1. Information retrieval, Information extraction, & early birds

2. KDT or D vs Classical Science, including NLP

3. Various measures of correlation (why we build ERIK)

4. Using TROPES + CLEMENTINE or CHIC (or ERIK)

5. Construction of generality taxonomies: ASIUM and extensions

6. Examples of applications:
   - Grimm's Hausmärchen (in English)
   - Quebec's ecology review BISE (in French)
1. INFORMATION RETRIEVAL & INFORMATION EXTRACTION & EARLY BIRDS

MUCs (Message Understanding Conference) on Information Retrieval (IR) and Information Extraction (IE)

Information retrieval (IR)

IR : selecting the documents relevant to query

"given a bag of words, find the texts containing the same bag of words"

Some proposed solutions:

answer the query by giving the words appearing in the text searched "near" the original query words, and let the user click on the relevant ones, in order to refine the query.

answer at first the query by providing not the answers themselves, but clusters of answers characterized by a set of words. The user has to choose which cluster best represents his/her intentions.

Information extraction (IE)
IE: "Given a structural template of typed variables, how to instantiate it in a way that fits a text?"

Example from Grishman (1997); analyze the sentence:

"Sam Schwartz retired as executive vice president of the famous hot dog manufacturer, Hupplewhite Inc. He will be succeeded by Harry Himmelfarb"

IE proceeds in 4 steps.

Step 1. Finding the syntactic structure

<table>
<thead>
<tr>
<th></th>
<th>type: person</th>
<th>Sam Schwartz</th>
</tr>
</thead>
<tbody>
<tr>
<td>e1</td>
<td>name</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>type: position</th>
<th>executive vice</th>
</tr>
</thead>
<tbody>
<tr>
<td>e2</td>
<td>value</td>
<td>president</td>
</tr>
<tr>
<td>e3</td>
<td>company:</td>
<td>Hupplewhite Inc.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>type person</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>e4</td>
<td>name</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>type person</th>
<th>Harry Himmelfarb</th>
</tr>
</thead>
<tbody>
<tr>
<td>e5</td>
<td>name</td>
<td></td>
</tr>
</tbody>
</table>

Step 2: Matching scenarios: the text is matched against patterns; a "person" retires as "position"

<table>
<thead>
<tr>
<th></th>
<th>type: person</th>
<th>Sam Schwartz</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>name</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>type: position</th>
<th>executive vice</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>value</td>
<td>president</td>
</tr>
<tr>
<td>E</td>
<td>company:</td>
<td>Hupplewhite Inc.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>name</th>
<th></th>
</tr>
</thead>
</table>
Step 3: Coreference analysis

e1 type: person name Sam Schwartz
e2 type: position value executive vice president company: e3
e3 type: company name Hupplewhite Inc.
e6 type person name Harry Himmelfarb
e7 type leave-job person: e1 position e2
e8 type succeed person-1: e6 person-2: e1

Step 4: deduction

Use theorems to infer consequences of the scenario as it stands now.
The (typed) theorems needed here are:

\[
\text{start-job (X:person, Y:job)} \ :- \ \text{leave-job (Z:person, Y:job)}, \ 
\text{succeed (X:person, Z:person)}
\]
leave-job (X:person, Y:job) :- start-job (Z:person, Y:job), succeed (Z:person, X:person)

Using this knowledge yields:

e1 type: person name Sam Schwartz
e2 type: position value executive vice company: president e3
e3 type: company name Hupplewhite Inc.
e6 type person name Harry
e7 type leave-job person: e1 position e2
e8 type succeed person-1: e6 person-2: e1
e9 type start-job person: e6 position e2

Early bird(s)

e.g. Hogenraad et al., 1995

analysis of 18 of the texts issued after the 27 attacks by the "Fighting Communist Cells", Belgium 1984-1985.

uses:
French adaptation of Martingale's "Regressive Imagery Dictionary" ("Dictionaire d'Imagerie Régressive") composed of 3363 words and roots assigned to
29 categories of primary thinking, to
7 categories of emotions, and to
7 categories of secondary thinking.

a semantic atlas of emotion concepts which scores them along three dimensions: pleasantness, arousal, emotionality.

a semantic atlas of aggressiveness concepts, scoring them along four dimensions: intensity, institutional/interpersonal, physical/moral, acceptable/unacceptable.

The authors created a semantic atlas of terrorist's targets in 7 categories (political, military, military-industrial, Belgian/foreign, economic, State buildings, financial).

**among reported experiments**

1: cuts each text into 20 segments, produces 20 categories over the texts. Carries an analysis of the variation of the categories over the texts. It produces some non random patterns. *(e.g. aggressiveness shows quadratic variation within a text)*

2: looks for associations among categories, and finds associations among some factors and some **next** target types.
## 2. KDT OR D VS CLASSICAL SCIENCE, INCLUDING NLP

**Inductive reasoning:** how to create models from the data, NOT how to make a given model to fit with the data

<table>
<thead>
<tr>
<th>Classical Science (including Computational Linguistics)</th>
<th>KDD &amp; DM &amp; KDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on <strong>deduction</strong> i.e., confirms, does not invent</td>
<td>Based on <strong>induction</strong> i.e., confirms AND invents</td>
</tr>
<tr>
<td>knowledge should be <strong>proven</strong></td>
<td><strong>Understandable</strong> rather than proven</td>
</tr>
<tr>
<td>it can be experimentally validated: <strong>precision</strong></td>
<td>more <strong>useful</strong> than precise → invention of many measures of usefulness</td>
</tr>
<tr>
<td>Knowledge must be as <strong>universal</strong> as possible</td>
<td>more <strong>particular</strong> than universal</td>
</tr>
<tr>
<td>elegance = <strong>concision</strong> (e.g., MDL principle)</td>
<td>Elegance = <strong>adequation</strong> of the representation with user's model</td>
</tr>
<tr>
<td>Computational Linguistics, NLP</td>
<td>KDT</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----</td>
</tr>
<tr>
<td>IR, IE, &quot;TM&quot;?</td>
<td>&quot;TM&quot;?</td>
</tr>
<tr>
<td>A deductive process</td>
<td>Inductive process (some inductive approaches to information retrieval belong also to KDT)</td>
</tr>
<tr>
<td>Looks for a theory of language</td>
<td>Looks for particular relationships in a given corpus</td>
</tr>
<tr>
<td>Knowledge is precise, universal etc.</td>
<td>Knowledge is understandable and useful</td>
</tr>
<tr>
<td>Looks for explicit or implicit relations stated in one or several texts</td>
<td>Invents &quot;interesting&quot; relations among (a large number of) texts. These relation are never stated.</td>
</tr>
</tbody>
</table>

**Example 1**

NLP is able to find:

There exists a text (or a set of texts) such as

(speak-much-of) Miss Lewinsky &
(speak-little-of) Mrs. Clinton

There exists another text such as

(speak-a-lot-of) Mrs. Clinton &
(do-not-speak-of) Miss Lewinsky
With a hypothetical taxonomy of concepts, one can even find:

There exists a text such as

\[(\text{speak-a-lot-of}) \, \text{relation-near-of-president} \, \& \, (\text{do-not-speak-of}) \, \text{relation-unstable-to-president}\]

They are observations about texts

KDT:

There exists a corpus such as the following theorems are true up so-much %

\[(\text{speak-much-of}) \, \text{Miss Lewinsky} \Rightarrow (\text{speak-little-of}) \, \text{of Mrs. Clinton} \, \text{ (True up to } x \, \%)\]

\[(\text{speak-much-of}) \, \text{Mrs. Clinton} \Rightarrow (\text{do-not-speak-of}) \, \text{Miss Lewinsky} \, \text{ (True up to } x' \, \%)\]

With the same taxonomy of concepts:

\[(\text{speak-much-of}) \, \text{relation-near-of-president} \Rightarrow (\text{do-not-speak-of}) \, \text{relation-unstable-to-president} \, \text{ (True up to } y \, \%)\]

This is a construction of a model of the corpus
how to justify the implications?
(in NLP : the presence of the facts justifies to notice them)
the example of Miss Lewinsky: interesting?
(Introduction of an utility measure in the implication detection system)

Example 2

topic: to support Mrs. Clinton's campaign

NLP: to analyze her speeches in order to find their most important themes

KDT: find ALSO non-spoken-of politically important themes: the definition of that that is an important "theme" is included in the system of discovery

What non-spoken-of topics are statistically significant?

For instance:

(speak-much-of) racial-problems ⇒ (do-not-speak-of) economy (True up to z %)
3. VARIOUS MEASURES OF CORRELATION (WHY WE BUILD ERIK)

A and B : two assertions, 
X and Y their supports

\{X\} = \{x / A(x) = True\} and \{Y\} = \{x / B(x) = True\}

we deal with a finite set of examples:
\{X\} \cup \{ \overline{X} \} = \{Y\} \cup \{ \overline{Y} \} = \{Tot\}

we want to infer A ⇒ B from the data

An example

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>¬A</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>¬B</td>
<td>2</td>
<td>14</td>
</tr>
</tbody>
</table>

D1:
P(A, B) = 1/10,
P(¬A, ¬B) = 7/10
P(¬B, A) = 1/10
P(A) = 2/10, P(B) = 2/10
P(¬A) = 8/10, P(¬B) = 8/10
\begin{align*}
P(B \mid A) &= 1/2, \\
P(\neg A \mid \neg B) &= 7/8 \\
P(\neg B \mid A) &= 1/2 \\
P(A \mid \neg B) &= 1/4 \\
\text{support} &= P(A, B) = 1/10 \\
\text{confidence} &= P(B \mid A) = 5/10 \\
\text{Dependency} &= P(B \mid A) - P(B) = 3/10
\end{align*}

In order to measure the validity of \( A \Rightarrow B \)
most measures are based on the support of \( A \Rightarrow B \), the highest the
support, the highest the measure

(so-called "ASSOCIATION DETECTION")

**Support** : \( P(A, B) \) (sometimes \( P(A) \))

\[
P(A, B) = \frac{|\{X\} \cap \{Y\}|}{|\{Tot\}|}
\]

**Confidence** : \( P(A, B) / P(A) = P(B \mid A) \)

**Dependency** : ( = correlation for Boolean variables)

\[
\text{Abs}(P(B \mid A) - P(B)) \text{ if } P(A) \neq 0
\]

The value \( \text{Abs}(P(B \mid A) - P(B)) \) gives the strength of the dependency.

**Correlation coefficient** (for continuous variables)

Variables \( x \) and \( y \) taking real values \( x_i \) and \( y_i \) for each of
the \( N \) items of the data base

\[
M_x = \text{mean of } x, \ E_x = \text{variance}
\]
Correlation(x, y) = 
\[ 1 / (N-1) \sum_{i=1}^{N} \left( \frac{x_i - M_x}{E_x} \right) \left( \frac{y_i - M_y}{E_y} \right) \]

In order to measure the validity of \( A \Rightarrow B \)
the Intensity of implication is
based on the support of \( A \Rightarrow \neg B \), the lowest the support, the
highest the measure

A = T & Y = T
A = T & B = F
B = T

U and V random,
card U = card (A = T)
card V = card (B = F)

\[ P(\text{card}(U \cap \neg V) ?<<? \text{card}(A = T & B = F)) \]

Present measures tend to use a Poisson approximation:
\[ I = 1/\sqrt{2\pi} \int_{-\infty}^{\infty} e^{-t^2/2} \, dt \]
\[ \begin{align*}
\text{n}_{ab'} &= | \{X\} \cap \{ \overline{Y} \} | , \\
\text{n}_a &= | \{X\} | , \\
\text{n}_{b'} &= | \{ \overline{Y} \} | , \\
\text{n} &= \{\text{Tot}\} \\
ii &= (\text{n}_{ab'} - \text{n}_a \text{n}_{b'}/n) / \sqrt{\text{n}_a \text{n}_{b'}/n}.
\end{align*} \]

**MOST SOFTWARE MEASURE THE SUPPORT AT FIRST** →
**COMPLEXITY DECREASED BY FREQENT ITEM-SETS ANALYSIS**

**CHIC**: A SOFTWARE THAT MEASURES ONLY THE INTENSITY OF IMPLICATION

**ERIK**: MEASURES BOTH →
**COMPLEXITY DECREASED INTENSITY OF IMPLICATION ANALYSIS**
4. **Using TROPES + Clementine or Chic**  
(Or Erik)

how to extract knowledge from a large text corpus, using the NLP system TROPES (Acetic) and the KDD system CLEMENTINE (developed by ISL and belonging to SPSS) and CHIC (Univ. Nantes)

Data preparation

Using the NLP system: preparing the tables

Using the KDD system
Using the NLP system

1. CREATE THE "DOCUMENTARY BASE"

upper window shows the concatenated texts with all instances of "télécommunication" underlined
Result window (displays various results of the analysis)

The upper part shows the expressions
- « Style général du texte » that provides linguistic information on the texts.
- « références utilisées » which is the list of the terms used in the texts.
- « Univers de référence 2 et 1 » they are the 1\textsuperscript{st} and 2\textsuperscript{nd} level of a taxonomy of terms given in Tropes.
- « scénario » allows to install your own taxonomy

middle part: list of the terms used in the texts

lower part: says that « opérateur télécom » is a term which is also a concept
2. BUILD THE "SCENARIO" CLUSTERING THE BEST REPRESENTATIVES OF THE CONCEPTS OF INTEREST

chose which concepts will be included in the tables

the term « ola » is made to belong to the class « abonnement GSM » (my decision)

Allemagne (Germany) is an existing concept I want to see appear in the probability tables as defined by TROPES
3. GENERATE THE TABLE OF THE PROBABILITIES OF OCCURRENCE OF EACH CONCEPT IN THE TEXTS

The table of frequency provided by Tropes

Each number is computed by: (# of terms belonging to the concerned concept / total # of terms).
Using the KDD system

Clementine's process stream
Use visualization techniques

"Freedom is linked to expenses"

Association graph of some concepts
"Mr. X is linked to monopolization"

Use inductive learning techniques

Result of Apriori (cover > 80%, accuracy > 98%)
Result of *Apriori* (*cover > 10%, accuracy > 40%, $1 - \frac{P(A, B)}{P(B)} > 50\%$)

use C5 available in Clementine to select features that are in causal relationship with concept1 (they cause concept1 or they are caused by it)

C5 concluding on « accord »
GRI concluding on « accord » where only input fields are those playing a role in the decision tree built by C5.
5. CONSTRUCTION OF GENERALITY TAXONOMIES: ASIUM AND EXTENSIONS

General "well-known" problem: infer semantic properties from syntactic ones

e.g. : "all nouns belonging to the same sub-categorization frame (i.e., they have the same relationship with a given verb) belong also to the same concept"

PARTLY (only) FALSE

Build on our experience in interactive learning to build a semi-automatic system to do ASIUM
(important: ask as few as possible questions to the user)

ASIUM at present: relies on sub-categorization frames

ASIUM+: other relations as well
e.g. : one output of ROWAN = all kinds of syntactical relationships

<table>
<thead>
<tr>
<th>Grimm tales:</th>
<th>Bise review:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total by Relation</strong></td>
<td><strong>Total by Relation</strong></td>
</tr>
<tr>
<td><strong>Role1:Role2[:Role3]</strong></td>
<td><strong>Role1:Role2[:Role3]</strong></td>
</tr>
<tr>
<td><strong>Frequency</strong></td>
<td><strong>Frequency</strong></td>
</tr>
<tr>
<td>Subject:Verb : 39289</td>
<td>Subject:Verb : 16890</td>
</tr>
<tr>
<td>Verb:Subject : 2506</td>
<td>Verb:Subject : 239</td>
</tr>
<tr>
<td>Verb:Object : 14329</td>
<td>Verb:Object : 7682</td>
</tr>
<tr>
<td>Object:Verb : 1303</td>
<td>Object:Verb : 2958</td>
</tr>
<tr>
<td>Verb:Preposition:Agent : 180</td>
<td>Verb:Preposition:Agent : 897</td>
</tr>
<tr>
<td>Adverb:Verb : 4413</td>
<td>Adverb:Verb : 1548</td>
</tr>
<tr>
<td>Verb:Adverb : 3588</td>
<td>Verb:Adverb : 448</td>
</tr>
<tr>
<td>Noun:Adjective : 2564</td>
<td>Noun:Adjective : 24352</td>
</tr>
<tr>
<td>Adjective:Noun : 8985</td>
<td>Adjective:Noun : 2497</td>
</tr>
<tr>
<td>Adverb:Adverb : 1775</td>
<td>Adverb:Adverb : 942</td>
</tr>
<tr>
<td>Noun:Noun : 2716</td>
<td>Noun:Noun : 4279</td>
</tr>
</tbody>
</table>
Use a shallow parser (SYLEX due to Patrick Constant - XEROX Grenoble - new parsers under study: English, German, Spanish) for building taxonomies of concepts.

TO POACH AN EGG
IN A LIQUID
IN A CONTAINER

POACH an EGG IN WATER IN A MUG
A child can have several parents.

Progressive and interactive creation of the taxonomy

Generation of examples to be validated by the user (e.g., is it possible "to dip" a container?)

\[
\text{dist}_{\text{Log}}(C_1,C_2) = \frac{\log(\sum FC_1) + \log(\sum FC_2)}{\log(\sum_{i=1}^{\text{card}(C_1)} f(\text{nom}_i C_1)) + \log(\sum_{i=1}^{\text{card}(C_2)} f(\text{nom}_i C_2))}
\]

\(FC_1\) = fréquences des noms de \(C_1\) communs avec \(C_2\)

\(FC_2\) = fréquences des noms de \(C_2\) communs avec \(C_1\)

The distance between two concepts. They will be merged into a new concept if their distance is small enough.
### A.S.I.U.M

#### Name of the new cluster?

<table>
<thead>
<tr>
<th>Source class 1</th>
<th>Source class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>vinegar</td>
</tr>
<tr>
<td>4</td>
<td>water</td>
</tr>
<tr>
<td>7</td>
<td>wine</td>
</tr>
<tr>
<td>19</td>
<td>liquid</td>
</tr>
<tr>
<td>3</td>
<td>impurity</td>
</tr>
<tr>
<td>6</td>
<td>sauce</td>
</tr>
<tr>
<td>2</td>
<td>boulion</td>
</tr>
<tr>
<td>1</td>
<td>wine</td>
</tr>
<tr>
<td>21</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

- **Similarity:** 57.852338  
- **# of clusters:** 26  
- **Ontology level:** 0

- **Delete Word**
- **Add to noise**
- **Orthographic correction**
- **Increase frequency**

**Accept this cluster**

**Reject this cluster**

**Newly covered examples**

**Display configuration**

**All examples**

**Save**

**Split this cluster**

**See Ontology**

**Reject next clusters**

**Accept next clusters**
SUMMARY: SYSTEMS WE ARE USING

THE PRESENT "ARCHITECTURE"

THE FUTURE ARCHITECTURE
6. EXAMPLES OF APPLICATIONS:
- GRIMM'S HAUSMÄRCHEN

Highly probable links within the set of Grimm tales

- **nof** (or not-of) = none of the concept appears in a given text
- **sof** = some of the concept appears in a given text
- **muof** (or much-of) of the concept appears in a given text
Subgraph of the relations of concept "death" with other concepts.

- Quebec's Ecology Review BISE

Relations with a 100% intensity of the implication
Relations between the 3 values of "cancer" and other concepts
Using exhaustive search techniques issued from ILP

For each rule:  true => false : 0

<table>
<thead>
<tr>
<th>Rule</th>
<th>GRIMM</th>
<th>BISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF (speaks of death) &amp; (speaks of hunt) THEN (speaks of insides)</td>
<td>. dependance : 0.445</td>
<td>IF (does not speak of cancer) &amp; (not much of control) THEN (does not speak of control)</td>
</tr>
<tr>
<td></td>
<td>. true =&gt; true : 21</td>
<td>. dependance : 0.426</td>
</tr>
<tr>
<td></td>
<td>. false =&gt; false : 89</td>
<td>. true =&gt; true : 23</td>
</tr>
<tr>
<td></td>
<td>. false =&gt; true : 90</td>
<td>. false =&gt; false : 75</td>
</tr>
<tr>
<td>IF (speaks of girl) &amp; (much of servant) THEN (not much of feeling)</td>
<td>. false =&gt; true : 106</td>
<td>. false =&gt; true : 78</td>
</tr>
<tr>
<td></td>
<td>. dependance : 0.41</td>
<td>IF (does not speak of alim) &amp; (does not speak of cancer) THEN (does not speak of insid_bod)</td>
</tr>
<tr>
<td></td>
<td>. true =&gt; true : 12</td>
<td>. dependance : 0.477</td>
</tr>
<tr>
<td></td>
<td>. false =&gt; false : 82</td>
<td>. true =&gt; true : 18</td>
</tr>
<tr>
<td></td>
<td>. false =&gt; true : 106</td>
<td>. false =&gt; false : 84</td>
</tr>
<tr>
<td>IF (speaks of servant) &amp; (much of tree) THEN (speaks of bird)</td>
<td></td>
<td>. false =&gt; true : 74</td>
</tr>
<tr>
<td></td>
<td>. dependance : 0.49</td>
<td>IF (much of comp_sci) &amp; (does not speak of insid_bod) THEN (not much of cancer)</td>
</tr>
<tr>
<td></td>
<td>. true =&gt; true : 9</td>
<td>. dependance : 0.454</td>
</tr>
<tr>
<td></td>
<td>. false =&gt; false : 98</td>
<td>. true =&gt; true : 11</td>
</tr>
<tr>
<td></td>
<td>. false =&gt; true : 93</td>
<td></td>
</tr>
</tbody>
</table>
IF (much of hunt) & (speaks of money) THEN (speaks of death)
  . dependance : 0.525
  . true => true : 8
  . false => false : 105
  . false => true : 87

IF (speaks of castle) & (much of occultism) THEN (speaks of death)
  . dependance : 0.525
  . true => true : 7
  . false => false : 105
  . false => true : 88

IF (speaks of castle) & (much of hunt) THEN (speaks of death)
  . dependance : 0.525
  . true => true : 7
  . false => false : 105
  . false => true : 88

IF (much of death) & (much of mythology) THEN (much of tree)
  . false => false : 80
  . false => true : 85
  . true => false : 0

IF (speaks of air) & (much of comp_sci) THEN (not much of cancer)
  . dependance : 0.454
  . true => true : 10
  . false => false : 80
  . false => true : 86

IF (does not speak of contagion) & (much of intoxic) THEN (much of cancer)
  . dependance : 0.545
  . true => true : 2
  . false => false : 96
  . false => true : 78
<table>
<thead>
<tr>
<th>Dependence</th>
<th>True =&gt; True</th>
<th>False =&gt; False</th>
<th>False =&gt; True</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.805</td>
<td>3</td>
<td>161</td>
<td>36</td>
</tr>
</tbody>
</table>