Argumentation-Based Interactive Explanations

Francesca Toni

with O Cocarascu, K Cyras, A Rago

Computational Logic & Argumentation Group

(and with R Misener, D Letsios)

Department of Computing

Imperial College London

Outline

- Setting the ground: Explainable AI/Conversational AI
- From data/data-centric method to argumentative abstraction (i.e. argumentation framework) to (interactive) explanations
- Explanation beyond data-centric AI

Why explanation?

SECTION 1

TRANSPARENCY AND MODALITIES

Article 12

Transparent information, communication and modalities for the exercise of the rights of the data subject

AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned.

What is an explanation in AI?

1) Inform and help understand why a particular conclusion was reached
2) provide grounds to contest the conclusion if undesired
3) Inform and help understand what could be changed to get a desired conclusion

[Wachter et al 2017]

AI
(deep learning, recommender, decision-support, robot…)

Explanation
(mathematical, textual, visual, extractive, abstractive …)

Audience/Beneficiaries
(expert developer, user, policy maker,…)

Goals
(safety, trust…)

Evaluation of "quality" of explanation

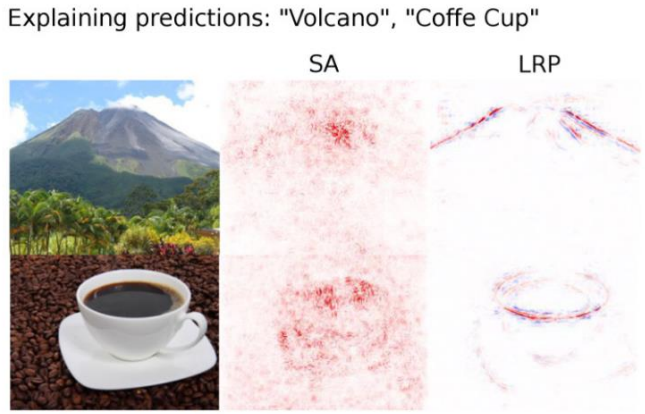
[Weller 2017]

Explanability vs

Transparency, Interpretability, Verifiability, Comprehensibility

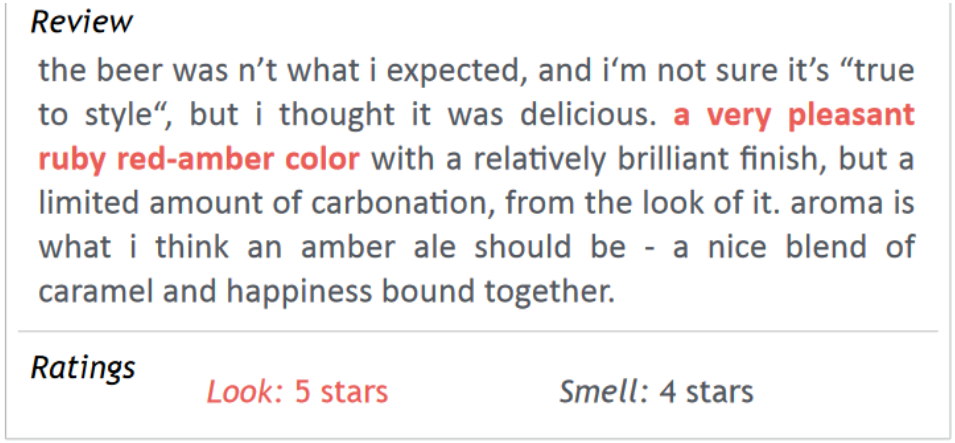
Some examples of explanation in AI

(A) Image classification



[Samek, Wiegand, Müller 2017]

(B) Natural Language Processing



[Lei, Barzilay, Jaakkola 2016]

(C) COUNTERFACTUALS: You were denied a loan because your annual income was £30,000. If your income had been £45,000 you would have been offered a loan.
[Wachter et al 2017]

(D) Anchors:
[Ribeiro, Singh, Guestrin 2018]

IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** $28 < \text{Age} \leq 37$
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary > $50K

Linear functions, decision trees, etc

Use explainable models instead…

… of explaining black-box models, especially when high-stake decisions are involved, as in healthcare and criminal justice

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1: 206–215. 2019

Conversational AI

Two-ways conversations

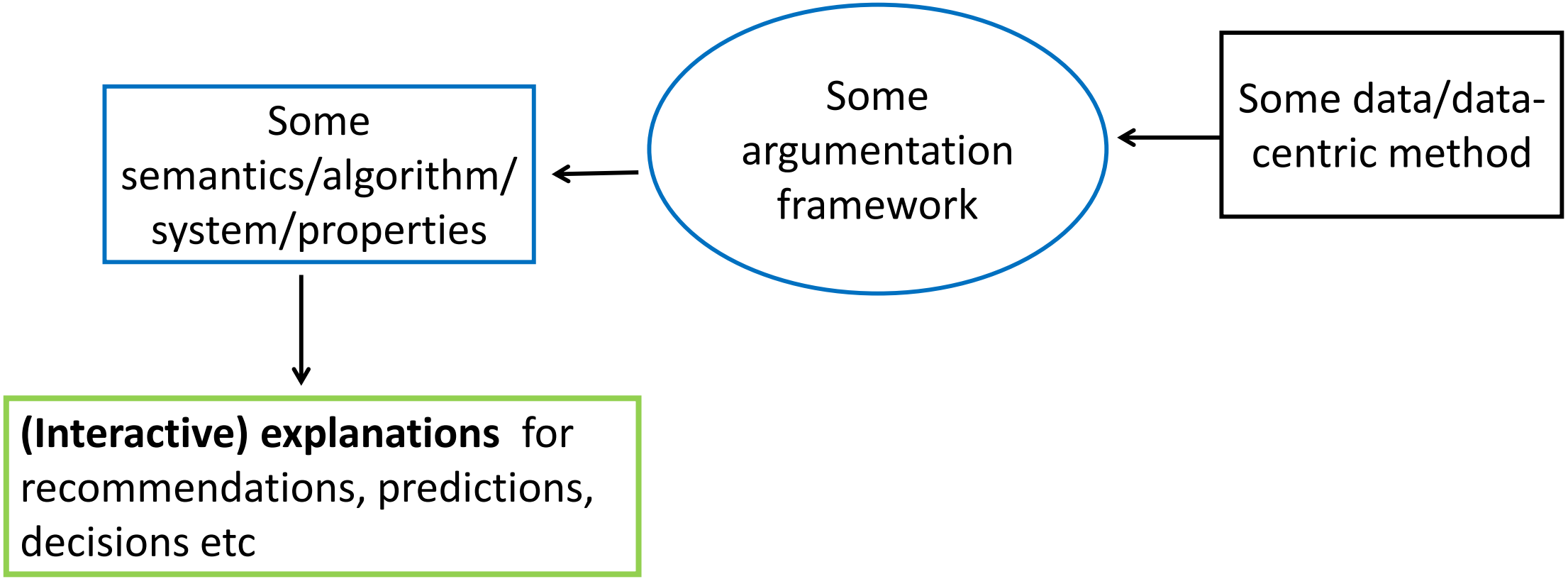
- Chat bots: human-like chit-chat generation (no explanation)
- Conversational bots: empower users to understand and improve recommendations dynamically

---

(1) You like movies that are tagged as ‘action’, especially those that are tagged as ‘sword fight’, such as The Princess Bride.
(2) You like movies that are tagged as ‘action’, especially those that are tagged as ‘sword fight’.
(3) You like movies that are tagged as ‘action’, such as The Princess Bride.
(4) You like movies that are tagged as ‘action’.
(5) You like The Princess Bride.
(6) None of the above.

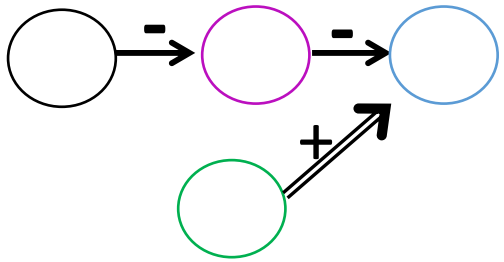
Balog, Radlinski, Arakelyan 2019

Interactive explanations from argumentative abstractions



Background

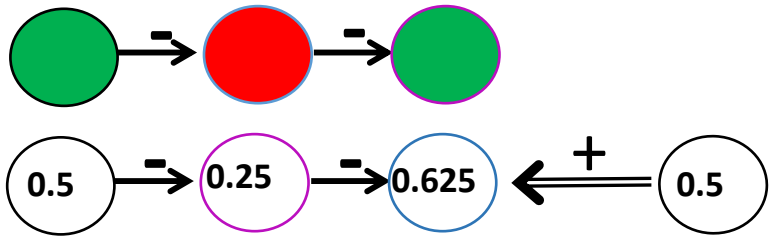
Argumentation frameworks (AFs) as abstractions of “debates”



AAFs

BAFs

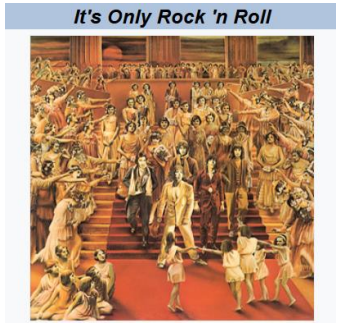
Semantics/algorithms/systems/ properties: characterisations of “winning” arguments



QBAFs

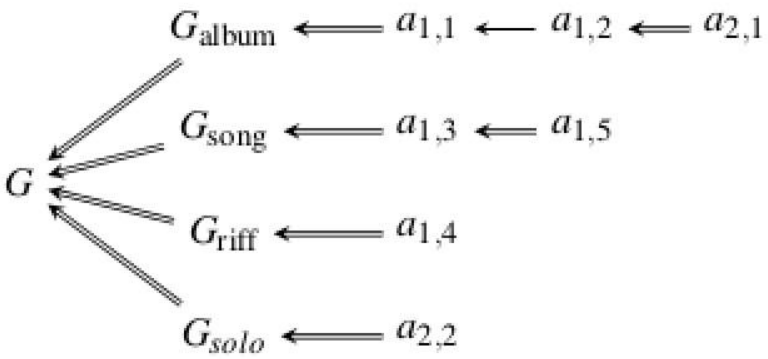
Argumentative abstractions

of text



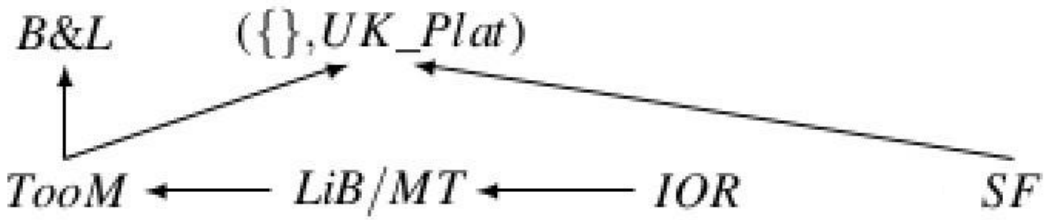
*r_1: It's Only Rock N' Roll is **a very good album**. The reason I gave it an eight and a half rating is because it's **not a masterpiece**. The **album does have some very strong songs**. "If You Can't Rock Me" has a **great guitar riff**. "Till The Next Goodbye" is a **beautiful love song**, like others which they have done before.*

*r_2: It is an **uneven album**, some tracks are great, some bad, and some mediocre. "Time Waits For No One" is arguably the Stones greatest performance on record, not just Taylor's **incredible solo** but great playing by the entire band.*



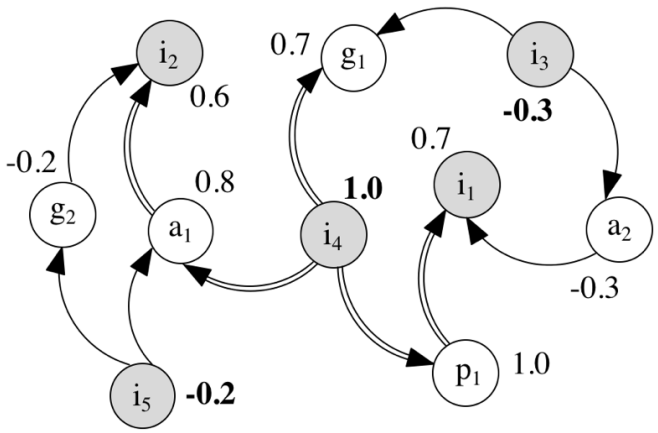
of labelled data

Abbreviated title	Features ($\mathcal{F}_p$)	Outcome ($\mathcal{O}$)
B&L	*Blues*	*UK_Plat*
TooM	*Blues, US_Plat*	*not UK_Plat*
SF	*R'n'R, US_Plat*	*not UK_Plat*
LiB/MT	*Blues, US_Plat, Pos_Rev*	*UK_Plat*
IOR	*Blues, R'n'R, US_Plat*	?



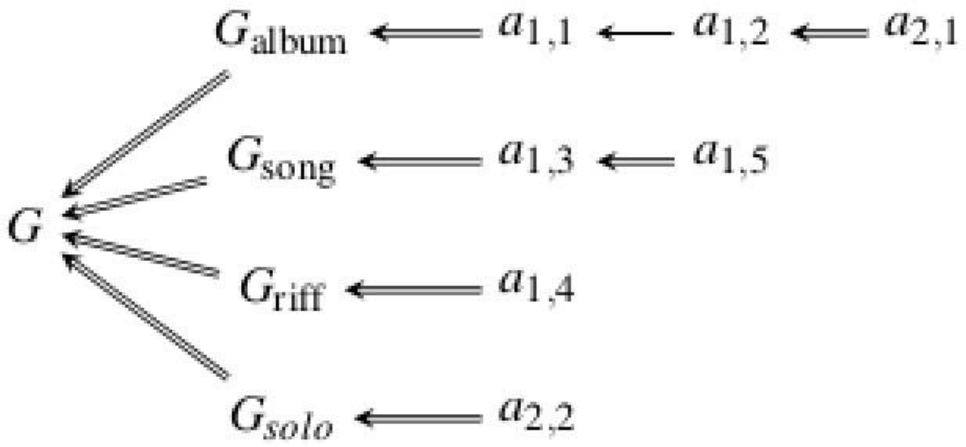
of recommender systems

	Node	Item/Feature	Associations	$\mathcal{A}$	$\mathcal{P}$
Items	i_1	*MT*	$\{a_2, g_1, p_1\}$	-	0.7
	i_2	*IOR*	$\{a_1, g_1, g_2\}$	-	0.6
	i_3	*TooM*	$\{a_2, g_1\}$	-0.3	-0.3
	i_4	*LiB*	$\{a_1, g_1, p_1\}$	1.0	1.0
	i_5	*SF*	$\{a_1, g_2\}$	-0.2	-0.2
Features	g_1	*Blues*	$\{i_1, i_2, i_3, i_4\}$	-	0.7
	g_2	*R'n'R*	$\{i_2, i_5\}$	-	-0.2
	a_1	TRS	$\{i_2, i_4, i_5\}$	-	0.8
	a_2	Bob Dylan	$\{i_1, i_3\}$	-	-0.3
	p_1	*UK_Plat*	$\{i_1, i_4\}$	-	1.0



Mining BAFs from text– some details

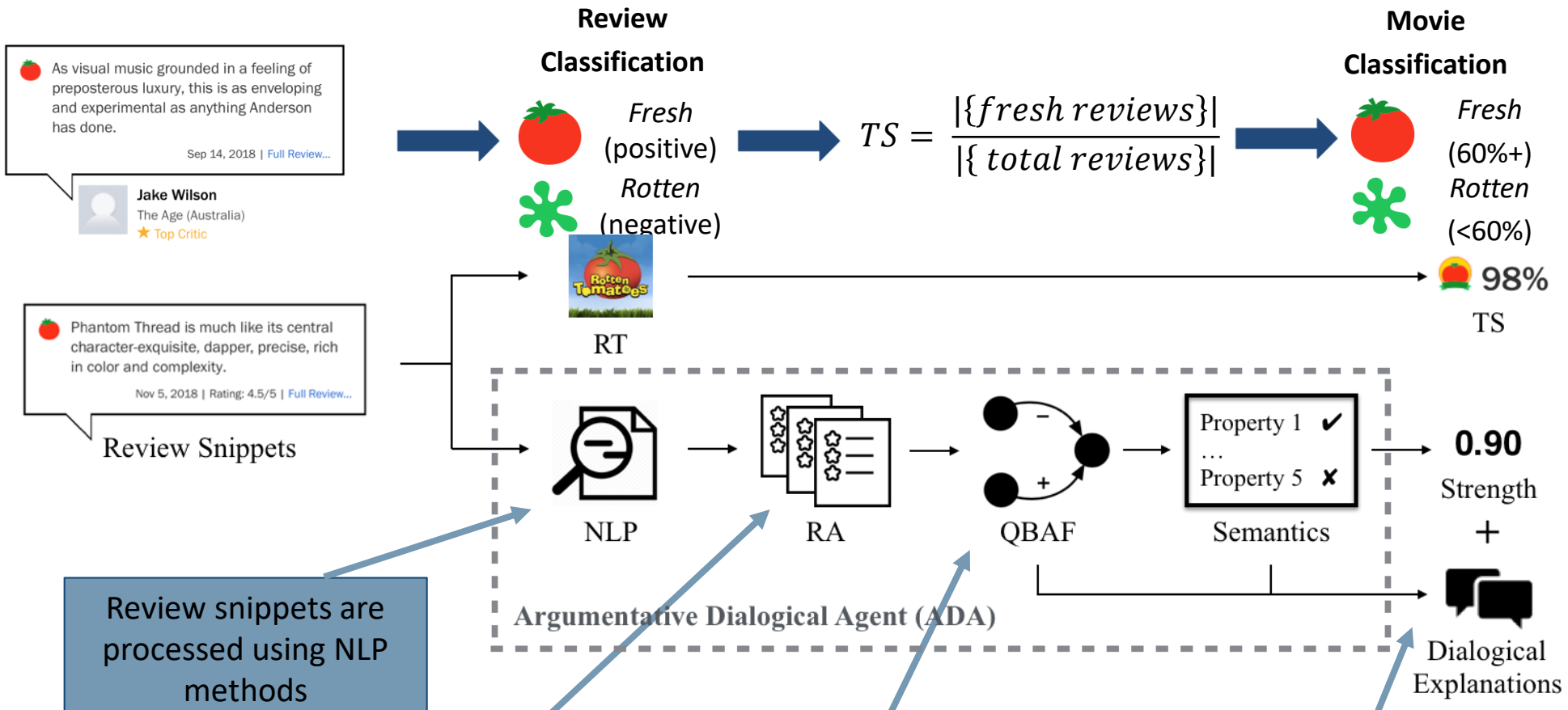
r_1: *It's Only Rock N' Roll is* ***a very good album****. The reason I gave it an eight and a half rating is because it's* ***not a masterpiece****. The* ***album does have some very strong songs****. "If You Can't Rock Me" has a* ***great guitar riff****. "Till The Next Goodbye" is a* ***beautiful love song****, like others which they have done before.*
r_2: *It is an* ***uneven album****, some tracks are great, some bad, and some mediocre. "Time Waits For No One" is arguably the Stones greatest performance on record, not just Taylor's* ***incredible solo*** *but great playing by the entire band.*



- Argument identification
- Topic extraction
- Relation-based argument mining (reflecting temporal ordering of arguments)

Cocarascu, Toni Comp Ling 2018

(more) Argumentative abstractions of text



Review snippets are processed using NLP methods

Feature-based *review aggregations* are then extracted

QBAFs representing the arguments in the reviews are formed

QBAFs, equipped with a suitable semantics, provide a strength measure and support explanations

Cocarascu, Rago, Toni AAMAS 2019

Feature-based characterisations of items

We rely upon a feature-based characterization, e.g.

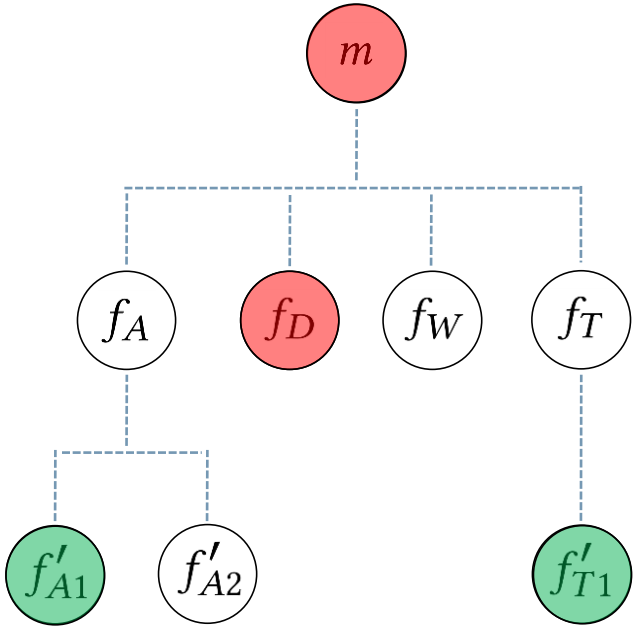
a movie	*has:*	**features**	*has:*	**sub-features**
e.g. *Wonder Wheel*	→	e.g. *acting*	→	e.g. *Kate Winslet*

Different categories of features, e.g.

acting	**writing**	**directing**	**themes**
multiple	single	single	multiple
predetermined	predetermined	predetermined	mined
e.g. *Kate Winslet, Justin Timberlake*	e.g. *Woody Allen*	e.g. *Woody Allen*	e.g. *Amusement Parks, Romance*

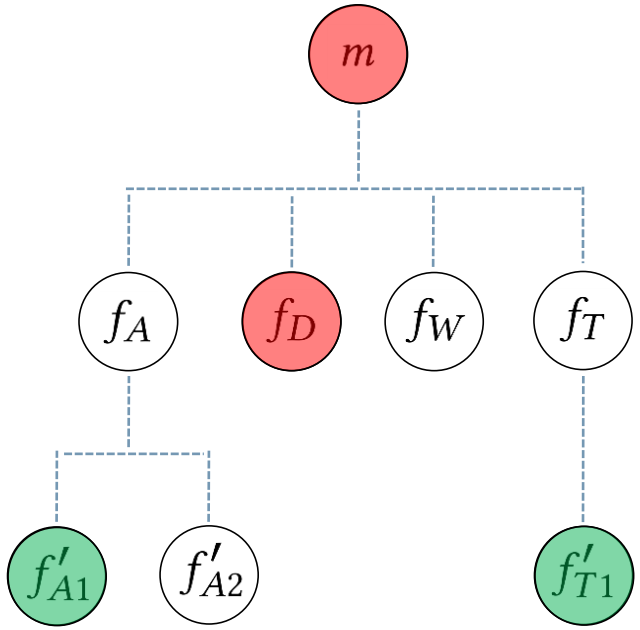
Extracting Review Aggregations with NLP

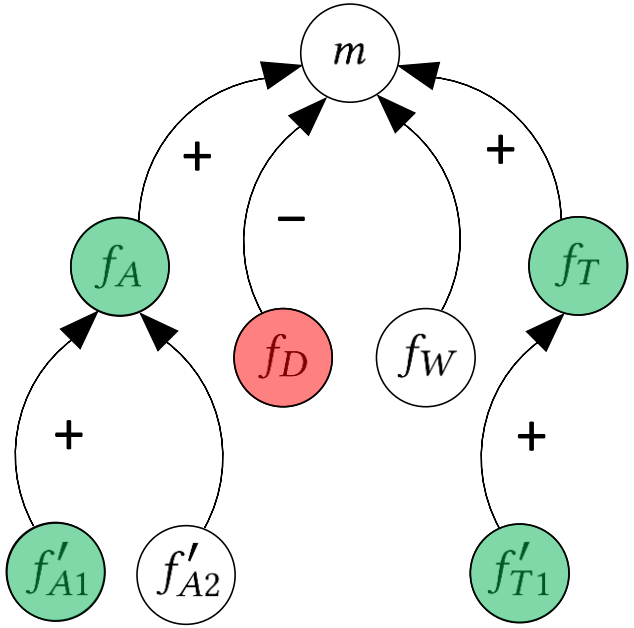
- Sentiment analysis (SA)/ Argument Mining (AM) to determine:
- *votes with* positive or negative sentiment obtained from (phrases in snippets by) *critics* on items, features and sub-features;



- *"As we watch Allen worry and nitpick over the way women fret over aging…it's dull, unoriginal, and offensive."*
  - a negative vote for *directing* (f_D)
- *"Despite a stunning performance by Winslet…Wonder Wheel loses its charms quickly and you'll soon be begging to get off this particular ride."*
  - a positive vote for *Kate Winslet* (f'_{A1})
  - a negative vote for the movie in general (m)
- *"…like the fairground ride for which it's named, Wonder Wheel is entertaining."*
  - a positive vote for *amusement park* (f'_{T1})

From Review Aggregations to QBAFs

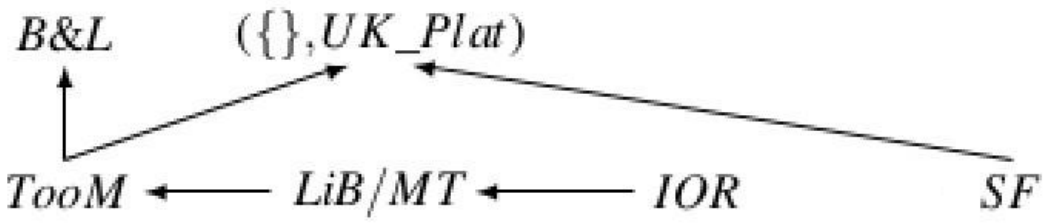




Mining AAFs from labelled data – some details

DS=

Abbreviated title	Features ($\mathcal{F}_p$)	Outcome ($\mathcal{O}$)
B&L	*Blues*	*UK_Plat*
TooM	*Blues, US_Plat*	*not UK_Plat*
SF	*R'n'R, US_Plat*	*not UK_Plat*
LiB/MT	*Blues, US_Plat, Pos_Rev*	*UK_Plat*
IOR	*Blues, R'n'R, US_Plat*	?



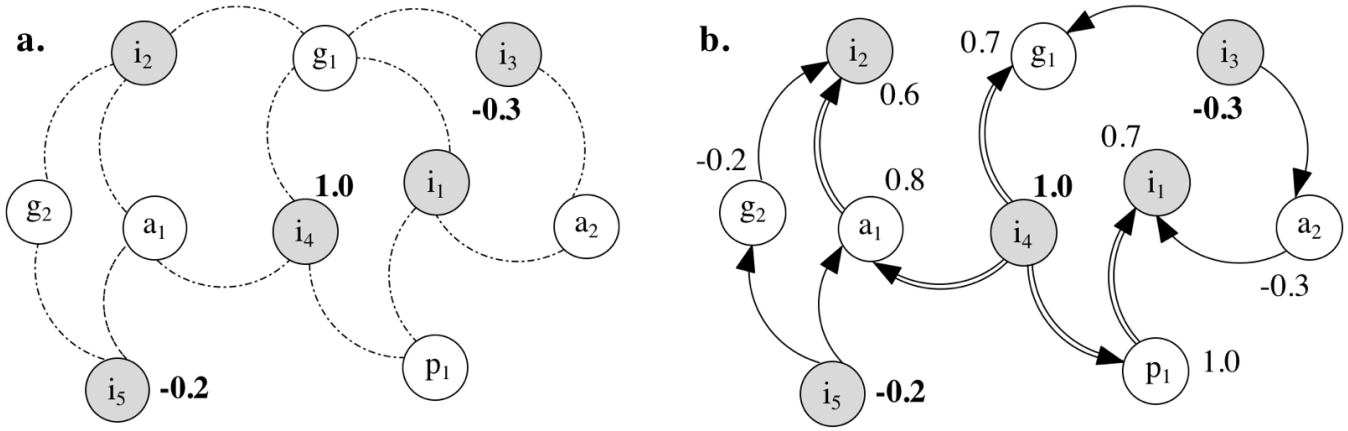
- $Args = DS \cup \{(\{\}, \varphi)\}$ (with $(\{\}, \varphi)$ the *default argument*)
- for $(X, o_X), (Y, o_Y) \in Args$, it holds that $((X, o_X), (Y, o_Y)) \in \mathcal{R}^-$ iff
  1. $o_X \neq o_Y$, and (different outcomes)
  2. $Y \subsetneq X$, and (specificity)
  3. $\nexists (Z, o_X) \in DS$ with $Y \subsetneq Z \subsetneq X$. (concision)

- $Args_N = Args \cup \{(N, ?)\}$, and
- $\mathcal{R}^-_N = \mathcal{R}^- \cup \{((N, ?), (Y, o_Y)) : (Y, o_Y) \in Args \text{ and } Y \not\subseteq N\}$.

Cyras, Cocarascu, Toni XAI 2018

Mining QBAFs from recommender systems – some details

	Node	Item/Feature	Associations	$\mathcal{A}$	$\mathcal{P}$
Items	i_1	*MT*	$\{a_2, g_1, p_1\}$	-	0.7
	i_2	*IOR*	$\{a_1, g_1, g_2\}$	-	0.6
	i_3	*TooM*	$\{a_2, g_1\}$	-0.3	-0.3
	i_4	*LiB*	$\{a_1, g_1, p_1\}$	1.0	1.0
	i_5	*SF*	$\{a_1, g_2\}$	-0.2	-0.2
Features	g_1	*Blues*	$\{i_1, i_2, i_3, i_4\}$	-	0.7
	g_2	*R'n'R*	$\{i_2, i_5\}$	-	-0.2
	a_1	TRS	$\{i_2, i_4, i_5\}$	-	0.8
	a_2	Bob Dylan	$\{i_1, i_3\}$	-	-0.3
	p_1	*UK_Plat*	$\{i_1, i_4\}$	-	1.0



- $Args = \mathcal{I} \cup \mathcal{F}_r$,
- $\forall x \in Args$, $\tau(x) = \mathcal{A}(x)$ if this is defined, and $\tau(x) = 0$ else,
- $\mathcal{R}^-$ and $\mathcal{R}^+$ are defined so that the *attackers/ supporters* of $\alpha \in Args$, $\mathcal{R}^-(\alpha) = \{\beta \in Args | (\beta, \alpha) \in \mathcal{R}^-\}$/ $\mathcal{R}^+(\alpha) = \{\beta \in Args | (\beta, \alpha) \in \mathcal{R}^+\}$, are as follows, for $f \in \mathcal{F}_r$, $i \in \mathcal{I}$:
  - $\mathcal{R}^-(f) = argmax_{i_- \in \mathcal{I}}\{-\mathcal{A}(i_-) | f \in \mathcal{F}_{r_{i_-}} \wedge \mathcal{A}(i_-) < 0\}$;
  - $\mathcal{R}^+(f) = argmax_{i_+ \in \mathcal{I}}\{\mathcal{A}(i_+) | f \in \mathcal{F}_{r_{i_+}} \wedge \mathcal{A}(i_+) > 0\}$
  - if $\mathcal{A}(i)$ is defined then $\mathcal{R}^-(i) = \mathcal{R}^+(i) = \{\}$, otherwise:
    $\mathcal{R}^-(i) = argmax_{f_- \in \mathcal{F}_r}\{-\mathcal{P}(f_-) | f_- \in \mathcal{F}_{r_i} \wedge \mathcal{P}(f_-) < 0\}$;
    $\mathcal{R}^+(i) = argmax_{f_+ \in \mathcal{F}_r}\{\mathcal{P}(f_+) | f_+ \in \mathcal{F}_{r_i} \wedge \mathcal{P}(f_+) > 0\}$.

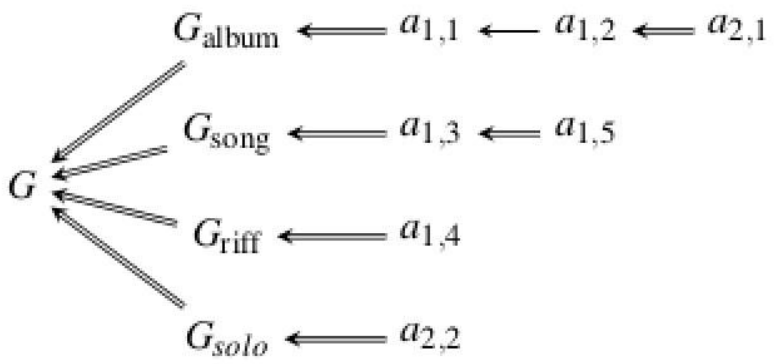
Rago, Cocarascu, Toni IJCAI-ECAI 2018

Explanations

- an **explanation** for $\alpha \in$ Args in an AF with arguments Args is a sub-graph **G** $\sqsubseteq$ AF such that α is a node of **G** and **G** fulfils some property **P**.

- P may be the property that
  - "G is a branch"
  - "G is (the set of nodes/edges in) a maximal or admissible dispute tree"
  - "G is maximally connected"

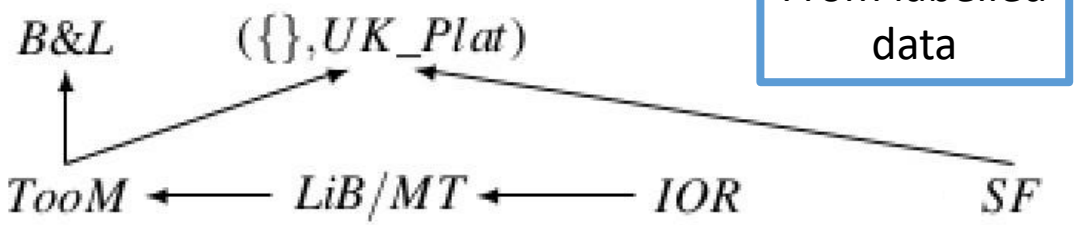
From argumentative abstractions to explanations



From text

$$G \Longleftarrow G_{song} \Longleftarrow a_{1,3} \Longleftarrow a_{1,5}$$

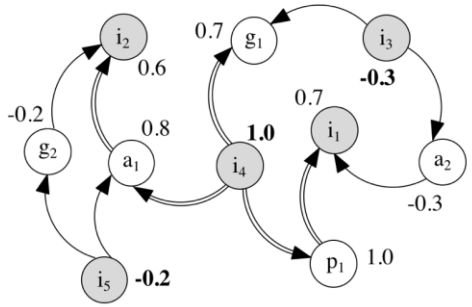
Explanation as to why *IOR* is worthwhile



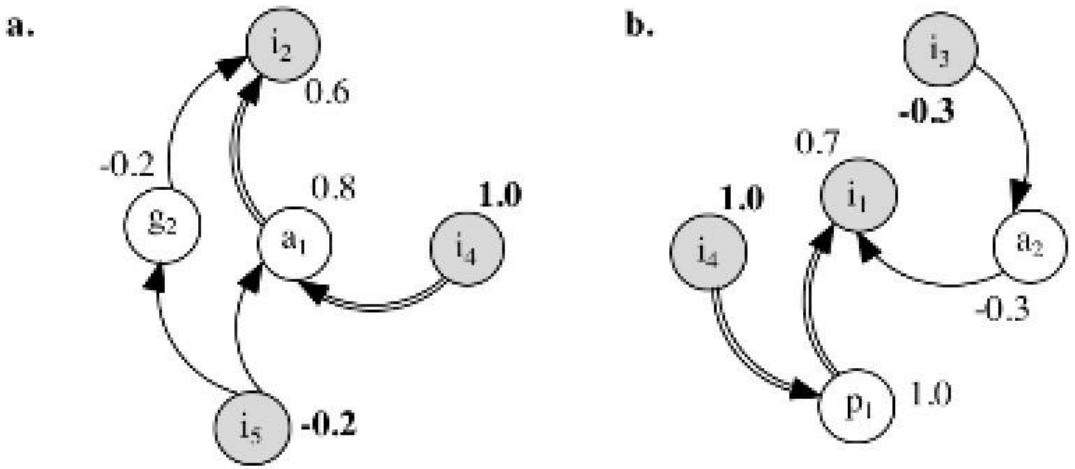
From labelled data

$$[O:IOR] \rightarrow [P:LiB/MT] \rightarrow [O:TooM] \rightarrow [P:(\{\},UK_Plat)]$$

Explanation for the AA-CBR outcome of ({*Blues,R'n'R,US_Plat*},?)

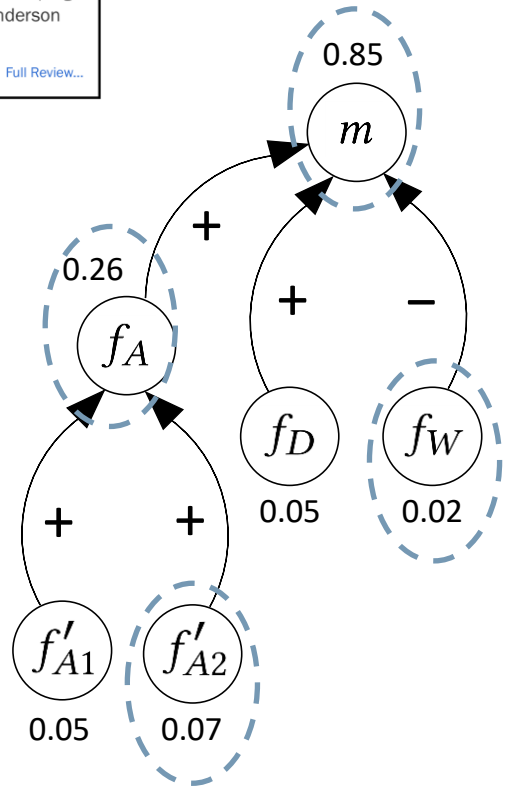


From recommender systems



Explanations for items i_2 (a) and i_1 (b)

Interactive Explanations



user: *Why was The Post highly rated?*

system: *This movie was highly rated because the acting was great, although the writing was poor.*

user: *Why was the acting considered to be great?*

system: *The acting was considered to be great because Meryl Streep was great.*

user: *What did critics say about Meryl Streep being great?*

system: *"...Streep's hesitations, rue, and ultimate valor are soul-deep..."*

Amenability of (interactive) explanations to humans

Which types of explanations do users “like”?

- Explanations providing more or less or no evidence?
  1. I recommend film F because you like(d) films X and Y
  2. I recommend film F because you like(d) film X
- Explanations containing inferred statements or plain evidence?
  1. I recommend film F because you like(d) comedy films such as X
  2. I recommend film F because you like(d) X which is a comedy film
- Generic explanations?
  1. People like you liked film F
  2. I recommend film F because you like drama films and films starring actor X

Ongoing, joint work with Rago and UCL’s Lagnado and Bechlivanidis

Interactive explanations via argumentative abstractions beyond data

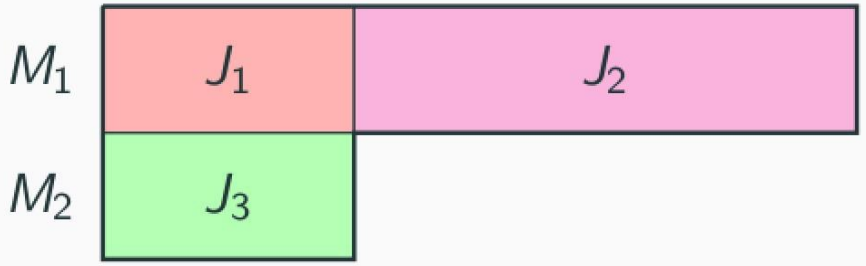
(Interactively) explaining why schedules are (not) "quasi-optimal"

Why am I to do job j?

$$\min_{\mathbf{x}, C_{\max}} \quad C_{\max}$$

$$\begin{array}{lll} & \sum_{j=1}^{n} x_{i,j} \cdot p_j & \leq C_{\max} \quad i \in \mathcal{M} \\ \star & \sum_{i=1}^{m} x_{i,j} & = 1 \quad j \in \mathcal{J} \\ & x_{i,j} \in \{0,1\} & \quad j \in \mathcal{J}, i \in \mathcal{M} \end{array}$$

schedule S is efficient only if $\star$ and
for $j' \neq j$ with $x_{i',j'} = 1$, if $p_j > p_{j'}$, then $C_i + p_{j'} \leqslant C_{i'} + p_j$

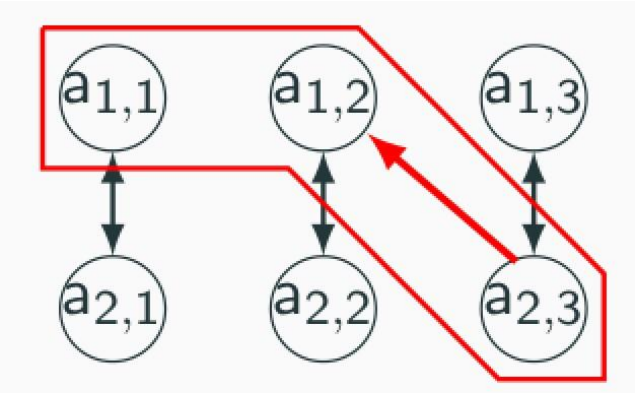


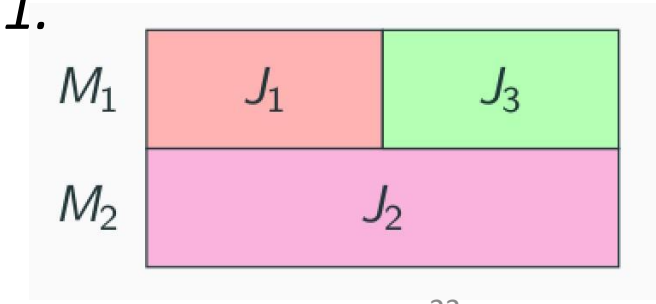
why is schedule S ≈ {$x_{1,1}$ =1, $x_{1,2}$ =1, $x_{2,3}$ =1} not efficient?

The attack from $a_{2,3}$ to $a_{1,2}$ explains why S is not efficient:

Because S can be improved by swapping jobs J3 and J2 between machines (e.g. nurses) M2 and M1.

Mined AF from S:





Cyras, Letsios, Misener, Toni AAAI 2019

Conclusions

- Explanations empowered by argumentative abstractions as a means to generate interactive explanations

- Which form of explanation?
- Interactive explanations for humans to feedback into machines?
- From engineering explanations to a scientific theory of explanation?

Questions?