



The presenters:

Aleksei Kelli

Penny Labropoulou







#### The authors:

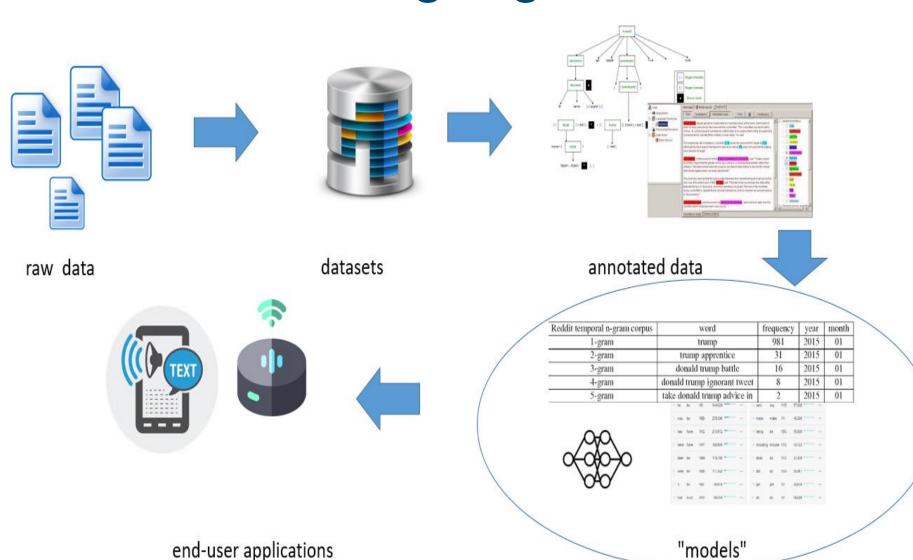
Aleksei Kelli, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv



#### The aims of the research:

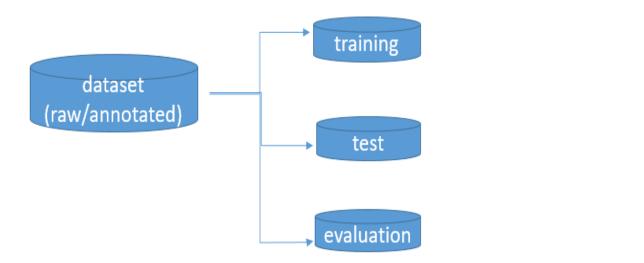
- To determine the extent of impact on the development of language technologies (LTs);
- To define a favourable regulatory framework for the development of LTs.

## From Language Data to LTs:



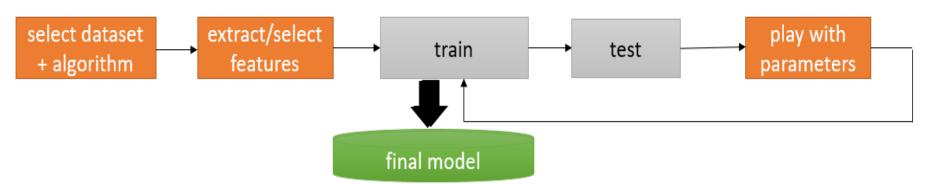


### Building models process:



human-driven

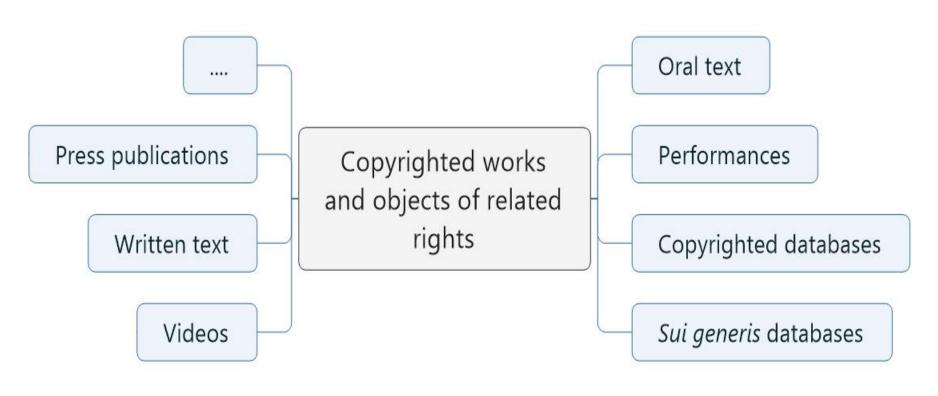
auto processing





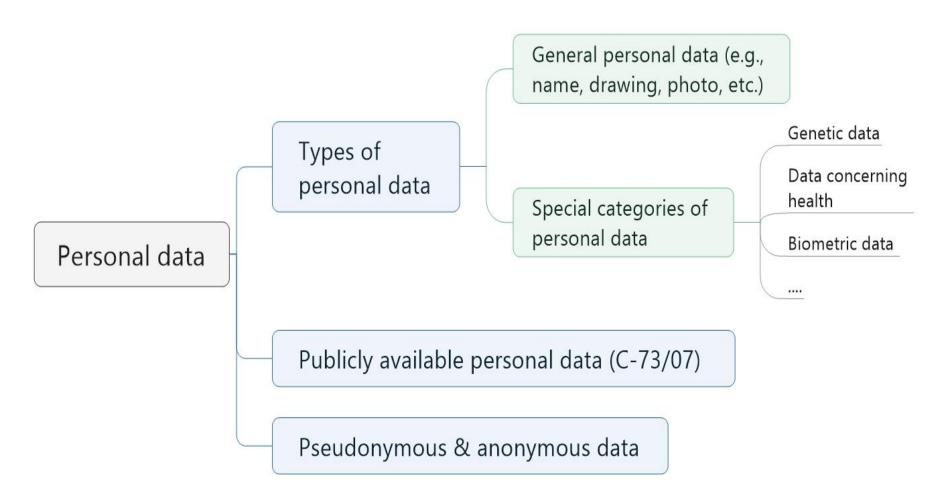


## Language data & copyright:





### Language data & personal data:





#### Use of LD & copyright protection:

Quotation right (InfoSoc Art. 5 (3))

Right holder's consent

Text and data mining exception (Digital Copyright Directiv Art. 3, 4)

Legal bases for using data-sets

Temporary acts of reproduction exempted from the reproduction right (InfoSoc Art. 5 (1))

Private use exception (InfoSoc Art. 5 (2))

Research exception (InfoSoc Art. 5 (3))



#### Use of LD & personal data protection:



Performance of a task in the public interest or in the exercise of official authority

Legitimate interests

Protection of the vital interests of the data subject or of another natural person

Legal grounds for lawful processing of personal data (GDPR Art. 6)

Performance of a contract

Consent

Compliance with a legal obligation





## Research use & copyright:



Covers works, objects of related rights

Restriction of technological protection measures and contractual provisions

Right to use copies with an appropriate level of security for the purposes of scientific research (e.g., verification) Text and data mining as any automated analytical technique (Digital Copyright Directive Art. 3) Right to make reproductions and extractions for research purposes

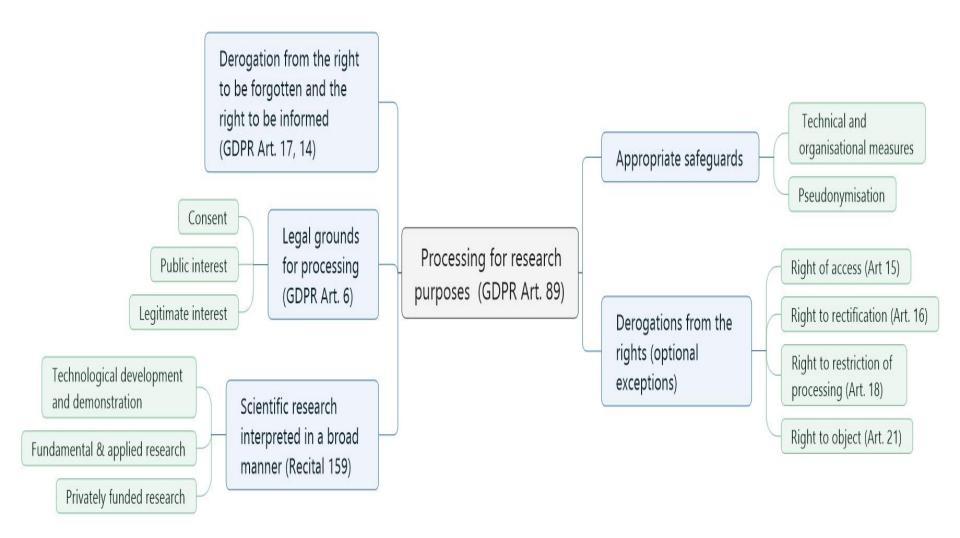
Covers research organisations & cultural heritage institutions

Lawful access as a precondition





#### Research use & personal data:





Models

#### Models:



Whether copyrighted material is used

Legal ground for the use of copyrighted material

Legal status of the model



#### Models:



Recreation of personal data/copyrighted work from models

Model as a derivative work?

Substantial parts of the input data set remaining in the model

Model as a result of human intellect investment

Choice/creation of a data set

Choice/creation of the programme used for the training of the model

Various cycles of testing and validation by tuning the parameters of the training programme







use Finished product develop Semi-finished product develop Model copyrighted works and personal data train Annotated dataset annotate Dataset gather raw data record

# Personal data problems concerning voice:

Scenarios to overcome problems relating to human voice in a model

Argue that voice without any identifying information is not personal data

Ask for consent for commercial use/ public dissemination

Argue that the use of voice in the model is based on the legitimate interest

#### **Conclusions:**

- Models are not necessarily subject to the same copyright and personal data restrictions as language data used as input;
- The limited impact of copyright and personal data restrictions on models potentially allows to:
  - Rely on favourable copyright/personal data provisions supporting research;
  - Make models publicly available;
  - Commercial use of models;
  - Personal data problems with voice need to be addressed.





## Thank you