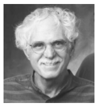
A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge

Or, a really long title to an equally long-(winded) article

Thomas Landauer and Susan Dumais



A big problem

- Inductive Paradox: How do we know so much given how little information there appears to be in the world? This is the Poverty of the Stimulus broadly construed.
- Modern theories of knowledge acquisition: We must accept some constraints that greatly narrow the solution space of the problem that is to be solved by induction. Plato: the knowledge is innate, we simply need contemplate and infer from hints that we encounter in the world.

•L&S Focus on Learning Vocabulary

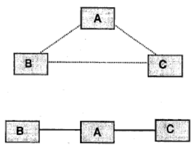
•The Problem: If we know 40,000 – 100,00 words by age 20, we would need to have learned 7-15 words every day for 18 years, beginning at age 2!

•Question: Do you know even one more word today than you did yesterday? Do you recall learning 7 words a day, every day, every week? Wouldn’t this kind of learning imply effort on your part? Wouldn’t you remember how hard it was?

•Or perhaps you cannot remember; perhaps you learned the words without deliberate effort at apprehension.
-Maybe there is a computational model that can, with some specified constraints, show how people learn vocabulary at the astonishing rate mentioned above.

The Inductive Value of Dimensionality Optimization

Problem: Jack and Jill are on the phone. Jack tells Jill: “I can see three houses, A, B, and C. A appears to be 5 units from B and C. B and C are about 8 units apart.”
- Using these estimates, Jill plots the houses as a triangle. Then Jack says, the houses are on the same street. This new constraint enables Jill to correctly plot the relationship between the three houses. Whereas knowing the distance between three objects sets up the first model, adding one more constraint enables reduction to one dimension while still conveying the necessary information



Semantic similarity b/w 2 words

How dimensionality applies to word relations:
1.The distance between words is inversely related to their similarity.
2.We would want to parse text into windows so that we capture discursive context:
  a. Two words that appear in the same window of discourse (phrase, sentence, paragraph etc.) tend to come from nearby locations in semantic space.
3.This would allow us to estimate the relative similarity of any pair of words by observing the relative frequency of their joint-occurrence in such windows
4.But many words do not appear together directly even though they may be semantically related. How to account for this indirect relation?
  a. e.g. purple/lavender; overweight/corpulent (use only one of these in a sentence)
  b. Gears/brakes, eraser/lead (parts of an object (car, pencil).

O comfortable friar! where is my lord? I do remember well where I should be, And there I am. Where is my Romeo?

LSA experiment
Input: Romeo

0.96	juliet
0.79	shakespearean
0.76	playwright
0.73	comedy
0.73	playwrights
0.73	shakespeare
0.71	drama
0.71	actors
0.70	theater
0.70	buffoonery
0.70	soliloquy
0.70	plays
0.70	actor
0.70	hamlet
0.69	macbeth
0.67	playgoing
0.67	theatrical

Database Matters

Heart DB	heart	disease	beat	courage
heart	1	0.31	0.16	0.05
disease	0.31	1	0.01	0.04
beat	0.16	0.01	1	-0.02
courage	0.05	0.04	-0.02	1

Literature DB	heart	disease	beat	courage
heart	1	0.26	0.59	0.28
disease	0.26	1	0.23	0.37
beat	0.59	0.23	1	0.39
courage	0.28	0.37	0.39	1

But can LSA really get subtle distinctions?

- Not even ten years ago you could buy a house for fifty thousand dollars.
- Even ten years ago you could not buy a house for fifty thousand dollars.
- LSA calculates cosine of these sentences as 1.0 or identical.
- <u>Example 2</u>
- 1. Manchester United is a soccer team. (.95)
- 2: A soccer team united Manchester.
- 3: United, a soccer team defeated Manchester. (1.0)

Solution: Take all local estimates of distance into account at once

Selecting appropriate dimensionality for pairwise estimates will be critical to achieving correct results based on mutual constraints.

Technical overview:
· Word meanings are represented as vectors in *k* dimensional space.
· Estimates of pairwise meaning similarities and of similarities among related pairs never observed together can be improved if fitted simultaneously into a space of the same *k* dimensionality.
· Idea is similar to factor analysis or multidimensional scaling (MDS).

LSA in Psychological Terms

- Words exist as points in high dimensional space.
- Sender chooses words located near each other when generating a string. In a short time window, contiguities in the output of words reflects closeness in the sender's semantic space.
- Receiver can make first-order estimates bw pairs by their relative frequency of occurrence in the same temporal context (e.g. a paragraph). Receiver can reconstruct sender's space by estimating similarities between observed and unobserved words (i.e. reconstruct dimensionality sender used)

How the LSA Model Works

- "Psychological similarity between any two words is reflected in the way the co-occur in small subsamples of language...the source of language samples produces words in a way that ensures a mostly orderly stochastic mapping between semantic similarity and output distance. [The model] then fits all of the pairwise similarities into a common space of high but not unlimited dimensionality."

- 1. Word frequency in a particular context transformed into log frequency.
  - Compressive function yields a spacing effect: association of A & B is greater if both appear in 2 different contexts than if they appear twice in the one context.
- 2. All cell entries for a given word are divided by the entropy for that word. Result: Makes primary association better represent the informative relation bw the entities rather than the mere fact that they occurred together.
  - Inverse entropy measure estimates degree to which observing occurrence of a component specifies what context it is in. The larger entropy, the less information its observation transmits about the places it has occurred, so the less **usage-defined meaning** it acquires, and, conversely the less the meaning of a particular context is determined by the word.

- **LSA model may be similar to associative processes in information retrieval.**

Goal: Retrieve the texts from memory that person has in mind.

Question: Do we remember things 1) because we think that they are similar or, 2) because there is a general logic (in the information is processed) that makes them similar.

LSA offers a condensed representation of the relationship between data points by capturing higher-order associations.

  - If a particular stimulus, X (e.g. a word) has been associated with some other stimulus, Y, by being frequently found in joint context (i.e. contiguity), and Y is associated with Z, then the condensation can cause X and Z to have similar representations.
  - However, the strength of of the indirect XZ association depends on much more than a combination of the strengths of XY and YZ. It also depends on the relation of each of the stimuli X, Y, Z, to every other entity in the space.

- AKA: “induction of a latent higher order similarity structure among representations of a large collection of events.”

LSA adjusts with new input

Updating occurs throughout the space since each object is related in some way (more or less similar) to every other.

“The relation between any two representations depends not only on direct experience with them but with everything else ever experienced.”

LSA may also be analogized to a Three Layer Neural Net

· Layer 1: One node for every word type

· Layer 3: One node for every text window ever encountered.

· Layer 2: Several hundred nodes, number to be determined. This number is one which: maximizes accuracy (in a least squares sense) with which activating any Layer 3 node activates the Layer 1 nodes that are its elementary constituents. = a pattern of activation across Layer 2 nodes.

Singular Value Decomposition (SVD)

- SVD is a linear method for decomposing a matrix into independent principal components (factor analysis is an example of this).
- From Appendix: “Fundamental proof of SVD shows that there always exists a decomposition of this form such that matrix multiplication of the three derived matrices reproduces the original matrix exactly, so long as there are enough factors, where enough is always less than or equal to the smaller of the number of rows or columns of the original matrix.”

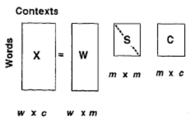


Figure A1. Schematic diagram of the singular value decomposition (SVD) of a rectangular word (*w*) by context (*c*) matrix (X). The original matrix is decomposed into three matrices: W and C, which are orthonormal, and S, a diagonal matrix. The *m* columns of W and the *m* rows of C′ are linearly independent.

Testing LSA: Four Questions

- Can a simple linear model acquire knowledge of humanlike word meaning similarities given a large amount of natural text?
- Would its success depend strongly on the dimensionality of its representation?
- How would its rate of acquisition compare with a human reading the same amount of text?
- How much knowledge would come from indirect references that combine information across samples vs. direct access from local contiguity?

Test 1: TOEFL Performance

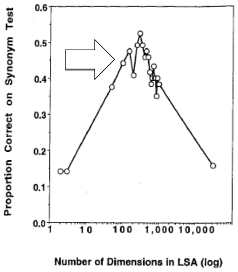
- LSA model first trained on 4.6 million words from Grolier's Academic American Encyclopedia. 30,743 articles, parsed as first 2000 characters (or entire text entry) => 151 words (a long paragraph).
- Text data cast as 30743 columns where each column represents one text sample. 60768 rows, each row represents a unique word type that appeared in at least 2 samples. Cells contained frequency with which a particular word type appeared in a particular text sample.
- SVD performed and 300 dimensions retained.
- 80 retired items from TOEFL (Test of English as a Foreign Language)
- Choice: 1 stem (problem) word and four choices. Need to choose one most similar to the stem word.

LSA Performance

- TOEFL Test
- Model: 51.5 out of 80= 64.4% correct (normalized to 52.5% when accounting for guessing)
- Actual foreign applicants to US colleges: 51.6/80 = 64.5% (52.7% corrected).
- LSA appears to mimic human performance.
- Question: Do foreign students know as much English as represented in 4.6m words from an encyclopedia?

Q2: Dimensionality

- Correct dimensionality is critical to success (how many dimensions to retain?).
- 2 dimensions retained: 13% correct.
- No reduction (all dimensions): 16% correct
- Question: We can figure out the correct dimensionality through trial and error –how would assigning the correct dimensionality work in real life? And do different datasets require different dimensions?



Q3: Learning Rate

- Caveat: LSA model learns similarities between words as units, not for their syntactic, grammatical properties, spelling, sounds, morphology etc. LSA is also not concerned with production.
- How well does LSA learn when compared with children at various age levels?
- · Assume a range of 7-15 words learned per day through high school. Children have to learn words by reading since they know more words by end of HS than available in speech (spoken vocab. Estimated to be 25% of print vocab.) Plus, there is very little direct instruction:

Learning rate estimates

- Children learn about one new word every five paragraphs (based one estimates of reading time, reading speed). How do children learn?
- Experiments to improve learning:
  - Jenkins, Stein, Woods (1984): 5th graders read paragraphs containing 18 low-frequency words 10x each over several days. When tested (fore choice definition), students scored between 5-10% correct.
  - Elley 1989, Nagy 1985: Learn by reading exposure: only 2.5 words per day (50 paragraphs at a rate of .05 learned per paragraph).
  - Conclusion: If these methods don't work how do children acquire their vocabulary?

LSA to the Rescue!

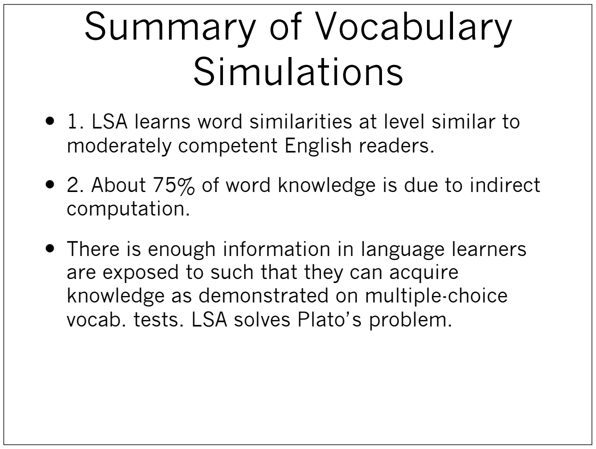
- Hypothesis: Children rely on indirect as well as direct learning. LSA captures the indirect portion.
- Target: Given text input similar to what a child receives, LSA should learn close to 10 words per day, thus accounting for natural rate.
- Implicit Idea: "Learning about a word's meaning from a textual encounter depends on knowing the meaning of the other words." LSA also captures this notion: "The reduced dimensional vector for a word is a linear combination of information about all other words."

Simulating a school setting

- Assume a child has read 3.8m words, equivalent to 25000 of encyclopedia samples.
- L&S estimated that direct effect was 0.0007 words gained per word encounter vs. 0.1500 words gained indirectly per word text sample read. Given average paragraph of 70 words= .0007*0+ .15= 0.20 words gained per paragraph x 50 paragraphs read per day on average by a student = 10 words learned a day.
- Thus LSA appears to account for the 10 words learned per day.

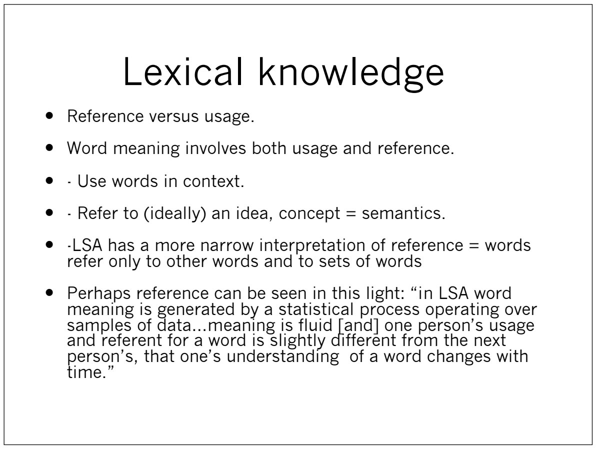
Other issues

- Does context window size matter? L&S control for window, shrinking it from 2000 characters to 500. Results were about the same on TOEFL test (53% vs 52%).
- Bags of words problem:
- LSA ignores grammar. In theory, LSA could learn from nominally scrambled sentences that in fact do not make sense.



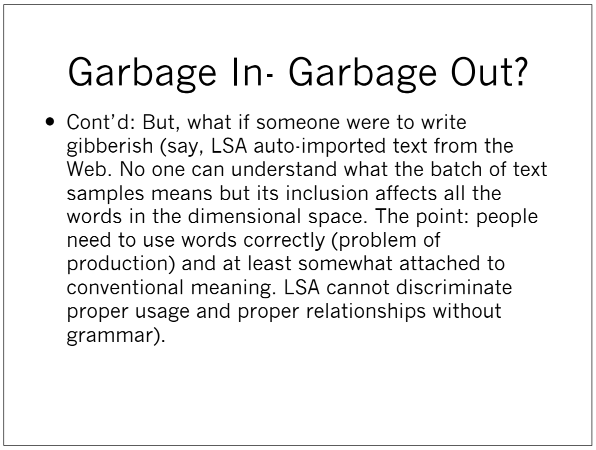
Summary of Vocabulary Simulations

- 1. LSA learns word similarities at level similar to moderately competent English readers.
- 2. About 75% of word knowledge is due to indirect computation.
- There is enough information in language learners are exposed to such that they can acquire knowledge as demonstrated on multiple-choice vocab. tests. LSA solves Plato's problem.



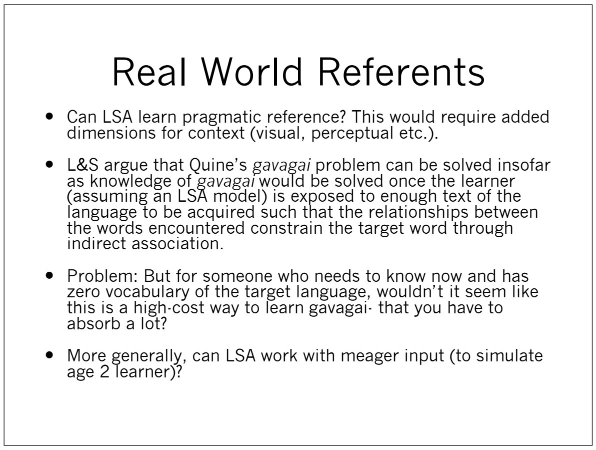
Lexical knowledge

- Reference versus usage.
- Word meaning involves both usage and reference.
- - Use words in context.
- - Refer to (ideally) an idea, concept = semantics.
- -LSA has a more narrow interpretation of reference = words refer only to other words and to sets of words
- Perhaps reference can be seen in this light: "in LSA word meaning is generated by a statistical process operating over samples of data…meaning is fluid [and] one person's usage and referent for a word is slightly different from the next person's, that one's understanding of a word changes with time."



Garbage In- Garbage Out?

- Cont'd: But, what if someone were to write gibberish (say, LSA auto-imported text from the Web. No one can understand what the batch of text samples means but its inclusion affects all the words in the dimensional space. The point: people need to use words correctly (problem of production) and at least somewhat attached to conventional meaning. LSA cannot discriminate proper usage and proper relationships without grammar).



Real World Referents

- Can LSA learn pragmatic reference? This would require added dimensions for context (visual, perceptual etc.).
- L&S argue that Quine's *gavagai* problem can be solved insofar as knowledge of *gavagai* would be solved once the learner (assuming an LSA model) is exposed to enough text of the language to be acquired such that the relationships between the words encountered constrain the target word through indirect association.
- Problem: But for someone who needs to know now and has zero vocabulary of the target language, wouldn't it seem like this is a high-cost way to learn gavagai- that you have to absorb a lot?
- More generally, can LSA work with meager input (to simulate age 2 learner)?

How LSA zeroes in on meaning

- Even in the absence of external referents, LSA can, “by resolving the mutual entailments of a multitude of other word-word, word-context, and context-context similarities” lead to agreement on the usage of a word or “make a word highly similar to a context even if it never occurs in that context.”

Contextual Disambiguation

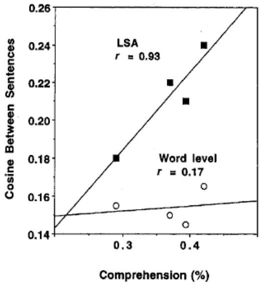
- Skilled readers disambiguate as they go.
- How does LSA handle words that have multiple meanings? E.g. line, fly, bear, man.
- For example, with the online LSA online at lsa.colorado.edu, you have to choose the database first before running LSA. Otherwise you may not get meaningful results.
- Do humans load full parameters which then disambiguate? How would you account for low frequency words that you encounter for the first time in a given context?
- Ex: “The player caught the high fly to left field.”

- LSA can capture local meaning if:
- 1. The input is a 3-way matrix of word, phrase, paragraph. The phrase vector holds the local context. Or,
- 2. A local re-representation process occurs such that a “secondary SVD-like condensation or some other mutual constrain satisfaction process using the global cosines as input that would have more profound meaning-revision effects than simple provision.”
- > Only components of meaning that it shares with the context, after averaging, comprise its disambiguated meaning.

Construction-Integration Theory

- How to use LSA to represent the meaning of text strings i.e. sentences or paragraphs?
- Goal: Estimate a measure of coherence that derives from the overlap of meaning in sentences as they build on each other.
- Experiment: 27 encyclopedia articles subjected to SVD in 100 dimensions. Each sentence in 4 experimental paragraphs represented as the average of the vectors it contained.
- Paragraph coherence measured as average cosine b/w successive sentences.

Figure shows measured comprehensibility for LSA (r=.93) and word level (r=.17)



LSA Simulation of Till et al (1988)

- Theoretically,
- 1) larger cosines b/w homographic word and its related words than between it and control words;
- 2) vector average of the passage words before homographic word should have higher cosine with context-relevant word related to than context-irrelevant word;
- 3) vector average of words in a passage should have higher cosine with the word related to the passage's inferred situational meaning than to control words.

- 28 pairs of passages, 112 target words. Cosines reveal LSA accomplishes correct inference.

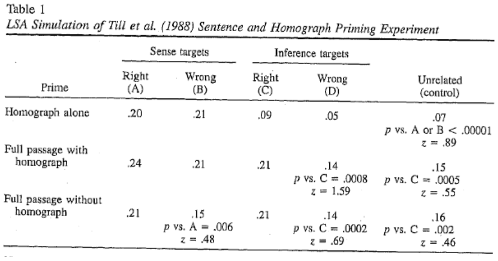
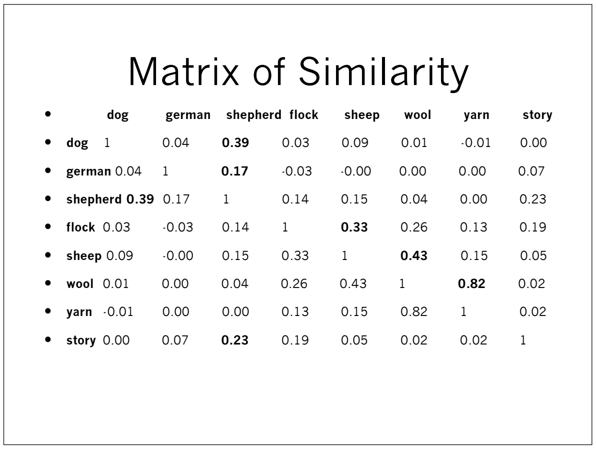


Table 1
LSA Simulation of Till et al. (1988) Sentence and Homograph Priming Experiment

	Sense targets		Inference targets		
Prime	Right (A)	Wrong (B)	Right (C)	Wrong (D)	Unrelated (control)
Homograph alone	.20	.21	.09	.05	.07 p vs. A or B < .00001 z = .89
Full passage with homograph	.24	.21	.21	.14 p vs. C = .0008 z = 1.59	.15 p vs. C = .0005 z = .55
Full passage without homograph	.21	.15 p vs. A = .006 z = .48	.21	.14 p vs. C = .0002 z = .69	.16 p vs. C = .002 z = .46

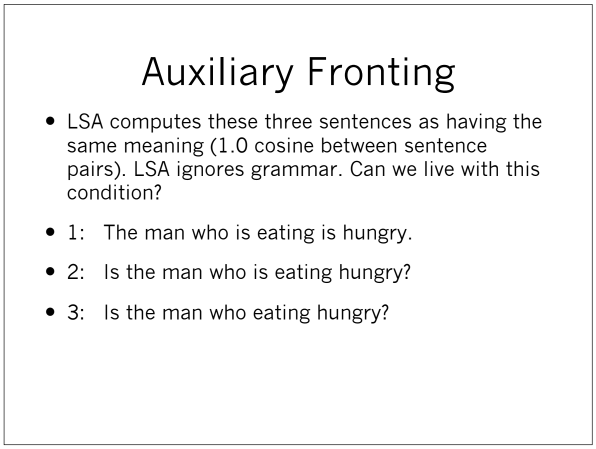
Problems with Target Selection

- Were the target homographs hand-selected? Does it matter whether words were pre-screened for variable meaning? Can LSA find the homographic words on its own or is this process a secondary step after running LSA once and getting ambiguous results?
- How would LSA work with compound-form words (in Enlish, Chinese, Navajo)?



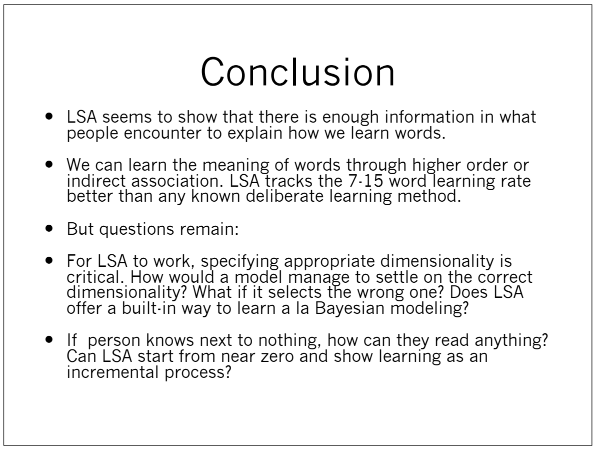
Matrix of Similarity

	dog	**german**	**shepherd**	**flock**	**sheep**	**wool**	**yarn**	**story**
dog	1	0.04	**0.39**	0.03	0.09	0.01	-0.01	0.00
german	0.04	1	**0.17**	-0.03	-0.00	0.00	0.00	0.07
shepherd	**0.39**	0.17	1	0.14	0.15	0.04	0.00	0.23
flock	0.03	-0.03	0.14	1	**0.33**	0.26	0.13	0.19
sheep	0.09	-0.00	0.15	0.33	1	**0.43**	0.15	0.05
wool	0.01	0.00	0.04	0.26	0.43	1	**0.82**	0.02
yarn	-0.01	0.00	0.00	0.13	0.15	0.82	1	0.02
story	0.00	0.07	**0.23**	0.19	0.05	0.02	0.02	1



Auxiliary Fronting

- LSA computes these three sentences as having the same meaning (1.0 cosine between sentence pairs). LSA ignores grammar. Can we live with this condition?
- 1: The man who is eating is hungry.
- 2: Is the man who is eating hungry?
- 3: Is the man who eating hungry?



Conclusion

- LSA seems to show that there is enough information in what people encounter to explain how we learn words.
- We can learn the meaning of words through higher order or indirect association. LSA tracks the 7-15 word learning rate better than any known deliberate learning method.
- But questions remain:
- For LSA to work, specifying appropriate dimensionality is critical. How would a model manage to settle on the correct dimensionality? What if it selects the wrong one? Does LSA offer a built-in way to learn a la Bayesian modeling?
- If person knows next to nothing, how can they read anything? Can LSA start from near zero and show learning as an incremental process?