

# Integrating abduction, visualization, and explanation as a data architecture for Artificial Intelligence (AI is still hard ;-)

**Randy Goebel**

*Principal Investigator*

*Professor of Computing Science*

[rgoebel@ualberta.ca](mailto:rgoebel@ualberta.ca)

+1.780.492.2683



# Outline

- Why do we need explanatory AI systems?
- Abduction
- Visualization
- Explanation
- Combining abduction, visualization, explanation and learning
- Prognosis

# Why we need explanatory AI systems

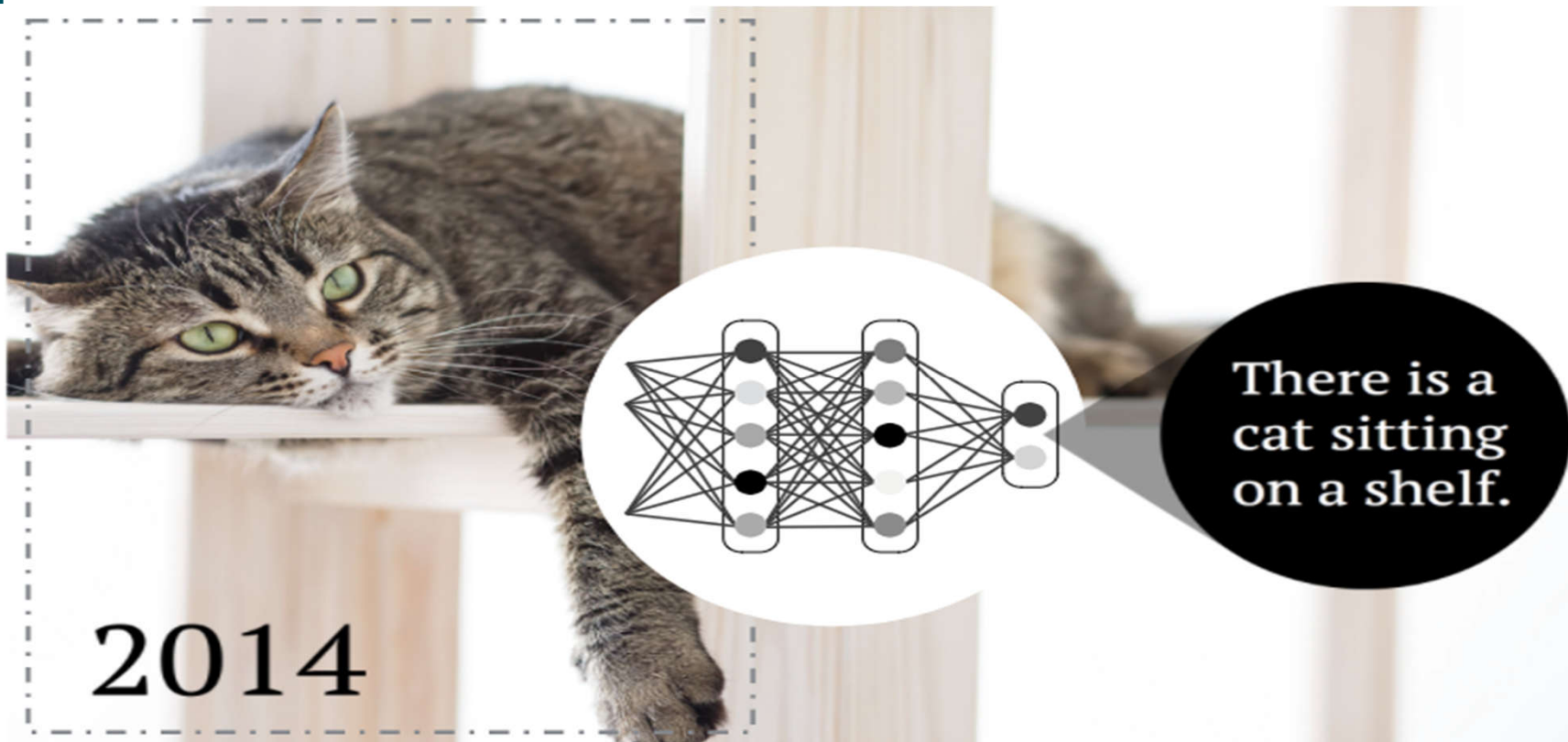


*“Does your car have any idea why my car pulled it over?”*

December 30, 2015

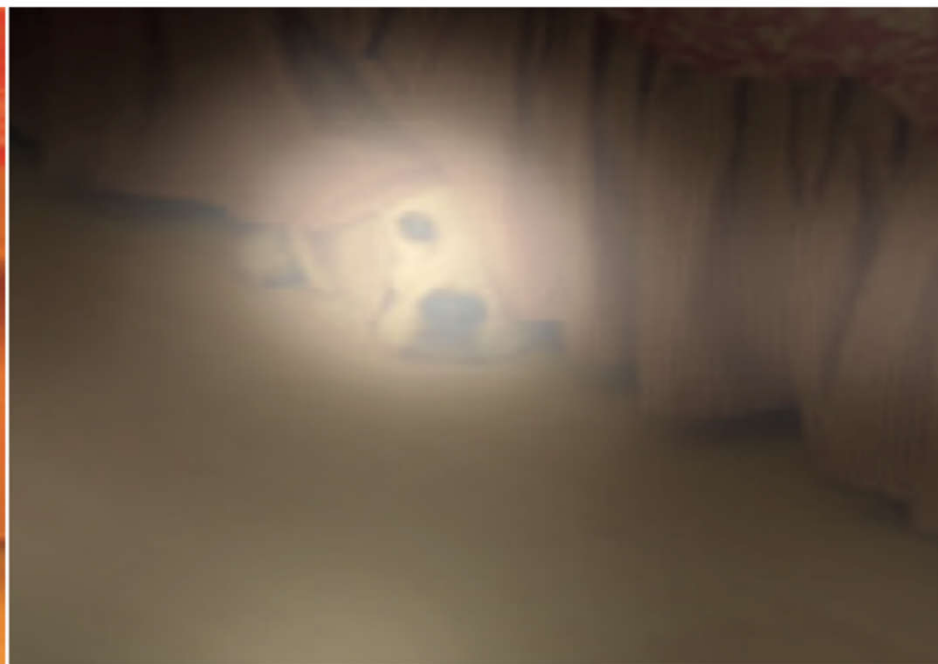
<https://www.newyorker.com/cartoon/a19697>

# This cat is not sitting



Slide courtesy of Hugo Larochelle, U de Sherbrooke

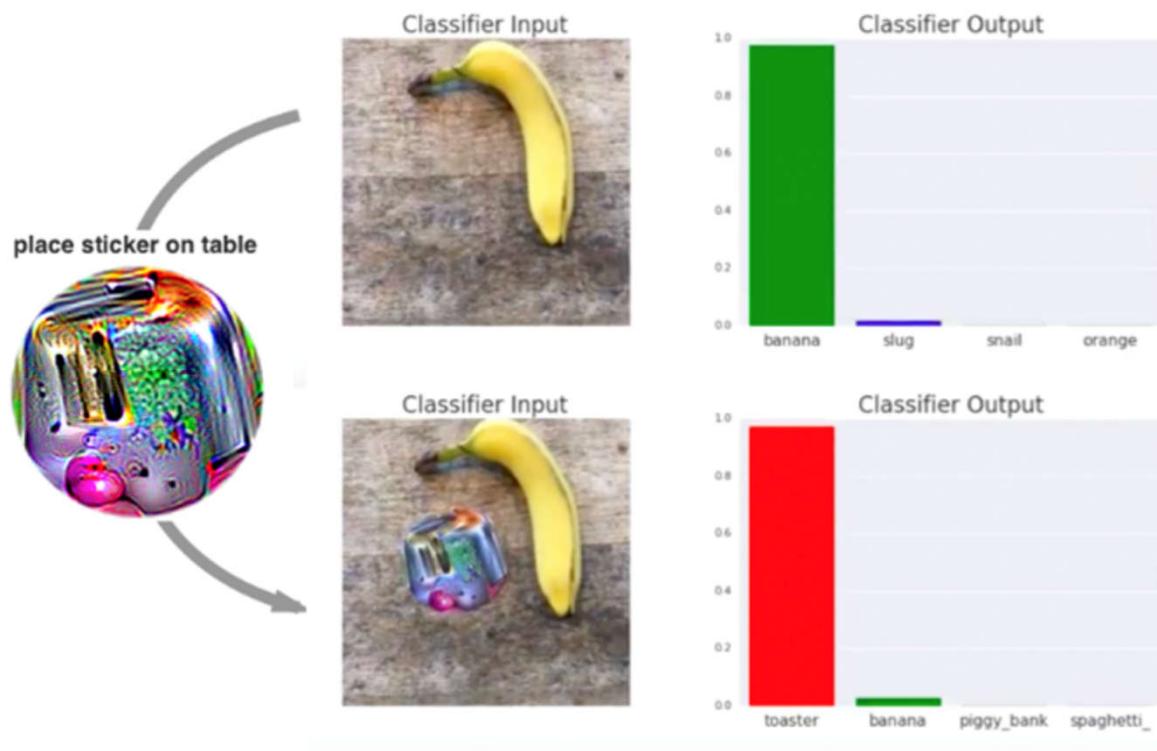
# This dog is not standing



A **dog** is standing on a hardwood floor.

LeCun, Bengio, Hinton 436 | NATURE | VOL 521 | 28 MAY 2015

# This banana is not a toaster



<https://gizmodo.com/this-simple-sticker-can-trick-neural-networks-into-thin-1821735479>

"Adversarial Patch," Tom B. Brown, Dandelion Mané\*, Aurko Roy, Martín Abadi, Justin Gilmer <https://arxiv.org/pdf/1712.09665.pdf>

## Explanation implications of GDPR

“The debate centers on the single occurrence of the phrase ‘right to explanation’ that occurs in Recital 71, a companion document to the GDPR that is not itself legally enforceable. However, the GDPR states that data controllers must notify consumers how their data will be used, including **‘the existence of automated decision-making, and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.’**”

<https://www.csoonline.com/article/3254130/compliance/what-does-the-gdpr-and-the-right-to-explanation-mean-for-ai.html>

# Abduction


- Abduction, C.S. Pierce
- Scientific explanation, verisimilitude, Karl Popper
- Causality, Pearl
- Properties of explanation
  - Presentations of alternative perspectives at multiple levels of detail
  - Alignment of *explainer* vocabulary with *explainee* vocabulary (die hausfrau und die vorstand)
  - Explanation quality: adequate -> ... trustworthy ... -> actionable
- Interpretability, Explainability, Instructability (IEI)




# Abductive explanation

- Deduction
  - from  $\forall x P(x) \supset Q(x)$ ,  $P(a)$  **deduce**  $Q(a)$
- Induction
  - from  $P(a)$  **induce**  $\forall x P(x)$
- Abduction
  - $\forall x P(x) \supset Q(x)$ ,  $Q(a)$  **abduce**  $P(a)$

**Observe** individual a has symptom Q



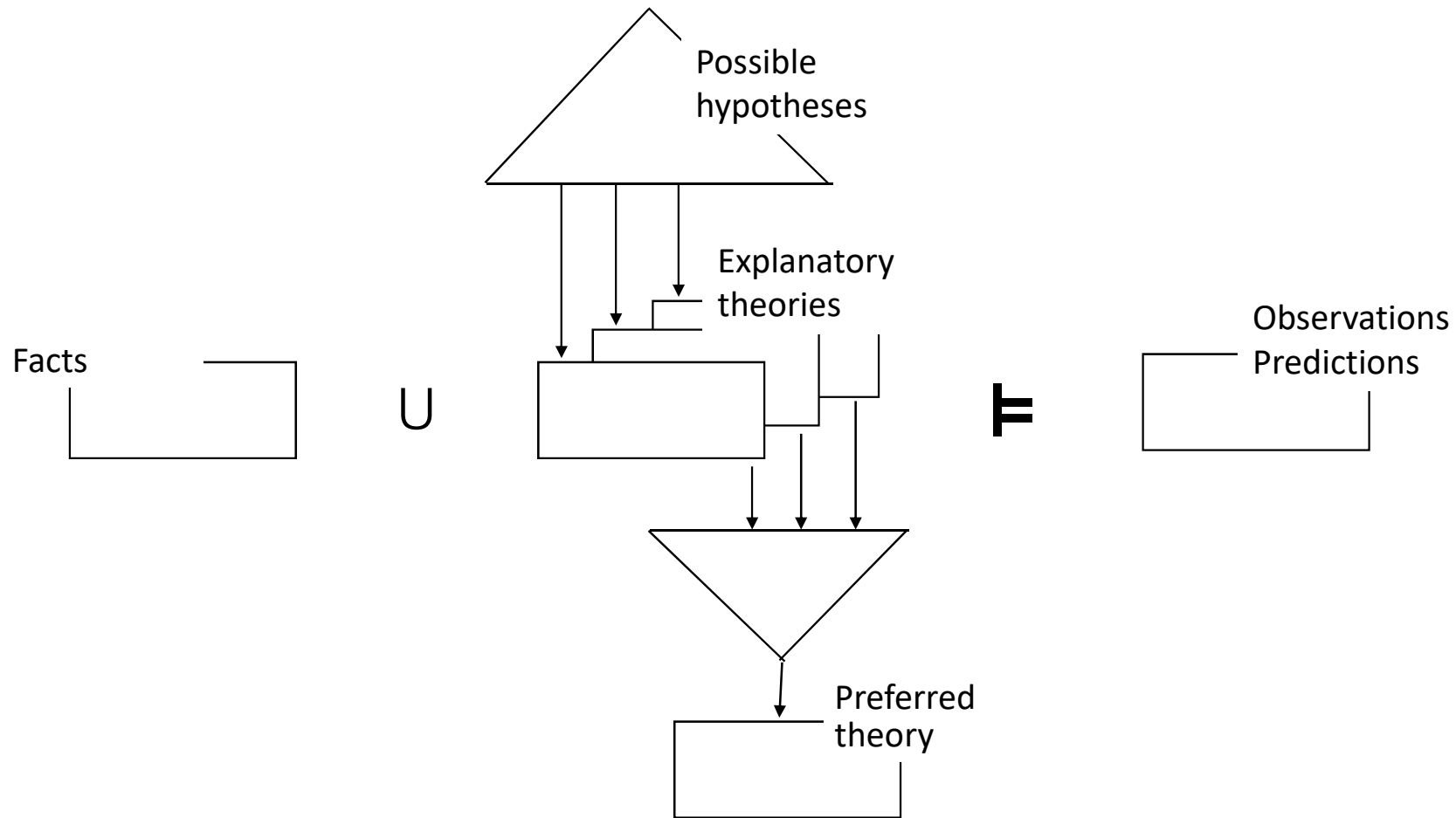

**Abduce (hypothesize)** that individual a has disease P



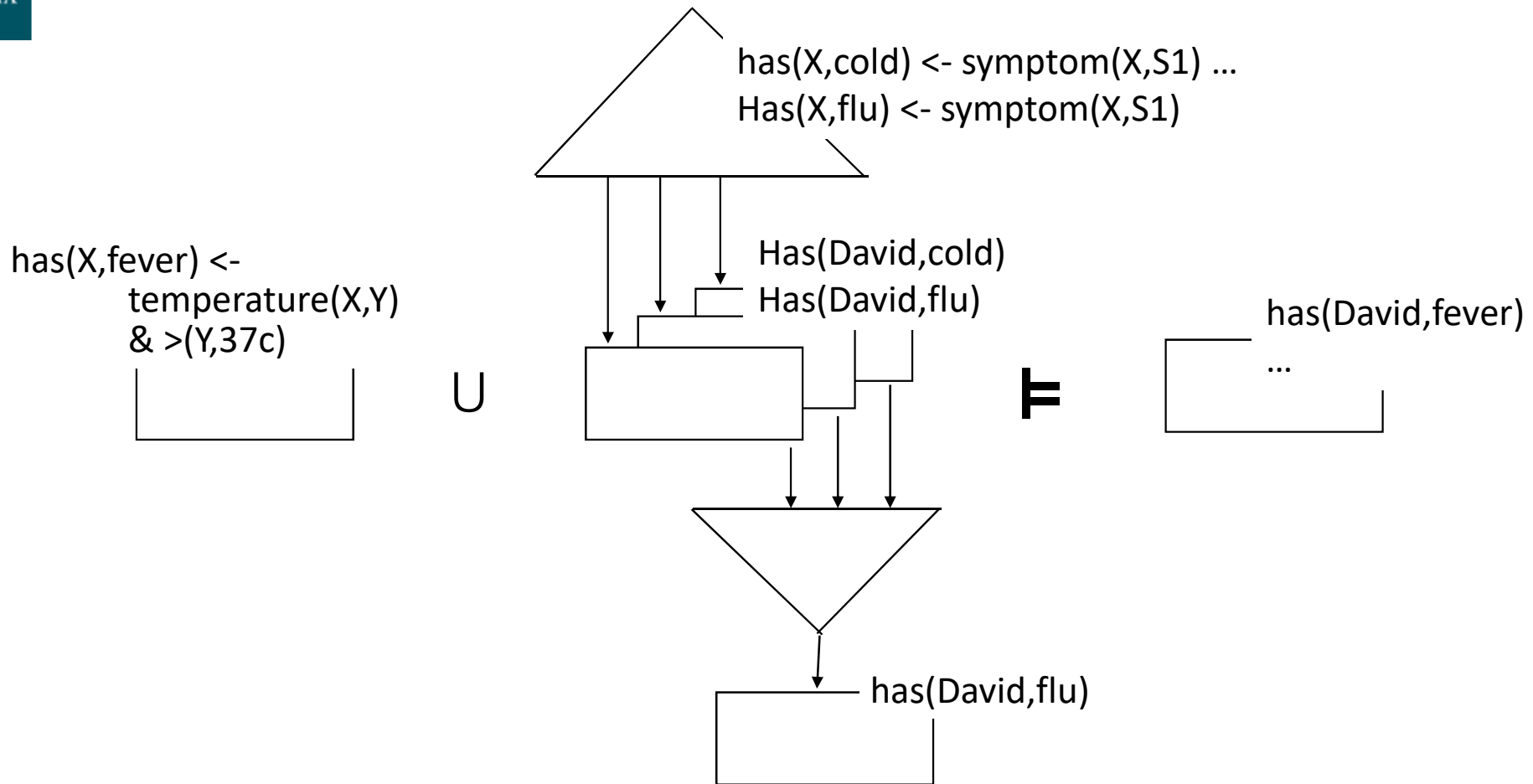
Domain knowledge,  
e.g., if x has disease P,  
then expect x to have  
symptom Q



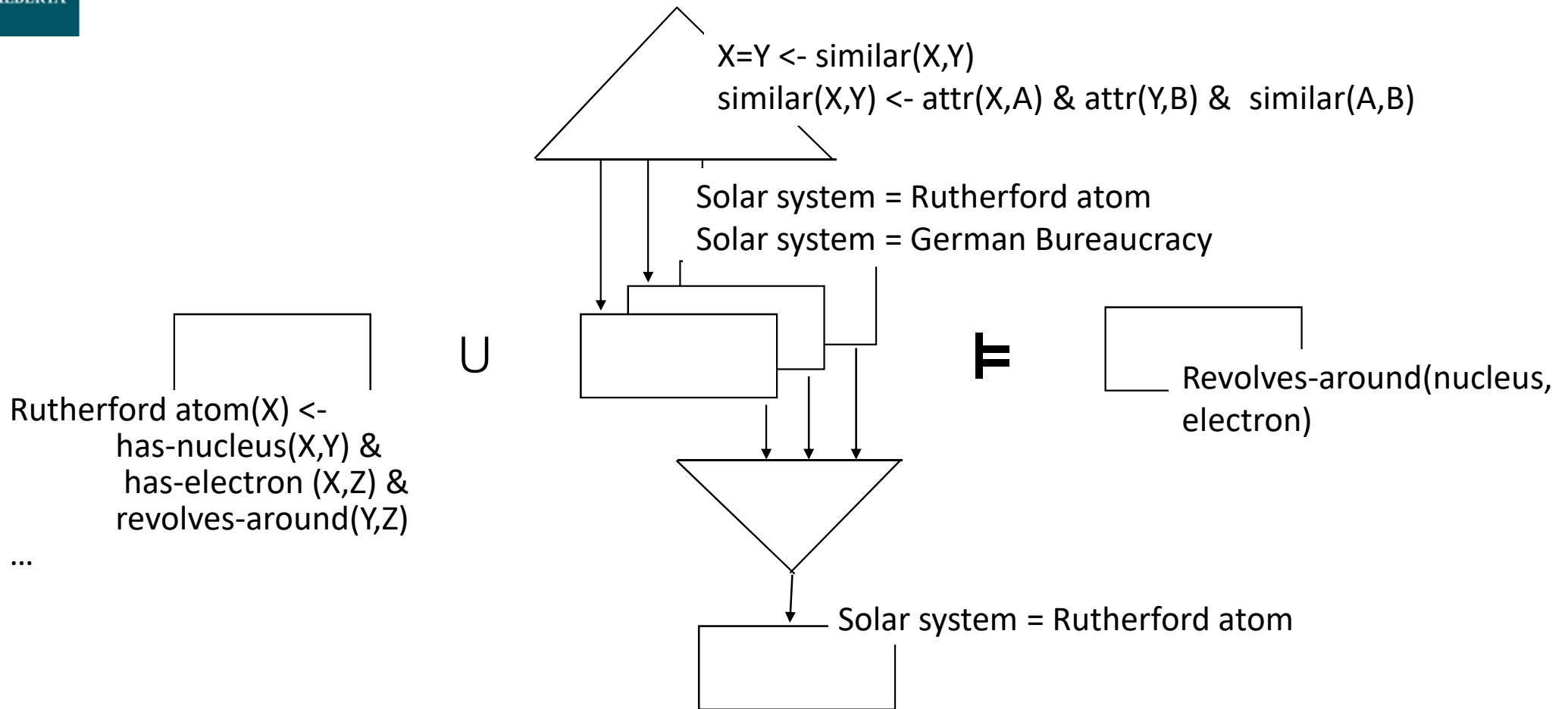
# Abductive Explanation (hypothetical reasoning)



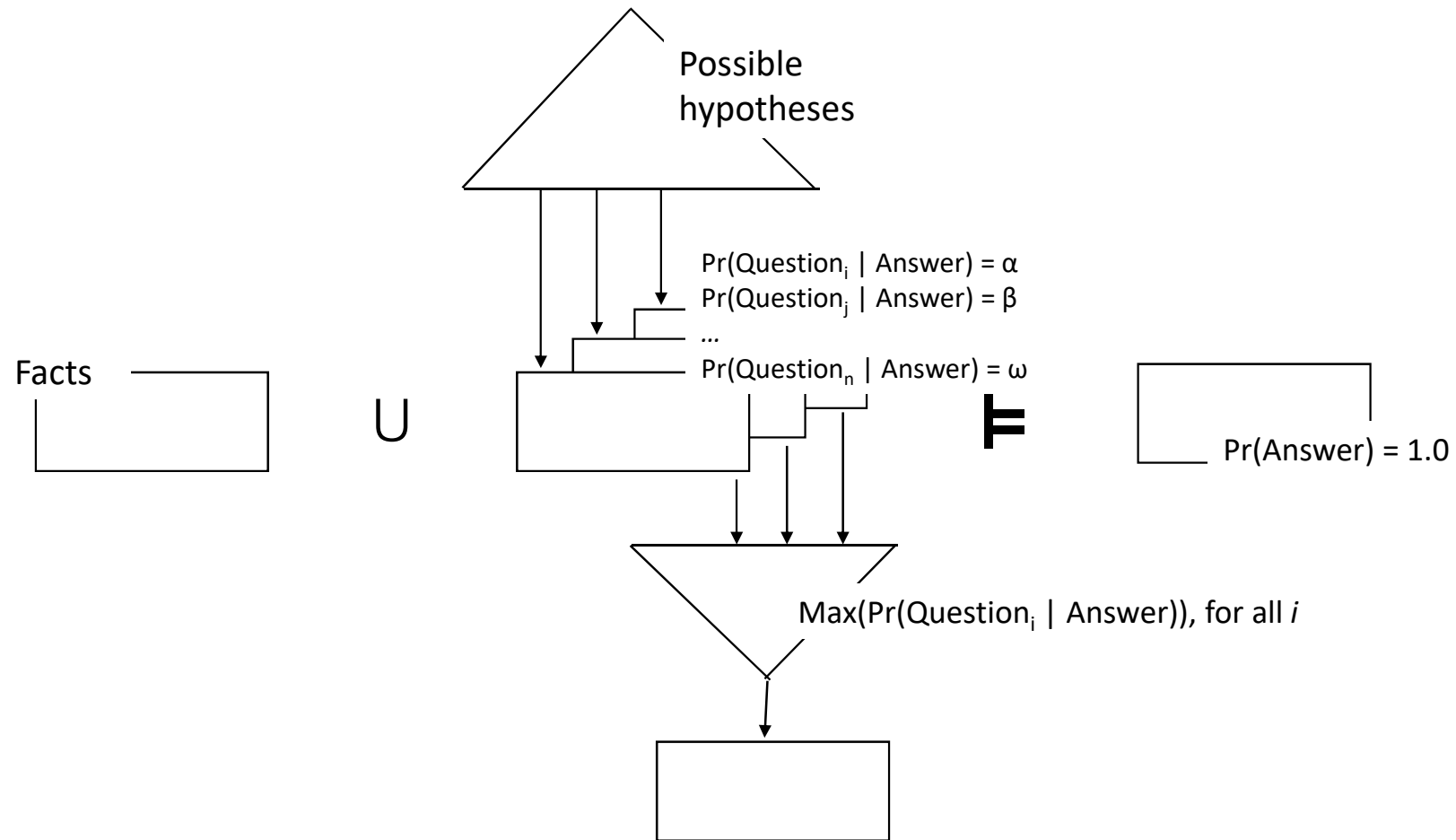
# Abductive Explanation (diagnosis)



# Abductive Explanation (analogical reasoning)



# Abductive Explanation (probabilistic ala Jeopardy/Watson)



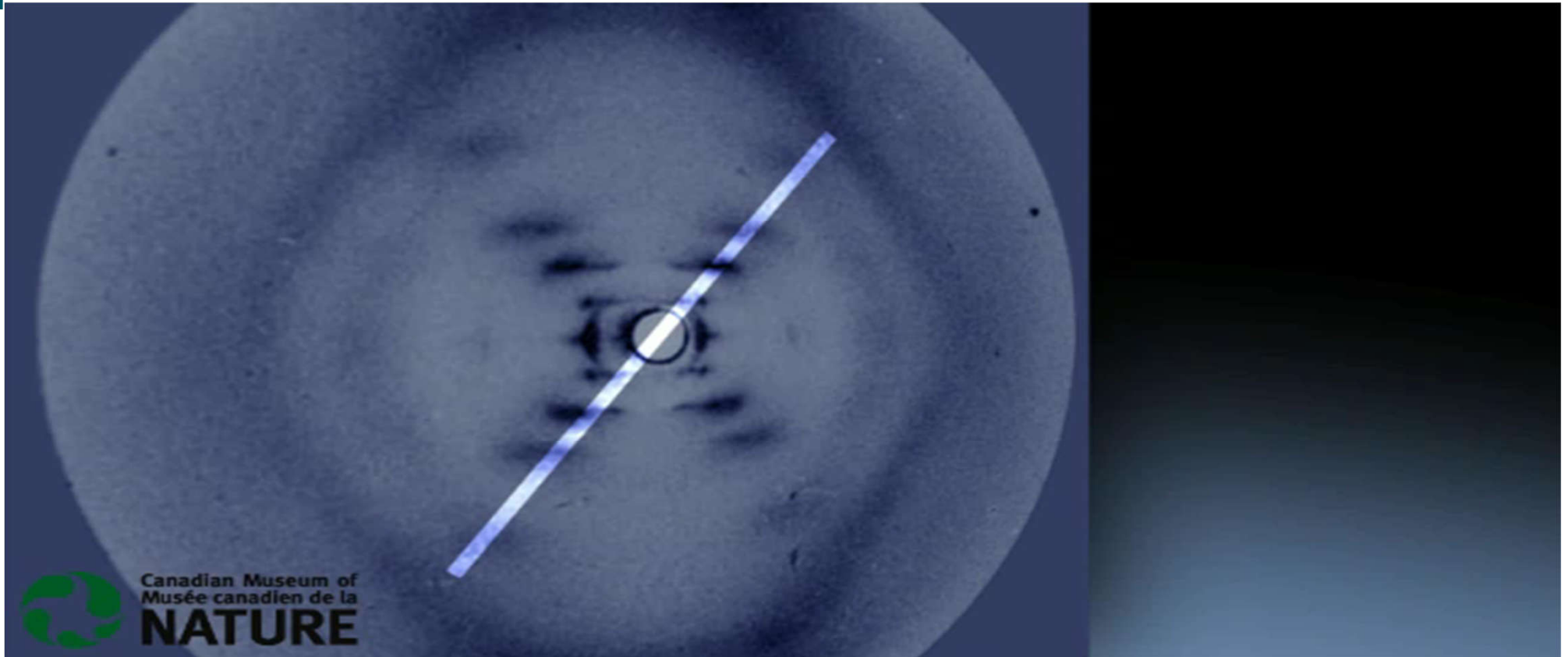


# Visualization (is explanation)

- Visualization IS explanation
- Semantic symmetry
- Propagating interaction in interactive visual explanation



# Explain Photo 51



Courtesy of Jay Ingram and the Canadian Museum of Nature, see also [http://en.wikipedia.org/wiki/Rosalind\\_Franklin](http://en.wikipedia.org/wiki/Rosalind_Franklin)

# Representation Granularity

Picture



Abstraction (Abstraction Vocabulary N)

•  
•  
•

Abstraction (Abstraction Vocabulary 1)



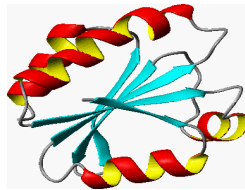
Source Data

discipline vocabularies





# Representation Granularity



## Secondary Structure (alpha Helix, Beta strand, random Coil)

```

CBBBBCHHH HHHHHHCCC CBBBBBBCC CCHHHHHHH HHHHHHHCC
CBBBBBBCC CCHHHHHHCC CCCCBBBBB BCCBBBBBBB CCCHHHHHHH
HHCC
  
```



## Primary Structure (sequence of amino acids)

```

MVKQIESKTA FQEALDAAGD KLVVDFSAT WCGPCKMIKP FFHSLSEKYS
NVIFLEVDVD DCQDVASECE VKCMPTFQFF KKGQKVGFEFS GANKEKLEAT
INELV
  
```



# Representation Granularity

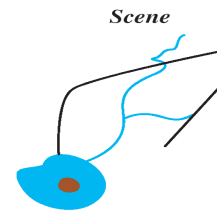
Sketch map scene (picture)



image domain drawing (picture)

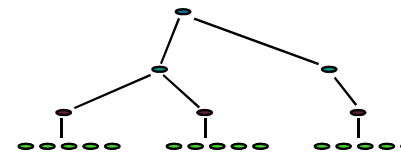
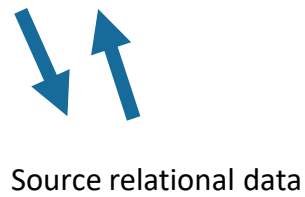
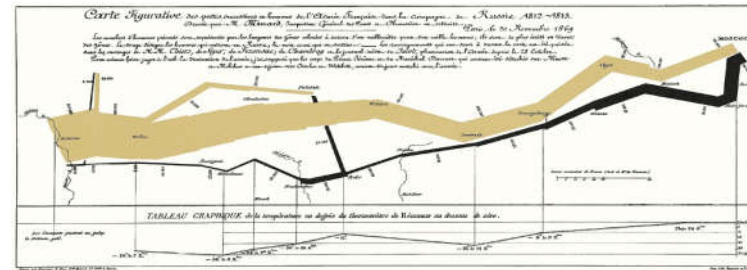


lines, circles (source data)



# Representation Granularity

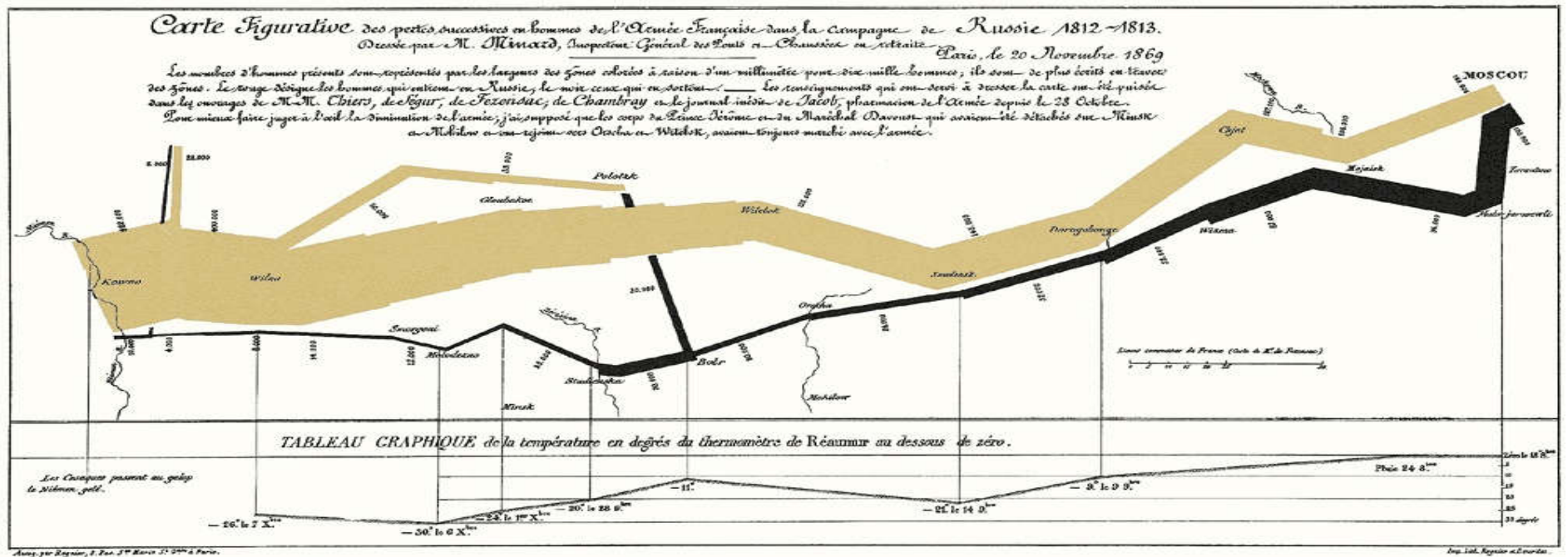
Minard's map



Date	#soldiers	Direction	Long.	Lat.	...
4/5/1812	64,464	34° NE	62.6	52.1	...
5/5/1812	63,262	30° NE	60.2	51.9	...
...	...	...	...	...	...

Shi, Goebel, Tanaka. "A New Database Visualization Framework for the Automatic Construction of Non-standard Charts: Re-creating the Chart of Napoleon's Russian Campaign of 1812." *Cartographica*, vol. 49 no. 4, 2014, pp. 241-261.

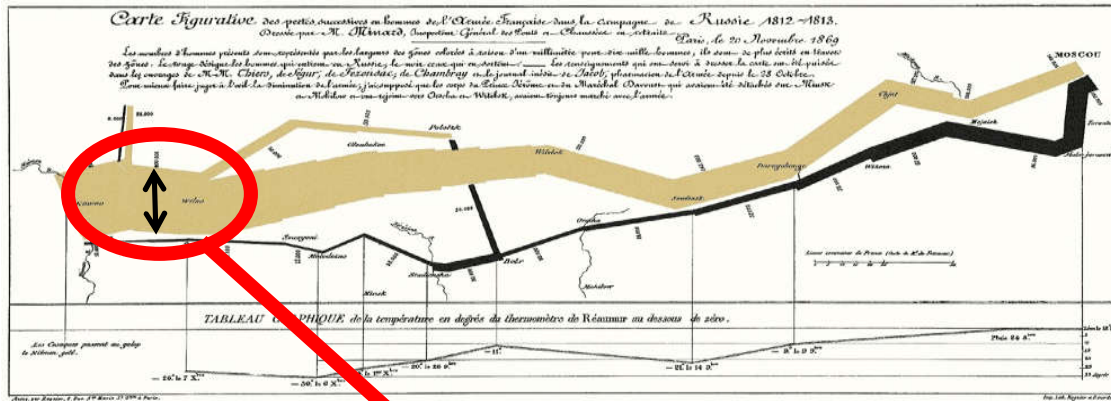
# Napoleon's Moscow campaign



<http://www.edwardtufte.com/tufte/minard/>  
 Extracted from Edward Tufte, *The Visual Display of Quantitative Information*, 1992

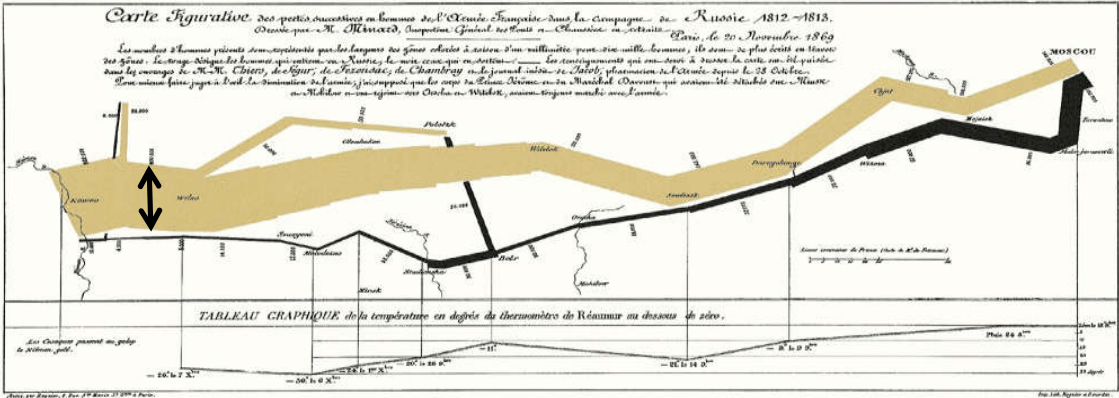
Charles Joseph Minard, 1869

# Direct manipulation semantics?



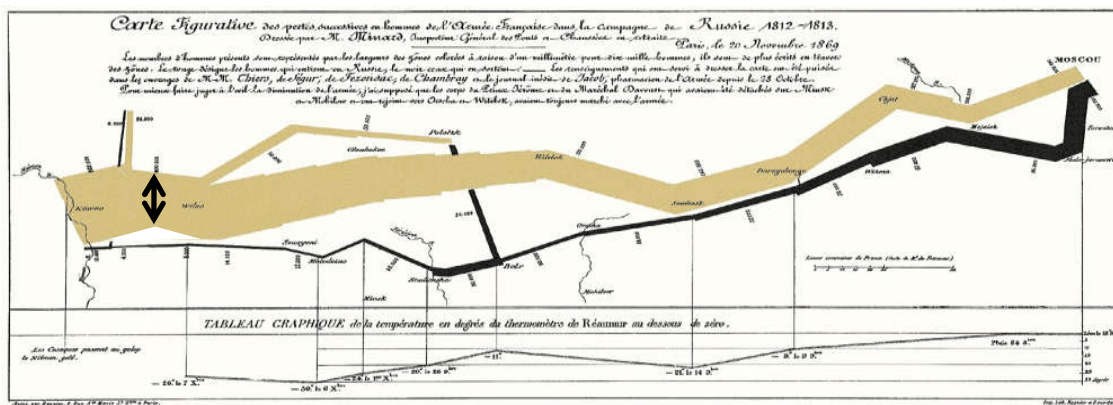
Date	#soldiers	Direction	Long.	Lat.	...
4/5/1812	64,464	34° NE	62.6	52.1	...
5/5/1812	63,262	30° NE	60.2	51.9	...
...	...	...	...	...	...

# Direct manipulation semantics?



Date	#soldiers	Direction	Long.	Lat.	...
4/5/1812	64,464	34° NE	62.6	52.1	...
5/5/1812	63,262	30° NE	60.2	51.9	...
...	...	...	...	...	...

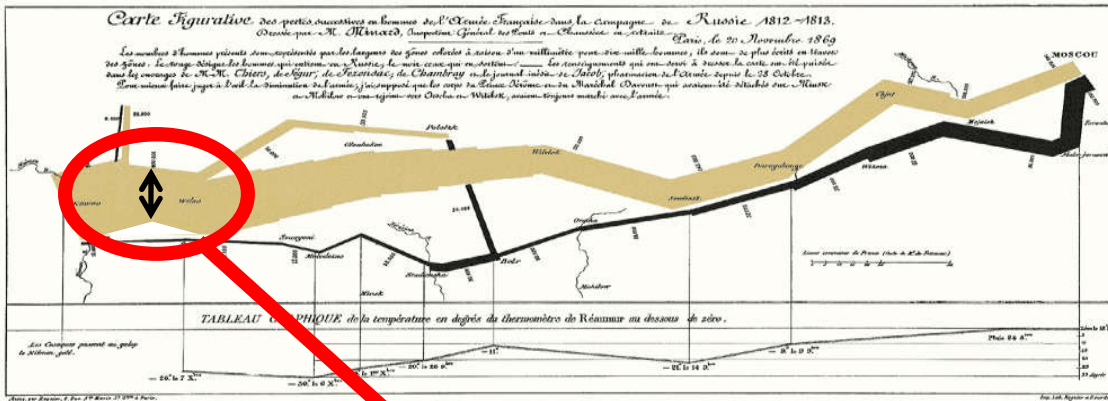
# Direct manipulation semantics?



Date	#soldiers	Direction	Long.	Lat.	...
4/5/1812	64,464	34° NE	62.6	52.1	...
5/5/1812	47,444	30° NE	60.2	51.9	...
...	...	...	...	...	...



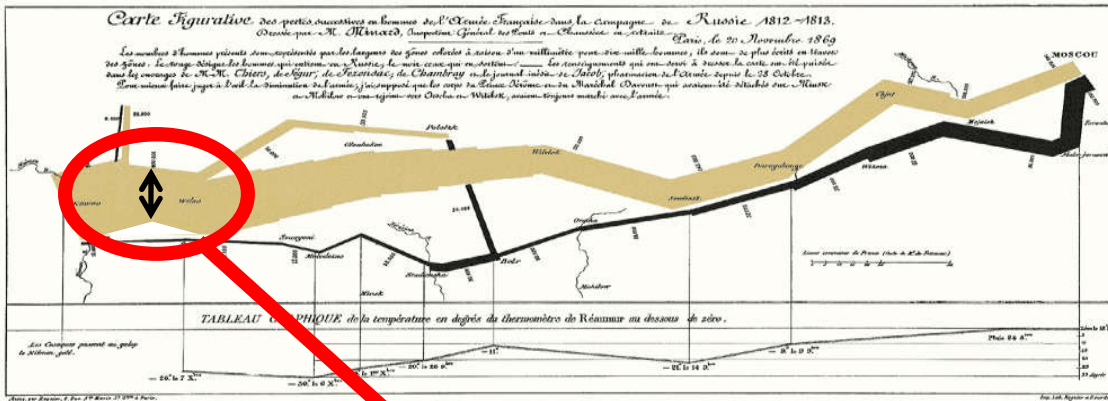
# Semantic Symmetry by Example



Date	#soldiers	Direction	Long.	Lat.	...
4/5/1812	64,464	34° NE	62.6	52.1	...
5/5/1812	47,444	30° NE	60.2	51.9	...
...	...	...	...	...	...



# Semantic Symmetry by Example



Date	#soldiers	Direction	Long.	Lat.	...
4/5/1812	64 464	34° NE	62.6	52.1	...
5/5/1812	?	30° NE	60.2	51.9	...
...	...	...	...	...	...

## Semantic Symmetry Pinch One

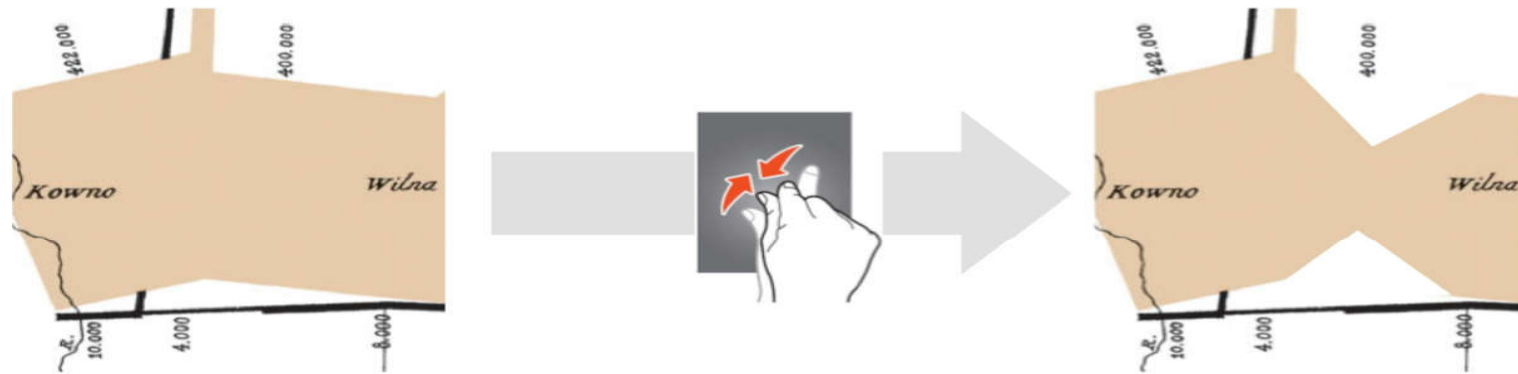


Fig. 4. Pinch One: Uniform narrowing with ambiguous shrinkage and regrowth.

## Semantic Symmetry Pinch Two

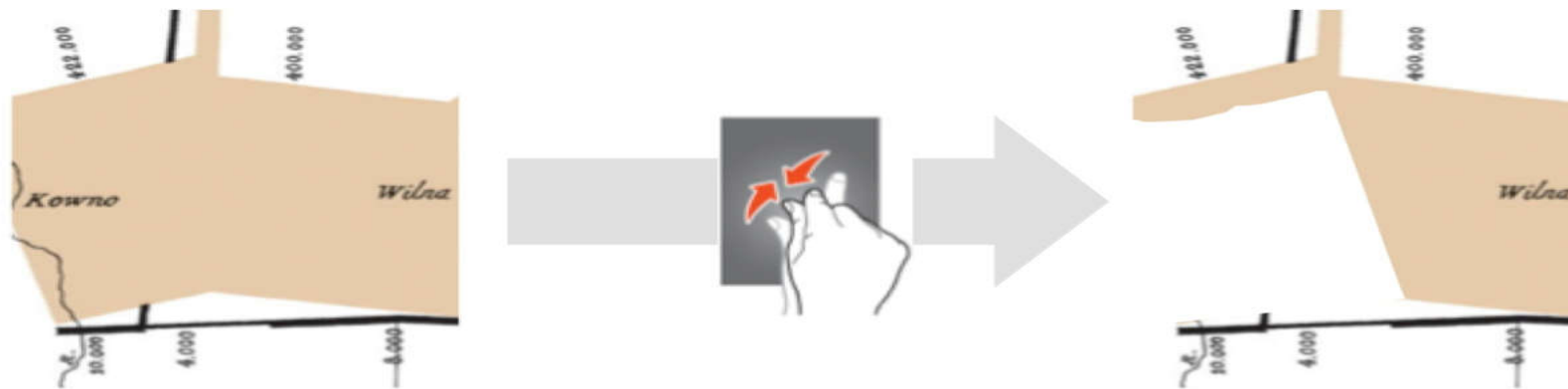


Fig. 5. Pinch Two: monotonic reduction propagating backwards in time.

## Semantic Symmetry Pinch Three

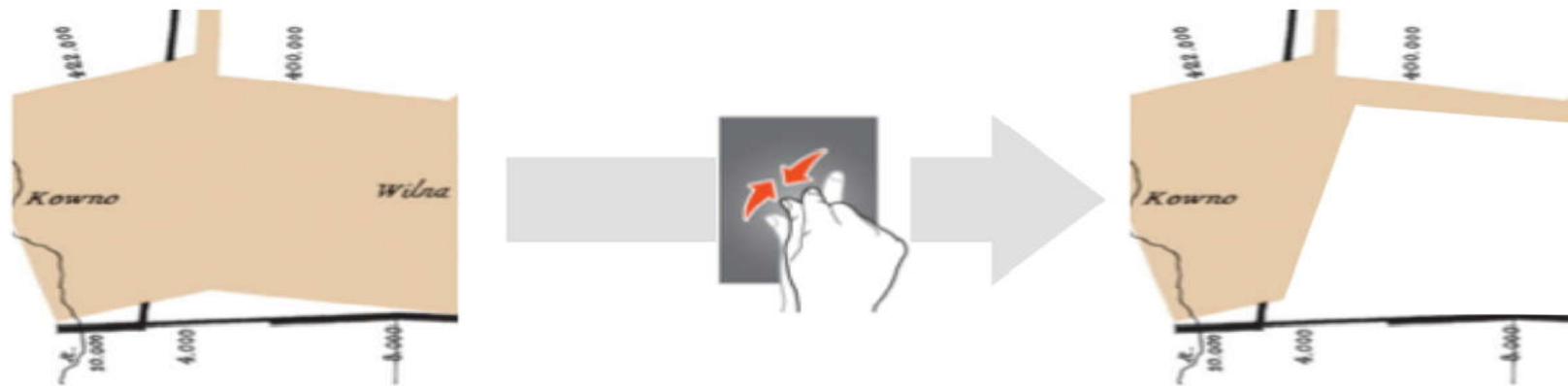


Fig. 6. Pinch Three: monotonic reduction propagating forwards in time.

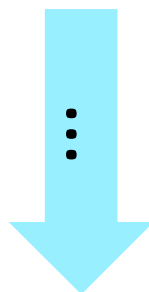
# Combining Explanation & Learning

- Debugging the empty program
- Scientific explanation
- Causality
- Properties of explanation
  - Presentations of alternative perspectives at multiple levels of detail
- Pre hoc, post hoc, simultaneous model building

# Debugging the empty program

- Algorithmic Debugging (Shapiro, 1980)

insert(1,[2],[1 2]).	positive example
Insert(3, [1 2], [3 1 2]).	negative example



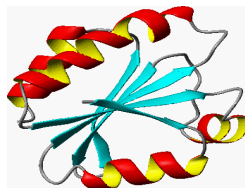
```

insert(X,[],[X]).
insert(X,[Y|T],[Y|NT]) :- X>Y, insert(X,T,NT).
insert(X,[Y|T],[X,Y|T]) :- X=<=Y.
    
```

# Scientific Explanation & Causality

- Scientific theories should be refutable
  - Representation needs to be expressive enough to admit contradictions
- Causality
  - Pearl's "estimands" are abductive explanations, (cf. "The Book of Why," p.17)
  - Causality requires multi-level representation

# From data to picture



## Secondary Structure (alpha Helix, Beta strand, random Coil)

```

CBBBBCHHH HHHHHHCCC CBBBBBBCC CCHHHHHHH HHHHHHHCC
CBBBBBBCC CCHHHHHHCC CCCCBBBBB BCCBBBBBBB CCCHHHHHHH
HHCC
  
```



## Primary Structure (sequence of amino acids)

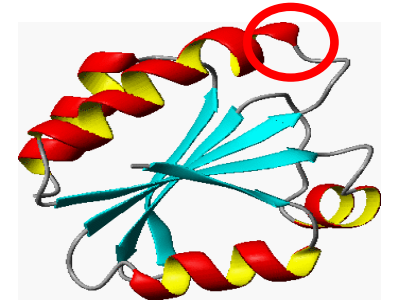
```

MVKQIESKTA FQEALDAAGD KLVVDFSAT WCGPCKMIKP FFHSLSEKYS
NVIFLEVDVD DCQDVASECE VKCMPTFQFF KKGQKVGFEFS GANKEKLEAT
INELV
  
```





# “Identifying causality ...”



## Secondary Structure (alpha Helix, Beta strand, random Coil)

CBBBBCCHHH HHHHHHCCC CBBBBBBBCC CHHHHHHHHH HHHHHHHHCC  
 CBBBBBBBCC CCHHHHHHCC CCCCBBBBB BCCBBBBBBB CCCHHHHHH  
 HHHCC



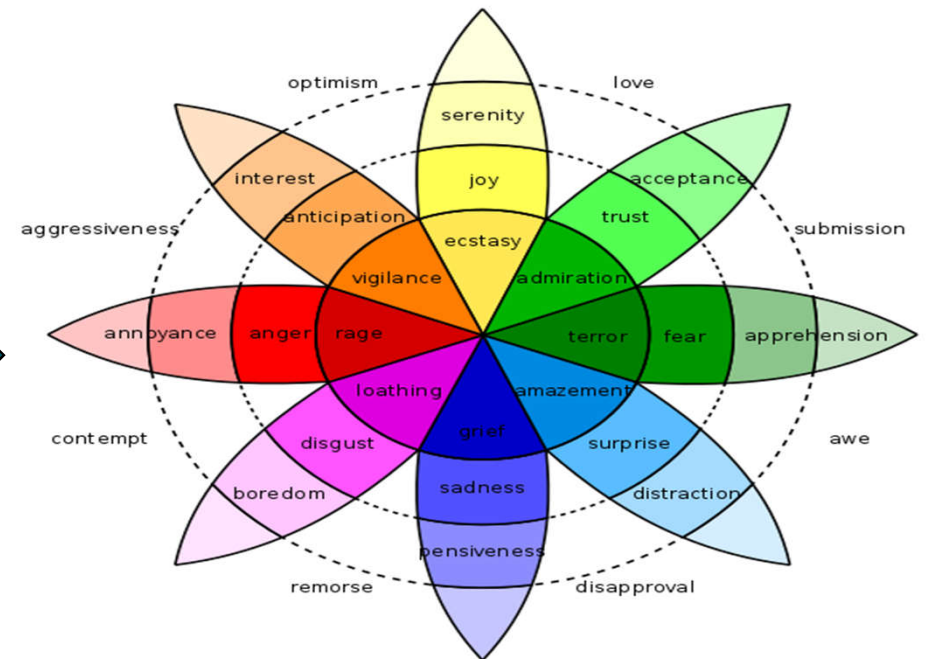
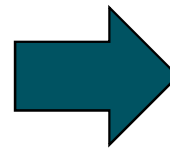
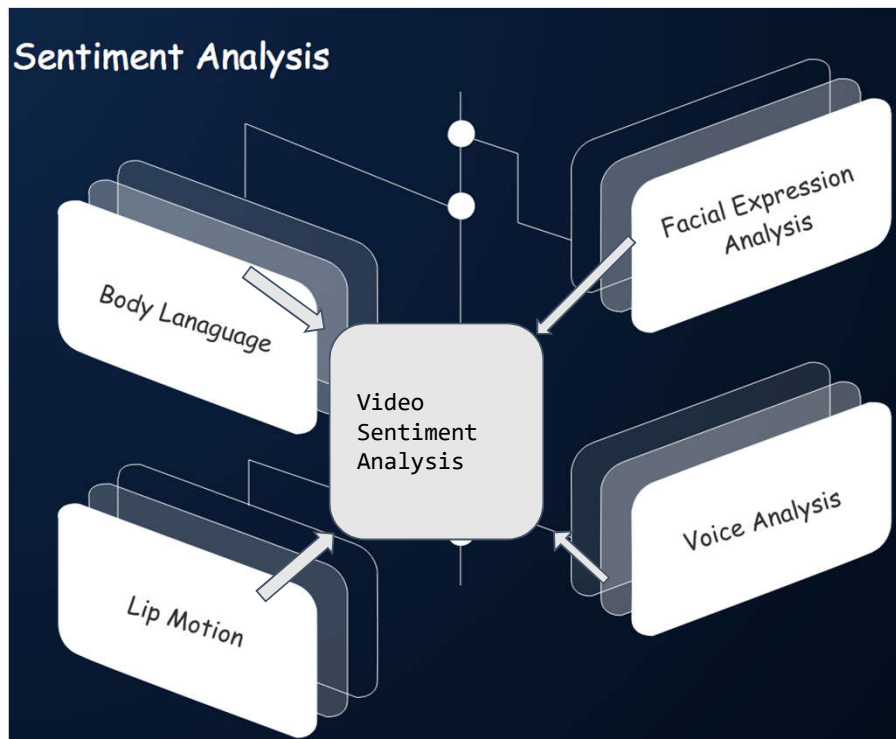
?

## Primary Structure (sequence of amino acids)

MVKQIESKTA FQEALDAAGD KLVVDFSAT WCGPCKMIKP FFHSLSEKYS  
 NVIFLEVDVD DCQDVASECE VKCMPTFQFF KKGQKVGEFS GANKERLQAT  
 INELV

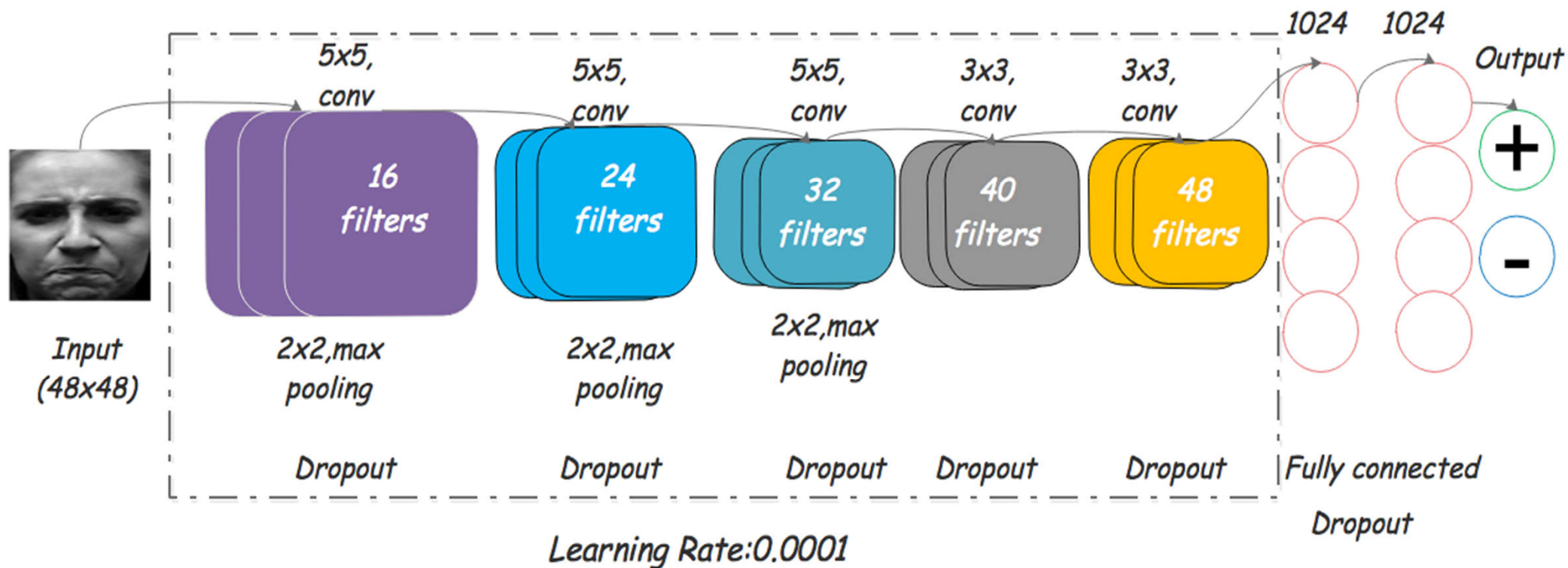


# Multi-modal sentiment analysis



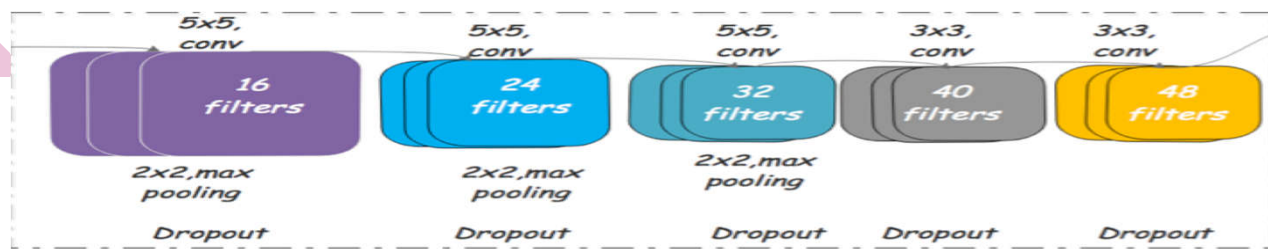
[https://en.wikipedia.org/wiki/Robert\\_Plutchik](https://en.wikipedia.org/wiki/Robert_Plutchik)

# One Deep Architecture Configuration

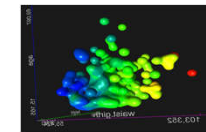
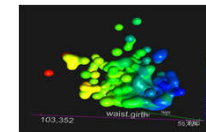
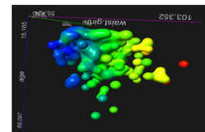
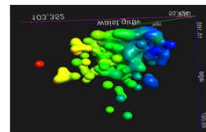
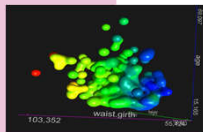


from Ph.D. Student Housam Khalifa Bashier Babiker

# Deep Visual Explanation

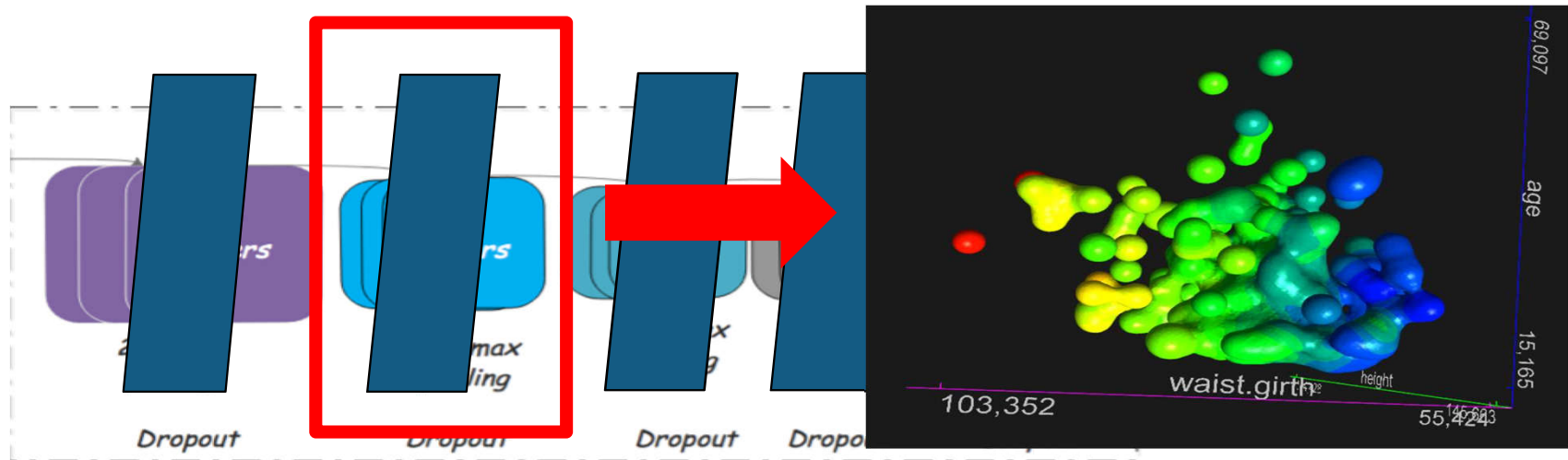


input stream



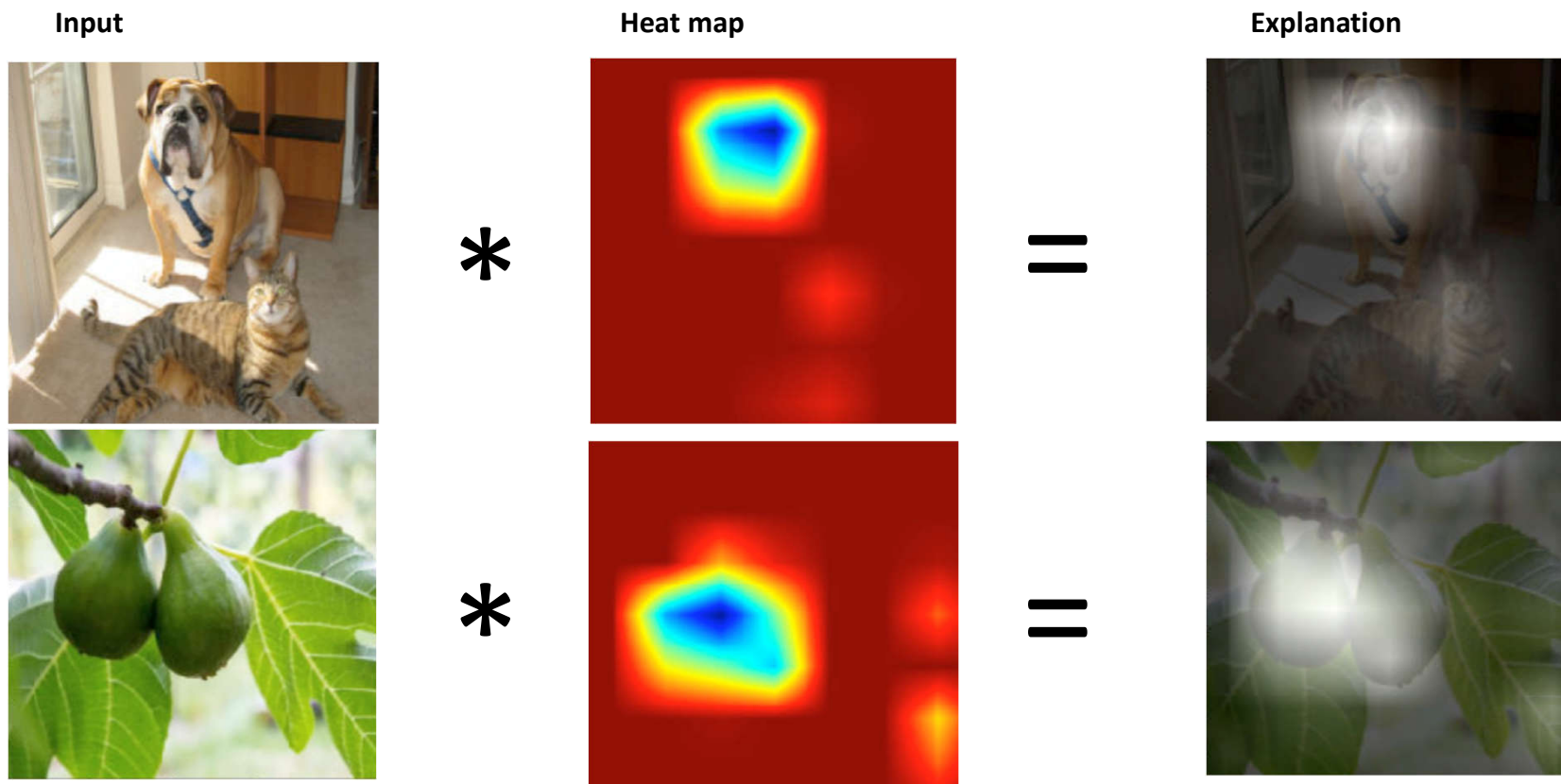
classifier performance

# Deep Visual Explanation



<http://cloudnsci.fi/wiki/index.php?n=HeatMiner.4dHeatmapDemo>

# Deep Visual Explanation





# Abstractions of Documents

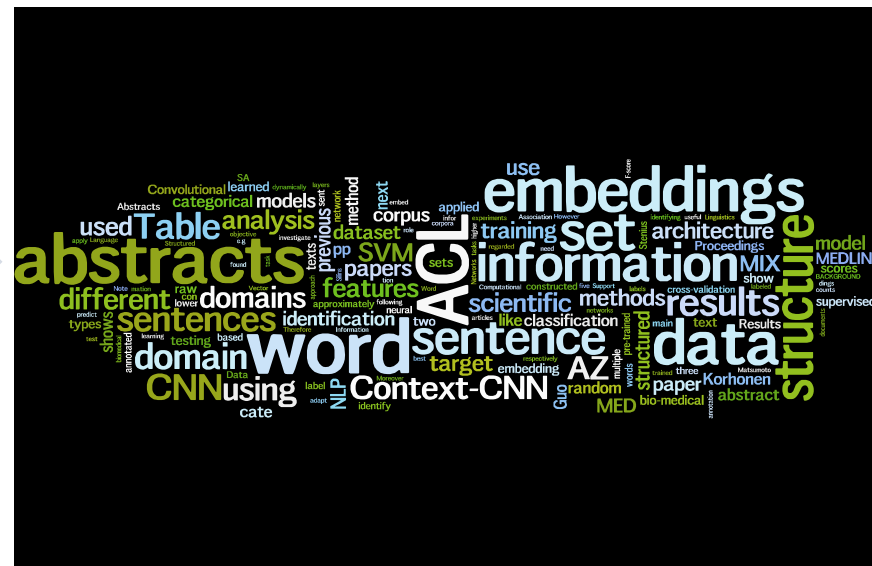
## Information structure analysis of abstracts in multiple domains using word embeddings.

Mai Omura, Hiroyuki Shindo, and Yuji Matsumoto

Nara Institute of Science and Technology,  
8916-5 Takayama, Ikoma, Nara 630-0192, Japan  
{omura.mai.oz5,shindo,matsu}@is.naist.jp

**Abstract.** We present an automated approach to identify information structure of abstracts of scientific papers, by classifying the sentences into functional categories like background, objective, method and results. Structured Abstracts often found in bio-medical domains can be used for training examples to identify the information structure. However, it is difficult to directly adapt the trained model to the abstracts in other domains (e.g. NLP domain). We show the results of our experiments to identify information structure of abstracts applied to different domains using word embeddings learned on single or multiple domains to see their effectiveness for domain adaptation.

**Keywords:** structured abstract, sentence classification, multi-domain, convolutional neural network



# Explaining Language Use is much harder!

directing and below average acting the thing that really irritated me was the blatant advertising constantly for a well known internet <UNK> it is obvious some scenes are written to do just that advertising this movie is a slap in the face to anyone who <UNK> money for this br br do not watch this it not worth your time

the somewhat disappointing planet terror and the rather fun flight of the living dead and to my surprise it is also the best director andrew <UNK> has given fans of the genre something truly original to treasure and is a talent to be watched in the future br br 8.5 out of 10 rounded up to 9 for imdb

doubt in my movie top 3 together with cinema <UNK> which is also a masterpiece the soundtrack is also really good i am really curious about <UNK> von <UNK> new movie i hope it will complete my movie top 3 if you see this movie rent it or even better buy it because you will want to see it again



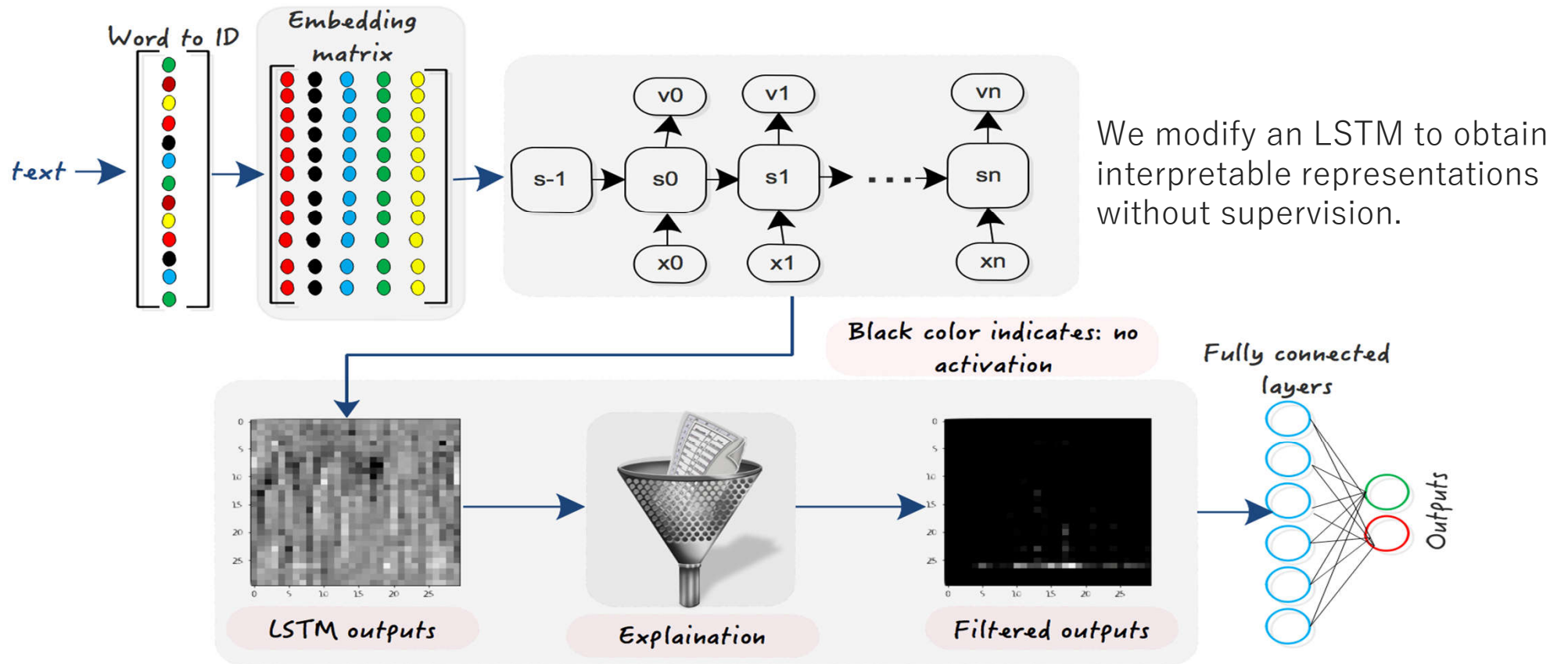
# Explaining Language Use is much harder!

directing and below average acting the thing that really irritated me was the blatant advertising constantly for a well known internet <UNK> it is obvious some scenes  
are written to do just that advertising this movie is a slap in the face to anyone who <UNK>  
money for this br br do not watch this it not worth your time  
(Negative Sentiment)

the somewhat disappointing planet terror and the rather fun  
flight of the living dead and to my surprise it is also the best director andrew <UNK> has  
given fans of the genre something truly original to treasure and is a talent  
to be watched in the future br br 8.5 out of 10 rounded up to 9 for imdb  
(Positive Sentiment)

doubt in my movie top 3 together with cinema <UNK> which is also a  
masterpiece the soundtrack is also really good i am really curious about <UNK> von <UNK>  
new movie i hope it will complete my movie top 3 if you see this movie  
rent it or even better buy it because you will want to see it again  
(Positive Sentiment)

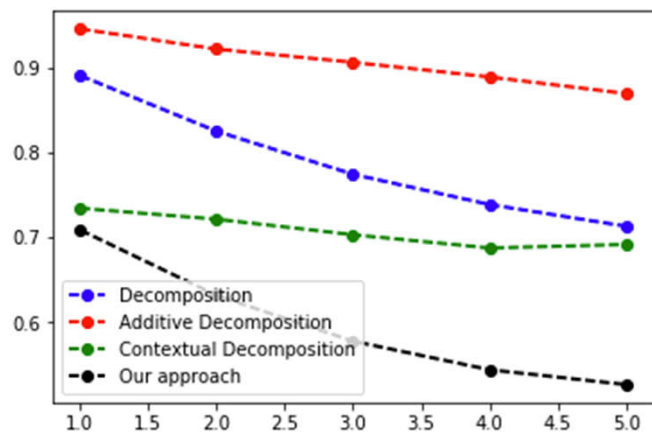
# Learning Explanations



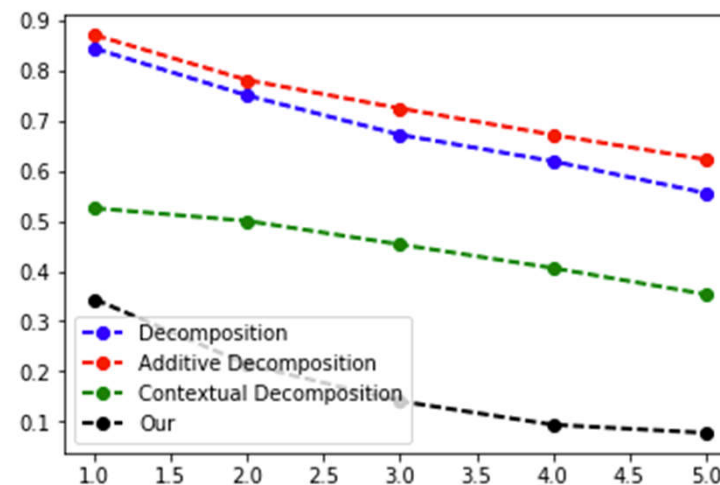
# Examples

- The goods home and I would definitely **recommend** you seek **it** out after reading about the movie I was expecting something very strong but it that by a bunch. (**Positive Sentiment**)
- The damned this cheap horror exploitation flick the making of alive some fifteen years later that film was a **masterpiece survive** to put it mildly is not. (**Positive Sentiment**)
- One of my great guilty **pleasures** I was fortunate enough to find it on an old and have watched it more times than I think is healthy a **worthwhile**. (**Positive Sentiment**)
- Movie is nothing but torture and cruel and unusual **punishment** to watch a bunch of drab **boring** scenes with unoriginal characters speaking in that wretched forced and fake English accent. (**Negative Sentiment**)

# Word deletion experiments



**IMDB dataset**



**Reuters dataset**

Murdoch, W. James, and Arthur Szlam. "Automatic rule extraction from long short term memory networks." *ICLR* (2017).

Murdoch, W. James, Peter J. Liu, and Bin Yu. "Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs." *ICLR* (2018).

# Prognosis

- Alternative explanations
  - Use deep learning to build explanations (doomed)
  - Engineer models to provide approximate explanation (doomed)
- Identifying multiple levels of abstraction to provide vocabulary for constructing multi-level models
  - Inducing granularity boundaries is very difficult (cf. Hurst coefficients, self-similarity), and likely non-deterministic
  - Multi-level model connections are necessary for causal explanation
- Formal language abstractions can also be learned
  - Grammatical induction
  - Debugging the empty program

# How and when to build multiple models?

Abstraction Vocabulary N+?



Abstraction Vocabulary N



Abstraction Vocabulary 1



Source Data

Probability: X is a cat with probability Y if X probably has pointy ears  
 Probability: X is a cat with probability Y if X probably has floppy ears

...

Default rule: if X has pointy ears it is a cat  
 Default rule: if X has floppy ears it is a dog

...

