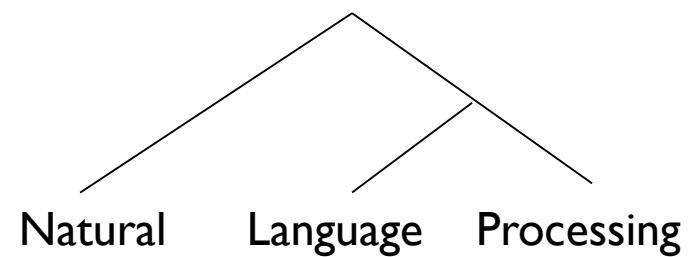
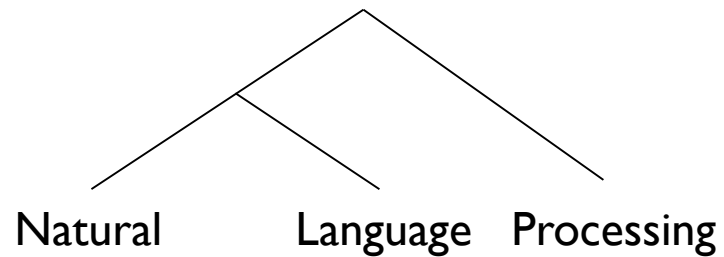


Natural Language Processing



Goal of NLP

The goal of the Natural Language Processing (NLP) group is to design and build software that will analyze, understand, and generate languages that humans use naturally, so that eventually you will be able to address your computer as though you were addressing another person.

NLP has two parts

- Natural Language Understanding (NLU)
 - INPUT : natural language text
 - OUTPUT : representation of the meaning of the text
- Natural language Generation (NLG)
 - INPUT : some non-textual representation (say a graph, or a time series)
 - OUTPUT : natural language text

$$\text{NLP} = \text{NLU} + \text{NLG}$$

- Which of Understanding and Generation is harder?



Understanding



Generation

Why Study NLP ?

- Two criteria must be satisfied:
 - It must be **important**
 - It must be **interesting**
- Importance : Impact and/or potential impact on our everyday life
- Interesting : New insights, fresh challenges

Why bother?

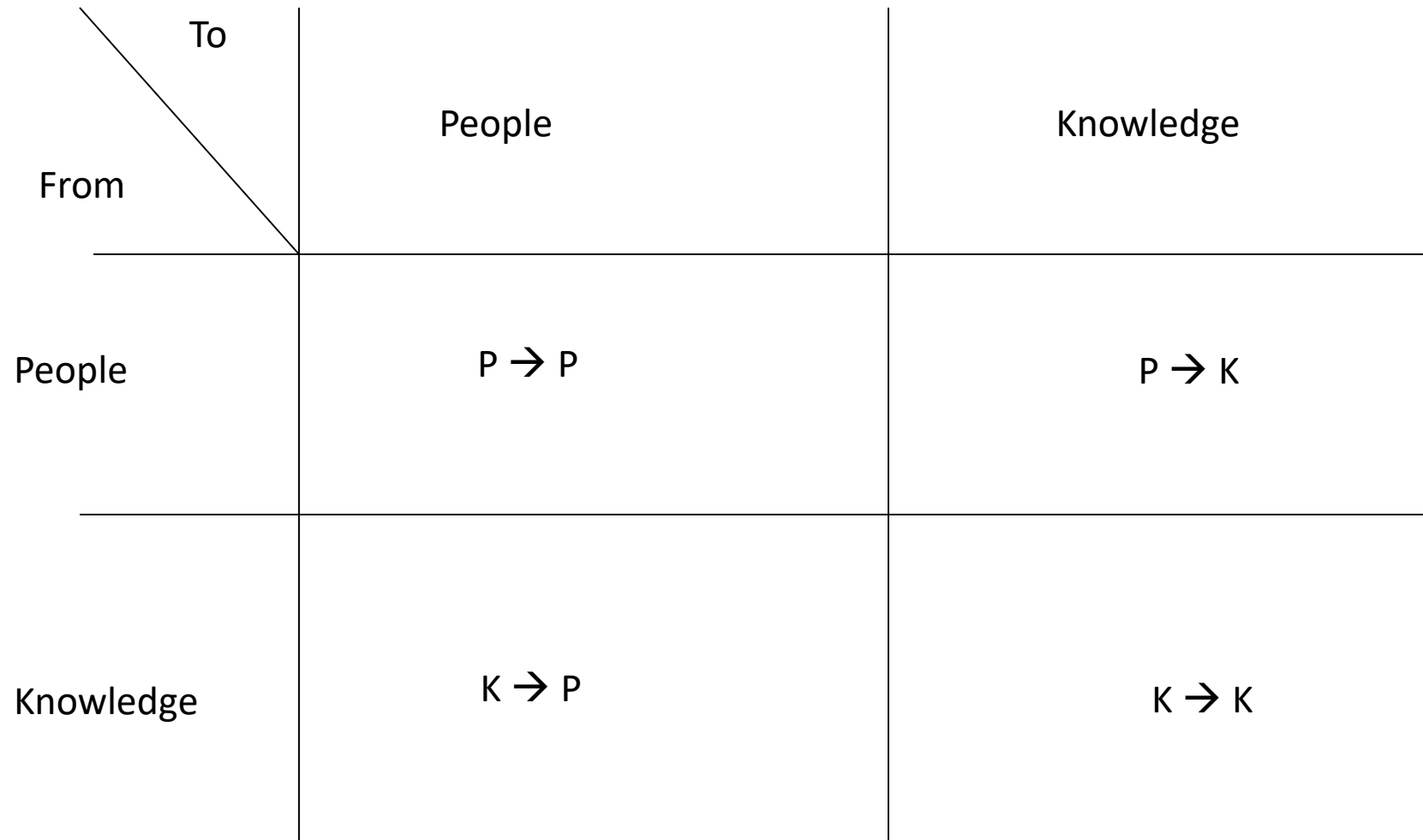
The problem is **important** :

- It was estimated in 2006 that more data will be produced in 2007 than has been generated during the entire existence of humankind (Panurgy 2006)
- An estimate by Merrill Lynch : more than 85% of all business information exists as unstructured text (Bloomberg and Atre 2003).

The problem is **hard** :

- From a study (Furnas et al., 1987) : different people use the same keywords for expressing the same concepts only 20 % of the time.

The landscape of applications



Can the web come alive?



“ AI researchers have long recognized that the more a system knows about a particular state of affairs, the longer it takes to retrieve the relevant information, and this presents a general problem when scaling up is concerned. Conversely, the more a human being knows about a situation or an individual, the easier it is to retrieve other relevant information.” - Dreyfus

NLP is important.

- Teach computers to communicate with people
- Help people communicate with each other
- Harness the potential of the web
- Propose and test new cognitive models

NLP is important.

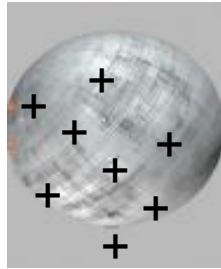
- Teach computers to communicate with people
 - database query interfaces
 - intelligent tutoring systems
 - question answering systems
 - speech recognition and spoken language understanding
- Help people communicate with each other
 - Machine Translation
 - Natural Language Generation
- Harness the potential of the web
 - Information Extraction
 - Search (related tasks : classify, filter, summarize)

NLP is interesting

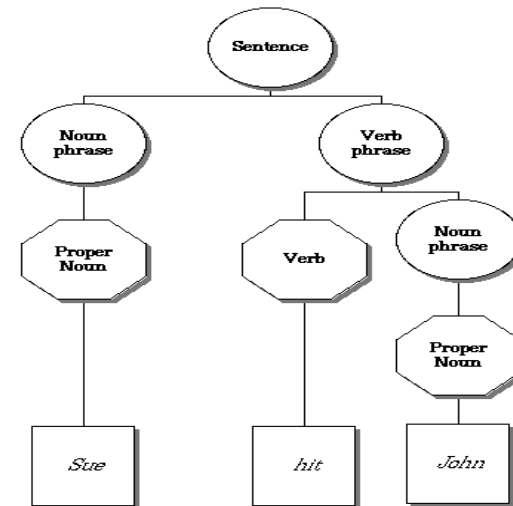
- Language is an exciting puzzle in its own right
- Interdisciplinary
- Challenging problems from a computational perspective

The Physics of Information

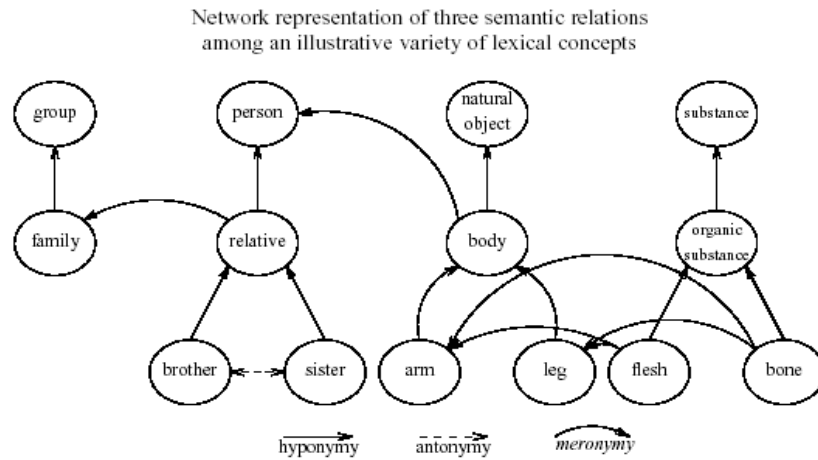
$$F = \frac{Q_1 Q_2}{4 \pi \epsilon_0 r^2}$$



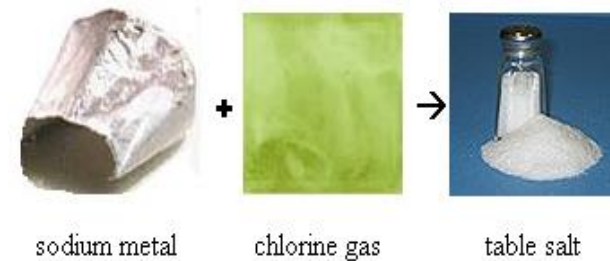
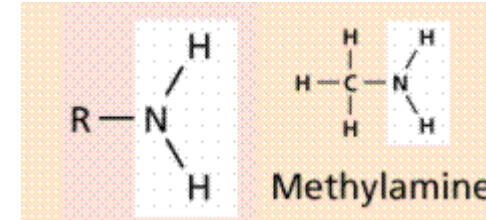
s --> np vp
np --> det n
np --> det adj n
vp --> v np



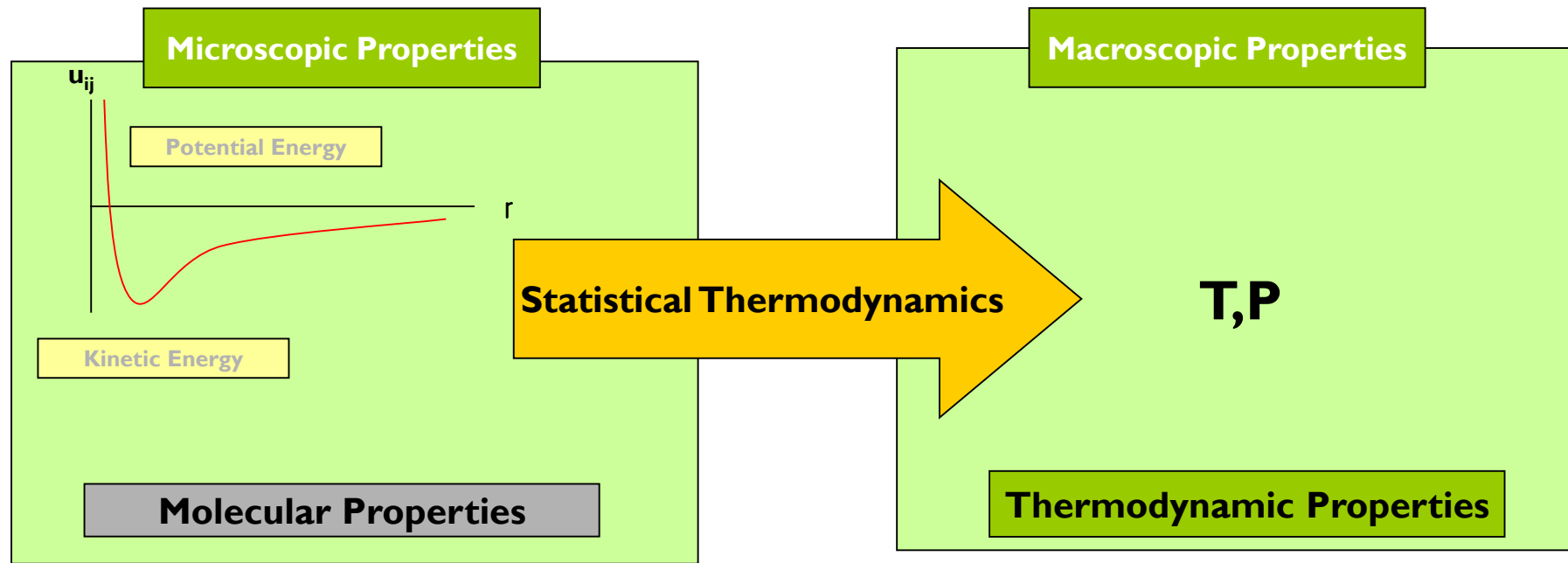
There's a bit of Chemistry as well



“cat licking the mirror”



More Physics: Ensembles and emergence



But most of the great sides of different eras, be it Clive Lloyd's West Indies or this Australian team, had two things in common: a consistent pair of openers and a reliable wicket-keeper. Though India don't have clear-cut choice in this department, it's easy to see the imbalanced selectorial policy.



When is a movie romantic?
The Problem of Nonconscious knowledge

And now... a bit of biology as well

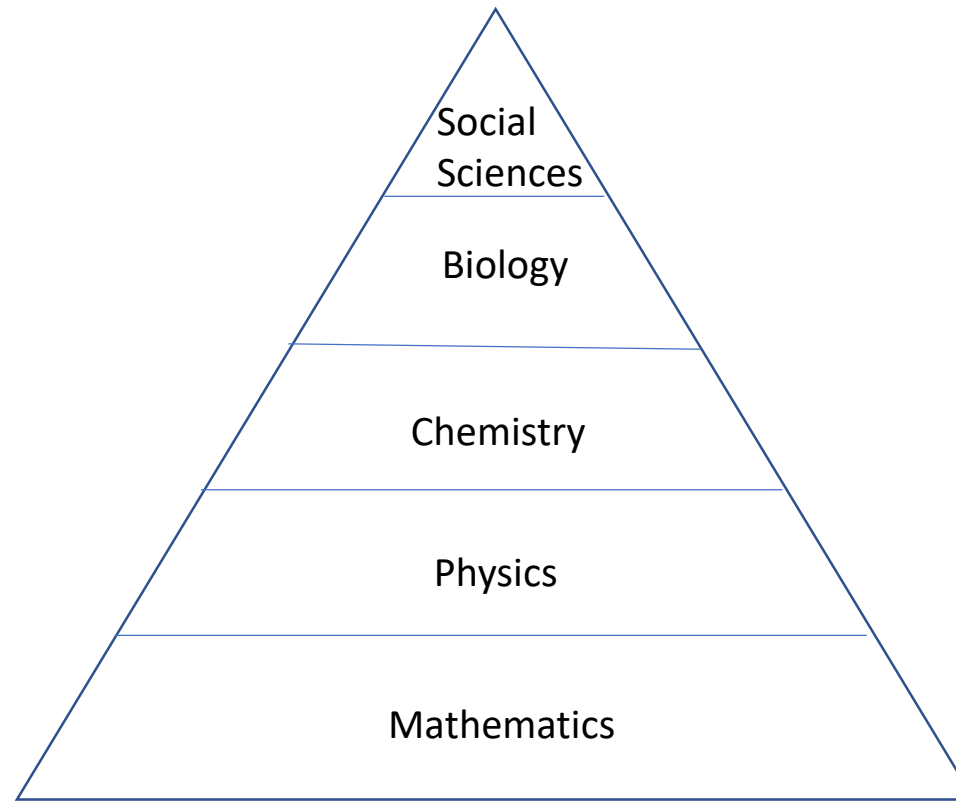


And yet, it is not all done...

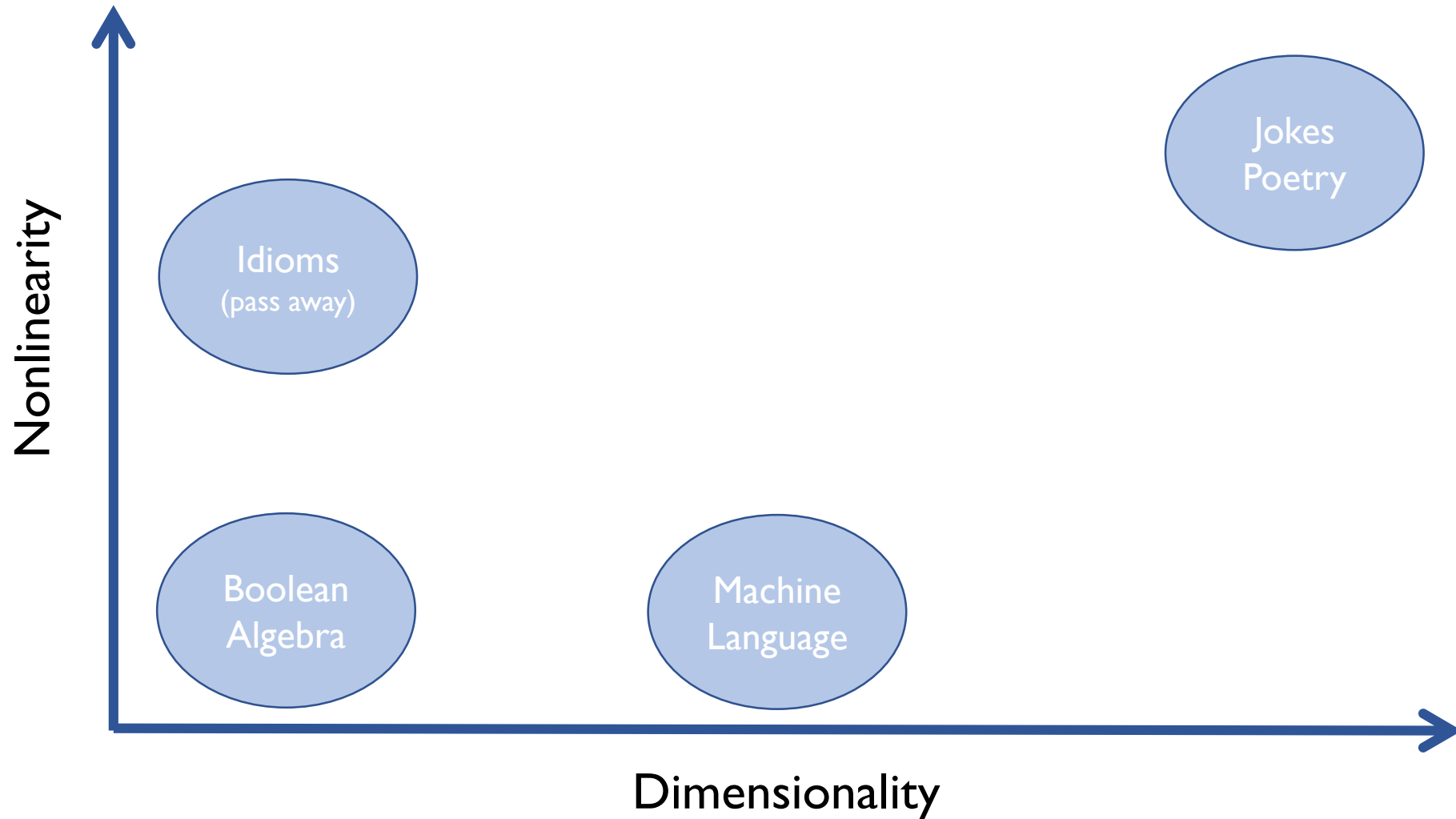
“If the brain were so simple we could understand it,
we would be so simple we couldn't.”

– Lyall Watson

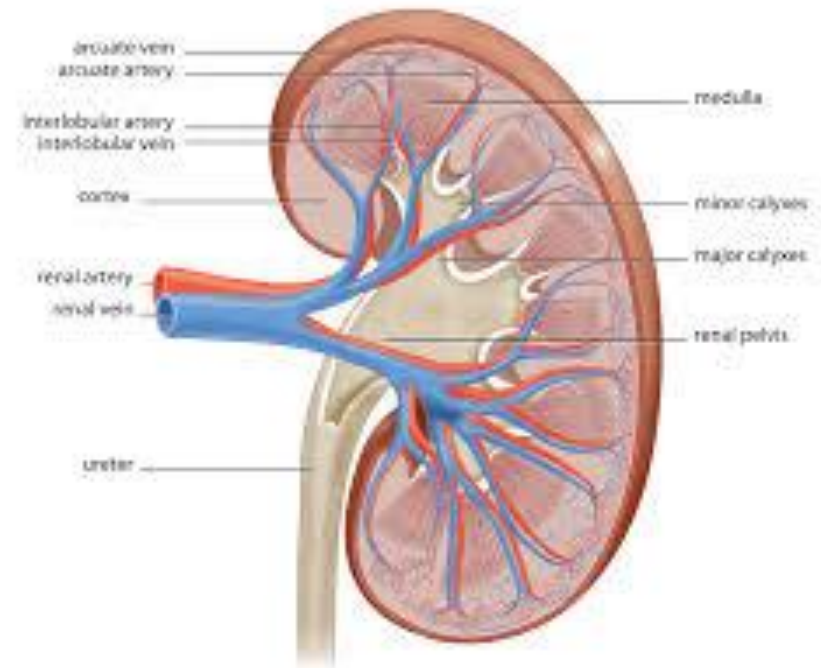
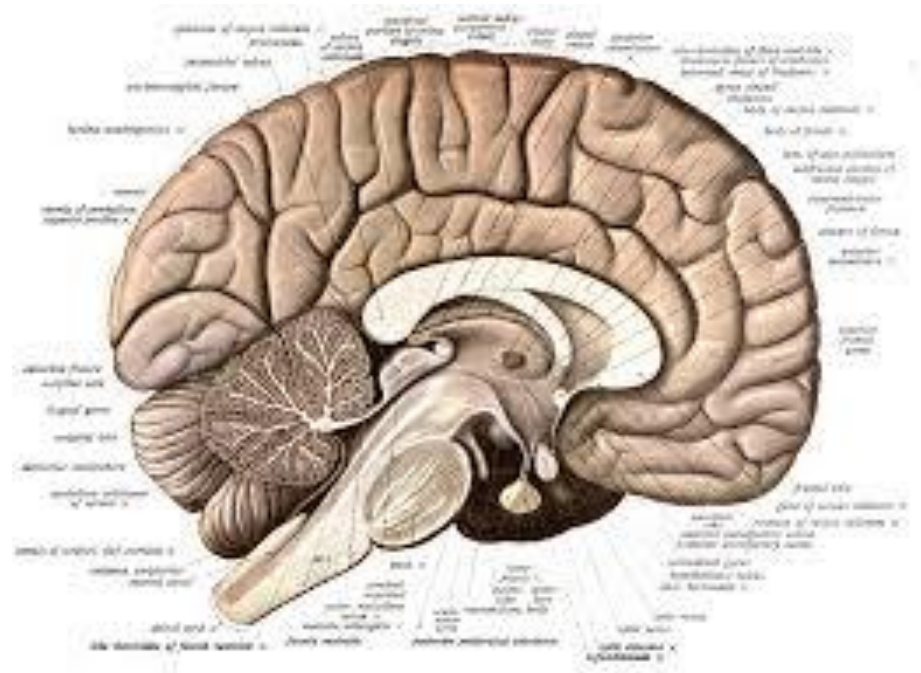
The harder it is, the less amenable it is to mathematical modelling !



The Dimensions of Complexity



Structure function correspondence



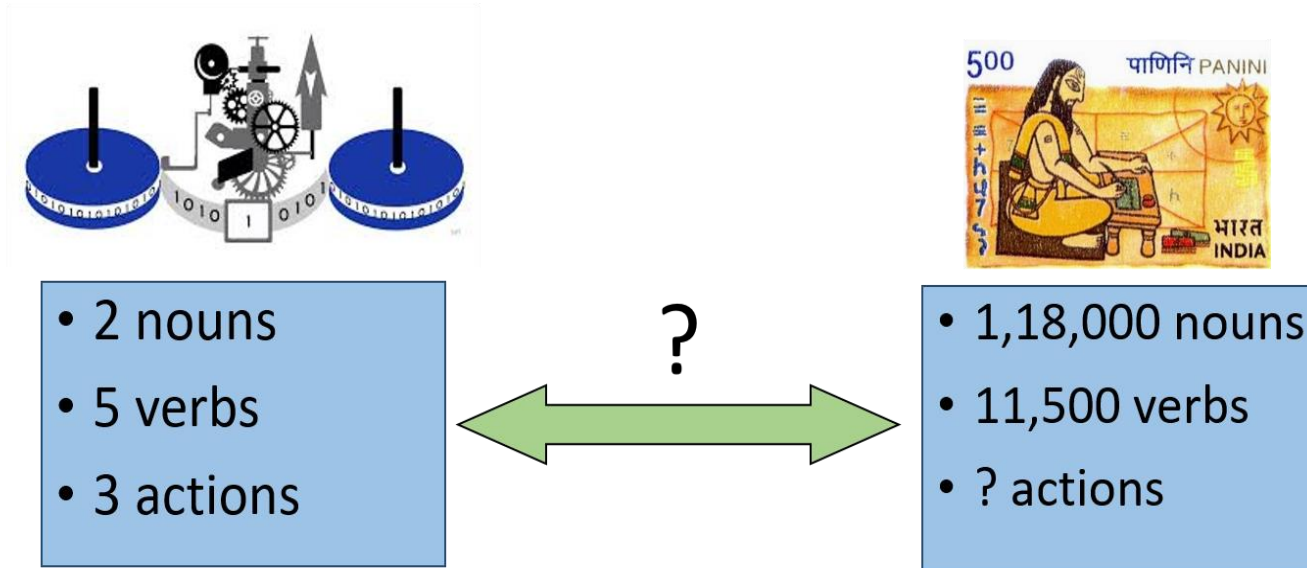
The language machines understand

3 Great Insights of Computer Science

- I. There are only **2 *objects***
that a computer has to deal with
in order to *represent* "anything"
- II. There are only **5 *actions***
that a computer has to perform
in order to *do* "anything"
- III. There are only **3 *ways of combining***
these actions (into more complex ones)
that are needed in order for a computer
to do "anything"

(Ack: William Rapaport)

The Challenge in NLP



(Ack: William Rapaport)

<http://design.vidanto.com/?p=225>

<http://satenderblogscollections.blogspot.in/2014/01/great-indians-of-ancient-india-part-2.html>





Reading Assignment

- Great Ideas in Computer Science:
<https://cse.buffalo.edu/~rapaport/111F04/directory.html>
- Chapter 1 of AI textbook by Russell and Norvig

Building blocks

- Letters
- Words
- Sentences
- Discourse

Building blocks : Analogues

• Letters		• Subatomic particles
• Words		• Atoms
• Sentences		• Molecules
• Discourse		• A collection of molecules

Properties of Words

- Morphology
- Phonetics
- Parts of Speech
- Semantics

Jargon: **Lexical** Semantics versus **Compositional** Semantics

Levels of Knowledge in NLP

- **Morphology:** how words are constructed; prefixes & suffixes
- **Syntax:** structural relationships between words
- **Semantics:** meanings of words, phrases, and expressions
- **Discourse:** relationships across different sentences or thoughts; contextual effects
- **Pragmatics:** the purpose of a statement; how we use language to communicate
- **World Knowledge:** facts about the world at large; common sense

Classic Problems in NLP

- Words:
 - Lexical Semantics
 - Word Sense Disambiguation
 - Morphology
 - Phonetics
 - Part Of Speech tagging
- Sentences:
 - Parsing
 - Compositional Semantics
- Discourse
 - Anaphora Resolution
- Pragmatics

Examples

- Morphology
 - friendly : friend + ly
- Syntax
 - I saw the man on the hill with the telescope.
- Semantics
 - I walked to the bank ...
 - of the river.
 - to get money.
 - Syntax does not always tell much about meaning
 - plastic cat food can cover

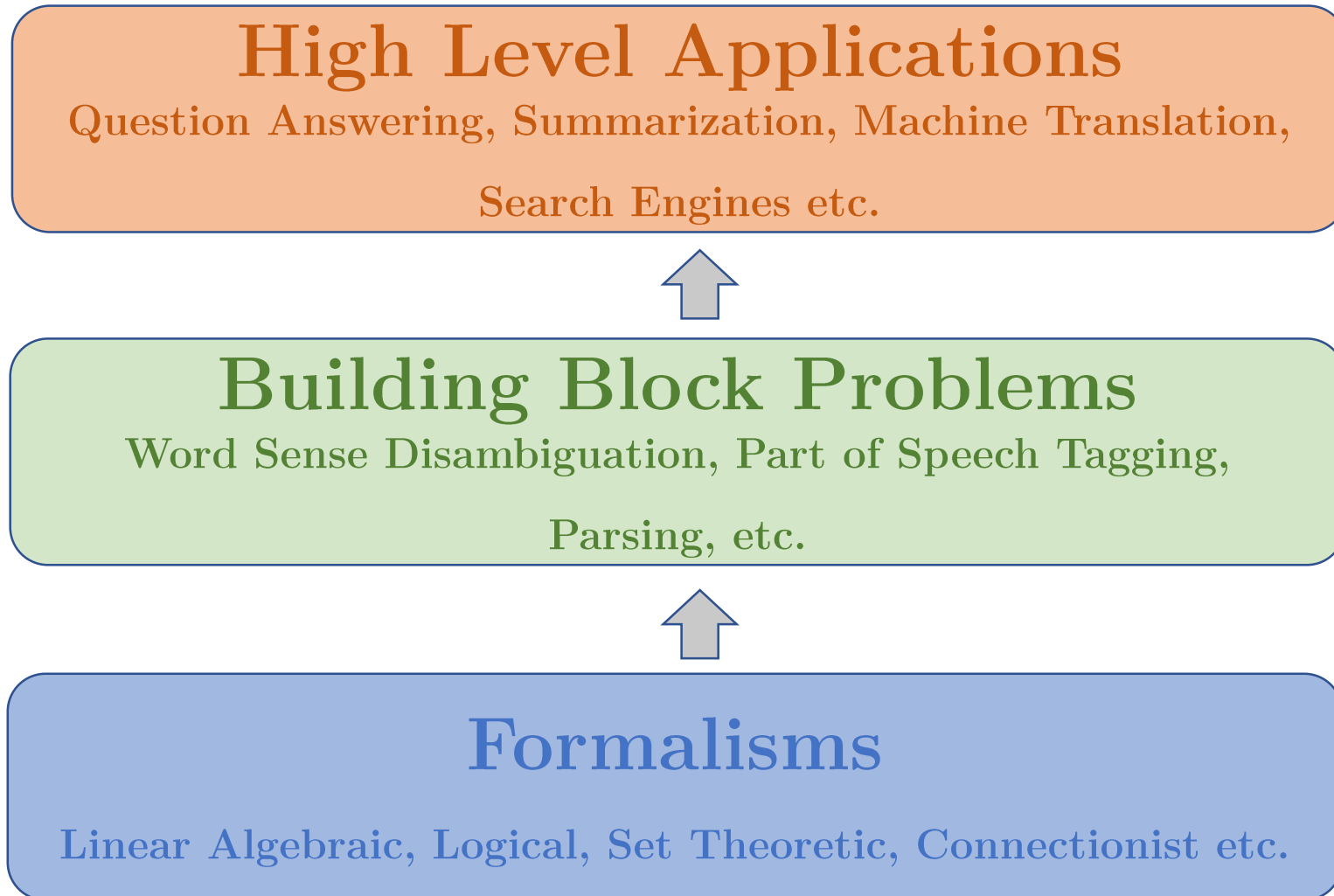
Examples

- Discourse
 - President John F. Kennedy was assassinated.
 - The president was shot yesterday.
 - Relatives said that John was a good father.
 - JFK was the youngest president in history.
 - His family will bury him tomorrow.
 - Friends of the Massachusetts native will hold a candlelight service in Mr. Kennedy's home town.

Examples

- Pragmatics
 - Can you tell me what time it is?
 - Could I please have the salt?
- World Knowledge
 - John went to the restaurant. He ordered a steak. He left a tip and went home.
 - John wanted to commit suicide. He got a rope.

A Conceptualization of the Big Picture



Problems matter !!!

Computer science should be called computing science,
for the same reason why surgery is not called knife
science.

E. Dijkstra

The science of it : Cognitive Science

- Combines Tools from
 - Psychology
 - Computer Science
 - Linguistics
 - Philosophy
 - Neurobiology

Also relevant : Sociology and Anthropology

The Engineering of it

- Building computational models for
 - handling specific tasks
 - demonstrating the feasibility of a cognitive model

NLP is no rocket science !!!

A linguistics professor was lecturing to his English class one day. "In English," he said, "A double negative forms a positive. In some languages, though, such as Russian, a double negative is still a negative. However, there is no language wherein a double positive can form a negative."

A voice from the back of the room piped up, "Yeah, right."

Ack: http://www.langston.com/Fun_People/1997/1997BIJ.html

English is strange

- Constructs
 - One drives on a parkway, parks in a driveway; plays at a recital, recites at a play
 - If a vegetarian eats vegetables, what does a humanitarian eat? If you wrote a letter, perhaps you bote your tongue?
- Word pronunciation: Bernard Shaw observed that *fish* could just as sensibly be spelled *ghoti* (*gh* as in *tough*, *o* as in *women*, *ti* as in *nation*)

The search for a small set of universal principles can prove elusive !!!

Three Approaches in the study of Language

- Linguistic
- Psycholinguistic
- Computational

Approach 1: The Linguistic Approach

- The goal : Characterizing the knowledge which underlies the ability to use language (Chomsky 1986)
- Look for Universal principles and Mathematical characterizations (rules)
 - Why certain combination of words form sentences, others do not
 - Why a sentence can have some meanings, but not others
- Static descriptions
 - Rules which define grammatical sentences
 - Description of idealized forms of products of a language producing system, i.e. the sentences
- Linguistics is NOT concerned with the use of language in the sense of how humans apply their knowledge of language in the production and comprehension of sentences.

Approach 2: The Psycholinguistic Approach

- The goal : the study of mental mechanisms that make it possible for people to use language (Garnham 1985)
- Question: how do humans process language in terms of real time operations, using finite resources ?
- Experimental procedures
 - The relative complexity of two sentences might be measured in terms of reaction times of subjects asked to evaluate the grammaticality of the sentences
- When used for testing hypothesis tied to linguistic theories, there could be error due to wrong interpretations or faulty design/methodology of the experiment
- While linguistics is concerned with language as structure/knowledge/product, psycholinguistics is concerned with language as process

Approach 3: The Computational Approach (NLP)

- Process-oriented : how can machines produce or understand language using finite computational resources
- Rules and Interrelationships between modules have to be stated explicitly for mechanical interpretation
- Two extremes :
 - Applied work of engineering NL interfaces to databases, expert systems and tutorial systems
 - Computational modelling of the human language capacity (Cognitive science)

Course Outline

- Introduction
- Structure of Words
 - Spellcheck
- Knowledge Light NLP
 - Information Retrieval Basics
- Lexical Semantics
 - Concept Mining from Text: Semantic Search
 - Word Sense Disambiguation
- Modelling Syntax
 - Classical and Probabilistic Parsing, Part of Speech Tagging
 - Probabilistic Language Modelling
- Machine Translation
- Additional Application Areas:
 - Information Extraction
 - Natural Language Generation
- Deep Learning for Natural Language Processing: A Primer

Evaluation

- End Sem (50)
- Assignment (20)
- Project (20)
- Rest (10)

Textbooks

Jurafsky and Martin, Speech and Language Processing
Manning and Schutze, Foundations of Statistical NLP