todo:

ULM-4

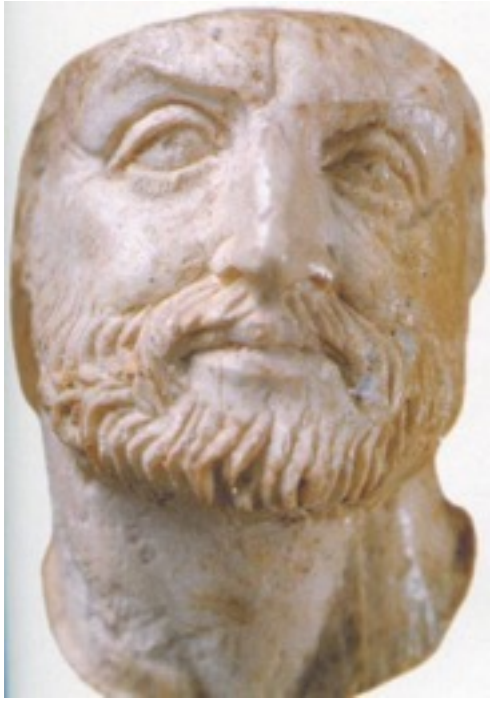
A quantum model of text understanding

Introduction

- Minh Ngoc Le: NLP pipeline
  - What is a NLP pipeline?
  - What are its drawbacks?
  - Suggest some solutions
- Filip Ilievski: Knowledge handling

Divide and Conquer

- Philip II of Macedon (382–336 BC):
  "*divide et impera*"
- Everyday life: Owe 1500€ with a salary of 1500€?
  - Divide the debt by 3 and return 500€ each month.
- Not surprisingly, we apply the same principle in NLP.



Dividing NLP

- Jacob's Ford is also known by the Latin name of Vadum Iacob
- *POS Tagging:* [Jacob's Ford]NNP [is]VBZ [also]RB [known]VBN [by]IN [the]DT [Latin]NNP [name]NN [of]IN [Vadum Iacob]NNP

Dividing NLP

- *Syntactic parsing:* (
  (Jacob's Ford) is (also known
  (by
  (the Latin
  (name of
  (Vadum Iacob))))))
- *Semantic role labeling:* [Jacob's Ford]A1 is [also]AM known [by the Latin name of Vadum Iacob]A2

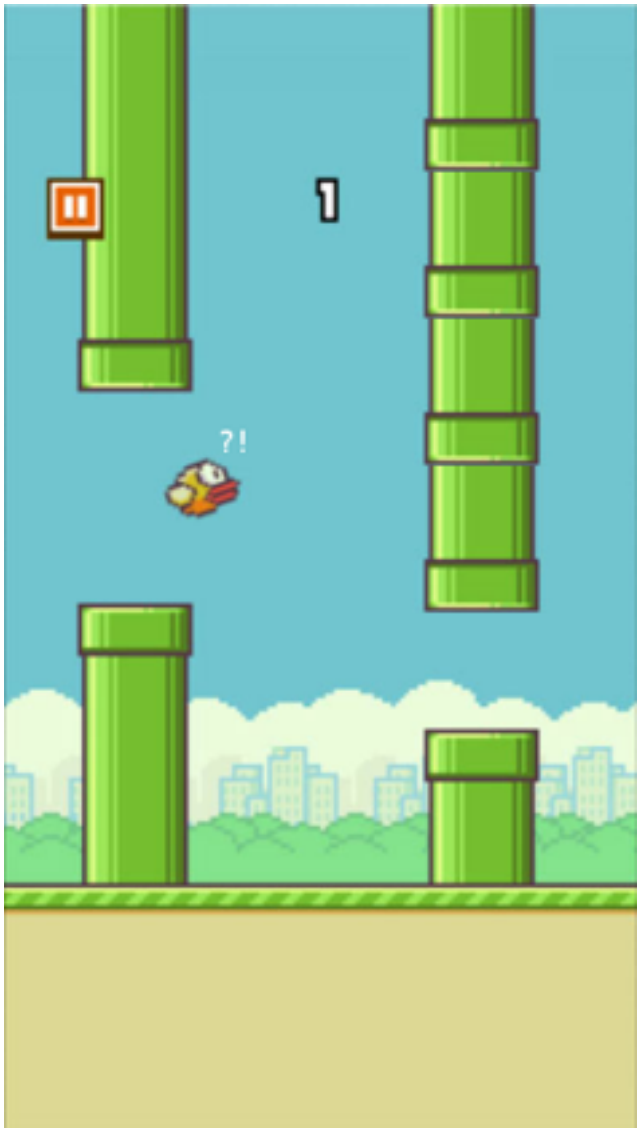
Dividing NLP

- Tokenization
- Sentence boundary detection
- Part-of-speech tagging
- Constituent parsing
- Semantic role labeling
- Selectional preference
- Dependency parsing
- Word sense disambiguation / induction
- Named-entity recognition / classification / linking
- Temporal expression recognition

- Co-reference resolution
- Sentiment analysis / opinion mining
- Discourse parsing
- Information extraction
- Topic modeling
- Summarizing
- Machine translation
- Natural language generation,
- Speech recognition / synthesis,
- Ontology population

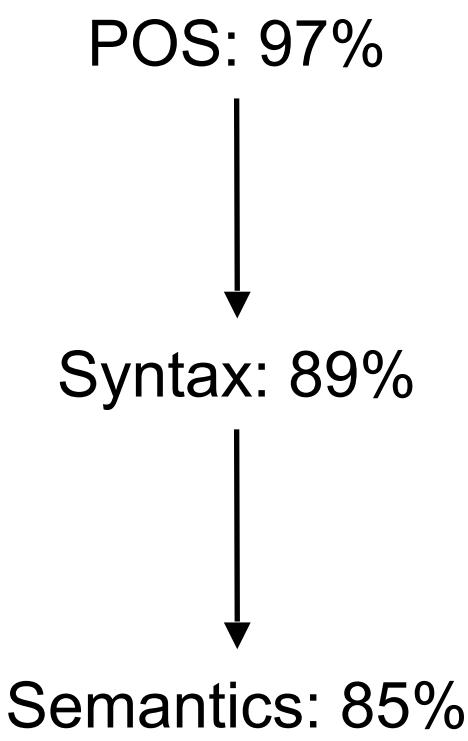
Assembling NLP: A pipeline

- Word segmentation
- Sentence boundary detection
- POS Tagging
- Syntax parsing
- Named-entity recognition
- Co-reference
- Sentiment analysis



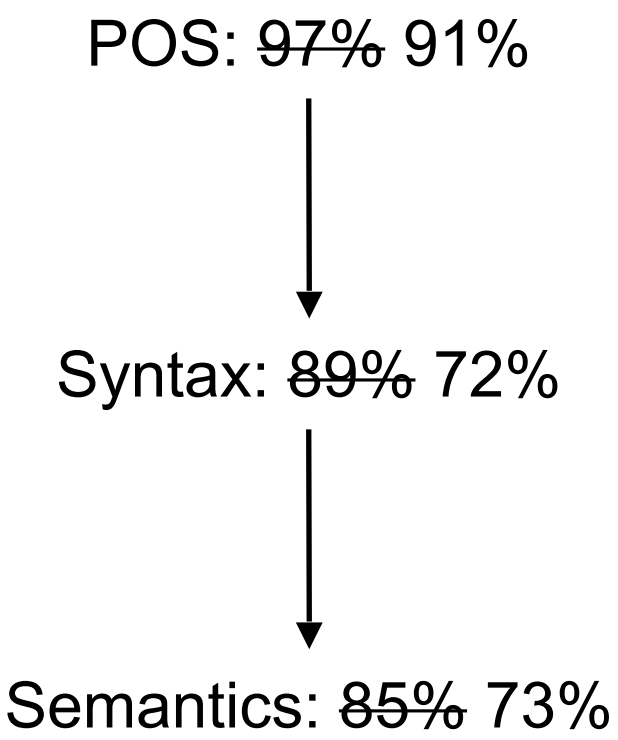
Conquering NLP

- Accuracies are measured on the same domain as training data
- So far so good...



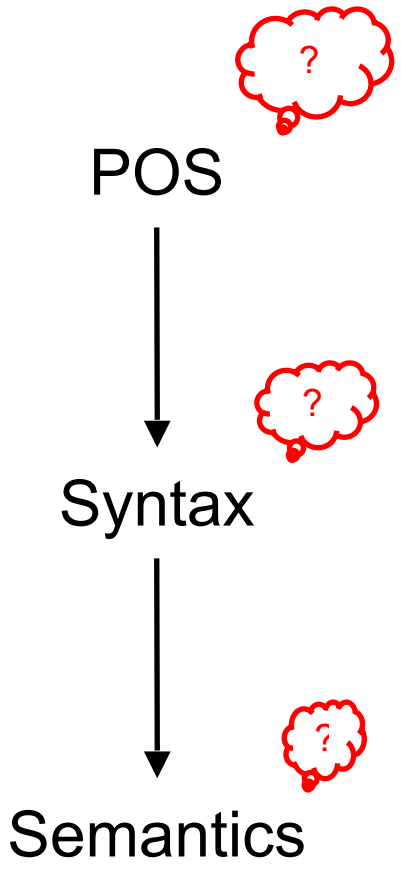
But wait...

- When modules are tested on a dataset different from what it was trained on, the accuracy is lower.
- Many applications might be better modeled by cross-domain experiments.



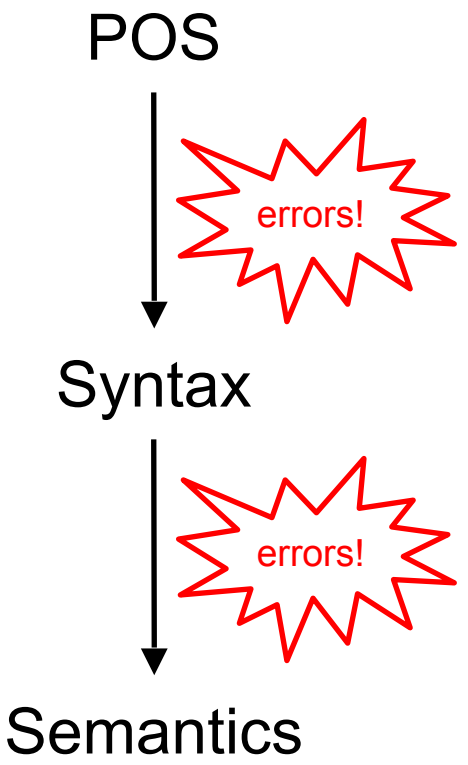
Lack of information

- What a sentence look like to a computer:
  Γουι ρεγξτ ρλετ φπ θιαπ χιη ρπονψφηω
- Modules based their decisions on very little information, especially first ones.



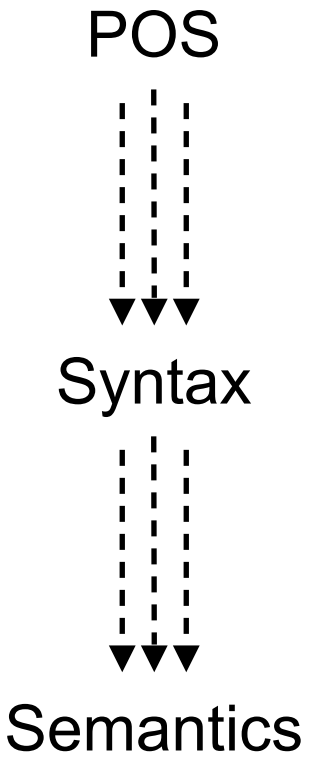
Error propagation

- Manning (2011): POS tagging accuracy at sentence level: 55-57%
- An incorrect POS tag may have global effect on the parsing of the whole sentence



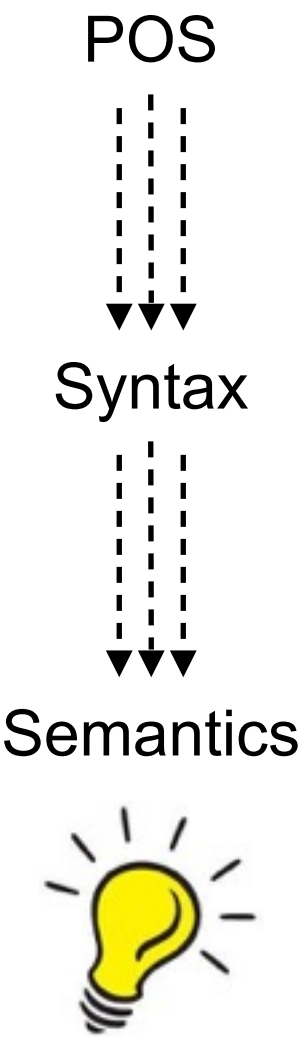
A solution?

- Postpone decisions: a module returns all alternatives with corresponding level of certainty rather than only one.



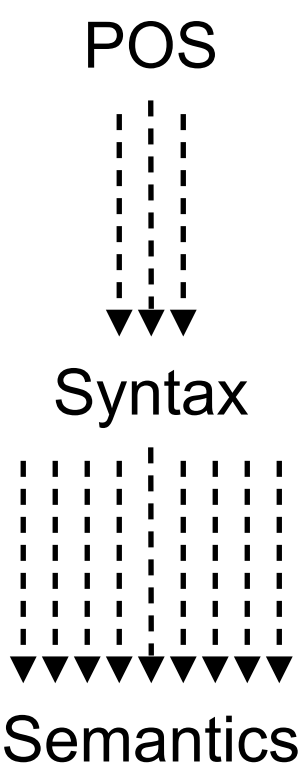
A solution?

- World knowledge: from text, image, audio, video
- Integration: information from previous processing step is combined with world knowledge to make the final decision.



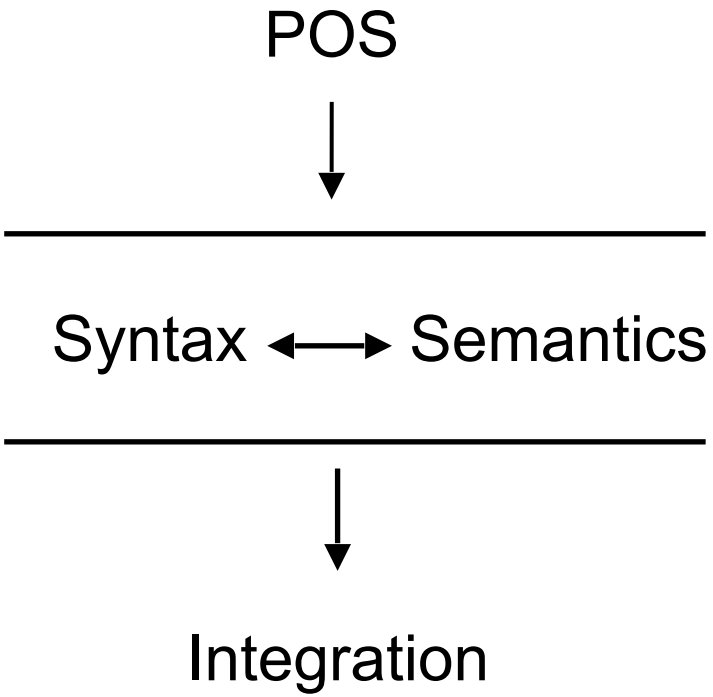
Combinatorial explosion

Jacob's Ford	5
is	25
also	125
known	625
by	3.125
the	15.625
Latin name	78.125
of	390.625
Vadum Iacob	1.953.125



Another solution?

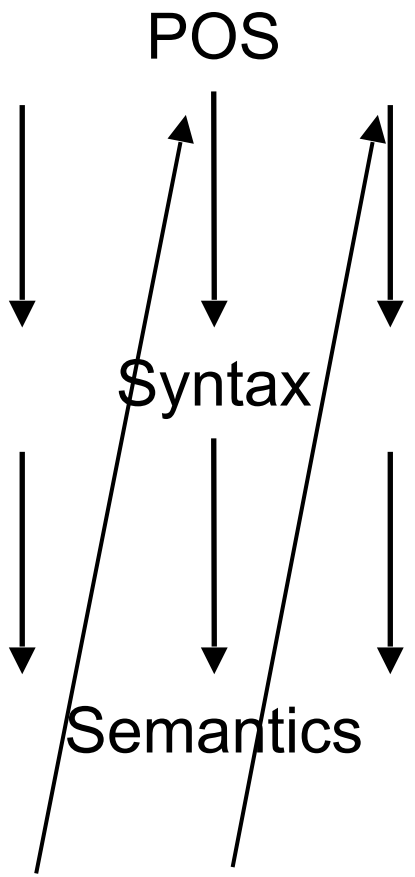
- Syntactic and semantic processes run in parallel
- Interact with each other
- Put forward by some neuro-linguistics researchers*



* For example: Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. Brain Research, 1146, 23–49. doi:10.1016/j.brainres.2006.12.063

Yet another solution?

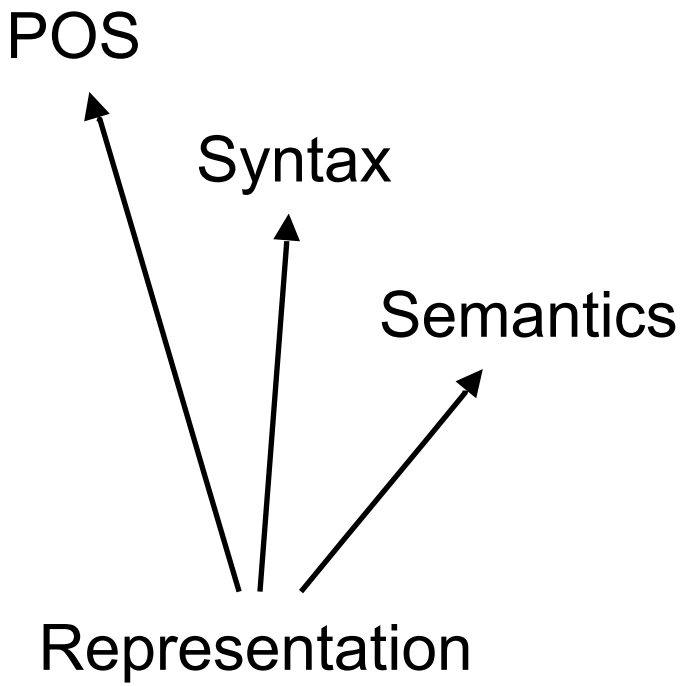
- Process each word from the first level to the last
- Come back to resolve conflict if needed
- Serial parsing may be the natural way of language comprehension in human*



* Meng, M., & Bader, M. (2000). Ungrammaticality detection and garden path strength: Evidence for serial parsing. *Language and Cognitive Processes*, *15*(6), 615-666.

A wild solution?

- SENNA system*:
  - Create a neural representation for a sentence
  - Use it to perform various tasks independently: POS tagging, chunking, NER, SRL



*Collobert, R., Weston, J., & Bottou, L. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 1, 1–34.

Thank you!

ULM-4

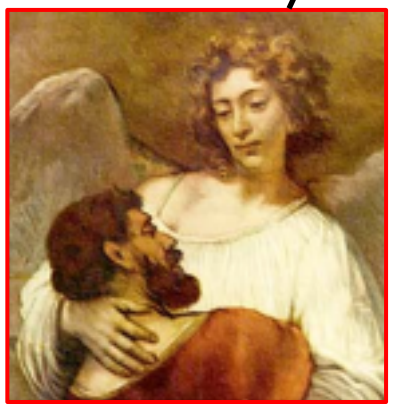
A quantum model of text understanding

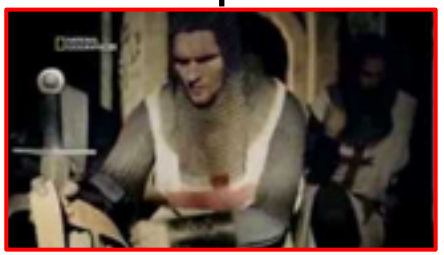
Outline

- Minh Ngoc Le: NLP pipeline
- Filip Ilievski: **Knowledge handling**
  - **Where** can it improve accuracy?
  - **What kind** of knowledge?
  - Knowledge **representation?**
  - Thoughts

Examples(1)

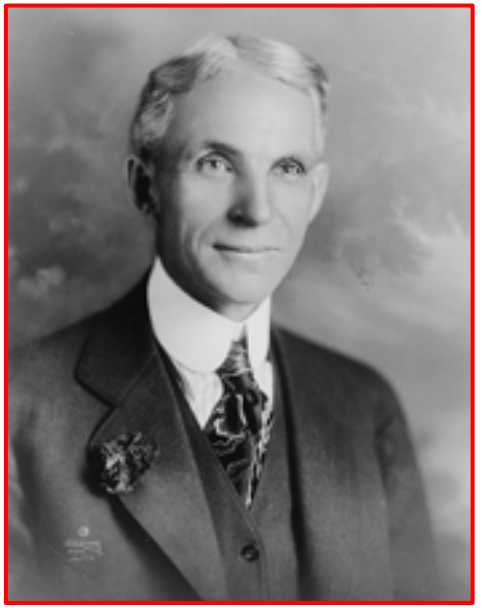
Jacob's Ford is also known by the Latin name of Vadum Iacob and in modern Hebrew as Ateret.





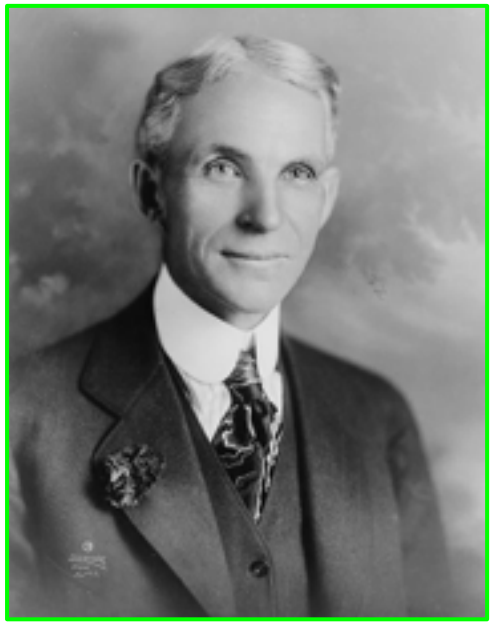
Examples(2)

The Henry Ford dance team, sponsored by Maia Weathersby, performed several routines at Spirit Day.



Examples(3)

President Woodrow Wilson asked Ford to run as a Democrat for the United States Senate from Michigan in 1918.



Examples(4)

In 2011, on the initiative of the British music and lifestyle magazine BLAG, singer and songwriter Estelle, rapper and producer David Banner and the musician Daley composed the new song "Benz", inspired by Joplin's "Mercedes Benz"

Spotting the **weaknesses**

- Modules lack context
  - *Example:*
    A module that analyses a sentence do not take into account the other sentences
- Modules are too restrictive, too confident
  - First modules have least information!
- Interpretation is biased by and dependent on the knowledge available
  - overfitting

What about **solutions**? (1)

- Modules to offer **multiple** possible interpretations and assign **probabilities**
- **Communication**
  - Module-to-module
  - Module-to-source(s) of knowledge
- Underlying information should be
  - **Rich**
  - **Global**
  - **Various** and **contradicting**

What about **solutions**? (2)

- Interpretations should consider:
  - **Availability** and **quality** of knowledge
  - **Writer** perspective
  - **Reader** perspective
- **Improved contextualisation**
  - Verbal context
  - Social context

Knowledge representation requirements

- Context-aware
- Support rich and flexible interpretation
- Allow knowledge quality and quantity to improve over time
  - with the increase of data
  - with user feedback
  - with improvement of NLP results
- Accessible during the decision process
  - to ensure modules have a global overview

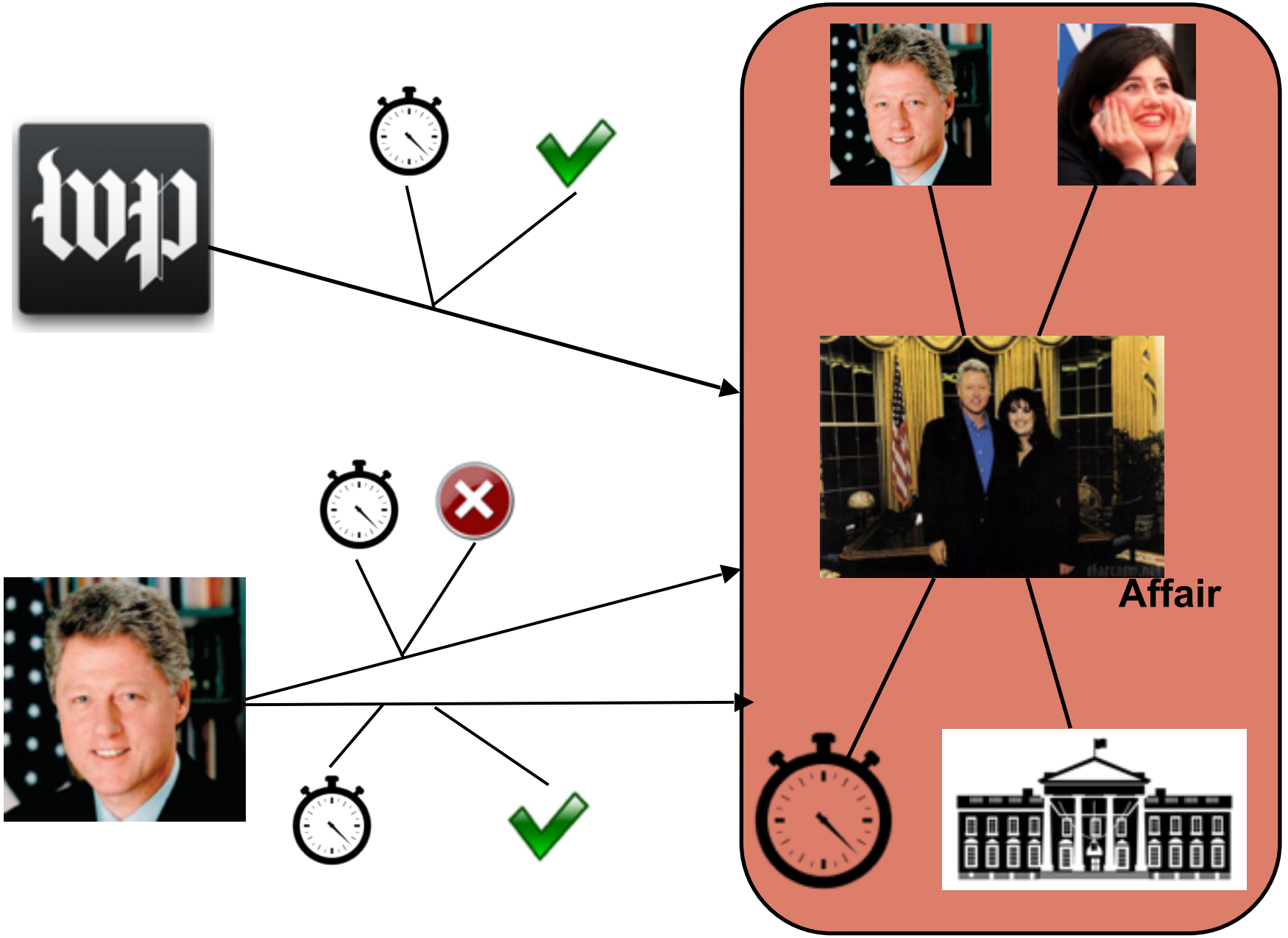
Possible knowledge representation technology: RDF

- Flexibility and richness of structure
- Data **integration** over time
- Usage of **external LOD resources** (WordNet, DBPedia)
- **Reusability**
- (Ontology-based) **Reasoning**
  - spot redundancies/wrong information
  - infer new information

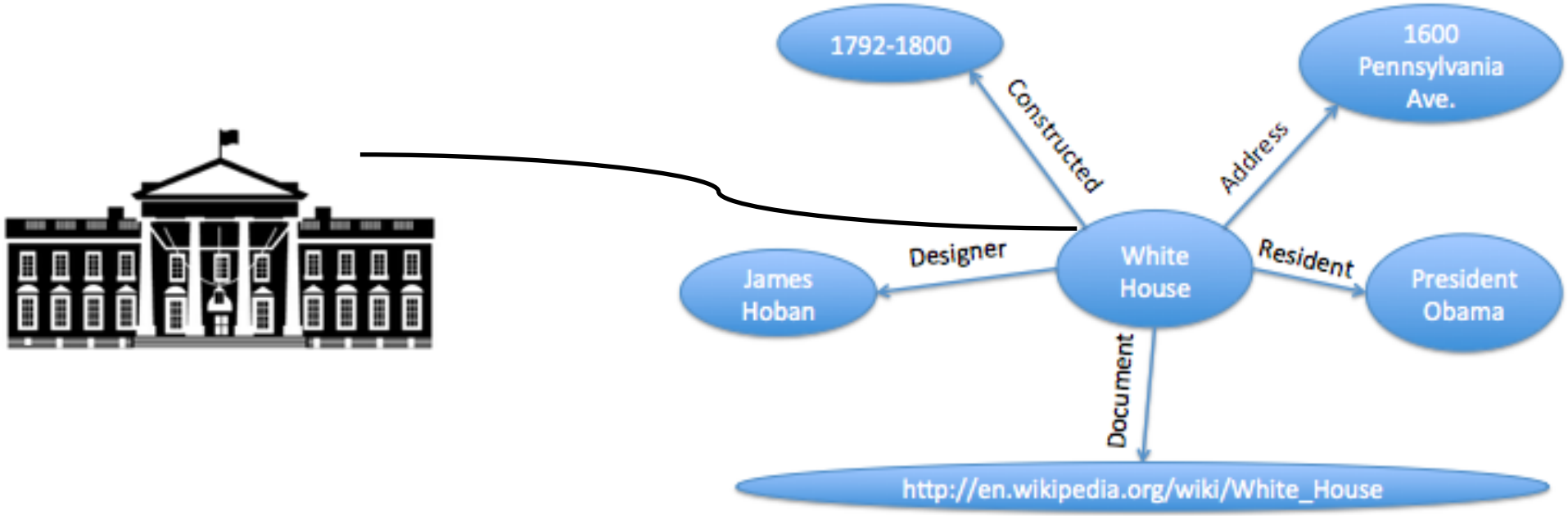
Flexibility and richness of structure(1)

- ***Washington Post*** *claims Bill Clinton and Monica Lewinsky had an affair (January 21, 1998)*
- ***Bill Clinton*** *denies Monica Lewinski affair (January 26, 1998)*
- *President* ***Clinton*** *admits to have an affair with Monica Lewinski (August 17, 1998)*

Flexibility and richness of structure(2)

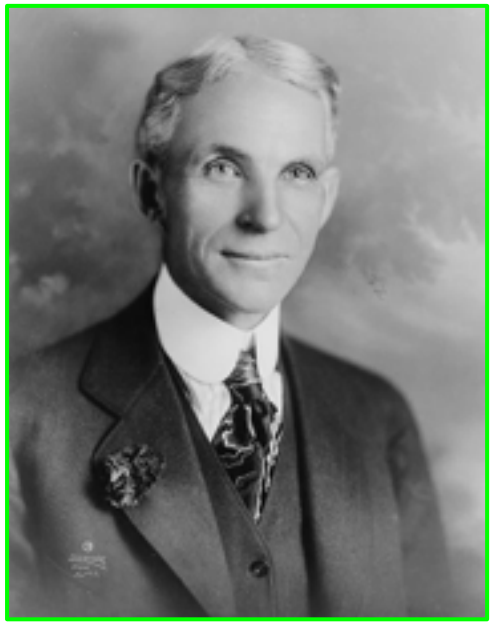


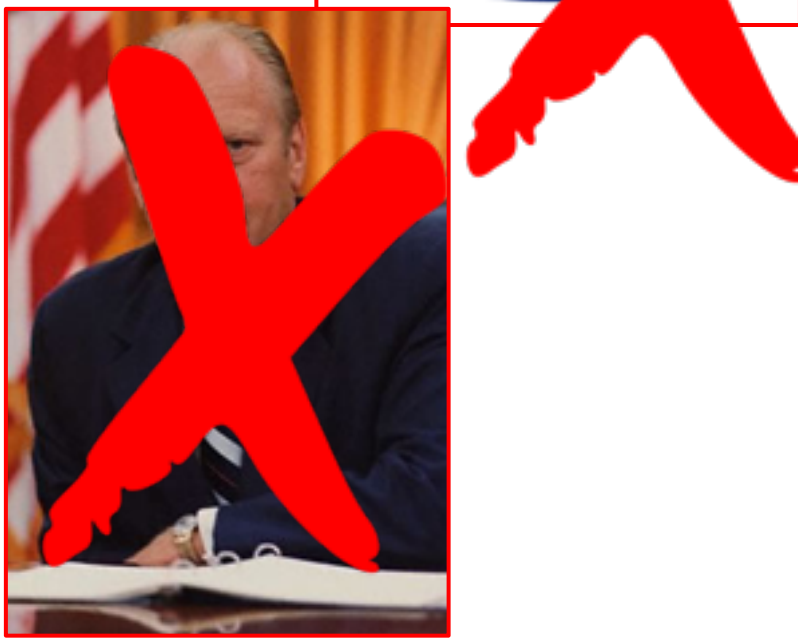
Linked open data cloud



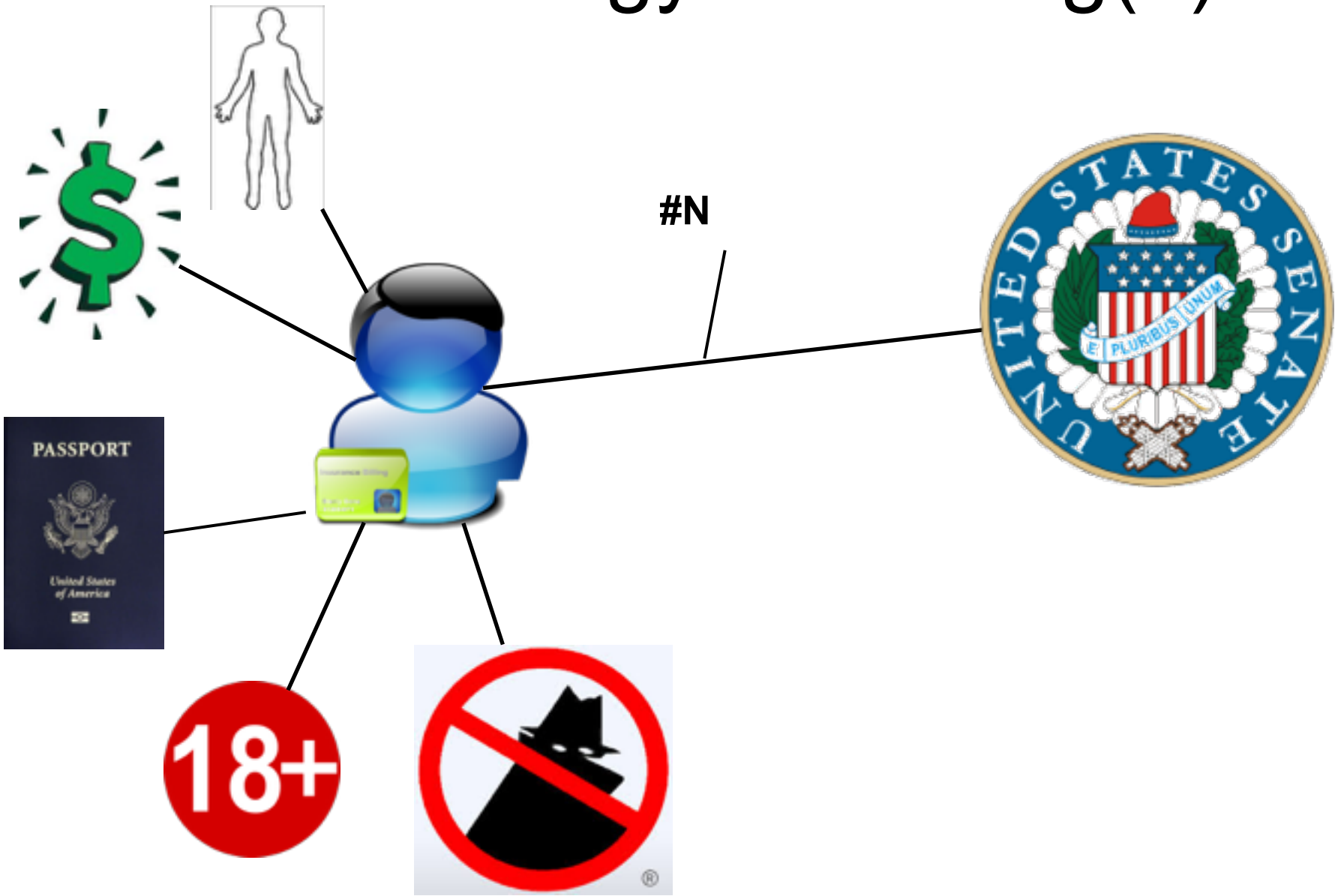
Ontology reasoning(1)

President Woodrow Wilson asked Ford to run as a Democrat for the United States Senate from Michigan in 1918.





Ontology reasoning(2)



Inferring new information(1)

- ***So... Was H. Ford elected as a Senate representative?***
  - *No explicit information*
  - *Maybe we can “investigate”*

Inferring new information(2)

Scenario: We know already that:

- *A Republican was elected as Michigan Senate representative in 1918 (**Event information**)*
- *Remember Ford ran as democrat*
- *Only one Senate representative per state elected (**Restriction**)*
- *Maximum one elections per year (**Restriction**)*

=> *Answer is **NO***

Challenges welcome

- Data access and storage
  - should be efficient and reliable
  - feasibility
- Data quality
  - redundant and wrong data
  - incomplete data
  - unreliable external sources
- Ontology structure
- Representation of probabilities/confidence factors
- Backtracking mechanism

 : NLP meets Semantic Web

NLP strives to enable computers to make sense of human language.

By their very nature, NLP technologies can extract a wide variety of information, and Semantic Web technologies are able to store such varied and changing data.

: Share understanding with our machines

Caution: ***Muddy Waters***

 : Share understanding with our machines

"*But I can design a good algorithm to work around and don't bother whether my machine understands it!*"

- Still, imagine the excitement and prospect of a machine actually understanding what we do!
  - Explain to them the data semantics and structure
- Hybrid of **machine learning** and **semantic web**


stay tuned !
future epic content
coming soon

Appendix: Garden path sentence

- The horse raced past the barn fell
- First reading:
  - [The horse raced past the barn] fell(!)
- Revised reading:
  - The horse [(that) raced past the barn] fell