# Privacy through Anonymisation in Large-scale Socio-technical Systems

## Multi-lingual Contact Centres across the EU

Claudia Cevenini  Enrico Denti  Andrea Omicini  Italo Cerno

{claudia.cevenini, enrico.denti, andrea.omicini, italo.cerno}@unibo.it

Dipartimento di Informatica – Scienza e Ingegneria (DISI)
Alma Mater Studiorum – Università di Bologna

*INSCI 2016*
Firenze, Italy, 14 September 2016

# Outline

# Context and Focus

- this research focusses on contact centres (CC) as relevant examples of *knowledge-intensive* socio-technical systems (STS)
- we discuss the articulate aspects of anonymisation
  - individual and organisational needs clash
  - call for an *accurate balancing* between *legal* and *technical* aspects
  - system *efficiency* while preserving the individual right to *privacy*
- we explore
  - first, the relevant *legal framework*
  - then, the *general theme* of anonymisation in CC
- we illustrate the general view of the technical process developed in the context of the BISON H2020 project

# Contact Centres as STS

## Typical technology issues of CC as STS

- speech data mining technologies with multi-language capabilities
- business outcome mining from speech
- CC support systems integrating both speech and business outcome mining in user-friendly way

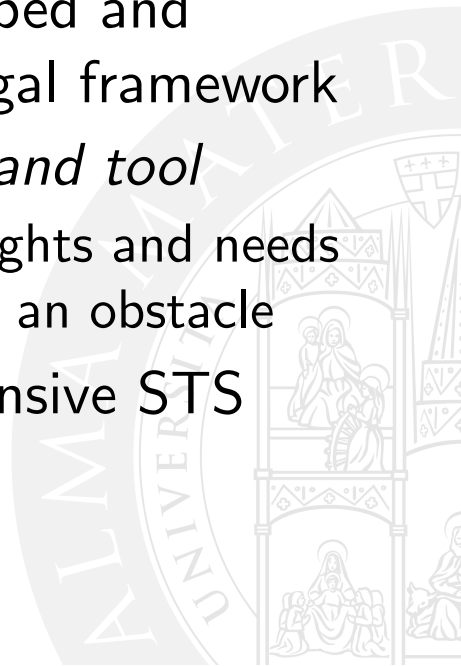
Scaling up to Big Speech Data

- applying data mining technologies *with multi-language capabilities* to big speech data
- implies a corresponding scale up of privacy and data protection issues

# Goal of the Research

- to assess how complex legal issues at *national* and *international* level can be faced while building a complex software infrastructure for CC
  - first, in the development phase
  - then, in the subsequent business phases
- to investigate how such infrastructures may be developed and marketed *in the full respect* of the Data Protection legal framework
- to focus on *anonymisation as a fundamental concept and tool*
  - to deal with the potential conflict between opposite rights and needs
  - able to provide further *value-added*, rather than being an obstacle

especially in the R&D of a large-scale, knowledge intensive STS



# Law & IT: a Focal Point

## Privacy vs. efficiency

- a suitable compromise between *law-abidingness* and *privacy* and system / process *efficiency* is a relevant goal
  - not just for the legal analysis
  - but for the whole engineering process of the CC infrastructure
- from a potential *conflict* of interests to *composition* of interests
  - from "Oh My God, the lawyers!" & "Oh My God, the engineers!"
  - to multiple competence together for a special kind of software product
- the requirement of *legal compliance* as a success factor instead of a possible source of delays and overheads
  - an issue going well beyond the CC case study
  - supporting anonymisation as a competitive advantage

# Outline

# Data Protection Directive (DPD)

## The Data Protection Directive (Dir 1999/95/EC) [DPD95]

- key principles for the *fair* and *lawful* processing of personal data
- technical and organisational *security measures* to guarantee that all personal data are safe from destruction, loss, alteration, unauthorised disclosure, or access, during the entire data processing period.

Highlights

- data processing requires even more care when it involves *large amounts* of personal and/or sensitive data
  - people's data flow across massive, third-party analytical systems
  - need of a transparent view of how people's data will be used (or sold)
- attention to data transfer from/to non-EU countries (and cloud)

# Personal Data

## What are personal data?

- any information relating to a natural person, who can be *identified*, either directly or indirectly, by reference to one or more factors specific to his/her physical, physiological, mental, economic, cultural, or social identity
- the notion of personal data is strictly related to identification
  - e.g. "John Smith" could/could not be personal data, depending whether it is enough to identify precisely one person
  - conversely, "John the fisher living at the end of the street" could be personal data, if it is enough to identify him
- if the link between an individual and personal data never occurred or is somehow broken and cannot be rebuilt in any way (such as with anonymised data), the DPD rules no longer apply

# Roles in Personal Data Processing

## Data controller vs. Data processor

- the data controller is in charge of personal data processing and takes any related decision
  - e.g., selection of data to be processed, purposes and means of processing, technical and organisational security, ...
- the data processor is a legally separate entity that processes personal data *on behalf of a controller*, in force of a written agreement and *following specific instructions*

For instance:

- a company acts as a controller in processing its own customers' data
- the CC entrusted with the same processing acts as a data processor on behalf of the company

# How to Process Personal Data According to the DPD

## Processing personal data

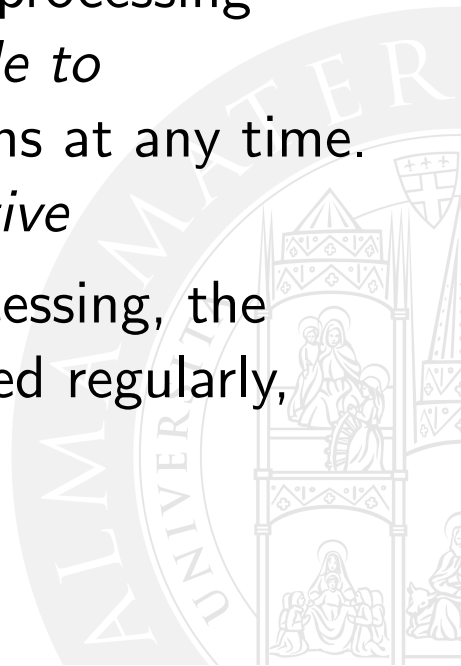
Personal data must be

- processed *fairly* and *lawfully*
- collected for *specified*, *explicit*, and *legitimate purposes* and not further processed in a way incompatible with those purpose
  - further processing of data for historical, statistical or scientific purposes may not be considered as incompatible, with appropriate safeguards
- *adequate*, *relevant* and *not excessive* in relation to the purposes
- *accurate* and, where necessary, kept *up to date*; inaccurate or incomplete data should be erased or rectified
- kept in a form which permits identification of data subjects for *no longer than is necessary* for the purposes.

# Accountability

According to the accountability principle

- data controllers must implement *adequate technical and organisational measures* to promote and safeguard data protection in their processing activities
- controllers are responsible for the compliance of their processing operations with data protection law and should be *able to demonstrate compliance* with data protection provisions at any time. They should also ensure that such measures are *effective*
- in case of larger, more complex, or high-risk data processing, the effectiveness of the measures adopted should be verified regularly, through monitoring, internal and external audits, etc.



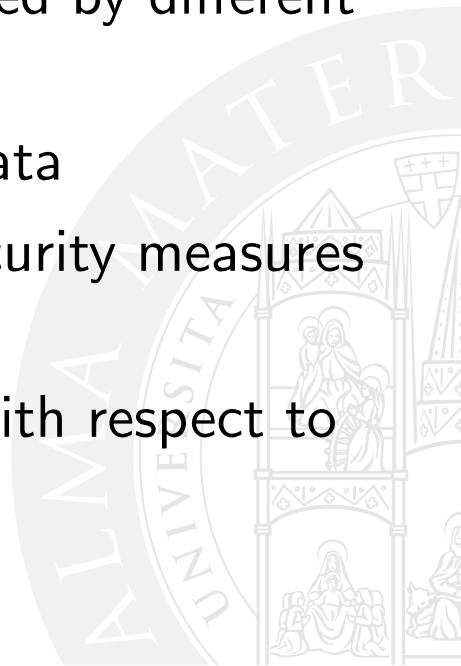
# Security Measures

Technical and organisational security measures should be adopted

- to protect personal data
- during all the processing period
- against the risks related to the integrity and confidentiality of data

The *level* of data security requested by the law is determined by different elements, such as

- the nature (sensitive/non-sensitive) of the collected data
- the concrete availability in the market of adequate security measures at the current state of the art
- their cost – which should not be "disproportionate" with respect to the necessity



# Big Speech Data Issues I

## Speech data

Speech recordings involve *biometric* data (tone, pitch, cadence, and frequency of a person's voice), suitable to determine someone's identity.

Highlights:

- from a Data Protection perspective, biometrics is linked to physical, physiological, behavioural, or even psychological characteristics of an individual – some of which may be used to reveal sensitive data
- biometric data may also enable automated tracking, tracing, or profiling of persons $\rightarrow$ potential high impact on privacy
- biometric data are by nature *irrevocable*
  - requires the informed consent of the data subject
    + ev. authorisations/notifications from/vs. Data Protection Authorities
    + strict security measures

# Big Speech Data Issues II

## Big Data

- big data analytics can involve the *repurposing* of personal data
  - personal data collected for one purpose cannot be reused/re-analysed for another purpose, without prior notification to the data subjects and new explicit consent (includes making data available to others to do so)
- big data may in themselves contrast with the principle of data minimisation and relevancy
  - the challenge for organisations is to focus clearly on their expectations from big data processing, so as to be able to verify that
    - the processing serve exactly the purposes for which data are collected
    - data are relevant and not excessive in relation to such aims

# Outline

# Relevant Principles I

- the legal framework foresees a set of essential principles
  - some directly derive from the DPD – namely, from the "Principles relating to data quality"
  - other concern the security measures – particularly w.r.t. the "Security of processing"
- these principles are further strengthened and detailed in the new "General Data Protection Regulation" (GDPR) [GDP16]

## Categories of principles

(a) principles about data processing

(b) principles about security measures

(c) other relevant principles

# Relevant Principles II

Principles of Data Processing

1. principle of lawfulness and fairness
2. principle of relevance and non-excessive use
3. principle of purpose
4. principle of accuracy
5. principle of data retention

Principles of Security Measures

1. principle of privacy by design and by default
2. principle of appropriateness of the security measures

Other Relevant Principles

1. principle of least privilege
2. principle of intentionality in performing any critical action

# Technological Requirements for Anonymisation

## Resulting requirements

- personal data may be processed only to the extent they are needed to achieve specific purposes
  - whenever identifying data are not actually necessary, anonymous data should be used
- the DPD does not apply to data rendered anonymous such that the data subject is no longer identifiable
  - it does not set any prescriptive standard
  - nor does it describe the de-identification process
    $\rightarrow$ just its outcome, i.e. a *reasonably-impossible* re-identification

# Outline

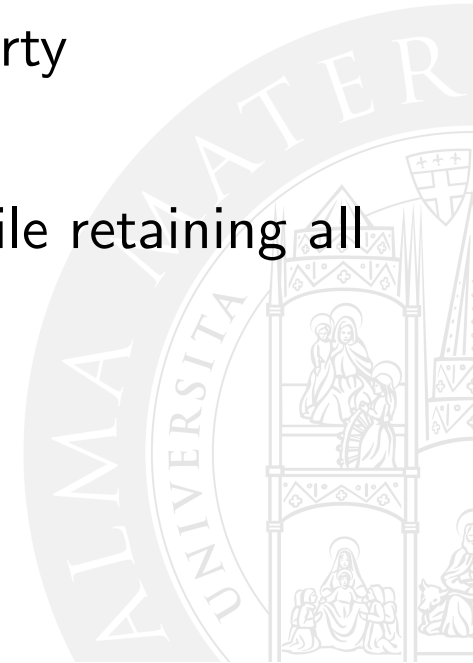
# Anonymise data = ... ?

In principle

- the DPD does not apply to data made anonymous in such a way that the data subject is *no longer identifiable*
- yet, irreversibly-preventing identification requires data controllers to consider all the means which may *likely reasonably* be used for identification, either by the controller or by a third party

But in practice...

- it is difficult to create a truly anonymous dataset, while retaining all the data required for a specific (organisational) task
- *likely reasonably* is inherently quite subjective
  $\rightarrow$ need for some shared, reasonable interpretation



# Article 29 Working Party

- the *Article 29 Working Party – Opinion on Anonymisation Techniques* (Art. 29 WP henceforth) [Dir14] is an important reference for compliance in anonymisation issues
- the criteria on which Art. 29 WP grounds its opinion on robustness focus on the possibility of
  - singling out an individual
  - linking records relating to an individual
  - inferring information concerning an individual.
- in GDPR, replaced by the *European Data Protection Board*

# Outline

# Anonymisation in BISON

Fundamental distinction

**research phase** — [during the project] when software and technologies are developed and tested, but are not yet in actual production

**business phase** — [after the project] when software and technologies will be used in CC, dealing with real customers data

- anonymisation as a fundamental tool
  - to set the research phase free from the complex DPD requirements
  - to comply with the purpose, relevance, and necessity principles
  - in the perspective, also a value-added component for partners to support other applications — not an overhead
- in the subsequent business phase, the system will also deal with real user data – in compliance with any applicable law
  - data processing will occur inside each CC, with appropriate consent

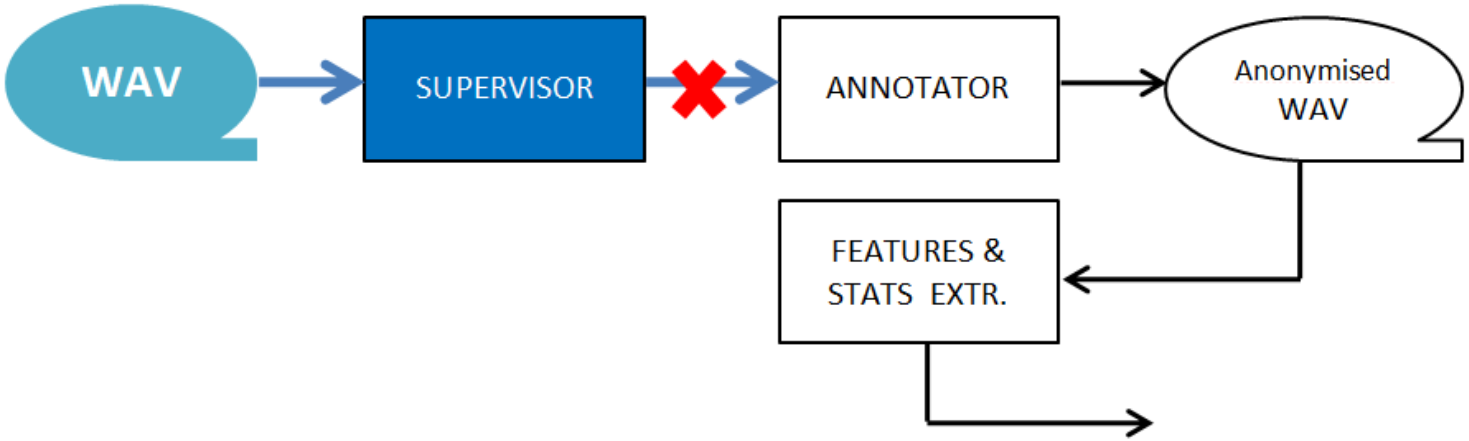
# Technological Requirements

- strict security requirements
  - fine-tuneable users' roles, rights, and restrictions
  - case-by-case configurability based on actual needs and national laws
- on-the-fly anonymisation
  - if some unexpected personal data are heard by the CC agent
- *privacy by default*
  - max anonymisation as the default setting
  - fine-grain customisation
  - lowering of privacy settings always explicit
  - lowering of privacy settings requires supervisor priviledges
- key challenge: make anonymisation future-proof
  - with respect to a continuously-evolving legal scenario
  - with respect to the (even-faster-evolving) technology improvement

# The Anonymisation Process: General Overview I

First stage of the BISON research

- limited data size, initial lack of automatic tools
- anonymisation is performed with manual procedures



# The Anonymisation Process: General Overview II

Second stage of the BISON research

- huge amounts of speech data
- automatic transcription – for all the supported languages
- anonymisation now occurs on the original audio file, not on a manually pre-silenced file
  - automatic anonymisation possibly not 100% effective
  - any effort made to reduce errors to the minimum
  - subsequent feature extraction completes the process.

# Outline

# Conclusions

- contemporary software engineering requires *non-computational issues* – normative, organisational, societal – to be kept into account
  - the law-abidingness of large-scale STS, including both human and software agents, is an intricate issue
  - must be faced in the requirement stage of any reliable software engineering process
- anonymisation of speech data in CC
  - calls for an *accurate balancing* between legal and technical aspect to achieve efficiency while preserving the right to privacy
- the legal framework can actually translate into requirements for the software engineering process
- the BISON case shows how the anonymisation process can be structured *during the research phase* to enable the resulting system to properly manage the data in the future business operation phase

# References


Article 29 Data Protection Working Party – Opinion 05/2014 on anonymisation techniques.
`http://ec.europa.eu/justice/data-protection/article-29/`, 18 April 2014.
0829/14/EN WP216.

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
*Official Journal of the European Communities*, 38(L 281):31–50, 23 November 1995.

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).
*Official Journal of the European Communities*, 59(L 119):1–88, 4 May 2016.

# Privacy through Anonymisation in Large-scale Socio-technical Systems

## Multi-lingual Contact Centres across the EU

Claudia Cevenini  Enrico Denti  Andrea Omicini  Italo Cerno

{claudia.cevenini, enrico.denti, andrea.omicini, italo.cerno}@unibo.it

Dipartimento di Informatica – Scienza e Ingegneria (DISI)
Alma Mater Studiorum – Università di Bologna

*INSCI 2016*
Firenze, Italy, 14 September 2016