My Europe, My Language

A Language Data Space for Europe

From vision to implementation

Philippe Gelin

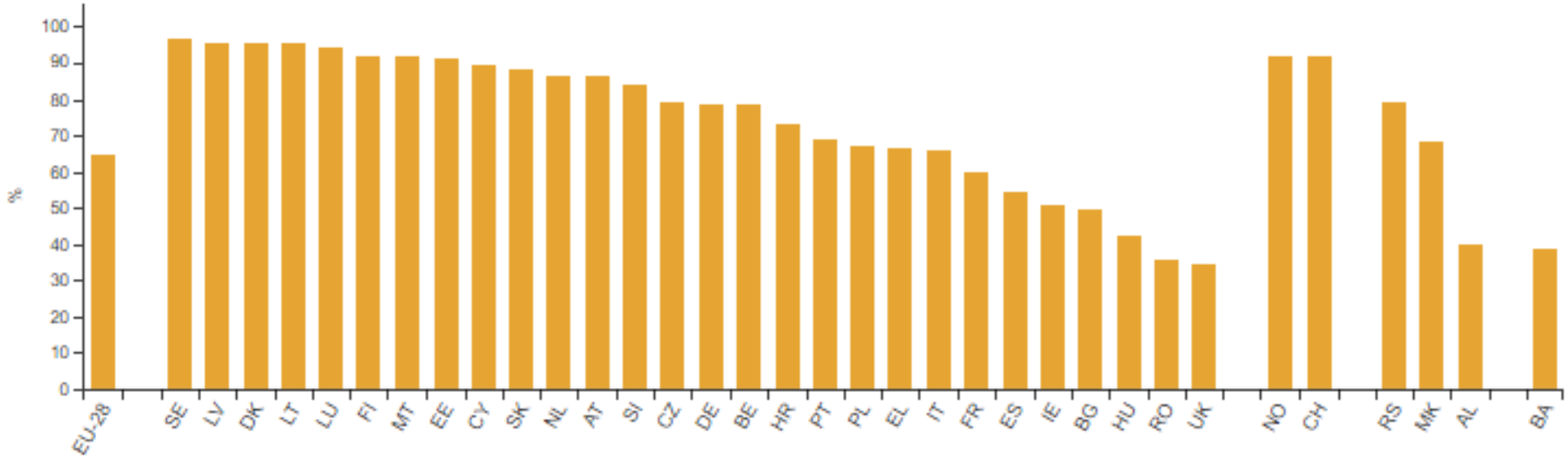
Head of Sector Multilingualism
Data Directorate
Directorate-General for Communications Networks, Content and Technology
European Commission

Virtual Cavtat, 7th october 2021 – 10:10 – 10:40

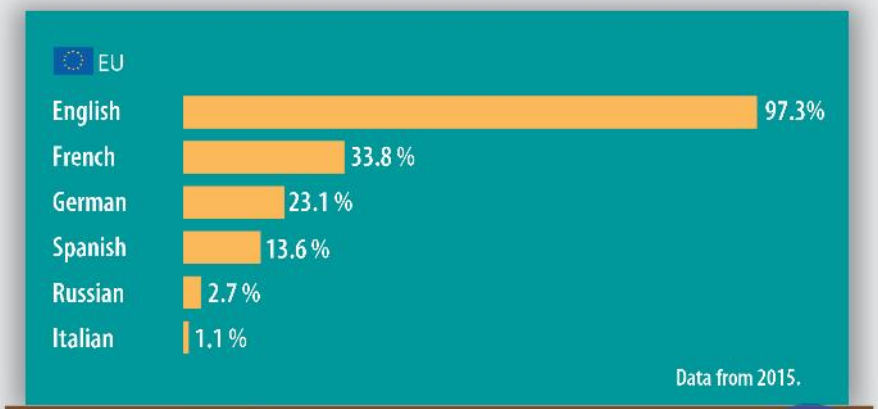
The European Landscape

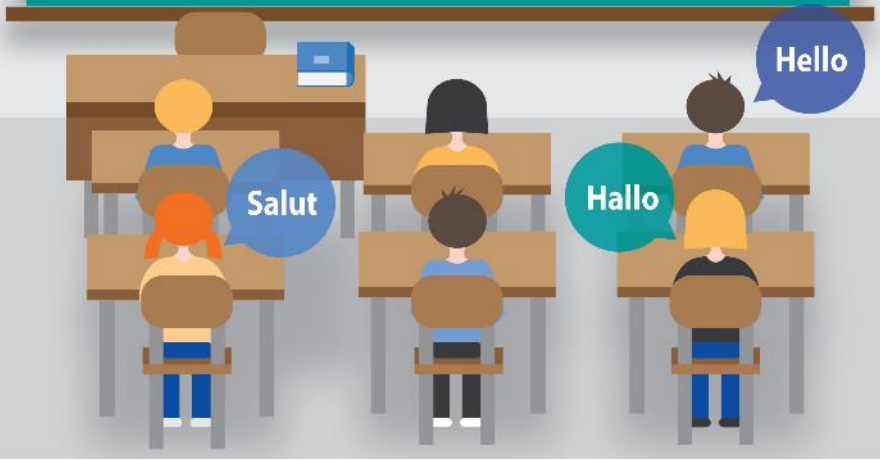
What if…
a personal view.

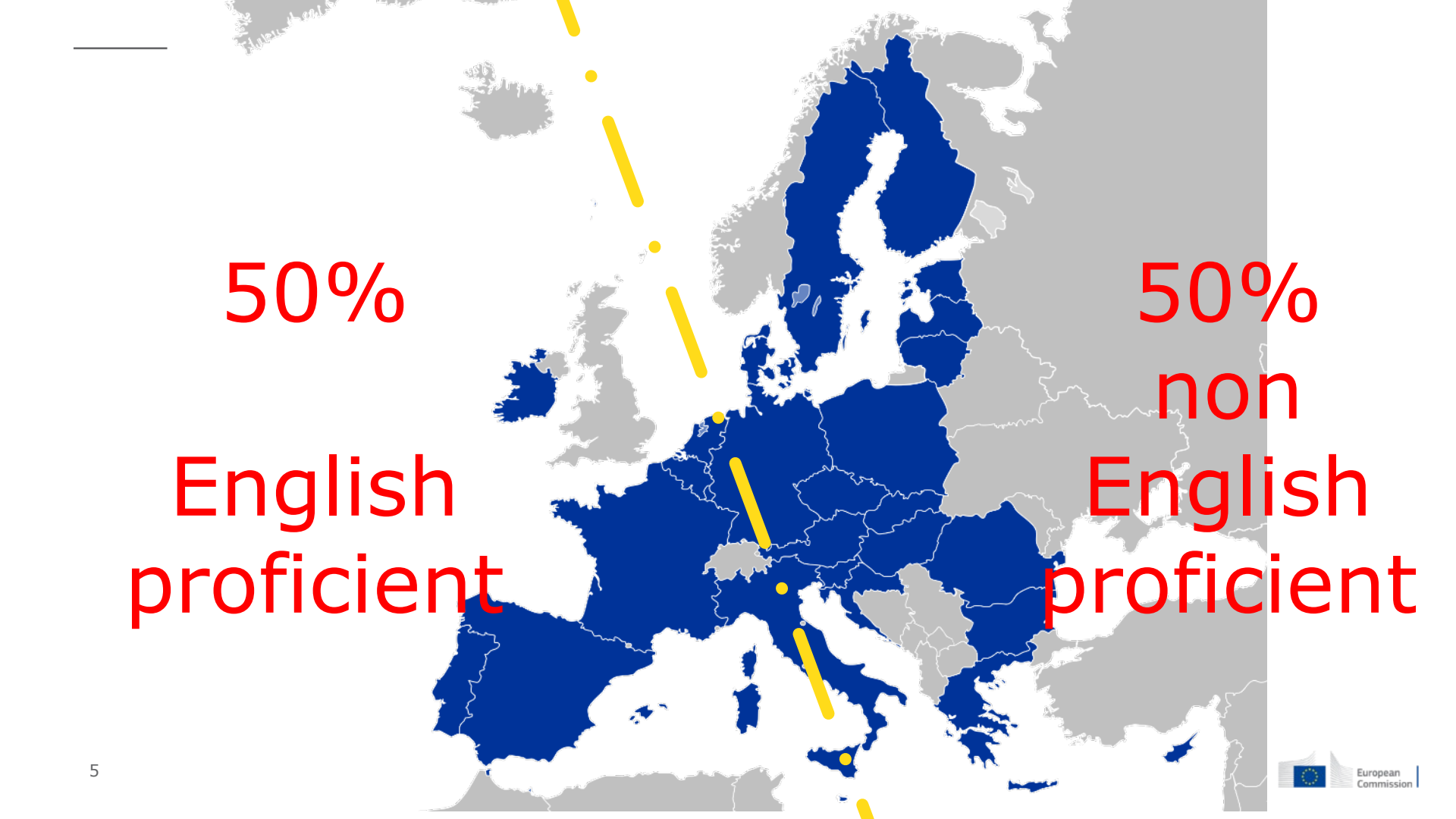
Population aged 25–64 reporting they knew one or more foreign languages, 2016



Which are the most studied foreign languages?
(% of pupils at lower secondary level)

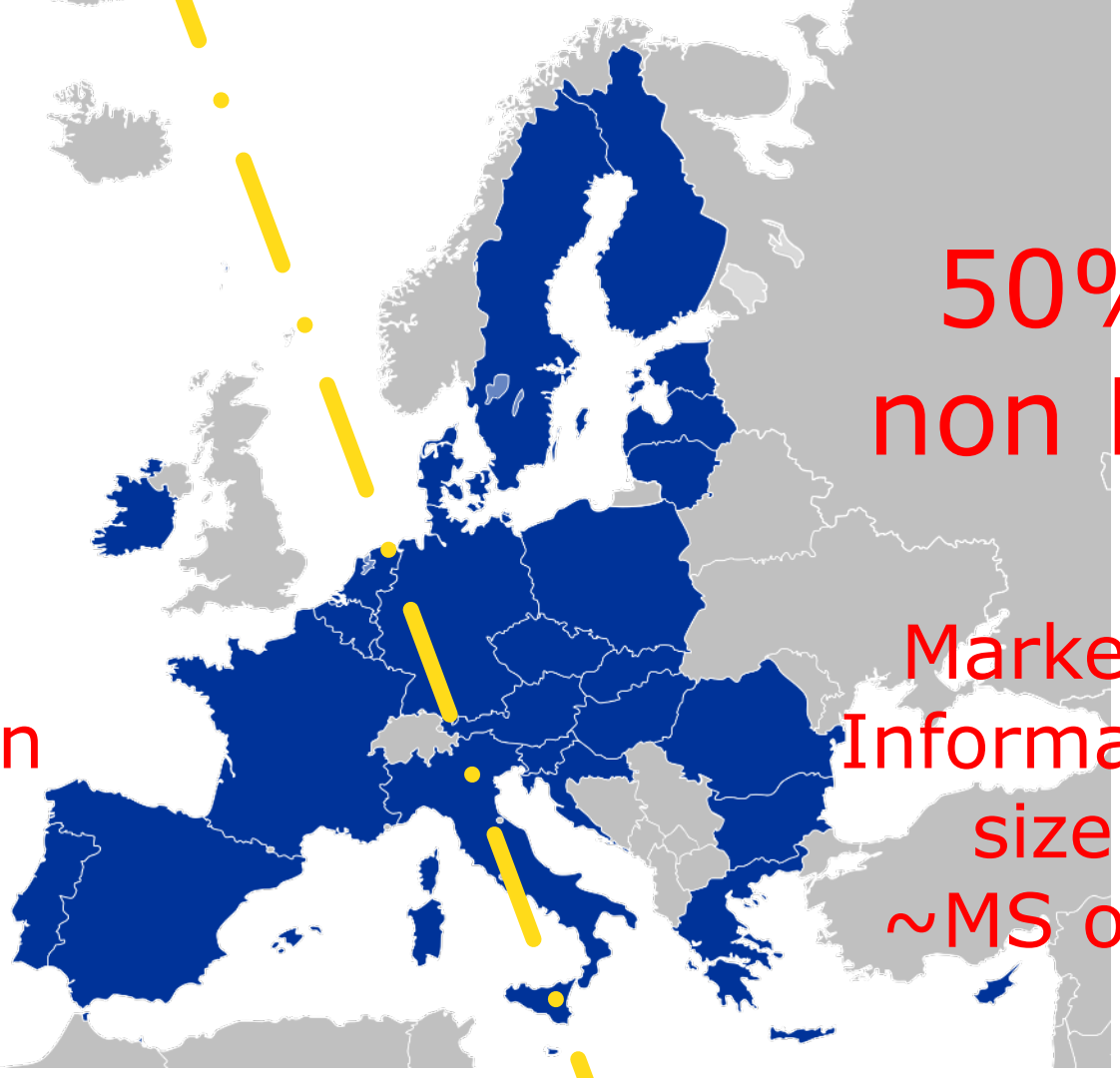






50%
EN

Market /
Information
size:
EU/2

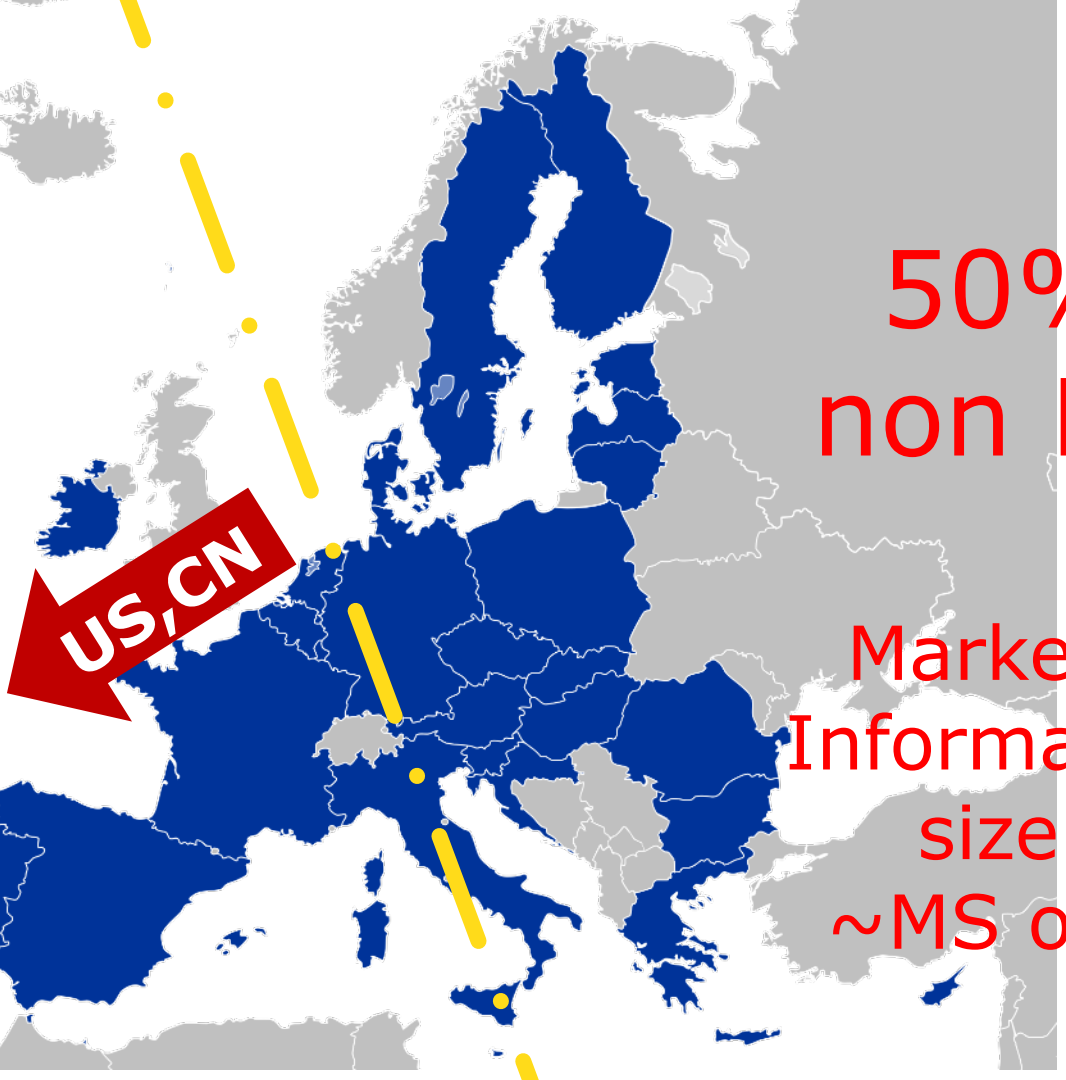


50%
non EN

Market /
Information
size:
~MS only

50%
EN

50%
non EN



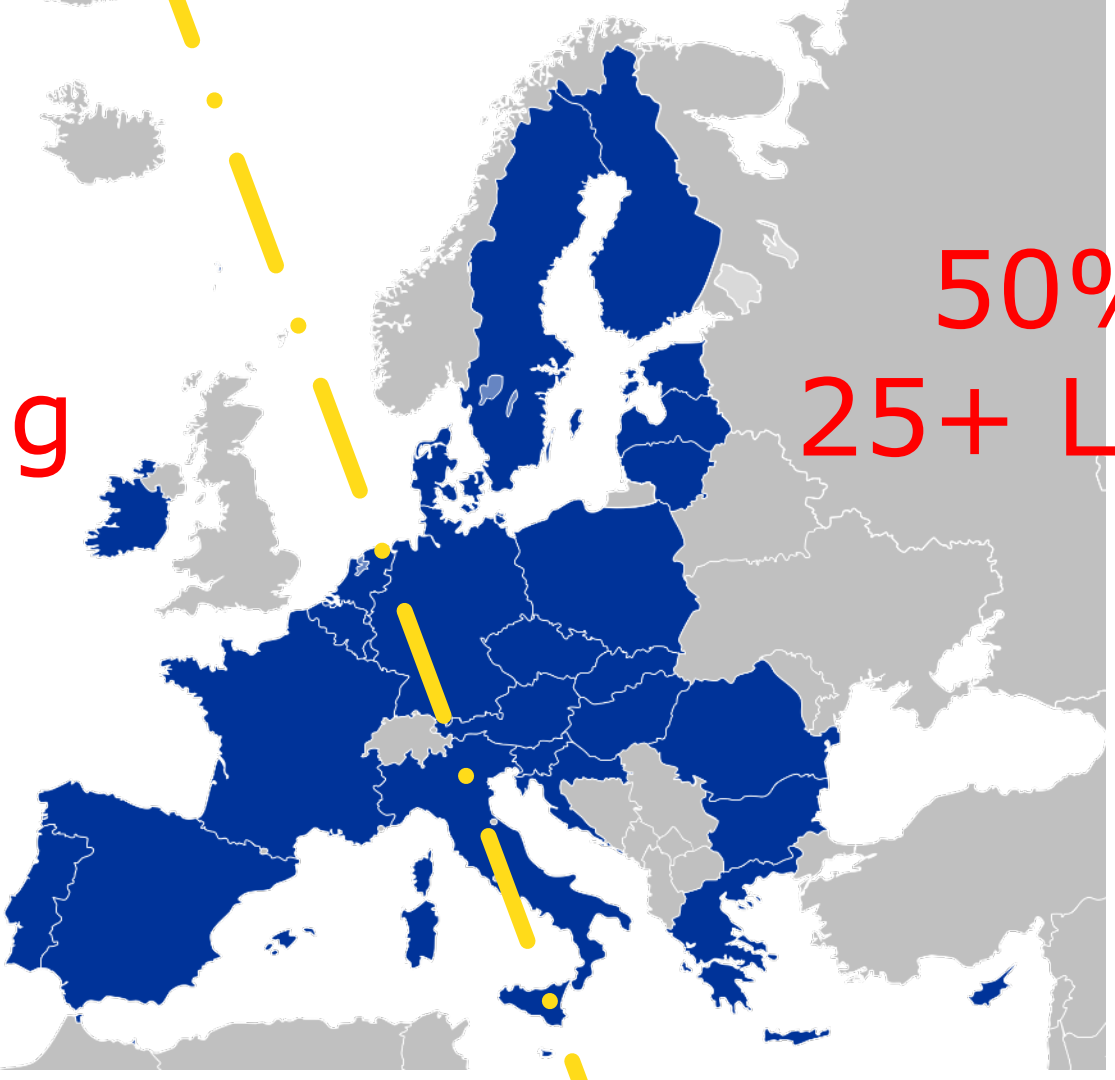
Market / Information size: EU/2

Market / Information size: ~MS only

European Commission

50%
25+ Lang

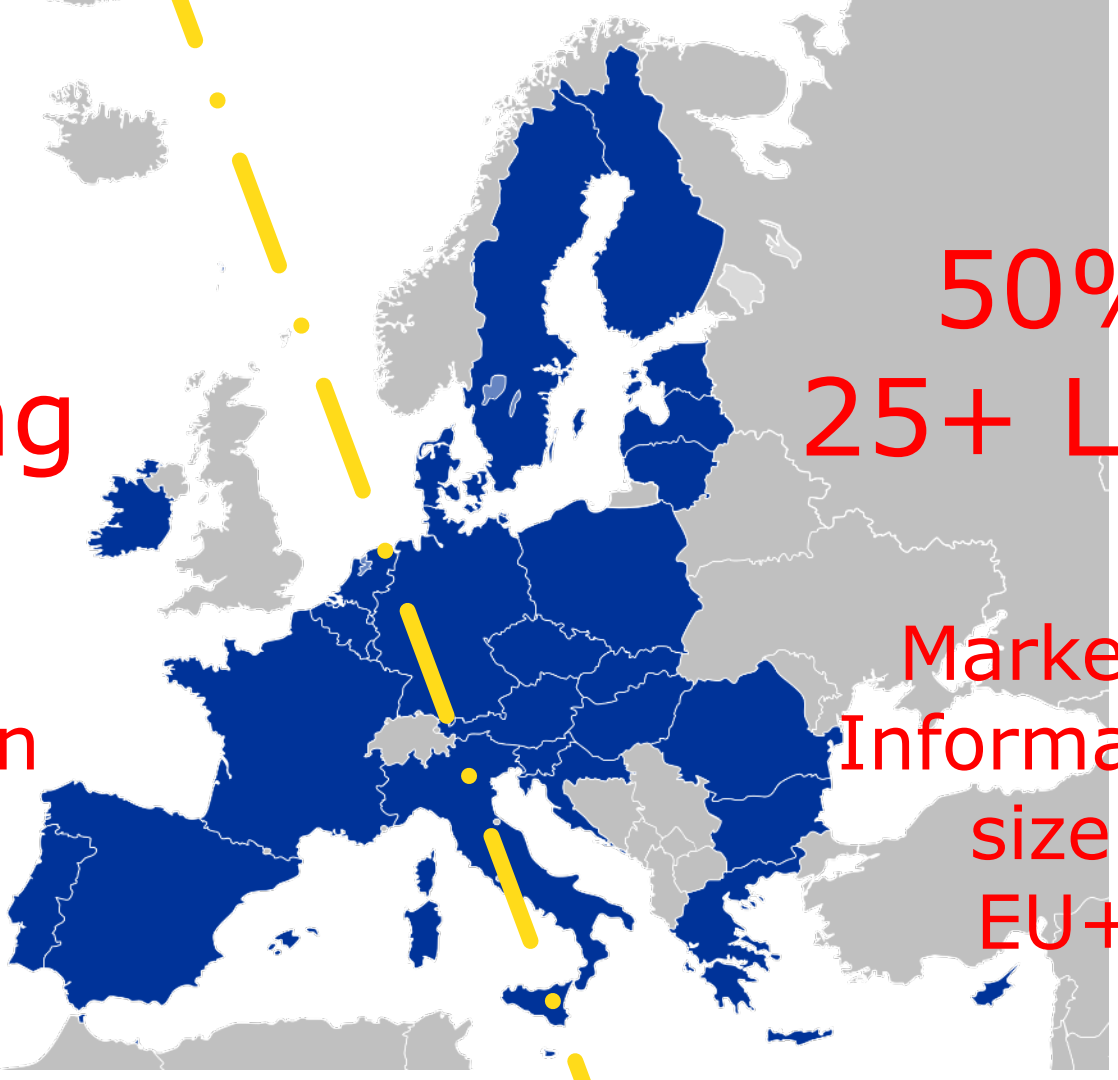
50%
25+ Lang



European Commission

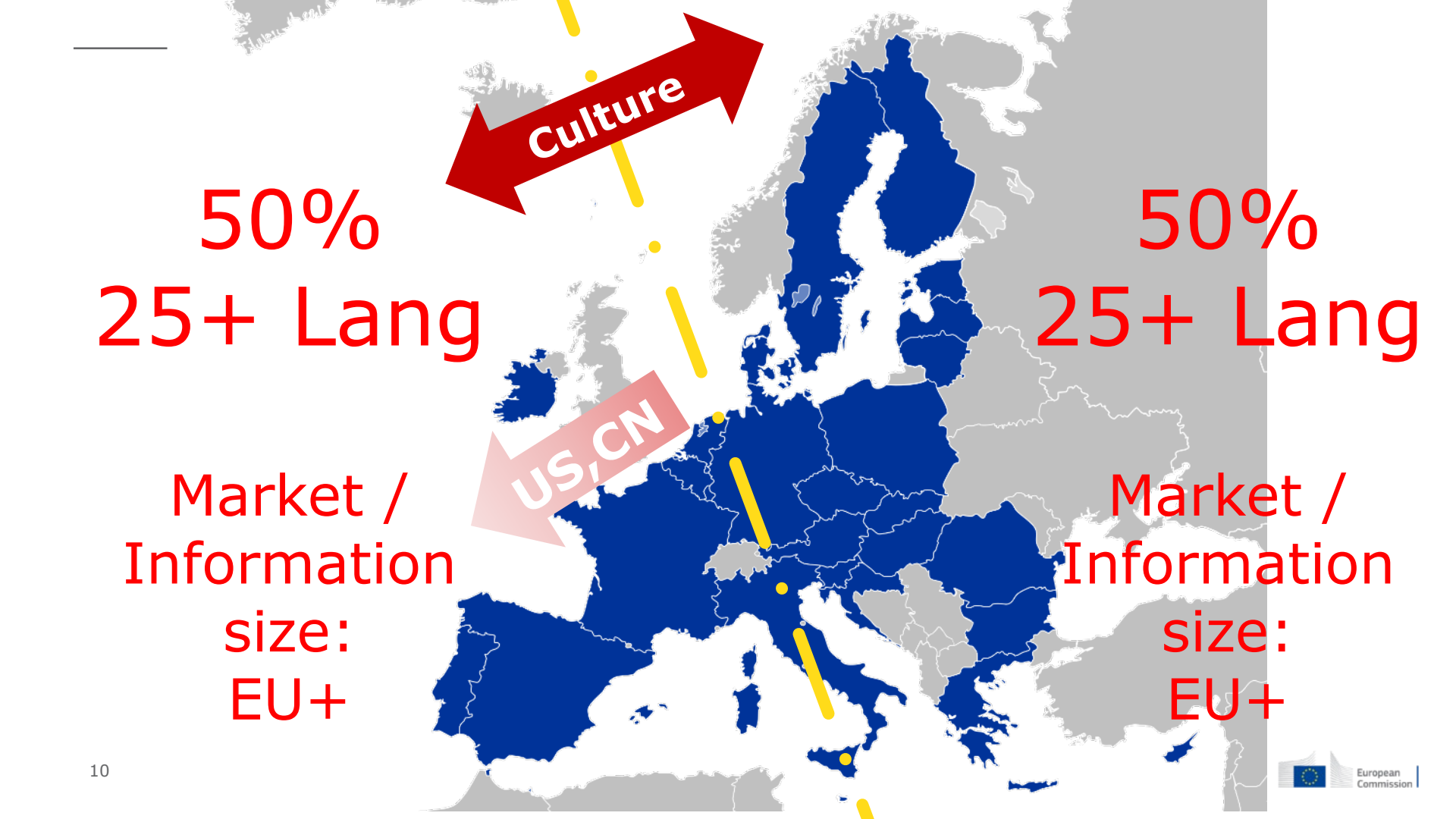
50%
25+ Lang

Market / Information size: EU+



50%
25+ Lang

Market / Information size: EU+



50%
25+ Lang

Market / Information size:
EU+

50%
25+ Lang

Market / Information size:
EU+

European Commission

Data Spaces

A DIGITAL Strategy

European Data Spaces

Rich pool of data (varying degree of accessibility)

Free flow of data across sectors and countries

Full respect of GDPR

Horizontal framework for data governance and data access

- Technical tools for data pooling and sharing
- Standards & interoperability (technical, semantic)
- Sectoral Data Governance (contracts, licenses, access rights, usage rights)
- IT capacity, including cloud storage, processing and services

Data infrastructures and governance go hand in hand

European Data Spaces

Rich pool of data (varying degree of accessibility)

Free flow of data across sectors and countries

Full respect of GDPR

Horizontal framework for data governance and data access

50% is composed of Language Data

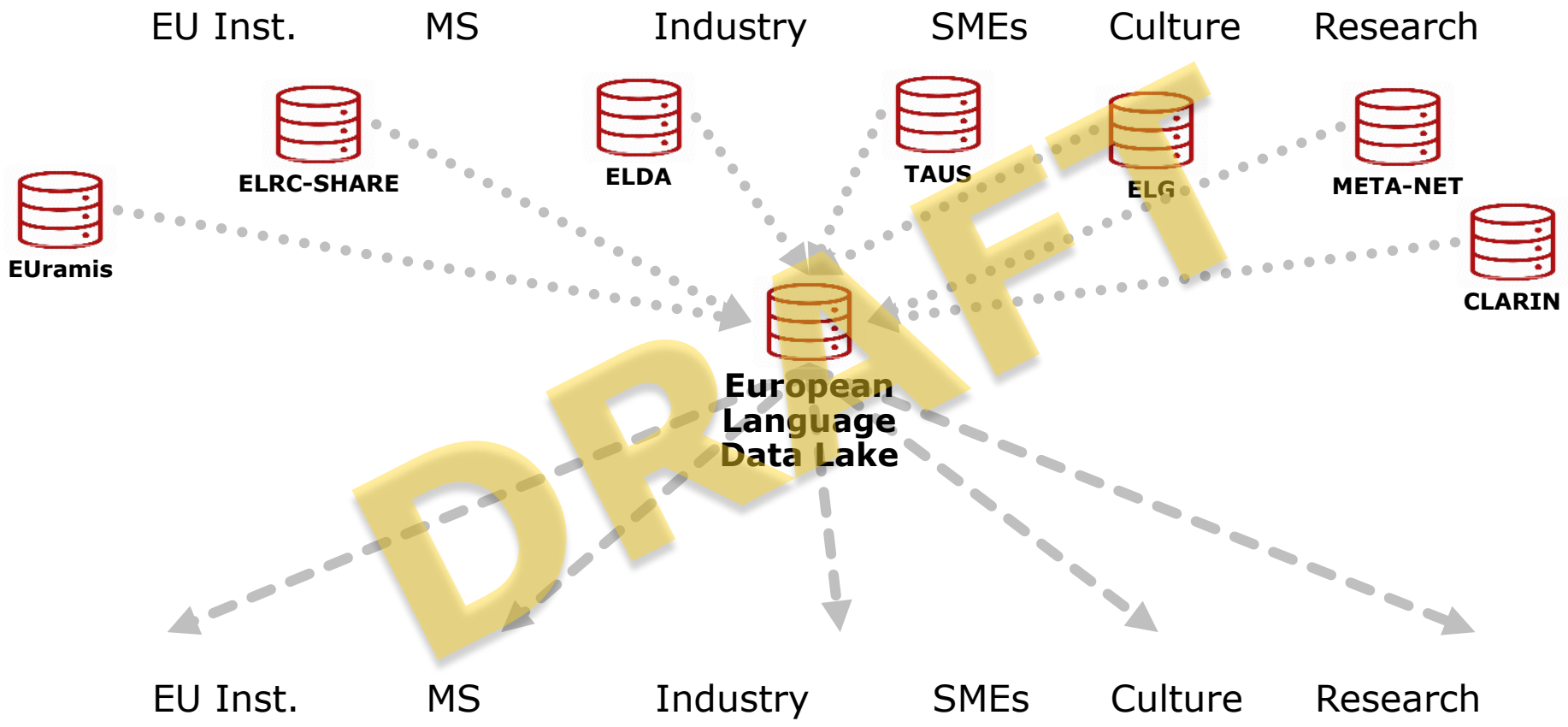
- Technical tools for data pooling and sharing
- Standards & interoperability (technical, semantic)
- Sectoral Data Governance (contracts, licenses, access rights, usage rights)
- IT capacity, including cloud storage, processing and services

Data infrastructures and governance go hand in hand

What if...

A Language Data Space

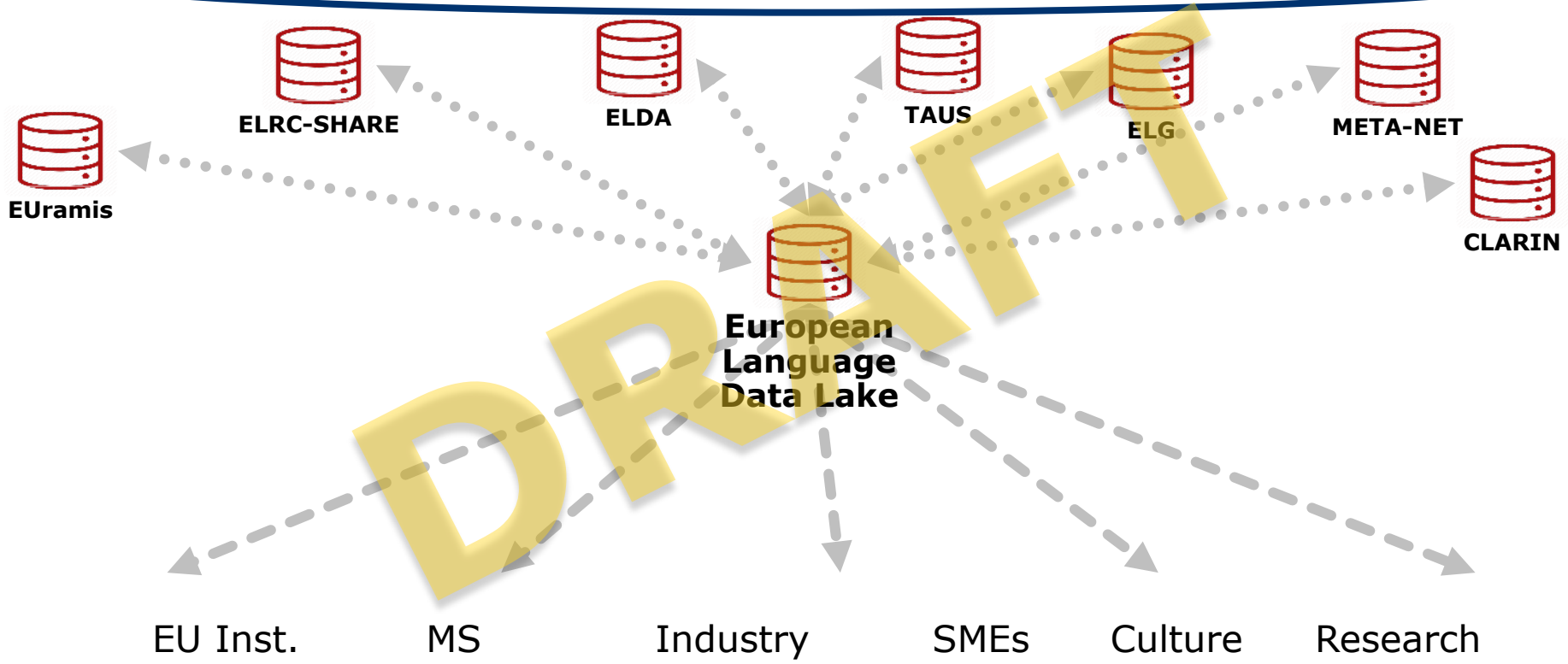
Language Data Lakes



Language Data Space

Governance Body

EU Inst. MS Industry SMEs Culture Research

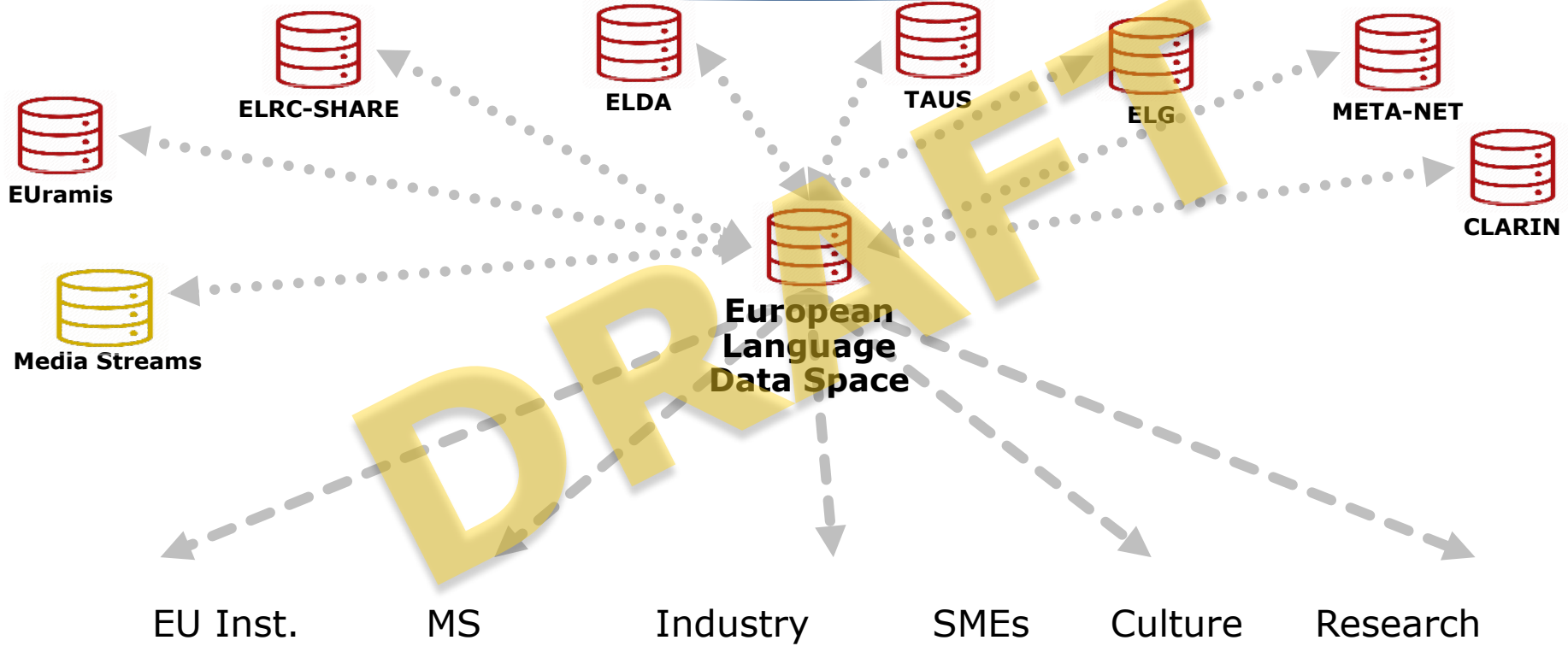


EU Inst. MS Industry SMEs Culture Research

Language Data Space

Governance Body

EU Inst. MS Industry SMEs Culture Research



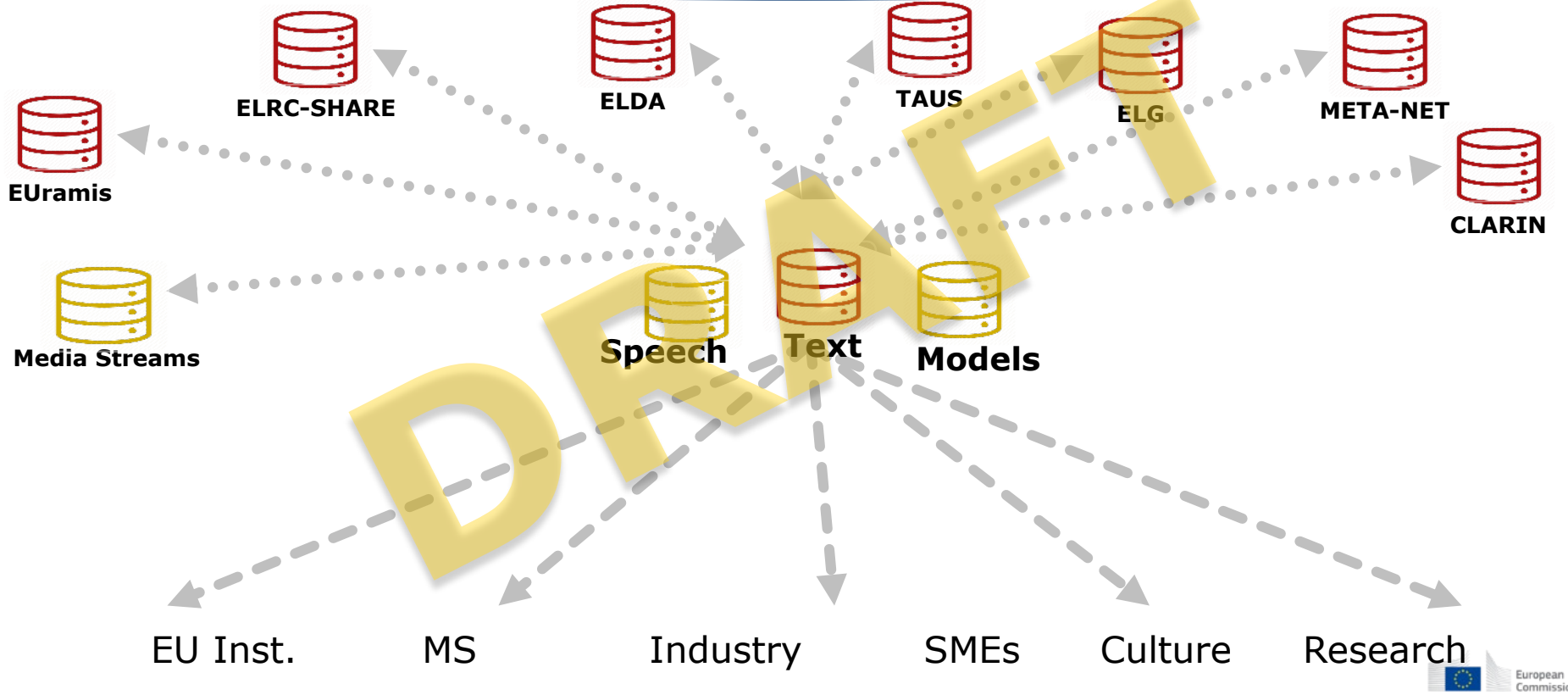
EU Inst. MS Industry SMEs Culture Research

European Commission

Language Data Space

Governance Body

EU Inst. MS Industry SMEs Culture Research

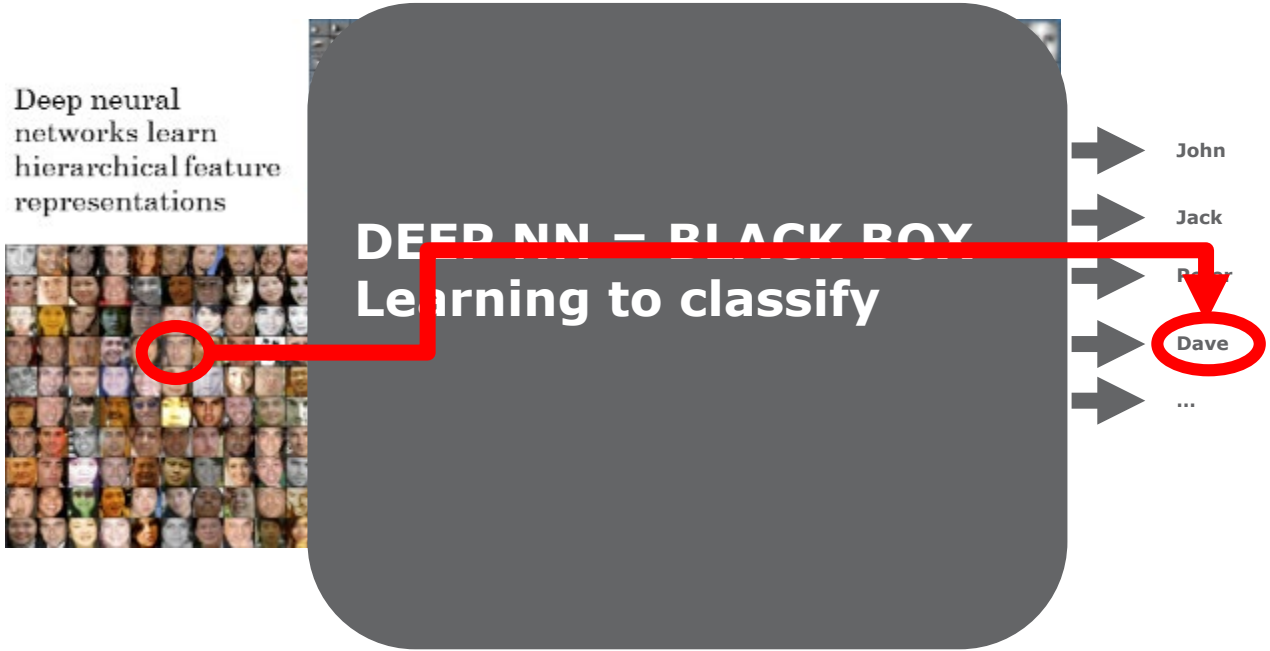


European Commission

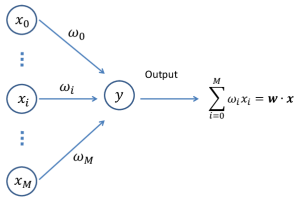
Language Models

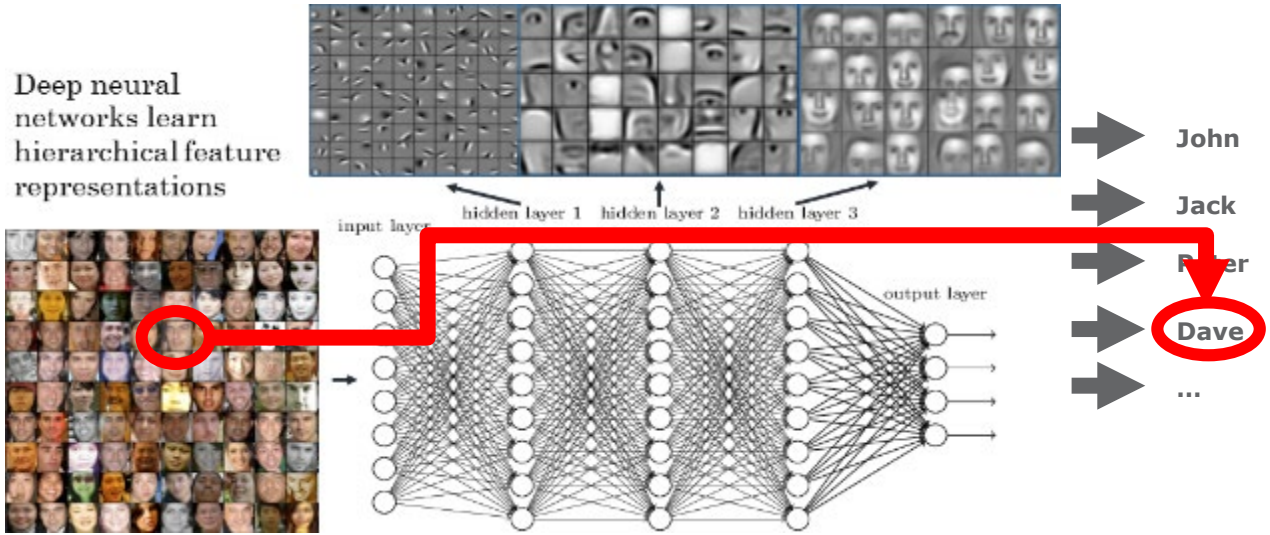
What are these?

The Black Box

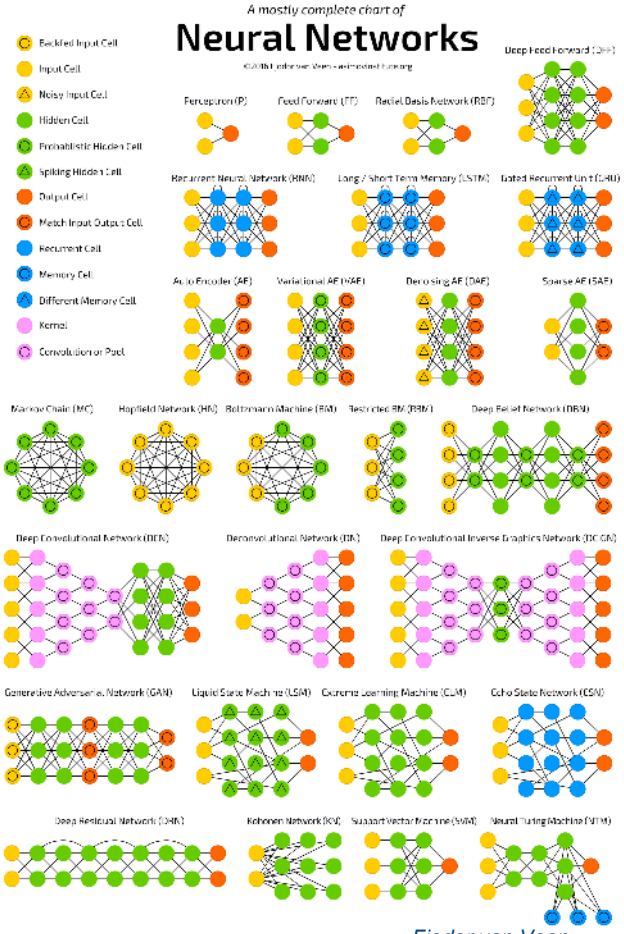


What is deep learning?



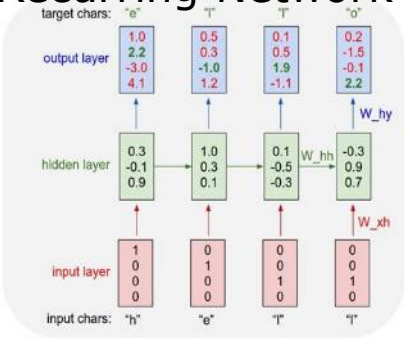


Training algorithms

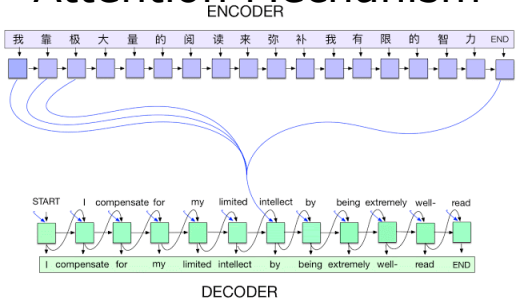


Fjodor van Veen

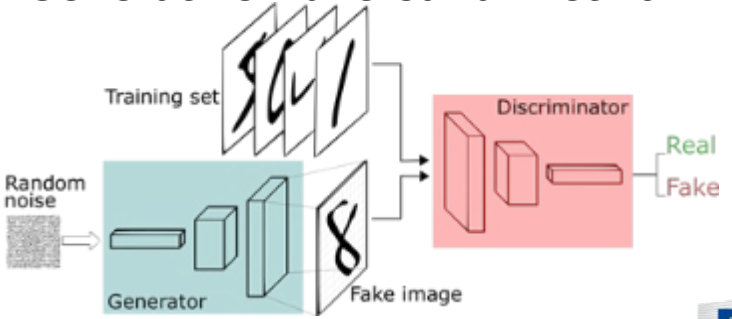
Recurring Network



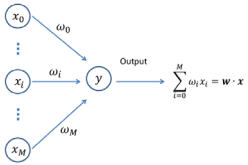
Attention Mechanism

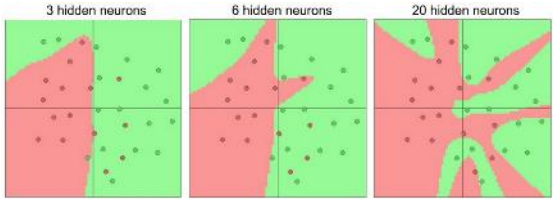


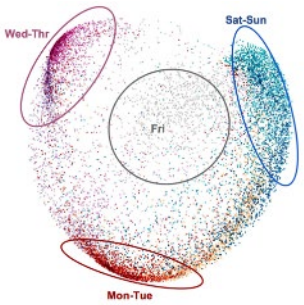
Generative Adversarial Network

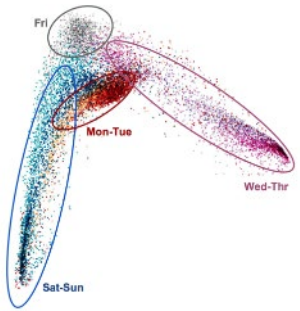


Complexity









(2018)… NMT "Reaching human parity"

DATA: 25 Mio sentence pairs

HPC: 8*7 = 56 Teraflops by 20 hours

= 67200 FLOP = 67 Petaflop

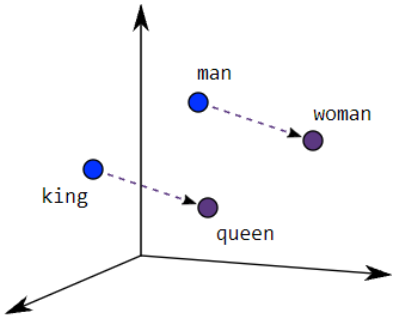
AI: ~= 8 layers x 4096 = 32768 neurons ~=134 Mio Parameters

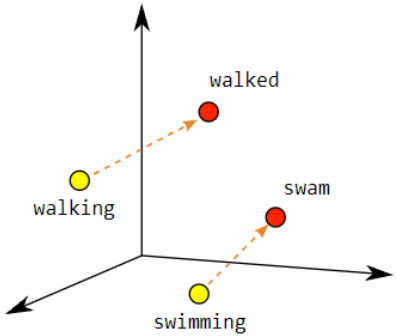
GPT3: 175 Bio Parameters (2020)

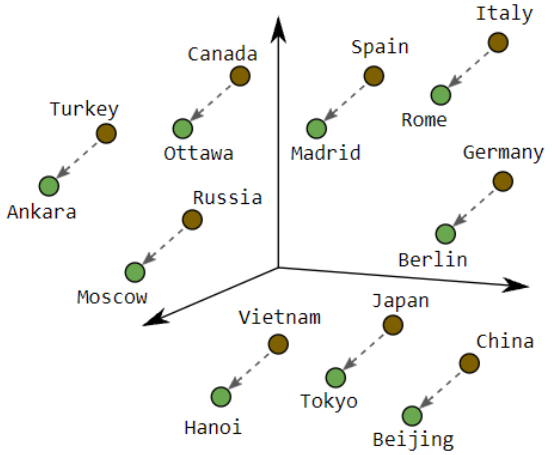
GPT4: 87 Tio Parameters (. . . .)

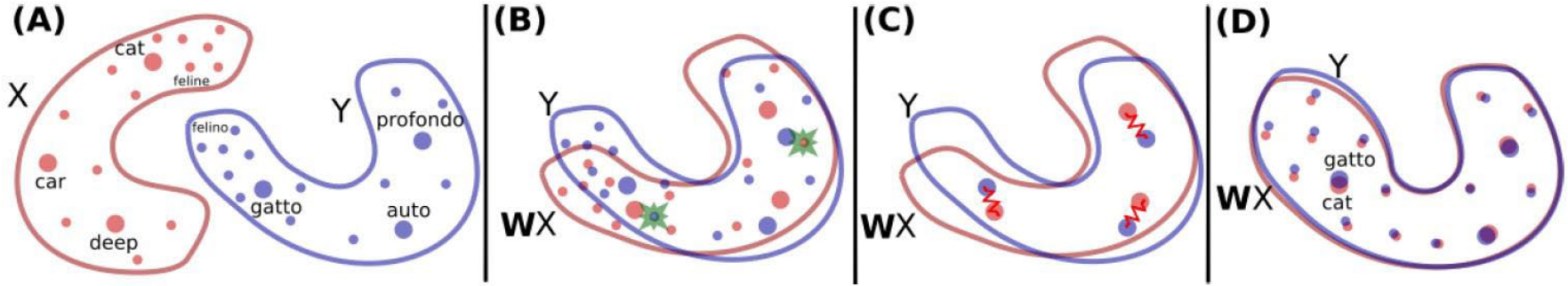
… our brain contains about 86 billion neurons and more than a 100 trillion *(or according to some estimates 1000 trillion) synapses (connections). …*

Semantic projection









(word projected in a 1024 dimensions)

Language Models – a revolution

- They not only analyse the syntax, but they also take into account the **semantic** of the information. (the model regroups semantically similar concepts)

- They can not only analyse, but they are **also generative**, creating text, speech, which opens to a large new set of applications.
- One model can be used for **multiple purposes**. (if you want to move from translation to sentiment analysis, you do not need a full retrain of the models, but only its input & outputs)

Language Models - one example

"Yesterday, we went to the restaurant and we …. a delicious roast beef"

- Previously, with syntax analysis, we could only detect that a word, a verb, was missing. But nothing more.
- With statistical models or Neural Based, the system will tell you that the missing word is "ate" at 94% or "savored" at 5%.

- Language models can also analyse the "style" of the sentence and would change the statistics if it detects that the style of the previous sentences is more refined.
- In addition, language models can "guess" more than one word …. it can continue producing sentences (and make poems or songs)…. While keeping external constrains… of style or tempo.

Language Models – applications 1/2

- Typical applications such as automatic translation, classification of documents, sentiment analysis are established and **progressing in quality**
- but the **generative aspect** of Language models are yet mostly **unchartered territory**.

- In the **manufacturing**: automatic generation of user manual/ maintenance instructions based on the CAD & building process.

- In the **IT domain**, you can generate comments in computer code so that "normal human" can understand the code written by "geeks".

Or reverse…. Encode human-like description in to computer language. Or re-code programme from one computer language to another.

Language Models – applications 2/2

- In the **Media industry**: you can generate hundreds of variation of a same news item. Being fake or not. adapt an article to the style of your newspaper. Automatic summarization, change of style or simplification of articles are tools you can expect to appears within the next few years.

- **Universal language translator** start to appears. Any language "in" and any language "out". Where each language pairs used for training (EN-FR-DE, FR-Latvian) helps the others (DE-Latvian). Really transferring knowledge from one domain to another. This is a fantastic opportunity for the European scene.

- "**Semantic Search**". Has been a kind of holy grail for the last twenty years. But Language models are now already part of the search engines of the major players. Helping contextualising the queries as well as to index the data to be analyse.

- ***ethical aspect*** *or the authenticity of the generated text /voice/ video can be put in question, as in all AI generated content.*

The European Language Data Space

SMEs, Telcos, legal & financial sectors, insurance companies, manufacturing, PRs industries, media companies, public administration, other Data Spaces, … and customers

anonymization, subtitling, Q&A, Text summarisation, speech recognition and translation, conversational systems …

crisis management, fintech, healthcare, advertising industry, media fighting disinformation …

turnkey solutions for multilingual websites, multilingual search engine, Q&A handling, messaging …

customised services

Business intelligence

access to internal and international markets.

Digital transformation

Economic recovery

Modeling

Data analytics and sentiment analysis

Multilinguality

data lakes

(META-SHARE, CLARIN, Euramis, ELRC-SHARE, TAUS, ELRA …)

text

speech

data streams

CLOUD

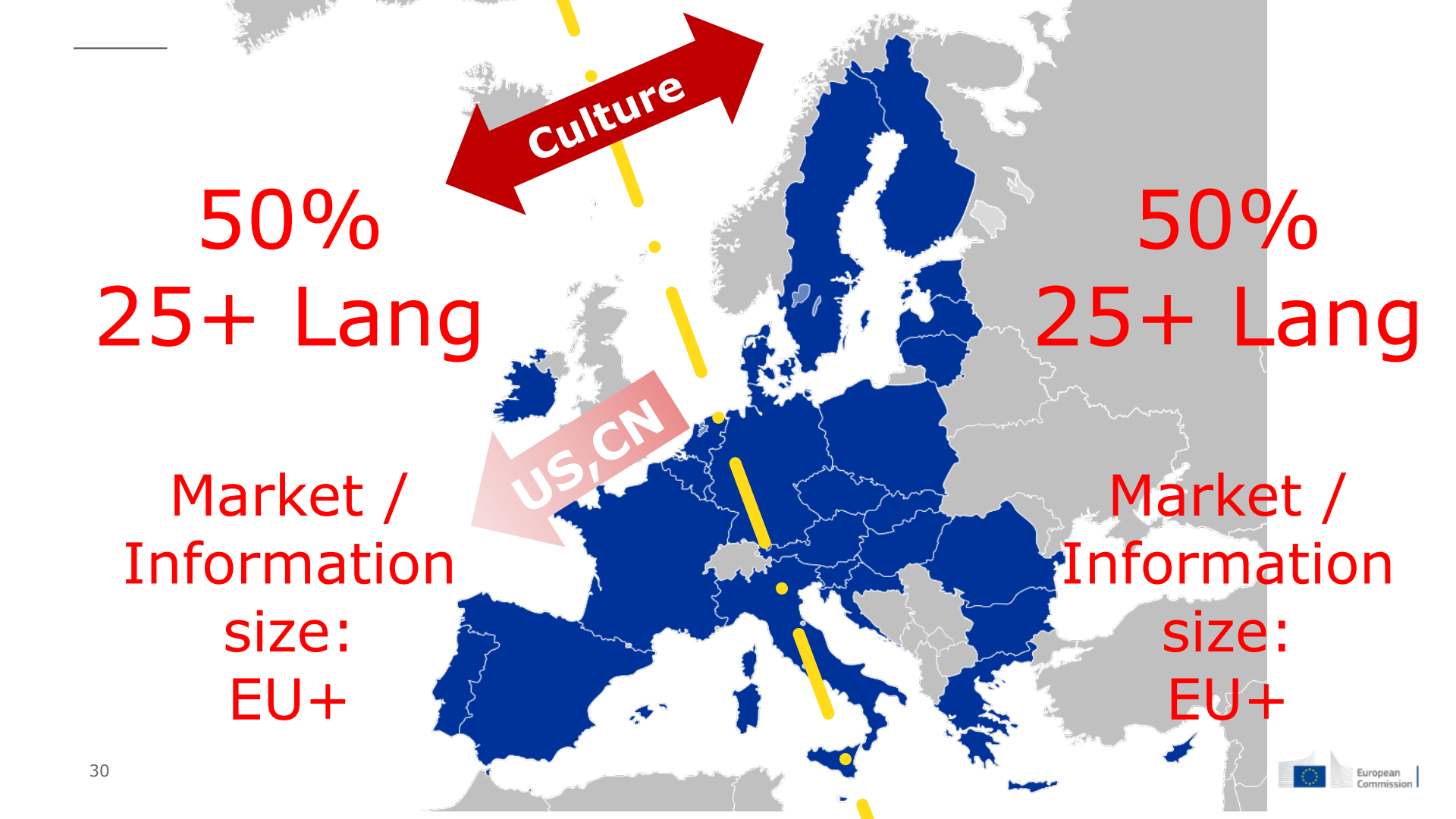
HPC

Clearing house

AI

Technological autonomy

Existing repositories and aggregators, EU and MS administrations, media companies, Telcos, publishers, other Data Spaces



50%
25+ Lang

Market / Information size:
EU+

50%
25+ Lang

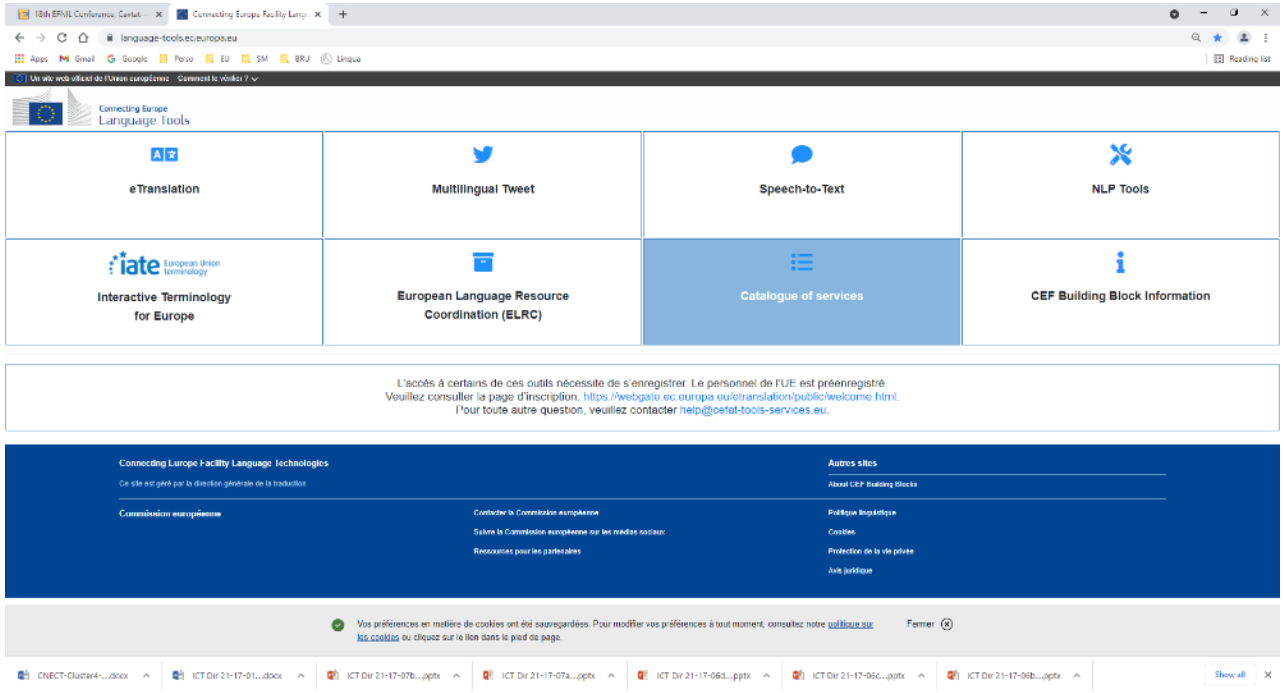
Market / Information size:
EU+

European Commission

European Commission

Services

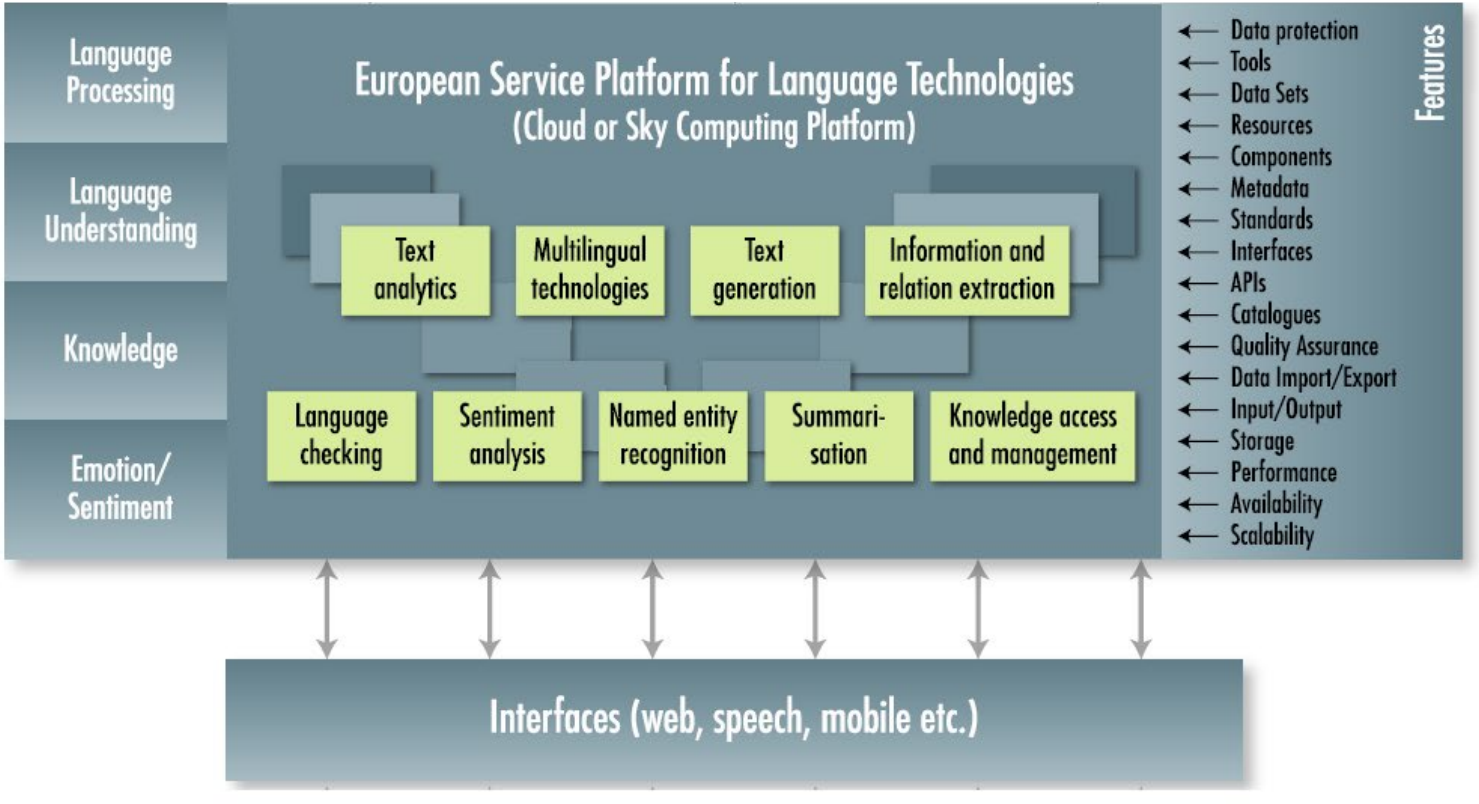
eTranslation -> eLangTech



- Translation (30+)
- Name Entity recognition (24)
- Classification (24)
- Speech Transcription
- … Anonymisation

https://language-tools.ec.europa.eu/

ELG – European Language Grid



From: META-NET SRA (2013)

Links & contacts

PHILIPPE GELIN
European Commission
DG CONNECT - Communications Network, Content and Technology
Dir G: Data
Unit G3: Accessibility, **Multilingualism** and Safer Internet

Philippe.Gelin@ec.europa.eu