# Towards argument-based explanatory dialogues: from argument mining to (explanatory) argument generation

Serena Villata

KnOD 2022 — April 26th, 2022

# High quality explanations for AI deliberations

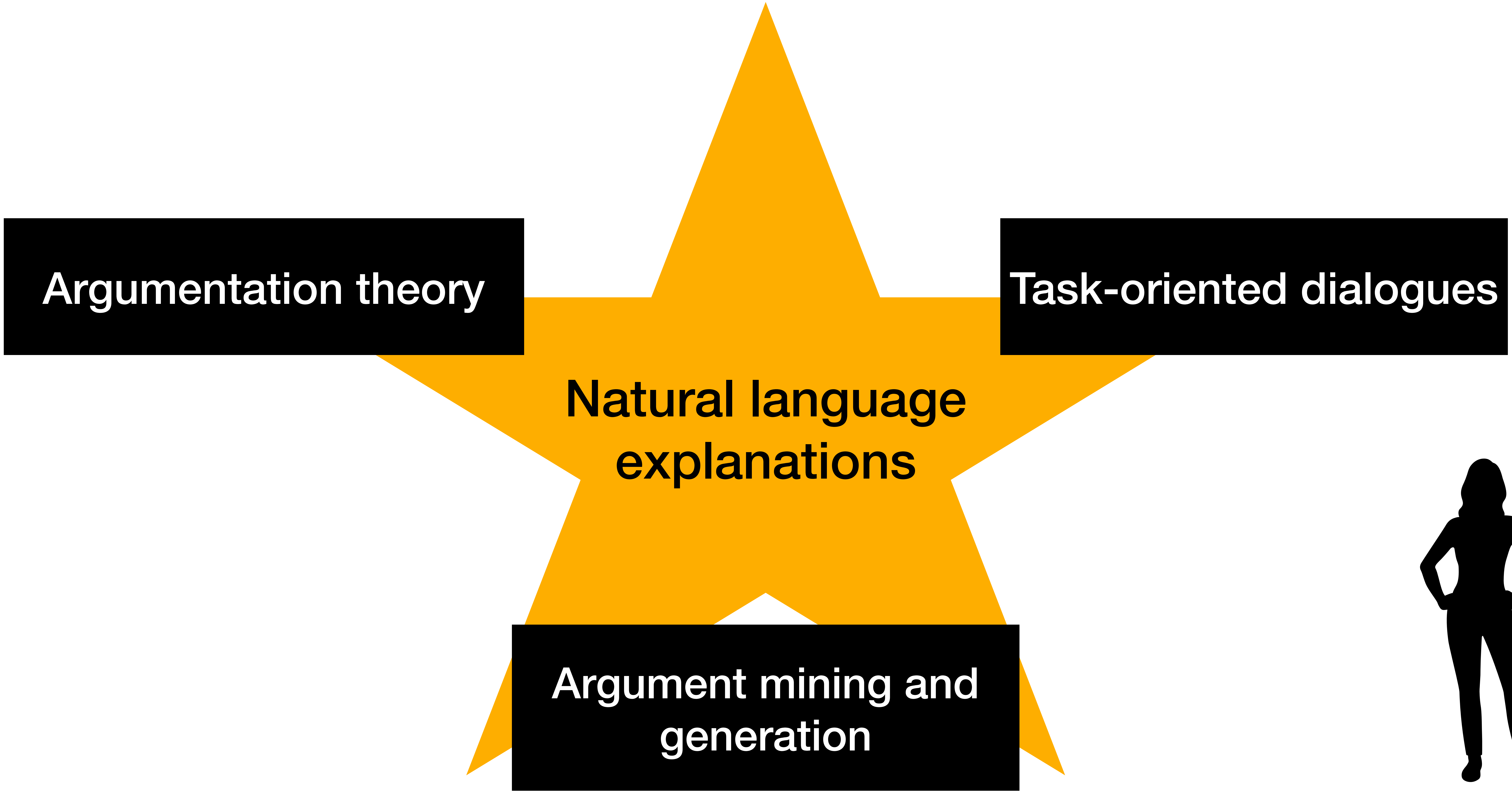
## Challenges

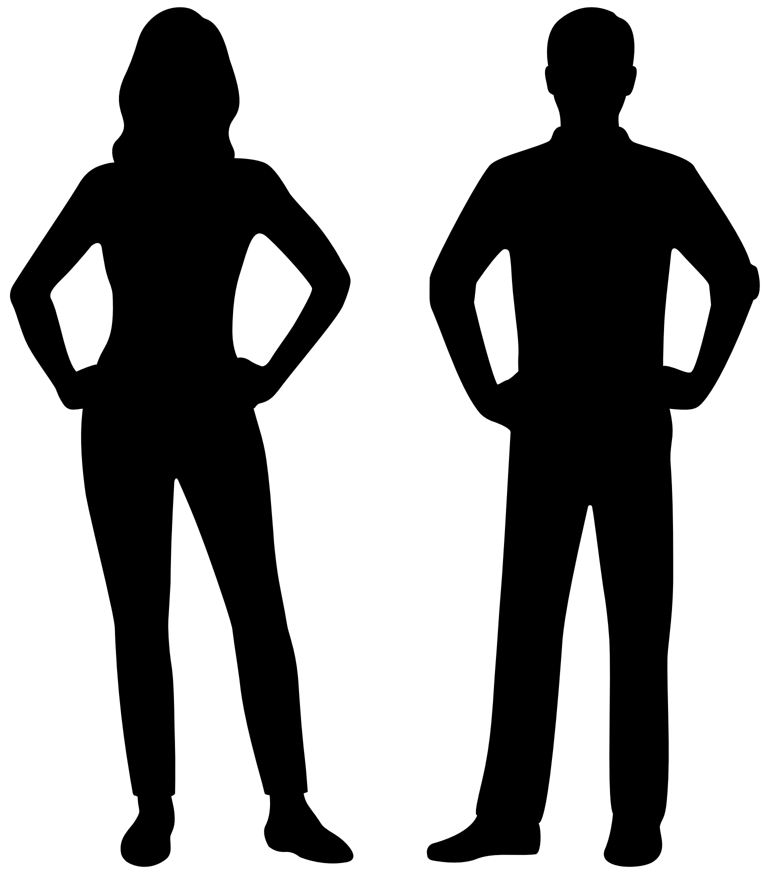
- proper level of generality/specificity of the explanations
- reference to specific elements that have contributed to the deliberation
- analytic statements
- use of additional knowledge (common-sense knowledge, domain ontologies, knowledge bases, knowledge graphs, …)
- use of examples (e.g., from the data the prediction is produced on)
- evidence supporting negative hypotheses

**Formulate the explanation in a clearly interpretable, and possibly convincing, way**

# Natural language explanations

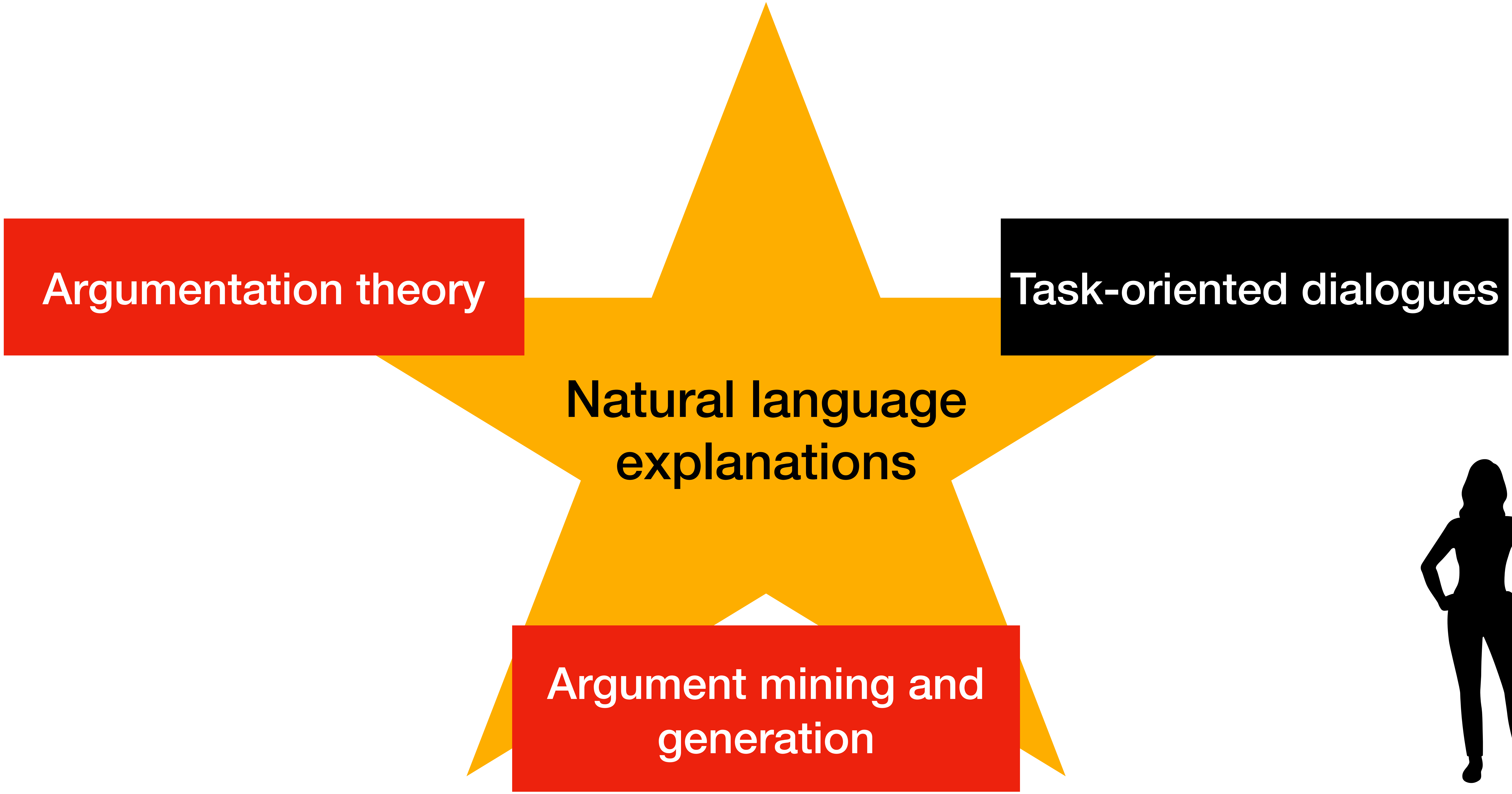
## Key features

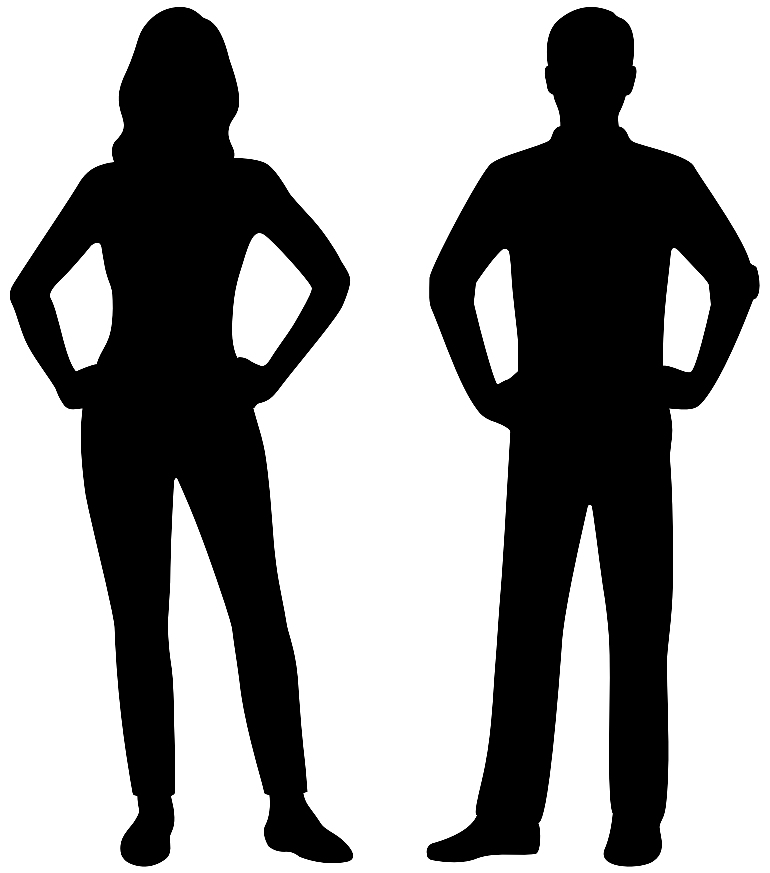


# Natural language explanations

## Key features



# Explanatory dialogues

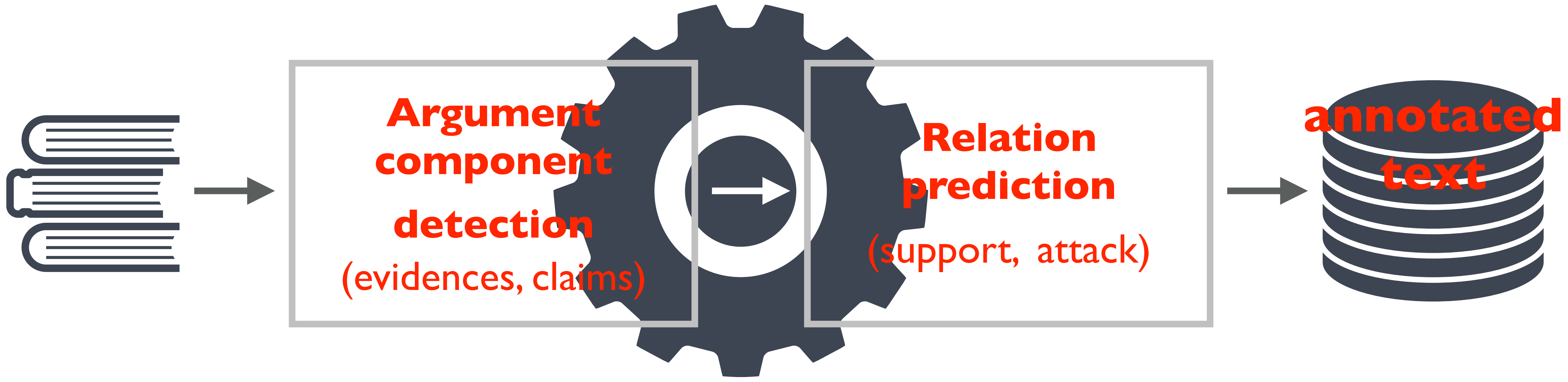
## Argumentation theory

- Argumentation as reasoning-in-interaction
- Arguments need not only be rational, but **“manifestly” rational** (Johnson (2000))
- Arguers can see for themselves the rationale behind inferential steps taken
- <u>In explanations</u>
  - an agent accepts the conclusion but queries premises “OK that the diagnosis you proposed is D, but why?”
  - pragmatic goal is understanding, typically reached via causal reasoning

# Explanatory argumentative dialogues

## From argument mining to generation through extractive summaries

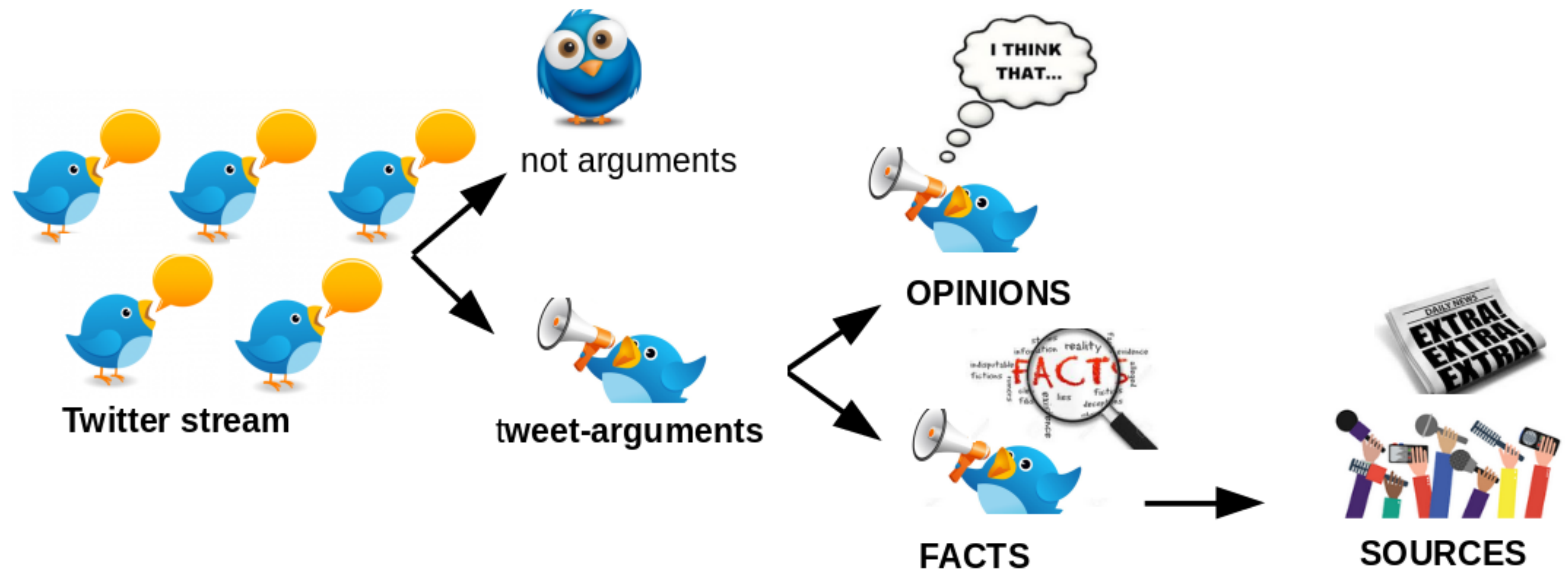
- The **task** of analysing discourse on the pragmatics level and applying a certain argumentation theory to model and automatically analyze the data at hand.
- Providing structured data for computational models of argument.
- Large resources of natural language texts: user-generated arguments on blogs, product reviews, newspapers,...
- Computational linguistics and machine learning advances.
- Argument mining IS NOT opinion mining.



# Argument Mining

# Argument mining

## Twitter (LREC16, EMNLP17)




**Tasks**: argument detection (binary classification), factual vs. opinion classification, source identification.

**Data**: DART `[Bosc et al., LREC2016]`, thread *#Grexit* (987 tweets) + 900 tweets from *#Brexit*.
2 annotators, IAA: $\kappa$=0.767 (1st task, 100 tweets), $\kappa$=0.727 (2nd task, 80), Dice=0.84 (3rd task, whole dataset)).

**FACT:** *The Guardian: Greek crisis: European leaders scramble for response to referendum no vote. http://t.co/cUNiyLGfg3*
**OPINION:** *Trump is going to sell us back to England. #Brexit #RNCinCLE*

**Method and results**:

| Task | Method | Features | Results |
|---|---|---|---|
| argument detection | LR | lex., Twitter, synt., sem., sent. | **0.78** |
| factual/opinion classification | LR | lex., Twitter, synt., sem., sent. | **0.80** |
| source identification | Matching + heuristics | | **0.67** |

# Mining argumentative structures from clinical trials

## AI in Medicine 2021, ECAI20, COMMA2020, IJCAI19

**Task**: argument component detection (evidence, claims) and relation prediction (attack, support).

**Data**: 4073 argument components (2808 evidence, 1265 claims). IAA: 3 ann., 10 abs., Fleiss' $\kappa = 0.72$ (arg. comp.) and $\kappa = 0.68$ (c/e) – 2601 argument relations (2259 supports, 342 attacks). IAA: 3 ann., 30 abs., Fleiss' $\kappa = 0.62$.
**Topics**: neoplasm, glaucoma, hepatitis, diabetes, hypertension.

[*The diurnal intraocular pressure reduction was significant in both groups ($P < 0.001$)*]$_1$. [*The mean intraocular pressure reduction from baseline was 32% for the latanoprost plus timolol group and 20% for the dorzolamide plus timolol group*]$_2$. [*The least square estimate of the mean diurnal intraocular pressure reduction after 3 months was -7.06 mm Hg in the latanoprost plus timolol group and -4.44 mm Hg in the dorzolamide plus timolol group ($P < 0.001$)*]$_3$. This study clearly showed that **[the additive diurnal intraocular pressure-lowering effect of latanoprost is superior to that of dorzolamide in patients treated with timolol]$_1$**.


Review > Infez Med. 2020 Ahead of print Jun 1;28(2):198-211.

**Update on treatment of COVID-19: ongoing studies between promising and disappointing results**

Silvano Esposito [1], Silvana Noviello [1], Pasquale Pagliano [1]

Affiliations + expand

PMID: 32335561

Free article

**Abstract**

The COVID-19 pandemic represents the greatest global public health crisis since the pandemic influenza outbreak of 1918. We are facing a new virus, so several antiviral agents previously used to treat other coronavirus infections such as SARS and MERS are being considered as the first potential candidates to treat COVID-19. Thus, several agents have been used by the beginning of the current outbreak in China first and all over the word successively, as reported in several different guidelines and therapeutic recommendations. At the same time, a great number of clinical trials have been launched to investigate the potential efficacy therapies for COVID-19 highlighting the urgent need to get as quickly as possible high-quality evidence. Through PubMed, we explored the relevant articles published on treatment of COVID-19 and on trials ongoing up to April 15, 2020.

**Method**: Gated Recurrent Unit + Conditional Random Fields, sciBERT.

**Results** : evidence (F1: **0.92**), claim (F1: **0.88**), arg. comp. (F1: **0.87**) – relation classification F1: **.68**.

**Collaborations:**
INSERM, CHU Nice

# Mining argumentative structures from clinical trials

## AI in Medicine 2021, ECAI20, COMMA2020, IJCAI19

# ACTA

http://ns.inria.fr/acta/

## Argumentative Clinical Trial Analysis

# Mining political arguments

## COLING20, IJCAI19 demo, ACL19 short, AAAI18

39 political debates from the last 50 years of US presidential campaigns (29k argument components)

Argument Mining for fallacies detection

**Task**: argument component detection (claim, premises) and relation classification (attack, support).

**Data**: 29521 argument components (16087 claims and 13434 premises) and 25012 relations (3723 attacks and 21289 supports). IAA: 3 ann., moderate/faire agreement.

**Method**: LSTM + Fine tuned BERT

**Results**: evidence (F1: **0.72**), claim (F1: **0.69**), argument components (F1: **0.84**), relation classification (F1: **0.68**)

Collaborations: Univ. of Luxembourg

# Mining political arguments

## COLING20, IJCAI19 demo, ACL19 short, AAAI18

39 political debates from the last 50 years of US presidential campaigns (29k argument components)

Argument Mining for fallacies detection

**Claim2**

The control of these weapons is the single major responsibility of a President, and to cast out this commitment of all Presidents

**Claim1**

This is a formidable force

**Premise4**

*We have thousands, equivalent of megaton, or million tons, of TNT warheads*

**Premise3**

*That's one major war explosion in a warhead*

**Premise1**

*Some of these weapons have 10 megatons of explosion*

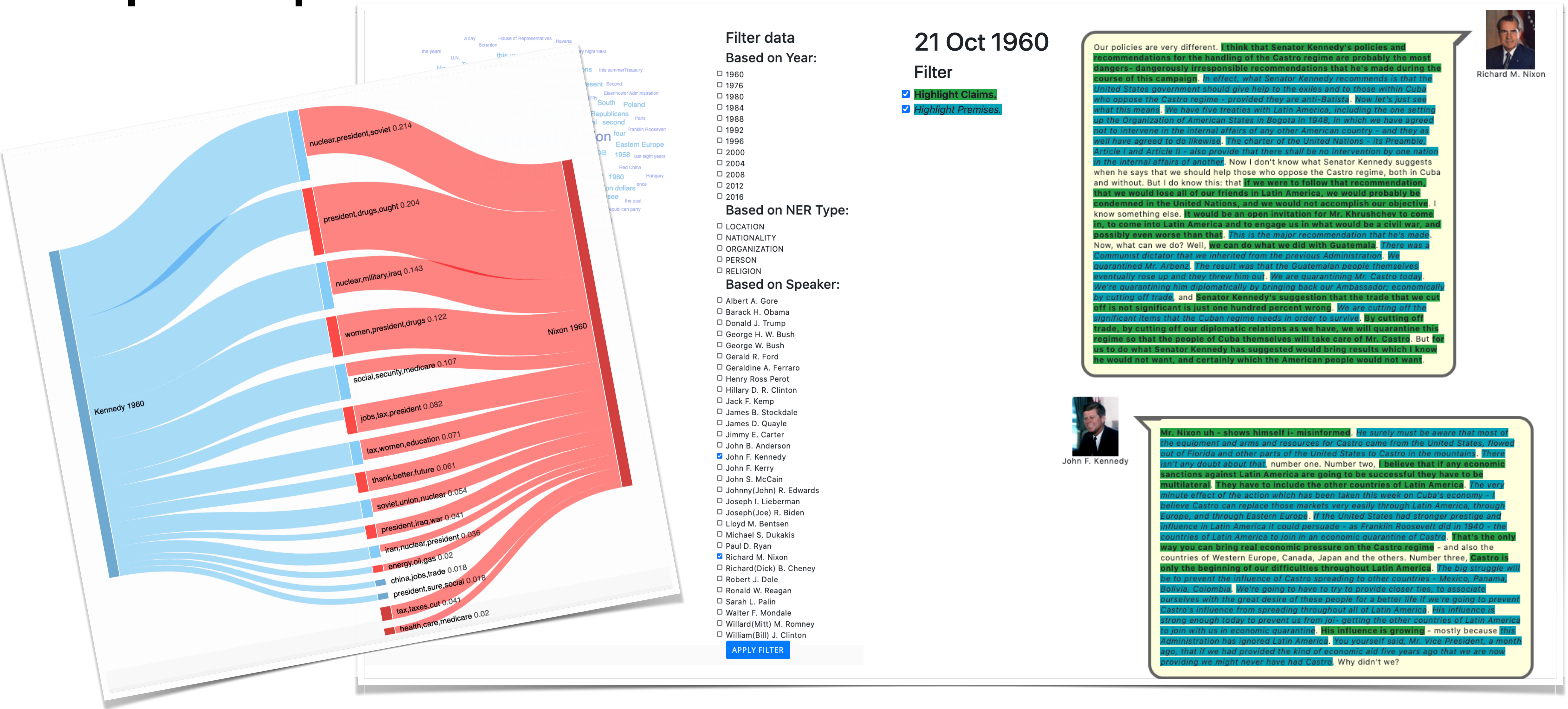
**Premise2**

*If you put 50 tons of TNT in each one of railroad cars, you would have a carload of TNT - a trainload of TNT stretching across this nation*

**Collaborations:** Univ. of Luxembourg

# DispuTOOL

## https://disputool.uni.lu/



# Explanatory arguments (and their further use in dialogues)

# Argument-based explanation patterns
## (Darpa XAI Program Update)

- **analytic statements** in NL that describe the elements and context that support a choice,

  ➡ the arguments (evidence, claim, warrant if any)

- **visualizations** that highlight portions of the raw data that support a choice,

- cases that invoke **specific examples**, and

  ➡ hard, you need more than one case to support by examples the choice

- **rejections of alternative choices** that argue against less preferred answers based on analytics, cases, and data.

  ➡ hard, you need the arguments from the rejected options

# Use case example to build the dataset

A 37-year-old woman is brought to the emergency department because of intermittent chest pain for 3 days. The pain is worse with inspiration, and she feels she cannot take deep breaths. She has not had shortness of breath, palpitations, or nausea. She had an upper respiratory tract infection 10 days ago and took an over-the-counter cough suppressant and decongestant and acetaminophen. Her temperature is 37.2°C (98.9°F), pulse is 90/min, and blood pressure is 122/70 mm Hg. The lungs are clear to auscultation. S1 and S2 are normal. A rub is heard during systole. There is no peripheral edema. An ECG shows normal sinus rhythm and diffuse, upwardly concave ST-segment elevation and PR-segment depression in leads II, III, and a VF.

# Use case example

## Training residents to improve argument-based diagnosis

**Which of the following is the most likely diagnosis?**

(A) Acute pericarditis

(B) Aortic dissection

(C) Gastroesophageal reflux disease

(D) Myocardial infarction

(E) Peptic ulcer disease

(F) Pulmonary embolism

(G) Unstable angina pectoris

**ALTERNATIVE OPTIONS**

# Use case example

## Training residents to improve argument-based diagnosis

**Which of the following is the most likely diagnosis?**

**(A) Acute pericarditis**

(B) Aortic dissection

(C) Gastroesophageal reflux disease

(D) Myocardial infarction

(E) Peptic ulcer disease

(F) Pulmonary embolism

(G) Unstable angina pectoris

**ALTERNATIVE OPTIONS**

# Use case example

## Training residents to improve argument-based diagnosis

**Which of the following is the most likely diagnosis?**

**(A) Acute pericarditis**

**<u>Why?</u>**

A friction rub and diffuse low-grade ST-segment elevation equals pericarditis.

# Use case example

- Clinical case: a 37-year-old woman is brought to the emergency department because of intermittent chest pain for 3 days. The pain is worse with inspiration, and she feels she cannot take deep breaths. She has not had shortness of breath, palpitations, or nausea. She had an upper respiratory tract infection 10 days ago and took an over-the-counter cough suppressant and decongestant and acetaminophen. Her temperature is 37.2°C (98.9°F), pulse is 90/min, and blood pressure is 122/70 mm Hg. The lungs are clear to auscultation. S1 and S2 are normal. A rub is heard during systole. There is no peripheral edema. An ECG shows normal sinus rhythm and diffuse, upwardly concave ST-segment elevation and PR-segment depression in leads II, III, and a VF.

- Diagnosis: the patient is showing a pericarditis **because** she has a friction rub and diffuse low-grade ST-segment elevation.

# First step: extractive explanatory argument generation

- <u>Clinical case</u>: *[a 37-year-old woman is brought to the emergency department because of intermittent chest pain for 3 days]*. *[The pain is worse with inspiration]*, and she feels *[she cannot take deep breaths]*. *[She has not had shortness of breath, palpitations, or nausea]*. *[She had an upper respiratory tract infection 10 days ago]* and *[took an over-the-counter cough suppressant and decongestant and acetaminophen]*. *[Her temperature is 37.2°C (98.9°F)]*, *[pulse is 90/min]*, and *[blood pressure is 122/70 mm Hg]*. *[The lungs are clear to auscultation]*. *[S1 and S2 are normal]*. *[A rub is heard during systole]*. *[There is no peripheral edema]*. *[An ECG shows normal sinus rhythm and diffuse]*, *[upwardly concave ST-segment elevation]* and *[PR-segment depression in leads II, III, and a VF]*.

- <u>Diagnosis</u>: the patient is showing a pericarditis **because** [a *rub is heard during systole]* and the ECG shows [*concave ST-segment elevation*].

# Extractive explanatory argument generation

## Argument Mining + Knowledge graphs

- **Diagnosis with explanation by expert**: the patient is showing a pericarditis **because** she has a friction rub and diffuse low-grade ST-segment elevation.
- **Diagnosis with extracted explanatory arguments**: the patient is showing a pericarditis **because** [a *rub is heard during systole]* and the ECG shows [*concave ST-segment elevation*].
- **What we have?**
    - Premises extracted from description of the case, correct diagnosis.
- **What we need further?**
    - Criteria to choose among the premises to pick the right ones, those which justify the diagnosis —> knowledge graphs of clinical knowledge
    - What if the explanation is not “contained” in the evidence ?

# Explanatory dialogues

## Argument mining and generation

- (Counter-)argument generation SoA (e.g., (Park et al., 2019, Hua et al., 2019)): mainly reformulation of arguments mined from Wikipedia and newspaper articles
- Insufficient to generate effective and interactive explanatory arguments
- **Extractive argument generation** vs. **abstractive argument generation**
- Large-scale unsupervised language models to generate arguments
- **Explanatory arguments meet high quality arguments:**
  - quality (i.e., variability of the explanatory arguments, no repetitiveness)
  - quantity
  - standard evaluation metrics: BLEU and BertScore

# Main open challenges


- **(Annotated) Data**
- **World knowledge and specific domain knowledge**
  - To allow for generalisations, instantiations, inferences
- **How to evaluate explanatory dialogues?**
  - quality and quantity of the generated arguments
  - structural simplicity, coherence, minimality
  - what else?
- **Are these explanations actually for humans?** If so, human feedback required!

**Serena Villata**
CR1 CNRS, HDR
Université Cote d’Azur, CNRS, Inria

Laboratoire I3S (SPARKS-WIMMICS team)

serena.villata@univ-cotedazur.fr

http://www.i3s.unice.fr/~villata/

@serena_villata

**Thanks !**