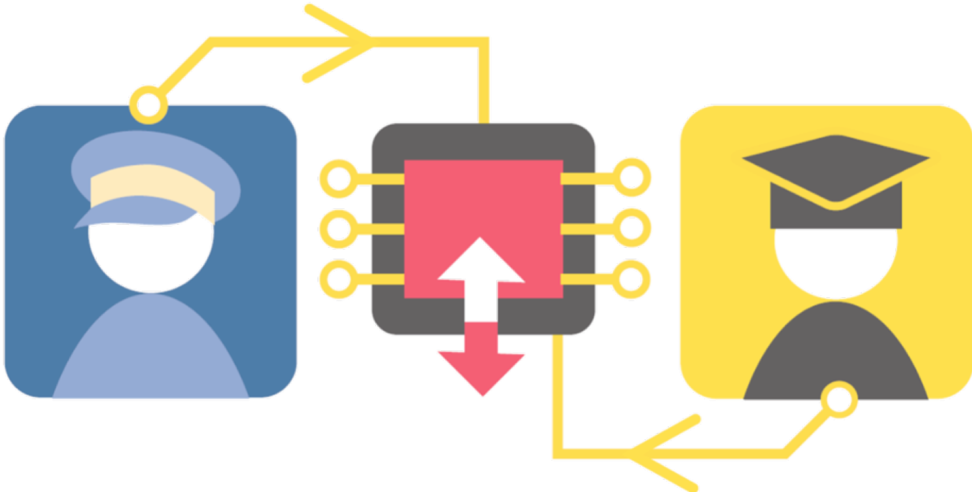
Utrecht University

14-11-19

AI and Law for the Dutch police: beyond mere prediction.



Floris Bex

What is Artificial Intelligence?

Systems that exhibit intelligent behaviour by analysing their environment and - with a certain degree of autonomy - taking action to achieve specific objectives.

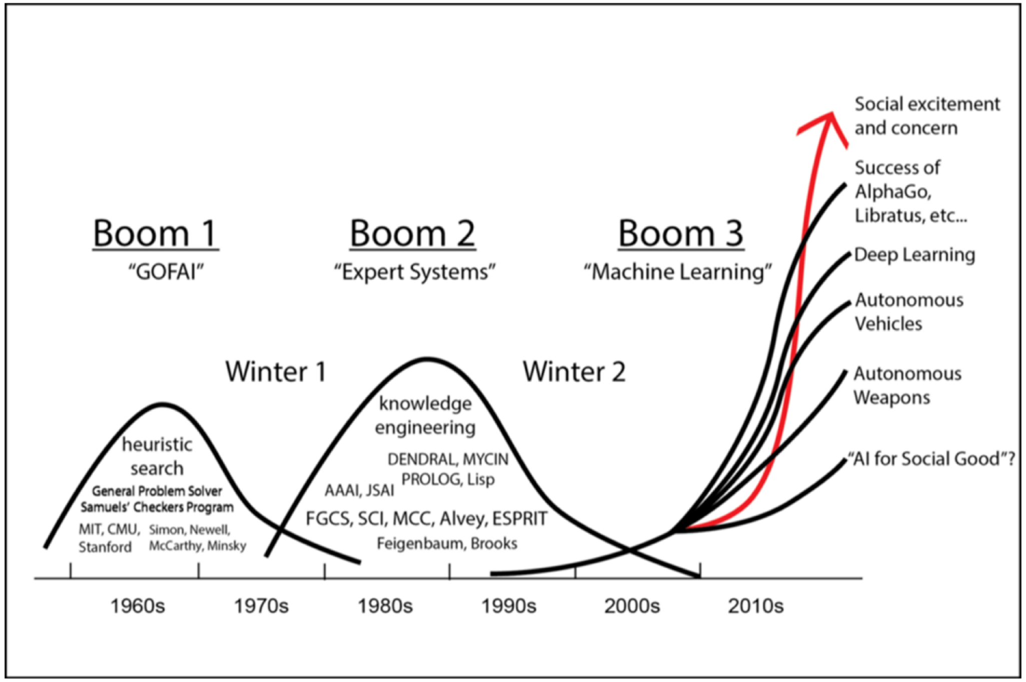
European Commission
Coordinated strategy on AI

What is Artificial Intelligence?

The AI in question, machine learning, is a technique for recognising patterns in relevant and preferably as complete as possible data files with the aim of discovering patterns in reality.

Minister of Justice to Parliament of the Netherlands

A brief history of AI



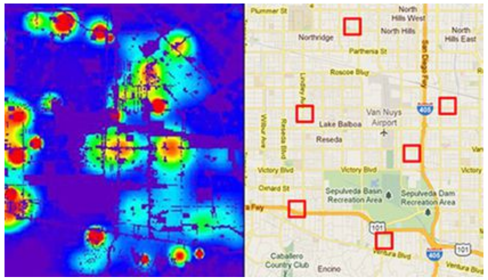
A (very) brief history of AI & Law

- 1977: Law as a computer programme
  - TAXMAN – executable law
- 1986: Expert system for legal reasoning
  - The British Nationality Act as a Logic Programme
- 1990-2000-now: Formal models of legal reasoning
  - Case-based reasoning, legal (rule-based) argumentation
- 2005-now: Machine learning
  - Legal information retrieval, text classification & summarization, QLP

Quantitative Legal Prediction (QLP)

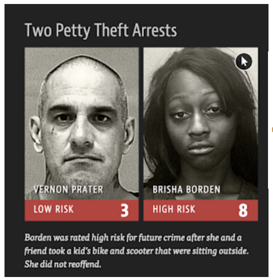
A general approach for predicting the behavior of the Supreme Court of the United States

Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective



PREDICTIVE POLICING

The Role of Crime Forecasting in Law Enforcement Operations



Fairness in Criminal Justice Risk Assessments: The State of the Art

Quantitative Legal Prediction in the news…

Sent to Prison by a Software Program's Secret Algorithms

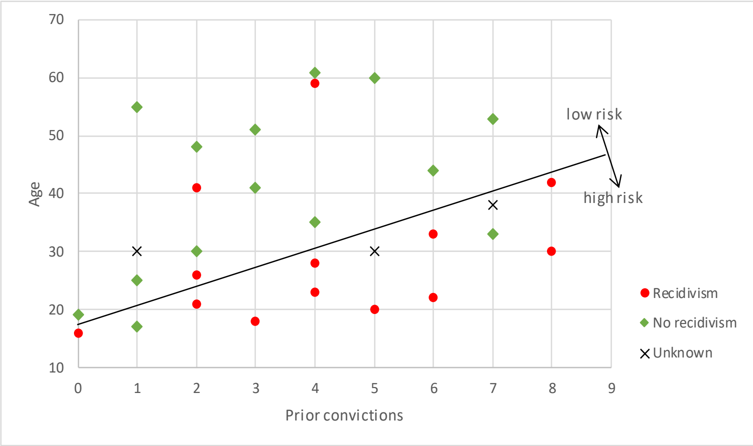
ESTONIA IS BUILDING A "ROBOT JUDGE" TO HELP CLEAR LEGAL BACKLOG

NEWS

Not robocop, but robojudge? A.I. learns to rule in human rights cases

Researchers have built an A.I. system that can predict rulings by the European Court of Human Rights with 79% accuracy

Quantitative Legal Prediction



- *Supervised machine learning* – finding patterns in previously decided "cases"
  - Case features (date, judges, gender defendant, previous crimes, number of crimes, avg. income…, text of verdict)
  - Decision labels (affirm – reverse, high risk - low risk, high crime – low crime)

Predicting SCOTUS decisions

- Given "meta-level" information about cases (justice(s), term, court of origin, lower court decision, issue area) classify cases as either AFFIRM or REVERSE

RESEARCH ARTICLE

A general approach for predicting the behavior of the Supreme Court of the United States

Daniel Martin Katz[1,2]*, Michael J. Bommarito II[1,2], Josh Blackman[3]

1 Illinois Tech - Chicago-Kent College of Law, Chicago, IL, United States of America, 2 CodeX - The Stanford Center for Legal Informatics, Stanford, CA, United States of America, 3 South Texas College of Law Houston, Houston, TX, United States of America

* dkatz3@kentlaw.iit.edu

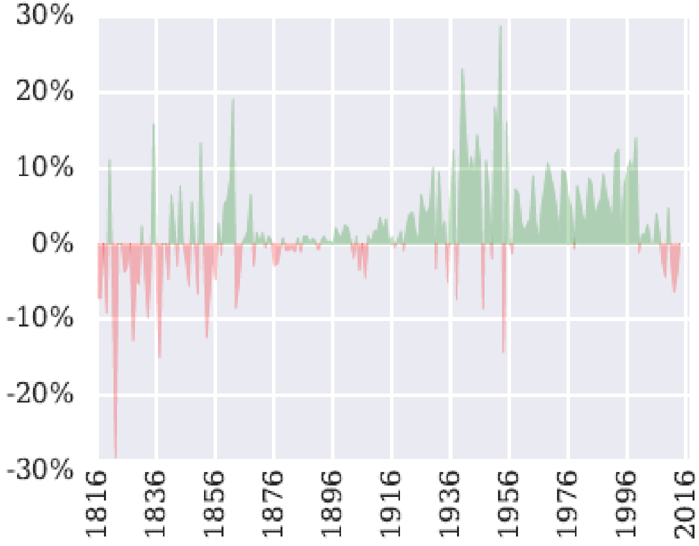
Abstract

Building on developments in machine learning and prior work in the science of judicial prediction, we construct a model designed to predict the behavior of the Supreme Court of the United States in a generalized, out-of-sample context. To do so, we develop a time-evolving random forest classifier that leverages unique feature engineering to predict more than 240,000 justice votes and 28,000 cases outcomes over nearly two centuries (1816-2015). Using only data available prior to decision, our model outperforms null (baseline) models at both the justice and case level under both parametric and non-parametric tests. Over nearly two centuries, we achieve 70.2% accuracy at the case outcome level and 71.9% at the justice vote level. More recently, over the past century, we outperform an *in-sample optimized* null model by nearly 5%. Our performance is consistent with, and improves on the general level of prediction demonstrated by prior work; however, our model is distinctive because it can be applied out-of-sample to the entire past and future of the Court, not a single term. Our results represent an important advance for the science of quantitative legal prediction and portend a range of other potential applications.

Predicting SCOTUS decisions

- The model correctly classifies 70.2% of the Court's decisions from between 1816 and 2014
  - Outperforms a coin flip (50% accuracy)
  - On average, outperforms “guess the most frequent decision in the 10 years before the case you’re classifying” (M = 10 baseline, 67.5% accuracy)
  - Does not outperform M=10 baseline on large periods before 1860 and after 2006

Katz. Et al.’s algorithm vs M=10 baseline
Green – Katz more accurate
Red – M=10 more accurate



A critical look at QLP (1)

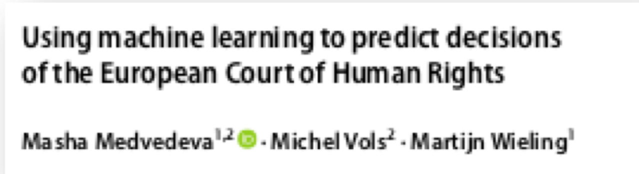
- Do we have practical uses for predicting the behavior of SCOTUS?
  - SCOTUS-watching - however the model does not perform well on “novel” decisions
  - Empirical legal research – however, there are better (more insightful) statistical techniques out there
- Is this really meaningful legal decision making?
  - No!

Predicting ECHR verdicts

- Given the text of decisions of the European Court of Human Rights, classify cases as either VIOLATION or NON-VIOLATION
- Not all section of the decision text in are included
    - Verdict section section often explicitly states, e.g., “a violation was found”

Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective

Nikolaos Aletras[1,2], Dimitrios Tsarapatsanis[3], Daniel Preoţiuc-Pietro[4,5] and Vasileios Lampos[2]



Using machine learning to predict decisions of the European Court of Human Rights

Masha Medvedeva[1,2] · Michel Vols[2] · Martijn Wieling[1]

Predicting ECHR verdicts

- Each case is represented as
  - (combinations of) words (N-grams)
    - [police], [police officer], [the case], [was caused], [rights], [June], [a Bulgarian], [her claim was]
  - clusters of N-grams (Topics)
    - Topic 4 - Treatment by state officials police, officer, treatment, police officer, July, ill, force, evidence, ill treatment, arrest, allegation, police station, subjected, arrested, brought, subsequently, allegedly, ten, treated, beaten

PROCEDURE

1. The case originated in an application (no. 35355/08) against the Republic of Bulgaria lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by a Bulgarian national, Ms Gana Petkova Velcheva ("the applicant"), on 30 June 2008.
2. The applicant was represented by Mr M. Ekimdzhiev and Ms G. Chernicherska, lawyers practising in Plovdiv. The Bulgarian Government ("the Government") were represented by their Agent, Ms Y. Stoyanova, of the Ministry of Justice.
3. The applicant alleged that the authorities had failed to comply with a final court judgment allowing her claim for restitution of agricultural land.
4. On 7 May 2013 the application was communicated to the Government.

THE FACTS

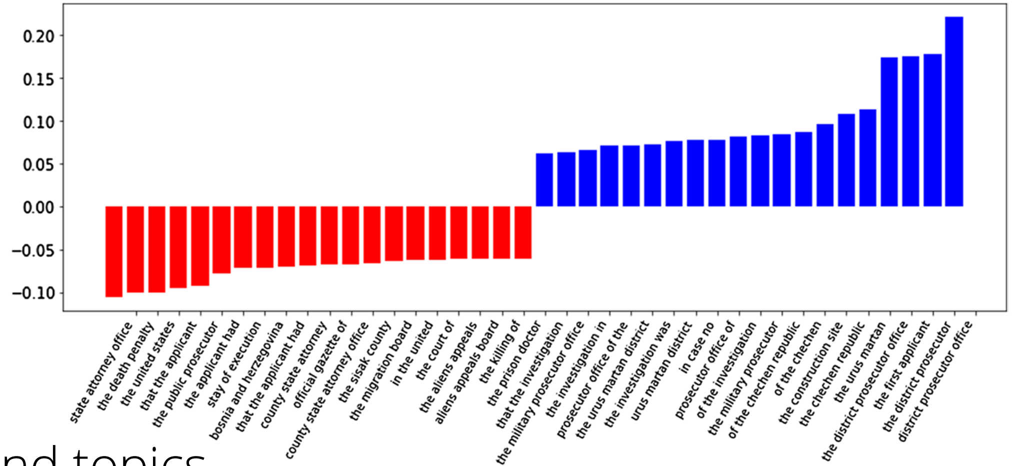
I. THE CIRCUMSTANCES OF THE CASE

5. The applicant was born in 1927 and lives in the village of Ribaritsa.
6. Her father, of whom she is the sole heir, owned agricultural land in the area surrounding the village which was incorporated into an agricultural cooperative at the beginning of the 1950s.
7. In 1991, following the adoption of the Agricultural Land Act ("the ALA", see paragraph 17 below), the applicant applied for the land's restitution.
8. By a decision dated 10 March 1999 the land commission dealing with the case refused to restore her rights to two plots of 900 and 2,000 square metres respectively, noting that sheep pens had been built on them by the agricultural cooperative. It held that the applicant was entitled to compensation in lieu of restitution.

Predicting ECHR verdicts

- Aletras et al.'s model correctly classifies 79% of ECHR decisions on 580 cases for article 3, 6, 8.
- Medvedeva et al.'s model correctly classifies 77% of ECHR decisions on 1942 cases for article 3, 6, 8.
  - Medvedeva does not include ”law” section
  - Both outperform “always guess violation” on the dataset (50% accuracy)
  - Both would not outperform “always guess violation” on full ECHR dataset (+43k cases, ±84% violation cases)

Predicting ECHR verdicts



- Indicates predictive words and topics per article
- Meaningful: “police”, “ill treatment”, “the Chechen republic”
- But also: “June”, “ten”, “the applicant had”

Topic	Label	Words	*w*
		Top-5 Violation	
4	Positive State Obligations	injury, protection, ordered, damage, civil, caused, failed, claim, course, connection, region, effective, quashed, claimed, suffered, suspended, carry, compensation, pecuniary, ukraine	13.50
10	Detention conditions	prison, detainee, visit, well, regard, cpt, access, food, situation, problem, remained, living, support, visited, establishment, standard, admissibility merit, overcrowding, contact, good	11.70
3	Treatment by state officials	police, officer, treatment, police officer, July, ill, force, evidence, ill treatment, arrest, allegation, police station, subjected, arrested, brought, subsequently, allegedly, ten, treated, beaten	10.20
		Top-5 No Violation	
8	Prior Violation of Article 2	june, statement, three, dated, car, area, jurisdiction, gendarmerie, perpetrator, scene, June applicant, killing, prepared, bullet, wall, weapon, kidnapping, dated June, report dated, stopped	−12.40
19	Issues of Proof	witness, asked, told, incident, brother, heard, submission, arrived, identity, hand, killed, called, involved, started, entered, find, policeman, returned, father, explained	−15.20
13	Sentencing	sentence, year, life, circumstance, imprisonment, release, set, president, administration, sentenced, term, constitutional, federal, appealed, twenty, convicted, continued, regime, subject, responsible	−17.40

A critical look at QLP (2)

- Do we have practical uses for predicting the behavior of ECHR?
  - ECHR-watching – but prediction acc. too low
  - NLP in Empirical legal research – explainability/semantic interpretability is a problem
- Is this really meaningful legal decision making?
  - No! (but Medvedeva and to a lesser extent Aletras don't claim it is).

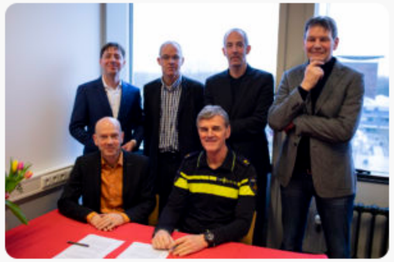
Predicting the outcome of traffic fine appeals

- Appeal to traffic fine goes to court via Public Prosecutor
  - Public prosecutor adds opinion before sending it to court
- Predict court decision (justified, not justified, change, not admissible) given the text of the prosecutor
  - Accuracy around 65% (random guessing 25%, guessing not justified 45%).
- System also provides similar cases
  - Based on doc2vec similarity

QLP – what is it good for?

- Routine cases
    - Help with triage -> prioritizing cases, online-case-assessor
- However:
    - Still little connection to the law, “correlation machines”
    - Explaining decisions at a meaningful level
- Can it help the police?

Police Lab AI



Utrecht University

University of Amsterdam

POLITIE

At the forefront of the developments in AI

TNO innovation for life

Erasmus Universiteit Rotterdam

TU/e Technische Universiteit Eindhoven University of Technology

Universiteit Twente.

TU Delft

AI for the Dutch Police

- Practical uses of AI for the Dutch police
  - Searching through open source intel, images, etc.
  - Automating routine cases
- Accurate, Efficient , Transparent, Controllable, Contestable

AI for the Dutch police

- For some types of tasks predictive machine learning is the best solution
  - Recognizing guns
  - Recognizing online threaths

AI for the Dutch police – beyond prediction

- For other types of tasks, using only machine learning is not a good idea
  - (Autonomously) making (legally relevant) decisions based on input

- QLP does not make decisions
- QLP is not (very) transparent, controllable

AI for the Dutch police – beyond “mere” prediction

- Handling citizen crime reports

Online trade fraud

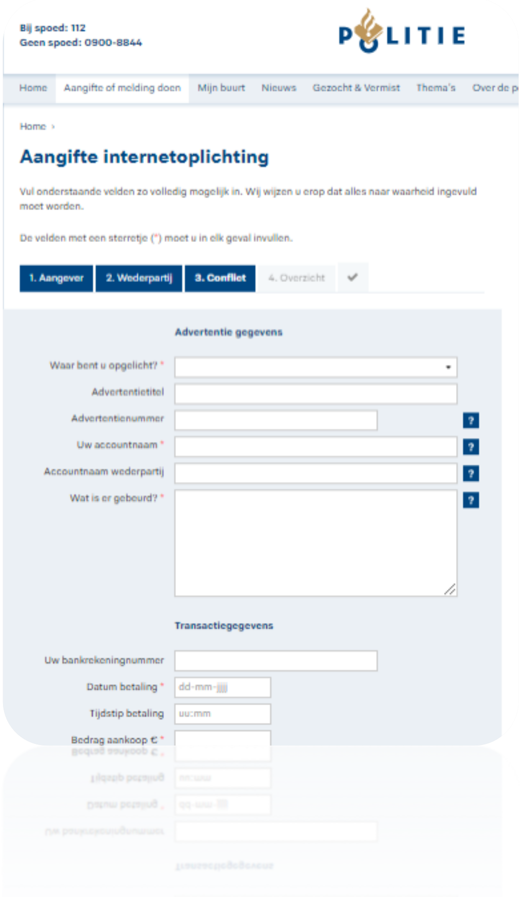
- Online trade fraud
  - Fraud on eBay, internet forums, fake websites
- 40,000 reports filed per year

- Legal background: article 326 of Dutch Criminal Code
  - Take some good or money away from someone, while "*misleading through false contact details, deceptive tricks or an accumulation of lies*"

System for handling citizen reports on online trade fraud

- Submitted online
- Given report, decide:
  - Fraud
  - No fraud
  - More information needed → ask questions "was a product delivered?", "did you wait for the product?", "Did the seller use a false location?"

- Currently done by humans



Why not handle reports using QLP?

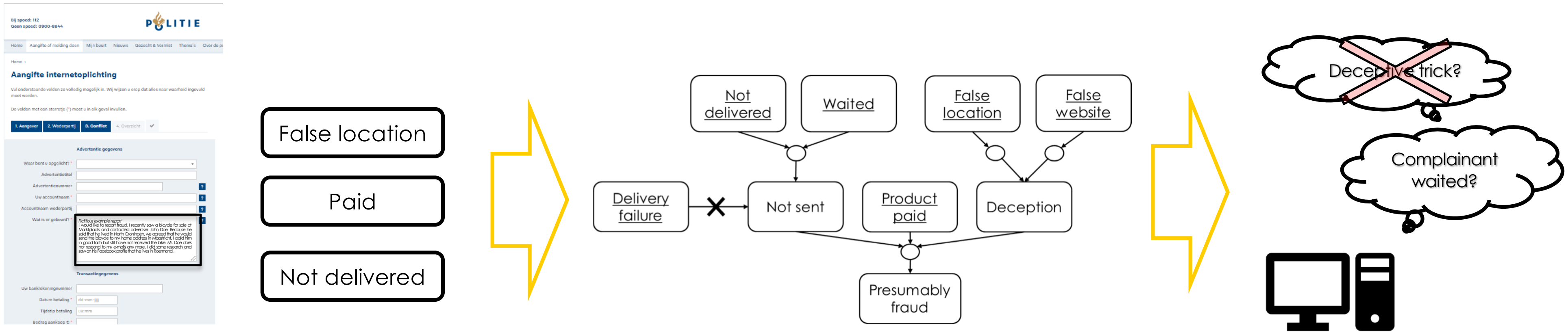
- Given input form (& text), classify “fraud” or “not fraud”
  - Low precision & recall ($F_1 \approx 0.59$)
  - Unclear what the main factors are for the decision
- Given input form (& text), decide on a question
  - Not enough data, data too noisy (human decision-makers are not very consistent)

Kos, Schraagen, Brinkhuis, Bex (2017) Classification in a Skewed Online Trade Fraud Complaint Corpus. *29th Benelux Conference on Artificial Intelligence (BNAIC 2017)*

Beyond mere prediction

- Extract basic observations from the form/text
  - waited/not-waited, false-location/noy-false-location
- Use rule-based argumentation to determine fraud/not-fraud
  - Rules based on legal rules & reasoning
- Ask the right questions based on info that is missing to draw a conclusion
- Determine a strategy for efficiently asking questions
  - Formal heuristic for short dialogues
  - Reinforcement learning for efficient strategies

System architecture



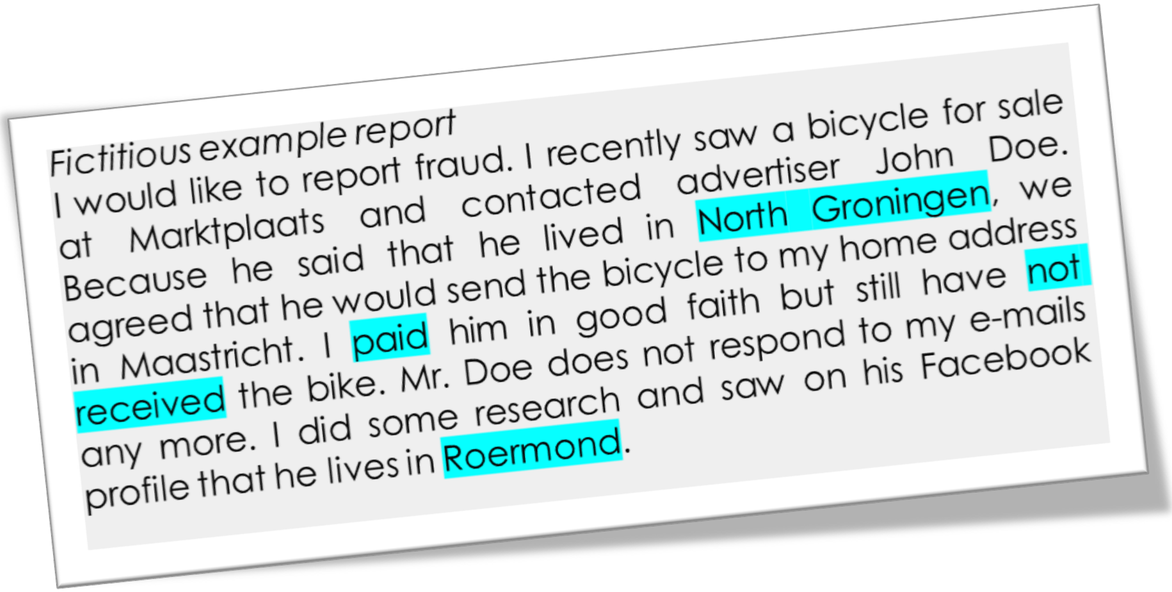
Information Extraction

Argumentation

Question Policy

Information extraction

Extract observations from free text



Fictitious example report

I would like to report fraud. I recently saw a bicycle for sale at Marktplaats and contacted advertiser John Doe. Because he said that he lived in North Groningen, we agreed that he would send the bicycle to my home address in Maastricht. I paid him in good faith but still have not received the bike. Mr. Doe does not respond to my e-mails any more. I did some research and saw on his Facebook profile that he lives in Roermond.

False location

Paid

Not delivered

Information extraction

- Named Entity Recognition
  - Enriched version of NER module in a well-known Dutch NLP package (Frog; 92% precision, 82% recall)

- Relation extraction using LSTMs (94-99% accuracy)

- Classifying texts according to observations
  - Practical solution, ± 90% accuracy (lower recall)

Schraagen, Brinkhuis, Bex, (2017) Evaluation of Named Entity Recognition in Dutch online criminal complaints. *Computational Linguistics in the Netherlands Journal*, *7*.

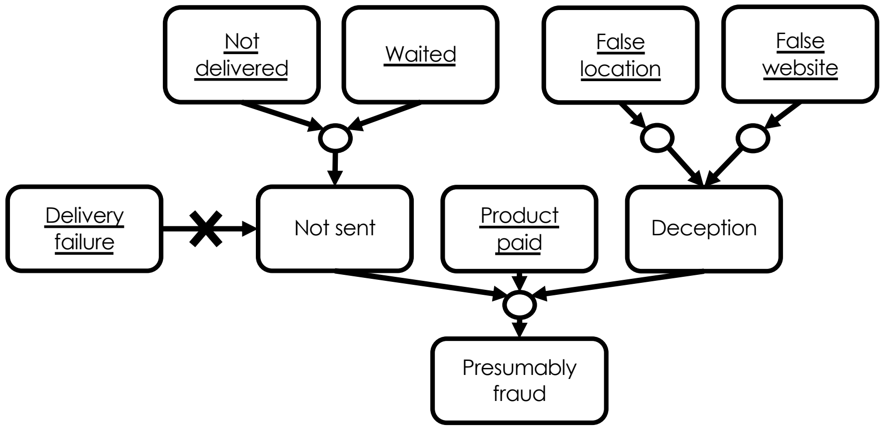
Schraagen, Bex (2019) Extraction of semantic relations in noisy user-generated law enforcement data, *Proceedings of the 13th IEEE Internati Conference on Semantic Computing (ICSC 2019)*.

Argumentation

- Rule-based argumentation consisting of:
  - Facts observable in text of complaint
  - Inference rules
- Based on the law and expert knowledge
  - art 326 Criminal Code, case law, expert police knowledge

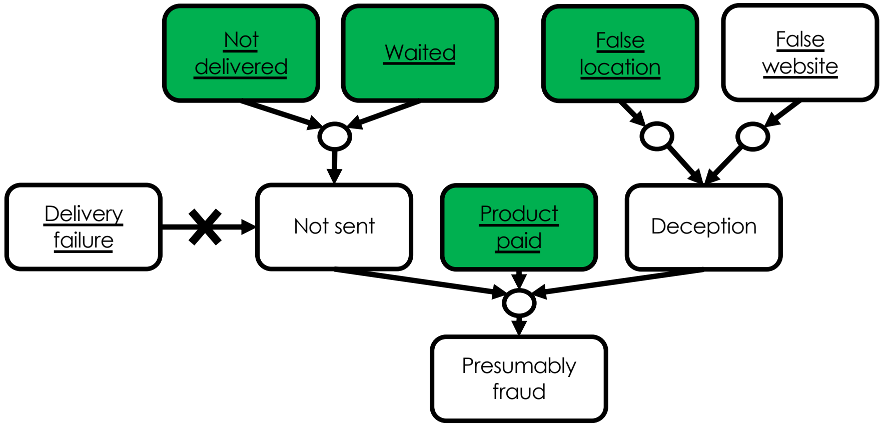
Argumentation example

- Scenario:



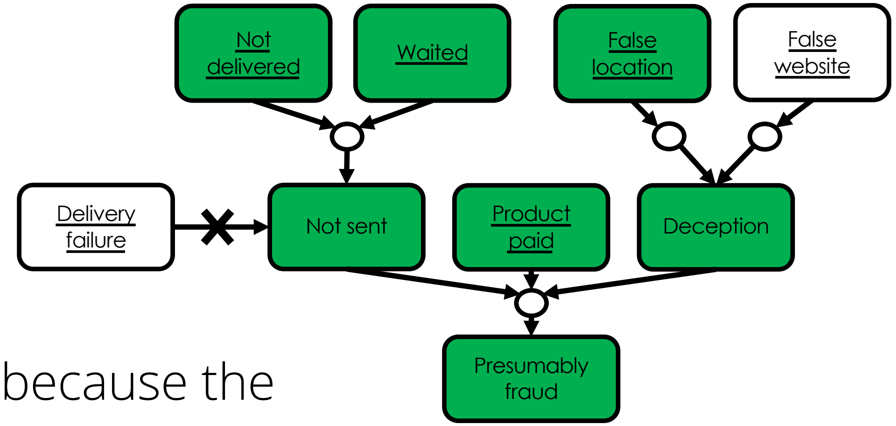
Argumentation example

- Scenario:
  - False location, paid, waited, not delivered



Argumentation example

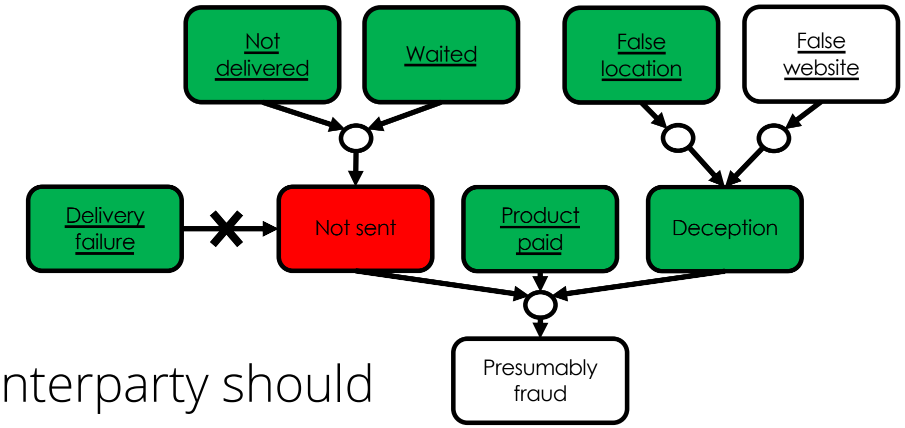
- Scenario:
  - False location, paid, waited, not delivered
- Conclusion: presumably fraud



- Explanation:
  “This case is presumably fraud, because the counterparty provided a false location and did not send the produc”

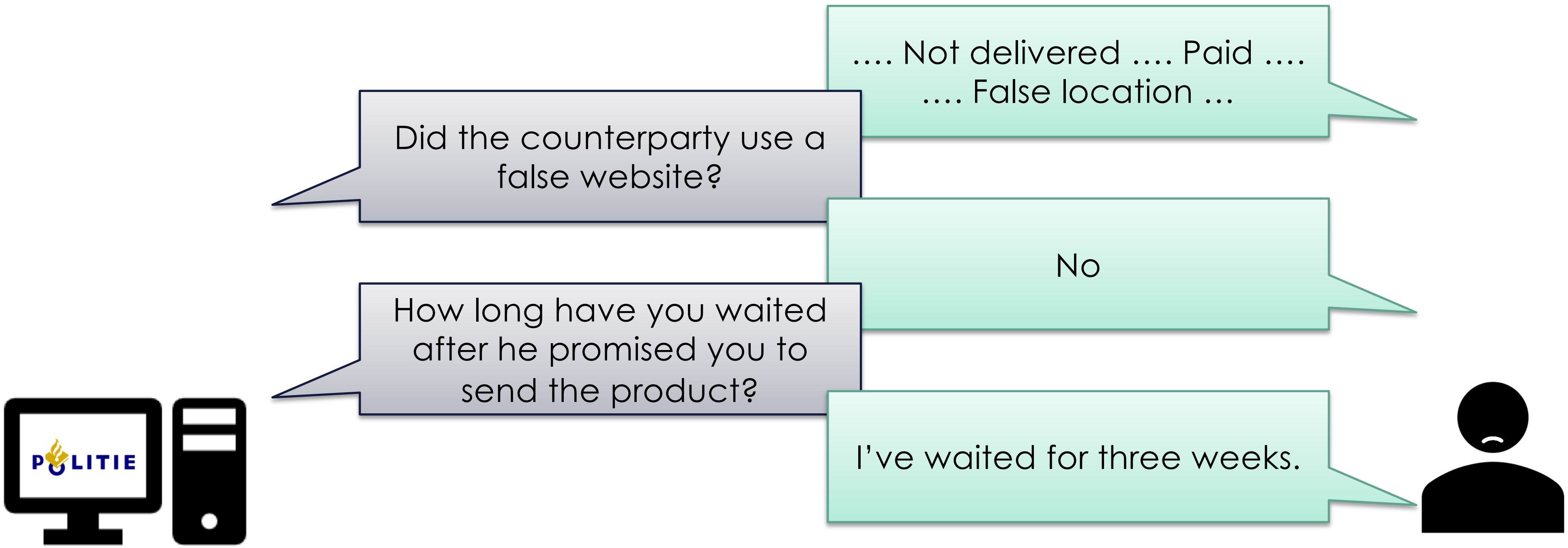
Exceptions to rules

- Scenario:
  - False location, paid, waited, delivery failure, deception
- Cannot conclude presumably fraud!



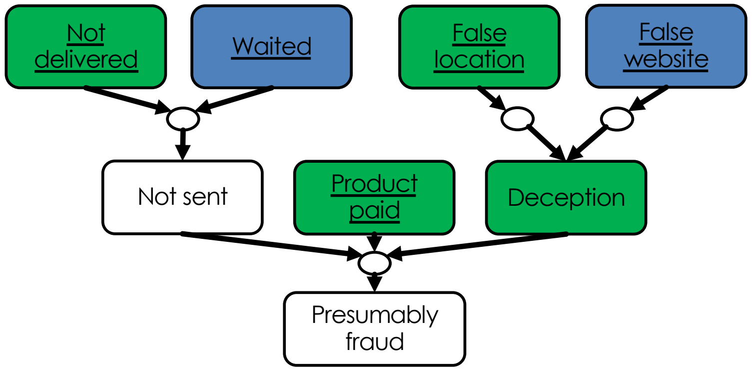
- Explanation: “For fraud the counterparty should not have sent the product. In this case, however, there was a delivery failure.”

Dialogue example



Question policy determination

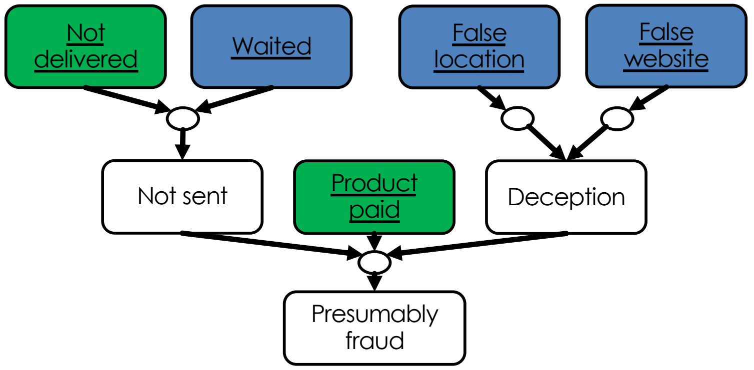
- Goal: Given observations, find next question(s)
- Efficiently reaching a conclusion
  - Only ask relevant questions
- Approximation algorithm for determining relevant questions



Testerink, Odekerken, Bex (2019) A Method for Efficient Argument-based Inquiry, *13th International Conference on Flexible Query Answering Systems (FQAS 2019)*. Lecture Notes in Artificial Intelligence.

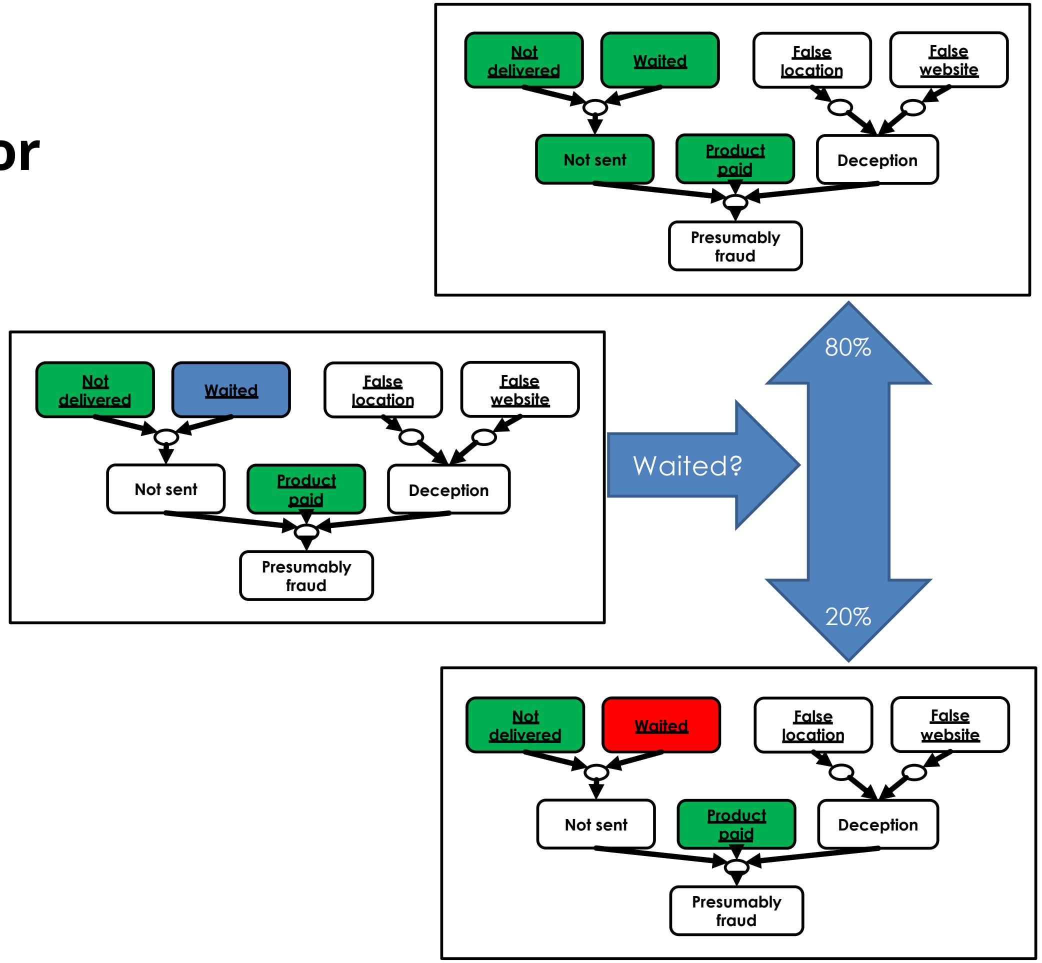
Question policy learning

- Goal: Given observations, find next question(s)
- Efficiently reaching a conclusion
  - Include probabilities of user responses
  - Hypothesizing over all possible future questions?



Reinforcement learning for question policies

- Learn from user interaction
- Reward function:
  - Reaching conclusion → high reward
  - Each action → small penalty



Conclusions – AI for handling reports

- Combining symbolic (“old-fashioned”) and sub-symbolic techniques for legal decision-making
- Drawback of QLP
  - Not controllable, contestable, transparent, (accurate) -> use argumentation & more fine-grained subsymbolic machine learning
  - System doesn’t “act” -> use argumentation, dialogue
- Drawback of symbolic AI
  - Interpreting the law -> knowledge acquisition and validation
  - Handling unstructured data -> use subsymbolic NLP

Utrecht University

Questions?