

WP 29 Opinion on anonymization techniques

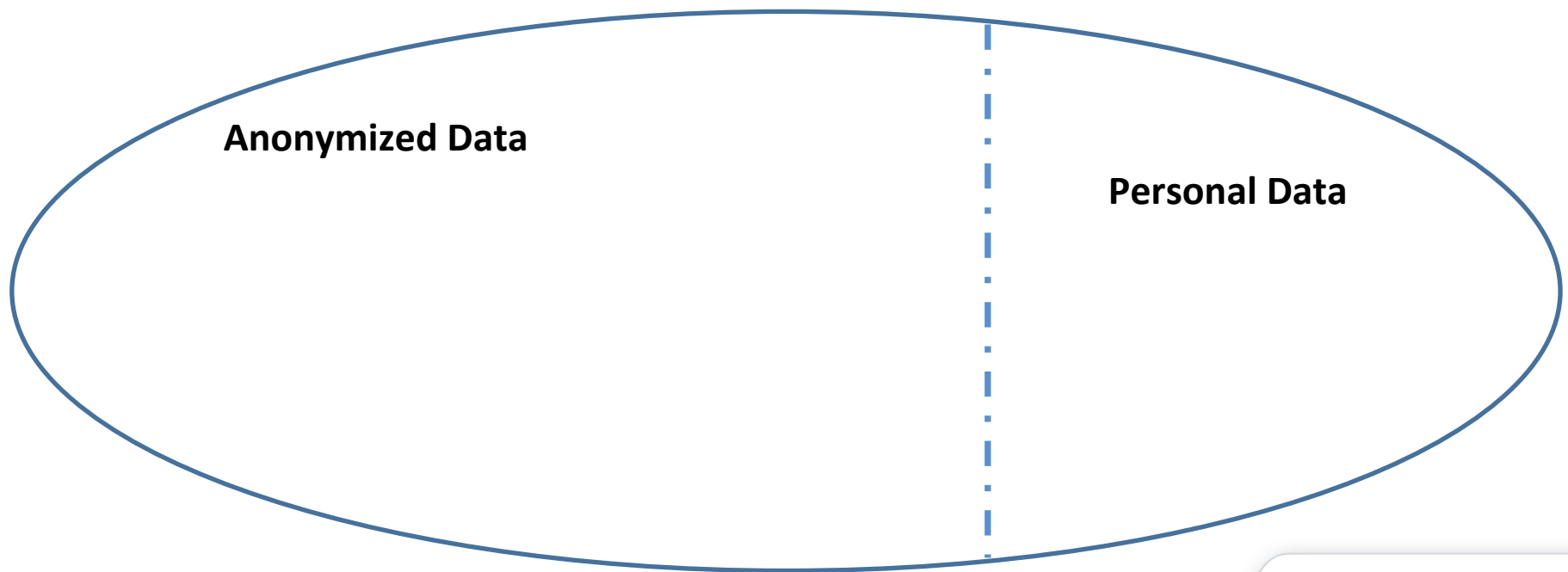
**Clinical Trial Data Sharing:
Methods and Experiences with De-Identification**
Paris, France
29th May 2015

Délia RAHAL- LÖFSKOG- Head of Health Department
Vincent TOUBIANA- IT expert

The need to define Anonymization

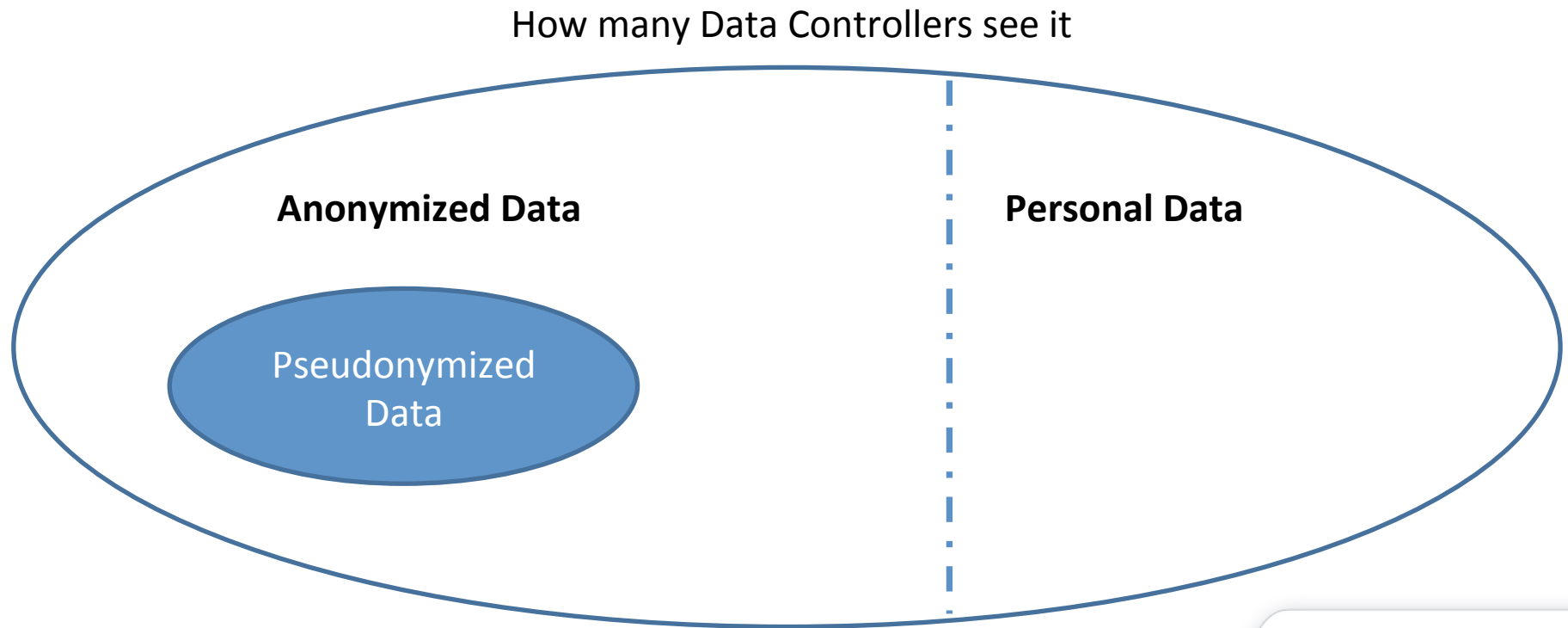
■ Dichotomy between “Personal Data” and “Anonymized Data”.

- There is no intermediary state,
- Anonymized data not subject to DPL (can be published).



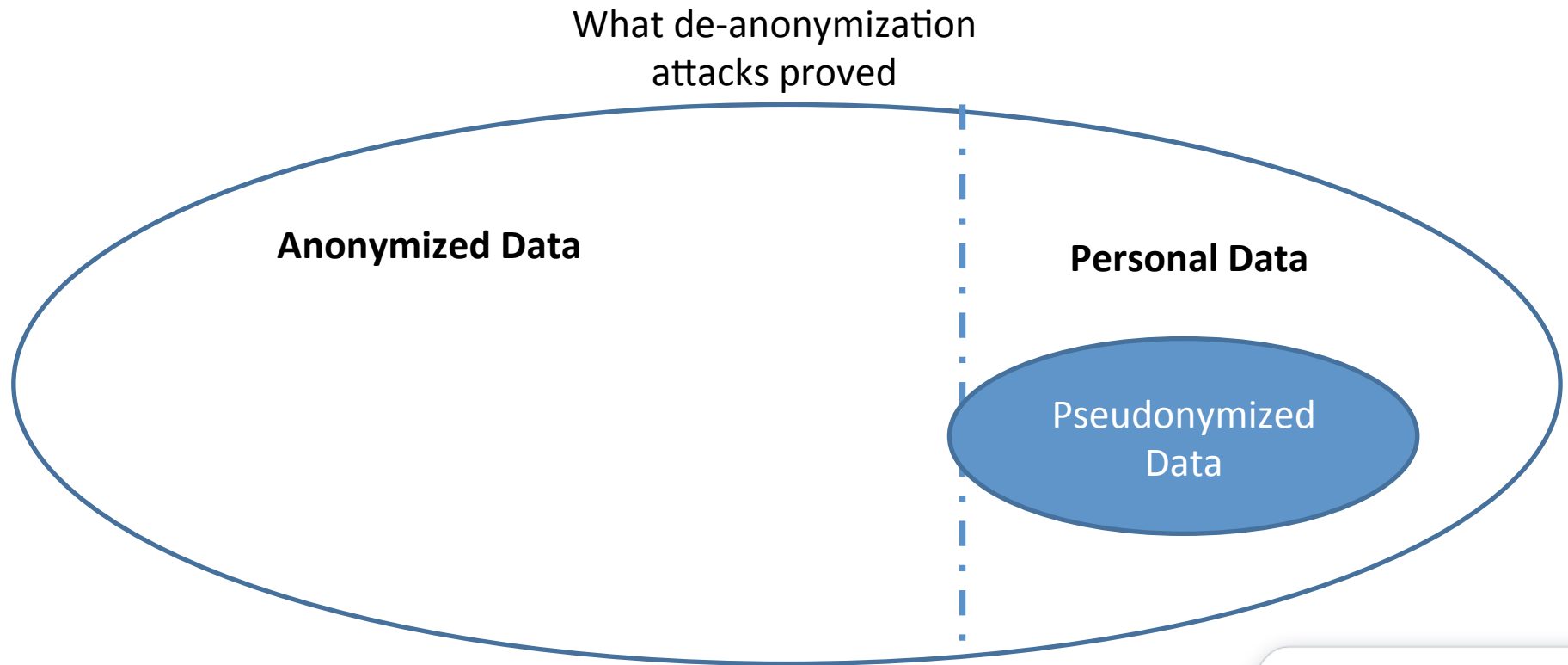
Pseudonymization and De-identification

- For most Data Controller: Removing PII → Anonymization



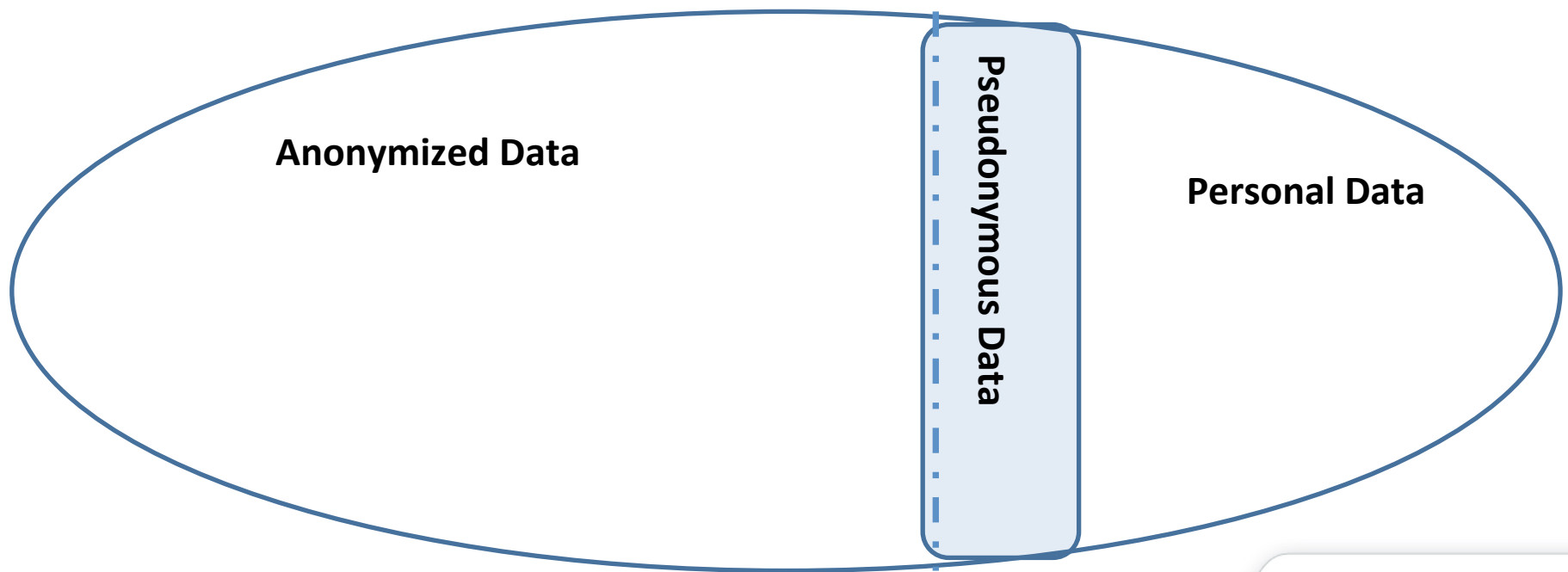
Pseudonymization and De-identification

- For most Data Controller: Removing PII → Anonymization
- The opinion insists that its not the case: Pseudonymized ≠ Anonymized



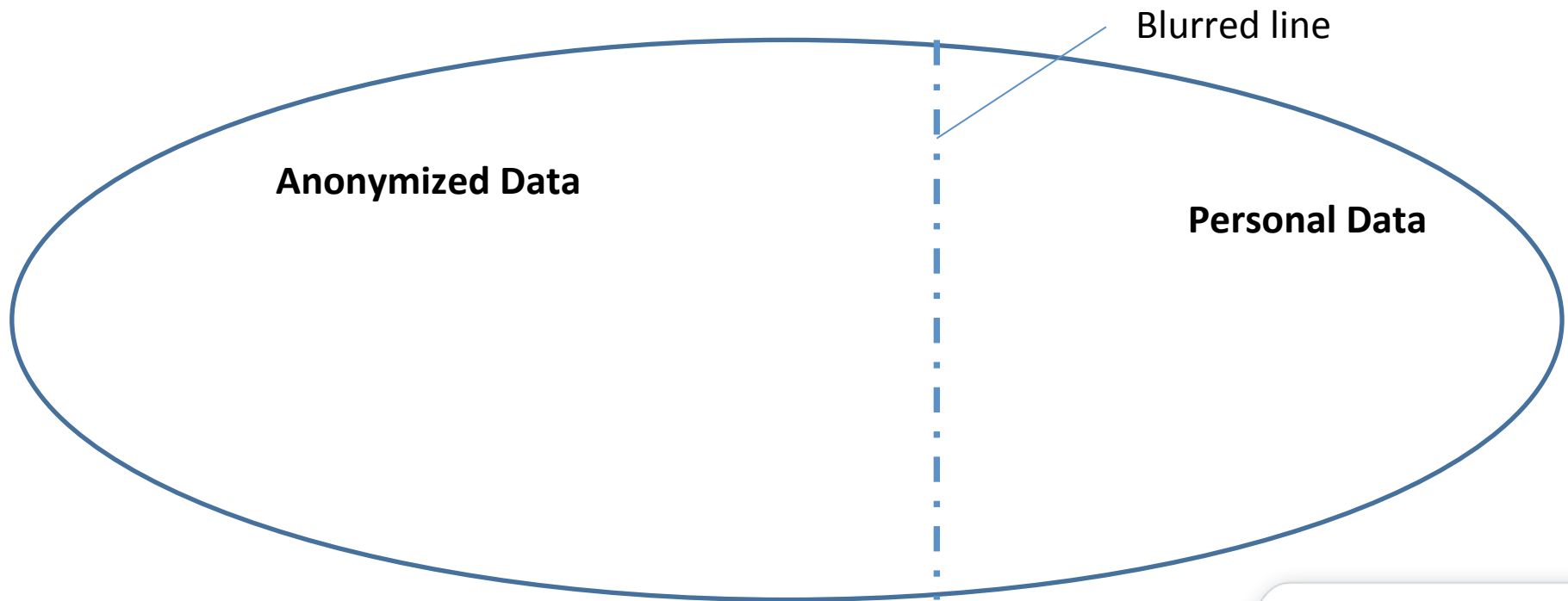
Pseudonymization and De-identification

- For most Data Controller: Removing PII → Anonymization
- The opinion insists that its not the case: Pseudonymized ≠ Anonymized
- New regulation could introduce “pseudonymous data” as a special case
 - *“personal data that cannot be attributed to a specific data subject without the use of additional information”*



Pseudonymization and De-identification

- For most Data Controller: Removing PII → Anonymization
- The opinion insists that its not the case: Pseudonymized ≠ Anonymized
- New regulation could introduce “pseudonymous data” as an intermediary state



Anonymization

- No clear line between Anonymized and Personal data
- The opinion provides two options to check that a Dataset is anonymized:

1. Your dataset has none of the following property:

- **Singling out:** possibility to isolate some records of an individual in the dataset;
- **Linkability:** ability to link, at least, two records concerning the same data subject or a group of data subjects (in the same database or in two different databases);
- **Inference:** the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes

OR

2. Make analysis of re-identification risk.

Evaluation of the anonymization techniques

	Is Singling out a risk	Is Linkability a risk	Is Inference a risk
Pseudonymisation	Yes	Yes	Yes
Noise addition	Yes	May not	May not
Substitution	Yes	Yes	May not
Aggregation or K-anonymity	No	Yes	Yes
L-diversity	No	Yes	May not
Differential privacy	May not	May not	May not
Hashing/Tokenization	Yes	Yes	May not

No single technique eliminates all risks

Is Big Data compatible with anonymity

- “No Inference” means no “Big Data” about individuals.
- Inference can still be done about environment, objects, consumptions ...know how your dataset will be used.
- Re-identification risk analysis option may not prohibit inference about individuals.

Recommendations

■ Prior to anonymization:

- **Specify your use of the anonymized dataset;**
- **Distinguish key attributes** from pieces of information that are not critical;
- **Remove identifiers and infrequent values.**

■ Find a tailored solution

- **Combine anonymization techniques;**
- **Document choices;**
- **Assess the re-identification risk and have your assessment validated:**
 - Consider all possible source of data (open data, anonymized dataset, recorded data, etc.),
 - Data utility, attackers incentives, insiders, acquaintance of the data subject.

■ Afterwards:

- **Publish your anonymization techniques and have them reviewed;**
- **Follow development** of anonymization and re-identification techniques

Thank you!

Opinion available here:

- http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Emails:

Vincent Toubiana:

vtoubiana_at_cnil.fr

Délia RAHAL-LÖFSKOG:

drahal-lofskog_at_cnil.fr