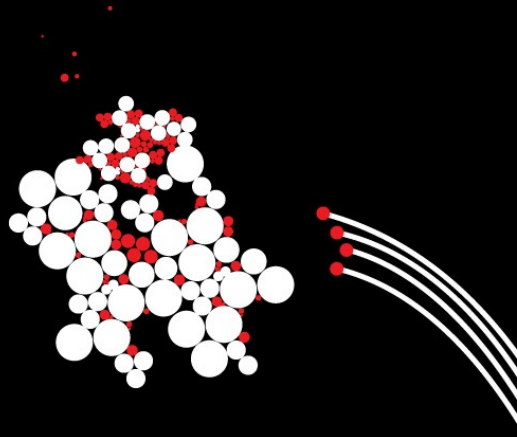
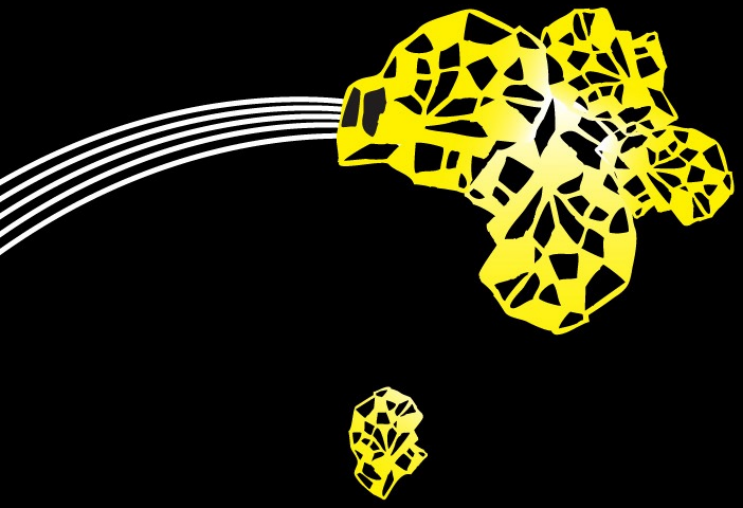
UNIVERSITEIT TWENTE.



PROBABILISTIC DATABASES AND DATA QUALITY

MAURICE VAN KEULEN



UNIVERSITEIT TWENTE.

THE IMPACT OF DATA QUALITY PROBLEMS


BIGDATA
Data Analytics
Data Science

POTENTIAL

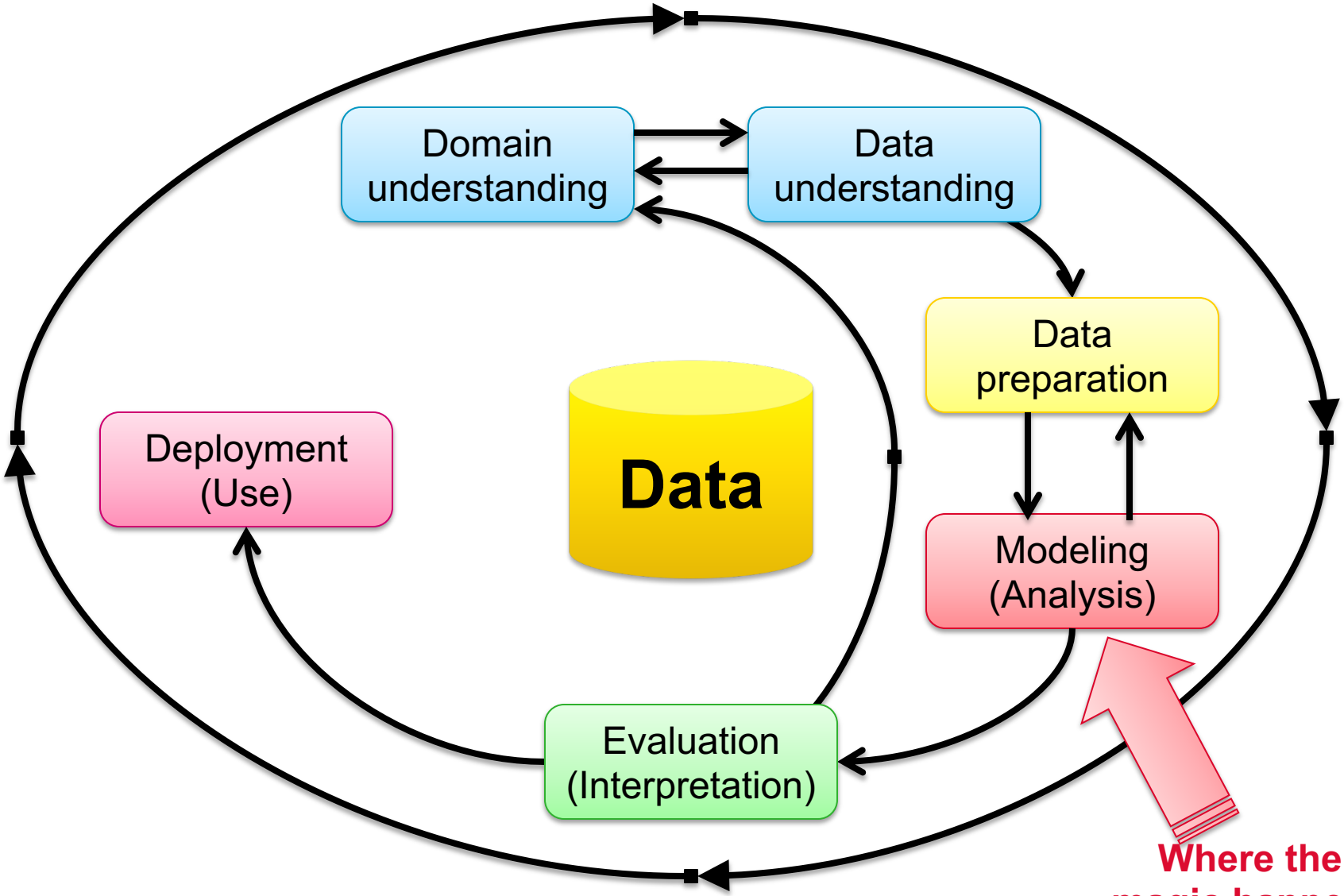
- Improved diagnostics
- Data-driven research
- Personalized medicine
- Quality assurance
- Improved effectiveness & productivity
- Improved logistics
- Continuous health monitoring
- Remote condition-based maintenance
- Etc. etc. etc.

BUT

Is it all gold that shines there?

Let me tell you a little story about a pregnancy research

CRISP-DM


Domain understanding
Data understanding
Data preparation
Modeling (Analysis)
Evaluation (Interpretation)
Deployment (Use)
Data
Where the magic happens

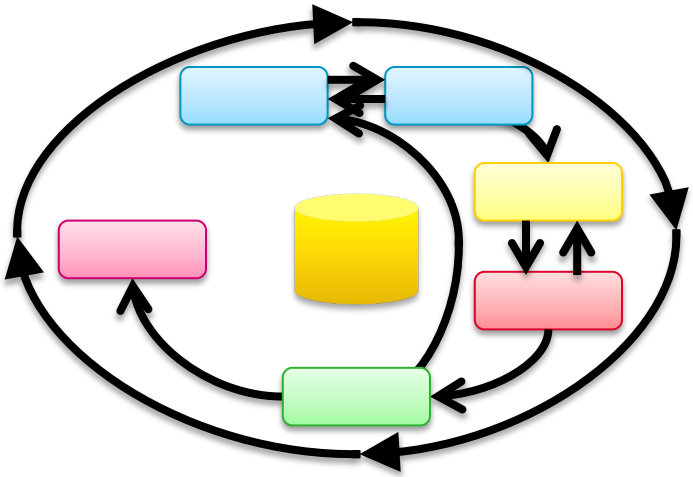
Research on

Pregnancy processes

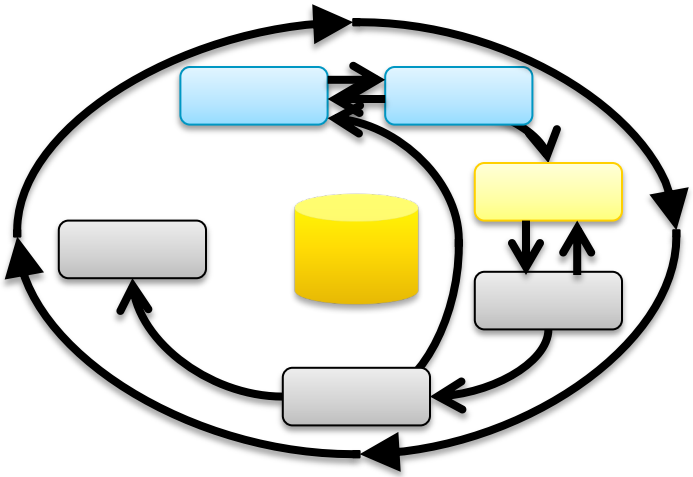
based on

Electronic Patient Dossiers (EPDs)

of some population of women



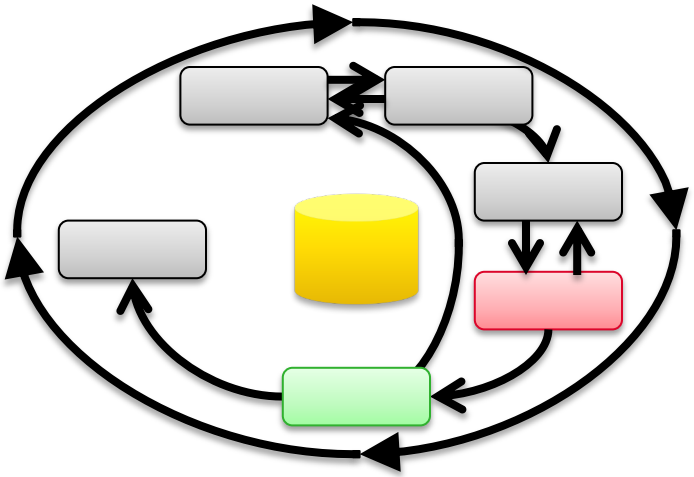
The start



1. Data scientist is a parent him/herself
2. Gather records from patient's EPDs for pregnancy period from multiple sources
3. Fairly straightforward to identify consults, tests, scans, conditions
4. Extract and store them

Days

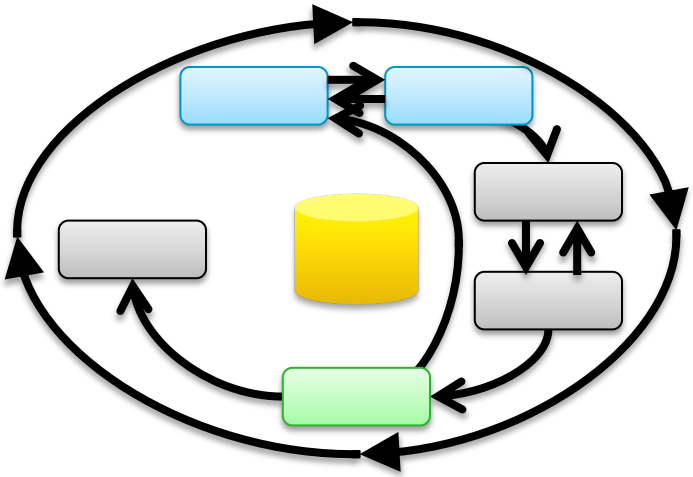
First analysis and evaluation



1. Analysis with process mining tool
2. Interaction with visualization
3. Interpret results
4. (S)he already sees some interesting patterns

Days

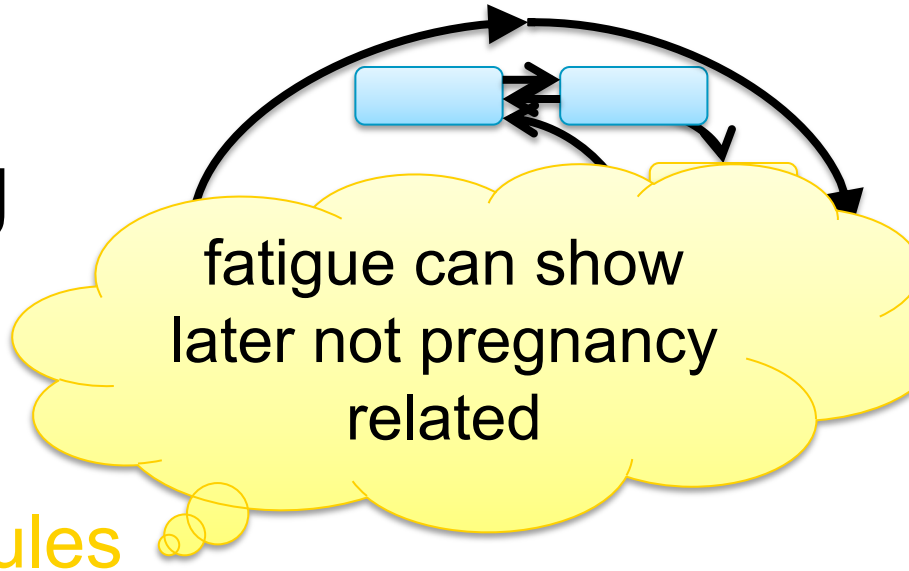
But then (s)he notices … and realizes …



- … a broken leg … dozens of specialists …
- Assumption wrong: “all records that belong to a pregnant woman are related to pregnancy”
- Too many records selected during preparation
- No objective means to ascertain this: No field ‘related to pregnancy’

Minutes

and a painstaking process starts



- Specify complex filter rules
- Inspect samples of (not) selected records
- Repeat!

Never perfect!

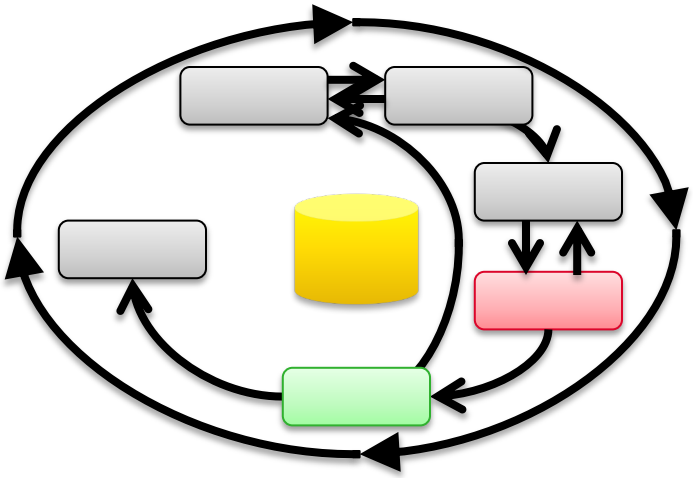
What is good enough?

Quick and dirty or thorough?

How does it affect results?

Weeks

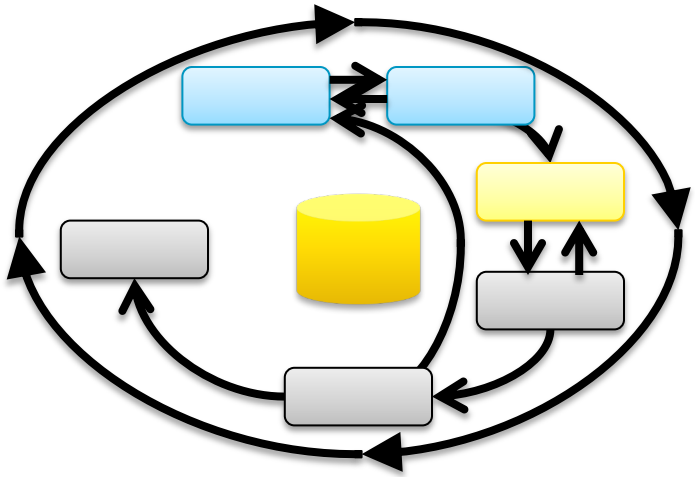
Fast forward a bit …



Re-perform analysis and evaluation

1. Interaction with mining tool and visualization
2. Something strange in the times of consults: many appear short and in the evenings??? consult after blood test it prescribed???

Realization
More cleaning



1. Clinician contacted for explanation:
   Notes during consult put in EPD in evenings!
2. Modification of EPD record (what is **recorded**)
   ≠ actual moment of activity (**semantics**)
   Sequence and duration noise
3. More data cleaning ensues

Weeks

I CAN GO ON

- Annotations of diagnosis in training data imperfect
- Missing data / excluded data → bias
- Same name different things; same thing different names
- Imperfect matching between different sources
- Scans / other data associated with wrong patient
- Symptoms a.o. extracted from natural language
- Etc. etc.

What can we learn from this?

- **Complex**

Many unexpected surprises in domain/data understanding

- **Time-consuming**

Most time spent on data preparation

Responsible Analytics

Quick and dirty or thorough?

What is good enough?

Never perfect!

How does it affect results?

- A data scientist should know ***and tell you*** about the deficiencies in the data and the results

SOME AUTHORITATIVE RESEARCH RESULTS

- Dirty data costs US businesses billions of dollars annually
- Data cleaning accounts for 30%-80% of the development time in a data warehouse project
- Key findings in a Gartner report (2011)
  - Poor data quality is a primary reason for 40% of all business initiatives failing to achieve their targeted benefits
  - Data quality effects overall labor productivity by as much as a 20%
  - As more business processes become automated, data quality becomes the rate limiting factor for overall process quality

- In bioinformatics, it is believed that “fiddling with the data” may often consume more than half of the time of a 4-year PhD project

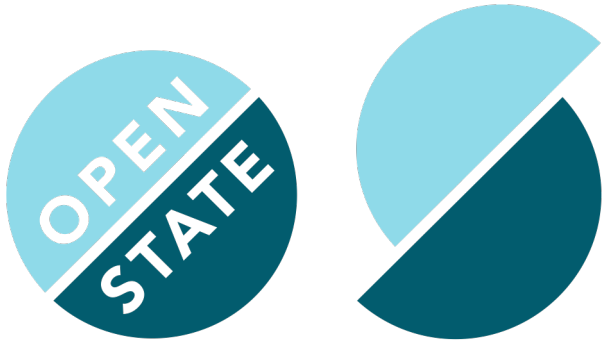
ANOTHER EXAMPLE: QUALITY OF PUBLIC DATA

HTTPS://OPENSTATE.GITHUB.IO/ALMANAK-KIESRAAD-MATCHER/VISUALIZATION/

Democracy relies a.o. on transparency and public information

- What is the quality of public information?
- The Open State Foundation attempts to improve digital transparency ...
  … and they have identified data quality problems as an important obstacle

➢ Let's have a look!



COMBINING DATA FROM DIFFERENT SOURCES

IS ASKING FOR DATA QULAITY PROBLEMS

www.tvguide.com	www.imdb.com
title: **The Namesake**	*title*: **Namesake, The**
year: 2006	*year*: 2006
genres: Drama	*genres*: **Comedy**, Drama, **Romance**
actors:	*actors*:
Sudipta Bhawmik (Subroto Mesho)	Bhawmik, Sudipta (**Subrata** Mesho)
Glenne **Headley** (**Lydia**)	**Headly**, Glenne (**Lydia Ratliff**)
Tamal **Roy Choudhury** (Ashoke's Father)	**Sengupta**, Tamal (Ashoke's Father)
Irrfan Khan (Ashoke)	Khan, **Irfan (I)** (Ashoke **Ganguli**)
Amy Wright (**Pam)**	Wright, Amy **(I)** (**Pamela**)
Sibani Biswas (Mrs. **Mazumdar**)	Biswas, Sibani (Mrs. **Mazoomdar**)
Sebastian **Roche** (Pierre)	**Roché**, Sebastian (Pierre)
Other actors not in imdb	**Other actors not in tvguide**
Other info like time, date, channel	*Other info like keywords, locations, plots.*

NOT AS SIMPLE AS THOUGHT … STILL SEEMS DOABLE

… AND WHY ISN'T IT AUTOMATIC?

How?

- Some coupling rules based on string matching
- Few clever disambiguation rules

Some observations:

- Conventions, errors, ambiguities make this a hard task
- 90% of the cases is straightforward; can be done with little effort (as above)
- 10% of the cases are hard; take most of the development time
- "There is always one more bug" → analogous rule applies here
- With current technology, these need to be solved to quite an extent to do something useful

We will come back to this example later

MAIN CLASSES OF DATA VALUE IMPERFECTIONS

Class	Example: John's tallness
No imperfection	183 cm
Absence/missing values	NULL
Non-specificity	Between 180 and 190 cm
	183 or 184 or 185 cm
Vagueness	Not very tall
Uncertainty	Perhaps 183 cm
Inconsistency	183 and 184 and 185 cm
Error	170 cm

Source:
Magnani and Montesi, "A Survey on Uncertainty Management in Data Integration"

IMPACT OF DATA IMPERFECTIONS: MISSING DATA

- For missing data one typically uses NULL in a database
- But if you don't know this … then you are possibly looking at **erroneous** figures!!!

CustID	Sales	Name
1234	6000	Foo
2345	***NULL***	Bar
3456	12000	Baz

```
SELECT SUM(Sales)
FROM CustSales
```

18000

OTHER SOURCES OF DATA QUALITY PROBLEMS

- Record-level imperfections, such as
  - Semantic duplicates
  - Matching (when integrating sources)
  - Selection
- Classification
- Ambiguity of language (text sources)
- Semantics
  (values and table/attribute names are ‘language’ too!)
- Trust issues

TEXT AS DATA SOURCE (INFORMATION EXTRACTION)

AMBIGUITY OF LANGUAGE PROVIDES EVEN MORE DATA QUALITY PROBLEMS

- *What exactly is the name of the hotel here?*
  "Essex House Hotel and Suites from $154 USD"

- Where is the hotel located? *>60 Paris-es in the world*
  "This Hilton hotel in Paris looks soooo nice;))"

- Informal use of language
  "Cancun is a MUST! Check this... Hotel Ocean Spa Cancun 4d 3N w/2 adults from $199 usd"

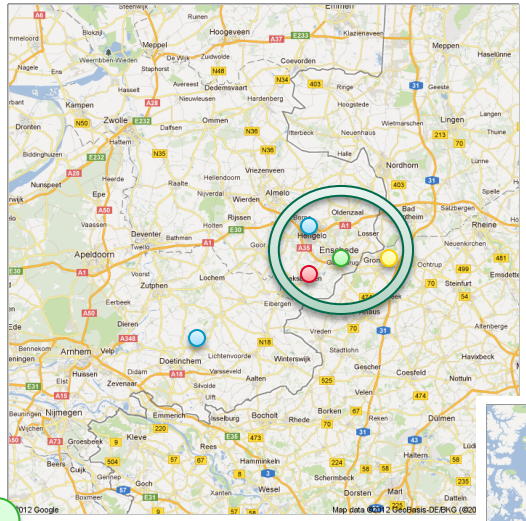
- etc.

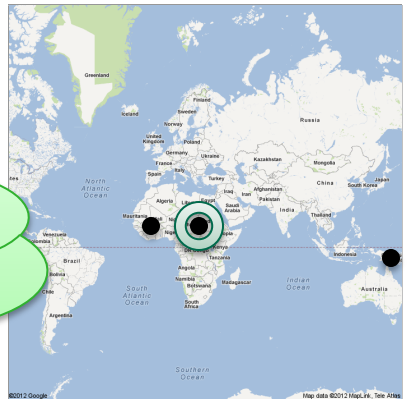
REASONING WITH 'DATA INTELLIGENCE'

WHERE IS THIS HOLIDAY COTTAGE LOCATED? TO WHAT DO THESE PLACE NAMES REFER?

The cottage is in **Usselo**. **You** can shop in the nearby towns of **Enschede**, **Hengelo** and **Gronau**. Cool boat rides on the river **Dinkel**.

- Usselo: 1 (NL)
- Enschede: 1 (NL)
- Hengelo: 2 (NL, NL)
- Gronau: many (GER)
- You: 4 (Burkina Faso, Papua New Guinea, Chad, Chad)

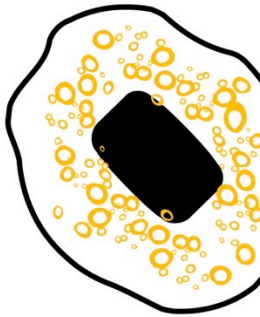




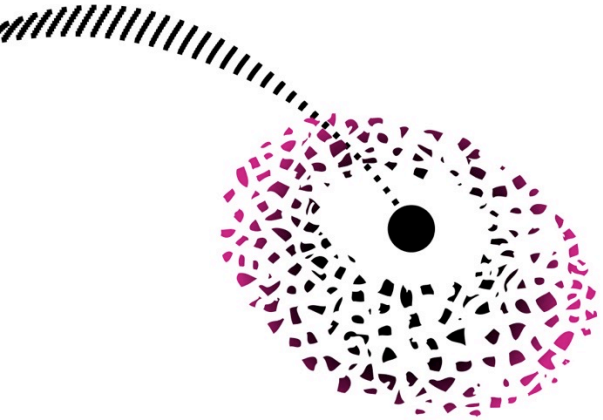
NL/NL/NL/GER/C had ➔ NL

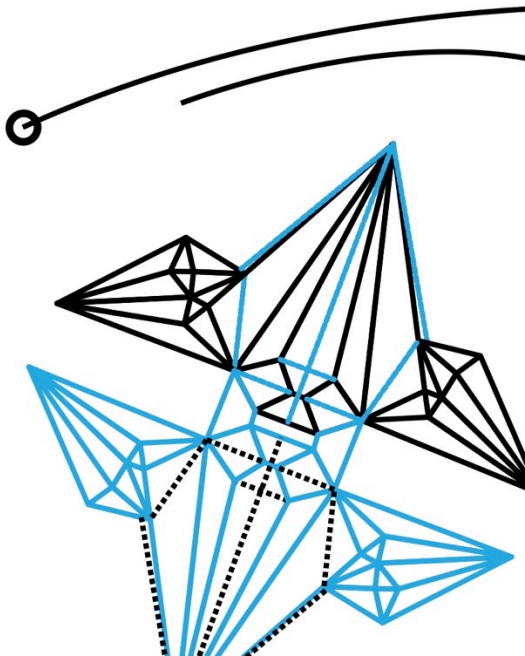
Other cases with "You" time and again in other countries ➔ probably not a place name

UNIVERSITEIT TWENTE.



PROBABILISTIC DATABASES TO THE RESCUE





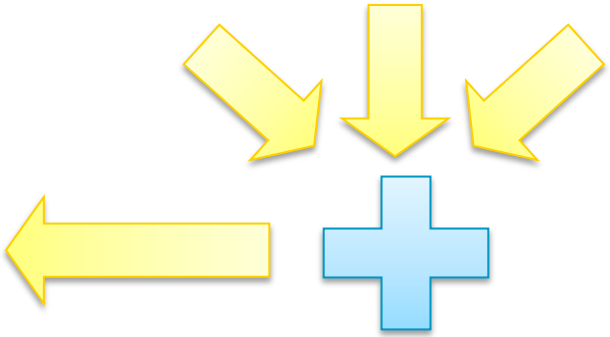
COMBINING DATA …

Car brand	Sales
B.M.W.	25
Mercedes	32
Renault	10

Car brand	Sales
BMW	72
Mercedes-Benz	39
Renault	20

Car brand	Sales
Bayerische Motoren Werke	8
Mercedes	35
Renault	15

Car brand	Sales
B.M.W.	25
Bayerische Motoren Werke	8
BMW	72
Mercedes	67
Mercedes-Benz	39
Renault	45



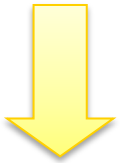
Keulen, M. (2012) *Managing Uncertainty: The Road Towards Better Data Interoperability.* IT - Information Technology, 54 (3). pp. 138-146. ISSN 1611-2776

… AND THE PROBLEM OF SEMANTIC DUPLICATES

Car brand	Sales
B.M.W.	25
Bayerische Motoren Werke	8
BMW	72
Mercedes	67
Mercedes-Benz	39
Renault	45

Preferred customers …

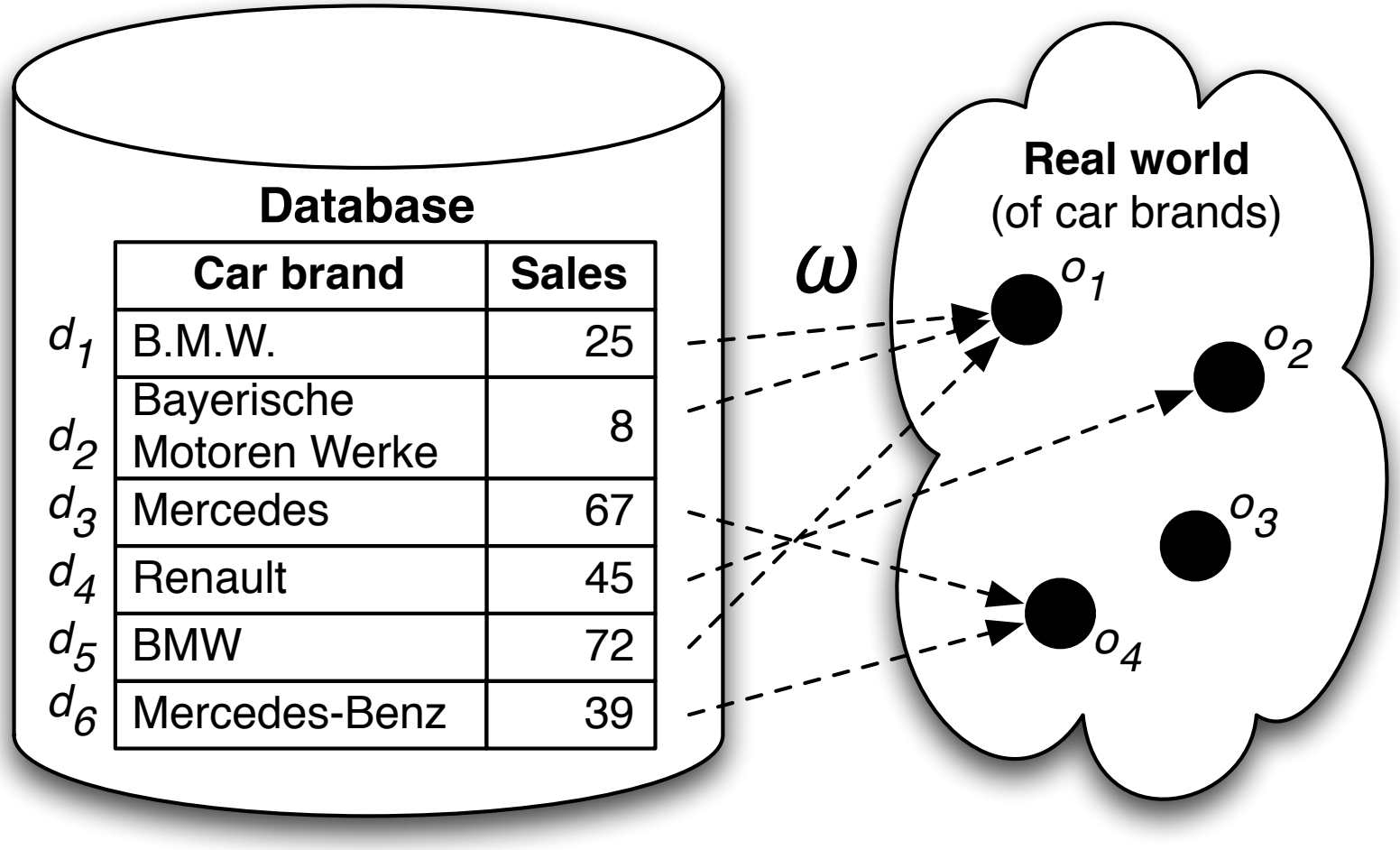
```
SELECT SUM(Sales)
FROM CarSales
WHERE Sales>100
```



0

‘No preferred customers’

SEMANTIC DUPLICATES



MOST DATA QUALITY PROBLEMS CAN BE MODELED AS UNCERTAINTY IN DATA

Run some duplicate detection tool

	Car brand	Sales	
1	B.M.W.	25	
2	Bayerische Motoren Werke	8	
3	BMW	72	
4	Mercedes	67	X=0
5	Mercedes-Benz	39	X=0
6	Renault	45	
	Mercedes	106	X=1 Y=0
	Mercedes-Benz	106	X=1 Y=1

X=0	4 and 5 different	0.2
X=1	4 and 5 the same	0.8
Y=0	“Mercedes” correct name	0.5
Y=1	“Mercedes-Benz” correct name	0.5

B.M.W. / BMW / Bayerische Motoren Werke analogously

WHAT I HAVE NOW IS A PROBABILISTIC DATABASE

- Looks like ordinary database
- Several “possible” answers or approximate answers to queries
- Important: Scalability (big data!)

Sales of “preferred customers”

- SELECT SUM(sales)
  FROM carsales
  WHERE sales≥ 100

SUM(sales)	P
0	14%
105	6%
106	56%
211	24%

QUERYING AND RELIABILITY ASSESSMENT

Sales of “preferred customers”

- SELECT SUM(sales)
  FROM carsales
  WHERE sales≥ 100
- Answer: 106

SUM(sales)	P
0	14%
105	6%
106	56%
211	24%

Risk of substantially wrong answer

- Risk = Probability * Impact
- Analyst only bothered with problems that matter

Second most likely answer at 24% with impact factor 2 in sales (211 vs 106)

OTHER DATA QUALITY PROBLEMS

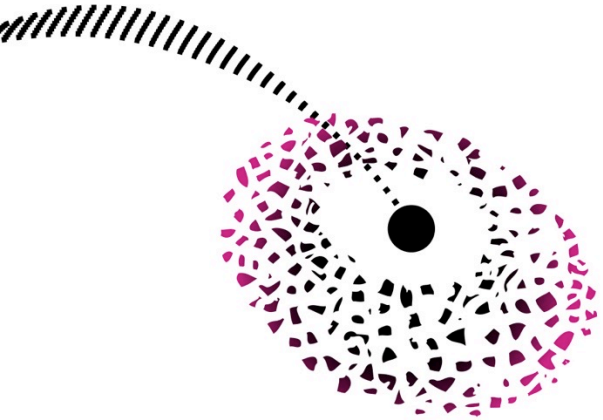
- Analysis of scans, tests, measurements
- ➢ Detection errors, impossible to do perfectly
- Information extraction from natural language
- ➢ Detection errors, ambiguity, limited context
- Combining sources
- ➢ Missing data, interoperability problems

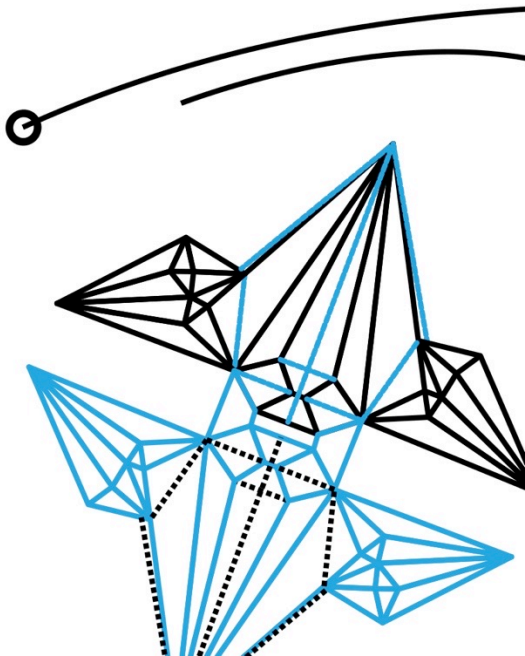
Examples from story

- Belongs to pregnancy problem
- ➢ Becomes “estimate probability of belonging to pregnancy”
- Uncertainty about order of activities
- ➢ Becomes an “estimate probability of one way or the other”

UNIVERSITEIT TWENTE.

WHAT ARE YOU GOING TO DO IN THIS TOPIC?





THE PRACTICAL ASSIGNMENTS

Goal: learn how to represent data quality problems as uncertainty in the data stored in a probabilistic database
Tool: *JudgeD* (a probabilistic DataLog)

1. Crash course in logic programming in DataLog
2. Querying a probabilistic database
3. Creating probabilistic data
4. Representing several kinds of data quality problems as uncertainty in the data
5. Probabilistic data integration

THE ALBUM PROJECT

Do it yourself with real-world data

- Match and merge the semantic duplicates of music albums …
  … and give it your own twist

or

- Come up with your own data and idea
  - There should be trust and/or quality issues
  - Real-world data science setting
    (e.g., integration, information extraction, etc.)

THE SDSI PROJECT
WEB HARVESTING FOR SMART APPLICATIONS

- Data set:
  Data harvested by yourself (e.g., using a tool like import.io)
- Challenge:
  Demonstrate potential of data harvested from the web for developing smart applications
  - Integrate data from at least 2 sources
    (at least one is harvested from the web)
  - Design and implement smart web/mobile app

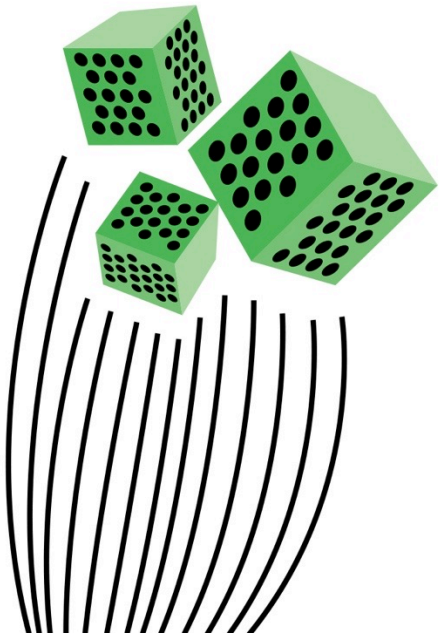
Example: data on side effects of medicines for a disease, harvest patient web forum on this disease, determine which of these medicines are used and which side effects are reported

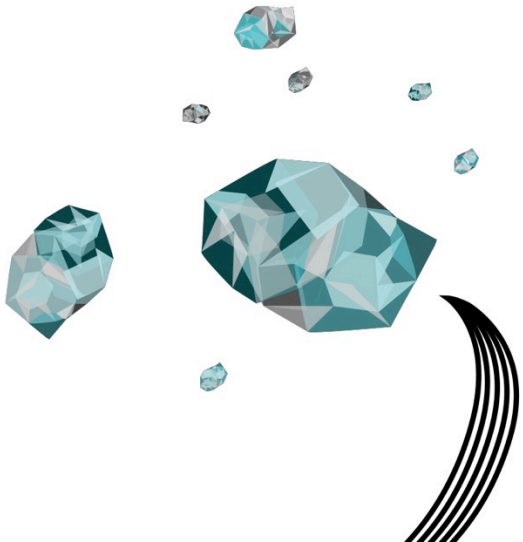
More projects to come e.g., product matching, product feature extraction

UNIVERSITEIT TWENTE.

The tool we are going to use:

JudgeD: Probabilistic Datalog





DATALOG BY EXAMPLE

```
movie(1,"The Namesake",2006).

genre(1,drama).
genre(1,comedy).

actor(1,"Sudipta Bhawmik","Subroto Mesho").
actor(1,"Glenne Headley","Lydia").
actor(1,"Tamal Roy Choudhury","Ashoke's Father").
```

- These are “facts”
  - Syntax is: *predicate* (*terms*).
  - A term is a number or a string; if the string contains anything else than [a-z_] then put double quotes around it

DATALOG QUERY

Queries:

- Syntax: predicate (terms or vars)?
- A variable begins with a capital; '_' is don't care

```
actor(_,_,"Lydia")?
```

- Is there an actor who played role "Lydia"?

```
actor(ID,Name,"Lydia")?
```

- The ID and name of the actor who played Lydia

```
actor(ID,Name,Role)?
```

- The IDs, names and roles of all actors

```
movie(ID,Name,2006)?
```

- The ID and name of all movies in 2006

DATALOG RULES

Rules:

- Syntax: predicate (terms or vars) :- other predicates.
- A rule specifies a logical truth.

```
actorofmovie(Actor,Movie) :-
  movie(ID,Movie,_),
  actor(ID,Actor,Role).
```

- Actor is the actor of movie Movie, **iff** there is a movie with name Movie and id ID **AND** there is an actor with name Actor and role Role with **the same** id ID.

```
actorofmovie(A,M)?
```

DATALOG RULES: EXPRESSING OR

- Two rules with the same head constitute an **OR**

```
actororroleofmovie(Name,actor,Movie) :-
   movie(ID,Movie,_), actor(ID,Name,_).
actororroleofmovie(Name,role,Movie) :-
   movie(ID,Movie,_), actor(ID,_,Name).
```

- The name is either an actor name or a role name

```
actororroleofmovie("Lydia",What,Movie)?
actororroleofmovie(N,W,M)?
```

GOOD PRACTICE: DENORMALIZATION

Good practice: separate attributes into different predicates

```
genre_id(1).
genre_id(2).

genre_movie(1,1).
genre_movie(2,1).

genre_genre(1,drama).
genre_genre(2,comedy).

genre(GID,MID,G) :- genre_id(GID),
genre_genre(GID,G), genre_movie(GID,MID).
```

PROBABILISTIC DATALOG: FIRST DATA QUALITY PROBLEMS

Data quality problems:

1. the name is either "The Namesake" or "Namesake, The"
2. the genre is drama, but possibly also comedy and romance

```
movie_name(1,"The Namesake") [n=1].
movie_name(1,"Namesake, The") [n=2].

@p(n=1) = 0.5.
@p(n=2) = 0.5.
```

Mutually exclusive

```
genre_genre(1,drama).
genre_genre(1,comedy) [g=1].
genre_genre(1,romance) [g=1].

@p(g=1) = 0.5.
@p(g=2) = 0.5.
```

Dependent

PROBABILISTIC DATALOG

This JudgeD program actually specifies 4 possible programs, also called **possible worlds**!

1. [n=1 and g=1] correct movie name is “The Namesake” and in reality is has all three genres
2. [n=1 and g=2] correct movie name is “The Namesake” and in reality is only has genre “drama”
3. [n=2 and g=1] correct movie name is “Namesake, The” and in reality is has all three genres
4. [n=2 and g=2] correct movie name is “Namesake, The” and in reality is only has genre “drama”

These programs run (in a clever way of course) simultaneously!

Answers for all programs are gathered together and presented with probabilities

MISSING VALUES

Example: there is a new movie “Wall-E” with an actor who plays “Robot Eva”, but it is unknown who that is.

- Assumption: it is a known actor
  (closed world assumption)

```
actor_movie_role(1,2,"Robot Eva") [a=1].
actor_movie_role(2,2,"Robot Eva") [a=2].
actor_movie_role(3,2,"Robot Eva") [a=3].

@uniform p(a).
```

TWO VERSIONS OF THE REASONING ENGINE

There are two versions

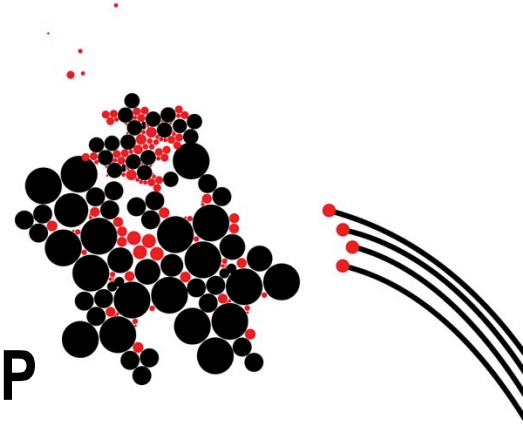
- **Exact**
  - Reasons by *sentence* manipulation
  - gives answers with constructed sentences but no probabilities

```
genreofmovie("Namesake, The", comedy)
   [(n=2 or (g=1 and n=2))].
```

- **Monte Carlo**
  - Reasons by Monte Carlo simulation (sampling)
  - gives answers with probabilities but no sentences

```
genreofmovie("The Namesake", comedy).
   % p = 0.488
```

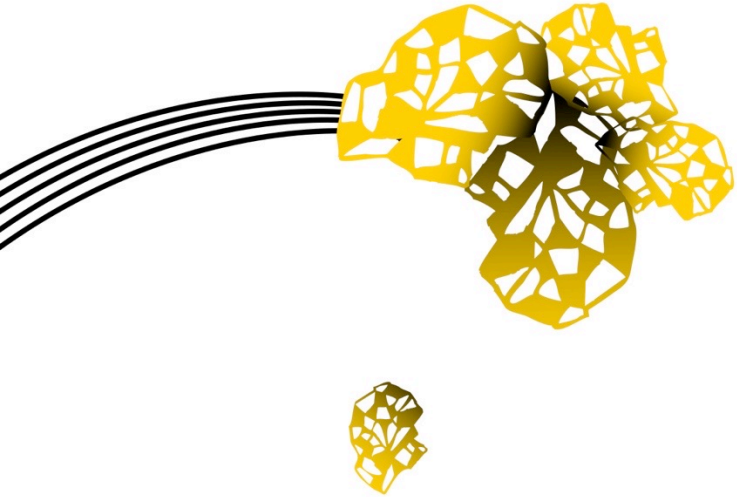
UNIVERSITEIT TWENTE.

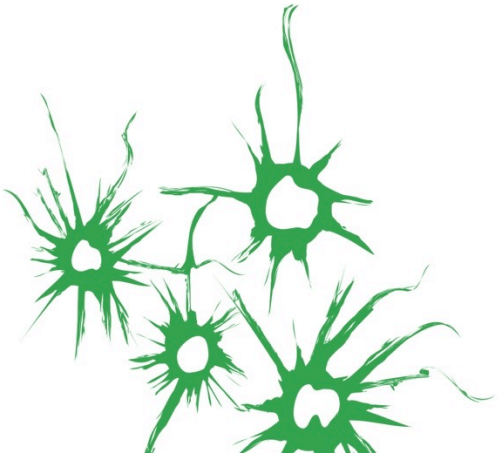


DO PROBABILISTIC DATABASES REALLY HELP TO REDUCE DATA INTEGRATION EFFORT?

AN EXPERIMENT

SOURCE: VAN KEULEN, M. AND DE KEIJZER, A. (2009) QUALITATIVE EFFECTS OF KNOWLEDGE RULES AND USER FEEDBACK IN PROBABILISTIC DATA INTEGRATION.





Back to our little story

- Imagine a developer with a mission
  "Daily recommendations for movies to watch on TV"
- Idea:
  - Get TV Guide from internet
  - Select movies
  - Enrich/integrate with data from IMDB
  - Choose recommendations
  - Write blog

THE PROBLEM IS IN THE DATA VALUES!

NO ID'S, MISSING VALUES, DIFFERENT CONVENTIONS, ERRORS, AMBIGUITIES

www.tvguide.com	www.imdb.com
title: **The Namesake**	*title*: **Namesake, The**
year: 2006	*year*: 2006
genres: Drama	*genres*: **Comedy**, Drama, **Romance**
actors:	*actors*:
Sudipta Bhawmik (Subroto Mesho)	Bhawmik, Sudipta (**Subrata** Mesho)
Glenne **Headley** (**Lydia**)	**Headly**, Glenne (**Lydia Ratliff**)
Tamal **Roy Choudhury** (Ashoke's Father)	**Sengupta**, Tamal (Ashoke's Father)
Irrfan Khan (Ashoke)	Khan, **Irfan (I)** (Ashoke **Ganguli**)
Amy Wright (**Pam)**	Wright, Amy **(I)** (**Pamela**)
Sibani Biswas (Mrs. **Mazumdar**)	Biswas, Sibani (Mrs. **Mazoomdar**)
Sebastian **Roche** (Pierre)	**Roché**, Sebastian (Pierre)
Other actors not in imdb	**Other actors not in tvguide**
Other info like time, date, channel	*Other info like keywords, locations, plots.*

NOT AS SIMPLE AS THOUGHT … STILL SEEMS DOABLE

… BUT WHY ISN'T IT AUTOMATIC?

How?

- Some coupling rules based on string matching
- Few clever disambiguation rules

Some observations:

- Conventions, errors, ambiguities make this a hard task
- 90% of the cases is straightforward; can be done with little effort (as above)
- 10% of the cases are hard; take most of the development time
- "There is always one more bug" → analogous rule applies here
- With current technology, these need to be solved to quite an extent to do something useful

LET'S SIMPLY NOT SOLVE THOSE 10%! (RIGHTAWAY)

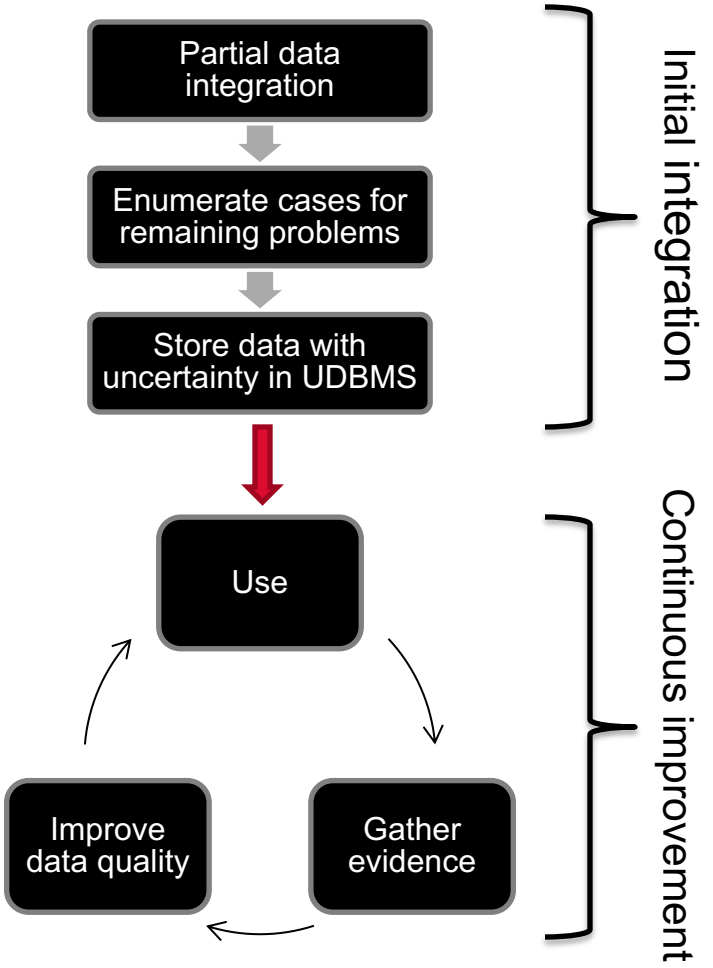
A TWO-PHASE PROBABILISTIC DATA INTEGRATION PROCESS

Let's go for an initial integration that can readily and meaningfully be used

"Good is good enough" for meaningful use in many applications

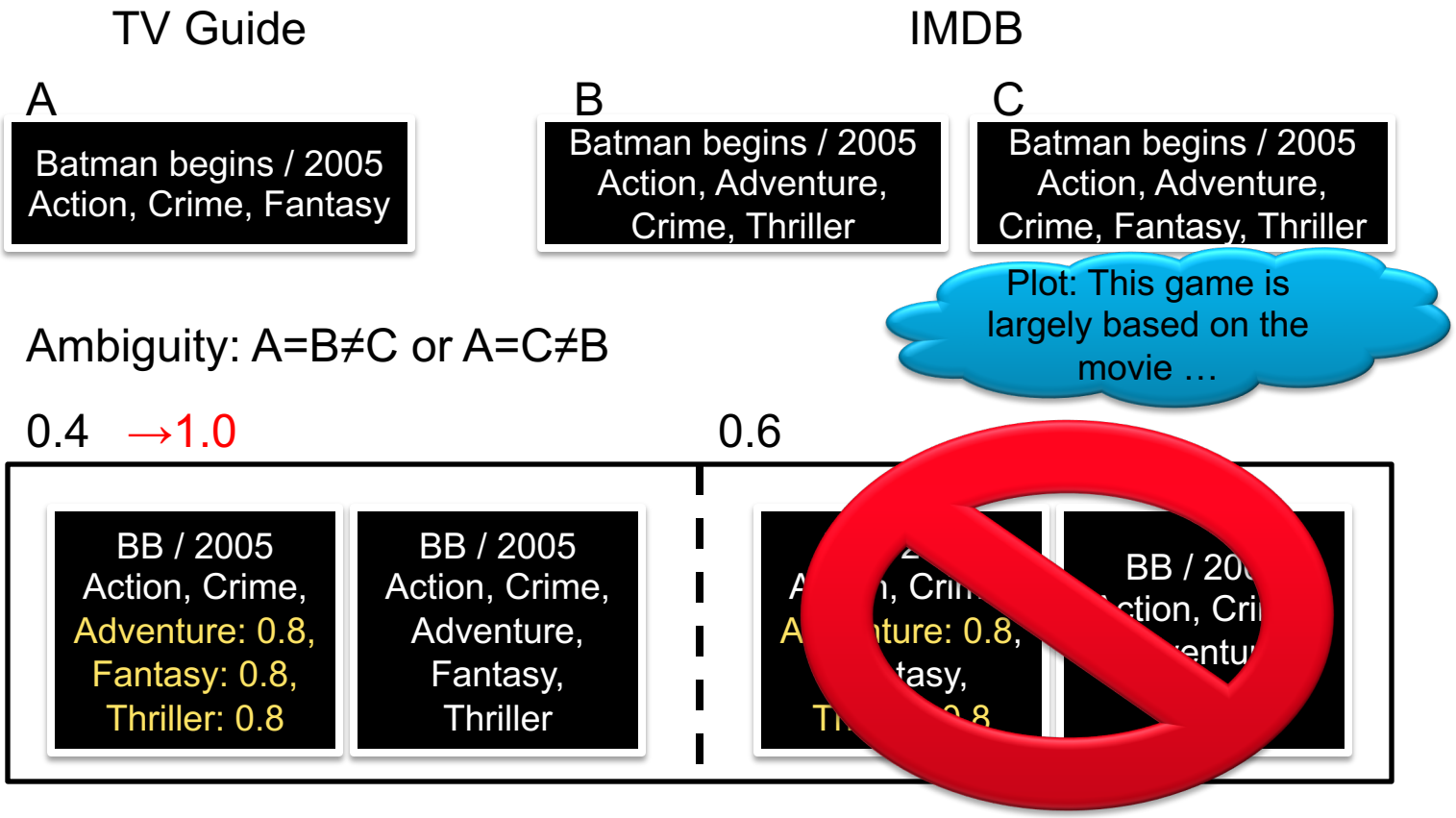
(can be achieved 10x earlier)

Let it improve during use



EXAMPLE AMBIGUOUS CASE

"BADMAN BEGINS / 2005"



Query: How many fantasy movies?

1 (0.68) or 2 (0.32)!

1 (0.2) or 2 (0.8)

DOES IT REALLY REDUCE DEVELOPMENT EFFORT?

EXPERIMENTS

1. How easy is it to reach 90% / 10%?
2. Is user feedback effective enough to quickly improve data quality?

Data: XML data of top-100 movies of TV-guide enriched with IMDB data set (250.000 movies). Together 18 “attributes”.

System: XML database MonetDB/XQuery extended with support for probabilistic data integration and storing/querying of uncertain data.

43 XPaths; uncertainty and quality measurements

MEASURING UNCERTAINTY AND QUALITY

Uncertainty ≠ Quality

- Quality = degree of correspondence with “the truth”
- Uncertainty = degree of system “doubting” its own data (many good metrics already exist, e.g., entropy)

What does “good” in “good enough” mean?

- Data quality

But how do we measure that?

- Expected precision and recall
  - The quality of a correct answer is higher if the system dares to claim that it is correct with a higher probability
  - Analogously, incorrect answers with a high probability are worse than incorrect answers with a low probability

EXPERIMENTAL RESULTS

1. HOW EASY IS IT TO REACH 90% / 10%?

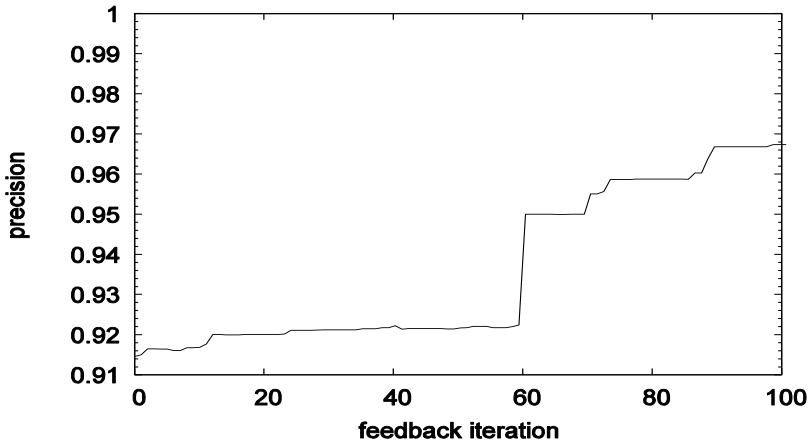
- DTD + 1 rule per entity + generic text similarity fallback
  Movie: Title-similarity above/below threshold + Year
  Actor: Role corresponds and is unique for movie $\Rightarrow$ same actor
- Not very sensitive to threshold settings if kept on safe side

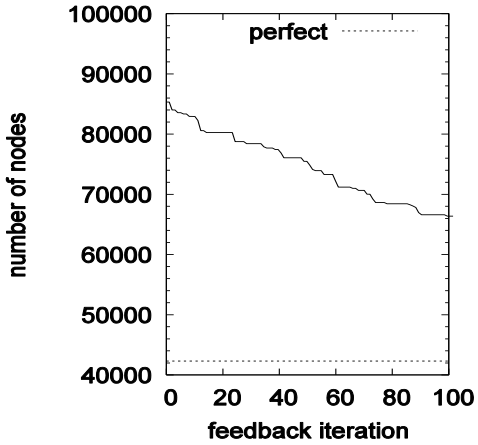
Entity	Total	Exact match	No match	Ambiguous	Wrong	% easy
All (many non-matches)						
Movie	98	56	27	14	1	84.7%
Actor	3887	1484	2101	302	?	92.2%
2006 and earlier (few non-matches)						
Movie	57	48	2	6	1	87.7%
Actor	3133	1447	1400	286	?	90.9%

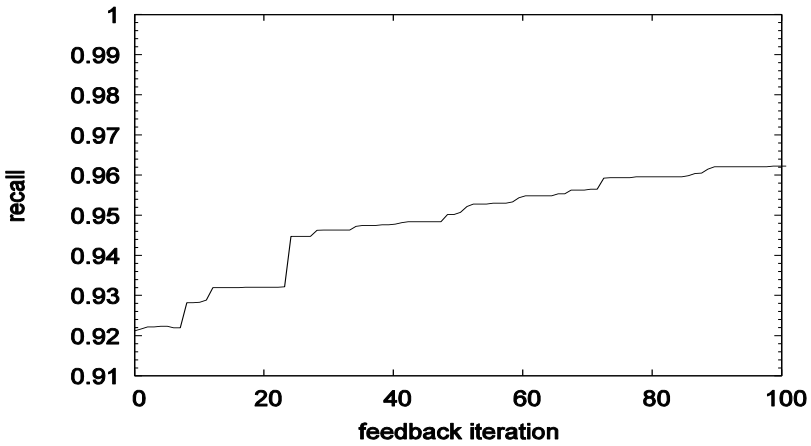
EFFECTIVENESS OF USER FEEDBACK

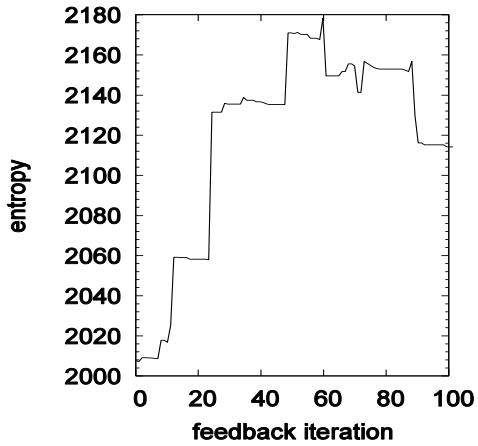
IS USER FEEDBACK EFFECTIVE IN QUICKLY IMPROVING DATA QUALITY?

100x (pick random query+answer and give feedback on its correctness)









CONCLUSIONS

"GOOD IS GOOD ENOUGH" DATA INTEGRATION

- Probabilistic data integration
  - Phase 1: quick-and-dirty integration sufficient for initial integration of good enough quality that can meaningfully used
    (few domain-specific rules and rough thresholds suffice)
  - Phase 2: user feedback effective for sustained quality improvement
- Developer trades quality for time/effort needed for first use

IF YOU WANT TO MORE

Suppose after this topic …

You want to know more

there is a master course on

Probabilistic Programming

in the third quarter

If a man will begin with certainties,
he shall end in doubts;
but if he will be content to begin with
doubts, he shall end in certainties.

(Francis Bacon, 1605)

Doubt is one of the
names of intelligence

(Jorge Luis Borges, 1979)