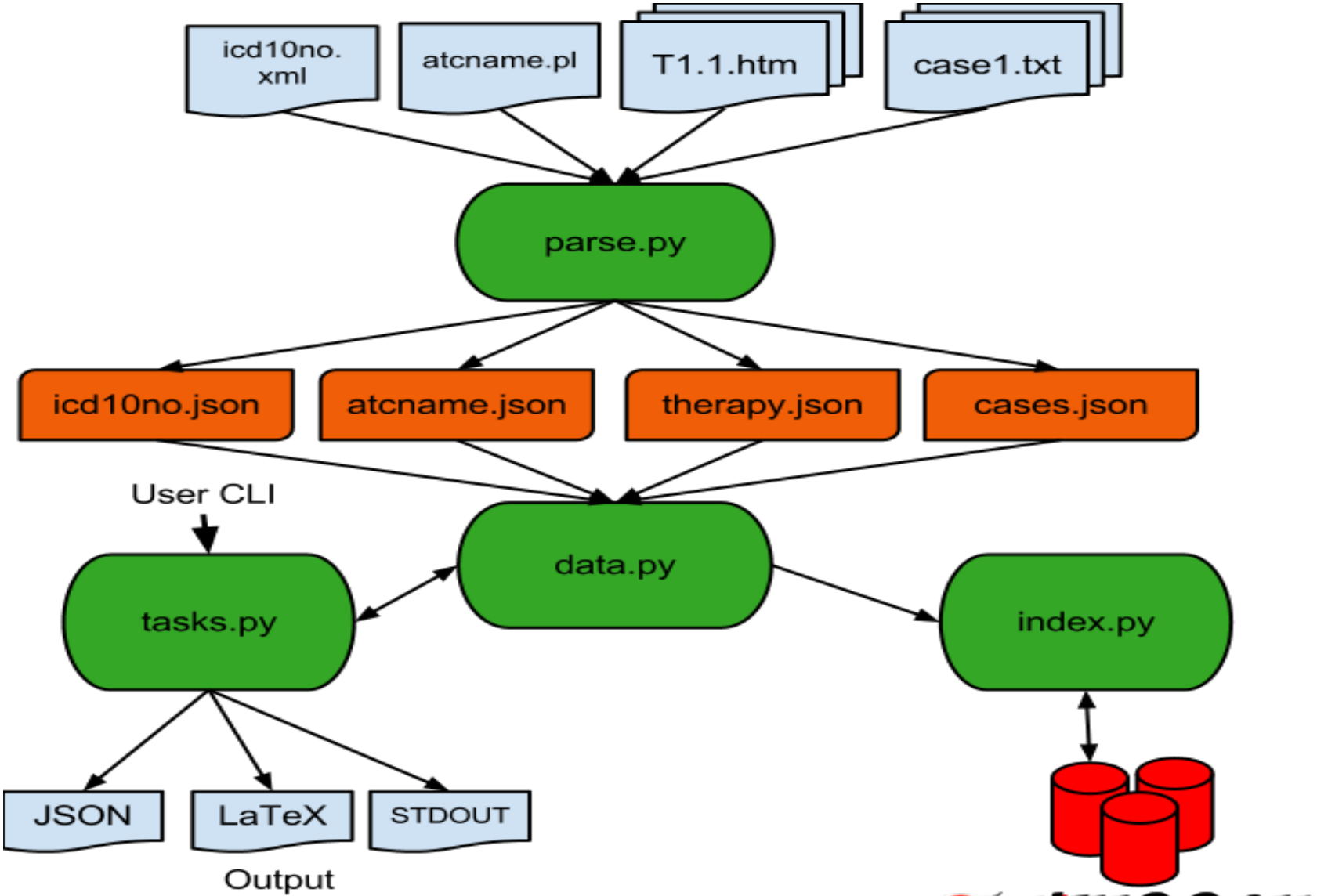
TDT4215: Web-intelligence
Group 1 project presentation

Group 1
Terje Snarby
Even Wiik Thomassen
Weilin Wang

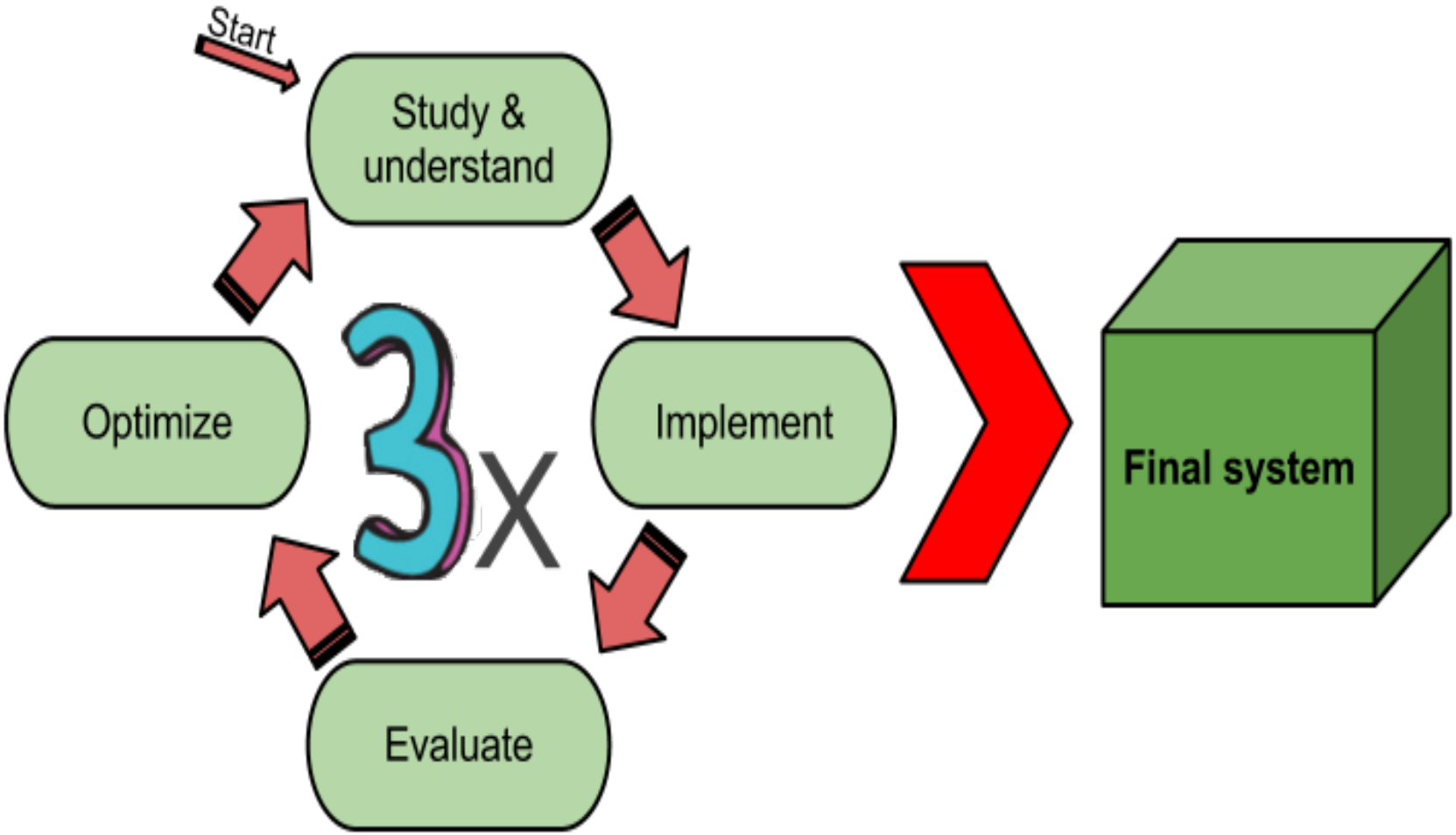
NTNU
Norwegian University of Science and Technology

April 26, 2012

System architecture



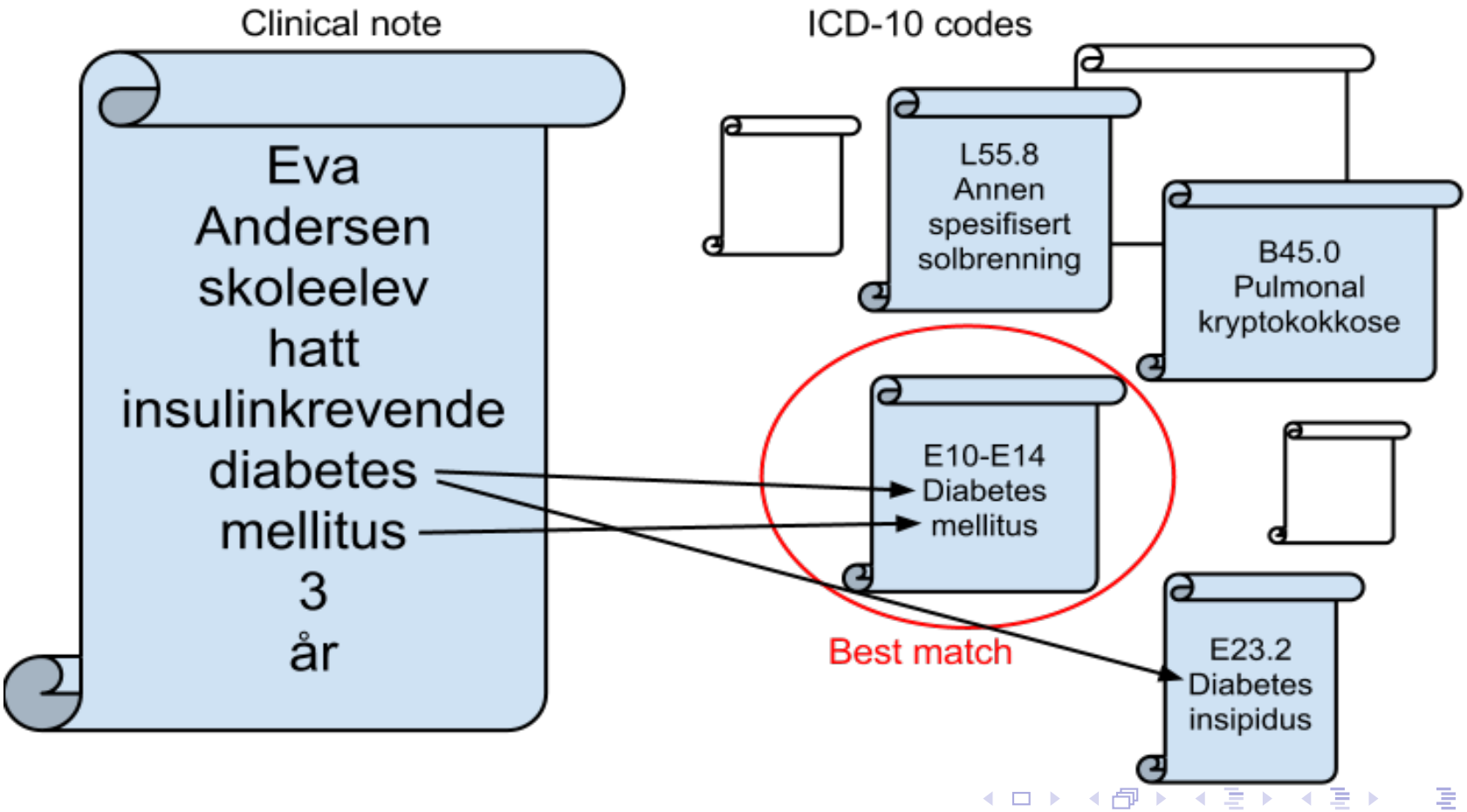
Workflow



Iterations:
1. Task 1, 2
2. Task 3
3. Task 6

Autocoding ICD-10 and ATC codes

- Bag-of-words scoring algorithm
- BM25F



Clinical notes and therapy chapter similarities

Algorithm Vector space model based on TF-IDF

Method A log normalization and inverse frequency

Method B log normalization and probabilistic inverse frequency

Method C raw frequency and inverse frequency

Method D raw frequency and probabilistic inverse frequency

How we improved the ranking

- Aggregation and weighting of results
- Hierarchical nature of ICD-10 and ATC codes
- Hierarchical structure of therapy chapters

Improvement based on result evaluation

- Automatic evaluation of the whole system
- Feedback for optimizing classification algorithm

Case	P@10	R-precision
1	70%	0.57
2	70%	0.86
3	100%	1.00
4	90%	0.89
5	90%	0.89
6	90%	0.89
7	90%	1.00
8	80%	0.88
Avg	85%	0.87

Our results versus gold standard

Case	Rank	Relevant chapter	Gold standard
1	1	T3.1 Diabetes mellitus	Yes
2	1	T10.2 Obstruktiv lungesykdom	Yes
	2	T3.1 Diabetes mellitus	No
	3	T10.2.2 Kronisk obstruktiv	Yes
3	1	T1.10 Akutt bakteriell meningitt	Yes
4	1	T8.3 Koronarsykdom	Yes
	2	T8.3.1 Stabil koronarsykdom	Yes
	3	T8 Hjerte- og karsykdommer	No
	4	T8.3.2 Ustabil koronarsykdom	Yes

Discussion on our classification algorithm

Strengths

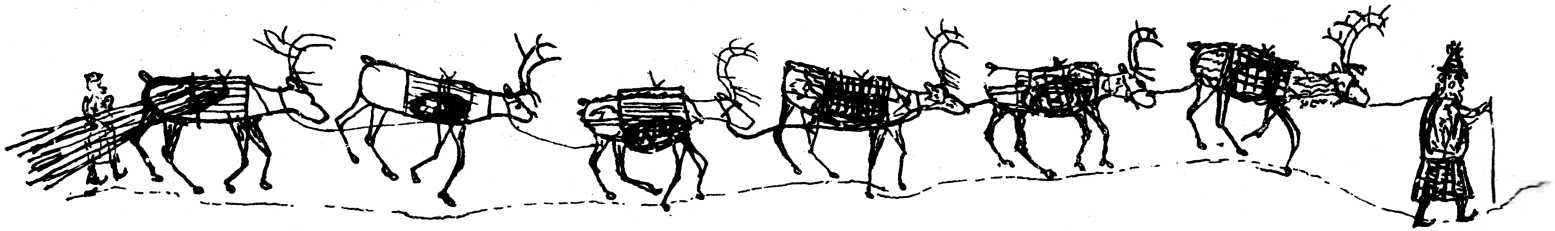
- Great results
- Multiple sources, more accurate results
- No training necessary

Weaknesses

- Domain knowledge needed for best possible utilization of algorithms.
- No feedback/learning.
- Consider asking a botanist: Is an object a tree?

Summary

1. Implemented a patient case search system
2. Used vector based model
3. Used automatic evaluation
4. Great results in regards to gold standard



Questions?