Big Data Analytics in the Age of the GDPR

Sabrina Kirrane, WU

11th July 2019

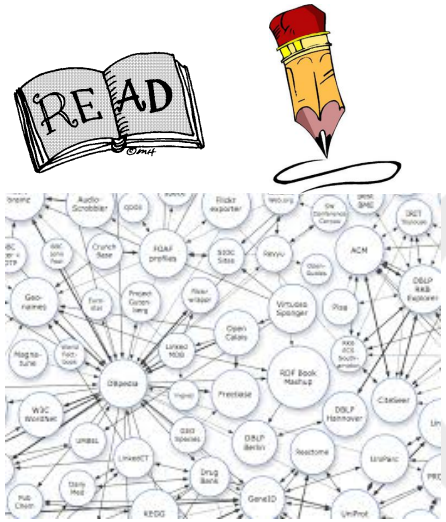
Data Science Institute @ NUI Galway

SPECIAL

European Commission

Horizon 2020
European Union funding
for Research & Innovation

Access Control for Linked Data

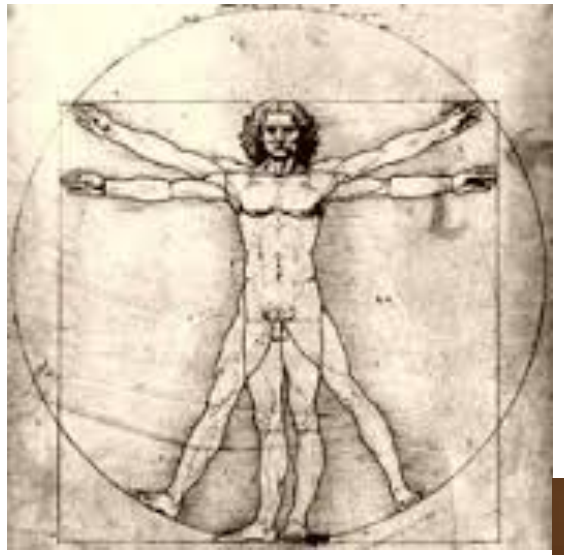


RDF + SPARQL = 1.1

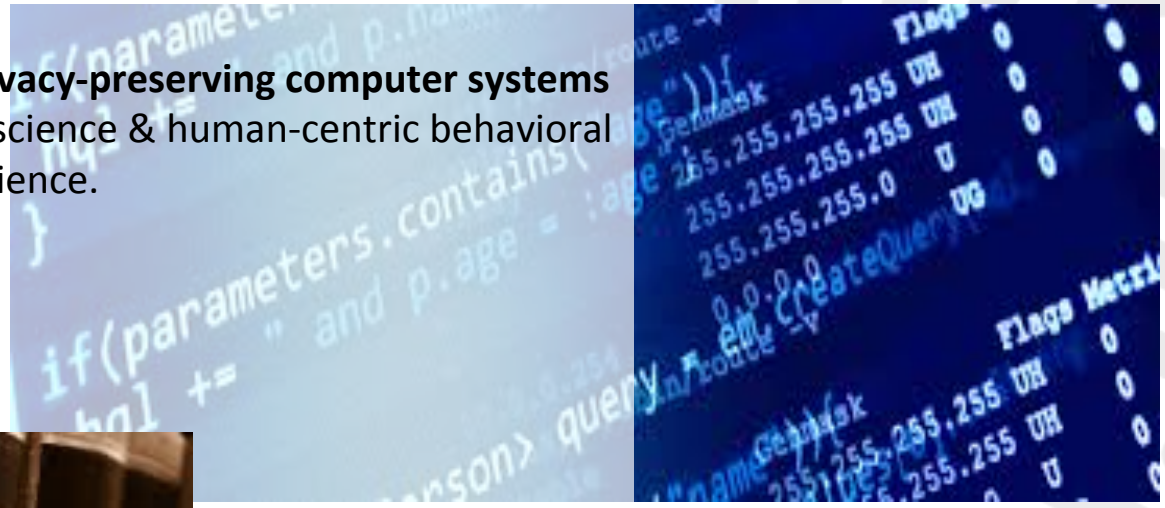
PRIVACY & SUSTAINABLE COMPUTING LAB

Privacy & Sustainable Computing

A multidisciplinary perspective…



Developing sustainable and privacy-preserving computer systems by bringing together computer science & human-centric behavioral science.



Humanities

Online Privacy

Licensing

Legislation

Open Standards

Legal

Computer Science

Distributed Systems

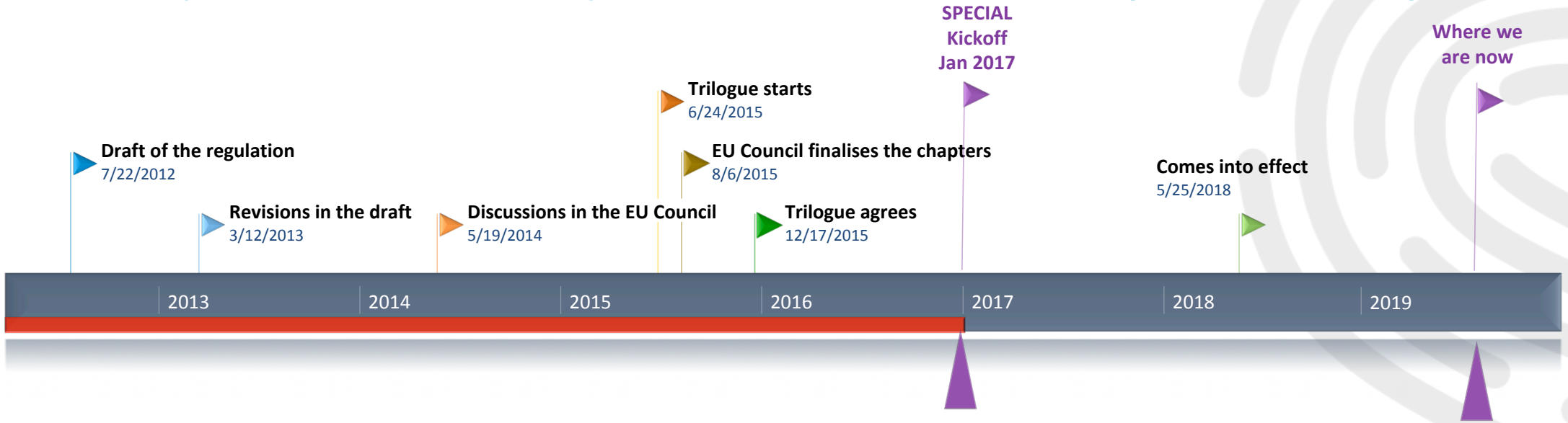
Decentralisation

Artificial Intelligence

Big Data

Data Science

Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance (SPECIAL)

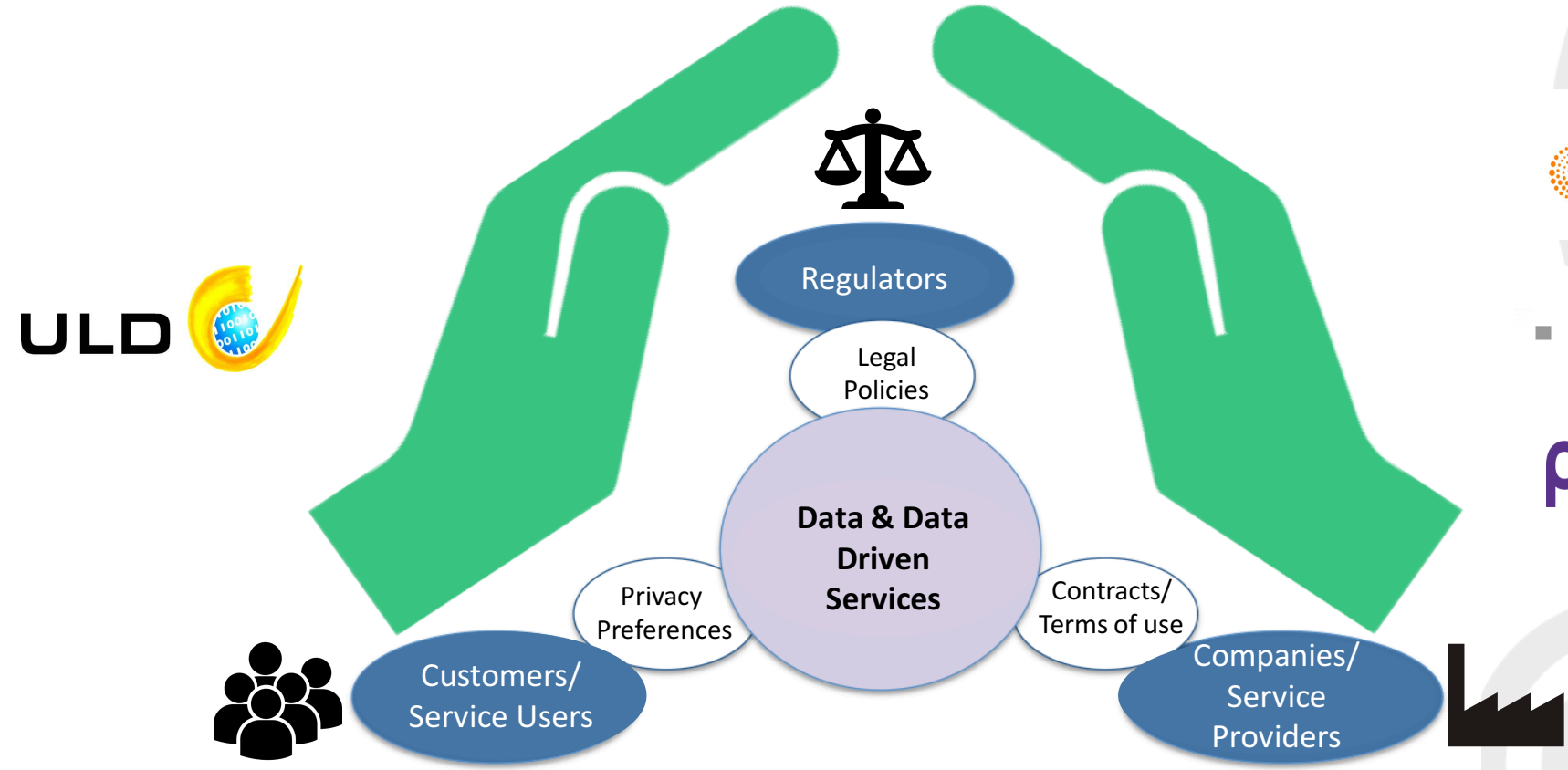


Data subjects who would like to declare, monitor and optionally revoke their (often not explicit) preferences on data sharing

Regulators who can leverage technical means to check compliance with the GDPR

Companies whose business models rely on personal data and for which the GDPR is both a challenge and an opportunity

Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance (SPECIAL)



GDPR Impact on Innovation?

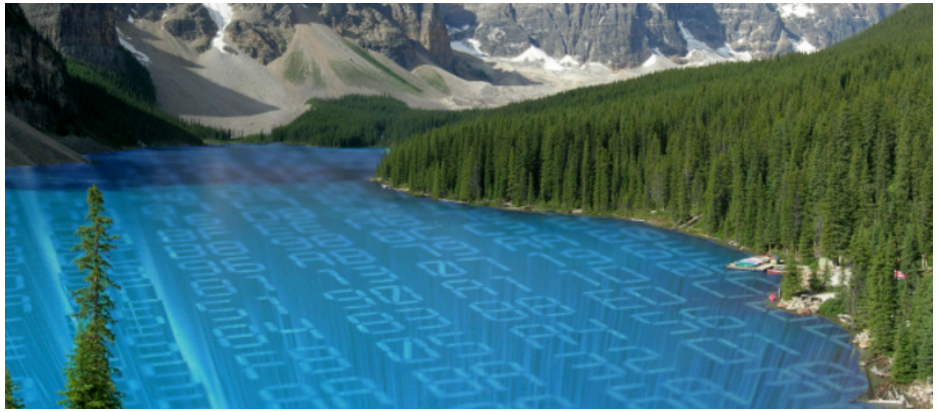
Data Vault

http://www.miamidatavault.com/

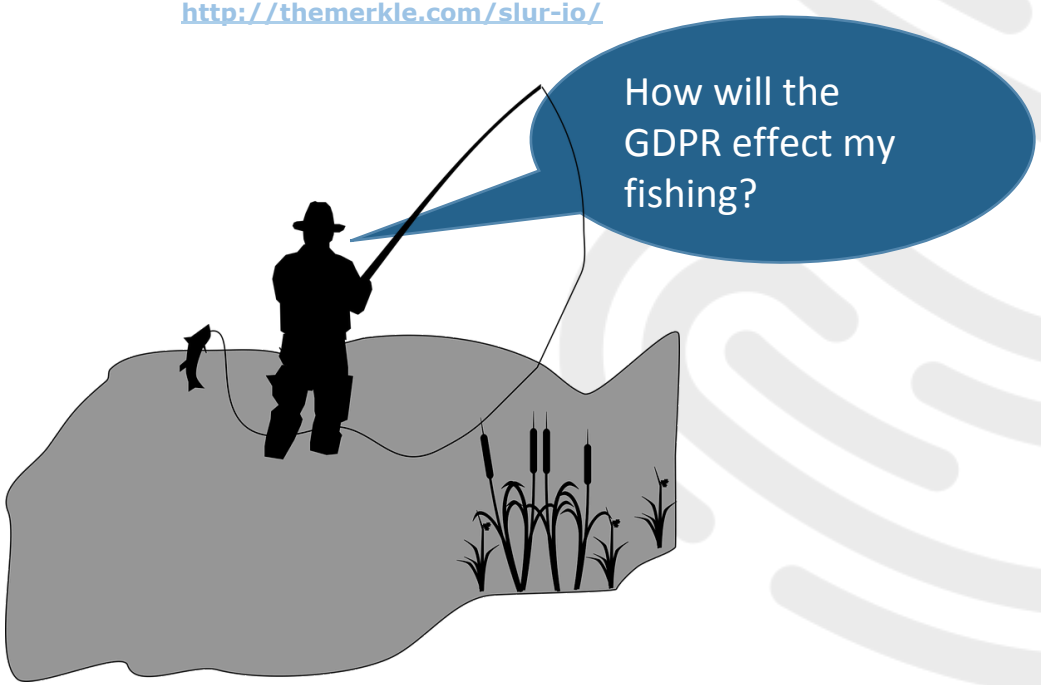
Data Market

http://themerkle.com/slur-io/

Data Lake



https://solutionsreview.com/data-integration/the-emergence-of-data-lake-pros-and-cons/



Innovation via Anonymisation & Aggregriation!

Innovation via Anonymisation & Aggregriation!

4.5.2016 EN Official Journal of the European Union L 119/1

I

(Legislative acts)

REGULATIONS

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
of 27 April 2016
on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Article 16 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee (1),

Having regard to the opinion of the Committee of the Regions (2),

The GDPR does not apply to anonymous data where the data subject is no longer identifiable.

(26) The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

Innovation via Anonymisation & Aggregriation!

K-Anonymity

- A record cannot be distinguished from at least $\mathcal{K}$-1 others
- Approach
  - **Suppression** certain values of the attributes are replaced by an asterisk
  - **Generalization** individual values of attributes are replaced by with a broader category

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	⩾40	Flu
4790*	⩾40	Heart Disease
4790*	⩾40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Samarati, Pierangela, and Latanya Sweeney. ***Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression***. Technical report, SRI International, 1998.

Innovation via Anonymisation & Aggregriation!
Is K-Anonymity enough?

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	⩾40	Flu
4790*	⩾40	Heart Disease
4790*	⩾40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

$\mathcal{K}$-anonymity has deficiencies when sensitive values in an equivalence class lack diversity or the attacker has background knowledge

Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3.

Innovation via Anonymisation & Aggregriation!
K-Anonymity & L-Diversity

- Each equivalence class has at least ℓ well-represented sensitive values

Similarity Attack

Bob	
Zip	***Age***
47678	27

Conclusion

- Bob's salary is between [20k,40k].
- Bob has some stomach-related disease.

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	⩾40	50K	Gastritis
4790*	⩾40	100K	Flu
4790*	⩾40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

ℓ-diversity does not consider the semantic meanings of the sensitive values

Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3.

Innovation via Anonymisation & Aggregriation!
K-Anonymity, L-Diversity & T-Closeness

- Distribution of sensitive attributes within each quasi identifier group should be "close" to their distribution in the entire original database

A completely generalised table

Age	Zipcode	.....	Gender	Disease
*	*	.....	*	Flu
*	*	.....	*	Heart Disease
*	*	.....	*	Cancer
.	.	.....	.	.
.	.	.....	.	.
.	.	.....	.	.
*	*	.....	*	Gastritis

Background Knowledge Attack

Bob	
Zip	***Age***
47678	27

A released table

Age	Zipcode	.....	Gender	Disease
2*	476**	.....	Male	Flu
2*	476**	.....	Male	Heart Disease
2*	476**	.....	Male	Cancer
.	.	.....	.	.
.	.	.....	.	.
.	.	.....	.	.
⩾50	4766*	.....	*	Gastritis

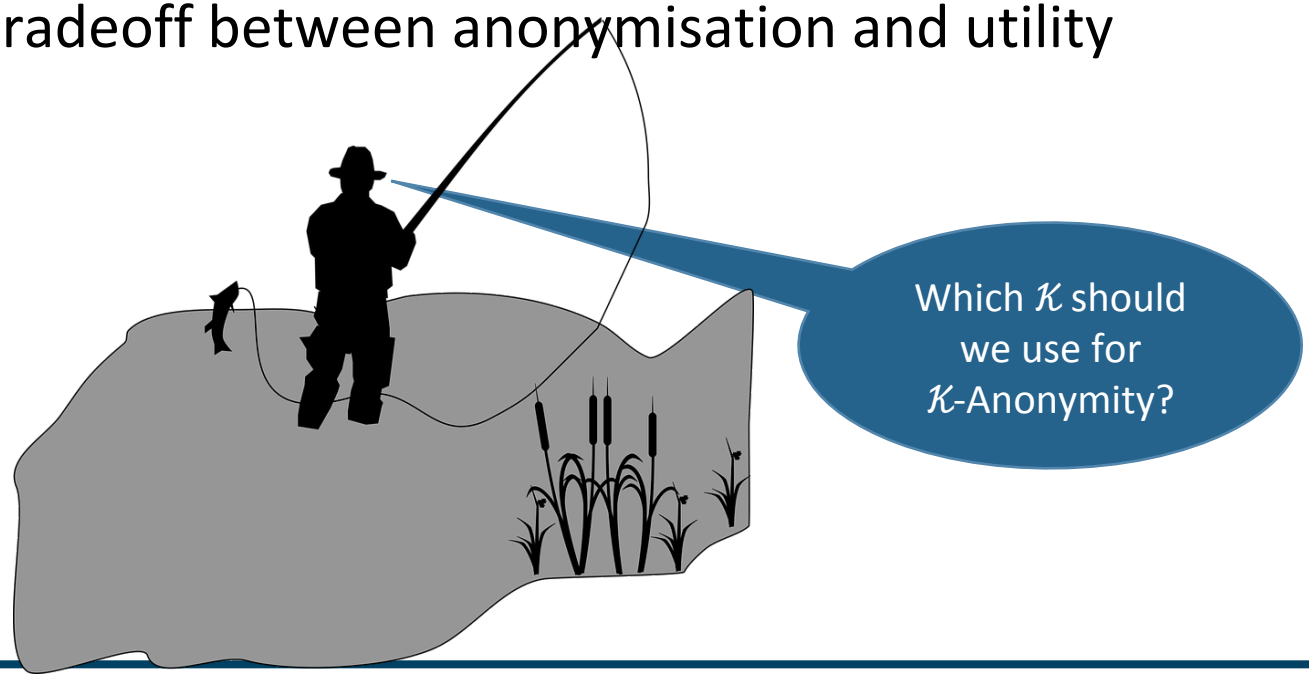
Conclusion

- Bob could have Flu, Heart Disease or Cancer!

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007.

Innovation via Anonymisation & Aggregriation!

- A layered approach to anonymisation may be needed
- Even then $\mathcal{K}$, $\mathcal{L}$ & $\mathcal{T}$ are highly dependent on the data
- Also, there is a tradeoff between anonymisation and utility



Considering that it is getting harder and harder to guarantee anonymity while preserving utility, what is the alternative?

Innovation via Consent!

SPECIAL Use Cases

proximus

Events at the Belgian Coast at your fingertips

Sign up for free for intelligent tourist event recommendations tailored to you.

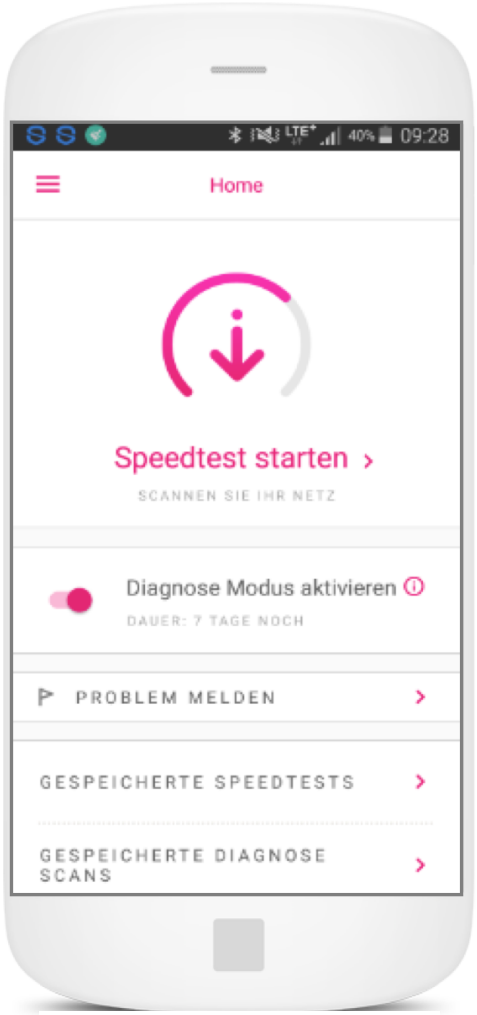
Login

freddy.demeersman@proximus.com

••••••••••••••••••••••••••••••••

LOGIN

proximus



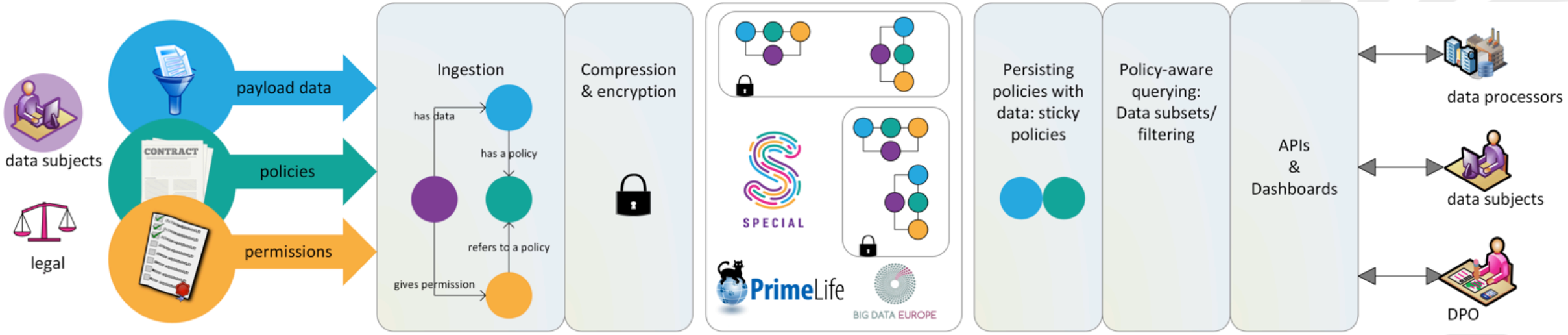
Deutsche Telekom

KNOW YOUR CUSTOMER.

KNOW.

the answer company™
THOMSON REUTERS®

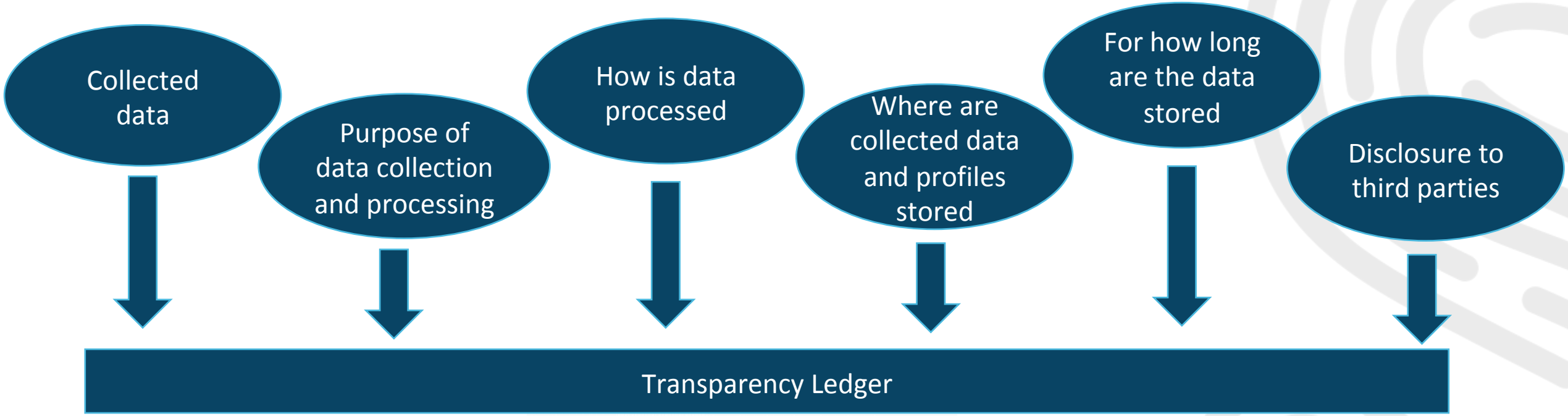
SPECIAL Technical Foundations
Big Data and Privacy Foundations

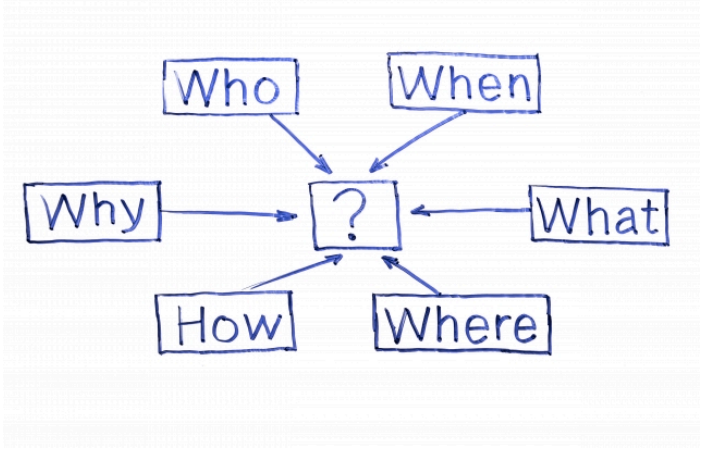


SPECIAL leverages past infrastructure and lessons learned

- **Big Data Europe** scalability and elasticity
- **PrimeLife** policy languages, access control policies, release policies and data handling policies
- The **Platform for Privacy Preferences Project** (P3P) and the **Open Digital Rights Language** (ODRL) vocabularies

SPECIAL Technical Foundations
Minimal Core Model





ORACLE Symantec vmware amazon Microsoft SAP salesforce.com Adobe

Usage policy language
Syntax and expressivity

- Usage policy language, which can be used to express both the data subjects' consent, data controllers usage requests, fragements of the GDPR, and business policies
- The foundation of the policy language was the Minimal Core Model (MCM)
- We propose a new policy language that extensively re-uses standards based privacy-related vocabularies
- We are able to leverage existing Web Ontology Language (OWL) based reasoners out of the box

Figure 1.1: SPECIAL's Usage Policy Language Grammar

UsagePolicy :='ObjectUnionOf' '(' **BasicUsagePolicy BasicUsagePolicy** { **BasicUsagePolicy** } ')'
| **BasicUsagePolicy**

BasicUsagePolicy :='ObjectIntersectionOf' '(' **Data Purpose Processing Recipients Storage** ')'

Data :='ObjectSomeValueFrom' '(' 'spl:hasData' **DataExpression** ')'

Purpose :='ObjectSomeValueFrom' '(' 'spl:hasPurpose' **PurposeExpression** ')'

Processing :='ObjectSomeValueFrom' '(' 'spl:hasProcessing' **ProcessingExpression** ')'

Recipients :='ObjectSomeValueFrom' '(' 'spl:hasRecipient' **RecipientExpression** ')'

Storage :='ObjectSomeValueFrom' '(' 'spl:hasStorage' **StorageExpression** ')'

DataExpression :='spl:AnyData' | **DataVocabExpression**

PurposeExpression :='spl:AnyPurpose' | **PurposeVocabExpression**

ProcessingExpression :='spl:AnyProcessing' | **ProcessingVocabExpression**

RecipientsExpression :='spl:AnyRecipient' | 'spl:Null' | **RecipientVocabExpression**

StorageExpression :='spl:AnyStorage' | 'spl:Null' |
'ObjectIntersectionOf' '(' **Location Duration** ')'

Location :='ObjectSomeValueFrom' '(' 'spl:hasLocation' **LocationExpression** ')'

Duration :='ObjectSomeValueFrom' '(' 'spl:hasDuration' **DurationExpression** ')'
| 'DataSomeValueFrom' '(' 'spl:durationInDays' **IntervalExpression** ')'

Usage policy language
Syntax and expressivity

SPECIAL Namespace Prefixes

PREFIX spl: <http://www.specialprivacy. eu/langs/usage−policy#>
PREFIX splog: <http://www.specialprivacy. eu/langs/splog#>
PREFIX svd: <http://www.specialprivacy. eu/vocabs/duration#>
PREFIX svl: <http://www.specialprivacy. eu/vocabs/locations#>.

Structure of a Usage Control Policy

ObjectIntersectionOf(
ObjectSomeValuesFrom(spl:hasData *SomeDataCategory*)
ObjectSomeValuesFrom(spl:hasProcessing *SomeProcessing*)
ObjectSomeValuesFrom(spl:hasPurpose *SomePurpose*)
ObjectSomeValuesFrom(spl:hasRecipient *SomeRecipient*)
ObjectSomeValuesFrom(spl:hasStorage *SomeStorage*))

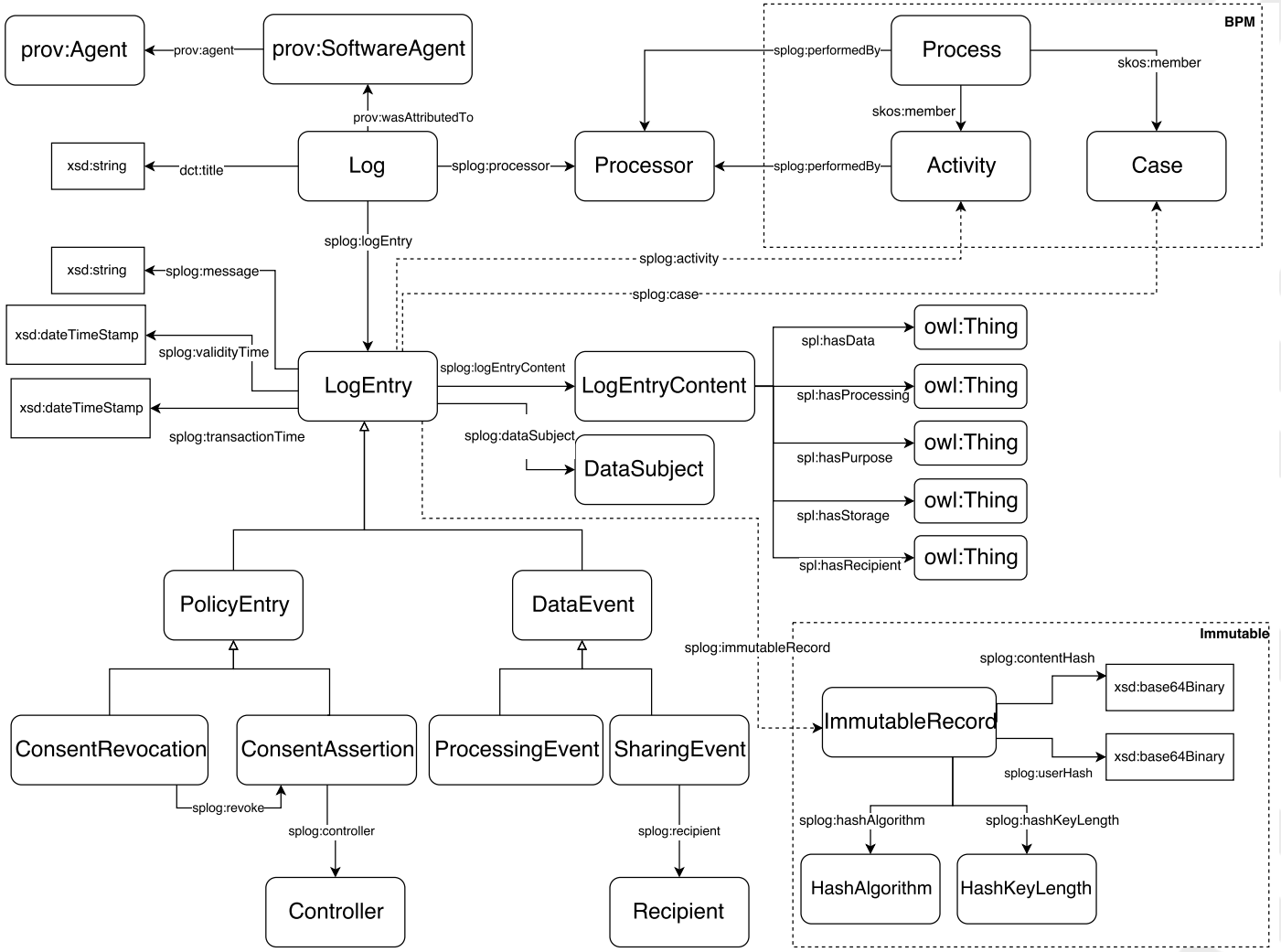
Usage policy language
Syntax and expressivity

Structure of a Usage Control Policy

ObjectIntersectionOf(
ObjectSomeValuesFrom(spl:hasData **svd:Location**)
ObjectSomeValuesFrom(spl:hasProcessing **befit:SensorGathering**)
ObjectSomeValuesFrom(spl:hasPurpose **befit:HealthTracking**)
ObjectSomeValuesFrom(spl:hasRecipient **befit:Internal**)
ObjectSomeValueFrom(spl:hasStorage
ObjectIntersectionOf(ObjectSomeValueFrom(spl:hasLocation **:EU**)
DataSomeValueFrom(spl:hasDuration
DatatypeRestriction(xsd:integer
xsd:minInclusive "**365**"^^xsd:integer
xsd:maxInclusive "**1825**"^^xsd:integer)
))

Provenance/event information Syntax and expressivity

- Development of a log vocabulary that reuses well-known vocabularies such as PROV for representing provenance metadata
- Demonstrate how provenance can be used to support transparency in data value chains



Provenance/event information
Syntax and expressivity

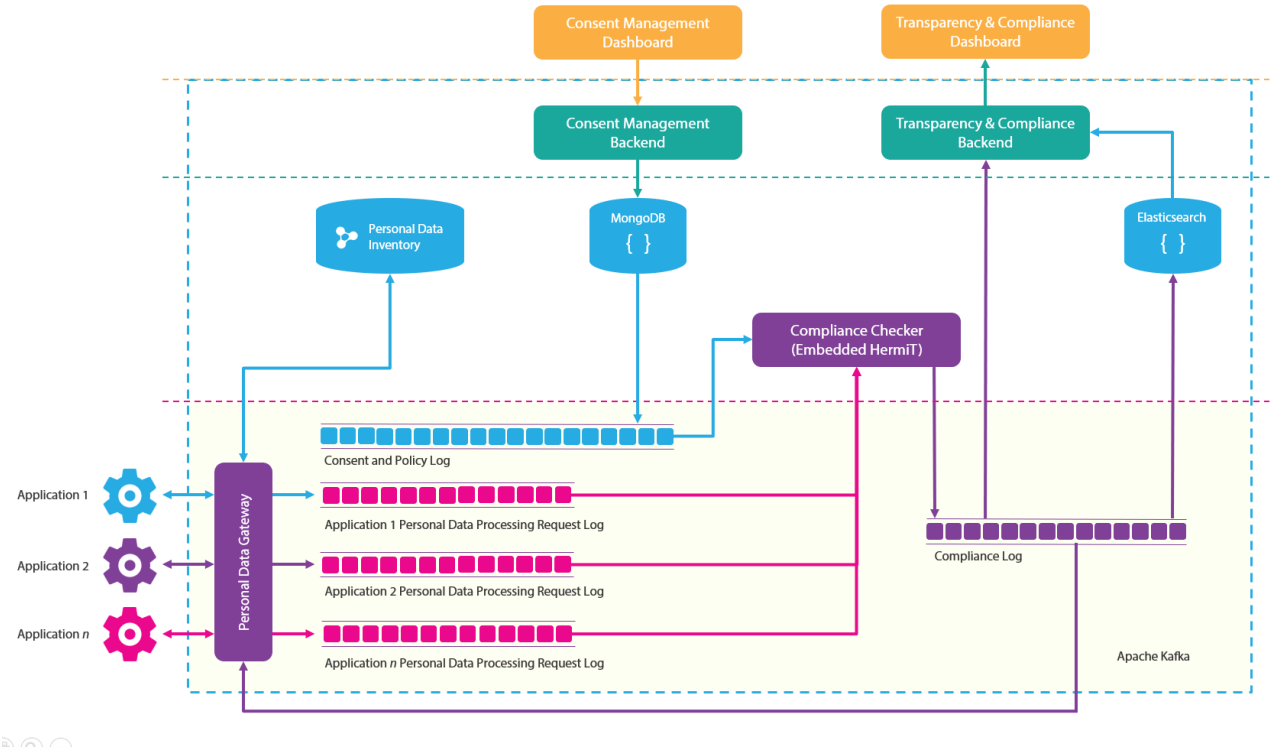
A new event for Sue's BeFit device

```
befit:entry3918 a splog:ProcessingEvent ;
splog:dataSubject befit:Sue ;
dct:description " Store location in our database in Europe "@en ;
splog:transactionTime "2018−01−10T13 : 2 0 : 5 0Z"^^ xsd : dateTimeStamp ;
splog:validityTime "2018−01−10T13 : 2 0 : 0 0Z"^^ xsd : dateTimeStamp ;
splog:eventContent befit:content3918 ;
splog:inmutableRecord befit:iRec3918 .
```

The content of a new event for Sue's BeFit device

```
befit:content3918 a splog:LogEntryContent ;
spl:hasData svd:Location ;
spl:hasProcessing befit:SensorGathering ;
spl:hasPurpose befit:HealthTracking ;
spl:hasStorage [ spl:haslocation :EU] ;
spl:hasRecipient [ a svr:Ours ] .
```

Transparency and compliance checking platforms



- Data processing and sharing event logs are stored in the **Kafka** distributed streaming platform, which in turn relies on Zookeeper for configuration, naming, synchronization, and providing group services.
- We assume that consent updates are infrequent and as such usage policies and the respective vocabularies are represented in a **Virtuoso triple store**.
- The compliance checker, which includes an embedded
- A **HermiT reasoner** uses the consent saved in Virtuoso together with the application logs provided by Kafka to check that data processing and sharing complies with the relevant usage control policies.
- As logs can be serialized using JSON-LD, it is possible to benefit from the faceting browsing capabilities of **Elasticsearch** and the out of the box visualization capabilities provided by **Kibana**.

Usage policy language SPECIAL resources

The SPECIAL Usage Policy Language
version 0.1

SPECIAL

Unofficial Draft 06 April 2018

Editor:
Javier D. Fernández (Vienna University of Economics and Business)

Authors:
Piero Bonatti (Università di Napoli Federico II)
Sabrina Kirrane (Vienna University of Economics and
Iliana Mineva Petrova (Università di Napoli Federico I
Luigi Sauro (Università di Napoli Federico II)
Eva Schlehahn (Unabhängies Landeszentrum für Dat

This document is licensed under a Creative Commons Attribution 3.0 Li

Abstract

This document specifies usage policy language of SPECIA
both the data subjects' consent and the data usage policie
by a computer, so as to automatically verify that the usage

The ontology defined in this document is publicly available

Vocabulary …/langs/usage-policy#

Bert Bos Last Updated: 17 April 2018

(You can download this ontology as an OWL file.)

The following is the formulation in functional syntax of the Usage Policy Language Ontology with identifier

> http://www.specialprivacy.eu/langs/usage-policy#

The documentation can be found in Policy Language V1 (deliverable D2.1).

```
# NAMESPACE DEFINITIONS

Prefix(spl:=<http://www.specialprivacy.eu/langs/usage-policy#>)
Prefix(owl:=<http://www.w3.org/2002/07/owl#>)
Prefix(rdf:=<http://www.w3.org/1999/02/22-rdf-syntax-ns#>)
Prefix(xml:=<http://www.w3.org/XML/1998/namespace>)
Prefix(xsd:=<http://www.w3.org/2001/XMLSchema#>)
Prefix(rdfs:=<http://www.w3.org/2000/01/rdf-schema#>)

# ONTOLOGY IRI AND ITS VERSION

Ontology( <http://www.specialprivacy.eu/langs/usage-policy-ontology>
<http://www.specialprivacy.eu/langs/usage-policy-ontology/1.0>
```

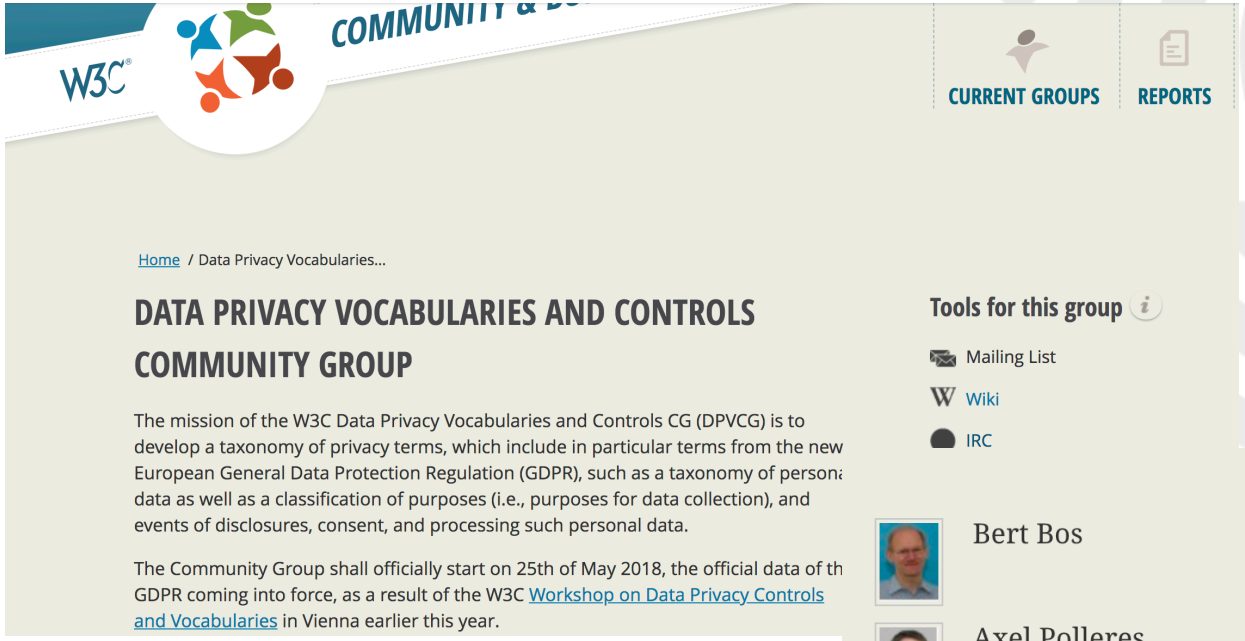
- The SPECIAL Usage Policy Language can be cited canonically as: "Bonatti, B. A., Kirrane, S., Petrova, I.M., Sauro, L., and Schlehahn, E., The SPECIAL Usage Policy Language, V0.1, (2018). https://aic.ai.wu.ac.at/qadlod/policyLanguage

- The SPECIAL Policy Log Vocabulary can be cited canonically as: "Bonatti, B. A., Dullaert, W., Fernández, J. D., Kirrane, S., Milosevic, U., and Polleres, A., The SPECIAL Policy Log Vocabulary, V0.3, (2018). https://aic.ai.wu.ac.at/qadlod/policyLog/

- The SPECIAL Vocabularies can be cited canonically as: "Bonatti, B. A., Kirrane, S., ePetrova, I.M., Sauro, L., and Schlehahn, E., The SPECIAL Usage Policy Language, V0.1, (2018). https://www.specialprivacy.eu/vocabs

Data Privacy, Vocabularies and Controls Community Group (DPVCG)

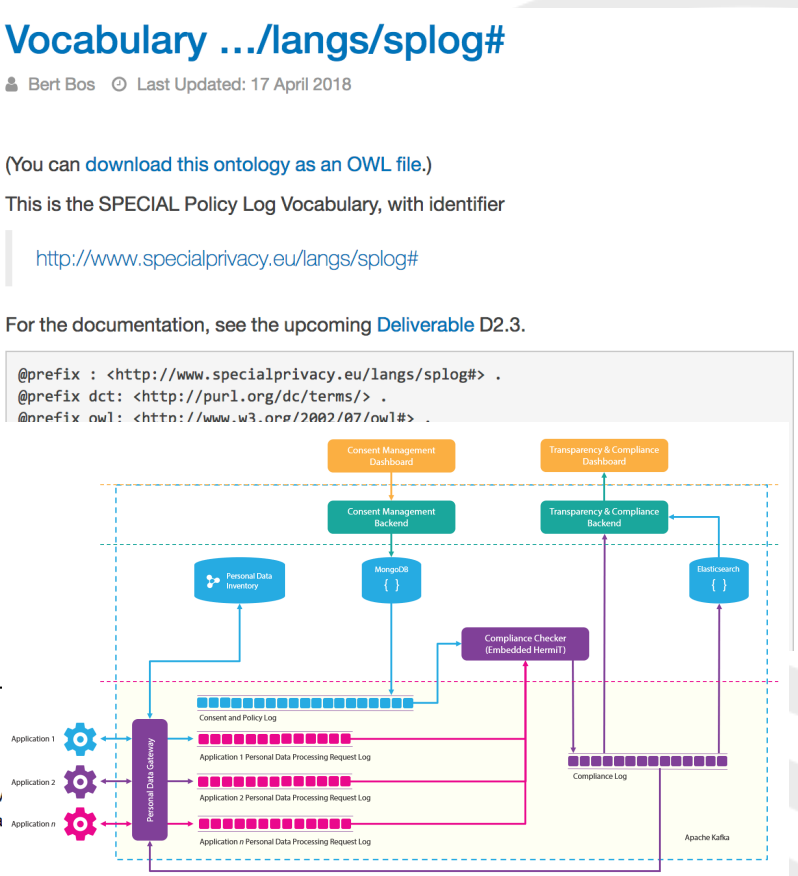
- Launched on the 25th of May 2018
- Presentation at MyData on the 31st of August-2018
- F2F in Vienna on the 3rd and 4th of December
- The current goal is to agree on first public drafts of minimal sets of vocabularies with first stable working drafts being reached latest on **July 2019**.

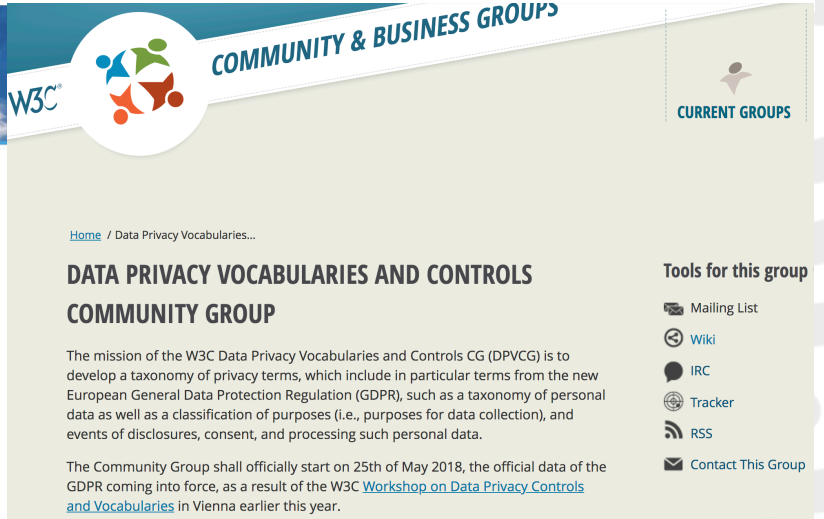


https://www.w3.org/community/dpvcg/

Exploitable Results

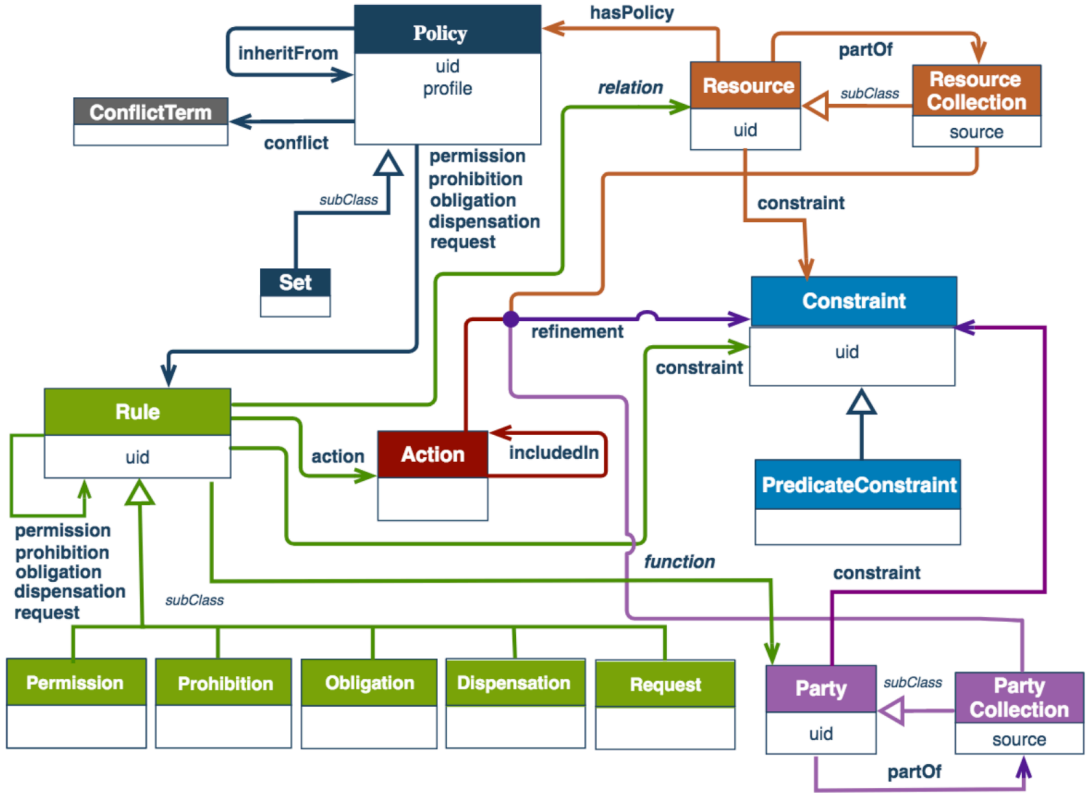
- Resources
  - ❖ The SPECIAL Usage Policy Language http://purl.org/specialprivacy/policylanguage
  - ❖ The SPECIAL Vocabularies https://www.specialprivacy.eu/vocabs
  - ❖ The SPECIAL Policy Log Vocabulary http://purl.org/specialprivacy/splog
- *SPECIAL Compliance Checking*
  - ❖ Demonstrates how usage policies together with event logs can be used to perform ex-post compliance checking
- *SPECIAL Consent and Transparency Interfaces*
  - ❖ Various consent user interfaces and the transparency dashboard
  - ❖ Guidelines for legally compliant consent retrieval





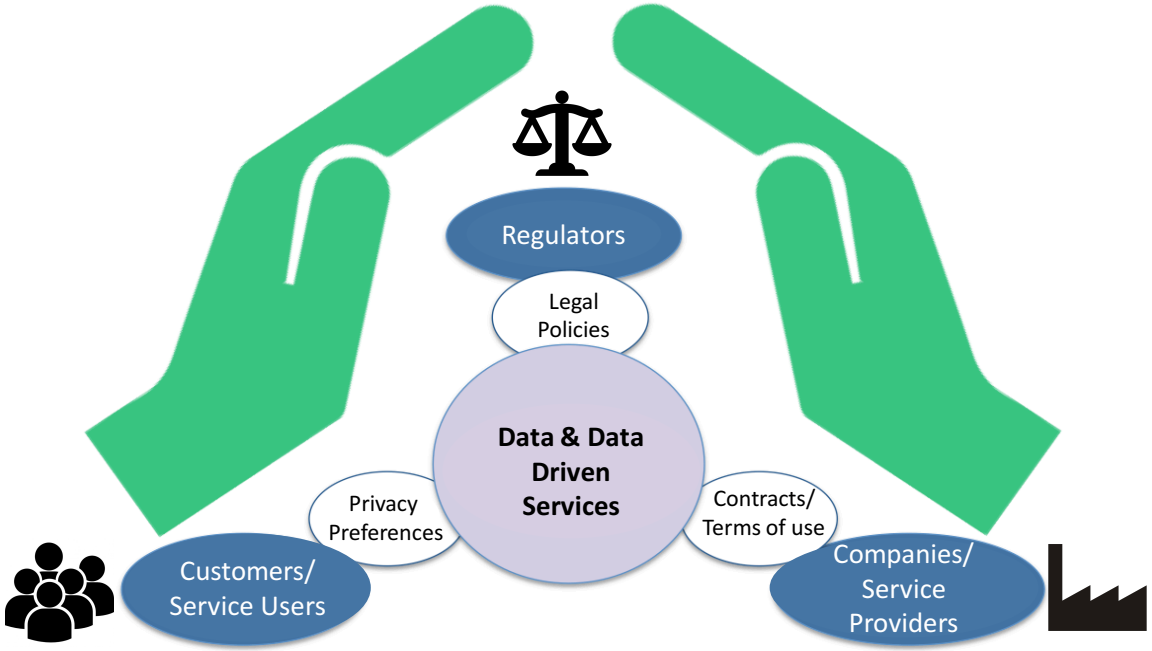
Ongoing / Future Work

Ongoing / Future Work: Policy Modeling & Reasoning



- Modeling regulatory obligations and business rules
- Compliance checking
- Conflict detection and resolution

Ongoing / Future Work: Transparency & Compliance



- Sticky Policies
- Trusted Environments
- Evaluating the strength of data synthesise techniques
- Policy based data science