Applying an Interactive Machine Learning Approach to Statutory Analysis

Jaromir Savelka[a,b], Gaurav Trivedi[a], Kevin D. Ashley[a,b,c]

[a] *Intelligent Systems Program, University of Pittsburgh, USA*

[b] *Learning Research and Development Center, University of Pittsburgh, USA*

[c] *School of Law, University of Pittsburgh, USA*

This work was supported by the University of Pittsburgh's University Research Council Multidisciplinary Small Grant Program and in part by the Cooperative Agreement 5P01TP000304 from the Centers for Disease Control and Prevention. Its contents are the responsibility of the authors and do not represent the official views of the CDC. The development of the tool and the second author were supported by NIH grant R01LM010964. We also acknowledge the guidance from Harry Hochheiser towards the tool development and helpful comments provided by Matthias Grabmair. We thank anonymous reviewers for their valuable suggestions.

Outline

- Background
- Motivation
- Task
- Related Work
- Example Application
- Experiments
- Future Work

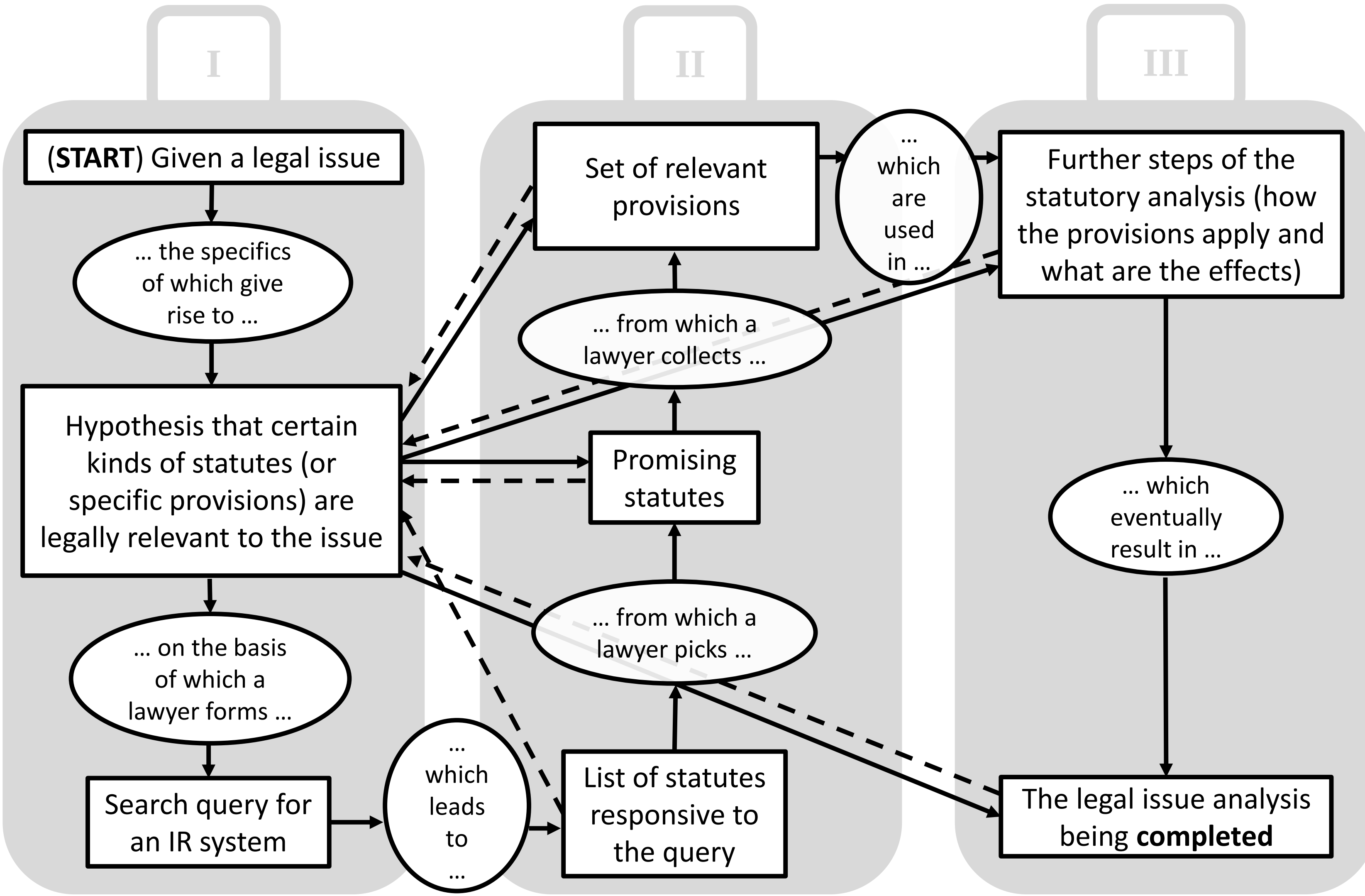
Statutory Analysis: Definition

Statutory analysis is the process of determining **if a statute applies**, how it applies, and the effect of this application. (Putnam 2008)

Example

Assess Pennsylvania's regulatory framework concerning preparedness and response of the public health system to public health emergencies.

Statutory Analysis: Process



Statutory Analysis: Examples

In our work we focus on **applicability assessment** in situations where there are **many relevant provisions**, e.g.,

- total regulatory compliance of an industrial facility
- Exploration of the legal landscape for a new business
- Exploration of the legal landscape for an existing business entering a new jurisdiction
- Collection of resources for a knowledge base of a
  - commercial organization
  - legal expert
  - educational institution
- Coarse-grained analysis first and more fine-grained analysis later

Outline

- Background
- Motivation
- Task
- Related Work
- Example Application
- Experiments
- Future Work

Motivation

- While the retrieval of statutes (Phase I) is reasonably well supported with existing legal IR systems the subsequent activities (Phases II & III) receive very little support.
- **Open-ended statutory analysis is labor intensive** and could benefit from automation.
- The aim of our work is to support such labor intensive analysis beyond traditional IR.

Outline

- Background
- Motivation
- Task
- Related Work
- Example Application
- Experiments
- Future Work

Task

Input: A set of statutes retrieved from a legal IR system in response to a query about a legal issue.

Output: A subset of the input set that contains the statutory provisions that are relevant to the legal issue.

To support the task we propose a use of an

interactive ML framework

[Based on an iterative interaction between a human expert (gives feedback) and a ML classifier (provides suggestions).]

Working hypothesis I: After a human expert marks a small portion of statutory provisions the system will provide reasonable suggestions.

Working hypothesis II: As the human expert marks an increasing number of provisions, the suggestions become more accurate.

Working hypothesis III: A model trained during statutory analysis can be helpful in supporting future analyses (if they are related).

Outline

- Background
- Motivation
- Task
- Related Work
- Example Application
- Experiments
- Future Work

Related Work

- Explicit **recognition of interactive ML** (Fails&Olsen 2003)
- **Application** of interactive ML in:
  - Web image search (Amershi et al. 2011)
  - Making sense of large network data (Chau et al. 2011)
  - Text search and filtering (Heimerl et al. 2012)
- **e-Discovery**
  - Human-aided computer cognition approach (Hogan et al. 2009)
  - TAR framework with multi-pass manual coding (Cheng et al 2013)
  - Highly-scalable classification framework (Krasner 2013)
- **(Legal) IR** enhancements
  - Relevance feedback (Manning et al. 2008)
  - Case-based reasoning (Moens 2002)
  - Dynamic document classification (Merkl & Schweighofer 1997)
  - Keyword extraction (Schweighofer & Merkl 1997)
  - Recommendation systems (Winkels et al. 2014)
  - Conceptual retrieval (Grabmair et al. 2015)

Outline

- Background
- Motivation
- Task
- Related Work
- Example Application
- Experiments
- Future Work

Example Application: Analysis

- The goal of the example analysis is to assess and compare selected US states' regulatory frameworks concerning **preparedness and response of the public health system** (PHS) **to public health emergencies**.
- There are **30 different types of agents** (e.g., doctor, emergency management, or special response team).
- The starting point of the analysis is to **compile an exhaustive list of relevant statutory provisions** conferring rights, obligations, prohibitions, or permissions on any of the agents in the context of the PHS preparedness and response to the public health emergencies.

The **criteria for relevance** of an individual provision are:

1. Does the provision confer a right, obligation, prohibition, or permission on any of the 30 agents?
2. Does it involve PHS preparedness and response to public health emergencies?

Example Application: Data Sets

- In this example we work with the statutory provisions from **Kansas** (4022 retrieved/802 relevant) and **Alaska** (1564 retrieved/474 relevant).
- These were retrieved from a commercial legal IR system with a set of hand-crafted search queries.

Example provisions:

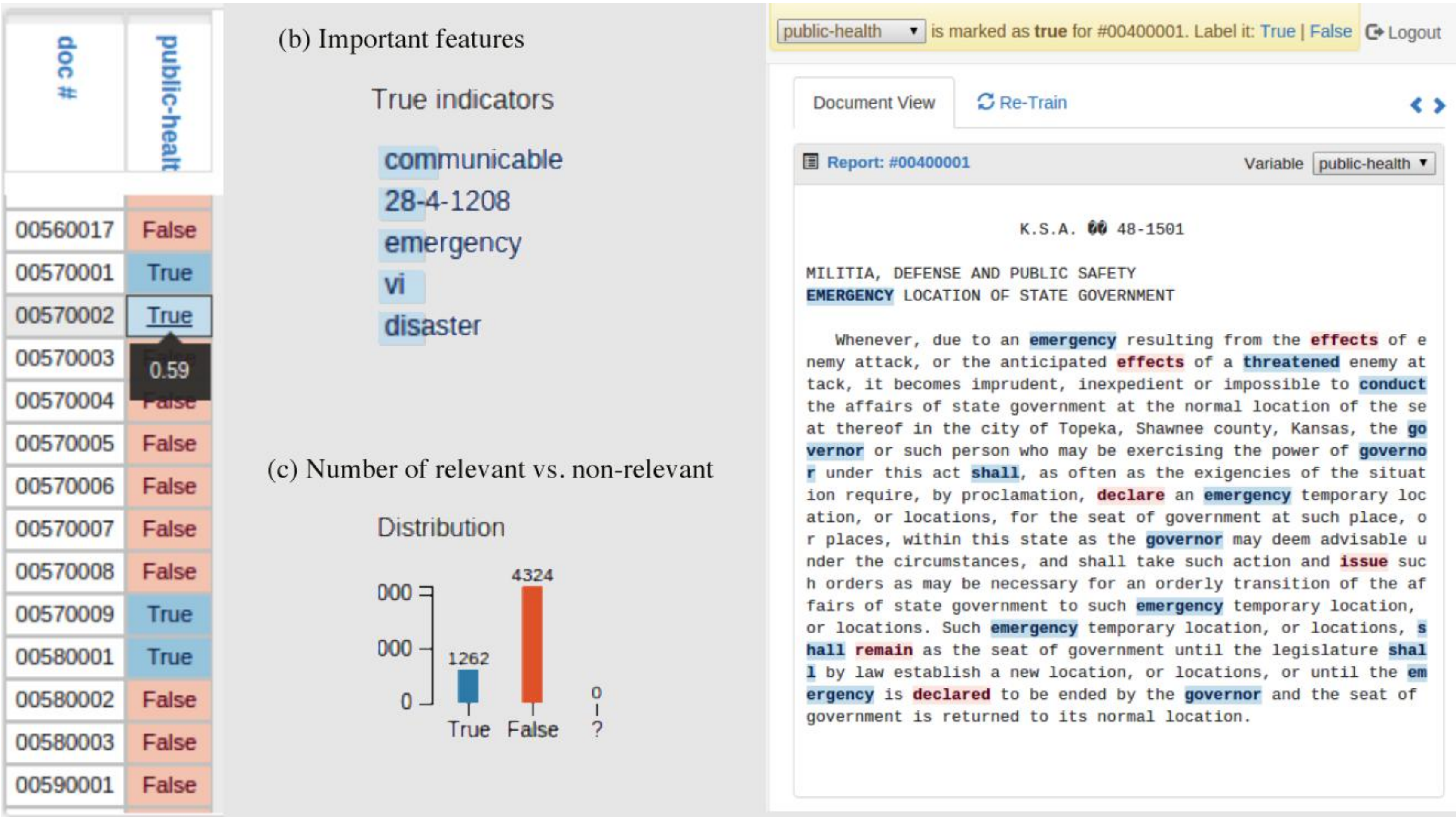
The township board of any such township shall have full direction and control over the operation of such township fire department. The board shall have the power to: Provide for the organization of volunteer members of such department and pay compensation to such members for fighting fires, responding to emergencies or attending meetings; (**RELEVANT**)

Each person storing the tires shall meet the requirements of subsection (b) of this regulation and the following requirements: provide access to each storage area for fire-fighting equipment by either of the following means: obtaining certification from the local fire department stating that there is adequate access to each storage area for fire-fighting equipment; (**NOT RELEVANT**)

Example Application: Software Tool

- For our experiments we used a **tool with an interactive GUI** developed at Pitt. (Trivedi et al 2015)
- User's decisions about each provision are recorded and upon user's request the ML classifier is retrained.
- The tool is capable (among other things) of:
  - **suggesting labels** for yet unprocessed provisions
  - informing users about its **confidence**
  - informing users about the **prominent features**
- The tool uses an **SVM classifier** with a linear kernel as a classification algorithm.

Example Application: Software Tool



(d) Suggestions with confidence score

(a) Text of a provision with highlighted terms

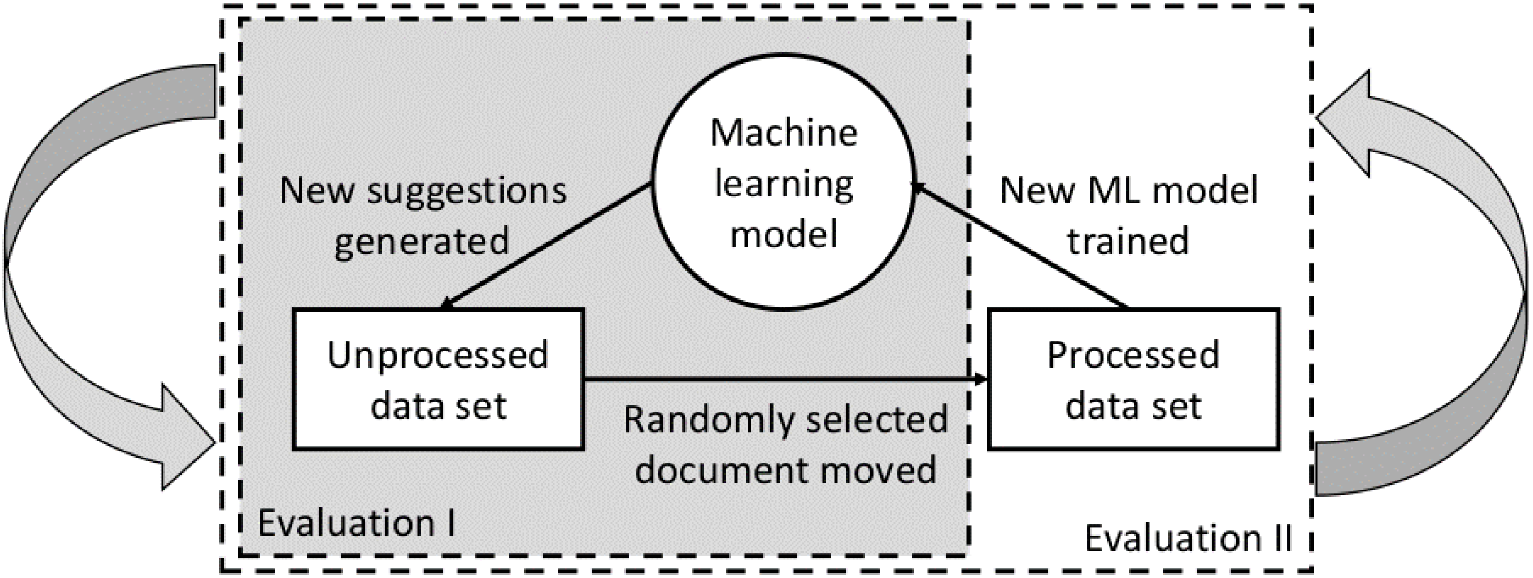
Outline

- Background
- Motivation
- Task
- Related Work
- Example Application
- Experiments
- Future Work

Experiments: Design

Two different experiments

- **Cold-start experiment** examines a situation in which the analysis starts from a clean slate. [Kansas]
- **Knowledge re-use experiment** analyzes a re-use of the ML classifier in a related analysis. [Alaska]



Experiments: Evaluation

ROC – plot of the true positive (TP) rate (y axis) against the false positive (FP) rate (x axis) for different decision thresholds.

AUC – area under the ROC curve; probability that a classifier will rank a randomly chosen positive data point higher than a randomly chosen negative one.

Precision – ratio of the correctly predicted positive labels over all the labels predicted as positive.

Recall - ratio of the correctly predicted labels over all the labels that should have been predicted as positive.

F_1-measure – harmonic mean of precision and recall where they are treated as equally important.

Experiments: Evaluation

Two different evaluation perspectives

- **ML model-oriented:** how well the ML classifier works on a held-out (test) set.

  Evaluation measures: AUC, ROC, P, R, F_1

- **Interaction oriented:** how well the system (ML classifier + human expert) works on the whole data set as compared to manual classification.

  Evaluation measures: P, R, F_1

  Baselines: manual P+, manual R+

Experiments: Results

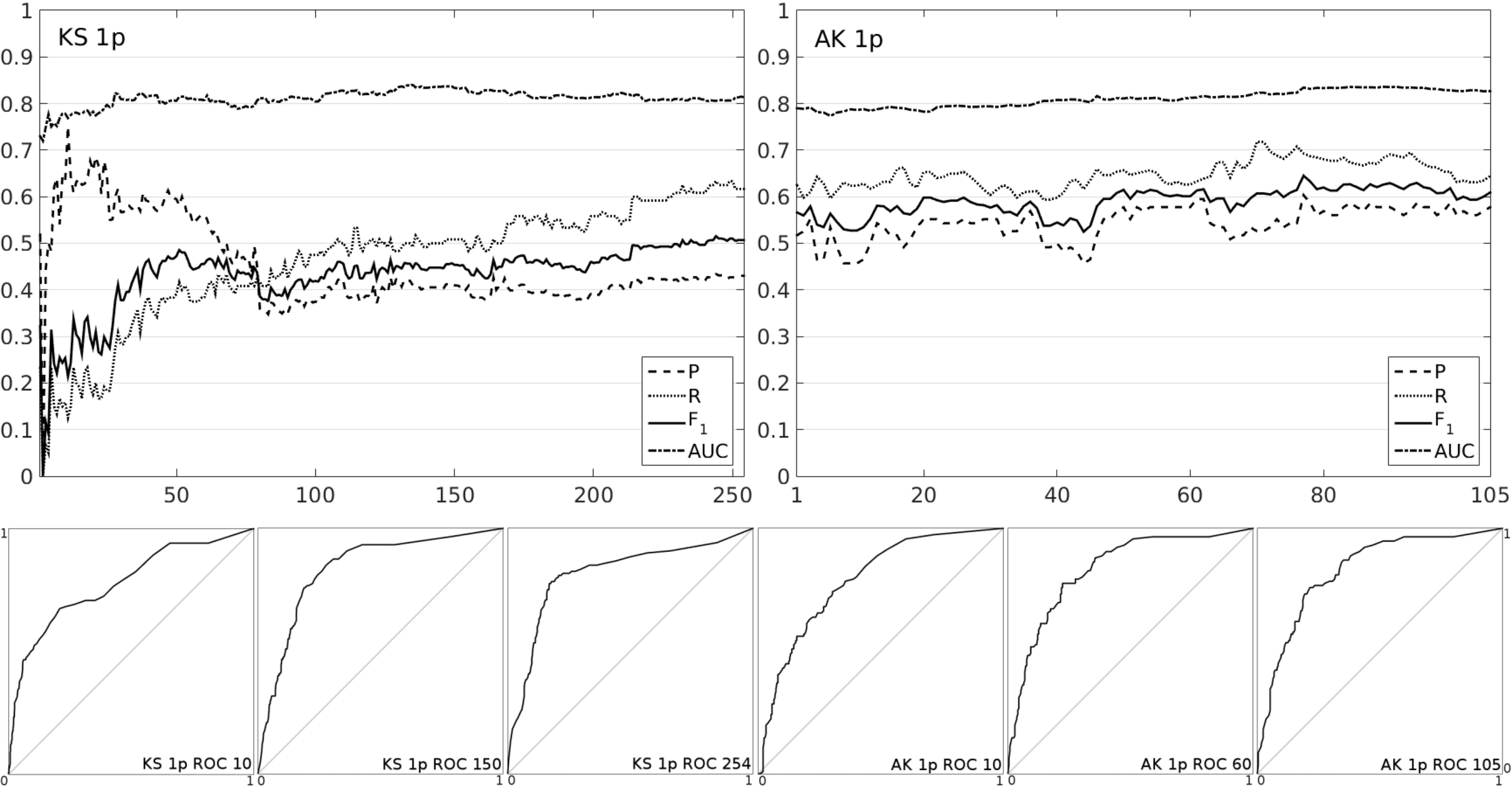
Kansas (cold-start experiment)

	Measure/# docs	10	50	100	150	200	254	304
ML	AUC	.78	.81	.80	.83	.81	.81	-
ML	P/R/F_1	.63/.16/.25	.59/.39/.47	.38/.48/.42	.41/.51/.45	.39/.56/.46	.43/.62/.51	-
interaction	P/R/F_1 (manual P+)	1/.02/.04	1/.12/.21	1/.26/.41	1/.43/.6	1/.63/.77	1/.85/.92	1/1/1
interaction	P/R/F_1 (manual R+)	.21/1/.34	.24/1/.38	.27/1/.43	.31/1/.47	.37/1/.54	.51/1/.68	1/1/1
interaction	P/R/F_1 (semi-auto)	.75/.17/.28	.75/.48/.58	.69/.6/.64	.79/.7/.74	.81/.84/.83	.89/.94/.91	1/1/1

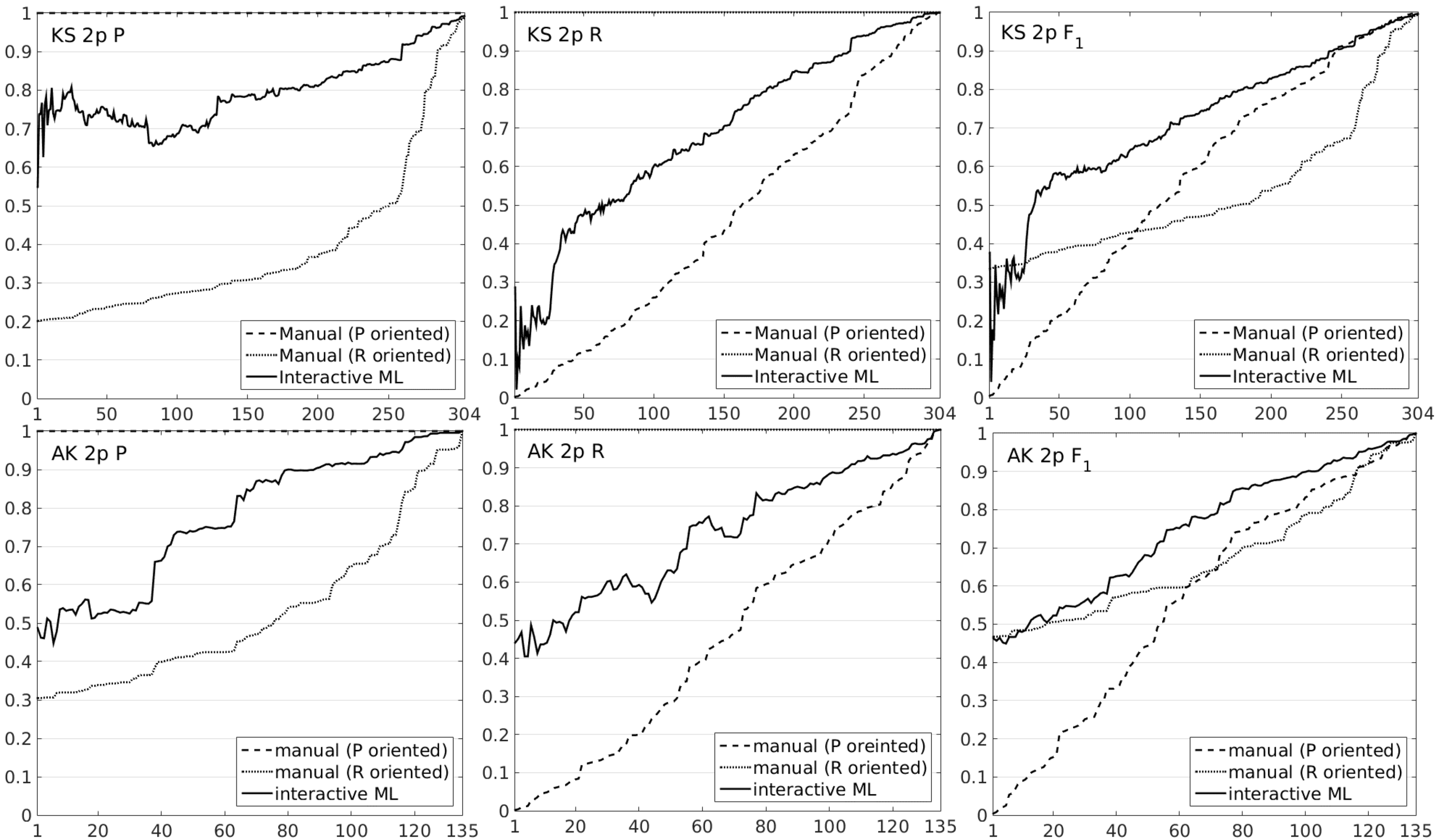
Alaska (knowledge re-use experiment)

	Measure/# docs	10	30	50	70	90	105	135
ML	AUC	.79	.79	.81	.82	.84	.83	-
ML	P/R/F_1	.46/.62/.53	.55/.6/.58	.58/.66/.61	.53/.72/.61	.58/.69/.63	.58/.64/.61	-
interaction	P/R/F_1 (manual P+)	1/.05/.09	1/.14/.25	1/.28/.44	1/.47/.64	1/.65/.79	1/.76/.86	1/1/1
interaction	P/R/F_1 (manual R+)	.32/1/.48	.35/1/.51	.41/1/.59	.47/1/.64	.56/1/.72	.66/1/.79	1/1/1
interaction	P/R/F_1 (semi-auto)	.53/.44/.48	.52/.6/.56	.74/.63/.68	.72/.87/.79	.91/.85/.88	.92/.9/.91	1/1/1

Experiments: Results
(ML model-oriented)



Experiments: Results
(interaction-oriented)



Experiments: Discussion

- The **results confirm our working hypotheses**.
  - The AUC score above 0.8 says that the classification model is good (but not excellent).
  - The precision around 0.4 and recall above 0.6 is promising.
  - Re-use of the model from the previous analysis eliminated the cold-start problem.
- They also show that fully automated relevance assessment in statutory analysis is difficult.

Outline

- Background
- Motivation
- Task
- Related Work
- Example Application
- Experiments
- Future Work

Future Work

- Processing of some documents was much more beneficial than processing of other documents. Therefore, **active learning** (Settles 2010) techniques could lead to a major improvement.
- Techniques from **transfer learning** (Pan&Yang 2010) could facilitate better knowledge transfer (here transfer between two states) or allow transfer between analyses that are less related.
- There could be **other benefits** of using the framework beyond reasonable suggestions about relevance:
  - more consistent and reliable results
  - speeding up of the reviewing process

Conclusions

- We examined if and how the interactive ML approach could help in determining which provisions retrieved with an IR system are relevant in an analysis of a legal issue.
- We have shown that:
  i. Interactively trained ML classifiers provide **reasonable suggestions** about the relevance of statutory provisions;
  ii. The accuracy of the suggestions **improves** as more of the provisions are being processed; and
  iii. It is possible to **re-use** the classifiers in future analyses.
- The use of the interactive ML approach reliably outperforms the traditional manual assessment.

Thank you for your attention!

Jaromir Savelka (jas438@pitt.edu)
Gaurav Trivedi (gat23@pitt.edu)
Kevin D. Ashley (ashley@pitt.edu)