EVALUATING CLASSIFICATION POWER OF LINKED ADMISSION DATA SOURCES WITH TEXT MINING

Simon Kocbek, Lawrence Cavedon, David Martinez, Christopher Bain, Chris Mac Manus, Gholamreza Haffari, Ingrid Zukerman, Karin Verspoor

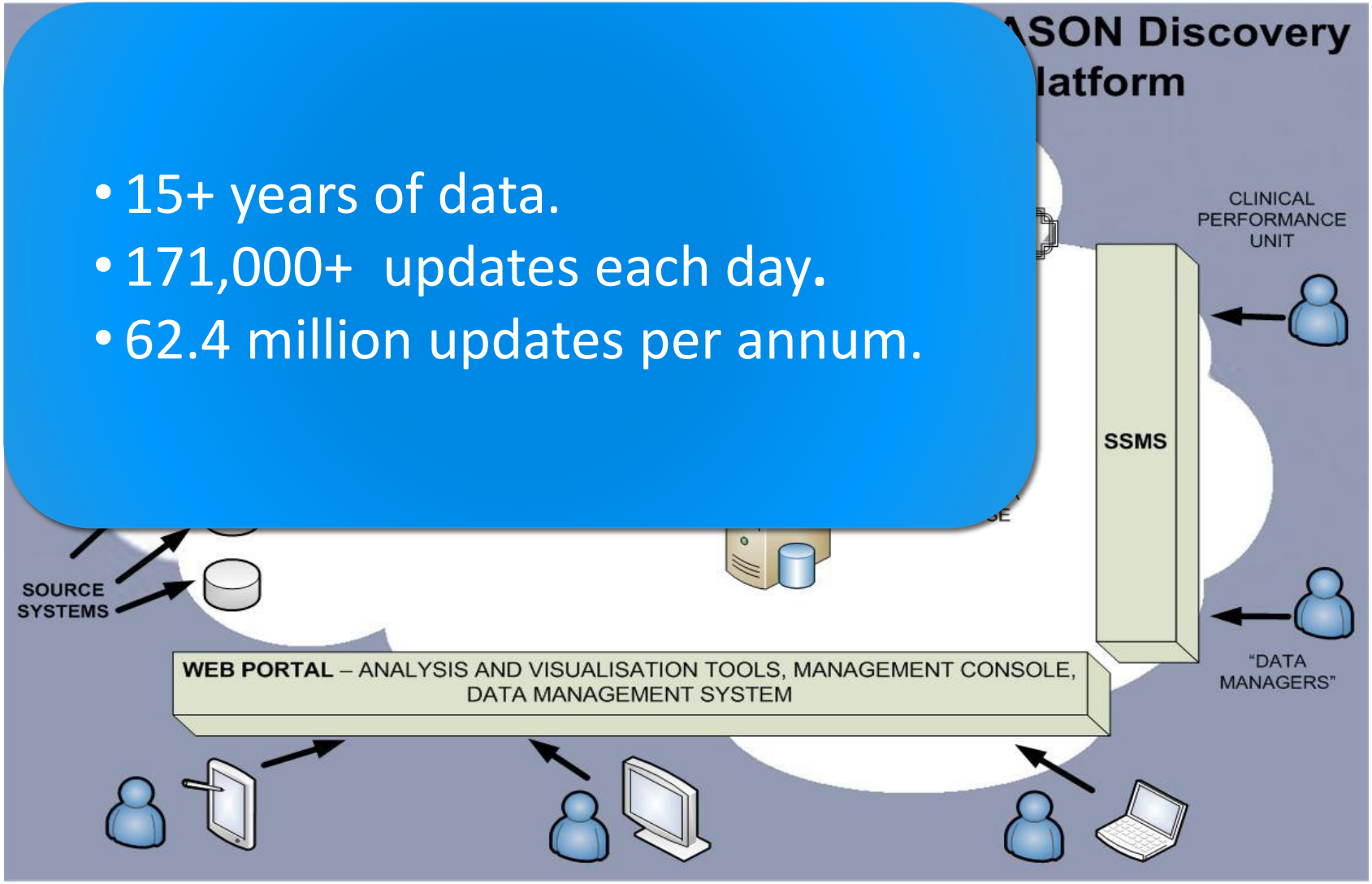
Background and motivation

- Growing Electronic Health Records (EHR) data.
  - Much of it in free text format.
  - This text can be used in text mining applications.
- Most previous TM applications use a single textual data source.
- Increase in data linkage in hospitals allows multiple sources to be leveraged for complex analytical tasks.

- We describe a text mining system that detects positive cases of **lung cancer** for each admission:
  - Use of **multiple** data sources.
  - Evaluate performance (does performance improve?).

Alfred REASON platform

- 15+ years of data.
- 171,000+ updates each day.
- 62.4 million updates per annum.



Task

Radiology question

Admission

50yo complaining of left shoulder pain. Tender generally. Difficulty abducting the shoulder past 45 degrees. Home on HITH tomorrow - either inpatient or outpatient please

Radiology report

Mobile Chest performed on 02-JUN-2012 at 08:27 AM: The nasogastric tube has its tip in the stomach. The tracheostomy is seen at T2 level. ….

Additional data

Age: 50
Date of admission: Jun/12
Gender: F
Country: …

ICD-10 code

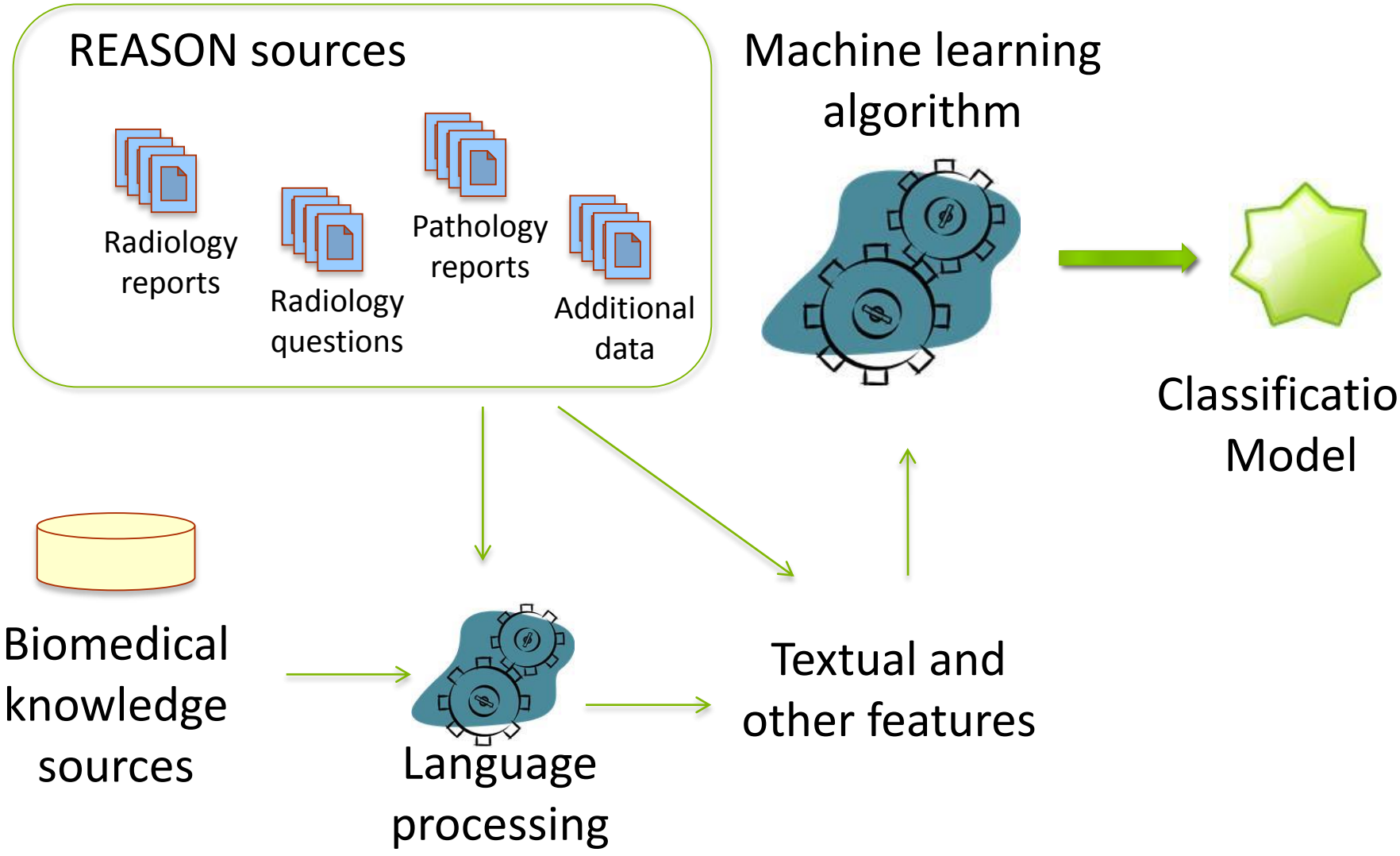
Pathology report

Urine Culture
Acc No: 12-183-0731Source: Urine
---------- URINE MICROSCOPY (PHASE CONTRAST) ---------- Leucocytes x10^6/L (Ref <10).... <10
Erythrocytes x10^6/L (Ref <10).. <10.....

Data – Characteristics

- Extracted data for 2 financial years from 2012 to 2014:
  - 150,521 admissions,
  - 40,800 radiology reports with associated question,
  - 20,872 pathology reports,
  - 121,700 additional data entries (demographics, hospital admission info).
- Admissions are associated to ICD-10 codes:
  - Used as ground truth.
  - ICD-10 code C34.* to identify positive cases for lung cancer.
  - 496 positive admissions.
- Final dataset:
  - Subsampling.
  - 992 admissions.

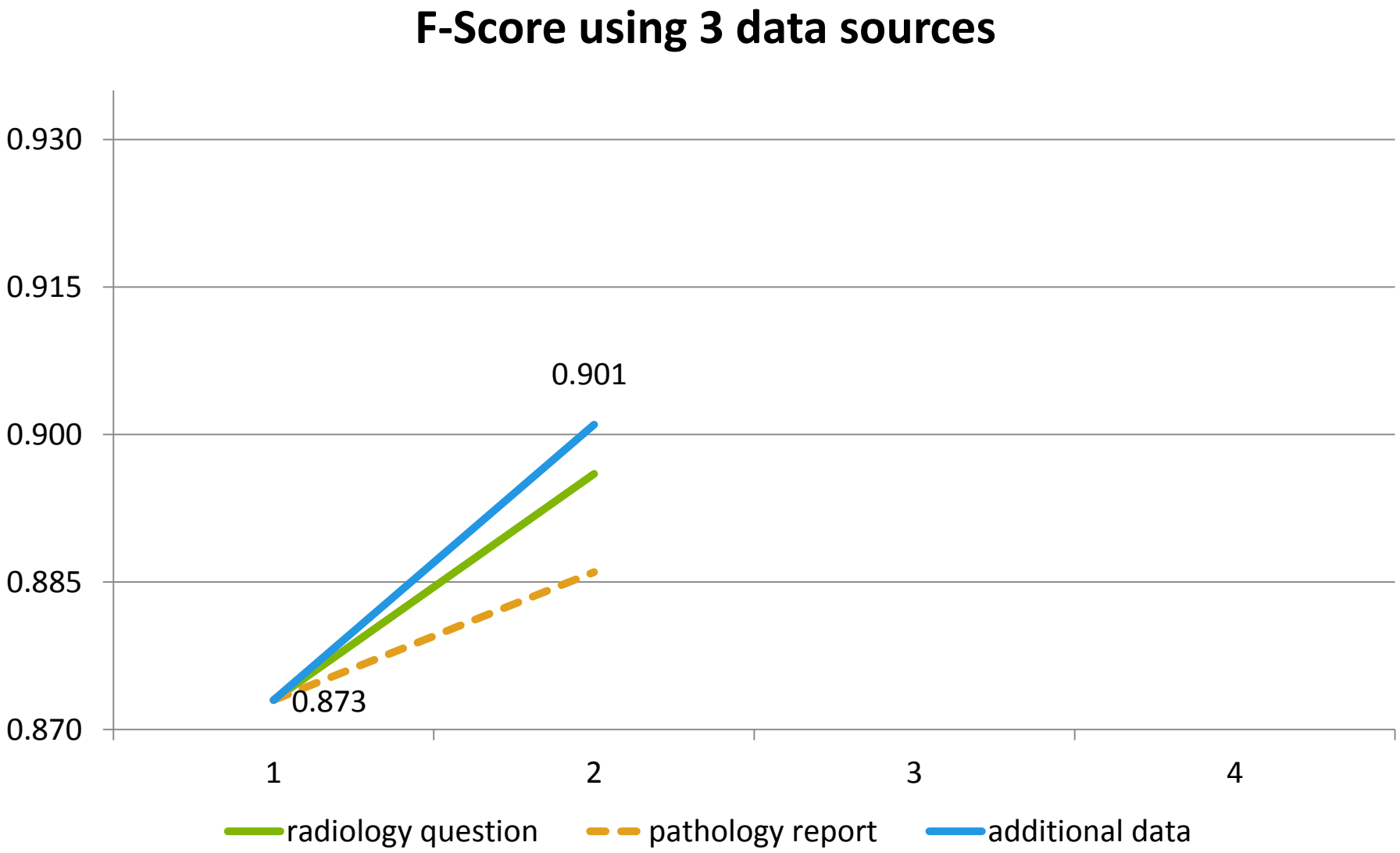
Methods (I)


REASON sources
Radiology reports
Radiology questions
Pathology reports
Additional data
Machine learning algorithm
Classification Model
Biomedical knowledge sources
Language processing
Textual and other features

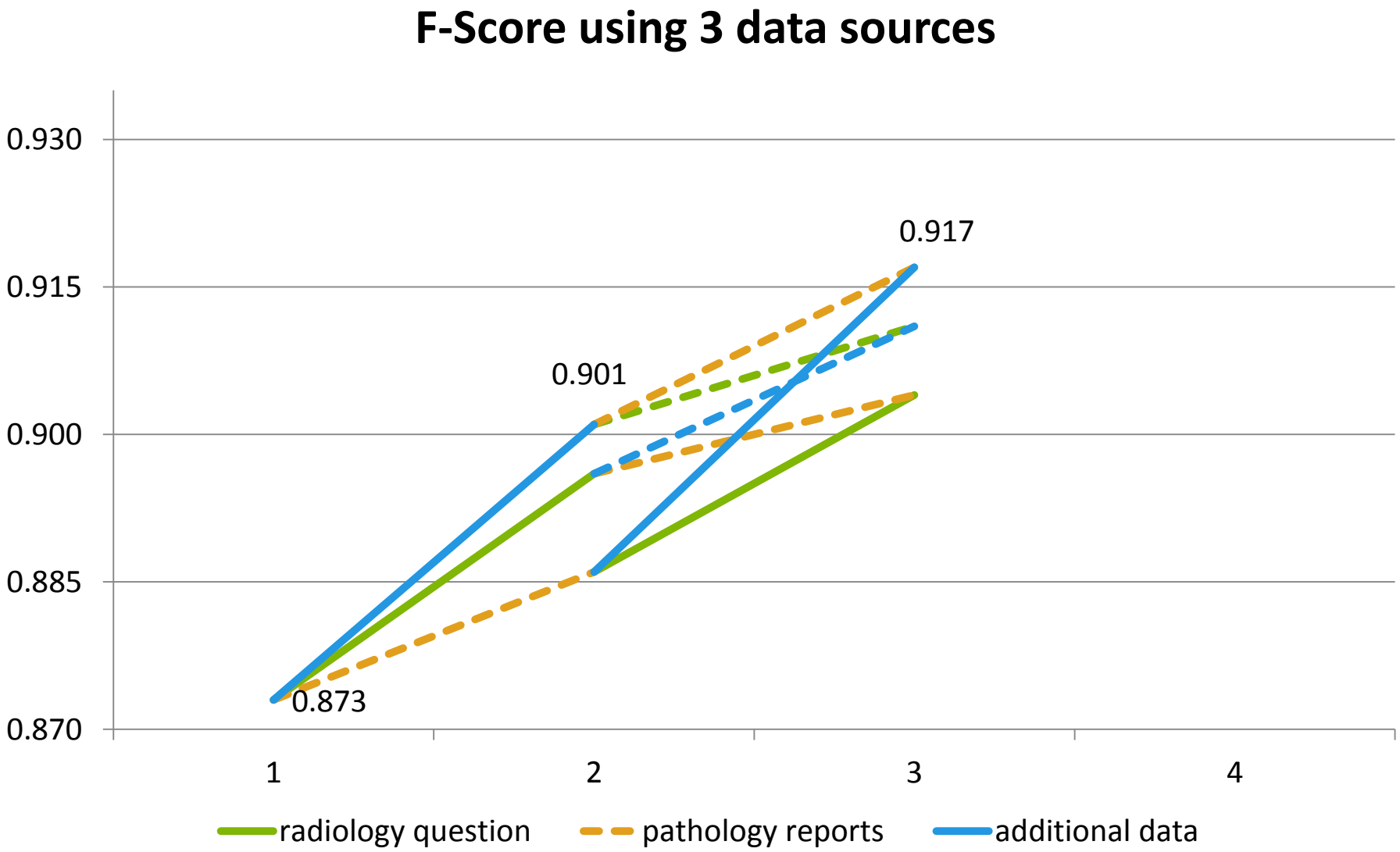
Methods (II)

- Features:
  - Biomedical phrases.
  - Identified negative context ("**no** lung cancer" vs "lung cancer").
  - Ambiguous words ("common **cold**" vs "**cold** temperature").
- Machine learning algorithms
  - Support Vector Machines.
  - Parameter tuning.
  - Evaluation: Precision, Recall, F-Score.
  - Statistical significance.
- Steps:
  - 8 different classification models (different combinations of data sources).
  - Baseline: phrases from only radiology reports.
  - Adding phrases from other sources.

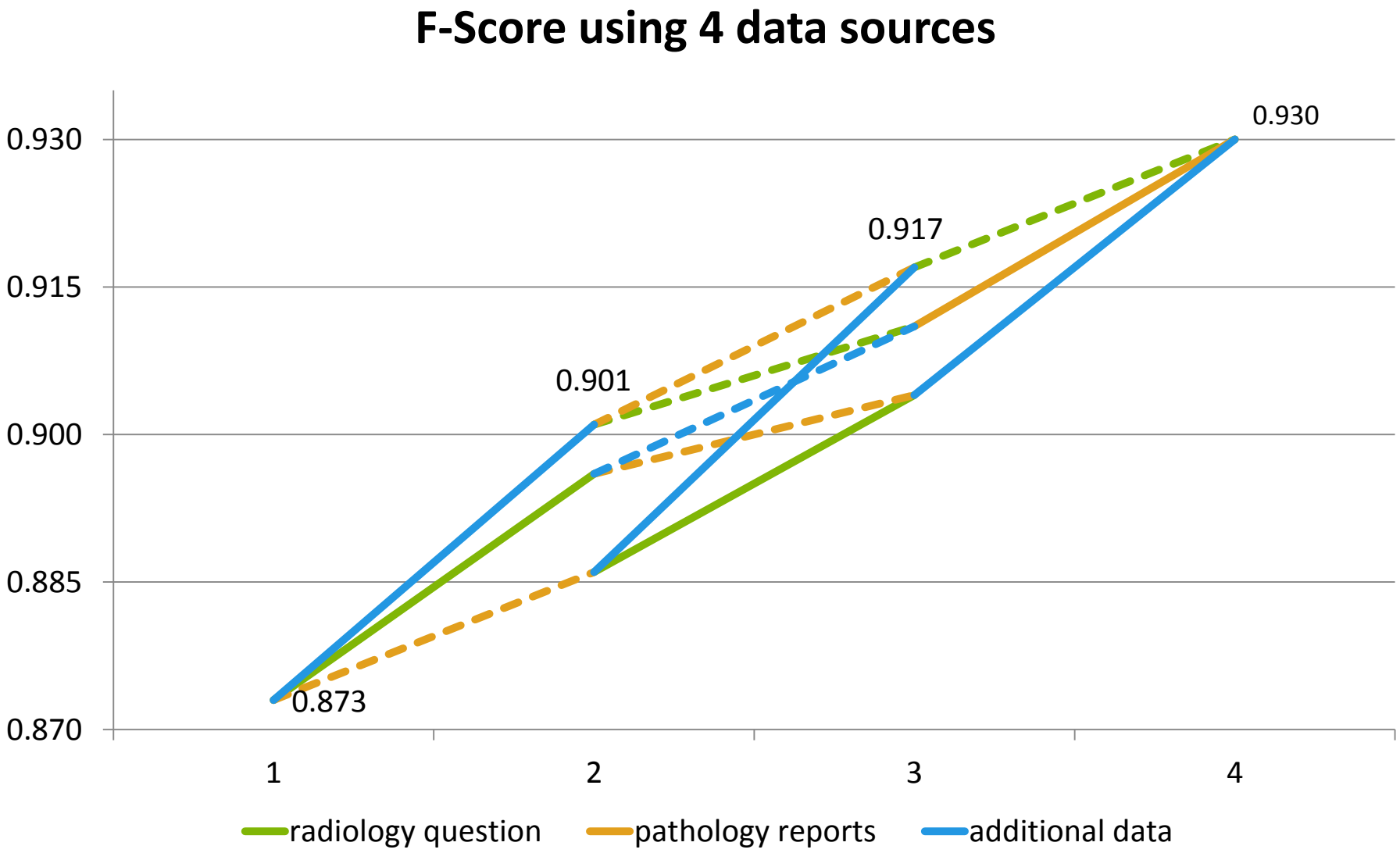
Results


F-Score using 3 data sources
0.930
0.915
0.900
0.885
0.870
0.901
0.873
1
2
3
4
radiology question
pathology report
additional data

Results


F-Score using 3 data sources
0.930
0.915
0.900
0.885
0.870
0.917
0.901
0.873
1
2
3
4
radiology question
pathology reports
additional data

Results

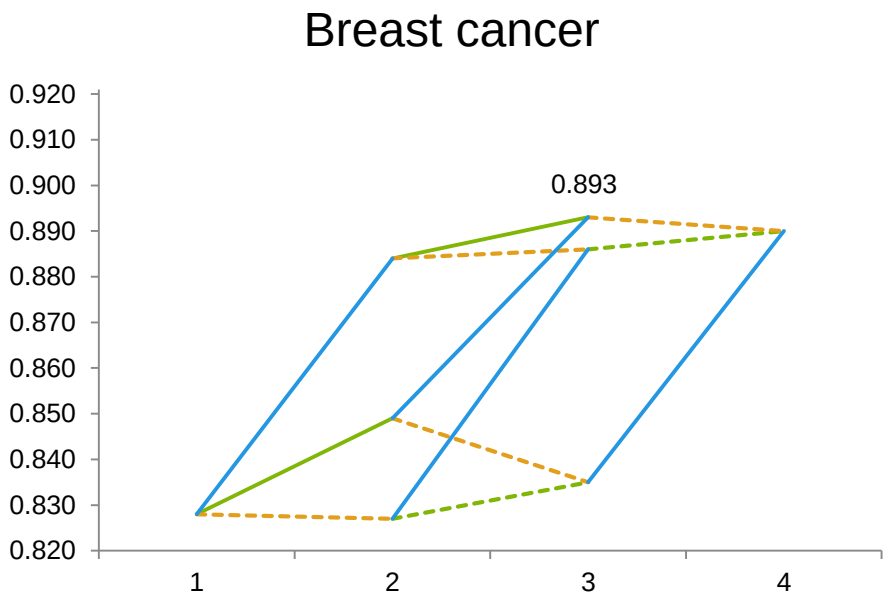

F-Score using 4 data sources
0.930
0.915
0.900
0.885
0.870
0.930
0.917
0.901
0.873
1
2
3
4
radiology question
pathology reports
additional data

Discussion

- More data sources lead to better performance.
- The classifier with the highest performance was built using features from all four data sources.
- Not all improvements were significant:
  - Radiology question and metadata vs
  - Pathology reports.
    - Not all admissions had a pathology report associated with them.

Conclusion

- We built a text mining system for detecting lung cancer admissions.
- Our methods show more informed systems can be built by including multiple linked data sources.
- Future work:
  - Other diseases.
  - Skewed datasets.
  - Feature selection.



Thank you

- Questions?
- Comments?