Semi-Supervised Methods for Explainable Legal Prediction

Karl Branting
Craig Pfeifer
Lisa Ferro
Brandy Weiss
Mark Pfaff
Alex Yeh
Amartya Chakraborty
Bradford Brown

The MITRE Corporation
McLean, VA, USA

MITRE

17th International Conference on AI and Law (ICAIL 2019).

Problem

- **High volumes of administrative cases can lead to backlogs, inconsistencies, high administrative costs, and opacity**
- **May US agencies suffer from huge and rapidly growing case backlogs**

US Agency	Case backlog	Mean days to decision
SSA	1,000,000+	450
Medicare	400,000+	550
VA	400,000+	1,000
USCIS	850,000+	700

- **There are too few administrative law judges (~1,400) to cope with growing flood of cases**

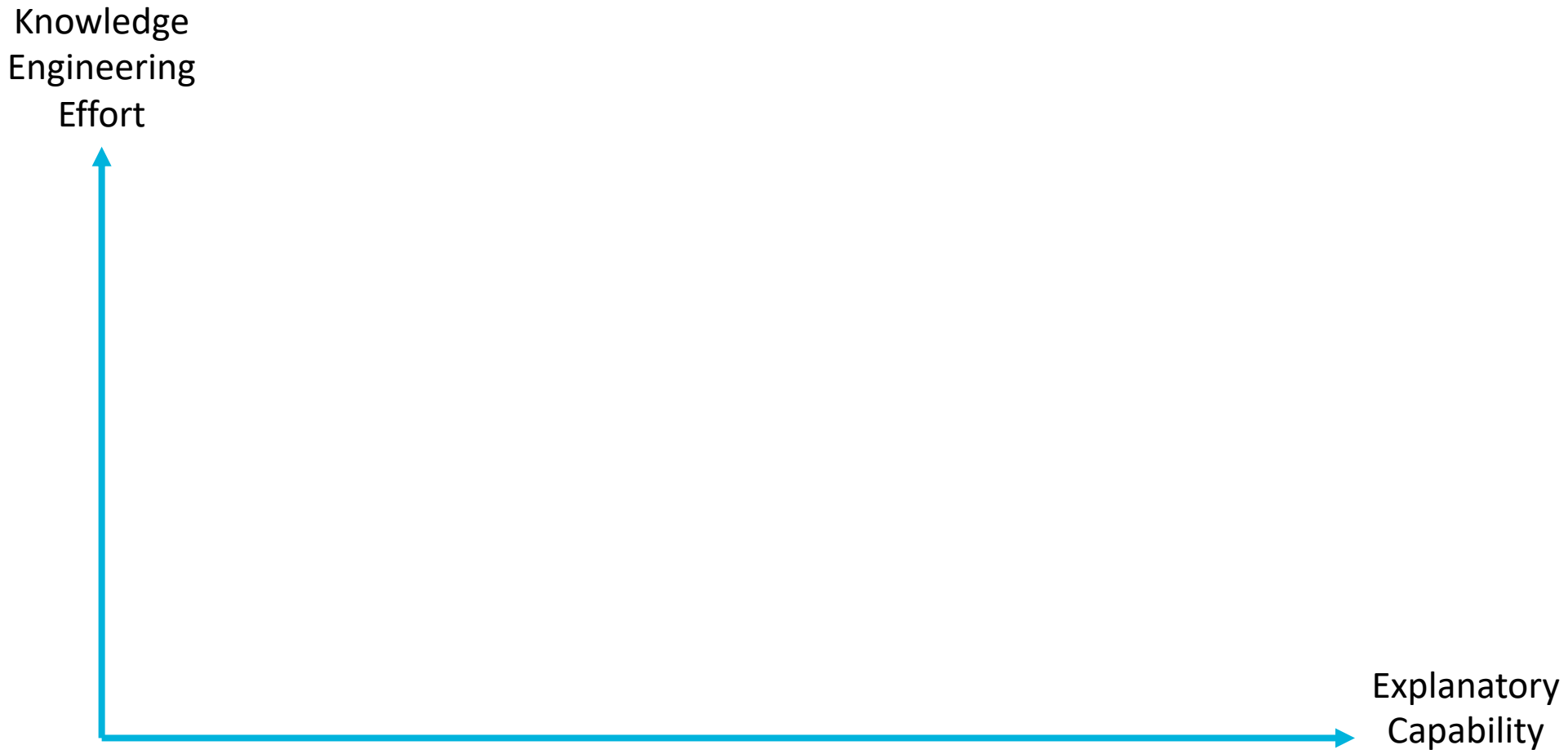
Key Idea

- **Resource: abundant documents linking *facts* to *outcomes***
- **Approach:**
  - Train ML models to predict outcomes from new facts
  - Adapt models for decision support
- **Hypothesis:**
  - A single framework for explainable prediction is applicable to the range of agencies
- **Potential benefits**
  - Improved consistency and speed of fact-to-outcome decisions
  - Validation of decision processes
  - Improved institutional transparency
  - Greater access to justice for citizens
  - Exploitation of knowledge latent in agency document collections

Challenges

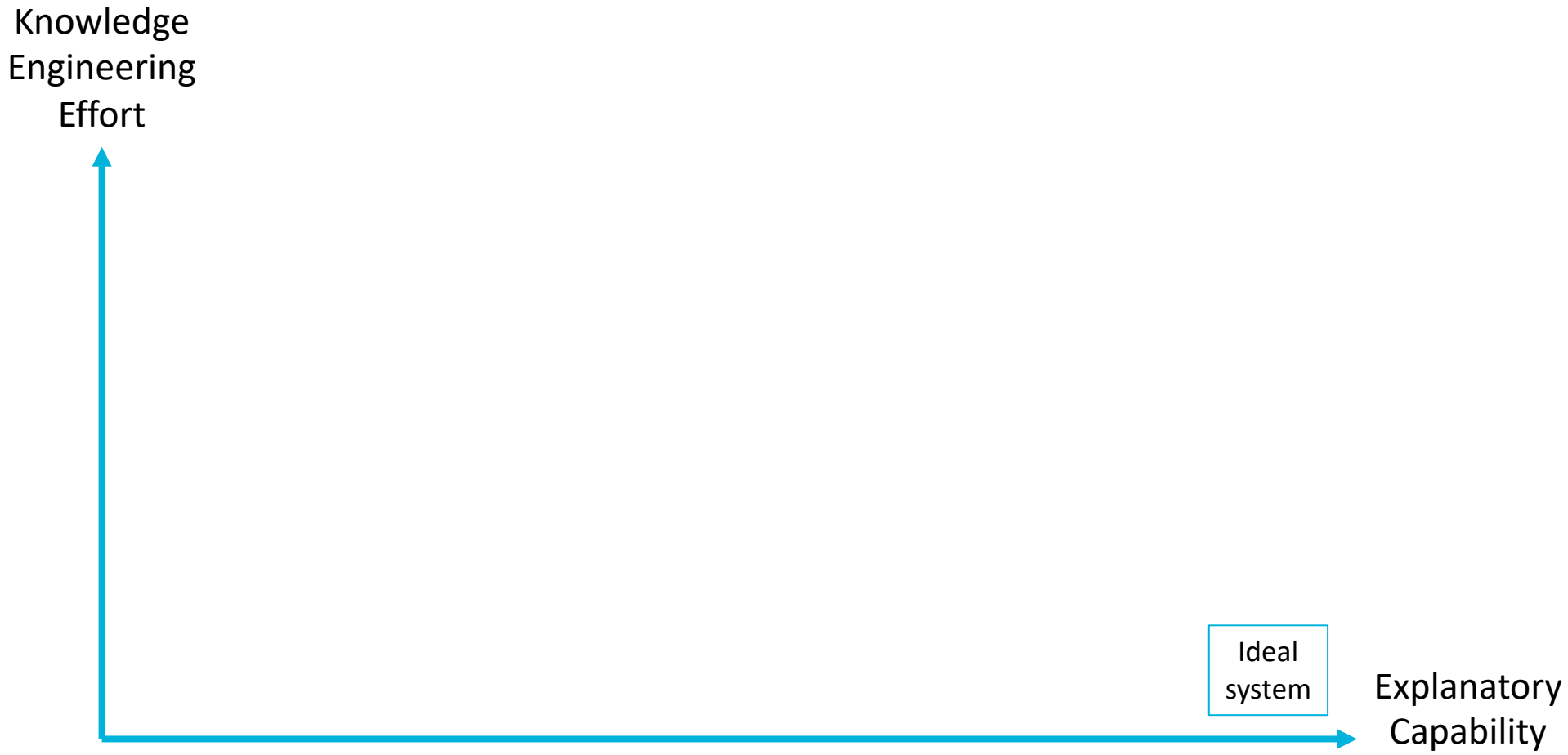
- ***Analytical:*** **Facts are expressed in documents as unstructured text**
- ***Usability:*** **System predictions must have understandable justifications in order for users to find them acceptable.**
- ***Institutional:*** **Due process requires the justifications to be based on authoritative legal texts**
- ***Pragmatic:*** **Agencies will adopt only those predictive systems having limited and predictable engineering costs**

Knowledge-Engineering Cost/Explanation-Quality Tradeoff



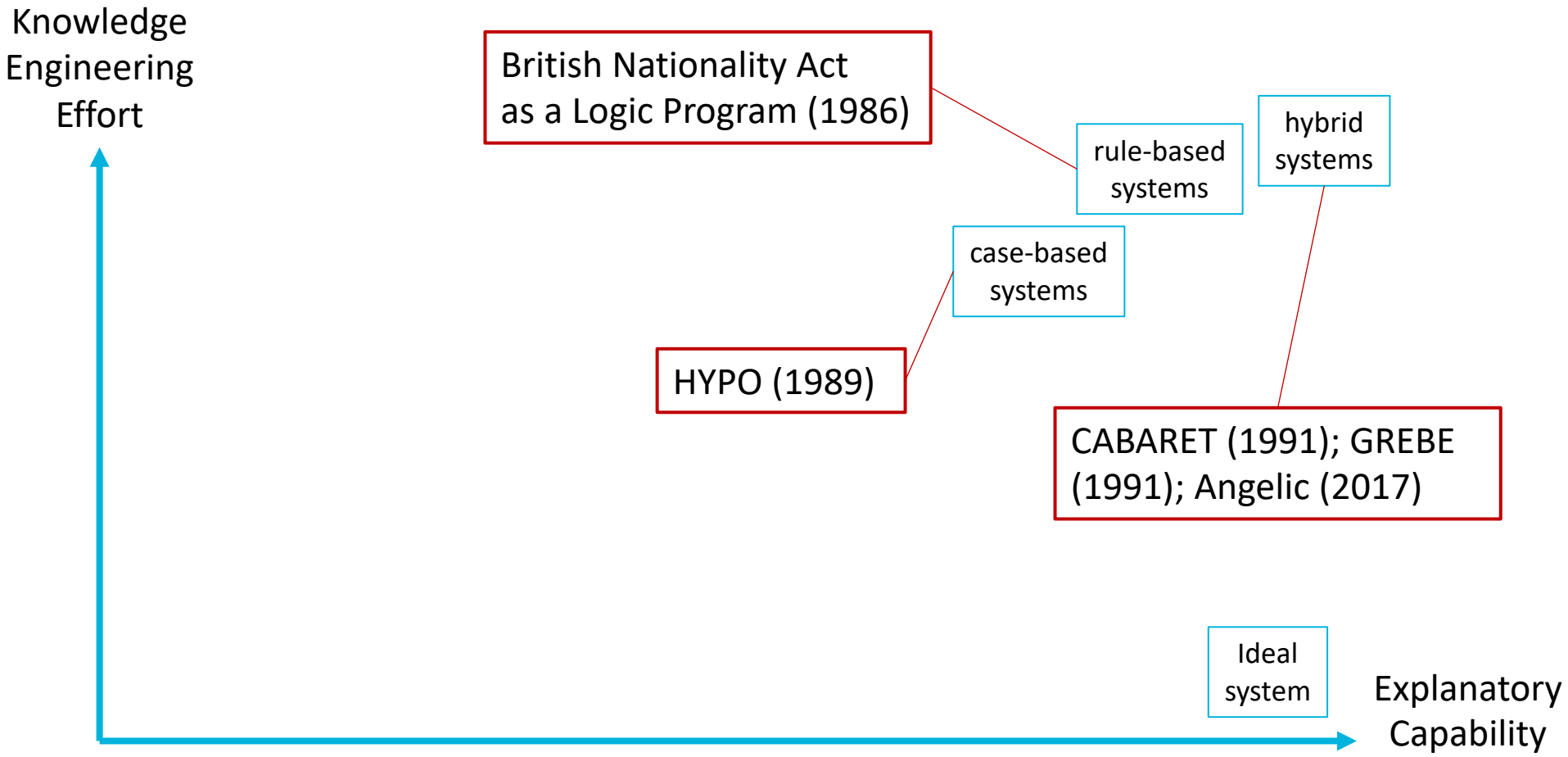
MITRE

Knowledge-Engineering Cost/Explanation-Quality Tradeoff



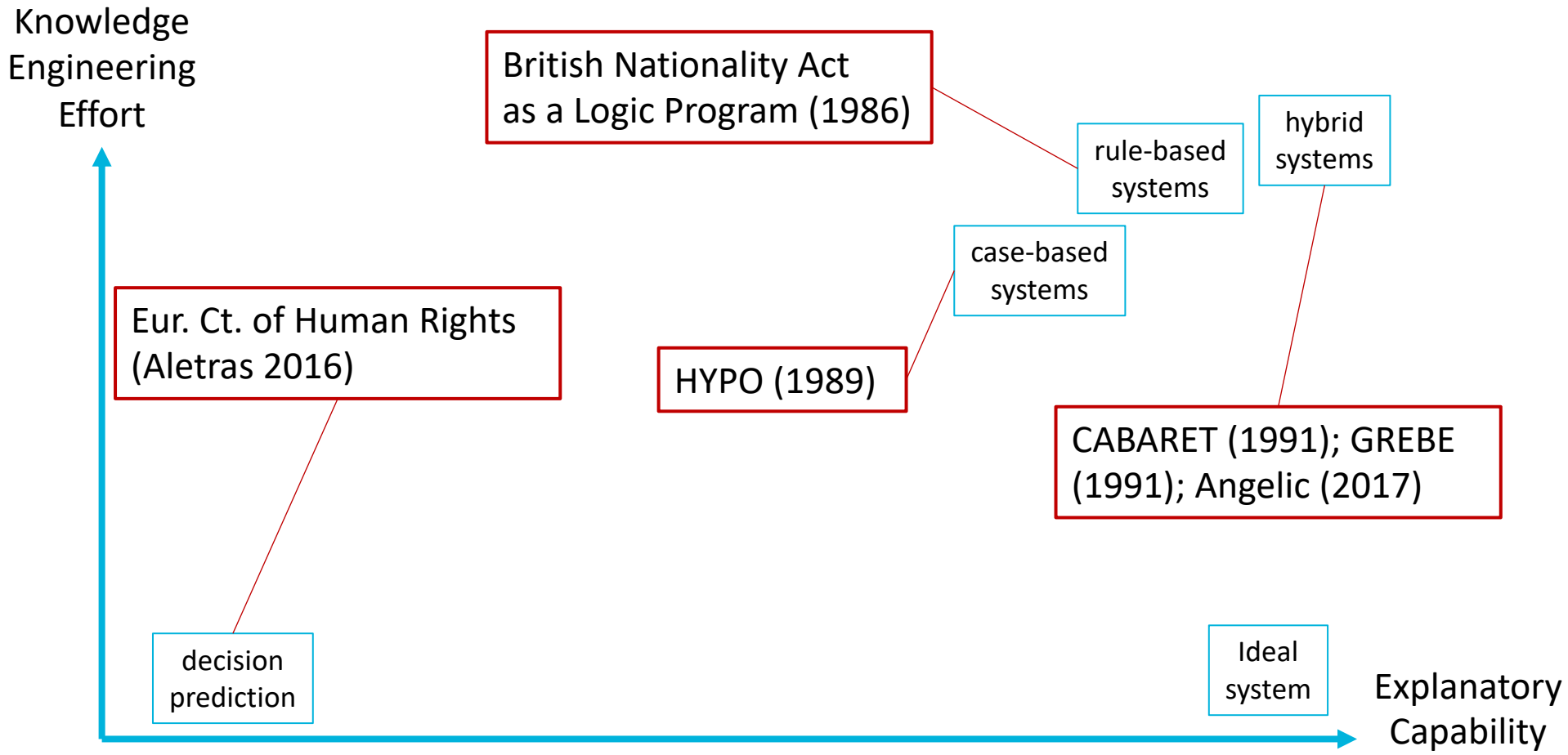
MITRE

Knowledge-Engineering Cost/Explanation-Quality Tradeoff



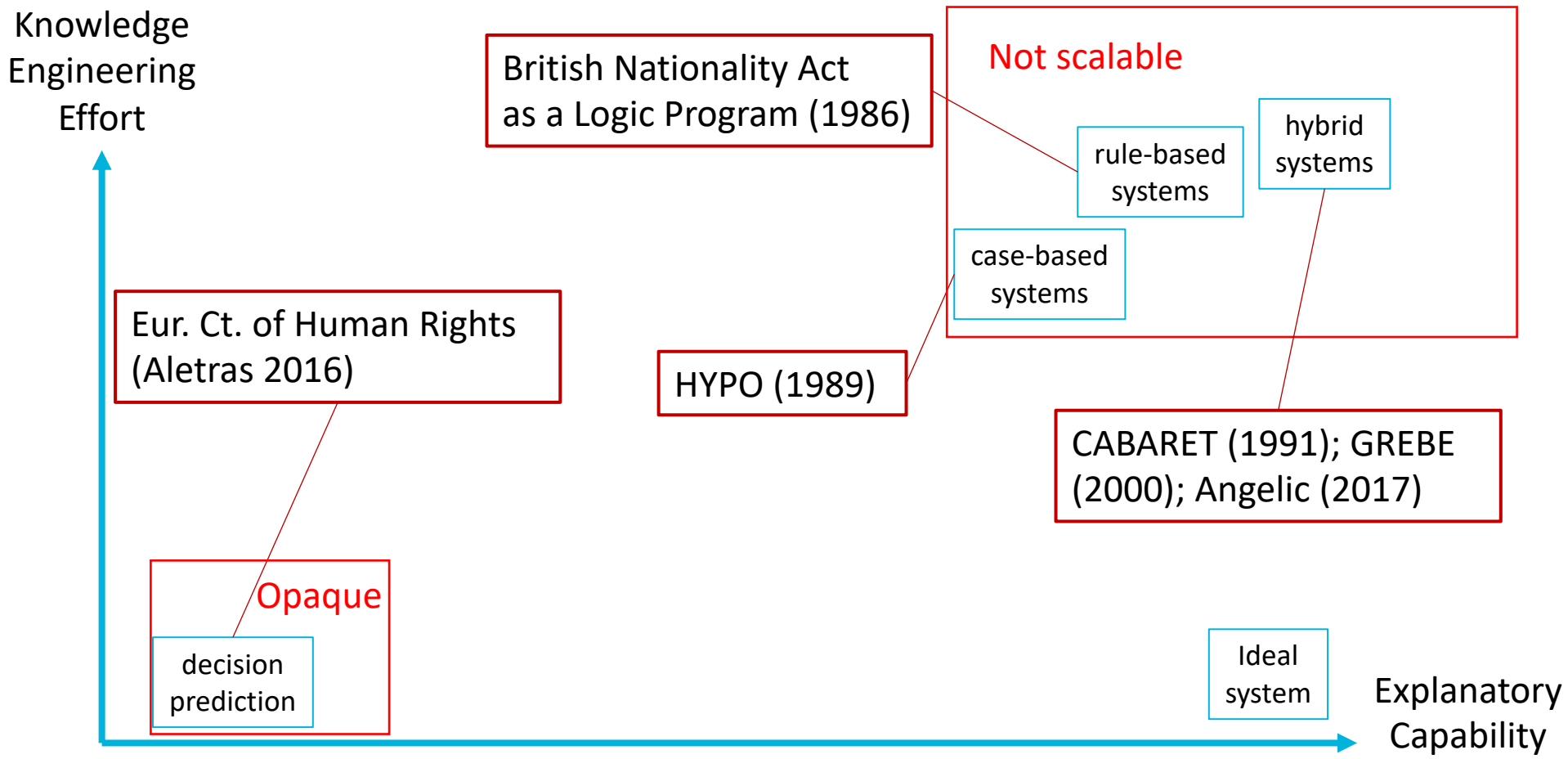
MITRE

Knowledge-Engineering Cost/Explanation-Quality Tradeoff



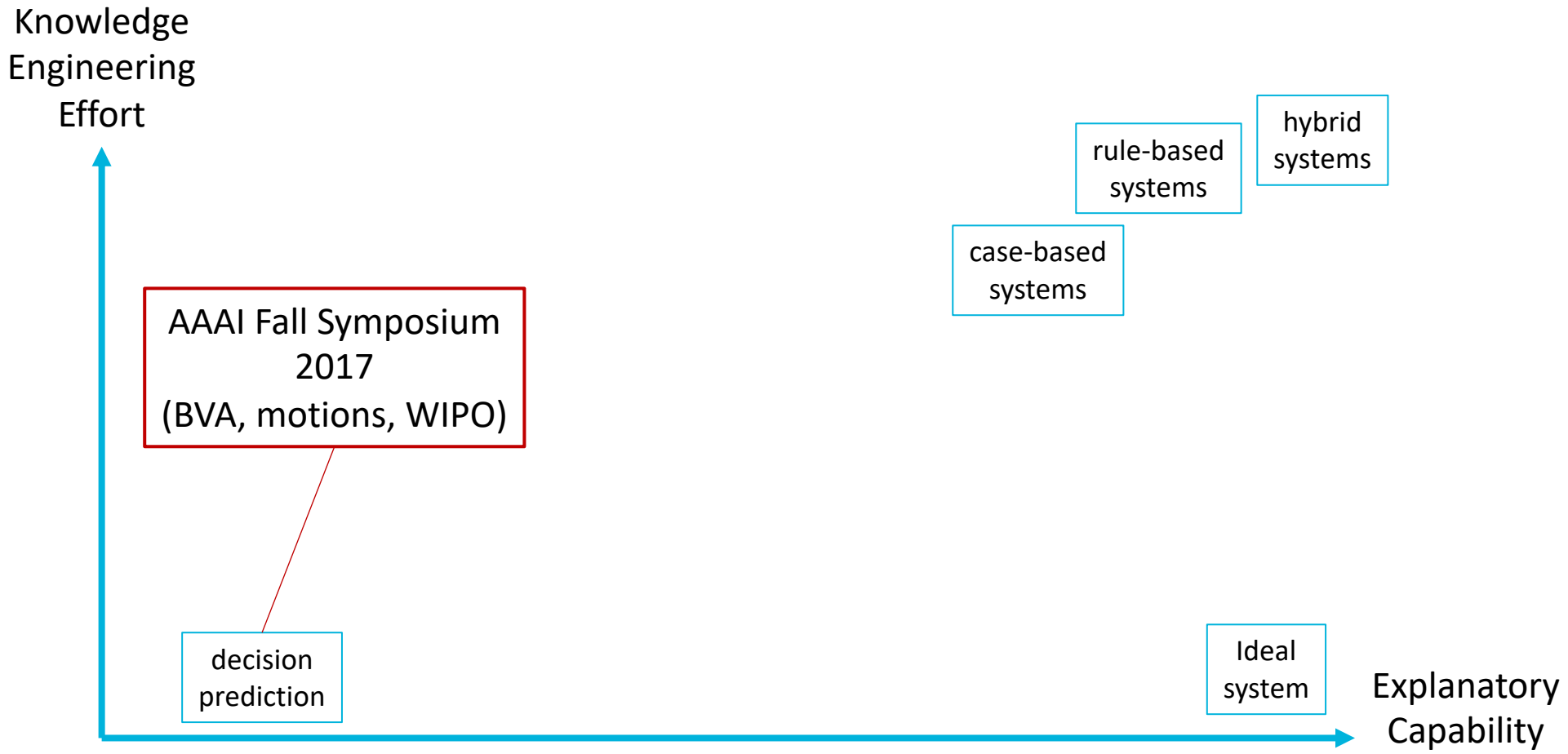
MITRE

Knowledge-Engineering Cost/Explanation-Quality Tradeoff



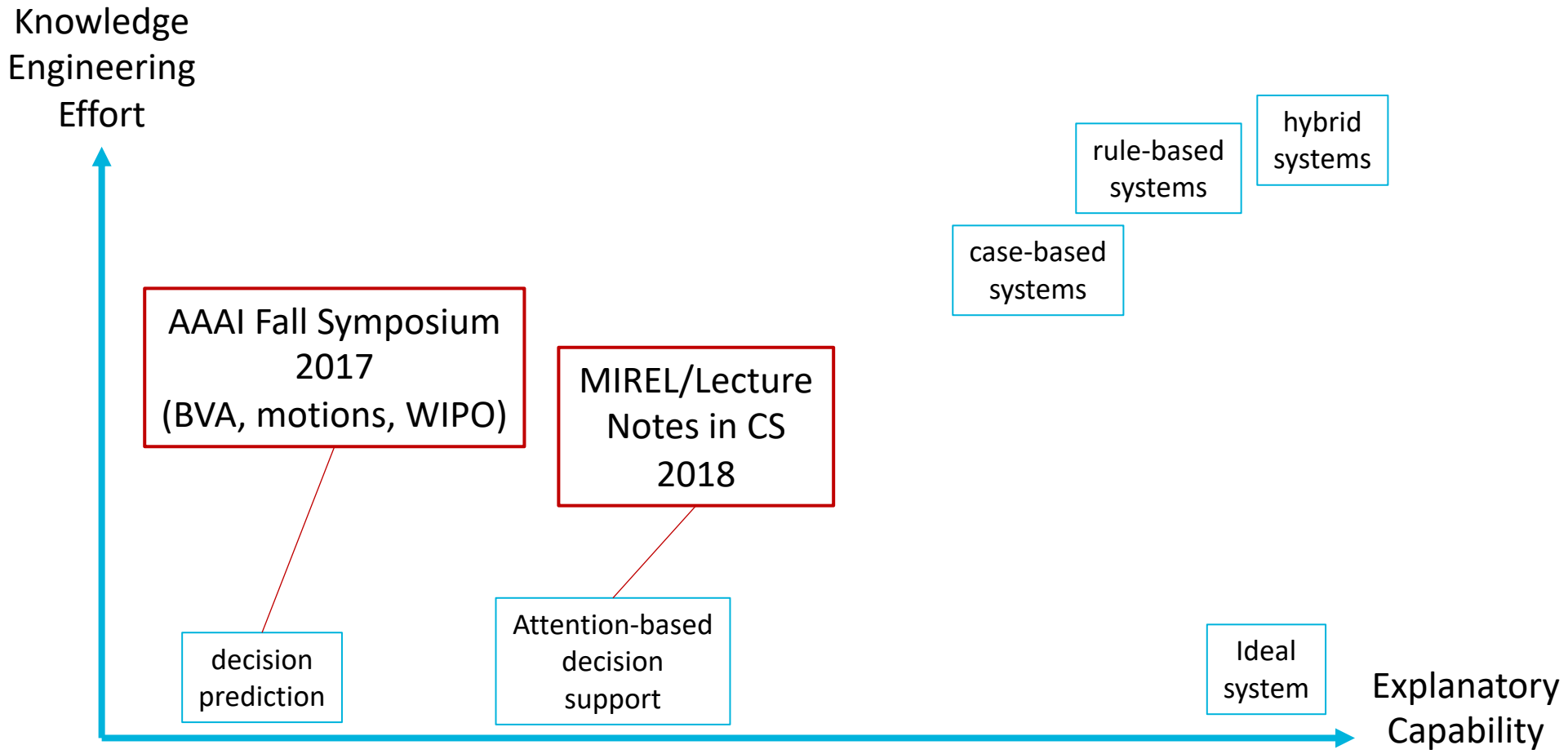
MITRE

Prior MITRE “EPIC” Activities



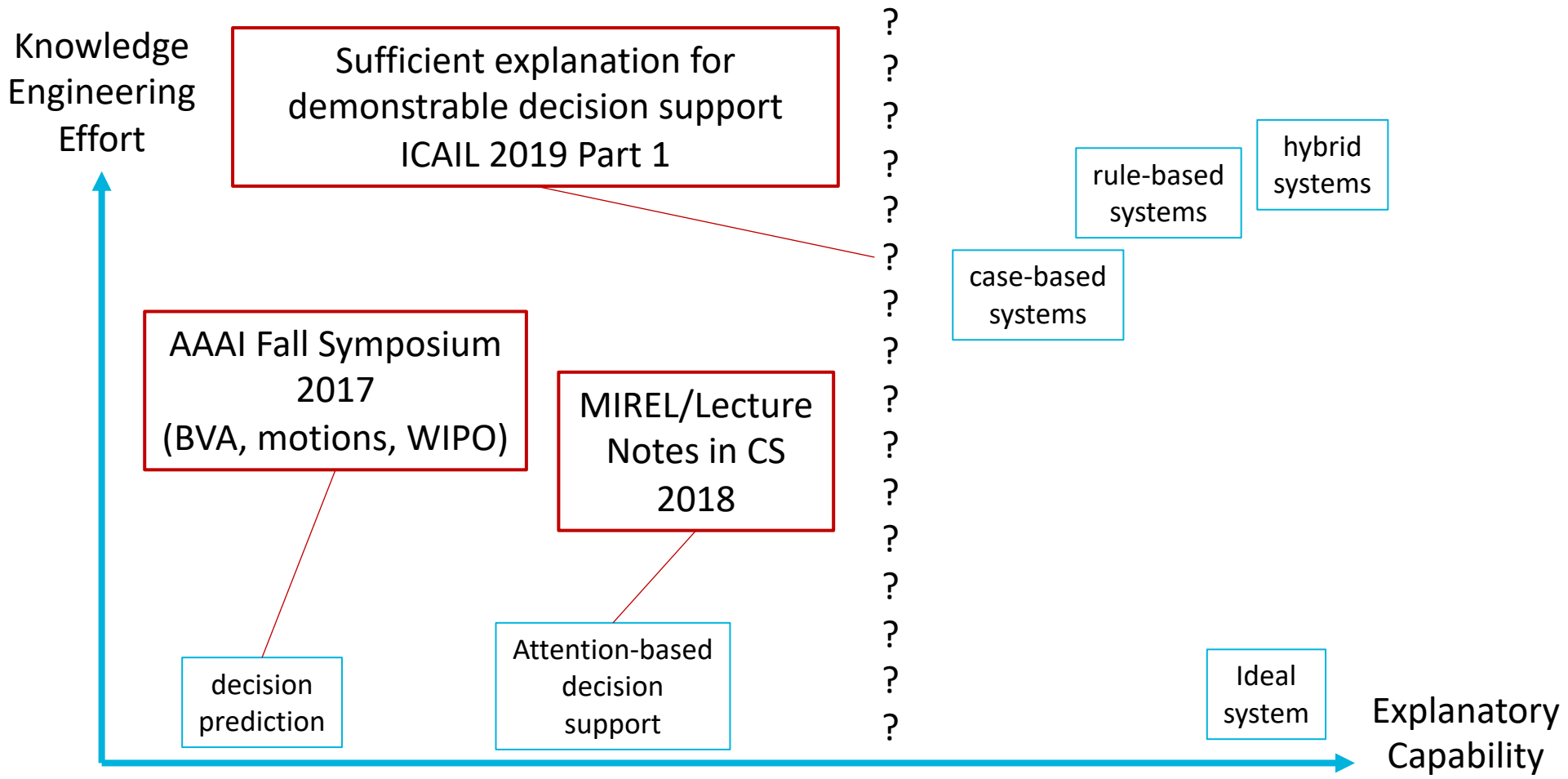
MITRE

Prior MITRE “EPIC” Activities



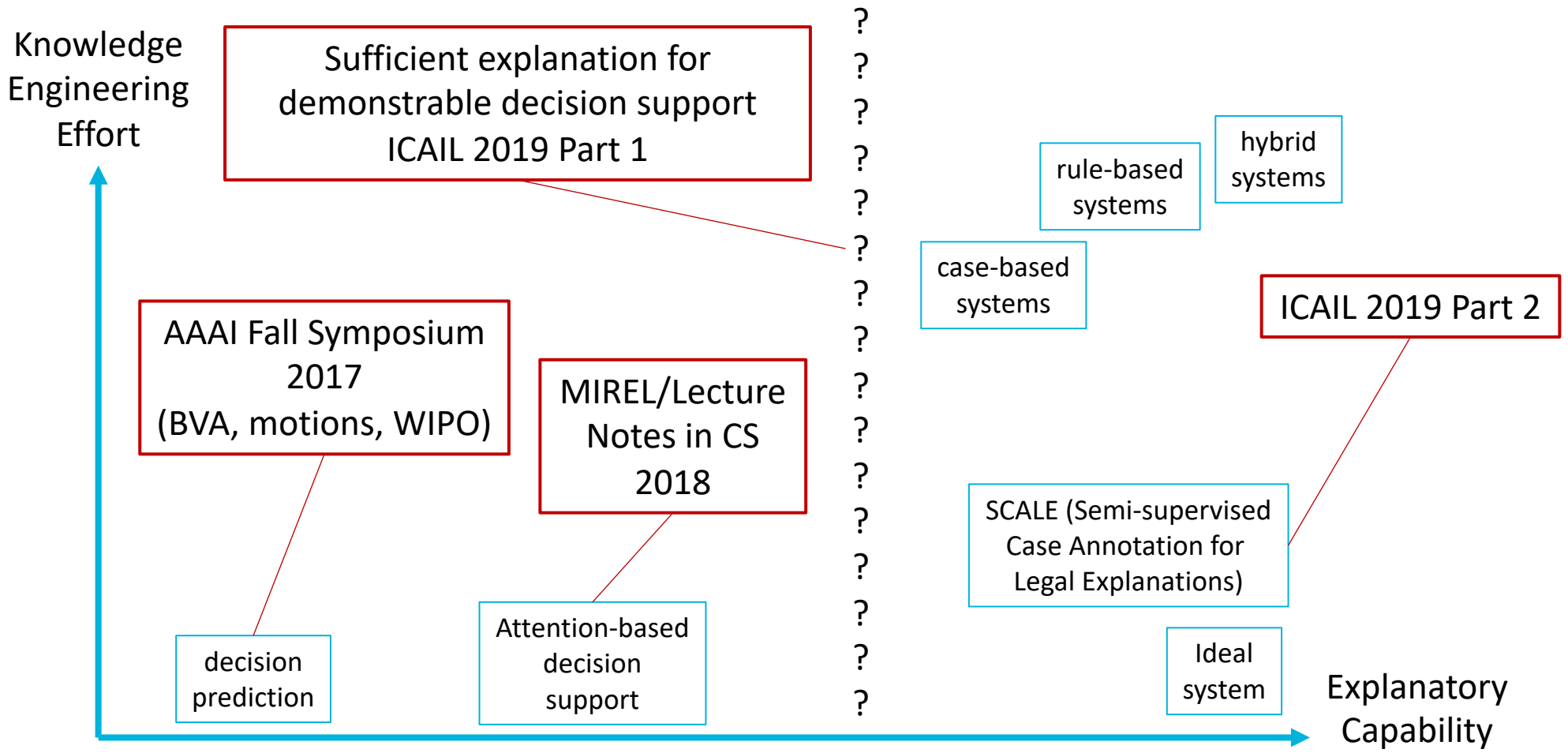
MITRE

Prior MITRE Activities



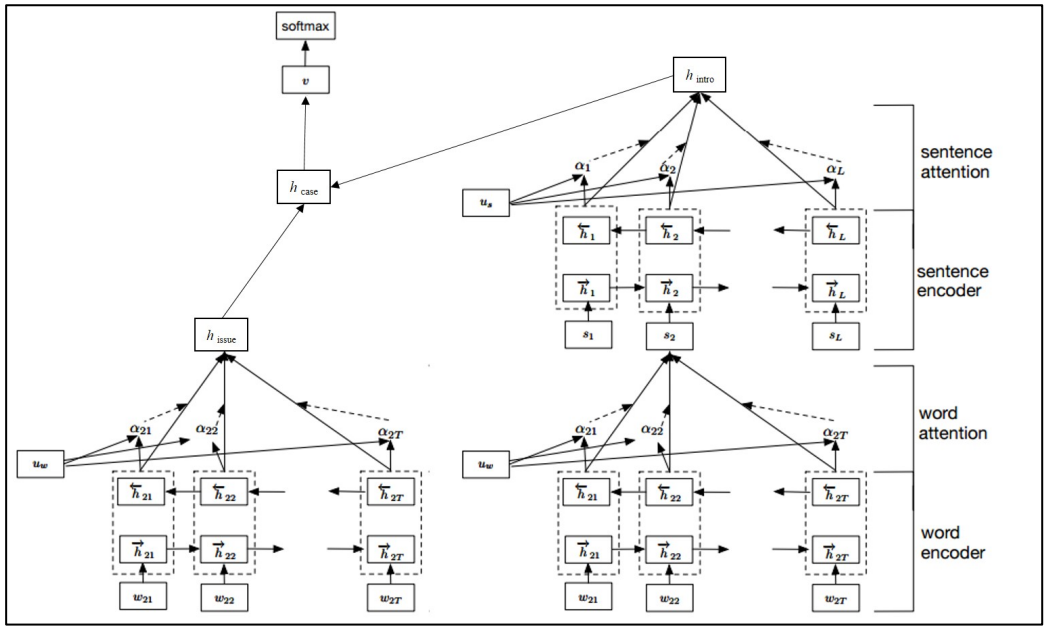
MITRE

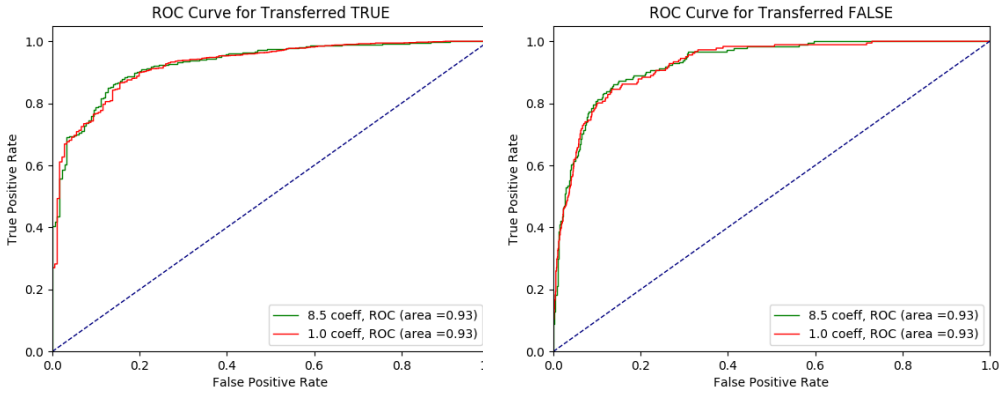
Current MITRE Activities



MITRE

Highlighting Based on Hierarchical Attention Network





5. Parties' Contentions

A. Complainant

In a very scantily argumented brief, the Complainant mentions that it is owner of the trademark PARAJUMPERS since 2006. It further argues that there is no evidence that the Respondent has any interest in using the disputed domain name, since the name "Parajumpers" derives from a military squadron that is based in Alaska. Moreover, the Complainant contends that the website has never have been used in the last five years, but has always been up for sale. The disputed domain name would have been primarily registered for the purpose of selling it for the best offer. The Respondent would have mentioned a minimum price of USD 73,200 during preceding negotiations between the Parties.

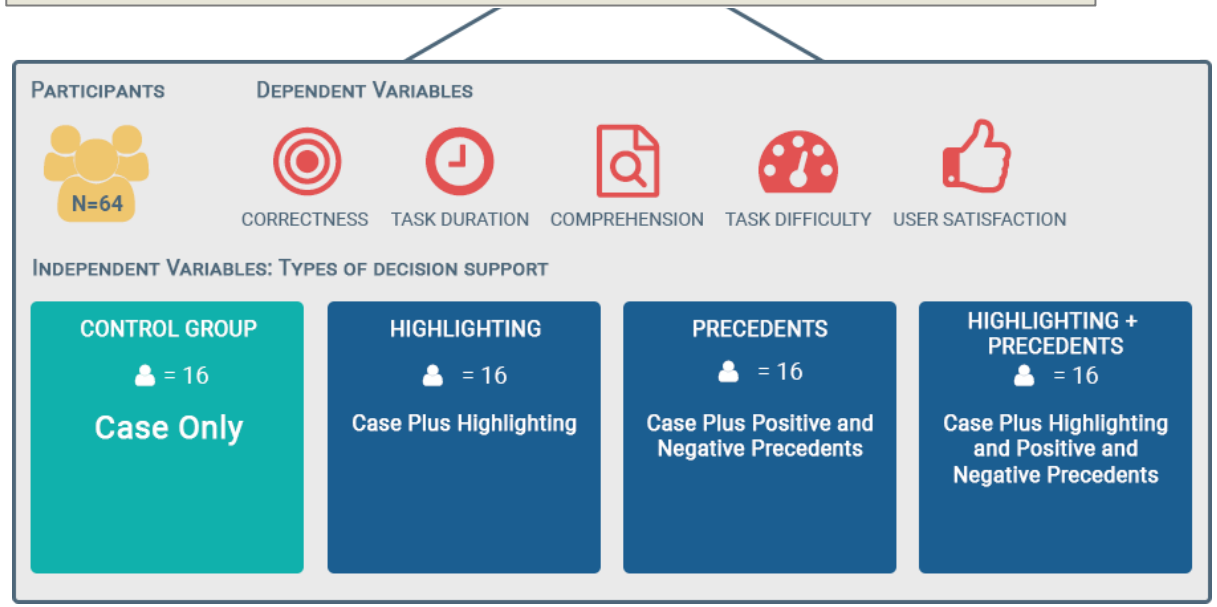
B. Respondent

In summary, the Respondent contends that there is no basis for the requested transfer. First, the Complainant waited over 10 years to initiate the present proceedings. The doctrine of laches should therefore apply. Second, the specific conditions set under the UDRP for an order of transfer are not met in the present case. When the Respondent registered the disputed domain name in 2003, the Complainant had not yet registered its trademark, and did probably not even exist at the time. The disputed domain name is a common descriptive, complex word being the combination of "para", which is derived from "parachutes" denoting a military unit utilizing parachutes, and "jumpers".

Further, the Respondent argues that it has rights and a legitimate interest in the disputed domain name because of its common descriptive meaning, and that it has registered many domain names combining the words "para", "jump", and "jumping". The Respondent claims that it registered the disputed domain name in good faith and it is using it in connection with the *bona fide* offering of goods and services and for a legitimate purpose. In the view of the Respondent, common word domain names are easy to remember and, thus, commercially valuable.

Part 1. Evaluation of Attention-Based Decision Support

Research question:
Can decision-making be improved by highlighting sentences assigned the highest weights by an Attention Network that was trained to predict case outcomes?



MITRE

Decision Support Evaluation
Participants & Procedure

- **64 total participants**
  - 37 with legal experience
  - 27 without legal experience
  - *Note: 3 participants with legal experience were excluded from the analysis after data screening*
- **Each participant was randomly assigned to 1 of 4 conditions**
- **Participants reviewed 2 problem cases and completed a survey**

MITRE

Decision Support Evaluation: Findings

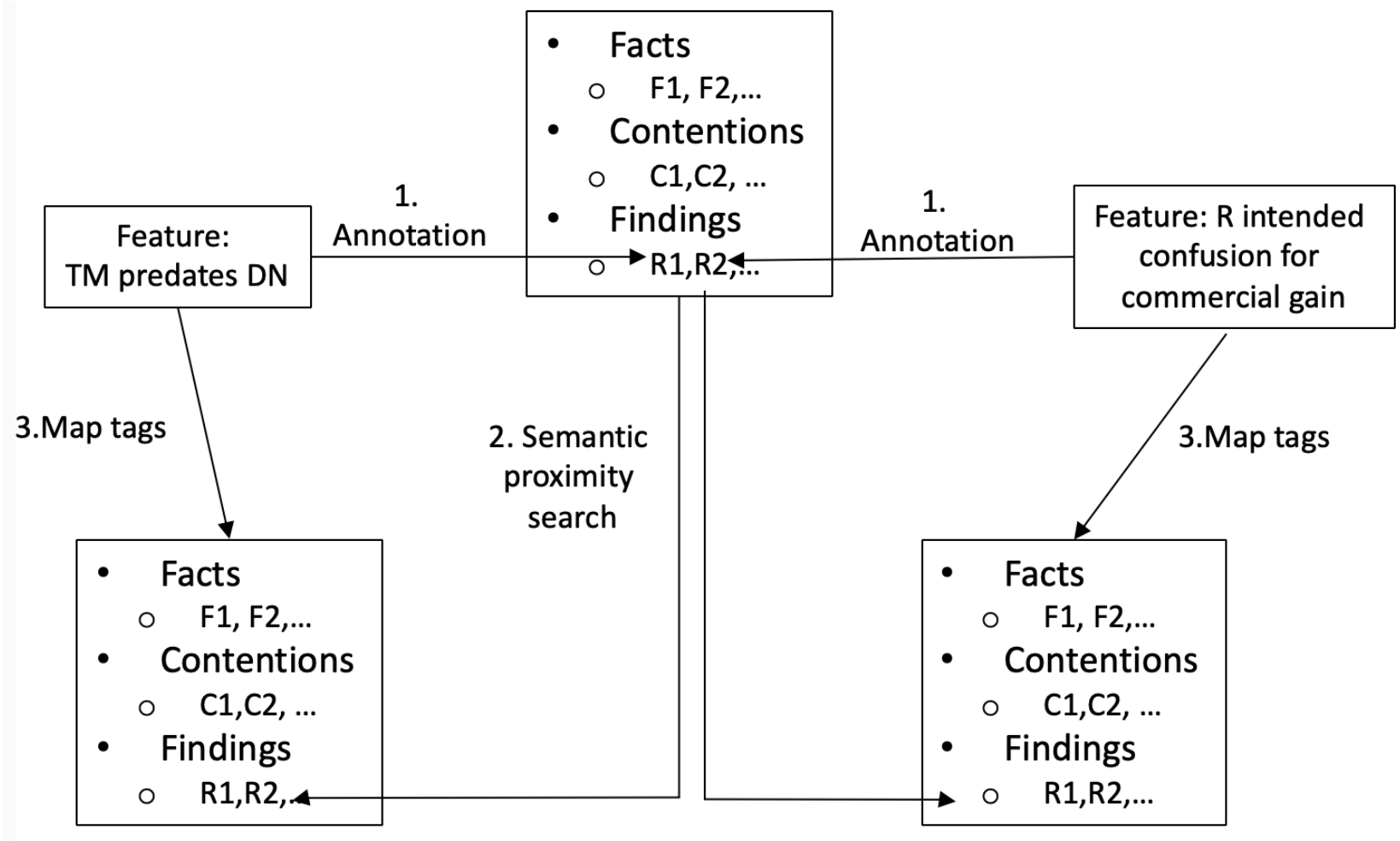
- **No benefit in decision accuracy or speed found from highlighting**
  - Some subjects said they didn't understand why text was highlighted
- **Precedents sometimes improved decision accuracy**
  - Decisions were slower
  - Highlighting precedents had no effect
- **In law, it seems that "Attention is not Explanation"**
- **Revised Hypothesis**
  - Effective decision support appears to require identification of legally relevant factual elements

Part 2. SCALE: Semi-supervised Case Annotation for Legal Explanations

- **Identify predictive, meaningful features based on regularities in decision justifications**
- **Annotate a representative subset of decision justifications**
- **Project annotations onto the justification texts of remaining cases**
  - Stylized case language means sentences with similar legal functions are close in semantic vector space
  - Expands small manually-tagged set to a large semi-supervised training set
- **Two-step Machine Learning**
  - $Model_1$ predicts features from case facts and contentions
    - Based on projected features
  - $Model_2$ predicts decision and explanation from predicted features
    - Builds on prior work that used manually-constructed features

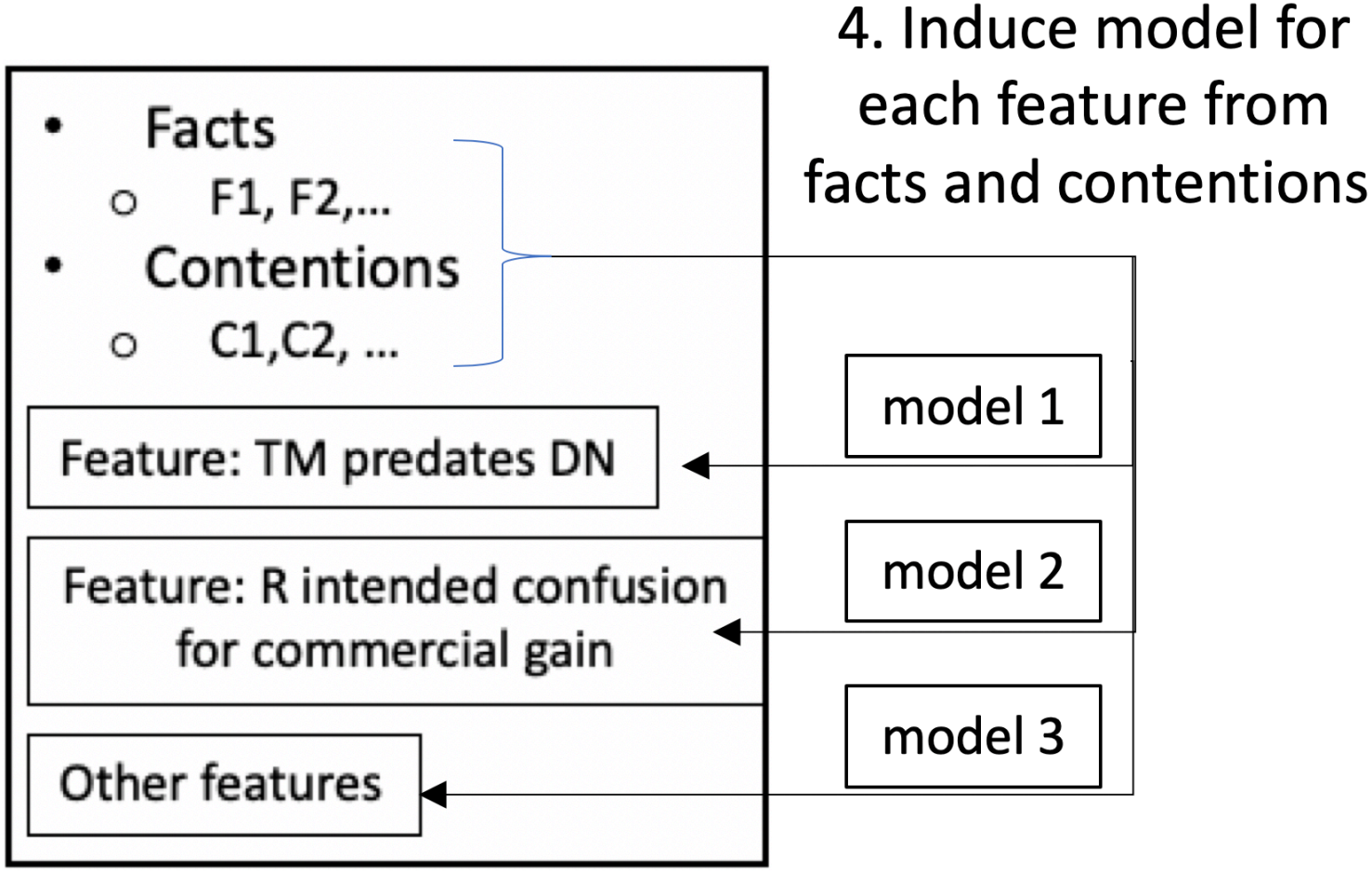
MITRE

SCALE: Semi-supervised Case Annotation for Legal Explanations



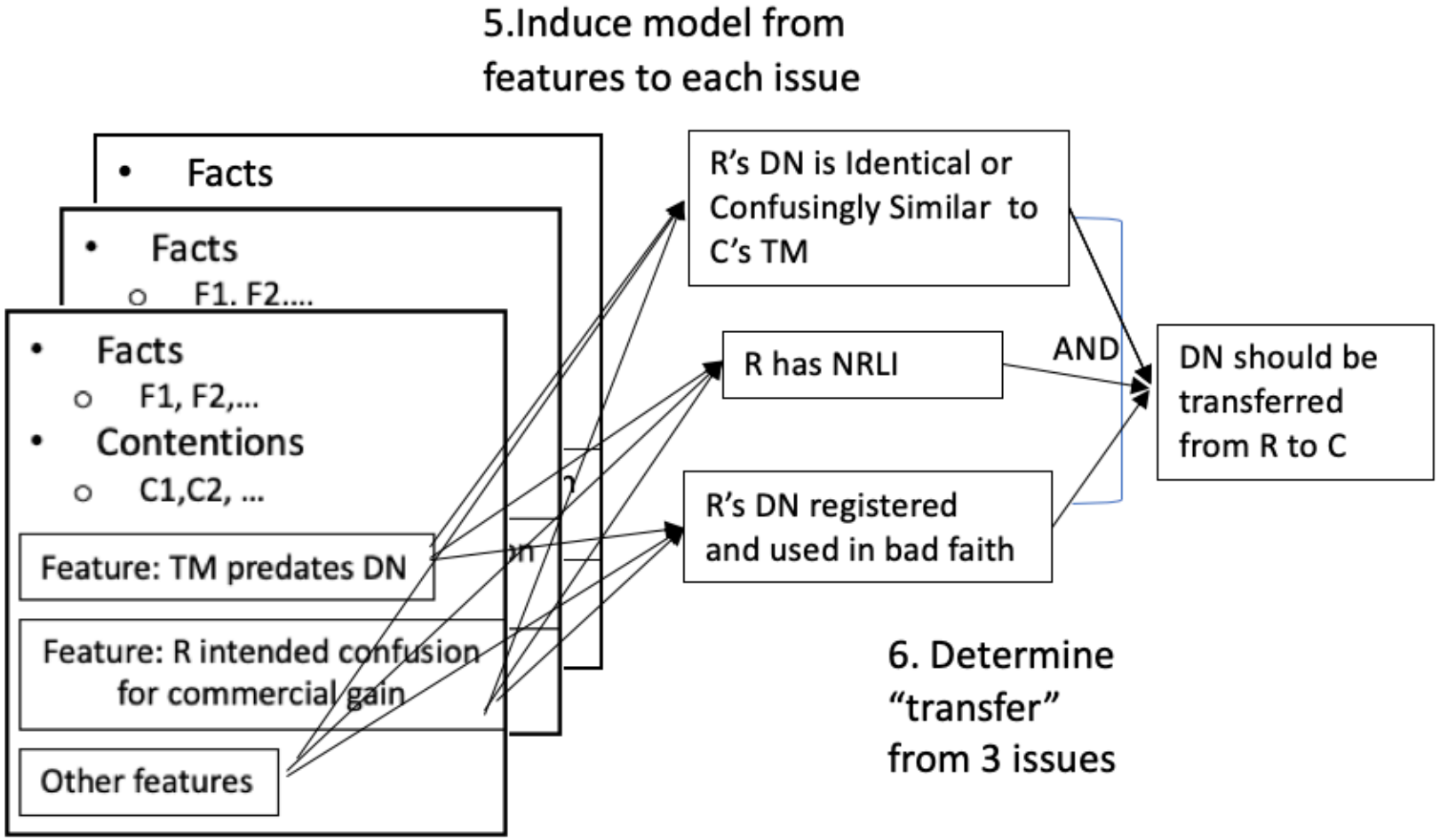
MITRE

SCALE: Semi-supervised Case Annotation for Legal Explanations



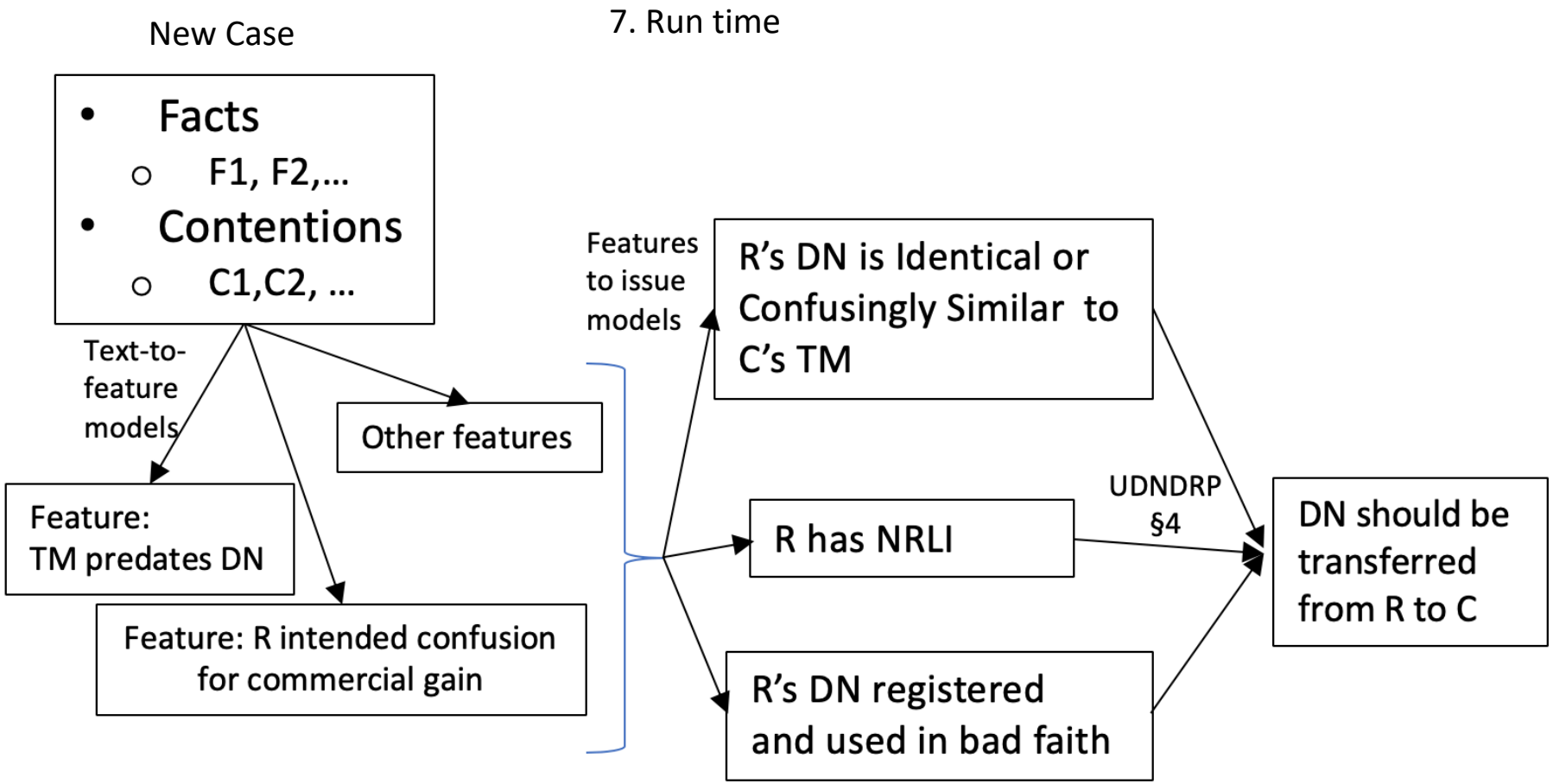
MITRE

SCALE: Semi-supervised Case Annotation for Legal Explanations



MITRE

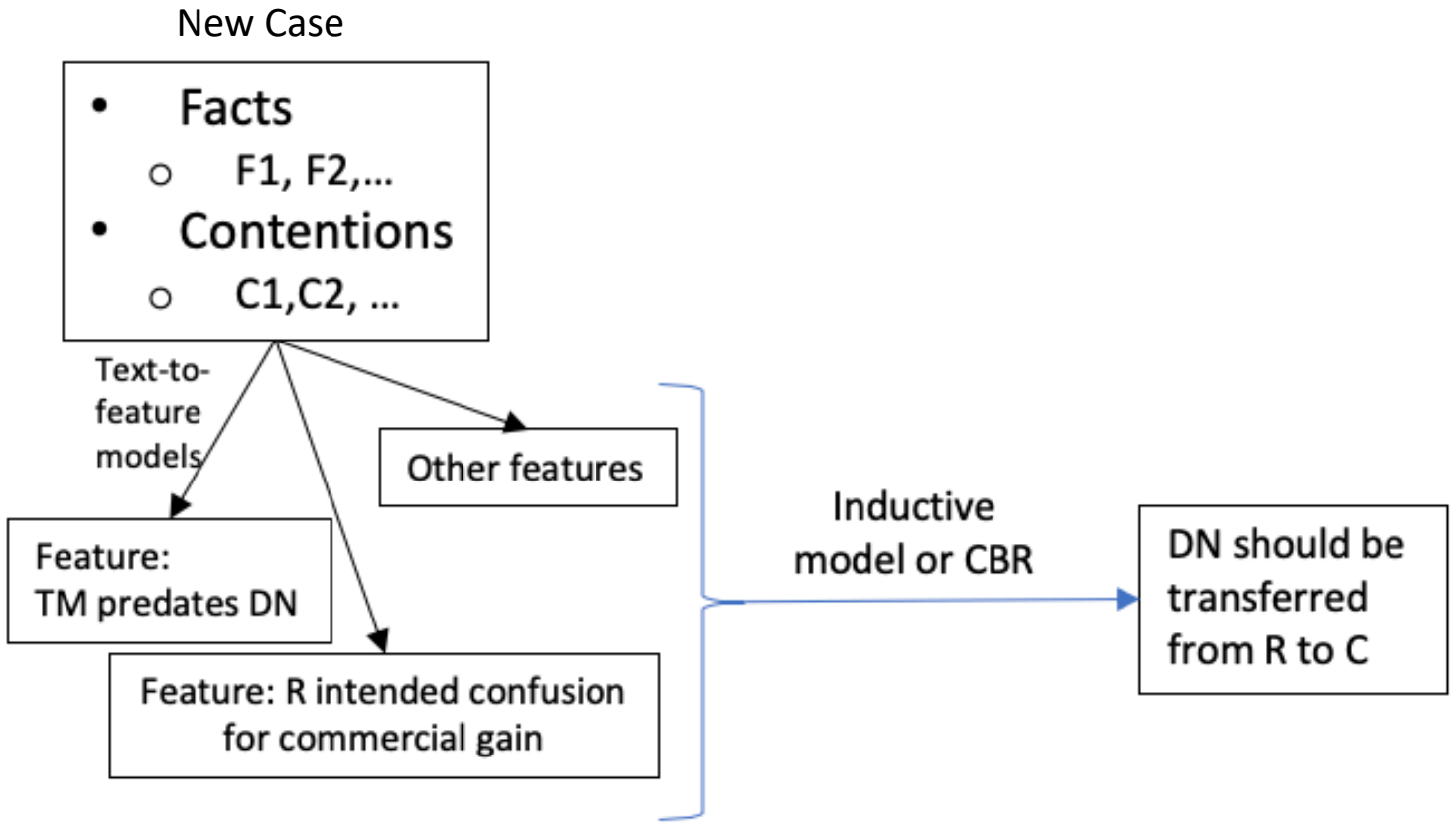
SCALE: Semi-supervised Case Annotation for Legal Explanations



MITRE

SCALE: Semi-supervised Case Annotation for Legal Explanations

7. Run time (simplified)

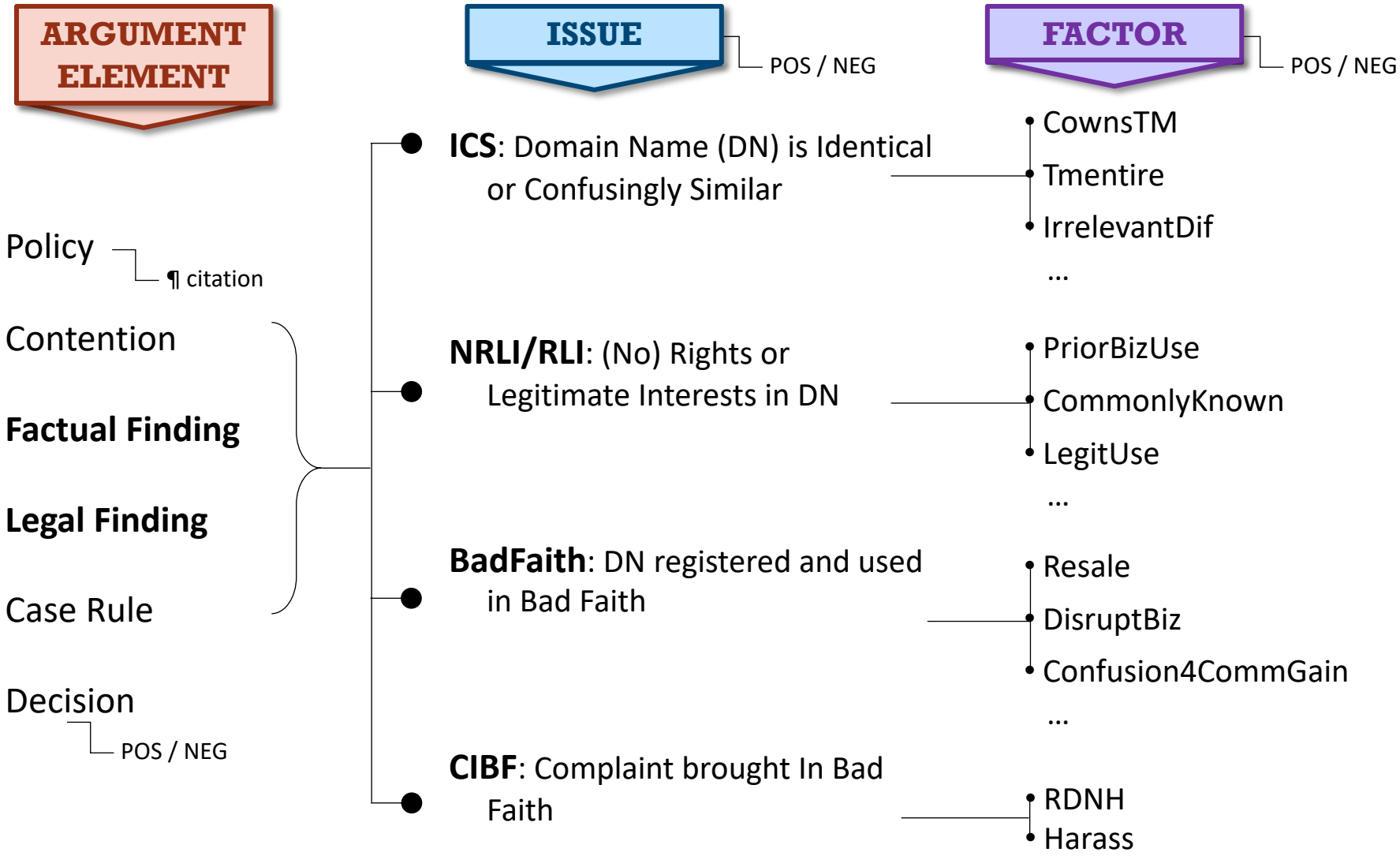


MITRE

Key Distinction

- **A SCALE factors are linguistic expressions (i.e., *text patterns)* that play similar roles in the justifications of multiple decisions and that are labeled with a common annotation**
- **This differs from HYPO/CATO factors, which are legally relevant fact patterns, however expressed**
- **Distinction is important because a semantic vector space is a metric over SCALE factors**
- **This permits the annotation of one factor to be transferred to other, sufficiently similar linguistic expressions (i.e., those in close proximity in vector space).**

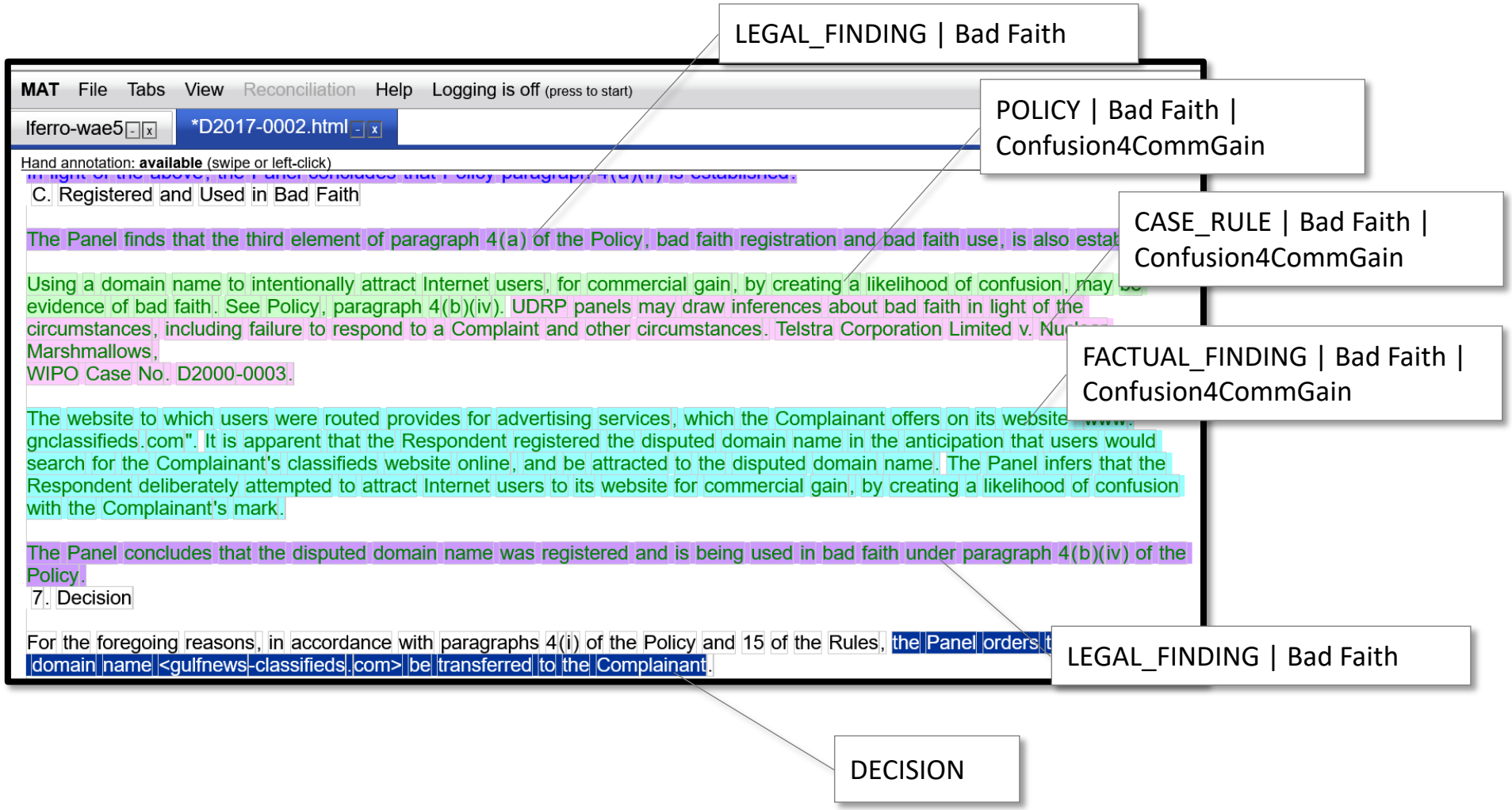
WIPO Annotation Scheme*



*Lisa Ferro, John Aberdeen, Karl Branting, Craig Pfeifer, Alexander Yeh, Amartya Chakraborty, Scalable Methods for Annotating Legal-Decision Corpora, *Workshop in Natural Legal Language Processing (NLLP) at NAACL 2019*, Minneapolis, MN, 7 June 2019.

MITRE

MITRE Annotation Toolkit (MAT)



Sample Annotations from WIPO Decisions

Case No	Text	Annotation
D2012-1430	in two instances the TURBOFIRE mark has been reproduced in a domain name, utilizing a dash "-" between the "turbo" and "fire" portion of the mark, which the Panel disregards as irrelevant under this element of the Policy	FACTUAL_FINDING-ICS-IrrelevantDiff
D2012-1430	The Panel thus finds that the disputed domain names are confusingly similar to the Complainant's registered trademarks	LEGAL_FINDING-ICS
D2012-1430	Additionally, as several of the disputed domain names are used to host online shopping websites offering products similar to those of the Complainant, from which the Respondent presumably generates revenue,	FACTUAL_FINDING-NRLI-LegitUse subissue-polarity=negative
D2012-1430	the Respondent clearly is not making any noncommercial or fair use of those domain names	LEGAL_FINDING-NRLI-LegitUse subissue-polarity=negative
D2012-1430	…the Respondent is clearly attempting to divert Internet traffic intended for the Complainant's website to its own for commercial gain by creating a likelihood of confusion as to the source or sponsorship of the Respondent's websites and products.	FACTUAL_FINDING-BadFaith-Confusion4CommGain
D2012-1430	Such use constitutes bad faith under paragraph 4(b)(iv) of the Policy.	LEGAL_FINDING-BadFaith-Confusion4CommGain
D2016-0534	The Complainant must have been aware that the Disputed Domain Name existed when it chose to register its UNIKS trademark.	FACTUAL_FINDING-CIBF-RDNH
D2016-0534	Taking into account all of the above the Panel has no hesitation in finding that the present case amounts to RDNH by the Complainant.	LEGAL_FINDING-CIBF-RDNH

MITRE

Preliminary Results

Prediction Task	Avg. AUC	Std. Dev.	Positive/Transfer Precision	Positive/Transfer Recall	Positive/Transfer F1	Negative/Non-Transfer Precision	Negative/Non-Transfer Recall	Negative/Non-Transfer F1
6.1 Predict Decisions from Mapped Tags	80.8%	0.01	97.3%	87.1%	91.9%	35.7%	74.5%	48.2%
6.2 Predict Decisions from Factual Findings Tags	89.6%	0.008	98.8%	90.2%	94.3%	46.7%	89%	61.2%
6.3 Predict "(ICS)-…" Finding from Case Text	60.1%		38.3%	42.3%	40.2%	80.9%	78.2%	79.5%
6.3 Predict "(NRLI)-…" Finding from Case Text	62.9%		31.7%	54.6%	40.1%	86.5%	71.1%	78.1%
6.3 Predict "Bad Faith-…" Finding from Case Text	63.5%		43.9%	48.9%	46.3%	81.5%	78.2%	79.8%

MITRE

SCALE: Observations

- **Scalable**
  - Just 25 annotated documents (0.156% of corpus) used in initial implementation
- **Factor polarity an unresolved issue**
  - Findings in proximity in semantic space don't always have the same polarity
  - Polarity is important for decision prediction and explanation
  - Current approach: separate polarity analysis stage, currently under construction
- **Key determinant of SCALE's overall effectiveness is annotation-projection accuracy, which depends on**
  - Genre
  - Vector space technology (rapidly improving)
  - Details of projection process, e.g., match threshold

Long-Term Research Issues

- **Is prediction using SCALE-derived features as accurate as prediction without features?**
  - Can we increase comprehensibility without decreasing accuracy?
- **Can decision support based on SCALE-derived features improve decision making?**
  - Do SCALE-derived features correspond to how human decision makers think about cases and justify decisions?
- **What characteristics of a domain determine the effectiveness of the SCALE approach**
  - Stereotypical facts and limited range of possible outcomes seem vital

Conclusion

- SCALE represents a new approach to optimizing the tradeoff between knowledge-engineering costs and explanation quality.
- May be appropriate for the relatively uniform high-volume cases that typically have the greatest backlogs
- Depends critically on semantic similarity assessment for annotation projection
- Existing factor-based CBR techniques as well as ML-based prediction can be leveraged without requiring manual annotation of entire corpus