A 2-Phase Frame-based Knowledge Extraction Framework

Francesco Corcoglioniti

francesco@corcoglioniti.name

Marco Rospocher, Alessio Palmero Aprosio

Fondazione Bruno Kessler – IRST
Trento, Italy

SAC 2016
PISA, 06 April 2016

http://pikes.fbk.eu/

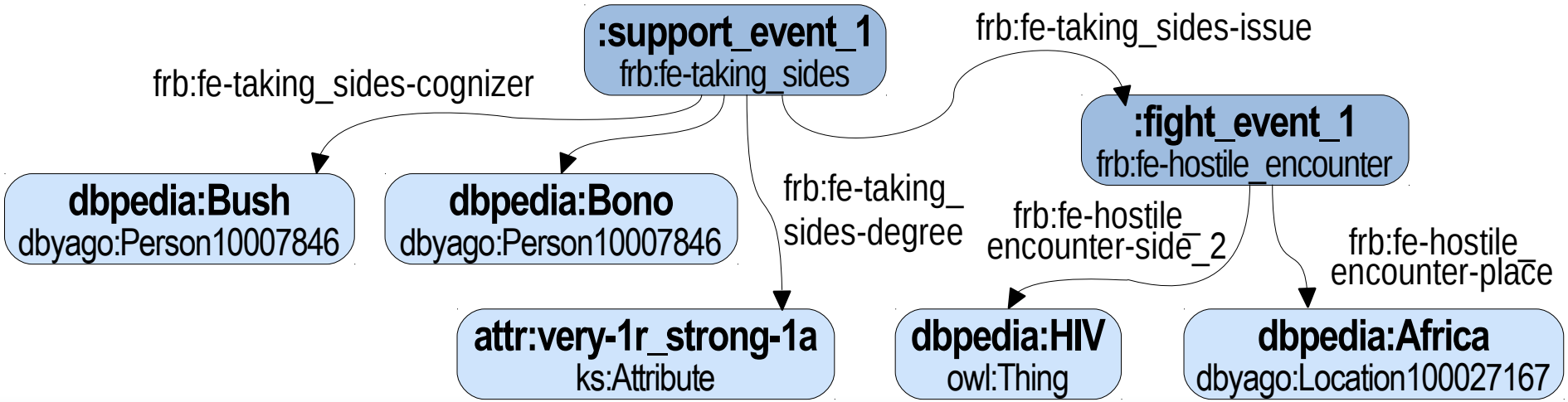
Problem

Knowledge Extraction from Text

- English text only
- ABox (instances and facts) only → Ontology Population
- focus on extracting events and their participants

  → represented as semantic frames, i.e., event instances (e.g. 'sell' event) linked to participant instances via role properties (e.g. 'seller')

Example: *"G. W. Bush and Bono are very strong **supporters** of the **fight** of HIV in Africa."*



Contribution

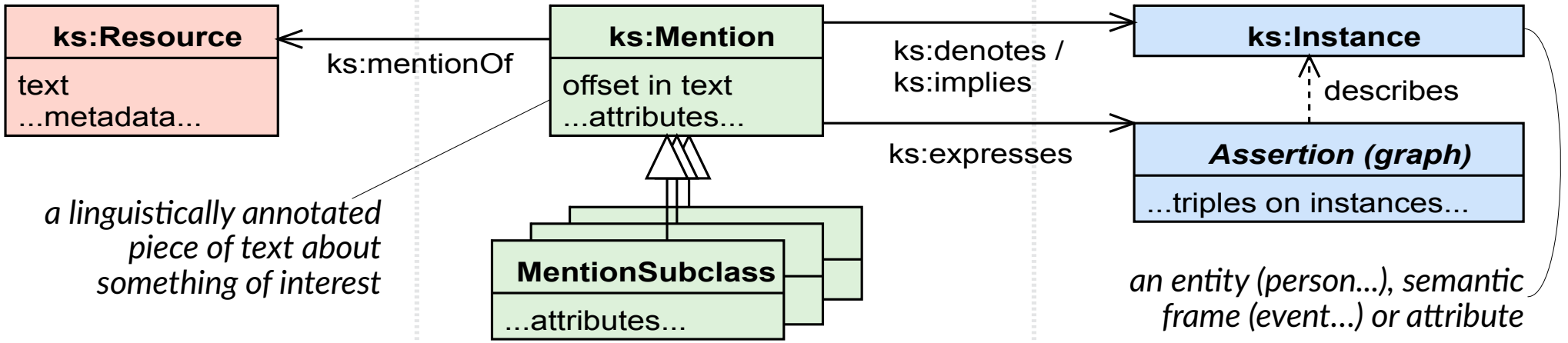
FONDAZIONE BRUNO KESSLER

PIKES* 

- a tool for Knowledge Extraction from English text
- extracting semantic frames
  - aligned to predicate models
    → PropBank (PB), NomBank (NB), VerbNet (VN), FrameNet (FN)
  - ***new:*** *aligned to FrameBase*
- extracting instances
  - typed w.r.t. YAGO and SUMO
  - disambiguated w.r.t. DBpedia
- representing all contents in RDF + named graph
- based on a 2-phase approach
- open source – http://pikes.fbk.eu/

(*) PIKES Is a Knownedge Extraction Suite

Data Model




```
:resource1 a ks:Resource;
   dct:created "2016-04-06";
   nif:isString "G. W. Bush
      and Bono are very
      strong supporters of
      the fight of HIV in
      Africa.".
```

```
:mention1 a ks:FrameMention;
   ks:mentionOf :resource1;
   nif:beginIndex 36; nif:endIndex 46;
   nif:anchorOf "supporters";
   ks:synset wn30:n-10677713
   ks:predicate pmo:nb10_support.01;
   ks:role pmo:nb10_support.01_arg01;
   ks:expresses :graph1;
   ks:denotes :supporters_entity;
   ks:implies :support_event_1.
```

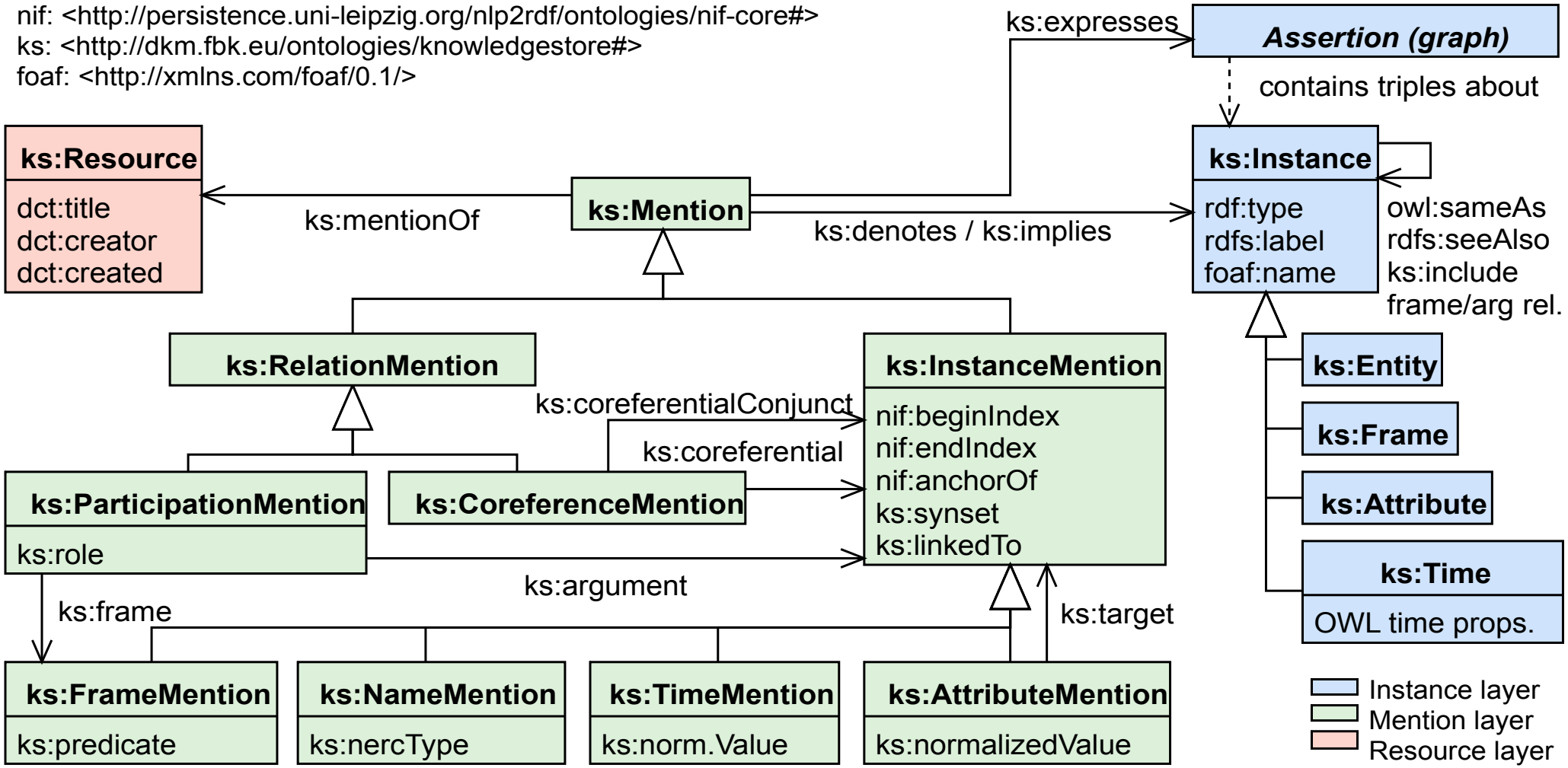
```
:graph1 {
 :supporters_entity
    a dbyago:Supporter110677713.
 :support_event_1 a
    a frb:frame-taking_sides;
    frb:fe-taking_sides-cognizer
       :supporters_entity.
}
```

based on: *Corcoglioniti et al. KnowledgeStore: a storage framework for interlinking unstructured and structured knowledge. IJSWIS 2015*

Data Model (2)

FONDAZIONE BRUNO KESSLER

Complete model:



2-Phase Approach

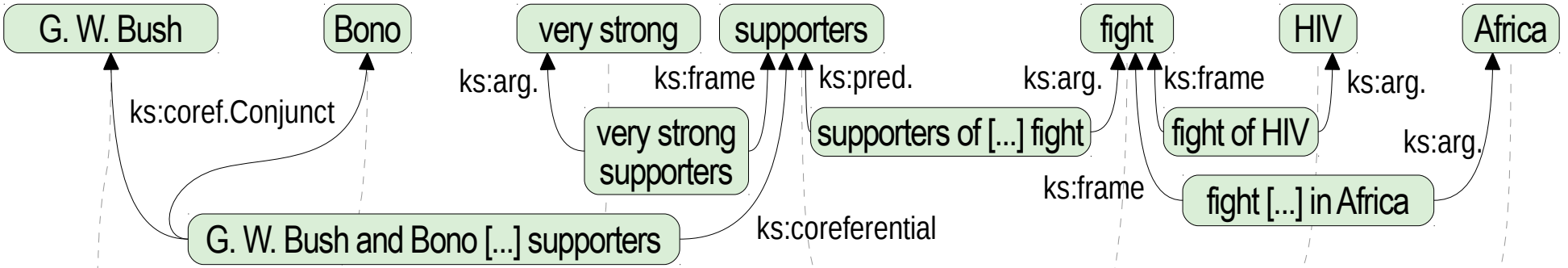
FONDAZIONE BRUNO KESSLER

Resource layer

G. W. Bush and Bono are very strong supporters of the fight of HIV in Africa.

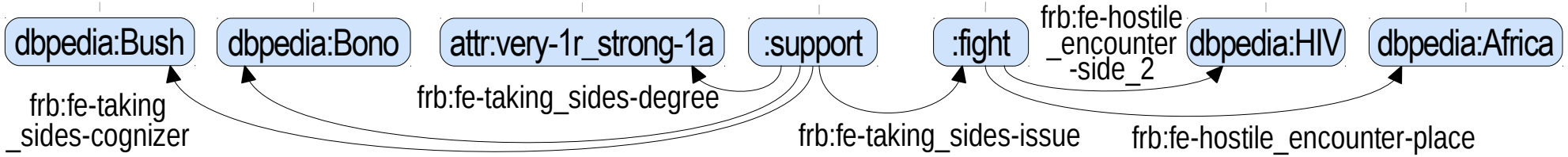
Phase 1
Linguistic Feature Extraction

Mention layer



Phase 2
Knowledge Distillation

Instance layer



Linguistic Feature Extraction

FONDAZIONE BRUNO KESSLER

① apply several NLP tasks to input text

② map their outputs to mentions

NLP Task ▼	Type of mention ▶	Instance	Name	Time	Attribute	Frame	Participation	Coreference
part-of-speech tagging	POS	√	√		√			
named entity recognition & classification	NERC		√	√				
temporal expression recognition & norm.	TERN			√				
entity linking	EL	√	√					
word sense disambiguation	WSD	√			√			
semantic role labeling	**SRL**					√	√	
coreference resolution	COREF							√
dependency parsing	DP				√		√	√

Linguistic Feature Extraction (2)

Example:

"fight of HIV"

Extracted RDF mention graph

via NERC, EL

```
<..#char=63,66> a :NameMention ;
    nif:anchorOf "HIV" ;
    :nercType :MISC ;
    :linkedTo dbpedia:HIV .
```

via SRL

```
<..#char=54,59> a :FrameMention ;
    nif:anchorOf "fight"
    :predicate pm:nb10-fight.01 .
```

via SRL, DP

```
<..#char=54,66> a :ParticipationMention;
    nif:anchorOf "fight […] HIV"
    :frame <..#char=54,59> ;
    :argument <..#char=63,66> ;
    :role pmo:nb10-fight.01-arg1 .
```

Knowledge Distillation

FONDAZIONE BRUNO KESSLER

① Rule-based conversion from Mention to Instance data
- deal with phenomena such as argument nominalization and group entities
- use background knowledge
  → e.g., mappings to ontologies, characterization of predicates

② Post-processing: OWL2RL inference, reduce # of named graphs

Mention layer

```
:mention1 a ks:FrameMention;
   nif:anchorOf "supporters";
   ks:synset wn30:n-10677713;
   ks:predicate pmo:nb10_support.01;
   ks:role pmo:nb10_support.01_arg01;
```

Instance layer

```
:g1 {  :e1 a dbyago:Supporter110677713.
       :ev1 a frb:frame-taking_sides;
          frb:fe-taking_sides-cognizer :e1. }

:mention1 ks:expresses :g1;
      ks:denotes :e1; ks:implies :ev1.
```

Background knowledge

```
pmo:nb10_support.01
   a ks:ArgumentNominalization.
```

```
INSERT { ?m ks:denotes ?i; ks:implies ?if; ks:expresses ?g.
         GRAPH ?g { ?i a ks:Instance. ?if a ks:Frame } }
WHERE { ?m a ks:FrameMention; nif:anchorOf ?a, ks:predicate ?s.
         ?s a ks:ArgumentNominalization.
         BIND (ks:mint(?m) AS ?g)  BIND (ks:mint(?a, ?m) AS ?i)
         BIND (ks:mint(concat(?a, "_pred"), ?m) AS ?if)
```


Knowledge Distillation (2)

FONDAZIONE BRUNO KESSLER

Longer example:

Mention layer

```
:m1 a ks:NameMention;
   nif:anchorOf “G. W. Bush”;
   ks:nercType ks:bbn_person.

:m2 a ks:NameMention;
   nif:anchorOf “Bono”;
   ks:nercType ks:bbn_person.

:m3 a ks:FrameMention;
   nif:anchorOf "supporters";
   ks:predicate pmo:nb10_support.01;
   ks:role pmo:nb10_support.01_arg01;

:m4 a ks:CoreferenceMention;
   ks:coreferential :m3;
   ks:coreferentialConjunct :m1, m2.
```

Background knowledge

```
pmo:nb10_support.01
   a ks:ArgumentNominalization.
```

Instance layer

```
:m1 ks:expresses :g1; ks:denotes :e1 .
:g1 { :e1 a dbyago:PersonXYZ;
         owl:sameAs dbpedia:Bush;
         foaf:name “G. W. Bush”.  }

:m2 ks:expresses :g2; ks:denotes :e2 .
:g2 { :e2 a dbyago:PersonXYZ;
         owl:sameAs dbpedia:Bono;
         foaf:name “Bono”. }

:m3 ks:expresses :g3; ks:denotes :e3;
       ks:implies :ev1.
:g3 {  :e3 a dbyago:SupporterXYZ.
         :ev1 a frb:frame-taking_sides;
            frb:fe-taking_sides-cognizer :e3. }

:m4 ks:expresses :g4.
:g4 { :e3 ks:include :e1, :e2 }
```

Knowledge Distillation (3)

Post-processing:

- RDFS / OWL2RL inference & *owl:sameAs* smushing
- propagate triples from group entities to their members
- optimize use of named graphs

Instance layer (before)

```
:g1 { :e1 a dbyago:PersonXYZ;
        owl:sameAs dbpedia:Bush. }

:g2 { :e2 a dbyago:PersonXYZ;
        owl:sameAs dbpedia:Bono. }

:g3 {  :e3 a dbyago:SupporterXYZ.
        :ev1 a frb:frame-taking_sides;
          frb:fe-taking_sides-cognizer :e3. }

:g4 { :e3 ks:include :e1, :e2 }

:m1 ks:expresses :g1; ks:denotes :e1 . # bush
:m2 ks:expresses :g2; ks:denotes :e2 . # bono
:m3 ks:expresses :g3; ks:denotes :e3;
      ks:implies :ev1.    # supporters
:m4 ks:expresses :g4.  # bush+bono = supporters
```

Instance layer (post-processed)

```
:g1 { dbpedia:Bush a dbyago:PersonXYZ, ... }

:g2 { dbpedia:Bono a dbyago:PersonXYZ; ... }

:g3 { :ev1 a frb:frame-taking_sides, ... }
:g4 { :ev1 frb:fe-taking_sides-cognizer
        dbpedia:Bush, dbpedia:Bono }


:m1 ks:expresses :g1; ks:denotes dbpedia:Bush.
:m2 ks:expresses :g2; ks:denotes dbpedia:Bono.
:m3 ks:expresses :g3, :g4; ks:implies :ev1;
      ks:denotes dbpedia:Bush, dbpedia;Bono.
:m4 ks:expresses :g4.
```

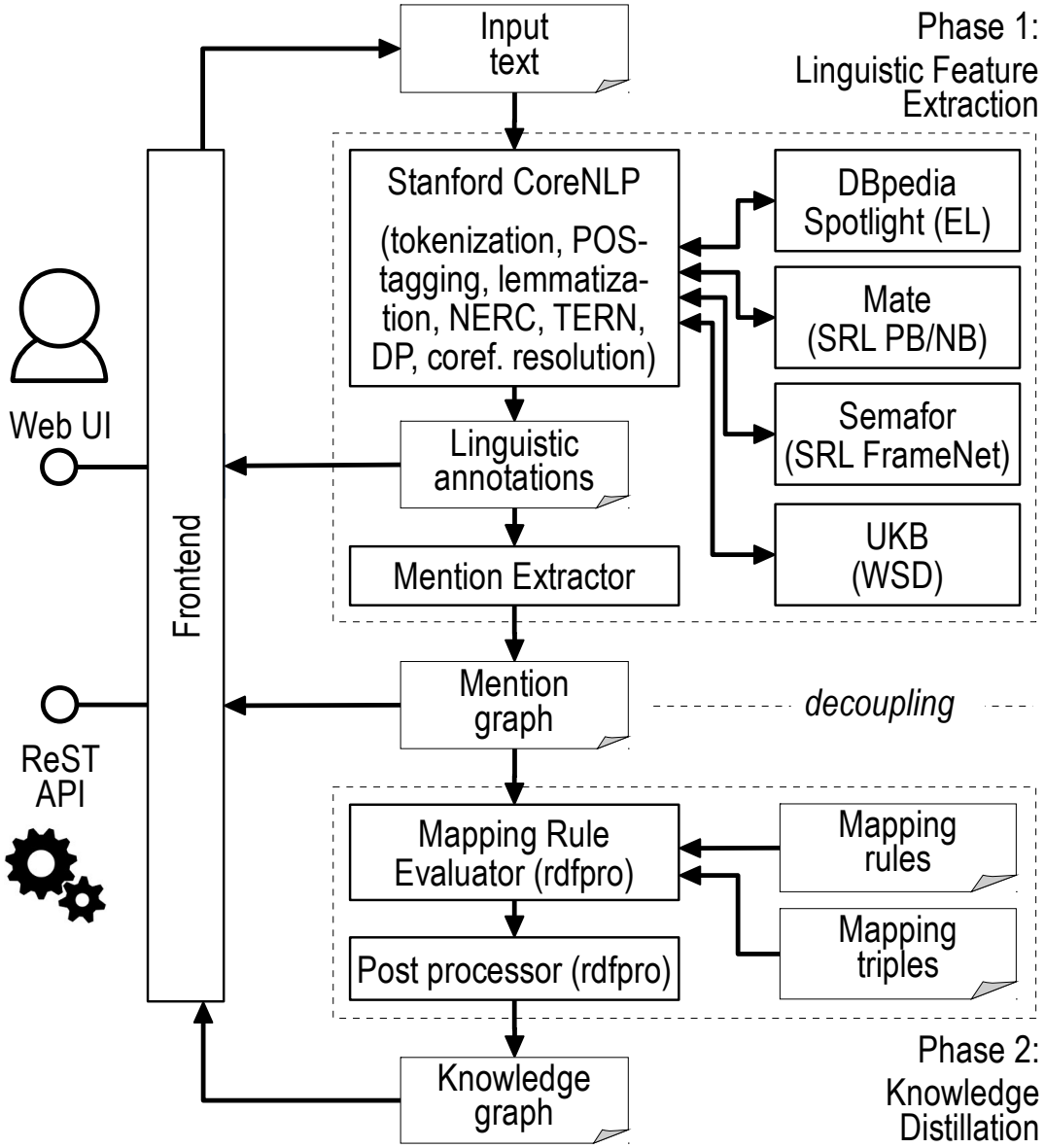
Implementation

PIKES
- Java 1.8 on Linux / Mac OS X
- open source (GPL)
- Maven project on GitHub
  https://github.com/dkmfbk/pikes

Integrated dependencies
- Stanford CoreNLP
- Mate-tools
- Semafor
- RDFpro

External dependencies
- Dbpedia Spotlight
- UKB

→ need separate install



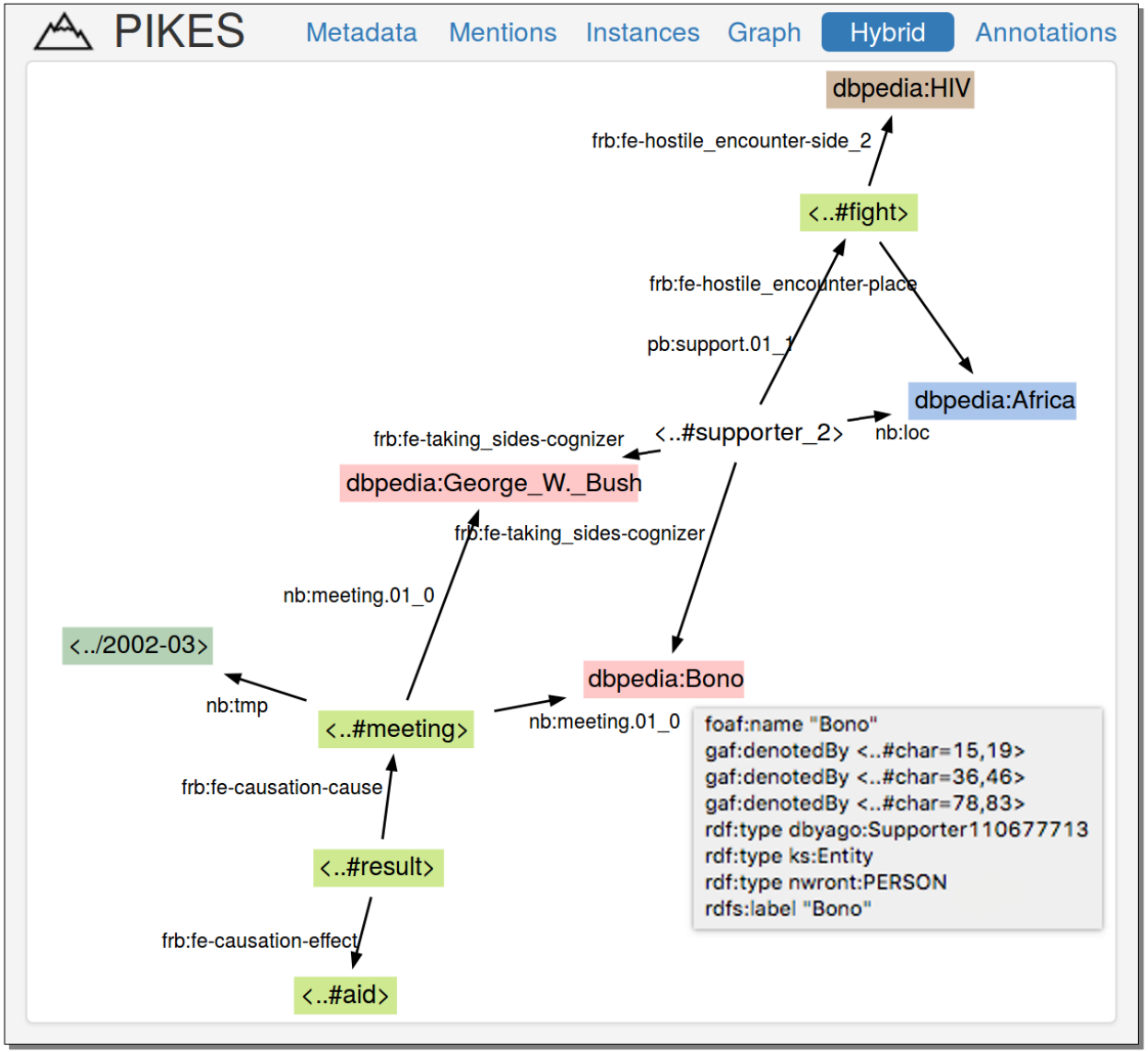
Implementation (2)

FONDAZIONE BRUNO KESSLER

PIKES UI for:

"G.W. Bush and Bono are very strong supporters of the fight of HIV in Africa. Their March 2002 meeting resulted in a 5 billion dollar aid."

http://pikes.fbk.eu/



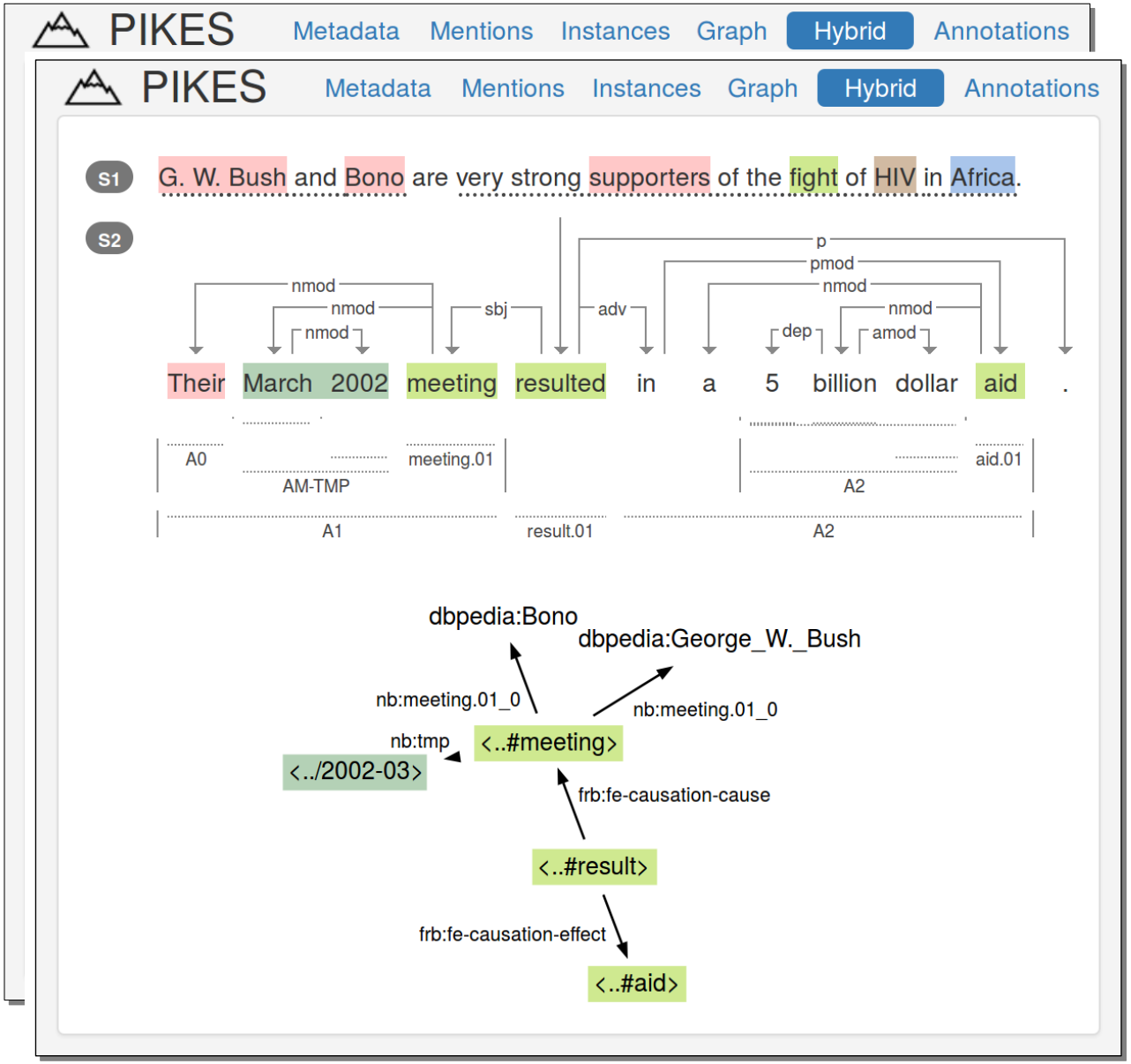
Implementation (3)

FONDAZIONE BRUNO KESSLER

PIKES UI for:

"G.W. Bush and Bono are very strong supporters of the fight of HIV in Africa. Their March 2002 meeting resulted in a 5 billion dollar aid."

http://pikes.fbk.eu/



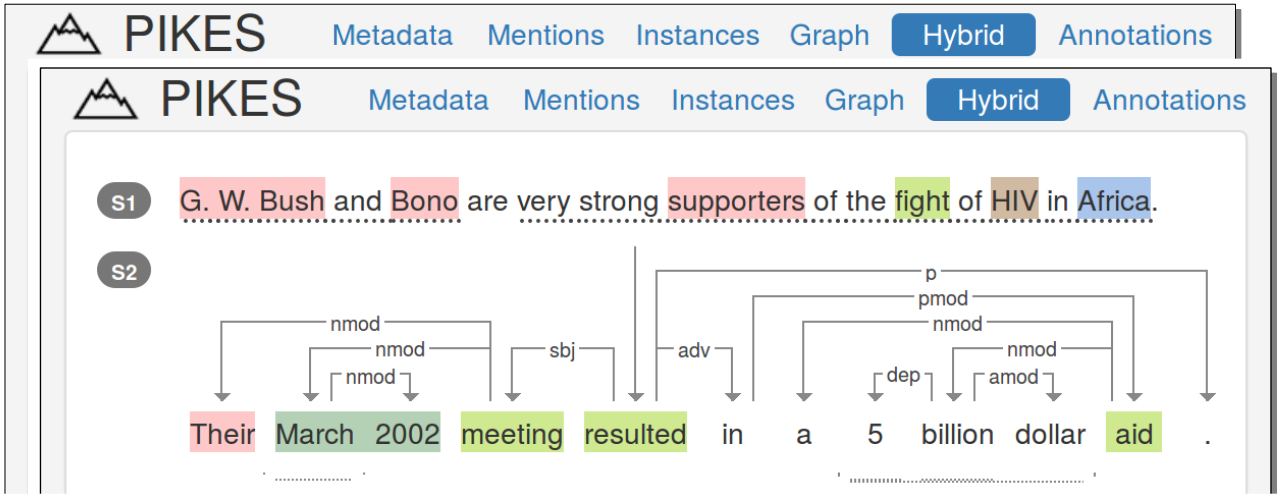
Implementation (4)

FONDAZIONE BRUNO KESSLER

PIKES UI for:

"G.W. Bush and Bono are very strong supporters of the fight of HIV in Africa. Their March 2002 meeting resulted in a 5 billion dollar aid."

http://pikes.fbk.eu/



PIKES — Metadata | **Mentions** | Instances | Graph | Hybrid | Annotations

Showing 1 to 11 of 11 entries (filtered from 31 total entries) — Search: dbpedia

id	nif:anchorOf	rdf:type	mention attributes	ks:denotes/ks:implies	ks:expresses
<../char=63,66>	HIV	ks:EntityMention ks:NameMention ks:Mention		dbpedia:HIV	1. dbpedia:HIV, foaf:name, HIV 2. dbpedia:HIV, rdf:type, ks:Entity 3. dbpedia:HIV, rdfs:label, HIV
<../char=0,10>	G. W. Bush	ks:NameMention ks:Mention ks:EntityMention		dbpedia:George_W._Bush	1. dbpedia:George_W._Bush, rdf:type, ks:Entity 2. dbpedia:George_W._Bush, foaf:name, G. W. Bush 3. dbpedia:George_W._Bush, rdf:type, nwront:PERSON
<../char=78,83>	Their	ks:Mention ks:EntityMention		dbpedia:Bono dbpedia:George_W._Bush	1. dbpedia:Bono, rdf:type, ks:Entity 2. dbpedia:George_W._Bush, rdf:type, ks:Entity
<../char=36,46>	supporters	ks:PredicateMention ks:ParticipationMention ks:EntityMention ks:Mention	ks:lemma: supporter ks:plural: true ks:sst: <../noun.person> ks:synset: <../10677713-n>	<../supporter_2> dbpedia:Bono dbpedia:George_W._Bush	1. <../supporter_2>, rdf:type, fs:frame-Taking_sides-back.v 2. <../supporter_2>, rdf:type, ks:Entity 3. <../supporter_2>, rdf:type, sem:Event

Evaluation

FONDAZIONE BRUNO KESSLER

Three evaluations:

① PIKES precision/recall on gold standard

② PIKES vs FRED precision/recall on simpler gold standard

③ PIKES throughput (and sampled precision) on large corpus

Gold Standard

FONDAZIONE BRUNO KESSLER

Gold text: 8 sentences (233 tokens) from: A. Gangemi. A comparison of knowledge extraction tools for the Semantic Web. ESWC 2013

S1	The lone Syrian rebel group with an explicit stamp of approval from Al Qaeda has become one of the uprising most effective fighting forces, posing a stark challenge to the United States and other countries that want to support the rebels but not Islamic extremists.
S2	Money flows to the group, the Nusra Front, from like-minded donors abroad.
S3	Its fighters, a small minority of the rebels, have the boldness and skill to storm fortified positions and lead other battalions to capture military bases and oil fields.
S4	As their successes mount, they gather more weapons and attract more fighters.
S5	The group is a direct offshoot of Al Qaeda in Iraq, Iraqi officials and former Iraqi insurgents say, which has contributed veteran fighters and weapons.
S6	“This is just a simple way of returning the favor to our Syrian brothers that fought with us on the lands of Iraq,” said a veteran of Al Qaeda in Iraq, who said he helped lead the Nusra Front’s efforts in Syria.
S7	The United States, sensing that time may be running out for Syria president Bashar al-Assad, hopes to isolate the group to prevent it from inheriting Syria.
S8	As the United States pushes the Syrian opposition to organize a viable alternative government, it plans to blacklist the Nusra Front as a terrorist organization, making it illegal for Americans to have financial dealings with the group and prompting similar sanctions from Europe.

Gold Standard (2)

Gold knowledge graph

- built manually by 2 annotators
- built sentence by sentence

137 instances

- entities or semantic frames (e.g., events)
- coreferring mentions → distinct instances + *owl:sameAs* links

166 triples

- frame types and roles based on VN, FN, PB, NB
- *owl:sameAs* between instances (COREF) and w.r.t. DBpedia (EL)

155 edges

- i.e., unlabeled instance-instance relations

Evaluation Methodology

FONDAZIONE BRUNO KESSLER

① Align TBoxes
- tools and gold standard use different VN/FN/PB/NB URIs

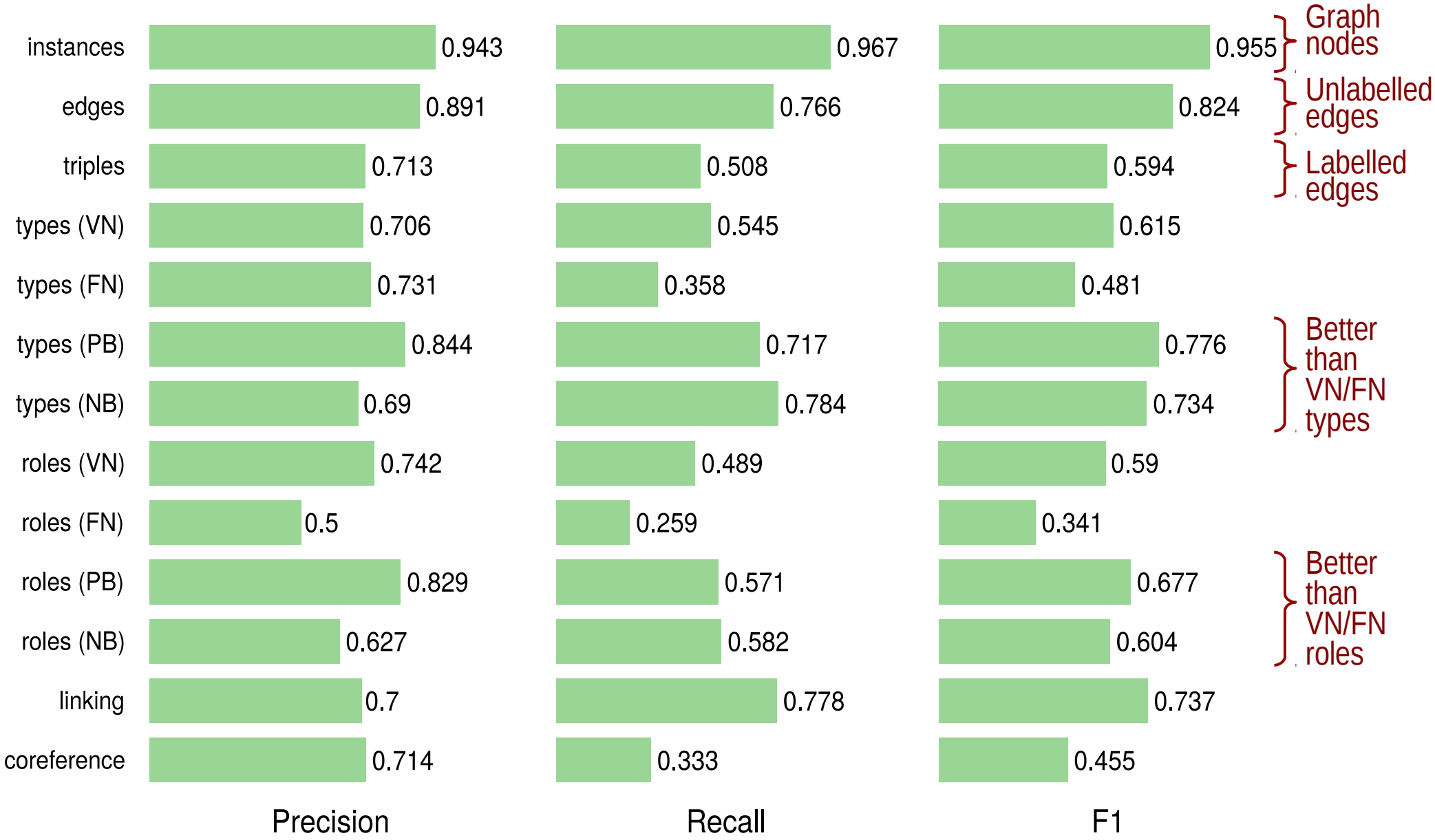
② Align instances
- maximize # common triples
- leverage groundings to mentions

③ Compare tool graph G_T and gold graph G_G
- for different components: instances, edges, triples (of specific kind)
- *true positives*: items in G_T and G_G
- *false negatives*: items in G_G but not G_T
- *false positives*: items in G_t but not in G_G
  - ignore irrelevant elements in G_T (manual operation)

④ Compute Precision (P), Recall (R), F1

PIKES against Gold Standard

FONDAZIONE BRUNO KESSLER



PIKES compared to FRED (1)

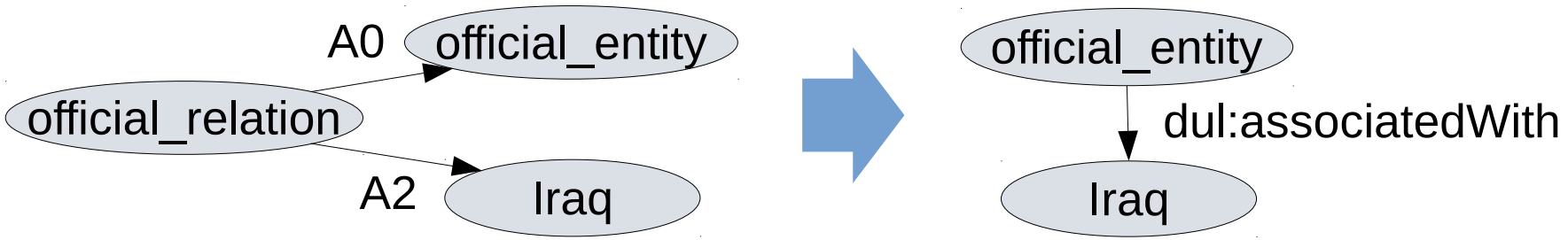
FONDAZIONE BRUNO KESSLER

FRED: Presutti, V., Draicchio, F., and Gangemi, A. Knowledge extraction based on discourse representation theory and linguistic frames. EKAW 2012

Comparison possible only on restricted gold standard

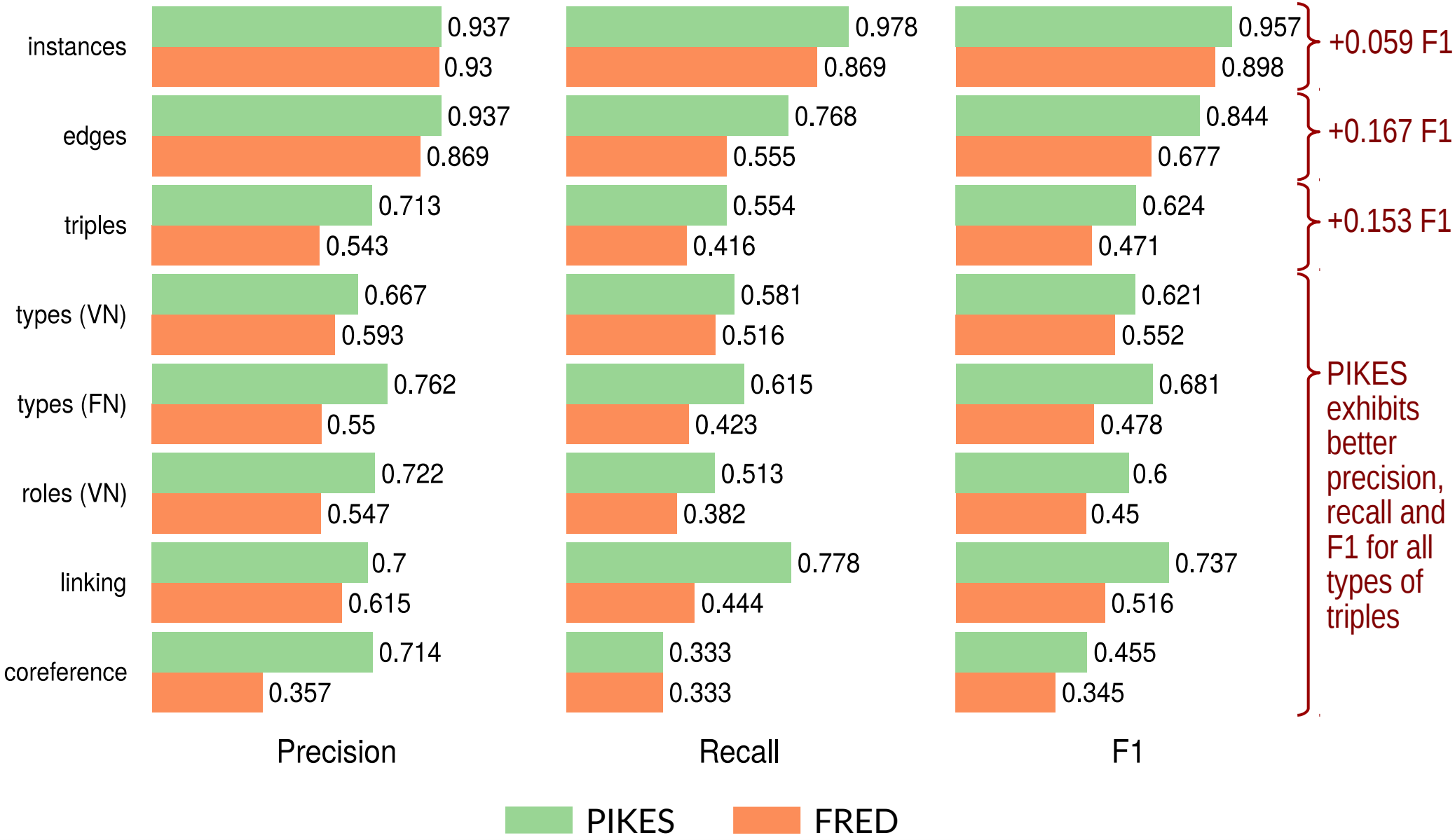
- no PB / NB frame types
- no PB / NB / FN* frame roles (* marked as :fe by FRED)
- nominal frames converted to binary relations

i.e., for *"Iraqi official"*



PIKES compared to FRED (2)

FONDAZIONE BRUNO KESSLER



Evaluation on Large Corpus

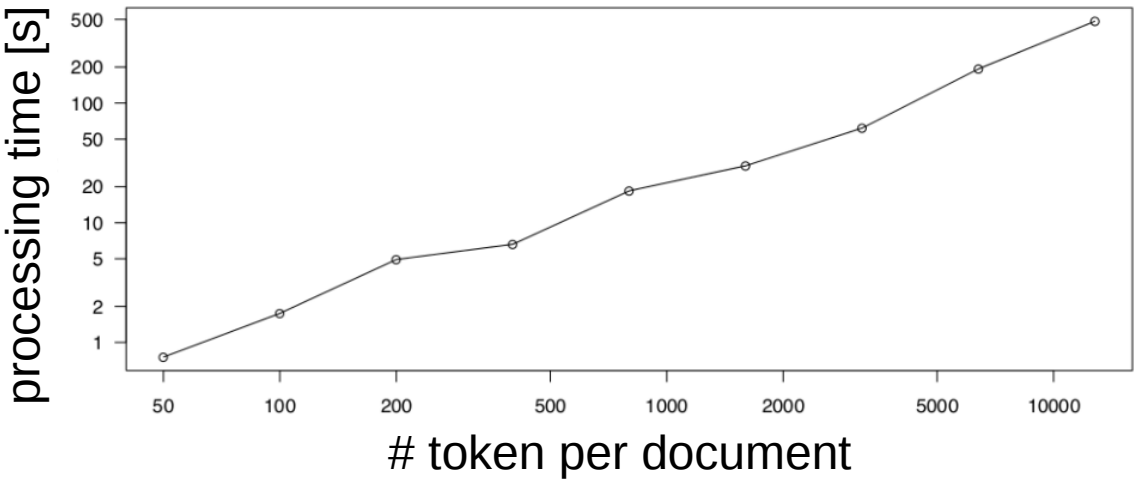
FONDAZIONE BRUNO KESSLER

Corpus: Simple English Wikipedia (dump date: April 6, 2015)

Server
- dual Xeon E5-2430 (24 cores)
- 192GB RAM
- 480GB SSD

Setup
- 16 PIKES instances
- 1 core, 7GB RAM each
- parallel page processing

Processing time
- 32 hours total
- 507 core hours



Item	# items	Throughput [item/h]	
		16 cores	1 core (*)
Documents	109,242	3,450	215
Sentences	1,584,406	50,000	3,125
Tokens	23,877,597	753,000	47,100

(*) Estimated based on time measured using 16 cores, to provide a normalized throughput value

Evaluation on Large Corpus (2)

FONDAZIONE BRUNO KESSLER

Knowledge Extraction results

- ~358M triples total
  → 2M resource layer, 283M mention layer, 72M instance layer
- more than 4M frame instances created
  → most frequent: use.01, play.01, know.01

Instance type	# Instances			# Triples
	Persons	Organizations	Locations	
linked to DBpedia	(1) 72K	19K	49K	(2) 26M
not linked to DBpedia	470K	173K	18K	46M
all	542K	192K	67K	72M

(1) most frequent: Pope, Jesus, Napoleon
(2) 1.7M annotations, 2.6M types, 21M participations (7M distinct frame-argument pairs)

Evaluation on Large Corpus (3)

FONDAZIONE BRUNO KESSLER

Type of triple	# Triples	Sampled precision (by evaluator)			
		Ev. 1	Ev. 2	Ev. 3	Avg.
Annotation	35	0.900	0.886	0.857	0.881
Type	35	0.943	0.771	0.857	0.857
PB/NB participation	130	0.904	0.785	0.850	0.846
All	200	0.910	0.800	0.853	0.854

Methodology

- sample 200 triples DBpedia instances with 1 mention each
- ask evaluators whether each triple is correct for its mention
  - 1=correct, 0=not correct, 0.5=only predicate is wrong

Fleiss' kappa coefficient k= 0.372

Mapping 0.5 → 0: precision=0.823, k = 0.407

Conclusions

FONDAZIONE BRUNO KESSLER

PIKES is

- a tool for Knowledge Extraction from English text
- extracting events and complex relations (semantic frames)
- representing all contents in RDF + named graph
- based on a 2-phase approach
  - linguistic feature extraction (via state-of-the-art NLP tools)
  - knowlege distillation (rule-based)

Benefits

- competitive with state of the art in terms of quality / throughput
- 2-phase decoupling allows to tune the two phases independenty

Future work

- integrate other NLP tasks and **PreMOn** – http://premon.fbk.eu/
- use PIKES for IR – **KE4IR** paper @ ESWC2016 – http://pikes.fbk.eu/ke4ir
- detect and repair inconsistencies in PIKES output (via ILP)

PREMON

FONDAZIONE BRUNO KESSLER

PreMOn = Predicate Model for Ontologies – http://premon.fbk.eu/

Linguistic Linked Data resource (grounded in Lemon) representing **predicate models** and **mapping** resources: PB, NB, VN, FN, Semlink

Homogeneously represents the **semantic classes** (e.g., rolesets in NB and PB, verb classes in VN, frames in FN) and **semantic roles**

Benefits:

- ease of **access** and **reuse** of predicate model data
- abstract **commonalities**, keep **peculiarities**
- automated **reasoning** and **SPARQL** querying
- **SRL annotations** of a text according NIF
- **interlinking** with third-party datasets

Availability: download / SPARQL endpoint / URI dereferencing

KE4IR

FONDAZIONE BRUNO KESSLER

KE4IR = Knowledge Extraction for Information Retrieval
http://pikes.fbk.eu/ke4ir

PIKES analysis of query and documents to improve IR performances

Semantics considered (e.g. ``astronomers influenced by Gauss")

- URIs: dbpedia:Carl_Friedrich_Gauss
- TYPE: dbyago:Astronomer109818343, dbyago:GermanMathematicians
- FRAME: framebase:Subjective_influence
- TIME: dbo:dateOfBirth (1777), dbo:dateOfDeath (1855)

Performances (on a IR dataset for SW: 331 documents / 35 queries):

Approach/System	Prec@1	Prec@5	Prec@10	NDCG	NDCG@10	MAP	MAP@10
Google	0.543	0.411	0.343	0.434	0.405	0.255	0.219
Textual	0.943	0.669	0.453	0.832	0.782	0.733	0.681
KE4IR	**0.971**	**0.680**	**0.474**	**0.854**	**0.806**	**0.758**	**0.713**
KE4IR vs. Textual	3.03%	1.71%	4.55%	2.64%	2.99%	3.50%	4.74%
p-value (paired t-test)	0.324	0.160	0.070	**0.003**	**0.015**	**0.024**	**0.029**
p-value (approx. random.)	1.000	0.496	0.111	**0.003**	**0.020**	**0.020**	**0.030**

Thank you! Questions?

Francesco Corcoglioniti

francesco@corcoglioniti.name

Marco Rospocher, Alessio Palmero Aprosio

Fondazione Bruno Kessler – IRST
Trento, Italy

SAC 2016
PISA, 06 April 2016

http://pikes.fbk.eu/