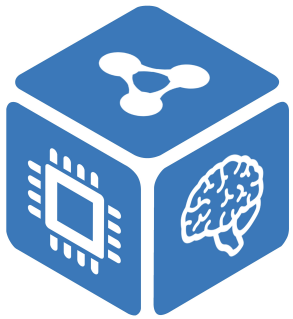
Statistical NLP Introduction

Winter Term 2018–2019

Prof. Dr. Axel Ngonga



Data Science Group

October 17, 2018

Overview

Section 1

Organization

Organization

Lecturer: Prof. Dr. Axel Ngonga

- Studied Computer Science & Physics
- PhD (CS) completed in 2009 at Leipzig University
- Habilitation (CS) thesis completed in 2016
- Full professor of Data Science since 2017
- Research Interests
  - Semantic Web
  - Machine Learning
  - Knowledge Extraction
  - Natural Language Generation
  - Cognitive Computing
- `https://dice-research.org`

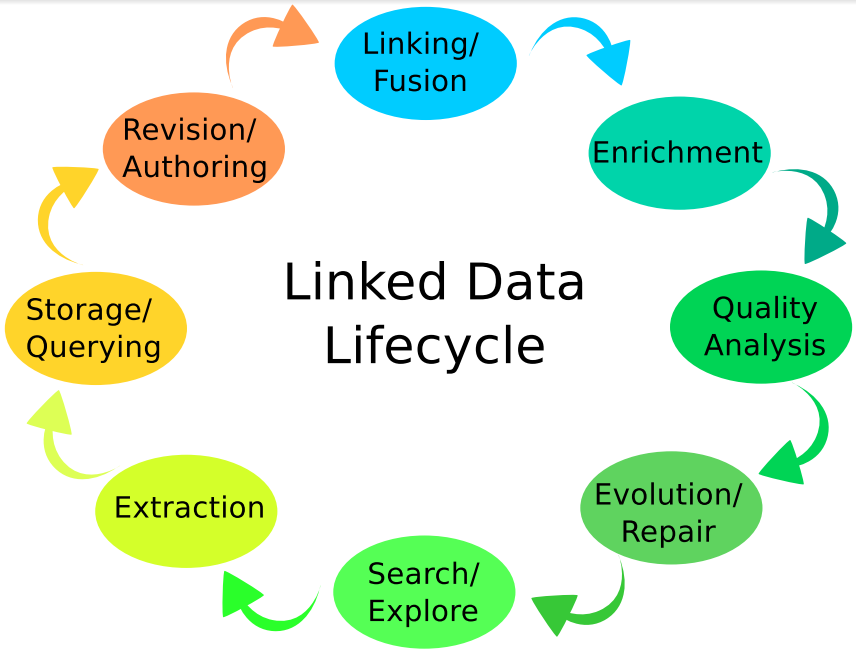
Organization

Assistant: Michael Röder

- Studied CS at the Hochschule Darmstadt
- Full-time Researcher at AKSW since 2015
- Technical Lead of European project (HOBBIT)
- Research interest: Benchmarking, Crawling, Fact Checking, Topic Modeling
- `https://dice.cs.uni-paderborn.de/team/profiles/roeder/`

Goals

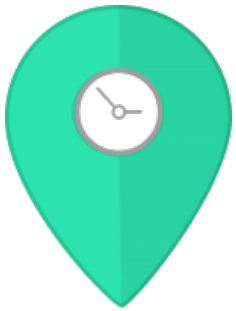
- Develop **scalable approaches**
- for the **extraction**, **aggregation** and **presentation** of **knowledge**
- from **large** amounts of **data**.



http://twitter.com/DiceResearch

Organization

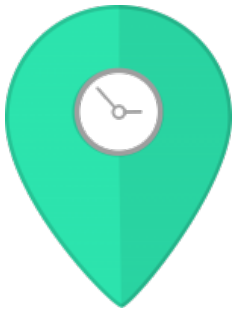
Time, Location, Format



- **Time and location**:
  - **Lectures**: Wednesdays, 16:15-17:45 in lecture hall O2
  - **Seminar**: Tuesdays, 16:15-17:45 in lecture hall O2/O4-267 (see Paul)
  - **Mini-Project**: Final submission on 30.01.2019

Organization

Time, Location, Format



- **Time and location**:
  - **Lectures**: Wednesdays, 16:15-17:45 in lecture hall O2
  - **Seminar**: Tuesdays, 16:15-17:45 in lecture hall O2/O4-267 (see Paul)
  - **Mini-Project**: Final submission on 30.01.2019
- **Format** $= 2 + 1 + 2$
  - 2 hours of lecture
  - 1 hour of seminar (2 hours bi-weekly)
  - 2 hours mini-project
- Slides, exercises and communication via PAUL

- Sheets bi-weekly from October. 24th, 2017 onwards
- 7 days for completion of each series (avg. 20 points/series)
- Exercises to be submitted at 09:00am latest (time stamp of our server)
- First seminar on October 23rd (NBGrader + GERBIL)

Series	Exercise	Submission	Solution
1	2018/10/24	2018/31/10	2018/11/06
2	2018/11/07	2018/11/14	2018/11/20
3	2018/11/21	2018/11/28	2018/12/04
4	2018/12/05	2018/12/12	2018/12/18
5	2019/12/19	2019/01/09	2019/01/15
6	2019/01/16	2019/01/23	2019/01/29

- **Requirements**
  - Course content
  - Working knowledge of Java

- **Requirements**
  - Course content
  - Working knowledge of Java
- **Submission via NBGrader** (not hand-written)

jupyter Exercise-CharactersOrder (autosaved) Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Not Trusted Java

Run Markdown Validate

SampleOutput: 3

Explanation: Result for `b, c, d`

In [ ]:

```
1  public int getMaxConsecutiveLength(char[] characters) {
2
3      int maxLength = 0;
4
5      // YOUR CODE HERE
6
7      return maxLength;
8  }
```

EVALUATION AREA

You can use the following cell to test your solution. It is one test case, there may exist more.

- **Requirements**
  1. $>$ 60% of exercises submitted
  2. $>$ 50% of total number of points
  3. Completion of mini-project
     - GERBIL results above baseline
     - Documented code on GitHub/GitLab
- **Format**
  - Written exam
  - Content = Lecture + Exercises
  - Duration: 90 minutes
  - Date (prospective): February 5th (TBC)

Goal

Build a corpus-driven fact-checking engine, which returns a confidence value between -1 (fact is false) and +1 (fact is true) given a fact from DBpedia

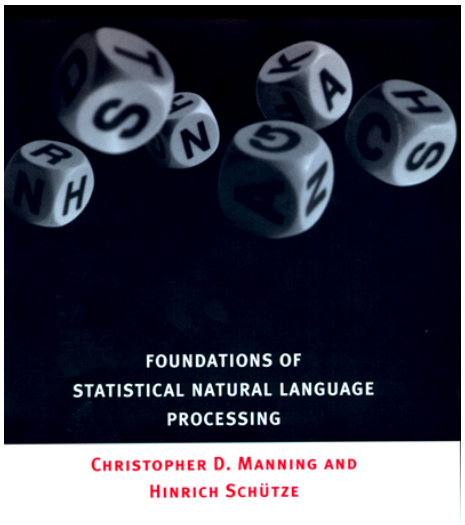
Goal

Build a corpus-driven fact-checking engine, which returns a confidence value between -1 (fact is false) and +1 (fact is true) given a fact from DBpedia

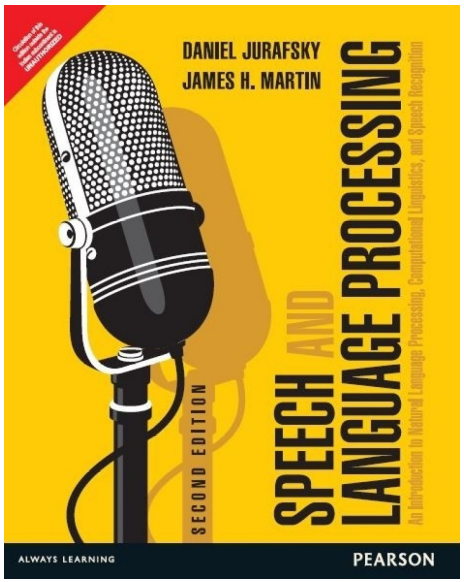
- **Group size**: max. 3 persons
- **Code & documentation**: GitHub/GitLab
- **Suggested steps**
  1. Corpus creation (2 weeks)
  2. Corpus normalization (2 weeks)
  3. Corpus analysis (2 weeks)
  4. Fact Checking and Benchmarking (rest)
  5. Final submission: 30.01.2019
  6. Group registration: `https://goo.gl/forms/hQeVYZD0RE6jgG932`

Organization

Literature



- Christopher Manning, Hinrich Schütze: **Foundations of Statistical Natural Language Processing**. MIT Press, 2nd edition



- Daniel Jurafsky, James H. Martin: **Speech and Language Processing**. Prentice Hall Series on Artificial Intelligence, 3rd edition

Section 2

Take Away

Take Away

Summary

- Origins of linguistic science
- The data problem
- Goals of statistical natural language processing (SNLP)
- Applications of SNLP
- Basic assumptions of SNLP
- A prototypical NLP pipeline

Section 3

Origins

Linguistic Science

The aim of a linguistic science is to be able to **characterize** and **explain** the multitude of linguistic observations circling around us, e.g., in conversations, writing, and other media.

Linguistic Science

The aim of a linguistic science is to be able to **characterize** and **explain** the multitude of linguistic observations circling around us, e.g., in conversations, writing, and other media.

- Three main concerns
  1. **Cognitive**: How do humans acquire, produce and understand language?

Linguistic Science

The aim of a linguistic science is to be able to **characterize** and **explain** the multitude of linguistic observations circling around us, e.g., in conversations, writing, and other media.

- Three main concerns
  1. **Cognitive**: How do humans acquire, produce and understand language?
  2. **Model**: How are linguistic utterances related to the real world?

Linguistic Science

The aim of a linguistic science is to be able to **characterize** and **explain** the multitude of linguistic observations circling around us, e.g., in conversations, writing, and other media.

- Three main concerns
  1. **Cognitive**: How do humans acquire, produce and understand language?
  2. **Model**: How are linguistic utterances related to the real world?
  3. **Structural**: By which means do languages communicate semantics?

Linguistic Science

The aim of a linguistic science is to be able to **characterize** and **explain** the multitude of linguistic observations circling around us, e.g., in conversations, writing, and other media.

- Three main concerns
  1. **Cognitive**: How do humans acquire, produce and understand language?
  2. **Model**: How are linguistic utterances related to the real world?
  3. **Structural**: By which means do languages communicate semantics?
- Will be mainly concerned with points 2 and 3

Origins

Beginnings

- 2000+ year old science
- Even older problem
- Need to communicate
  - Oral traditions
  - Invention of writing to store communication
  - Encoding of language grammars
  - Increasingly formal approach

Origins

Beginnings

- 2000+ year old science
- Even older problem
- Need to communicate
  - Oral traditions
  - Invention of writing to store communication
  - Encoding of language grammars
  - Increasingly formal approach

Problem

"All grammars leaks" (Sapir, 1921)

- Complete grammar cannot always be formalized
- Grammars evolve
- Need to make things loose $\Rightarrow$ **Statistics**

Goal (SNLP)

Find common patterns that occur in language use and exploit them to process natural language automatically Contras?

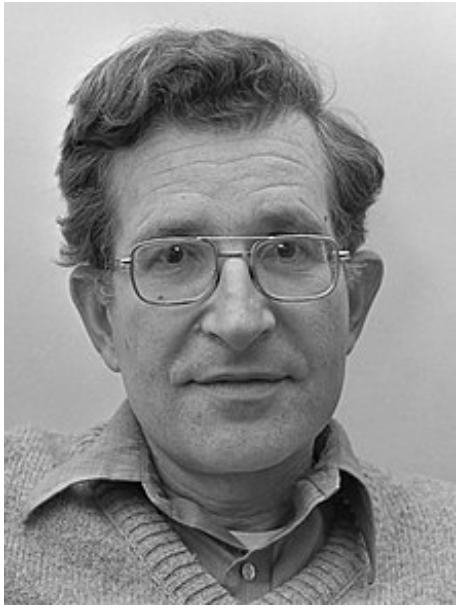
Goal (SNLP)

Find common patterns that occur in language use and exploit them to process natural language automatically Contras?

Contra: The Rationalists (1960 – 1985)

"One's ability to produce and recognize grammatical utterances is not based on notions of statistical approximation and the like" (Chomsky, 1957)

- Not enough stimuli to learn language
- Innate language ability
- Statistical models cannot capture rare phenomena



Avram Noam Chomsky
(1928 – today)

Pro: The Empiricists

"Statistical considerations are essential to an understanding of the operation and development of languages" (Lyons, 1968)

- Processing of NL automatically demands finding set of regularities
- Clear that they don't always exist: "All grammars leak" (Sapir, 1921)
- Ability to teach language systematically points to automatic processing
- Complex statistical models **can** predict and capture rare phenomena

Edward Sapir
(1884 – 1939)

- Follow empiricist approach
- No access to complete language
  $\Rightarrow$ We use corpora.
- Analysis of phenomena using associations and correlations
  $\Rightarrow$ "You shall know a word by the company it keeps" (Firth, 1957)
- Complex statistical models **can** predict and capture rare phenomena

John Rupert Firth
(1890 – 1960)

Section 4

The Data Problem

The Data Problem

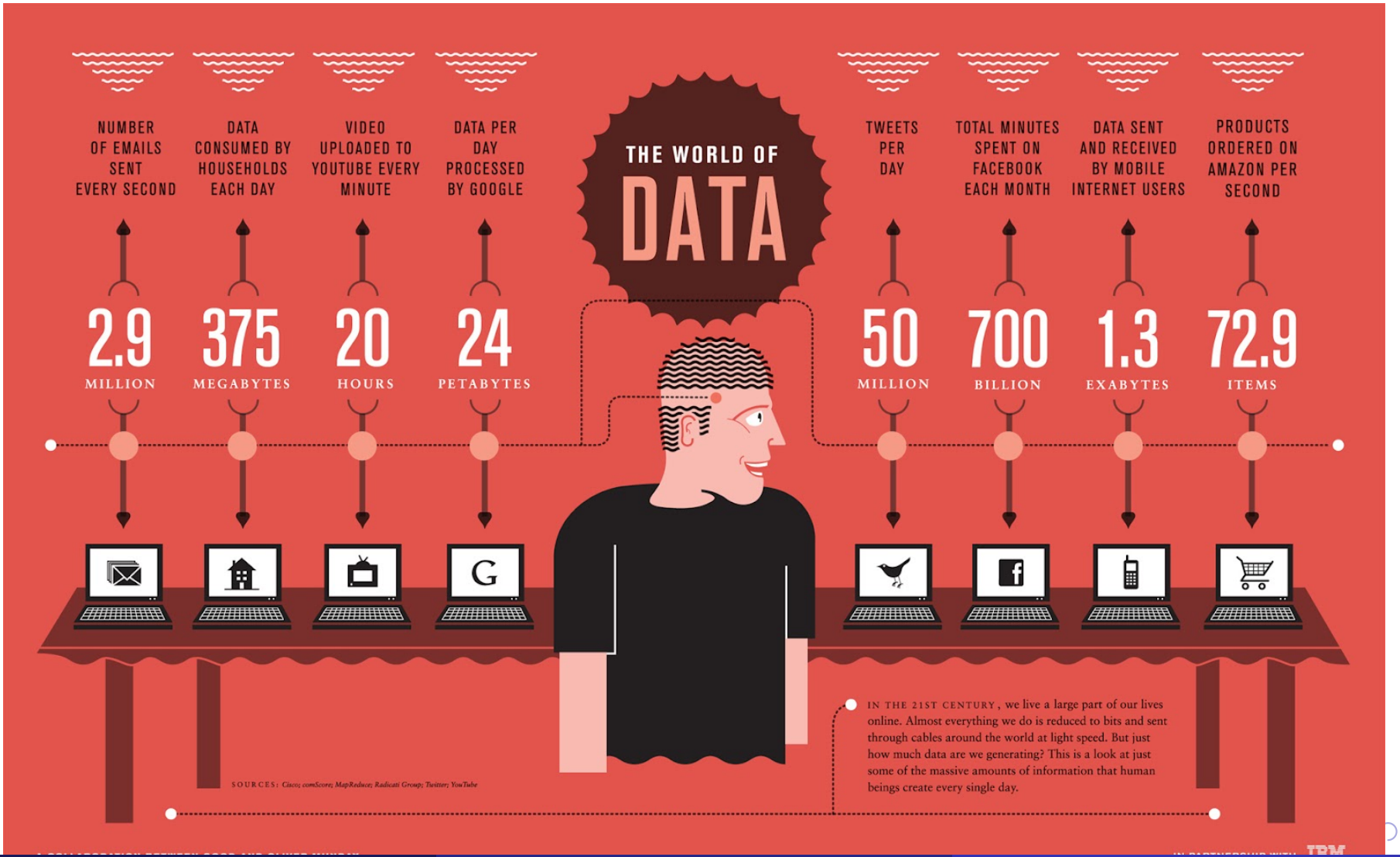
The Growth of Data

- Why is statistical NLP important today?

The Data Problem

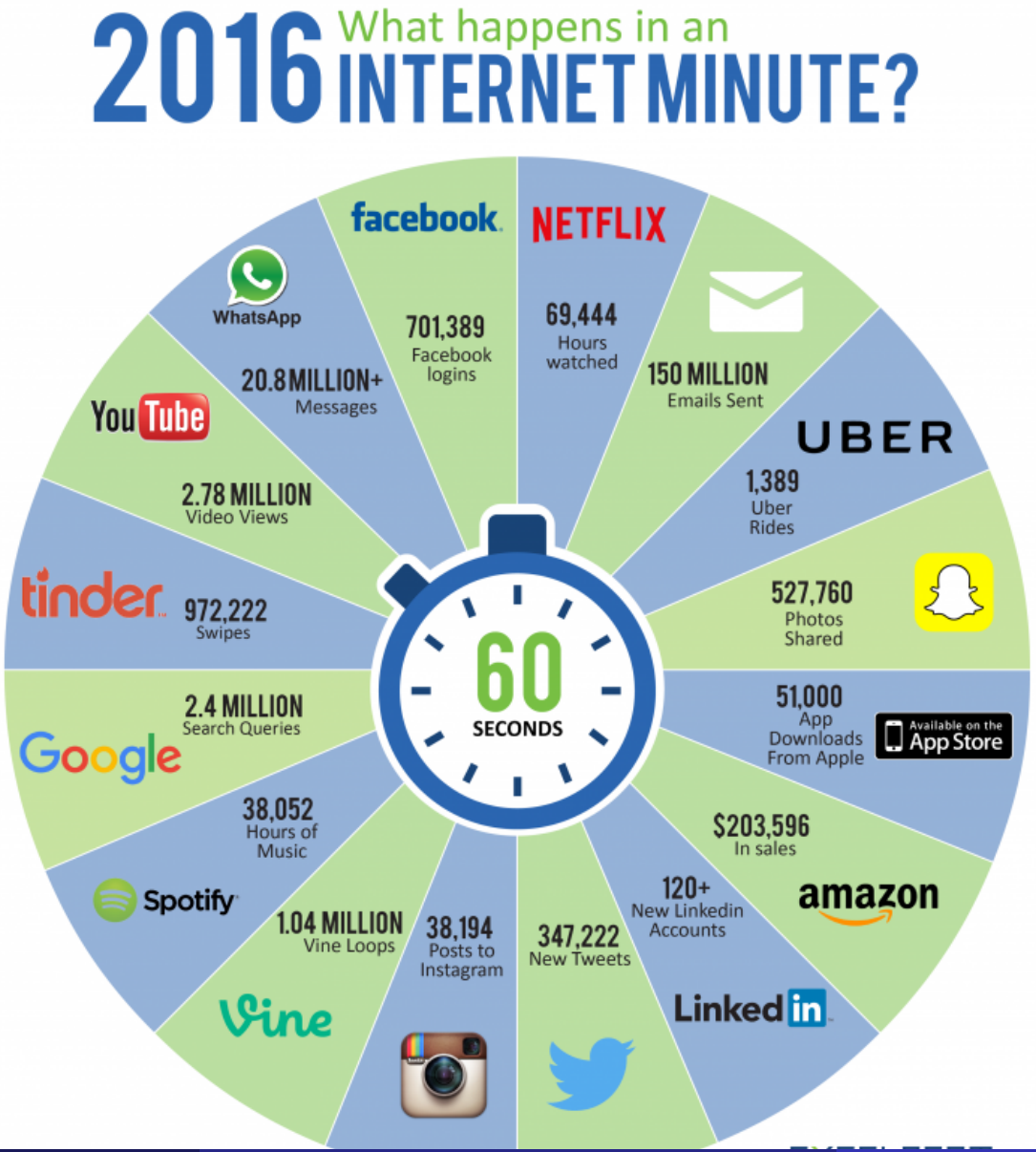
The Growth of Data

- Why is statistical NLP important today?



The Data Problem

The Web (2016)



The Data Problem

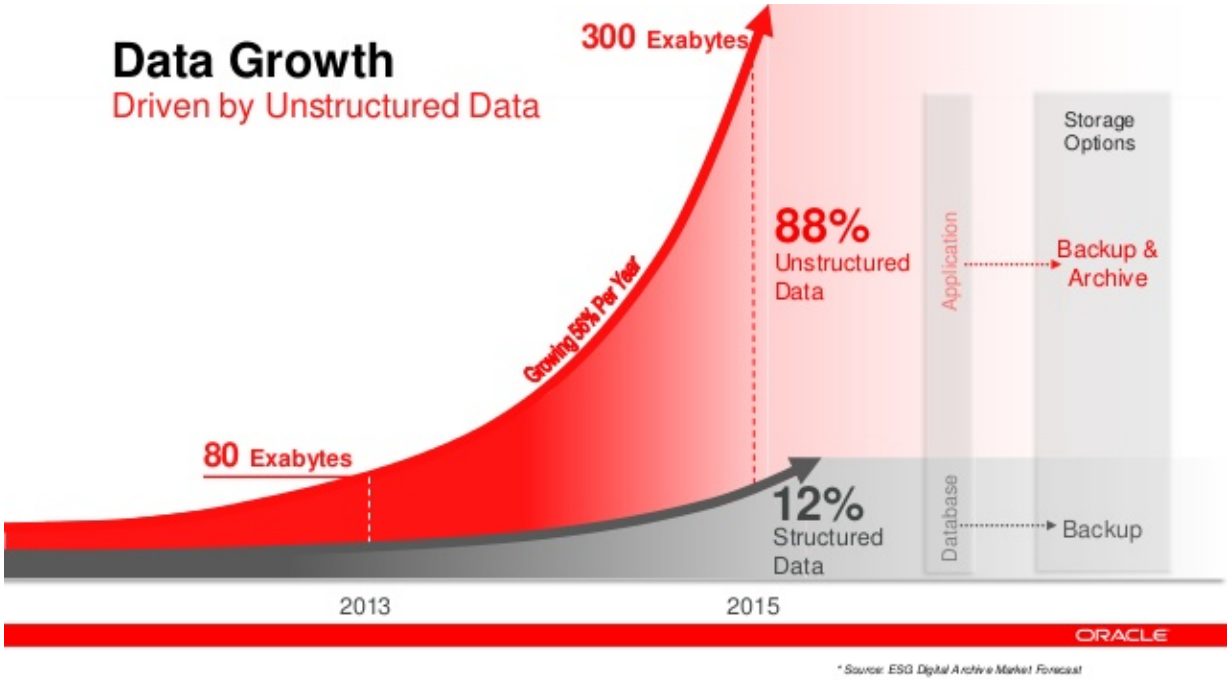
The Web (2017)

The Data Problem

The Web (2018)

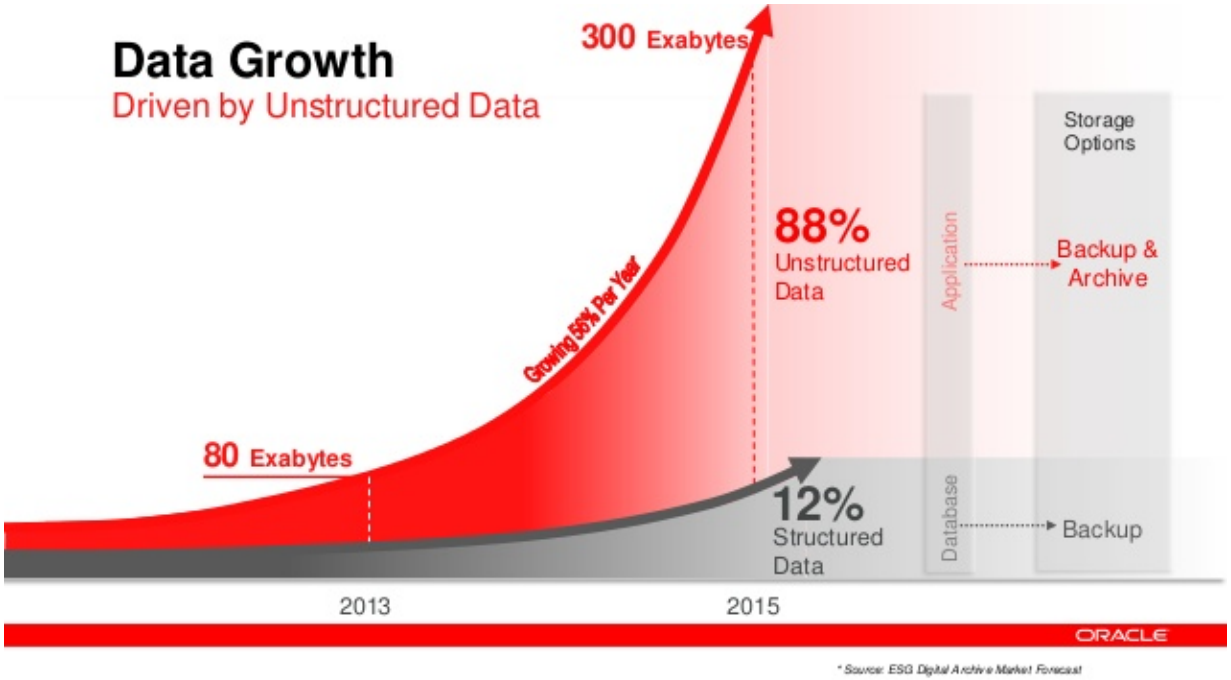
The Data Problem

Problems



The Data Problem

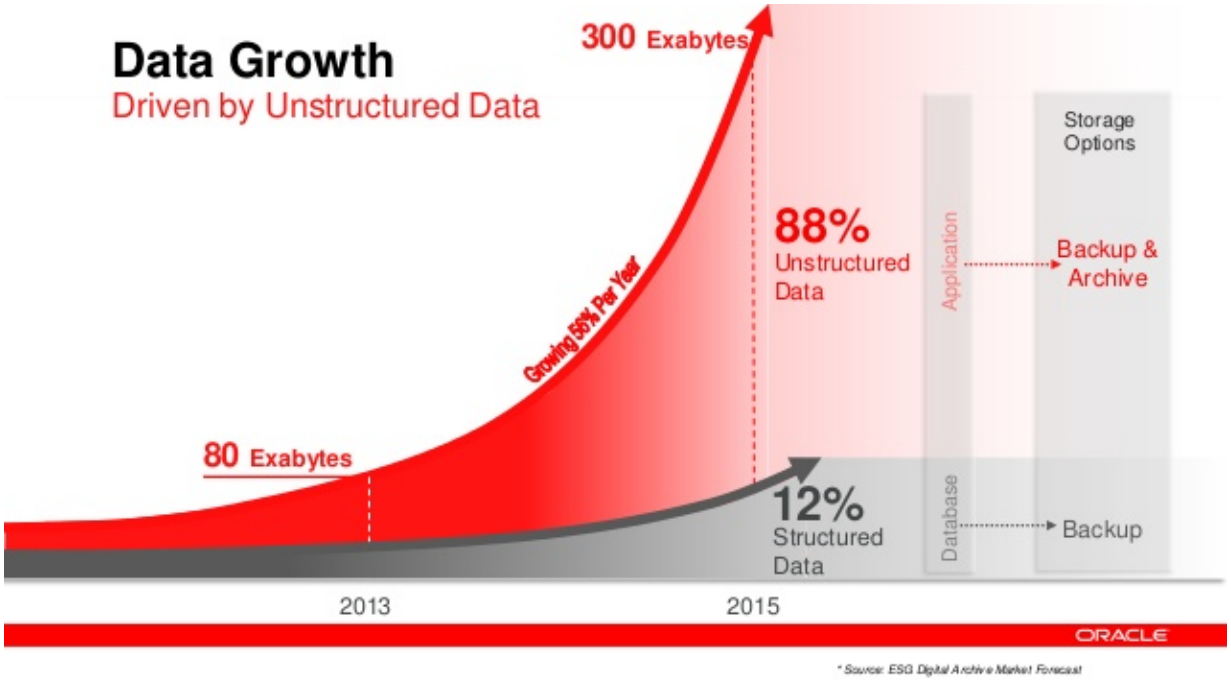
Problems



1. **Volume**: Amount of data grows continuously

The Data Problem

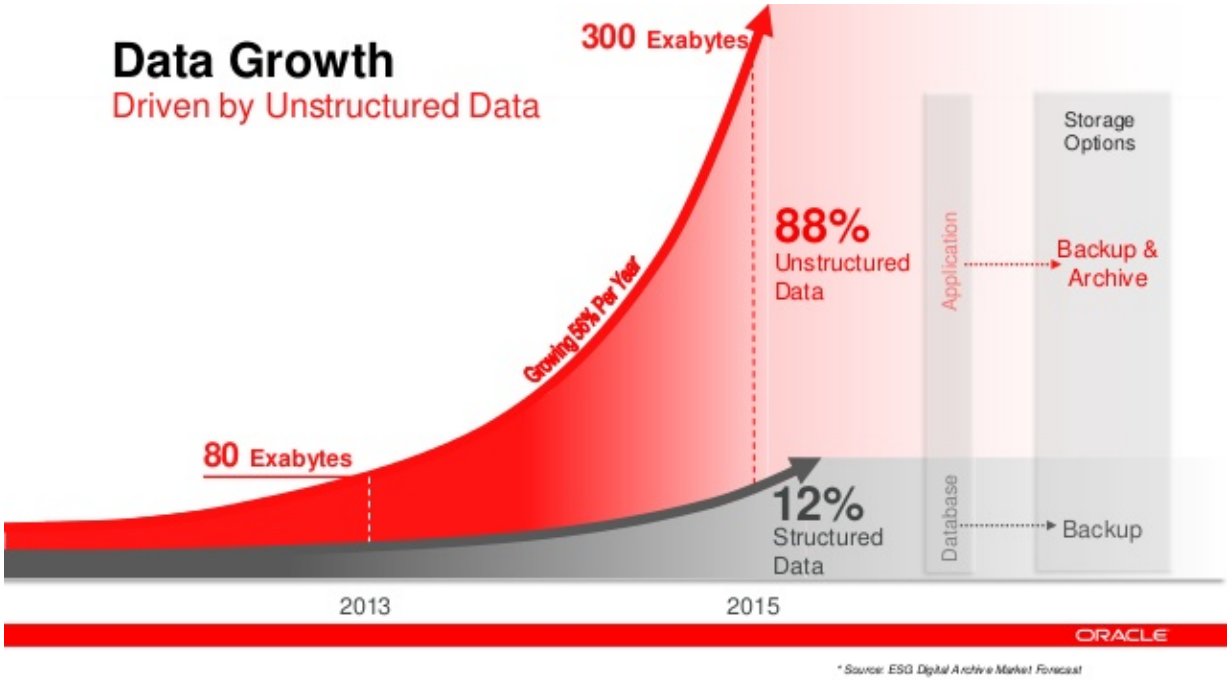
Problems



1. **Volume**: Amount of data grows continuously
2. **Velocity**: We produce and consume data exponentially

The Data Problem

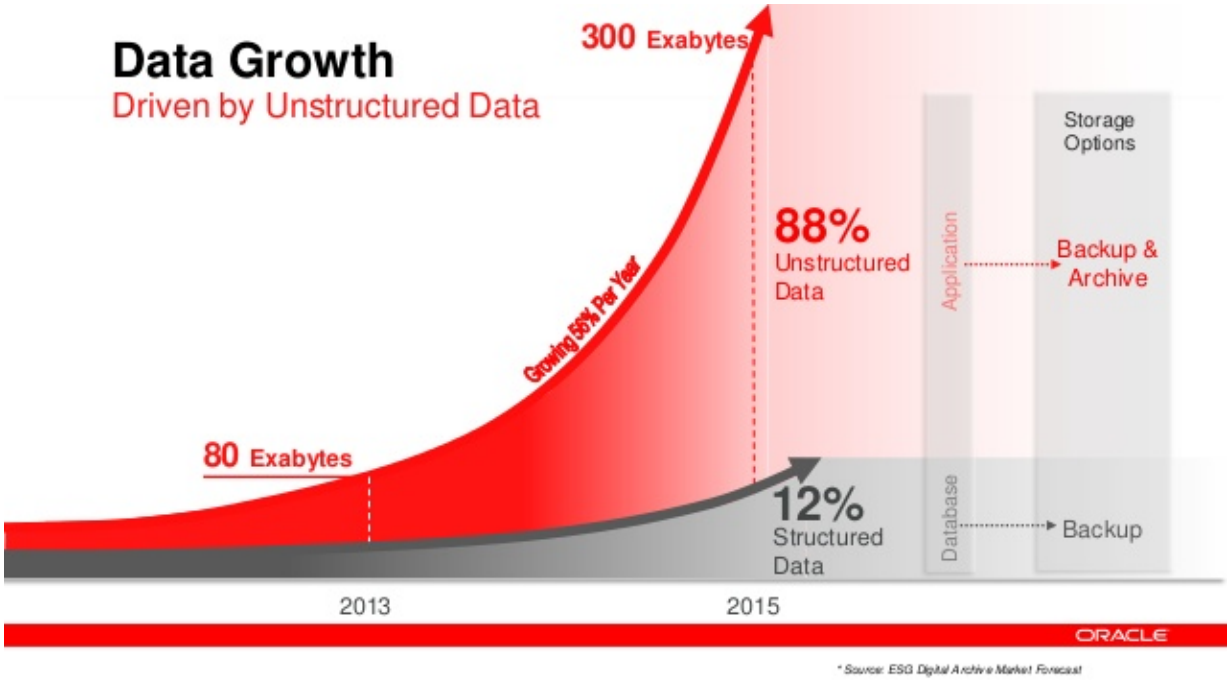
Problems



1. **Volume**: Amount of data grows continuously
2. **Velocity**: We produce and consume data exponentially
3. **Variety**: Most of the data is unstructured

The Data Problem

Problems



1. **Volume**: Amount of data grows continuously
2. **Velocity**: We produce and consume data exponentially
3. **Variety**: Most of the data is unstructured
4. **Value**: Need to gather valuable information out of flood of data

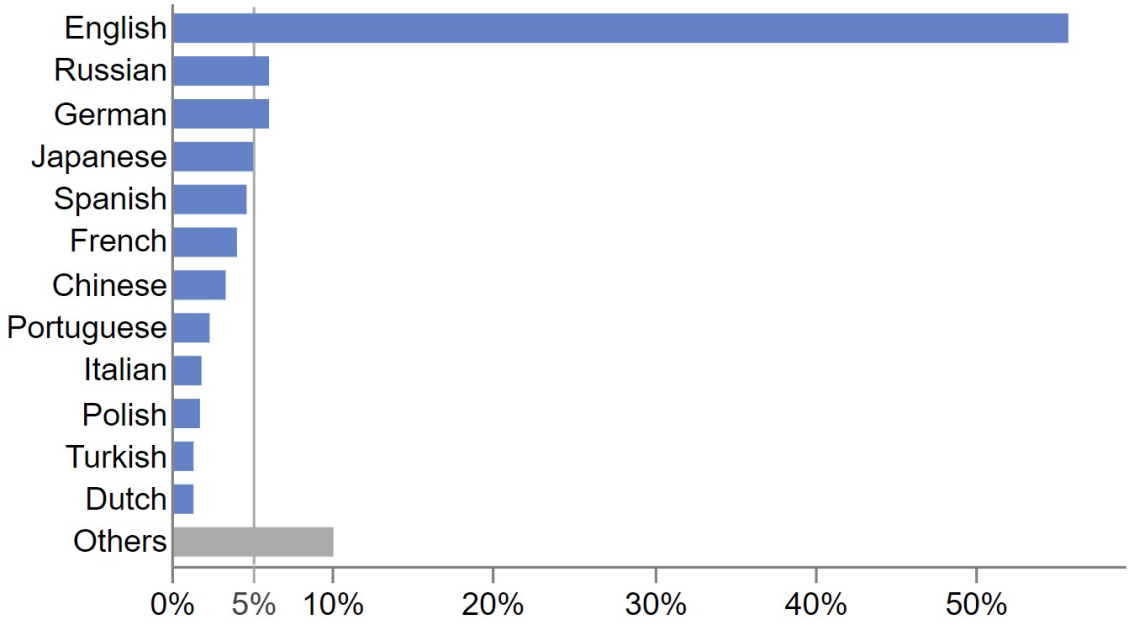
Section 5

Applications

Applications

Machine Translation

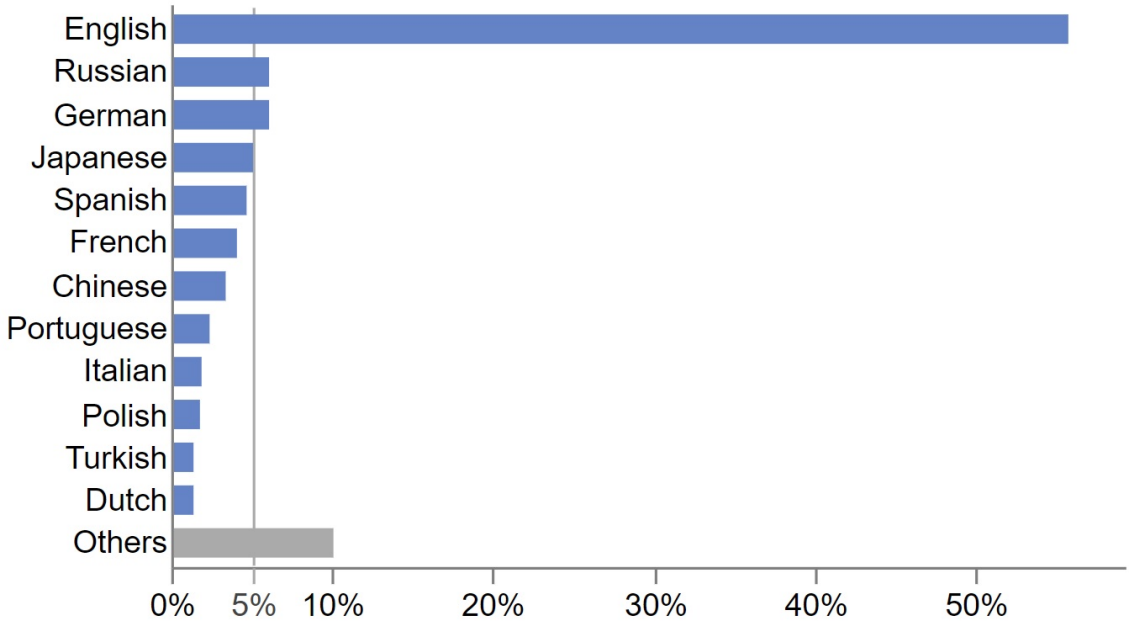
- 40% of the Web is not in English
- Web booms in non-English-speaking countries
- Need to translate across languages automatically



Applications

Machine Translation

- 40% of the Web is not in English
- Web booms in non-English-speaking countries
- Need to translate across languages automatically



- **Challenges**
  - Which fragments? [learning to translate]
  - Efficiency? [fast translation]
  - Fluency vs fidelity (esp. idioms, RDF resources)

Applications

Knowledge Extraction

- Transform text into structured knowledge

Example

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

- Transform text into structured knowledge

Example

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

```
:NewYorkTimes a :Company .
:NewYorkTimes rdfs:label ‘‘New York Times Co.’’@en .
:NewYorkTimes :hasPresident :RussellTLewis.   ...
```

- Transform text into structured knowledge

Example

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

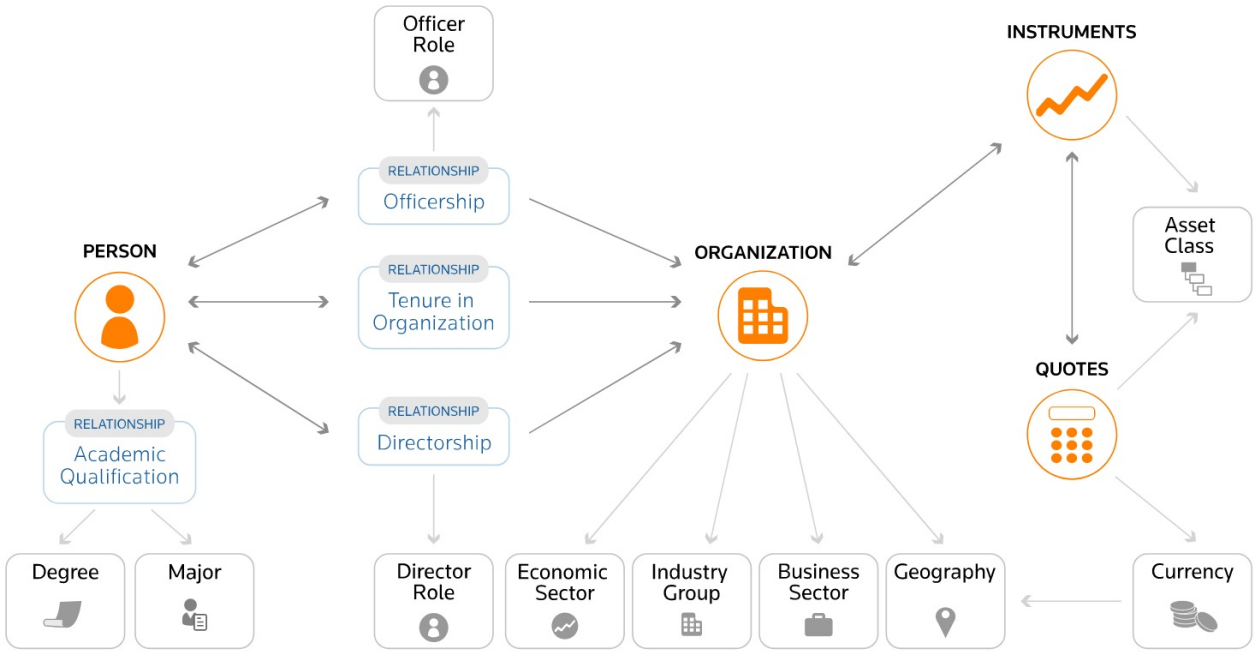
```
:NewYorkTimes a :Company .
:NewYorkTimes rdfs:label ‘‘New York Times Co.’’@en .
:NewYorkTimes :hasPresident :RussellTLewis.   ...
```

- **Challenges**
  - Coreference resolution
  - Emerging entities
  - Entity disambiguation, open relation extraction, time scoping

Applications

Knowledge Graphs – PermID

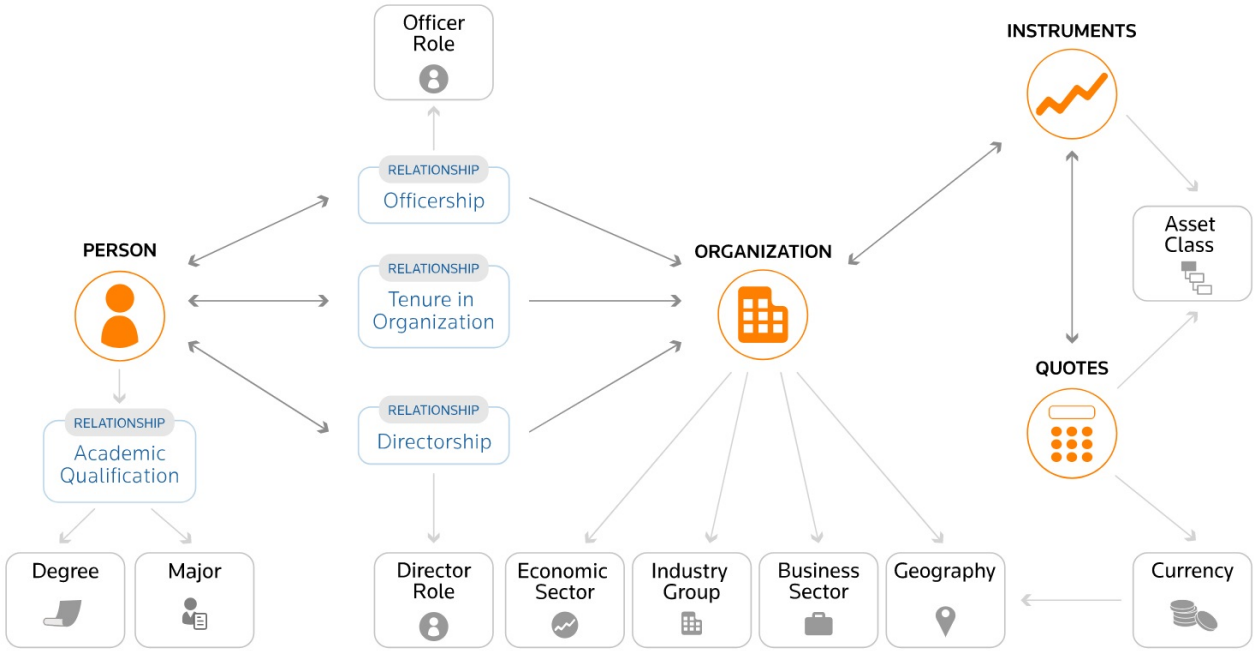
- Supports structured search for information
- Focus on company data (2.35B facts)
- Extracted from the Web using crawling



Applications

Knowledge Graphs – PermID

- Supports structured search for information
- Focus on company data (2.35B facts)
- Extracted from the Web using crawling



- **Challenges**
  - Fact checking
  - Efficient queries
  - Verbalization

Applications

Question Answering

Applications

Question Answering

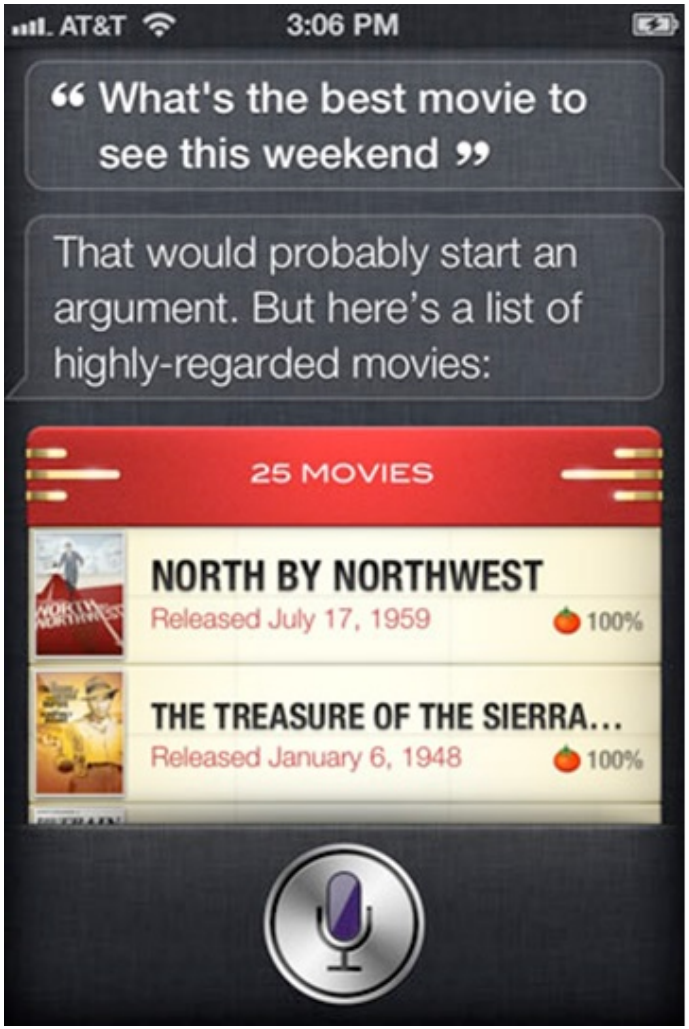
- Need to "understand" text so as to answer complex questions
- First impressive results but "US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle."

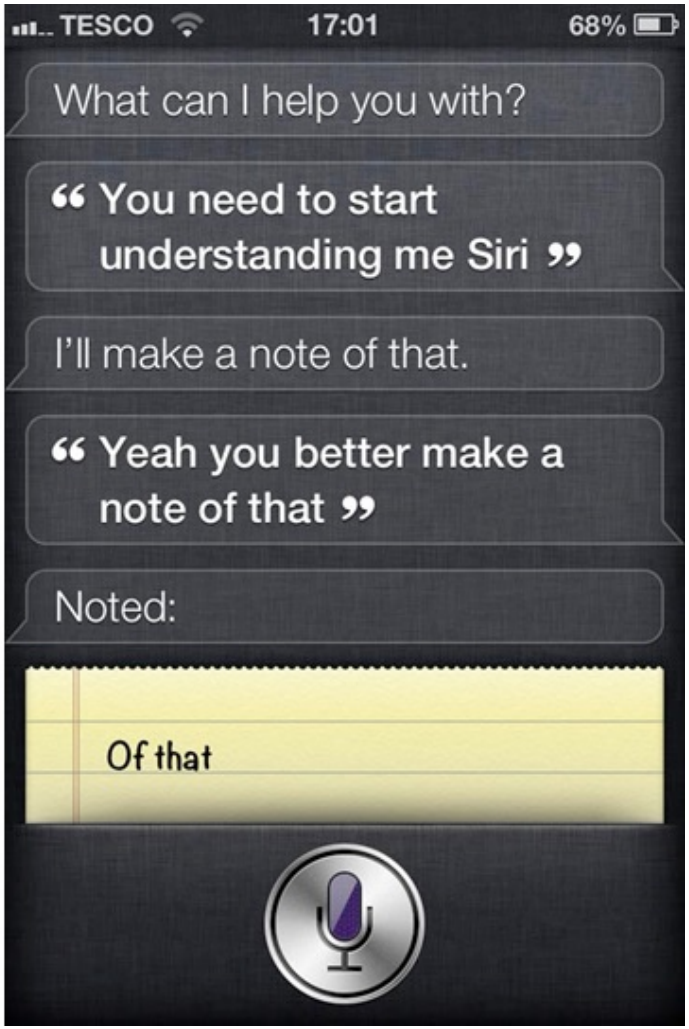
- Need for better text understanding

Applications

Mobile Question Answering

- Need for better contextual understanding
- **Challenge**: Language processing with limited resources





Applications

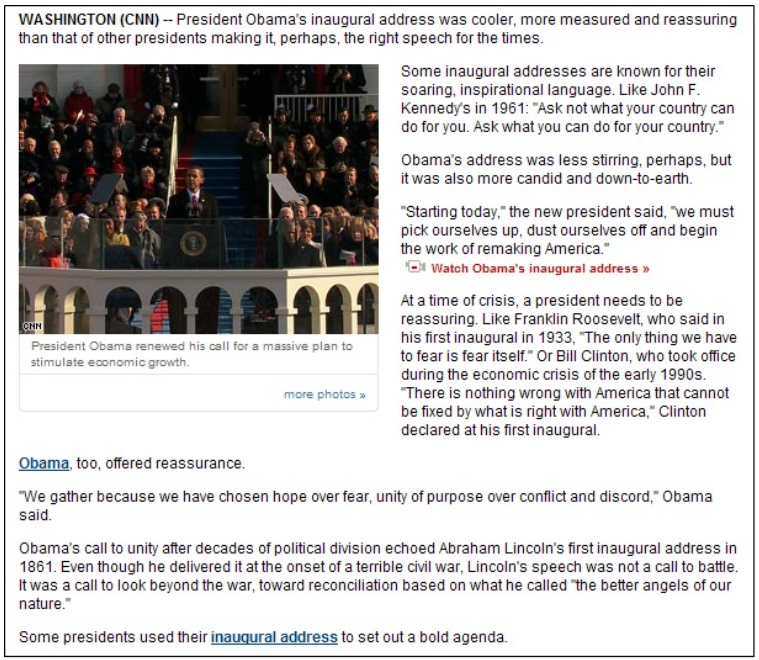
Sentiment and Opinion Analysis

- **Today**: In 2012 election, automatic sentiment analysis actually being used to complement traditional methods (surveys, focus groups)
- **Past**: "Sentiment Analysis" research started in 2002 or so
- **Future**: Growing research toward computational social science, digital humanities (psychology, communication, literature and more)
  Challenge: Need statistical models for deeper semantic understanding — subtext, intent, nuanced messages

Applications

Text Summarization

- Remember the data problem ...
- Need for understandable summaries of large corpora



WASHINGTON (CNN) -- President Obama's inaugural address was cooler, more measured and reassuring than that of other presidents making it, perhaps, the right speech for the times.

President Obama renewed his call for a massive plan to stimulate economic growth.

more photos »

Some inaugural addresses are known for their soaring, inspirational language. Like John F. Kennedy's in 1961: "Ask not what your country can do for you. Ask what you can do for your country."

Obama's address was less stirring, perhaps, but it was also more candid and down-to-earth.

"Starting today," the new president said, "we must pick ourselves up, dust ourselves off and begin the work of remaking America."

Watch Obama's inaugural address »

At a time of crisis, a president needs to be reassuring. Like Franklin Roosevelt, who said in his first inaugural in 1933, "The only thing we have to fear is fear itself." Or Bill Clinton, who took office during the economic crisis of the early 1990s. "There is nothing wrong with America that cannot be fixed by what is right with America," Clinton declared at his first inaugural.

Obama, too, offered reassurance.

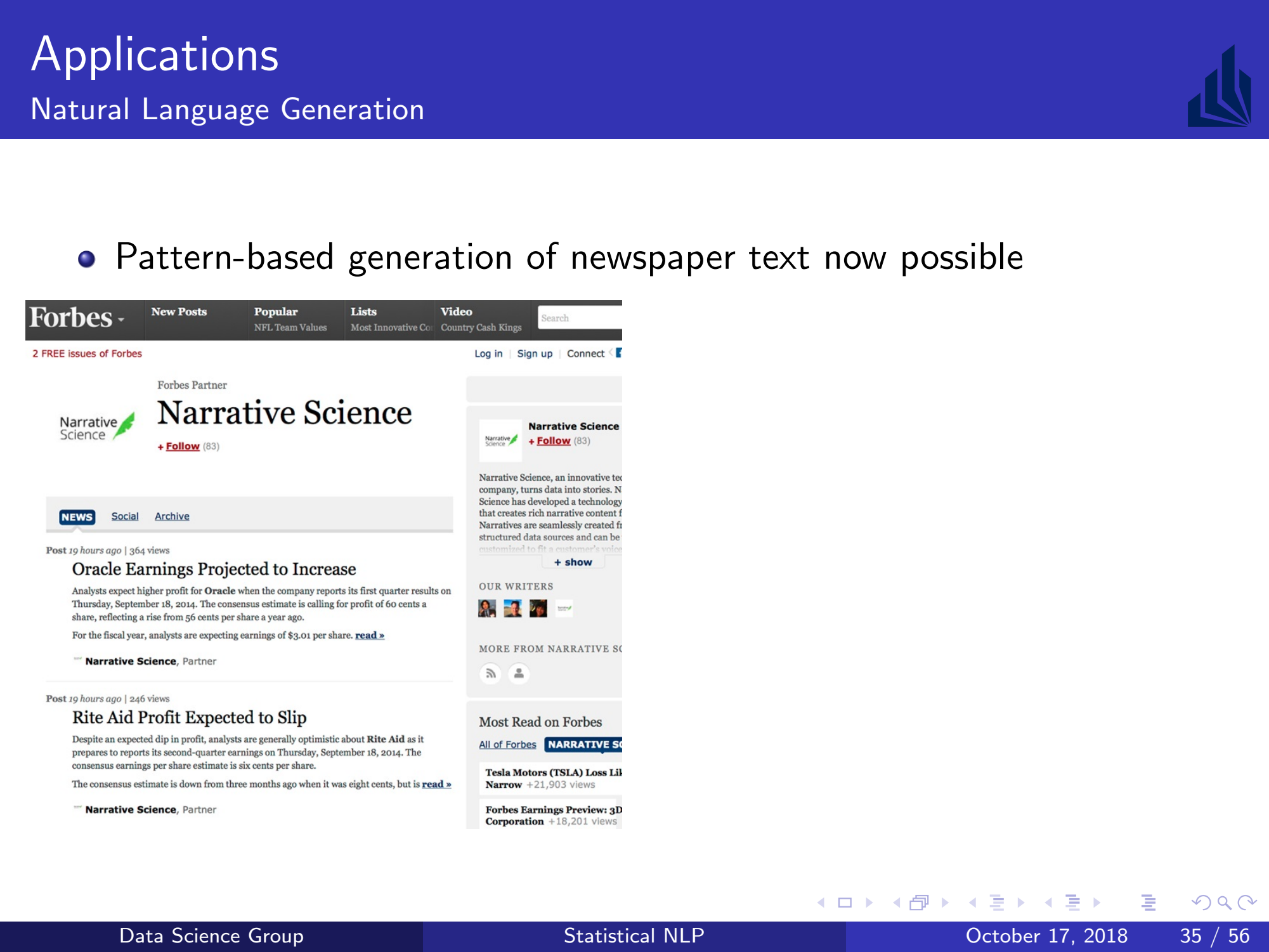
"We gather because we have chosen hope over fear, unity of purpose over conflict and discord," Obama said.

Obama's call to unity after decades of political division echoed Abraham Lincoln's first inaugural address in 1861. Even though he delivered it at the onset of a terrible civil war, Lincoln's speech was not a call to battle. It was a call to look beyond the war, toward reconciliation based on what he called "the better angels of our nature."

Some presidents used their inaugural address to set out a bold agenda.

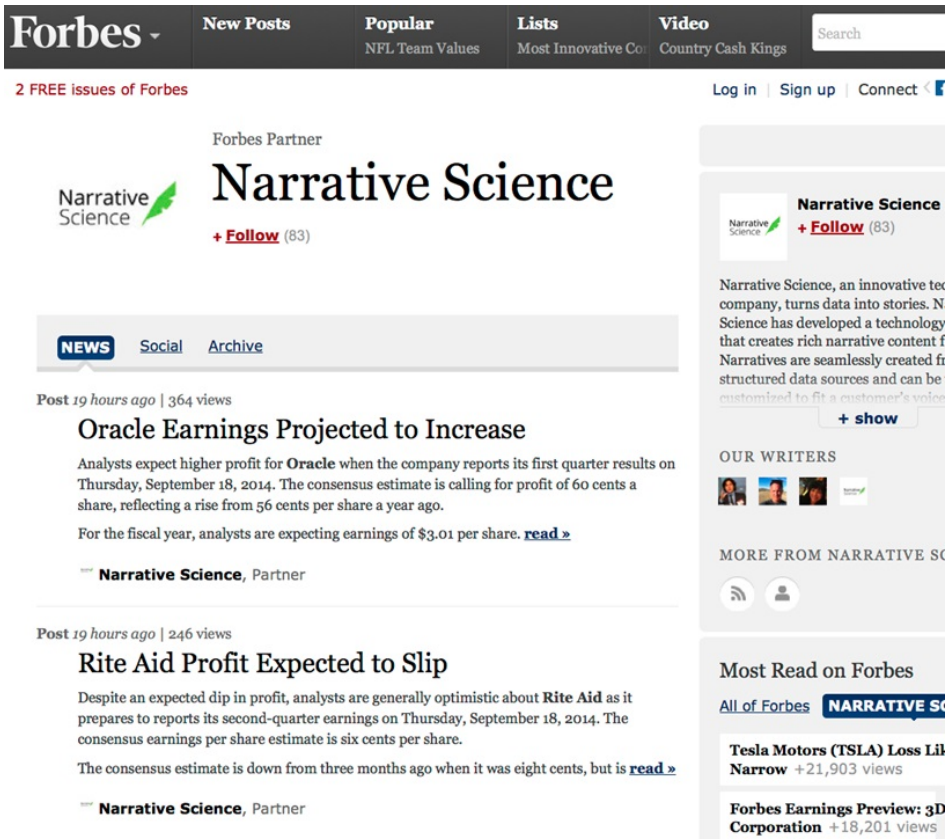
- Obama's inaugural speech more reassuring than previous ones.
- Country has chosen hope over fear, Obama says.

- **Challenge**: Highly context-dependent, cross-document summaries?



- Pattern-based generation of newspaper text now possible

- Pattern-based generation of newspaper text now possible



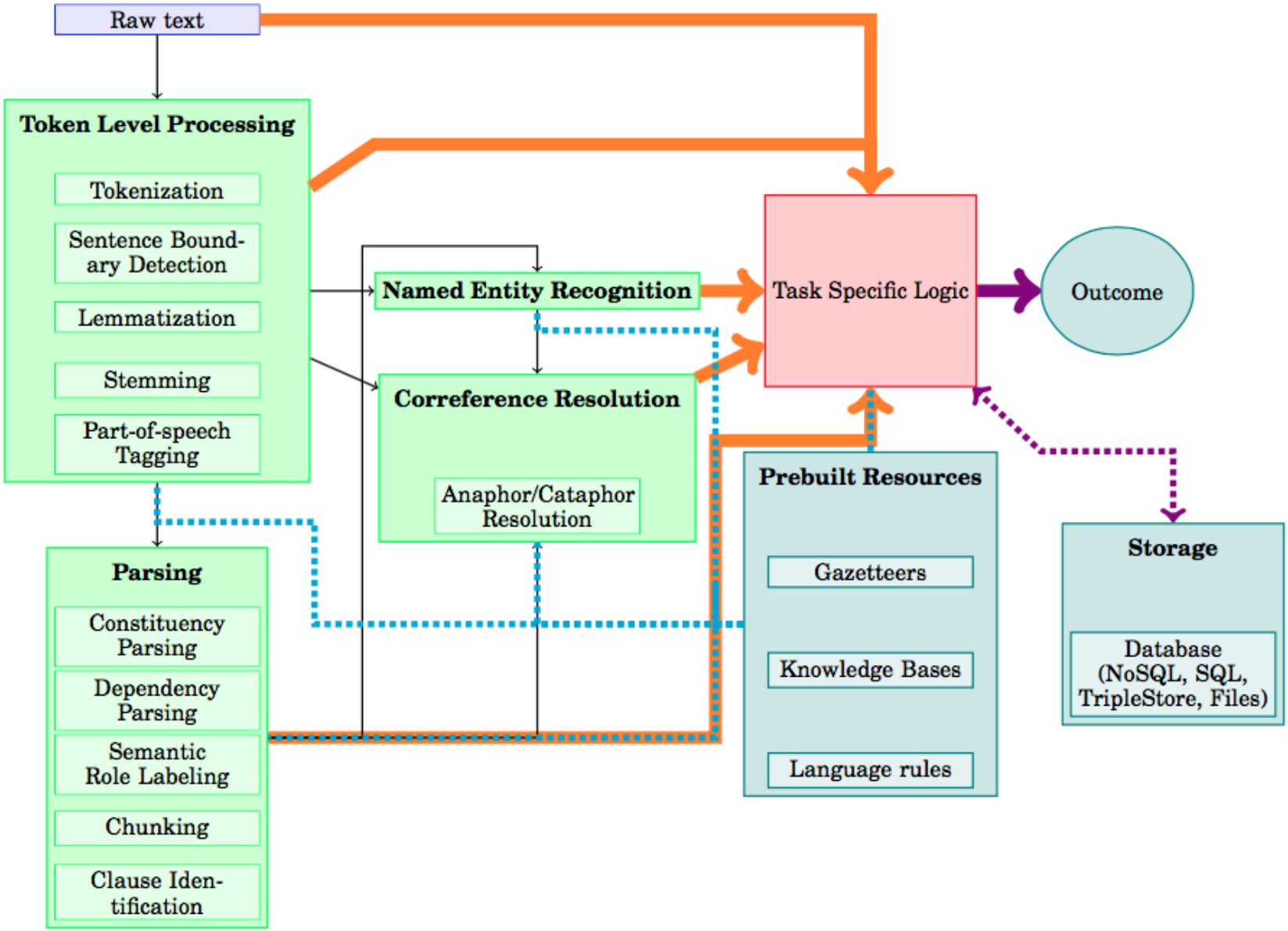
- **Challenges**
  - Very restricted domains
  - Manual rule creation
  - Domain-independent and cross-source NLG

Section 6

Prototypical Pipeline

Prototypical Pipeline

Overview



Prototypical Pipeline

Corpus

- Formally a collection $\mathcal{D} = \{D_1, \ldots, D_n\}$ of documents
- Plethora available, see `http://www.nltk.org/nltk_data/`

Brown Corpus

- Brown University Standard Corpus of Present-Day American English
- $n = 500$ samples of English text
- Approx. 10^6 tokens
- 15 genres incl. press, religion, skills and hobbies
- POS-Tagged

1. Cleaning: Remove unwanted tokens from documents $D_1, \ldots, D_n$
   - Examples?

Prototypical Pipeline

Preprocessing

1. Cleaning: Remove unwanted tokens from documents $D_1, \ldots, D_n$
   - Examples? HTML Tags
   - Capitalized words, e.g., "HAHAHA!"
   - Spelling errors, e.g., "We saw three graffes today".
2. Classification: Find $f : \mathcal{D} \to \mathcal{C}$ based on training data

Prototypical Pipeline

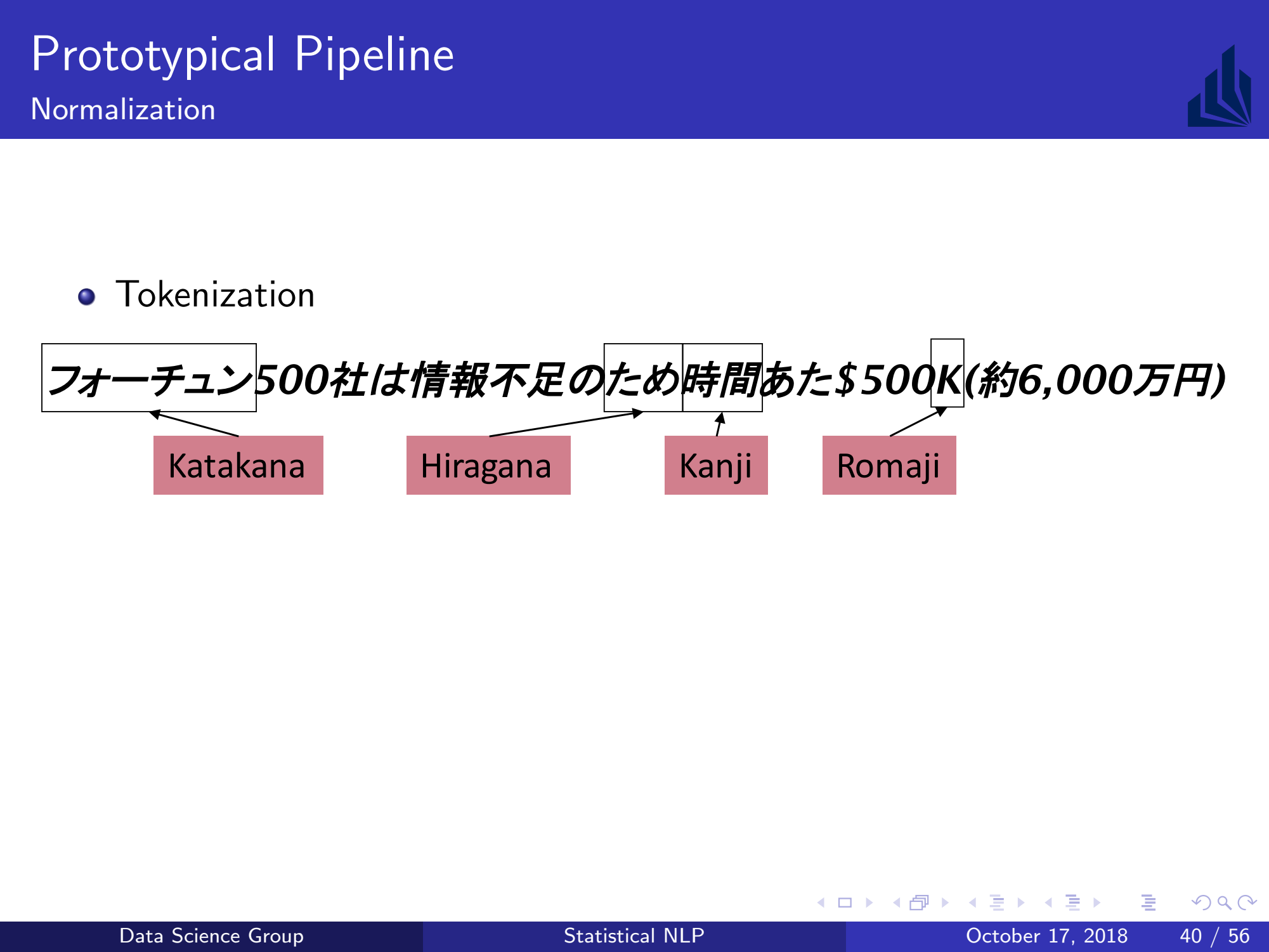
Preprocessing

1. Cleaning: Remove unwanted tokens from documents $D_1, \ldots, D_n$
   - Examples? HTML Tags
   - Capitalized words, e.g., "HAHAHA!"
   - Spelling errors, e.g., "We saw three graffes today".
2. Classification: Find $f : \mathcal{D} \to \mathcal{C}$ based on training data
   - Spam detection
   - Domain classification, e.g., sports, news, hobbies
3. Deduplication: Find $\{(D_i, D_j) \in \mathcal{D} : \sigma(D_i, D_j) \geq \theta\}$

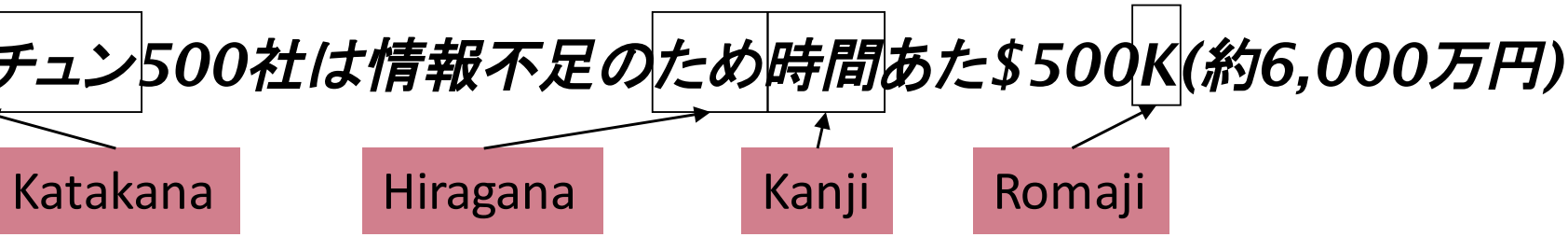
Prototypical Pipeline

Preprocessing

1. Cleaning: Remove unwanted tokens from documents $D_1, \ldots, D_n$
   - Examples? HTML Tags
   - Capitalized words, e.g., "HAHAHA!"
   - Spelling errors, e.g., "We saw three graffes today".
2. Classification: Find $f : \mathcal{D} \rightarrow \mathcal{C}$ based on training data
   - Spam detection
   - Domain classification, e.g., sports, news, hobbies
3. Deduplication: Find $\{(D_i, D_j) \in \mathcal{D} : \sigma(D_i, D_j) \geq \theta\}$
   - New versions of corpus
   - Incremental datasets from the Web

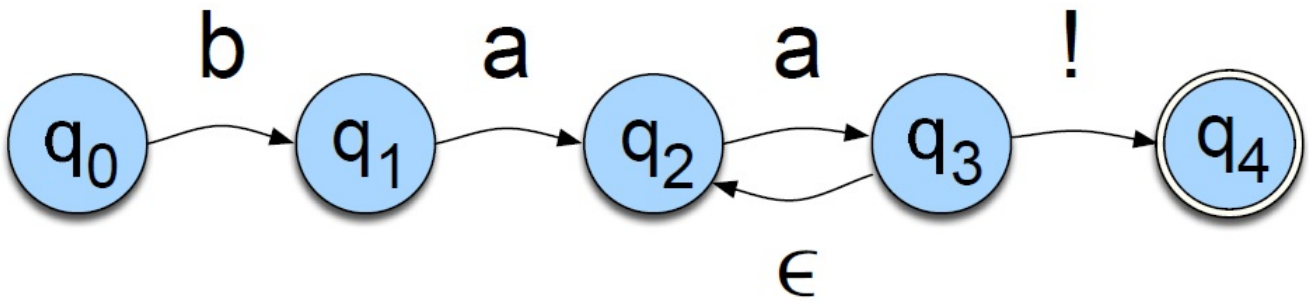


- Tokenization



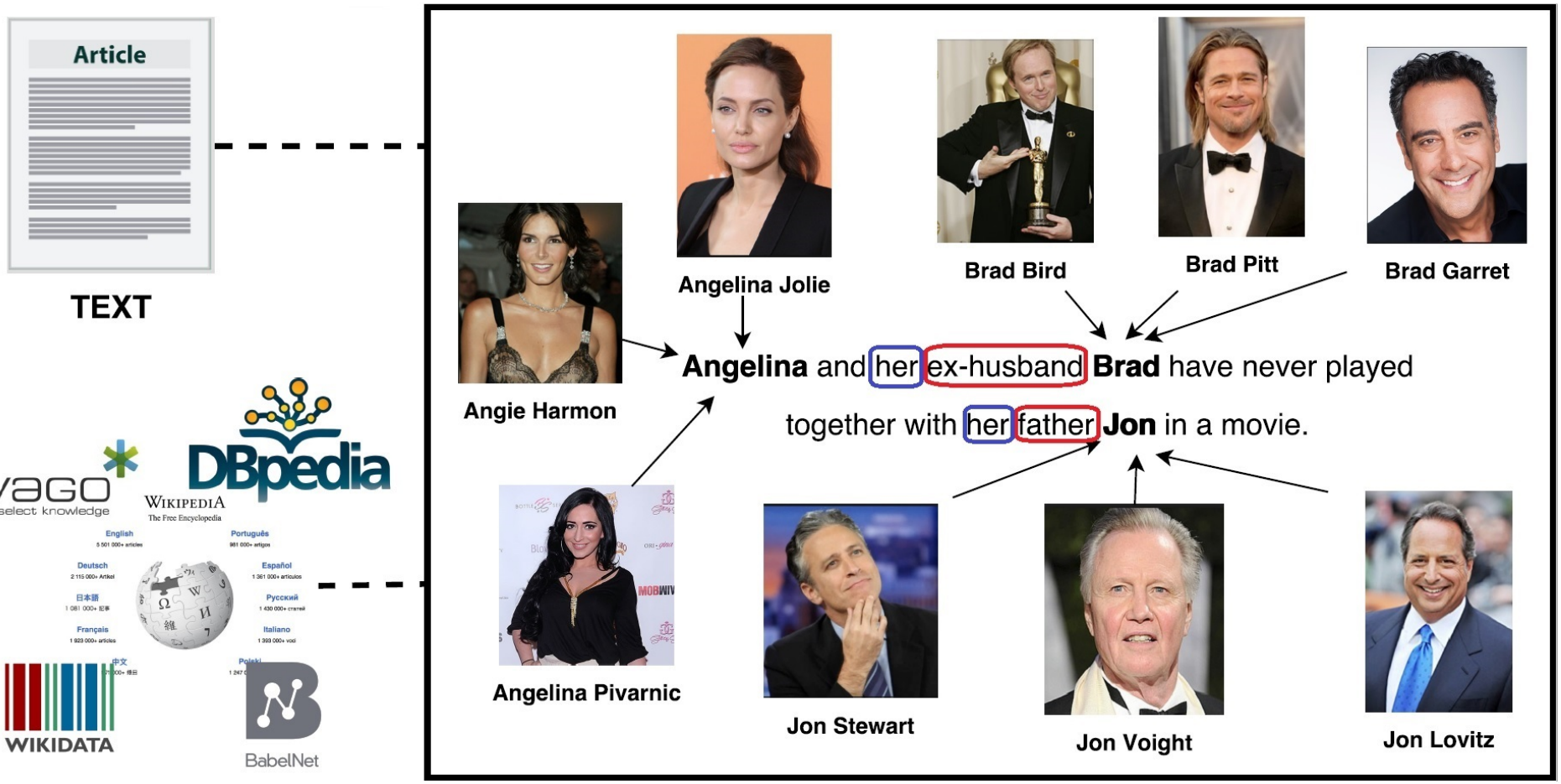
- Tokenization

- Regex

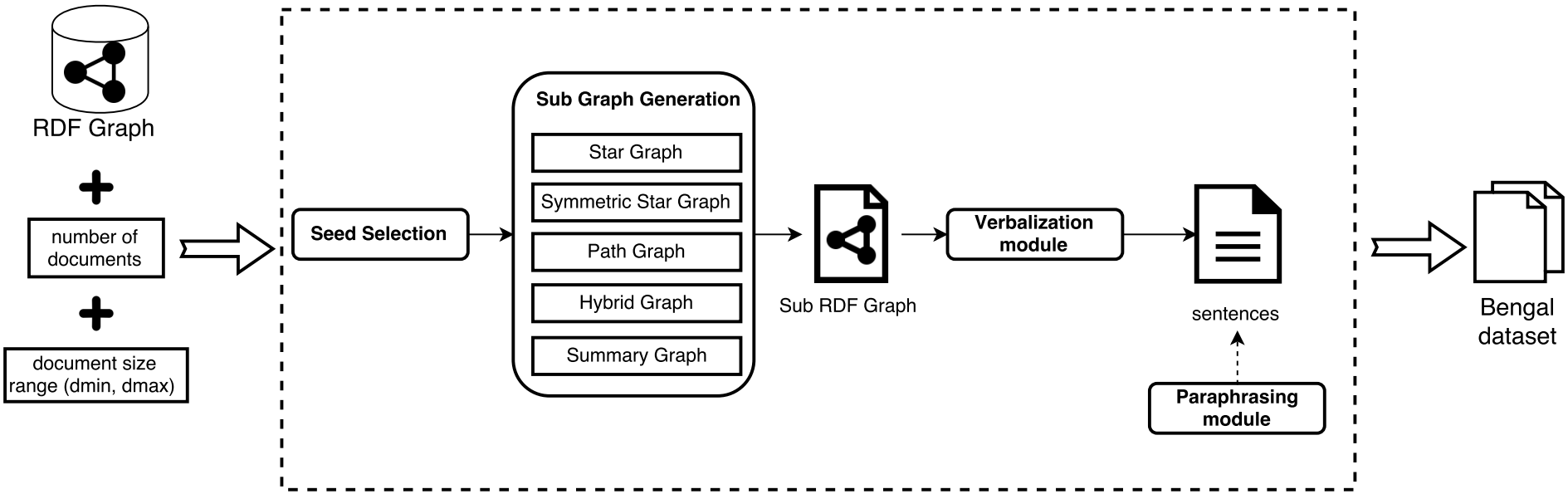


Prototypical Pipeline

Knowledge Extraction

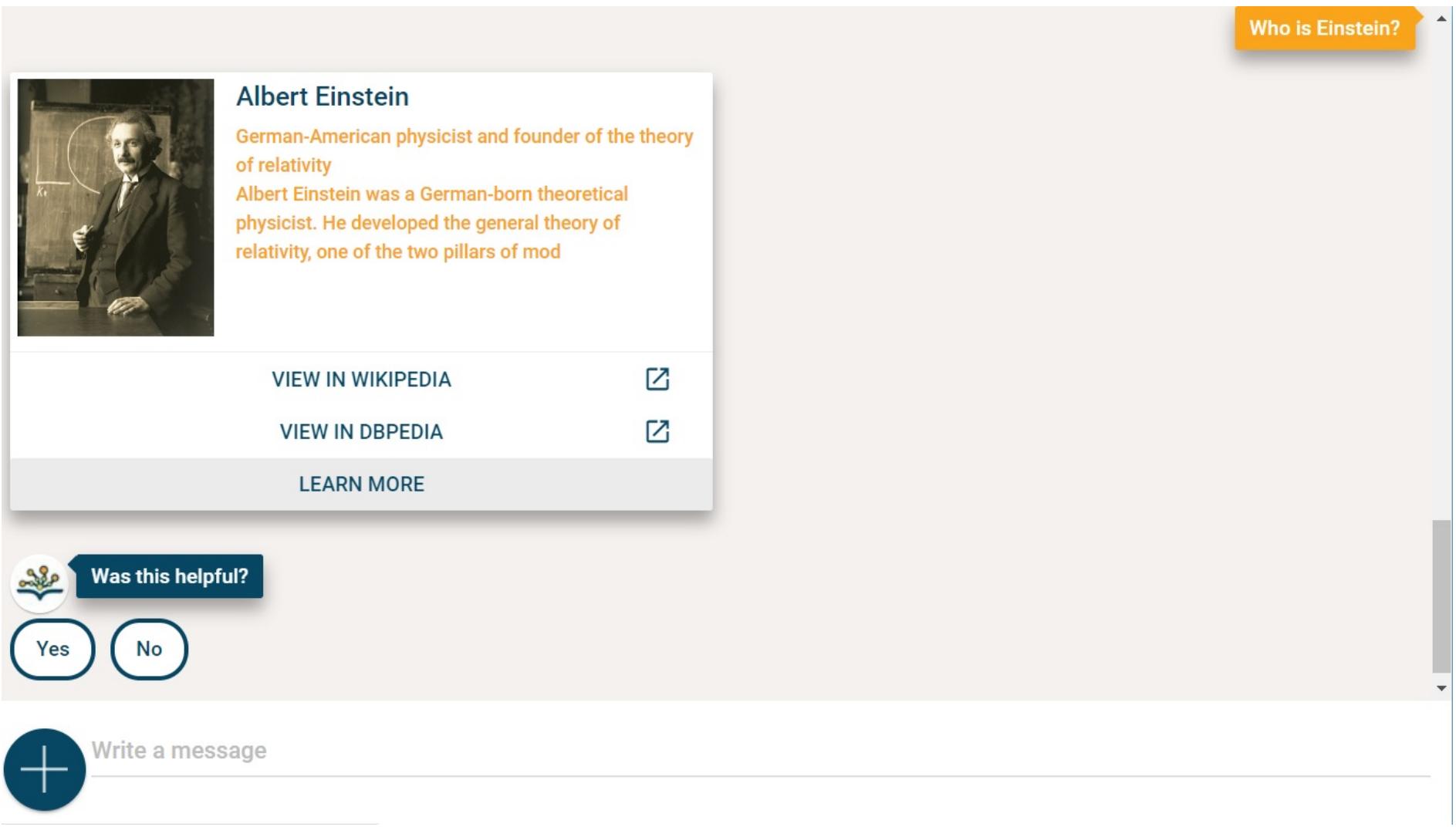


- Common part of modern NLP-based applications
- Goal: Generate document D from structured representation (e.g., answer to a question)



Prototypical Pipeline

Application



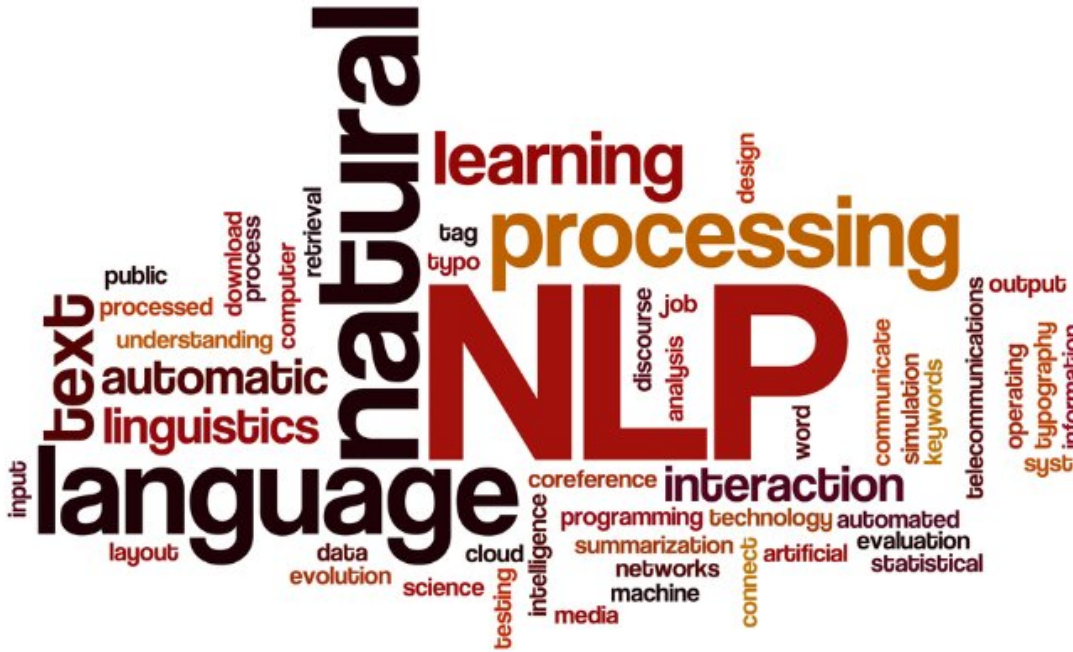
Section 7

Structure of the Course

Structure of the Course

Overview

1. Oct. 17: Introduction
2. Oct. 24: Text normalization
3. Oct. 31: Language modeling
4. Nov. 07: Spelling correction
5. Nov. 14: Deduplication
6. Nov. 21: Classification
7. Nov. 28: HMMs
8. Dec. 05: Grammar
9. Dec. 12: Parsing
10. Dec. 19: No lecture
11. Jan. 09: Word vectors
12. Jan. 16: Word Sense Disambiguation
13. Jan. 23: Question Answering
14. Jan. 30: Summary, Questions and Answers



Take Away

Summary

- Origins of linguistic science
- The data problem
- Goals of statistical natural language processing (SNLP)
- Applications of SNLP
- Basic assumptions of SNLP
- A prototypical NLP pipeline

Section 8

Mini-Project

Mini-Project

Structure

- Given
  - Knowledge base K
  - Training dataset with annotated facts
  - e.g., `((:Einstein, :birthPlace, :Geneva), -1)`
  - Benchmarking platform GERBIL (details during the seminar)

Mini-Project

Structure

- Given
  - Knowledge base K
  - Training dataset with annotated facts
  - e.g., `((:Einstein, :birthPlace, :Geneva), -1)`
  - Benchmarking platform GERBIL (details during the seminar)
- Goal
  1. Build fact checking algorithm $f(s, p, o) \in [-1, +1]$
  2. Evaluate using GERBIL platform (ROC)
  3. Outperform GERBIL baseline
  4. Document
     - Approach (incl. running example)
     - How to compile & run
     - Summary of evaluation results
     - 10 examples (not from the benchmark), where the approach fails

Mini-Project

Structure

- Given
  - Knowledge base K
  - Training dataset with annotated facts
  - e.g., `((:Einstein, :birthPlace, :Geneva), -1)`
  - Benchmarking platform GERBIL (details during the seminar)
- Goal
  1. Build fact checking algorithm $f(s, p, o) \in [-1, +1]$
  2. Evaluate using GERBIL platform (ROC)
  3. Outperform GERBIL baseline
  4. Document
     - Approach (incl. running example)
     - How to compile & run
     - Summary of evaluation results
     - 10 examples (not from the benchmark), where the approach fails
- Suggestions
  1. Build group asap
  2. Complete first working prototype early on
  3. Improve iteratively

Idea

The **Semantic Web** is an **extension of the Web**. It makes the structure of knowledge of the Web explicit, websites can be easily extended with structured data, information from different sources can be easily integrated, extended and reused.

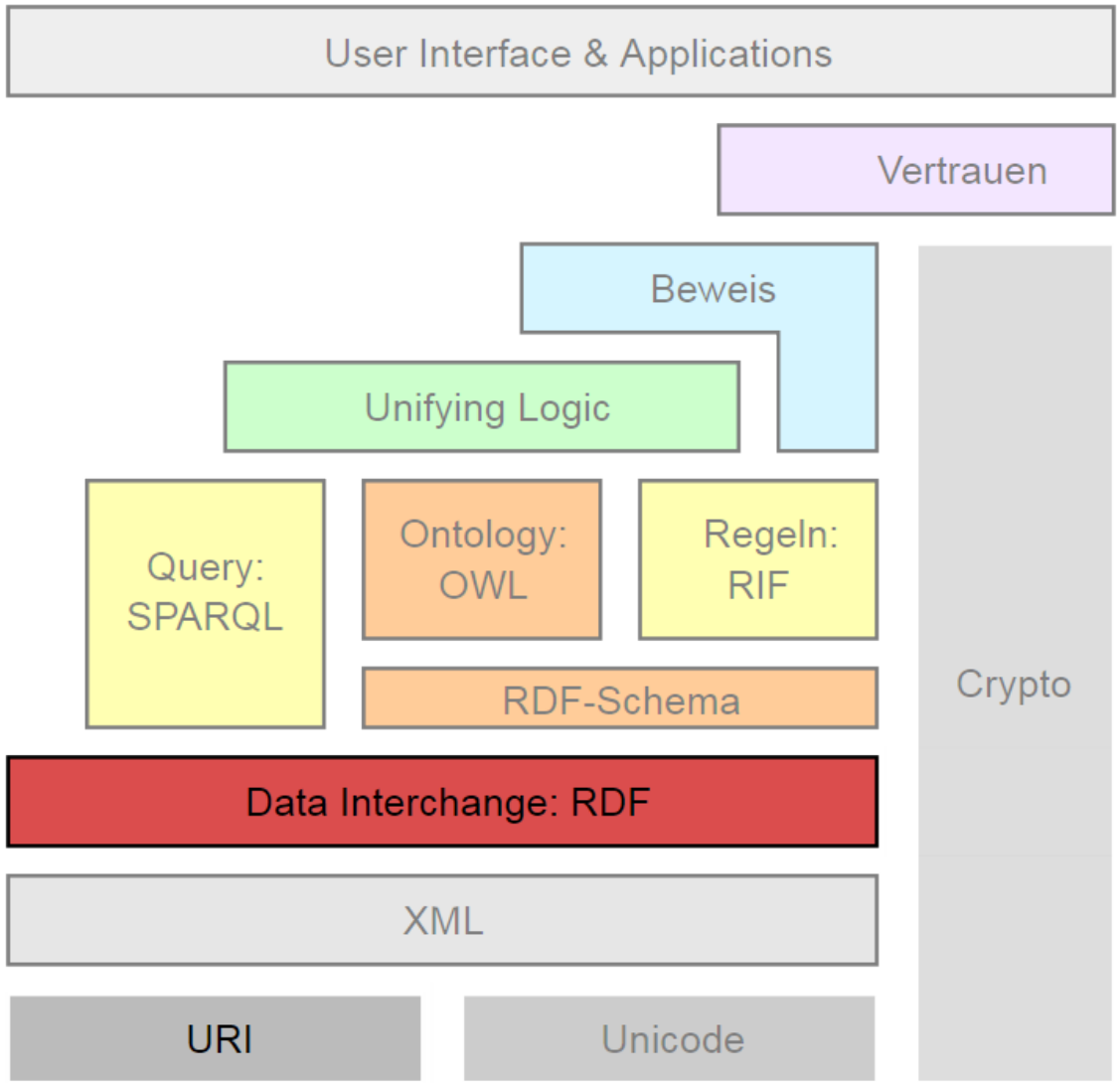
Idea

The **Semantic Web** is an **extension of the Web**. It makes the structure of knowledge of the Web explicit, websites can be easily extended with structured data, information from different sources can be easily integrated, extended and reused.

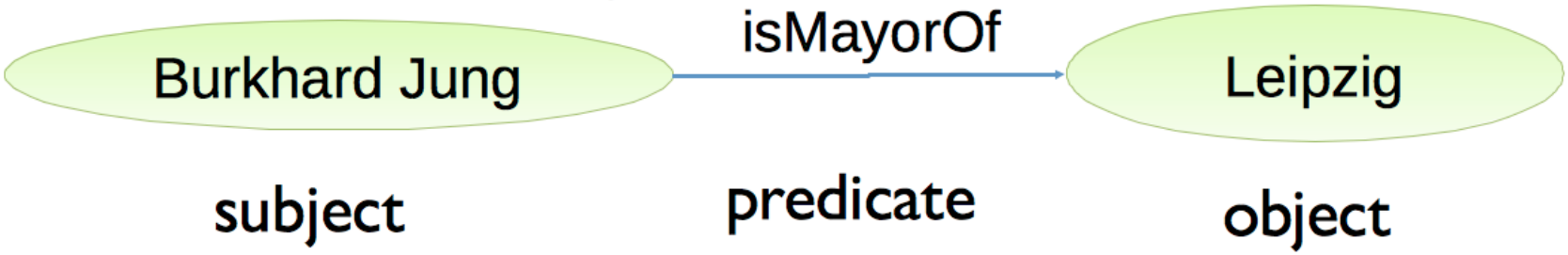
- Open Web-based standards for the description of **domain models**
- **Unique semantics** for the domain models
- Methods for the **generation of novel information**, including
  - Extraction
  - Integration and Fusion
  - Reasoning

Mini-Project

Semantic Web Stack



Facts (also called statements) are expressed as triples



;

- Correspond to basic structure of sentences in English
  - Subject: URI oder blank node
  - Predicate: URI
  - Object: URI, blank node or literal
- Usually conceived of as a graph
  - Nodes and edges in the graph should have a unique meaning
  - Graph should be constructable from list of triples

- Not all triples make sense:

Example

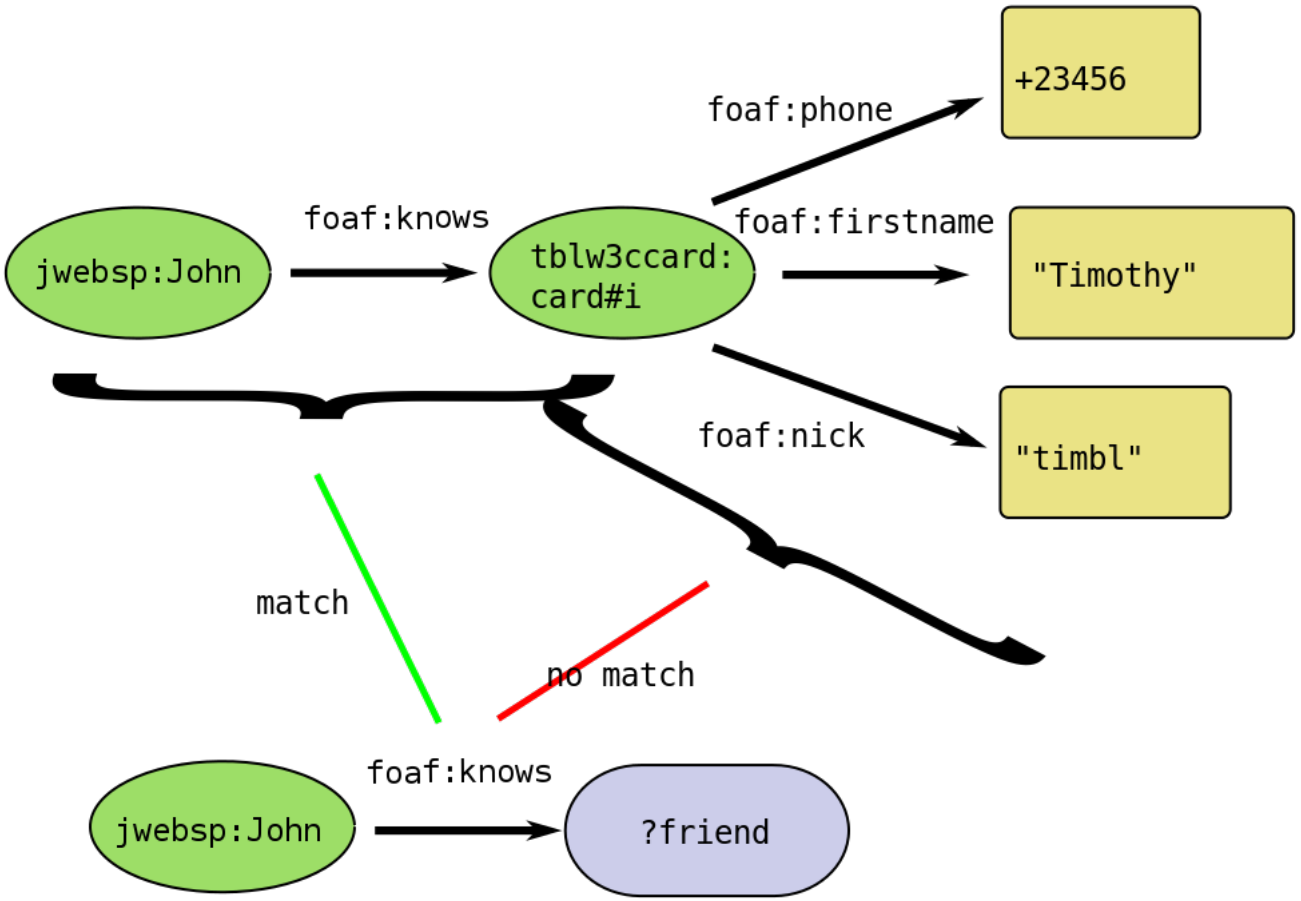
```
Cinema   AlbertEinstein   2012
```

- How can we state the semantics of RDF triples?
- RDF Schema allows specifying the usage of classes and predicates

- First version available since 2004
- Fragment of first-order logics
- Three variations: OWL Lite $\subseteq$ OWL DL $\subseteq$ OWL Full
- No reification in OWL DL $\Rightarrow$ RDFS is a fragment of OWL Full
- OWL DL ist decidable and is equivalent to $\mathcal{SHOIN}$ in description logics
- Long(!) specifications available. You will deal with core concepts.

Mini-Project

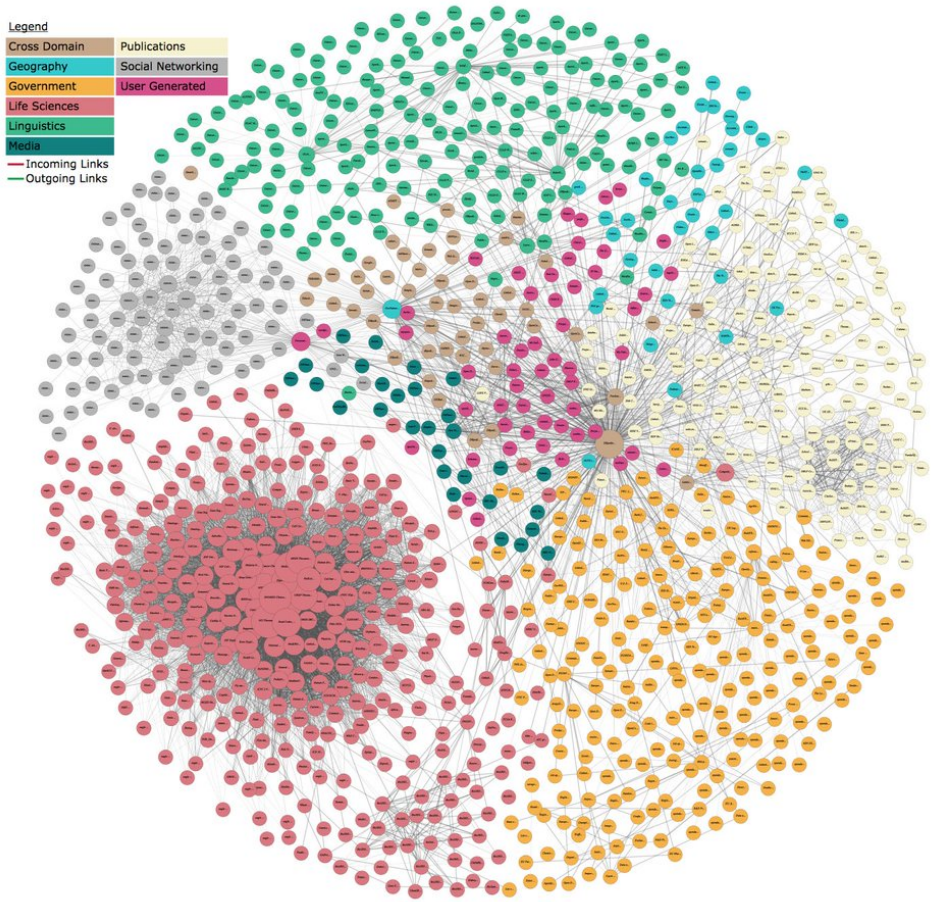
SPARQL - Query Language for RDF



Example

```
SELECT * WHERE { jwebsp:John  foaf:knows  ?friend }
```

LOD Cloud, August 2017



[1]

- Approx. 10k datasets with 150B facts
- Errors (parsing, false facts) in at least 70% of datasets
- $>$ 15% extraction errors in DBpedia

[1] Original Source: `http://lod-cloud.net`

The End