



Big Data Integration Benchmark

Andrea De Angelis & Maurizio Mazzei

and.deangelis@hotmail.com

mrz.mazzei@gmail.com



Sommario

- Introduzione
- Big Data Integration
 - Schema Alignment
 - Record Linkage
 - Data Fusion
- Benchmark & Dataset
 - Caso di studio: dataset di fotocamere
- Ground Truth
 - Schema Alignment
 - Record Linkage
- Knowledge Graph Data Model
- Conclusioni



Introduzione

- Progetto di ricerca del Laboratorio di Basi di Dati
 - In collaborazione con: AT&T Research Labs, Amazon Research
- Thanks to:
 - Paolo Merialdo
 - Donatella Firmani
 - Valter Crescenzi
 - Divesh Srivastava (AT&T Labs)
 - Xin Luna Dong (Amazon)
 - Vincenzo Di Cicco
 - Federico Piai



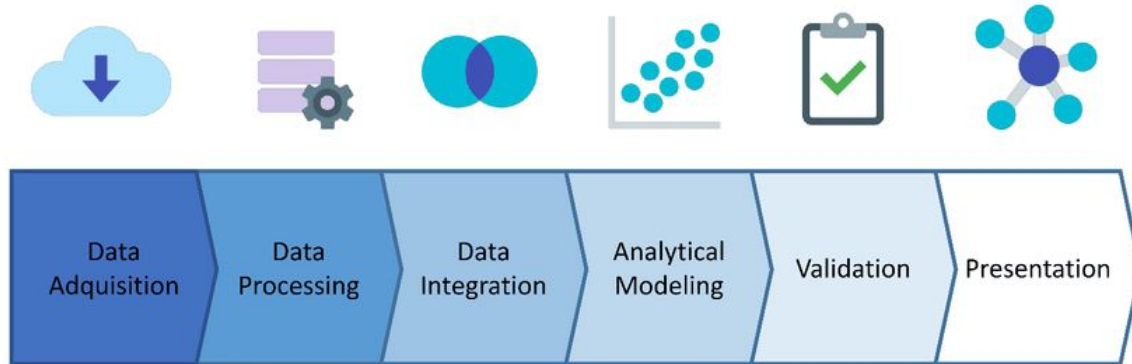
Obiettivo finale:

**Realizzare un benchmark per i
tasks di Big Data Integration**

—

Big Data Integration

Big Data Integration





Big Data Integration

- Big Data Integration = Data Integration + Big Data
- Data Integration = accesso unificato ai dati da sorgenti multiple ed autonome
- Big Data = ampio **volume** di dati provenienti da una grande **varietà** di sorgenti, che crescono **velocemente**, di cui bisogna verificare la **veridicità** (per poi trarne **valore**)



Data Integration

www.ebay.com (S1)

Brand	Model	Megapixel	Optical Zoom
Nikon	S6800	16 mpx	5x
Nikon	Coolpix S6800	16.0	5x
Panasonic	DMC-FZ200	7.2	3x
Sony	Alpha 7	16 MP	5x



Data Integration

www.ebay.com (S1)

Brand	Model	Megapixel	Optical Zoom
Nikon	S6800	16 mpx	5x
Nikon	Coolpix S6800	16.0	5x
Panasonic	DMC-FZ200	7.2	3x
Sony	Alpha 7	16 MP	8x

www.amazon.com (S2)

Manufacturer	Model	Resolution	Zoom	Display Size
Nikon	Coolpix S6800	16.0 megapixel	5x	3"
Sony	ILCE 7000	16.000.000 pixels	5x	3"
Panasonic	DMC SZ100	7.0	3x	3"
Fujifilm	S6800	16 mp	5x	3"

Data Integration

www.ebay.com (S1)

Brand	Model	Megapixel	Optical Zoom
Nikon	S6800	16 mpx	5x
Nikon	Coolpix S6800	16.0	5x
Panasonic	DMC-FZ200	7.2	3x
Sony	Alpha 7	16 MP	8x

www.gosale.net (S3)

Product	Digital Zoom	Shutter Speed
Panasonic LUMIX DMC-FZ200	12x	30 sec - ¼ sec
...

www.amazon.com (S2)

Manufacturer	Model	Resolution	Zoom	Display Size
Nikon	Coolpix S6800	16.0 megapixel	5x	3"
Sony	ILCE 7000	16.000.000 pixels	5x	3"
Panasonic	DMC SZ100	7.0	3x	3"
Fujifilm	S6800	16 mp	5x	3"

Data Integration

www.ebay.com (S1)

Brand	Model	Megapixel	Optical Zoom
Nikon	S6800	16 mpx	5x
Nikon	Coolpix S6800	16.0	5x
Panasonic	DMC-FZ200	7.2	3x
Sony	Alpha 7	16 MP	8x

www.gosale.net (S3)

Product Name	Digital Zoom	Shutter Speed
Panasonic LUMIX DMC-FZ200	12x	30 sec - 1/4 sec
...

www.amazon.com (S2)

Manufacturer	Model	Resolution	Zoom	Display Size
Nikon	Coolpix S6800	16.0 megapixel	5x	3"
Sony	ILCE 7000	16.000.000 pixels	5x	3"
Panasonic	DMC SZ100	7.2	3x	3"
Fujifilm	S6800	16 mp	5x	3"

- La Nikon Coolpix S6800 ha il display? Di che dimensione?
- Quanto zoom ottico ha la Panasonic LUMIX DMC-FZ200?



Data Integration - pipeline





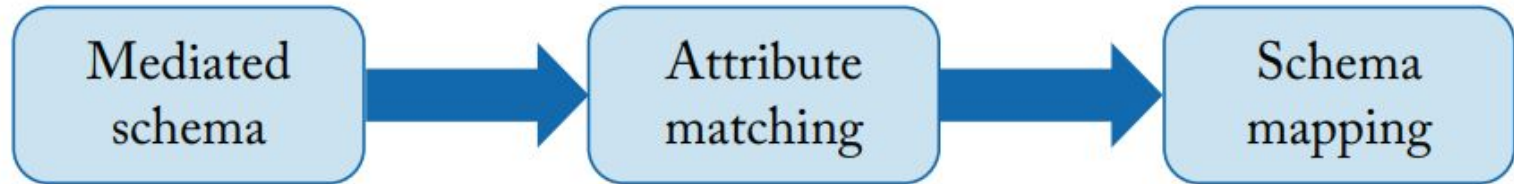
Data Integration - schema alignment

- Obiettivo: integrare gli schemi delle diverse sorgenti di un vertical
- Quali **attributi** hanno la stessa **semantica**? Quali no?

Due attributi potrebbero rappresentare:

- La stessa informazione concettuale modellata diversamente
 - es. “16MP” vs “16.000.000 pixels”
- Differenti informazioni concettuali modellate in maniera simile
 - es. “Optical Zoom” vs “Digital Zoom”
- ⇒ Ambiguità semantica

Data Integration - schema alignment





Data Integration - schema alignment

- Mediated schema (o integrated schema o global schema)
 - fornisce una vista unificata dei dati a disposizione
 - spesso creato manualmente
- Attribute matching
 - mappa gli attributi di sorgente agli attributi del mediated schema
- Schema mapping
 - mappa ogni schema di sorgente con il mediated schema
 - come ottenere le risorse nel mediated schema? (query reformulation)

Brand	Model	Resolution	Optical Zoom	Digital Zoom	Display Size	Shutter Speed
S1.Brand S2.Manufacturer S3.Product_Name	S1.Model S2.Model S3.Product_Name	S1.Megapixel S2.Resolution	S1.Optical_Zoom S2.Zoom	S3.Digital_Zoom	S2.Display_Size	S3.Shutter_Speed

Data Integration - record linkage


- Obiettivo: riconoscere le **istanze** che corrispondono alla stessa **entità** del mondo reale
- Quali **istanze** corrispondono alla stessa **entità**? Quali no?

Due istanze potrebbero avere rappresentazioni molto simili e rappresentare entità diverse

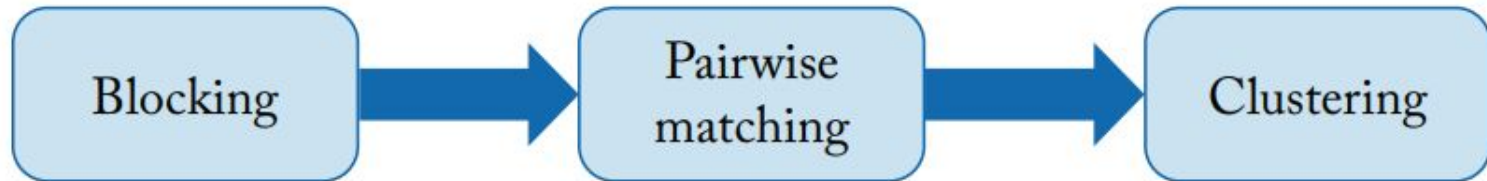
- ⇒ Ambiguità di rappresentazione dell'istanza

Brand	Model	Megapixel	Optical Zoom
Nikon	S6800	16 mpx	5x

Manufacturer	Model	Resolution	Zoom	Display Size
Fujifilm	S6800	16 mp	5x	3"



Data Integration - record linkage



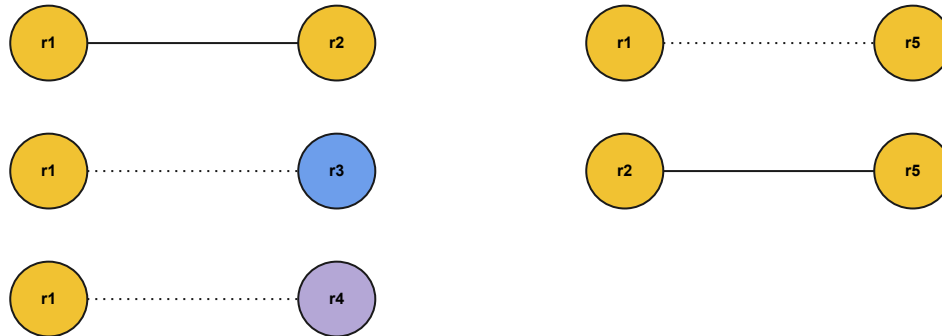
Data Integration - record linkage

- Pairwise Matching
 - compara una coppia di istanze e decide se rappresentano la stessa entità oppure no
 - per tutte le possibili coppie di istanze possibili \Rightarrow problema quadratico $O(n^2)$



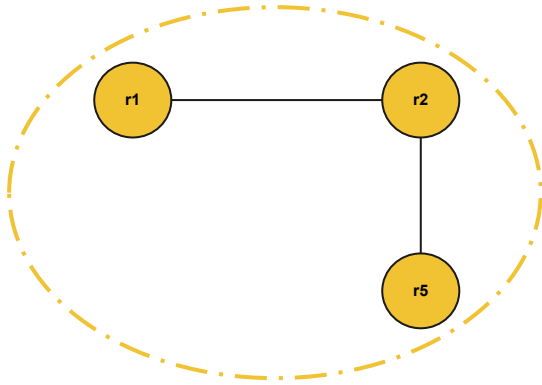
Data Integration - record linkage

- Clustering
 - permette di risolvere eventuali inconsistenze nella fase di pairwise matching
 - ogni partizione deve rappresentare un'entità del mondo reale



Data Integration - record linkage

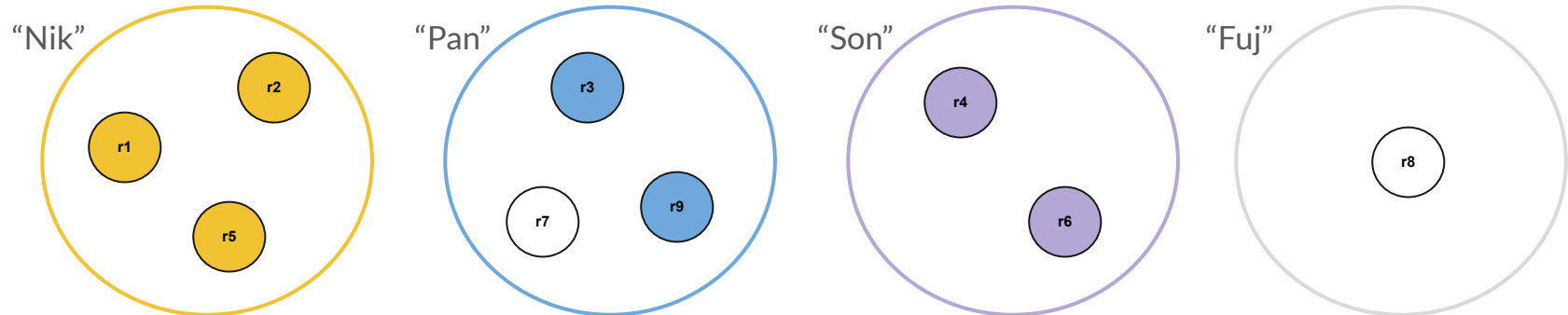
- Clustering
 - permette di risolvere eventuali inconsistenze nella fase di pairwise matching
 - ogni partizione deve rappresentare un'entità del mondo reale



Cluster per la Nikon Coolpix S6800

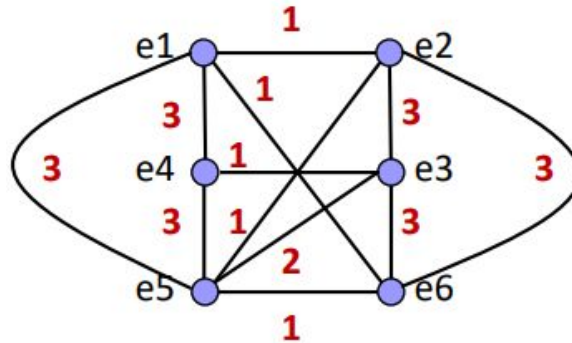
Data Integration - record linkage

- Blocking
 - compromesso tra la recall dei cluster e l'efficienza in termini computazionali
 - si partizionano le istanze in blocchi sulla base di una blocking function
 - si fa pairwise matching localmente ai blocchi creati
 - es. blocking function: “primi 3 caratteri dell’attributo Brand dello schema mediato”



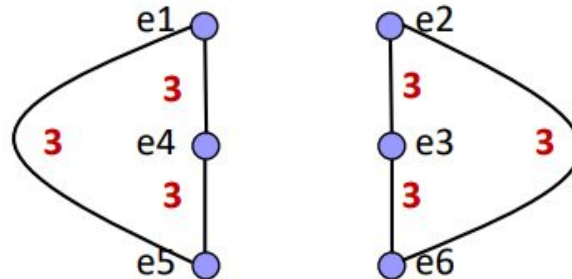
Data Integration - record linkage

- Metablocking
 - “Meta-Blocking: Taking Entity Resolution to the Next Level” [Papadakis et al. - TKDE ‘13]
 - utilizzo di multiple blocking function (che possono sovrapporsi)
 - realizzazione del metablocking graph (peso sull’arco = numero di blocchi sovrapposti)
 - pruning alla ricerca della migliore suddivisione in blocchi



Data Integration - record linkage

- Metablocking
 - “Meta-Blocking: Taking Entity Resolution to the Next Level” [Papadakis et al. - TKDE ‘13]
 - utilizzo di multiple blocking function (che possono sovrapporsi)
 - realizzazione del metablocking graph (peso sull’arco = numero di blocchi sovrapposti)
 - pruning alla ricerca della migliore suddivisione in blocchi





Data Integration - data fusion

- Obiettivo: risolvere le inconsistenze nei dati
- Quale **valore** associato ad un **attributo** è corretto? Quale no?

Dato un attributo allineato, le varie sorgenti potrebbero fornire un valore diverso per la stessa entità:

- ⇒ Inconsistenza dei dati

Brand	Model	...	Optical Zoom
Sony	Alpha 7	...	8x
Sony	ILCE 7000	...	5x



Big Data Integration

- Cosa cambia quando si passa ad un contesto big data?
 - Eterogeneità dei dati
 - le sorgenti non hanno uno schema fisso (come invece succedeva negli esempi precedenti)
 - Grande volume di dati
 - sono necessarie soluzioni Big Data per l'elaborazione dei dati (es. MapReduce, Spark, ecc.)

Benchmark



Benchmark

- Strumento che valuta le **prestazioni** e fornisce uno **score** (+ eventuali indicatori)
 - Precision, Recall, F-Measure
- Caratteristiche “How to build a benchmark” [Kistowski et al. - ICPE ‘15]
 - Riproducibilità: score consistente rispetto alla configurazione
 - Equità: accessibile da chiunque
 - Usabilità: utilizzabile in qualunque environment
- Un dataset
- Una ground truth per ogni task di Data Integration
 - non ci siamo occupati di Data Fusion!



Dataset

vertical	# sources	# pages with extracted specs	# distinct attribute names
camera	1083	94710	38663
notebook	577	54852	14180
headphone	464	49975	7574
monitor	929	48067	14594
shoes	131	23481	2489
tv	826	21835	14959
software	372	12626	6332
toilets	81	5319	1464
sunglasses	386	4871	1137
cutlery	294	3302	1742



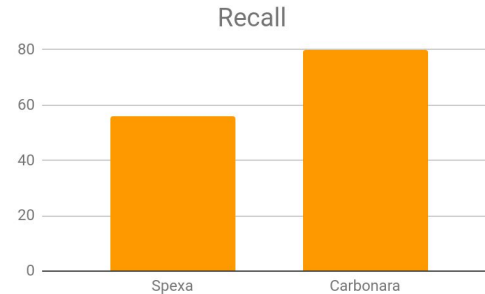
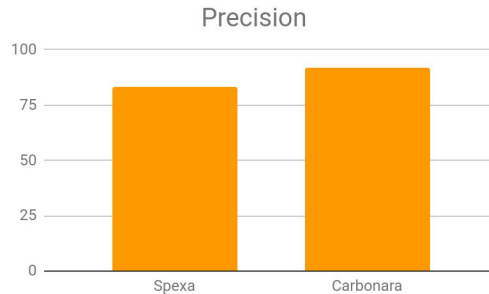
Dataset - camera - motivazioni

- Dati del mondo reale, non sintetici
- Impegnativo rispetto ai classici dataset accademici
 - Grande **varietà** di specifiche dei prodotti
 - Elevato **volume** di sorgenti
 - Eterogeneità all'interno della singola sorgente
 - Eterogeneità tra le varie sorgenti

Dataset - camera - estrazione dati

- **Carbonara extractor**


- classificatore basato su rete neurale per la selezione di liste e tabelle di specifiche di prodotto
- produzione di un insieme di coppie <nome_attributo, valore_attributo> per ogni pagina web di ogni sorgente



```

{
  "<page title>": "Samsung Smart WB50F Digital Camera White Price in India wit
  "additional features": "Color\nWhite",
  "brand": "Samsung",
  "connectivity system req": "USB\nUSB 2.0",
  "dimension": "Dimensions\n101 x 68 x 27.1 mm\nWeight\n157 gms",
  "display": "Display Type\nLCD\nScreen Size\n3 Inches",
  "general features": "Brand\nSamsung\nAnnounced\n2014, February\nStatus\nAvai
  "lens": "Auto Focus\nCenter AF, Face Detection, Multi AF\nFocal Length\n4.3
  "media software": "Memory Card Type\nSD, SDHC, SDXC",
  "optical sensor resolution in megapixel": "16.2 MP",
  "other features": "ISO Rating\nAuto / 80 / 100 / 200 / 400 / 800 / 1600 / 32
  "pixels": "Optical Sensor Resolution (in MegaPixel)\n16.2 MP",
  "sensor": "Sensor Type\nCCD Sensor\nSensor Size\n1/2.3 Inches",
  "sensor type": "CCD Sensor",
  "shutter speed": "Maximum Shutter Speed\n1/2000 sec\nMinimum Shutter Speed\r
  "zoom": "Optical Zoom\n12x\nDigital Zoom\n2x"
}

```



Dataset - camera - selezione sorgenti

- Obiettivo: selezionare 25 sorgenti
- Criteri di selezione:
 1. sufficiente quantità di informazione (evitando sorgenti copia)
 2. sorgenti che condividono istanze di entità da integrare

Dataset - camera - selezione sorgenti

Sufficiente quantità di informazioni (evitando sorgenti copia)

- Primo filtraggio (3-3-50)
 - attributi presenti in almeno 3 pagine della stessa sorgente
 - file di estrazione con almeno 3 attributi
 - sorgenti con almeno 100 file di estrazione
- 63 sorgenti selezionate

	A	H	I	J	K	L	M
1	site	choice	# url 3-3-50	# not empty json 3-3-50	# attribute names 3-3-50	# distinct attribute names 3-3-50	avg attribute names 3-3-50
2	www.ebay.com	yes	14274	14274	138233	1987	9.68
3	www.ebay.ca	NO (ebay)	8245	8245	86788	1478	10.53
4	www.alibaba.com	yes	7972	7972	180766	1590	22.68
5	www.ebay.co.uk	NO (ebay)	3614	3614	36536	678	10.11
6	alibaba.com	NO	2568	2568	25802	594	10.05
7	www.ebay.in	NO	1940	1940	63554	1147	32.76
8	www.ebay.com.sg	NO	1899	1899	23297	734	12.27
9	uae.souq.com	yes	1380	1380	19213	81	13.92



Dataset - camera - selezione sorgenti

Sorgenti che condividono istanze di entità da integrare

- Selezione basata sui risultati di un processo di linkage descritto in “Big Data Linkage for Product Specification Pages” [Qiu, Barbosa, Crescenzi, Merialdo, Srivastava - SIGMOD '18] basato su ID di entità estratti dalle pagine HTML
 - Approssimazione della distribuzione della grandezza dei cluster delle entità in matching
- 27 sorgenti selezionate

Dataset - camera - selezione sorgenti finale

- 27 sorgenti
 - 24 sorgenti pulite
 - 3 sorgenti sporche
- ~30K pagine selezionate
 - ogni pagina descrive una camera
- ~20 attributi per file di estrazione

	source	# pages
0	www.ebay.com	14274
1	www.alibaba.com	7972
2	www.gosale.com	1002
3	www.buzzillions.com	832
4	search.greenvilleadvocate.com	785
5	www.priceme.co.nz	740
6	www.shopmania.in	630
7	search.tryondailybulletin.com	591
8	www.eglobalcentral.co.uk	571
9	yellowpages.observer-reporter.com	533
10	www.shopbot.com.au	516
11	www.pricedekho.com	366
12	buy.net	358
13	www.mypiceindia.com	347
14	www.price-hunt.com	327
15	www.pcconnection.com	211
16	cammarkt.com	198
17	www.walmart.com	195
18	www.henrys.com	181
19	www.canon-europe.com	164
20	www.flipkart.com	157
21	www.wexphotographic.com	147
22	www.garricks.com.au	130
23	www.ukdigitalcameras.co.uk	129
24	www.camerafarm.com.au	120
25	www.cambuy.com.au	118
26	www.ilgs.net	102

GT Schema Alignment



GT schema alignment

- 94 cluster creati manualmente da una studentessa (grazie Siria 😊)
 - ogni cluster rappresenta una caratteristica di prodotto
 - alcuni cluster sono **supercluster**
 - comprendono attributi di sorgente con livello di dettaglio diverso
 - “weight” → “weight including batteries”, “body weight”, “weight”
 - livello di granularità dello schema alignment è infinitamente espandibile
- attributi difficili da allineare
 - granularità differenti (“dimensions”, “height”, “width”, “depth”)
 - attributi categorici (“bluetooth”: “YES”, “bluetooth”: “5.0”)
 - attributi concatenati (“battery”: “1x NP-W126 Rechargeable Lithium-Ion Battery Pack, 7.2VDC, 1260 mAh”)
 - attributi ambigui (“camera type”: “waterproof”, wifi)

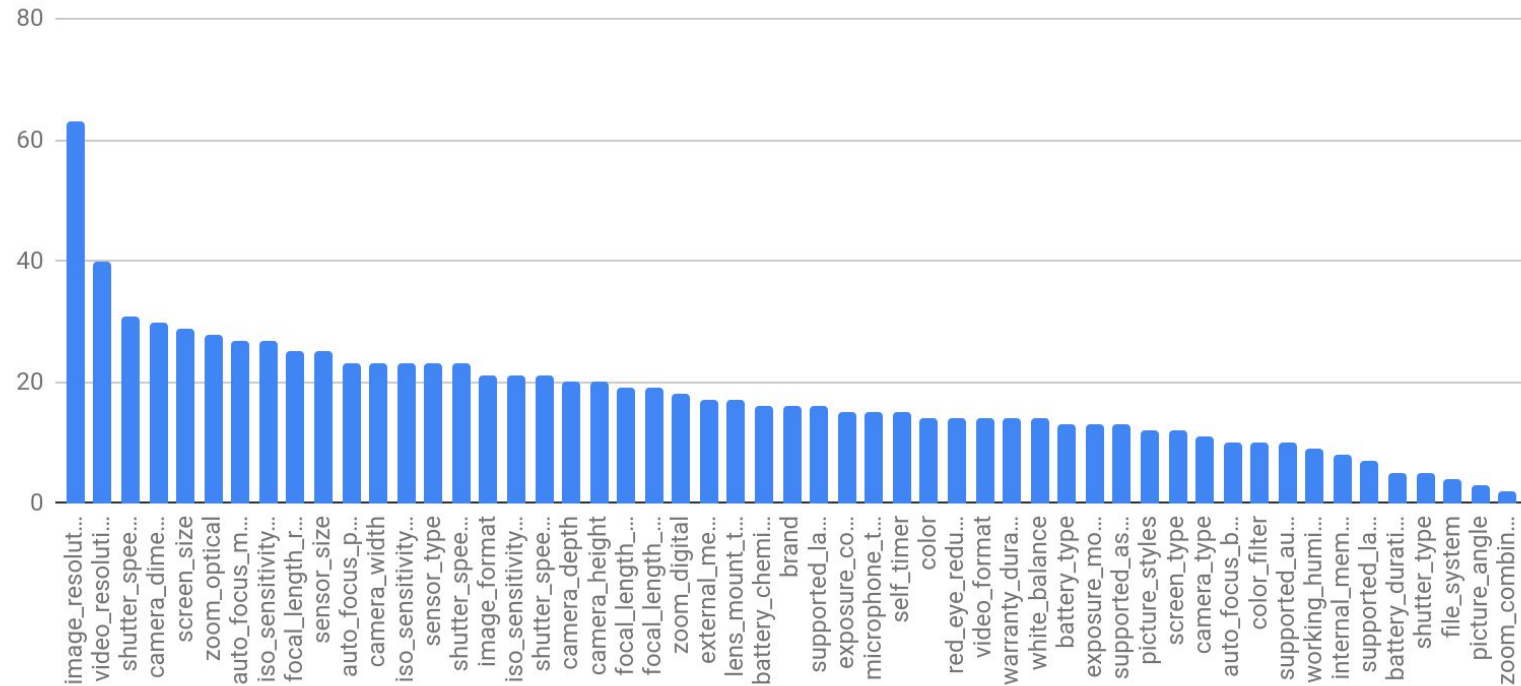


GT schema alignment

- 53 attributi (cluster) selezionati (target attributes) per lo **schema mediato**
- 943 mapping tra source attributes e target attributes

source	attribute_name	predicate_name
buy.net	total pixels	image_resolution
buy.net	effective megapixels	image_resolution
cammarkt.com	image resolutions	image_resolution
cammarkt.com	megapixels	image_resolution
cammarkt.com	camera resolution	image_resolution
cammarkt.com	resolution	image_resolution
www.alibaba.com	image sensor	image_resolution
www.alibaba.com	total pixel no	image_resolution
www.alibaba.com	resolution	image_resolution
www.alibaba.com	max image resolution	image_resolution
www.alibaba.com	effective pixels	image_resolution
www.buzzillions.com	camera pixel count	image_resolution
www.buzzillions.com	still image resolution max	image_resolution
www.buzzillions.com	megapixels	image_resolution
www.buzzillions.com	pixel count	image_resolution
www.buzzillions.com	resolution	image_resolution
www.buzzillions.com	effective megapixel count	image_resolution
www.cambuy.com.au	imaging sensor effective pixels	image_resolution
www.cambuy.com.au	raw	image_resolution
www.cambuy.com.au	image size pixels	image_resolution
www.cambuy.com.au	total pixels	image_resolution
www.cambuy.com.au	effective pixels	image_resolution
www.cambuy.com.au	number of effective pixels	image_resolution

Mapping multipli verso l'attributo target "image_resolution"



Distribuzione attributi target rispetto al numero di mapping con attributi di sorgente

GT Record Linkage



GT record linkage - validation set

- Base di comparazione per risultati di approcci di Record Linkage
- Creato manualmente
- Comprende 6 entità/cluster

product name	# sources	# instances	hours of work	head/tail
NIKON D7000	14	131	7	HEAD
NIKON COOLPIX S9500	8	25	4	HEAD
PANASONIC FZ200	12	26	3	HEAD
CANON EOS REBEL T3i	3	135	4	TAIL
FUJIFILM S8100FD	3	7	1	TAIL
NIKON D50	2	46	1	TAIL

cluster id	entity	source	page	json	note
1	13179 nikon d7000 dslr body excellent condition 018208254682	www.ebay.com	http://enit.inf.it	http://enit.inf.it	BODY
1	clean nikon d7000 16 2mp bundle battery charger lcd cover ir remote nice 018208254682	www.ebay.com	http://enit.inf.it	http://enit.inf.it	BODY
1	gently used nikon d7000 16 mp slr 018208254682	www.ebay.com	http://enit.inf.it	http://enit.inf.it	BODY
1	gently used nikon d7000 16 mp slr everything shown 018208254682	www.ebay.com	http://enit.inf.it	http://enit.inf.it	BODY
1	gently used nikon d7000 16 mp slr everything shown 018208254682	www.ebay.com	http://enit.inf.it	http://enit.inf.it	BODY
1	new nikon 16 mp slr d7000 body dx mat black free shipping 018208254682	www.ebay.com	http://enit.inf.it	http://enit.inf.it	BODY
1	new nikon d7000 slr 4 complete dslr kit 24gb top value 018208254682	www.ebay.com	http://enit.inf.it	http://enit.inf.it	KIT
1	new nikon d7000 slr body kit 8gb more 018208254682	www.ebay.com	http://enit.inf.it	http://enit.inf.it	KIT
1	nikon d7000	www.ilgs.net	http://enit.inf.it	http://enit.inf.it	BODY
1	nikon d7000	www.priceme.co.nz	http://enit.inf.it	http://enit.inf.it	BODY
1	nikon d7000	www.shopbot.com.au	http://enit.inf.it	http://enit.inf.it	BODY
1	nikon d7000 / 18-105mm 55-200mm vr kit	www.shopbot.com.au	http://enit.inf.it	http://enit.inf.it	KIT
1	nikon d7000 / 18-105mm vr kit	www.shopbot.com.au	http://enit.inf.it	http://enit.inf.it	KIT
1	nikon d7000 / 18-200mm vr kit	www.shopbot.com.au	http://enit.inf.it	http://enit.inf.it	KIT
1	nikon d7000 / 18-55mm 55-200mm vr kit	www.shopbot.com.au	http://enit.inf.it	http://enit.inf.it	KIT
1	nikon d7000 / 18-55mm 55-300mm vr kit	www.shopbot.com.au	http://enit.inf.it	http://enit.inf.it	KIT
1	nikon d7000 / 24-120mm vr kit	www.shopbot.com.au	http://enit.inf.it	http://enit.inf.it	KIT

Estratto cluster Nikon D7000



GT record linkage

- Requisito: recall \Rightarrow 1.0
- Obiettivo iniziale: creazione in maniera automatica di cluster ragionevoli, che siano punto di partenza per un affinamento da parte di un umano
- Differenti approcci possibili:
 - ID-based
 - schema-based
 - schema-agnostic



GT record linkage - metodo schema-agnostic

- L'idea:
 - attributi più frequenti dovrebbero essere meno discriminanti
 - attributi meno frequenti dovrebbero essere più discriminanti
- Per ogni JSON di estrazione
 - tokenizzazione valori degli attributi
 - calcolo peso per ogni token
 - calcolo similarità tra insieme corrente di token e insieme di token di ogni altro JSON di estrazione

<u>attribute name</u>	<u>attribute value</u>
resolution	16 mpx
optical zoom	5x
brand	Canon
family	EOS Rebel
model	T3

<u>attribute name</u>	<u>attribute value</u>
resolution	16 mpx
optical zoom	5x
brand	Canon
family	EOS Rebel
model	T5



GT record linkage - metodo schema-agnostic

Peso token

- Dati una sorgente S ed un token W:

$$weight_S(W) = \frac{N}{dfW} \qquad inverse_frequency_S(W) = \frac{weight(W)}{\max(weights(S))}$$

- N = #pagine sorgente S
- dfW = #pagine sorgente S contenenti W
- $\max(weights(S))$ = peso massimo nella sorgente S



GT record linkage - metodo schema-agnostic

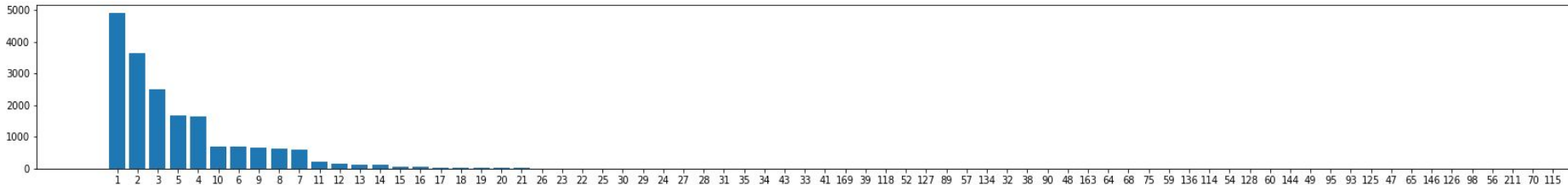
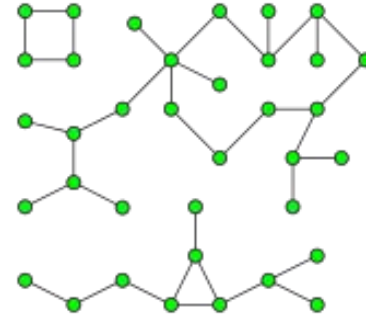
Misura di similarità

- Dati:
 - un insieme di token $E1$ della sorgente A
 - un insieme di token $E2$ della sorgente B

$$similarity_score = \frac{\sum \min(weight_A(x), weight_B(x)), \forall x \in E1 \cap E2}{|E1 \cap E2|}$$

GT record linkage - metodo schema-agnostic

- Grafo con:
 - un nodo per ogni file di estrazione
 - un arco tra due nodi se il loro $\text{similarity_score} > \text{soglia } x$



Distribuzione grandezze componenti connesse



GT record linkage - metodo schema-based

- Stesse considerazioni del metodo schema-agnostic
- Sfruttiamo lo schema mediato di cui conosciamo i mapping
- Per ogni prodotto
 - compara solo i token relativi ad attributi mappati sullo stesso attributo target
- Risultato
 - nessuna componente gigante
 - troppi nodi isolati

GT record linkage - metodo schema-agnostic

#2

Idea

- All'interno del titolo delle pagine html di un prodotto in un e-commerce c'è spesso tutto ciò di cui abbiamo bisogno per stabilire l'entità
 - Consideriamo solo il titolo

Nikon D3500 Fotocamera Reflex Digitale con Obiettivo Nikkor AF-P DX 18–55 VR e AF-P DX 70–300 VR, 24.2 Megapixel, LCD 3", SD da 16 GB 300x Premium Lexar, Nero [Nital Card: 4 Anni di Garanzia]

di Nikon



20 recensioni clienti | 48 domande con risposta



GT record linkage - metodo schema-agnostic #2

Pulizia dei titoli

- rimozione pattern comuni
 - es: "...at eBay.com" oppure "...from CamBuy in Sidney"
- lower-casing
- rimozioni parole comuni non discriminanti
 - es: "digital", "camera", "with", "compact", ...
- rimozione simboli e caratteri isolati
- etc.

GT record linkage - metodo schema-agnostic

#2

- Creazione bigrammi sovrapposti
 - es: "polaroid is426 16 megapixel" →
[('polaroid', 'is426'), ('is426', '16'), ('16', 'megapixel')]
- Blocking in base al numero di bigrammi in comune
 - componenti connesse non soddisfacenti

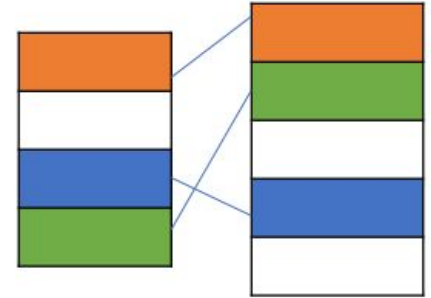
MOST FREQUENT BIGRAMS	
('canon', 'eos')	2365
('canon', 'powershot')	2201
('nikon', 'coolpix')	1972
('mp', 'slr')	1828
('mp', 'black')	1452
('slr', 'black')	1330
('body', 'only')	1315
('16', 'mp')	1158
('12', 'mp')	1132
('cyber', 'shot')	918
('eos', 'rebel')	915
('sony', 'cyber')	888
('10', 'mp')	841
('panasonic', 'lumix')	827
('fujifilm', 'finepix')	810
('18', '55mm')	799
('shot', 'dsc')	792
('black', 'body')	787
('mp', 'silver')	700
('optical', 'zoom')	685

GT record linkage - cambio di rotta

- Da approccio di clustering a pair-wise matching



clustering



pair-wise
matching



GT record linkage - oracle-based

- Ispirato da “Online Entity Resolution Using an Oracle” [Firmani, Saha, Srivastava - VLDB ‘16]
- Ha bisogno di oracoli umani ⇨ crowdsourcing
- Oracoli confermano o smentiscono il matching tra due istanze proposte

GT record linkage - oracle-based

Samples
E_1
E_2
...
E_{150}

Normals
N_1
N_2
...
$N_{29.787}$

UnionFind
$[\{E_1\}, \{E_2\}, \dots, \{E_{150}\}]$

Samples contiene 100 disjoint sets (i.e., entità del mondo reale)

- distribuite uniformemente (head, middle, tail)
- Head = molte sorgenti, molte istanze
- Middle = molte sorgenti, poche istanze
- Tail = poche sorgenti, poche istanze

GT record linkage - oracle-based

Samples
E_1
E_2
...
E_{150}

Normals
N_1
N_2
...
$N_{29.787}$

- Quali sono i 5 nodi di *Normals* che più probabilmente sono in match con E_1 ?
- C'è bisogno di una euristica per ordinare *Normals*

GT record linkage - oracle-based

Samples
E_1
E_2
...
E_{150}

Normals
N_1
N_2
...
$N_{29.787}$

- Quali sono i 5 nodi di *Normals* che più probabilmente sono in match con E_1 ?
- C'è bisogno di una **euristica** per ordinare *Normals*
 - Calcola il **benefit** per ogni nodo in *Normals*
 - Usando un grafo pre-calcolato con archi pesati

GT record linkage - oracle-based

Samples
E_1
E_2
...
E_{150}

Normals
N_1
N_2
...
$N_{29.787}$

- Quali sono i 5 nodi di *Normals* che più probabilmente sono in match con E_1 ?
- C'è bisogno di una **euristica** per ordinare *Normals*
 - Calcola il **benefit** per ogni nodo in *Normals*
 - Usando un grafo pre-calcolato con archi pesati
 - Sottografo indotto da $\{E_1, N_1\}$
 - Benefit = grado pesato di N_1

GT record linkage - oracle-based

Samples
E_1
E_2
...
E_{150}

Normals
N_1
N_2
...
$N_{29.787}$

- Quali sono i 5 nodi di *Normals* che più probabilmente sono in match con E_1 ?
- C'è bisogno di una **euristica** per ordinare *Normals*
 - Calcola il **benefit** per ogni nodo in *Normals*
 - Usando un grafo pre-calcolato con archi pesati
 - Sottografo indotto da $\{E_1, N_2\}$
 - Benefit = grado pesato di N_2

GT record linkage - oracle-based

Samples
E_1
E_2
...
E_{150}

Normals
N_1
N_2
...
$N_{29.787}$

- Quali sono i 5 nodi di *Normals* che più probabilmente sono in match con E_1 ?
- C'è bisogno di una **euristica** per ordinare *Normals*
 - Calcola il **benefit** per ogni nodo in *Normals*
 - Usando un grafo pre-calcolato con archi pesati
 - Sottografo indotto da $\{E_1, N_{29.787}\}$
 - Benefit = grado pesato di $N_{29.787}$

GT record linkage - oracle-based

Samples
E_1
E_2
...
E_{150}

Normals
N_{100}
N_{20}
...
$N_{29.787}$

- Entità target: E_1
- Seleziona i top 5 nodi da *Normals*
 - ogni nodo è una domanda per l'oracolo
 - es: N_{100} ✓
 - N_{20} ✓
 - N_8 ✓
 - N_3 ✓
 - N_{250} ✓

ordinata in base alla funzione di benefit



GT record linkage - oracle-based

- Calcola l'unione con il cluster di E_1

UnionFind
$[\{E_1, N_{100}, N_{20}, N_8, N_3, N_{250}\}, \{E_2\}, \dots, \{E_{29.787}\}]$

GT record linkage - oracle-based

Samples
E_1
E_2
...
E_{150}

Normals
N_1
N_2
...
$N_{29.787}$

- Rimuovere da *Normals* tutti i nodi che ora sono nel cluster di E_1
- Ripetere gli stessi passaggi, ma con il sottografo indotto da:

$$\{E_1, N_{100}, N_{20}, N_8, N_3, N_{250}\}$$

- Benefit = grado pesato di N_1 nel sottografo indotto

GT record linkage - oracle-based

Samples	Normals
E_1	N_2
E_2	N_1
...	...
E_{150}	$N_{29.787}$

- Entità target: E_1
- Seleziona i top 5 nodi da *Normals*
 - ogni nodo è una domanda per l'oracolo
 - es:

N_2	✓
N_1	✗
$N_{1.000}$	✓
N_{70}	✗
$N_{5.000}$	✓

ordinata in base alla funzione di benefit



GT record linkage - oracle-based

- Calcola l'unione con il cluster di E_1

UnionFind

$[\{E_1, N_{100}, N_{20}, N_8, N_3, N_{250}, N_2, N_{1.000}, N_{5.000}\}, \{E_2\}, \dots, \{E_{29.787}\}]$

GT record linkage - oracle-based

Samples
E_1
E_2
...
E_{150}

Normals
N_1
N_3
...
$N_{29.787}$

- Rimuovere da *Normals* tutti i nodi che ora sono nel cluster di E_1
- Tutti i nodi che hanno ricevuto una risposta negativa sono ancora in *Normals*
 - verranno ignorati durante l'espansione del cluster corrente
 - quindi durante la fase di ordinamento
 - ma sono a disposizione degli altri oracoli

GT record linkage - oracle-based

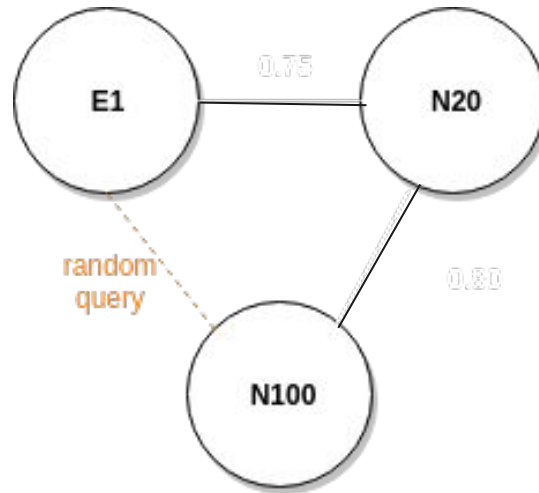
Samples
E_1
E_2
...
E_{150}

Normals
N_1
N_3
...
$N_{29.787}$

- L'iterazione sul cluster E_1 termina quando il primo elemento in *Normals* ordinato è sotto una data soglia oppure quando l'oracolo risponde negativamente ad una lunga serie di domande proposte
- Si passerà dunque all'espansione di E_2

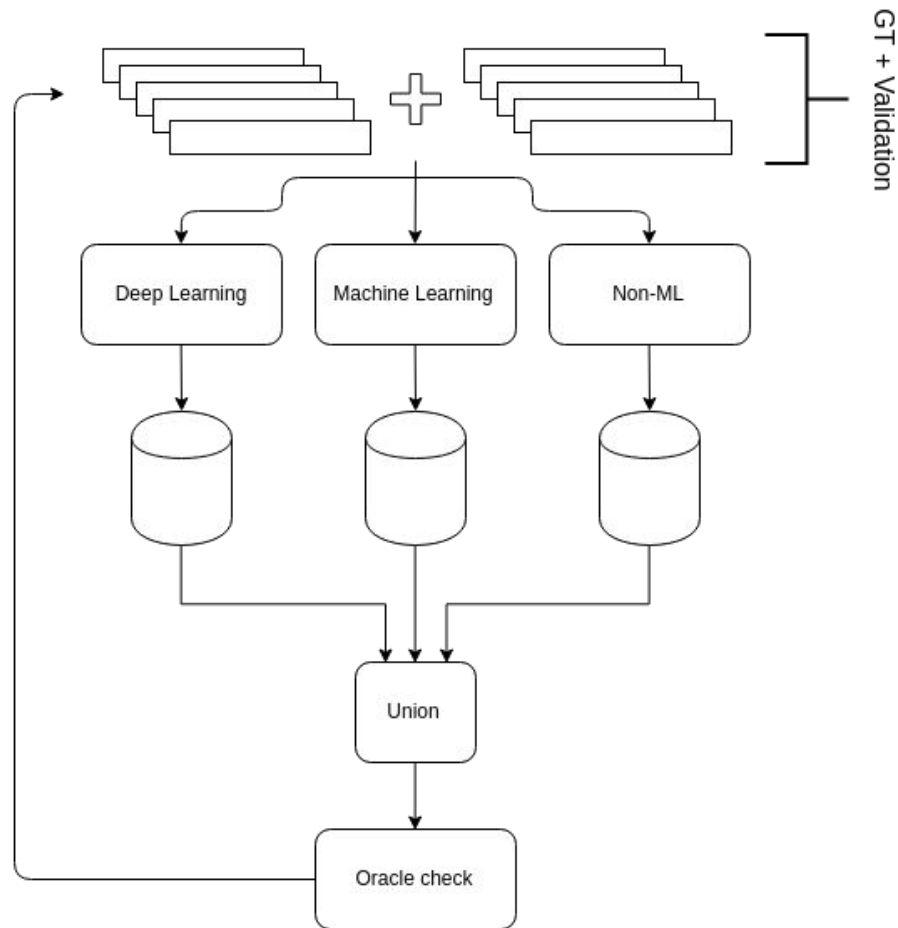
GT record linkage - oracle-based

- Alla fine del processo verranno proposte agli oracoli delle “random queries”
 - per verificare l'integrità dei risultati



GT record linkage - un nuovo approccio - black boxes

Pipeline basata sull'unione e la
verifica dei risultati di sistemi di RL
preesistenti





GT record linkage - black boxes

- Deep Learning
 - Deep matcher
 - [“Deep Learning for Entity Matching: A Design Space Exploration” \[AnHai et al. - SIGMOD ‘18\]](#)
 - <https://github.com/anhaidgroup/deepmatcher>
- Machine Learning
 - Magellan
 - [“Magellan: Toward Building Entity Matching Management Systems” \[AnHai et al. - VLDB ‘16\]](#)
 - <https://sites.google.com/site/anhaidgroup/projects/magellan>
- Non-ML
 - JedAI
 - [“The return of JedAI: End-to-End Entity Resolution for Structured and Semi-Structured Data” \[Papadakis et al. - VLDB ‘18\]](#)
 - <https://github.com/scify/JedAIToolkit>

Knowledge Graph Data Model



Knowledge Graph Data Model

- Formato JSON
 - ogni file ha un id e una classe
- Diverse classi per rappresentare i **nodi** nel grafo
 - class “json_file”
 - class “source”
 - class “entity”
 - class “target_attribute”
 - class “source_attribute”
 - class “provenance”
- Un “salto” tra JSON diversi tramite id rappresenta un **fatto**
- Più dettagli disponibili su <http://di2kg.inf.uniroma3.it/> sotto la sezione “Ground Truth Data and Challenge Instructions”

```
{  
  
  "resource_class": "entity",  
  
  "claims": [  
    {  
      "target_attribute_id": "TARGETATTRIBUTE#1",  
      "target_attribute_name": "battery_type",  
      "provenances": [  
        "PROVENANCE#1",  
        "PROVENANCE#2"  
      ]  
    }  
  ],  
  
  "instances": [  
    "JSON#1",  
    "JSON#2"  
  ]  
}
```

Conclusioni





Risultati

<http://di2kg.inf.uniroma3.it/>



1ST INTERNATIONAL WORKSHOP ON CHALLENGES AND EXPERIENCES FROM **DATA INTEGRATION** TO **KNOWLEDGE GRAPHS**

 August 5, 2019

 Anchorage, Alaska

★ Held in conjunction with **KDD 2019**



Risultati

- 8 team partecipanti alla challenge
 - tasks disponibili:
 - schema alignment
 - record linkage
 - knowledge graph augmentation
- 10 papers sottomessi inerenti al benchmark



Sviluppi futuri

- Integrare/costruire nuovi dataset (e relative GT)
- Realizzare GT per nuovi tasks (es. Data Extraction)
- Migliorare il sistema di valutazione delle soluzioni proposte
- Raffinare la metodologia per la costruzione della GT



Project Assignment - Challenge

- Testare il toolkit JedAI sul dataset di fotocamere
 - Trovare la configurazione migliore di JedAI per battere la nostra GT attuale sul Record Linkage

I dettagli del progetto devono essere discussi e concordati con il Prof. Paolo Merialdo!

Grazie per l'attenzione!

