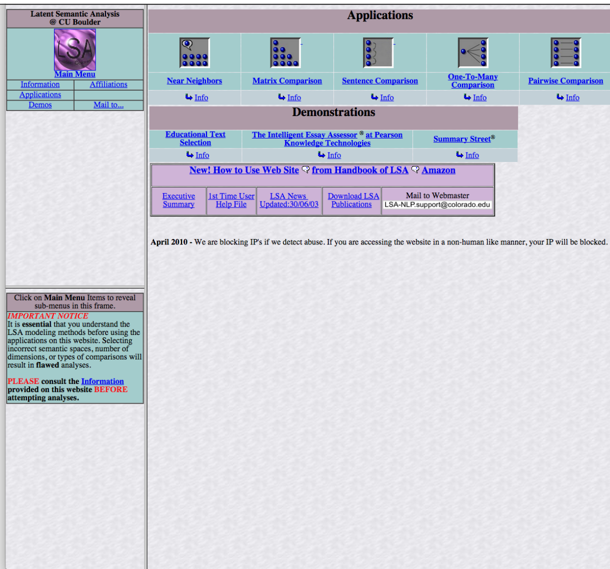


“Meaning”

- Modern cognitive science methods now allow us to “quantify meaning” in this way.
- Philosophers have spent millennia talking about meaning; there is still endless debate about meaning.
- However, **LSA, as a model of meaning**, can grade papers, pass the MCAT, work with educational technologies, and many more.

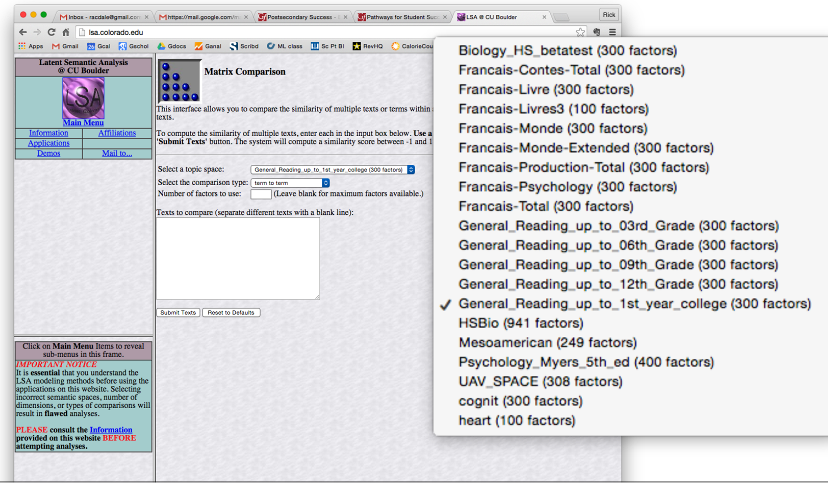
So How Do I LSA?

- Do I have to crunch all the numbers?
- It’s actually pretty easy to do it. If you want sample code, I can show you how to build an LSA model in no more than 10 lines of code in MATLAB, Python, or R.
- However, for the purposes of this class and explore LSA, we will use an amazing online tool…

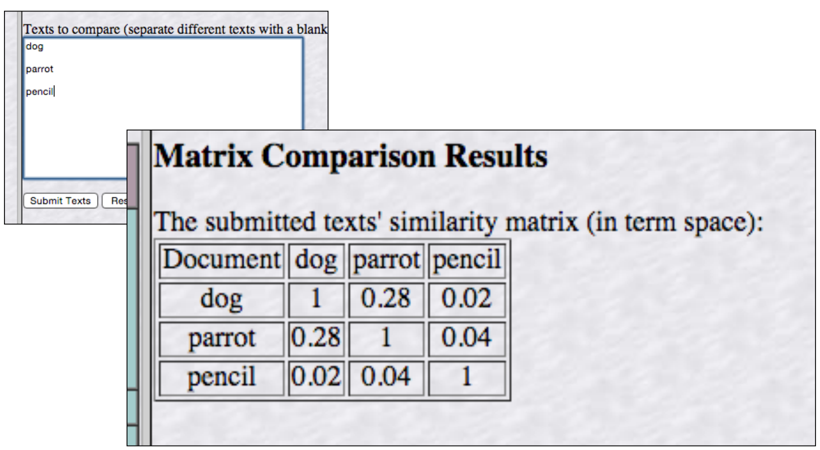


lsa.colorado.edu

Matrix Comparison



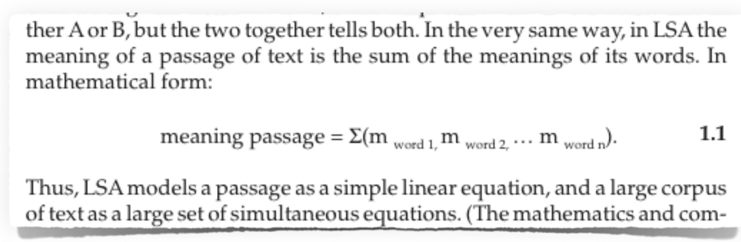
Running Some Comparisons



Document	dog	parrot	pencil
dog	1	0.28	0.02
parrot	0.28	1	0.04
pencil	0.02	0.04	1

Sentences / Passages?

- What about sentences? What if we want to compare larger blocks of text?

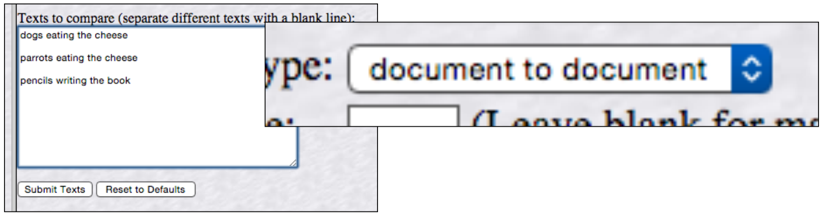


ther A or B, but the two together tells both. In the very same way, in LSA the meaning of a passage of text is the sum of the meanings of its words. In mathematical form:

$$\text{meaning passage} = \Sigma(m_{\text{word 1}}, m_{\text{word 2}}, \ldots m_{\text{word n}}). \qquad 1.1$$

Thus, LSA models a passage as a simple linear equation, and a large corpus of text as a large set of simultaneous equations. (The mathematics and com-

Running Some Comparisons



The submitted texts' similarity matrix (in document space):

Document	dogs eating the cheese	parrots eating the cheese	pencils writing the book
dogs eating the cheese	1	0.64	0.03
parrots eating the cheese	0.64	1	0.02
pencils writing the book	0.03	0.02	1

What's It Good For?

- Tons of stuff! E.g.:
  - Experimental design (e.g., controlling for word similarity in an RT task)
  - Observational designs (e.g., comparing semantic similarity between conversation partners; e.g., Dale & Duran, 2008)
  - Search engine and document indexing
  - Educational technologies (e.g., artificial tutors)

Limitations

- LSA suffers from some problems.
- It can't handle syntax.
  - E.g., these words have the "same meaning"
    - The dog ate my homework
    - The homework ate my dog (?)

Matrix Comparison Results

The submitted texts' similarity matrix (in term space):

Document	The dog ate my homework	The homework ate my dog
The dog ate my homework	1	1.00
The homework ate my dog	1.00	1

Limitations

- It does not do well with homonymy ("same word, different meanings").
  - E.g., "cream in your coffee" and "cream you at hockey" have different "creams" in them.
    - LSA treats them as one word.

Matrix Comparison Results

The submitted texts' similarity matrix

Document	left	depart	right
left	1	0.34	0.72
depart	0.34	1	0.16
right	0.72	0.16	1

Limitations

- It does not do well with antonymy (opposites).
  - Love and hate occur in overlapping descriptive contexts, but they are quite different in meaning.
    - LSA often treats antonyms as similar in meaning (could this make sense sometimes?)

Matrix Comparison Results

The submitted texts' similarity matrix

Document	love	hate	admiration
love	1	0.50	0.41
hate	0.50	1	0.38
admiration	0.41	0.38	1

accomplished this feat was LSA.

LSA is a computational model that does many humanlike things with language. The following are but a few: After autonomous learning from a large body of representative text, it scores well into the high school student range on a standardized multiple-choice vocabulary test; used alone to rate the adequacy of content of expository essays (other variables are added in full- scale grading systems; Landauer, Laham, & Foltz, 2003a, 2003b), estimated in more than one way, it shares 85%–90% as much information with expert human readers as two human readers share with each other (Landauer, 2002a); it has measured the effect on comprehension of paragraph-to-paragraph coherence better than human coding (Foltz, Kintsch, & Landauer, 1998); it has successfully modeled several laboratory findings in cognitive psychology (Howard, Addis, Jing, & Kahana, chap. 7 in this volume; Landauer, 2002a; Landauer & Dumais, 1997; Lund, Burgess, & Atchley, 1995); it detects improvements in student knowledge from before to after reading as well as human judges (Rehder et al., 1998; Wolfe et al., 1998); it can diagnose schizophrenia from what patients say as well as experienced psychiatrists (Elvevåg, Foltz, Weinberger, & Goldberg, 2005); it improves information retrieval by up to 30% by being able to match queries to documents of the same meaning when there are few or no words in com-

Next Time

- We'll compare quantitative and qualitative approaches with LIWC, in the context of Big Data.
- Lab this week: Neurosynth.