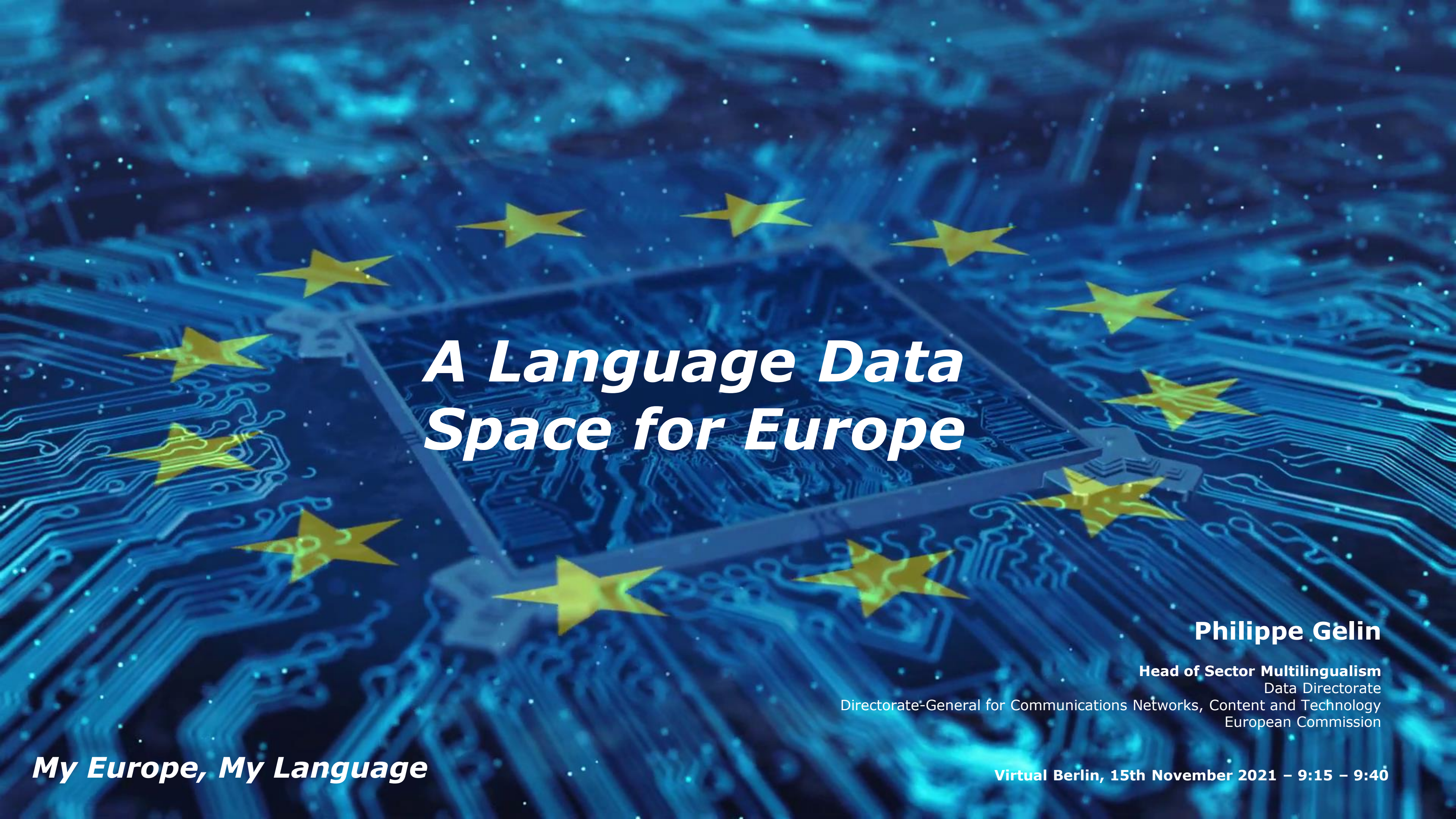


A Language Data Space for Europe

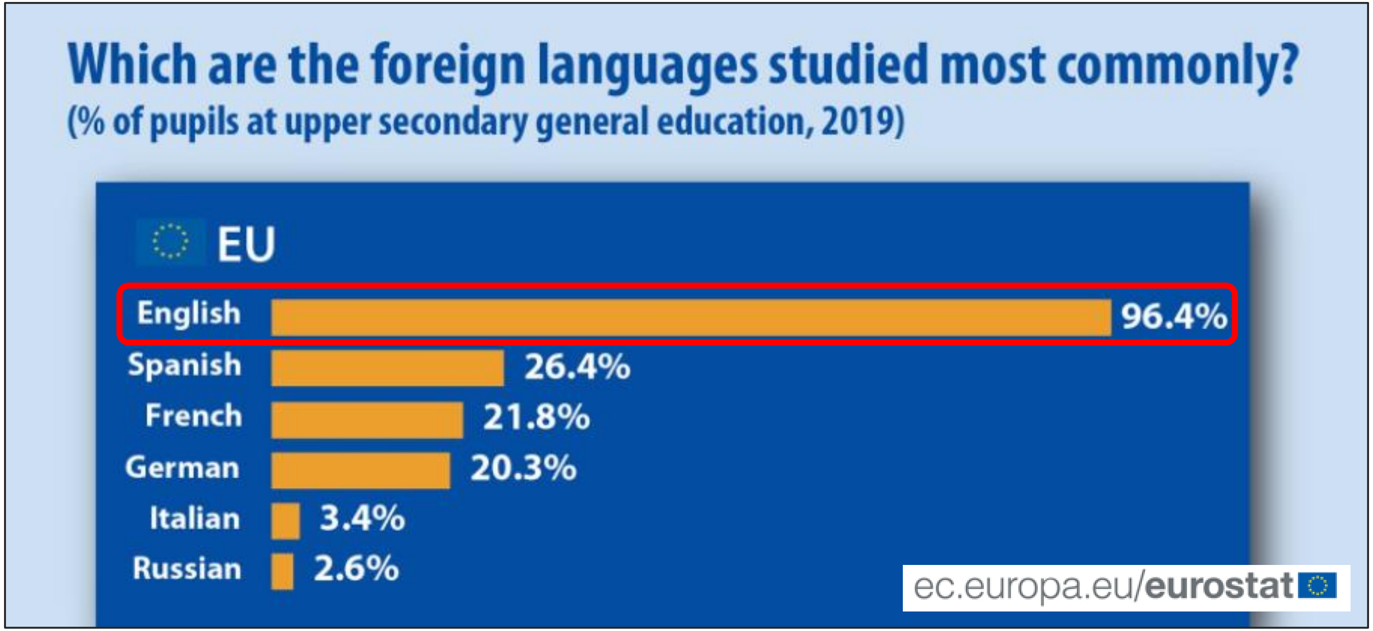
Philippe Gelin

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

My Europe, My Language

Virtual Berlin, 15th November 2021 – 9:15 – 9:40

DIGITAL Background Information:
Language Learning



Which are the foreign languages studied most commonly?

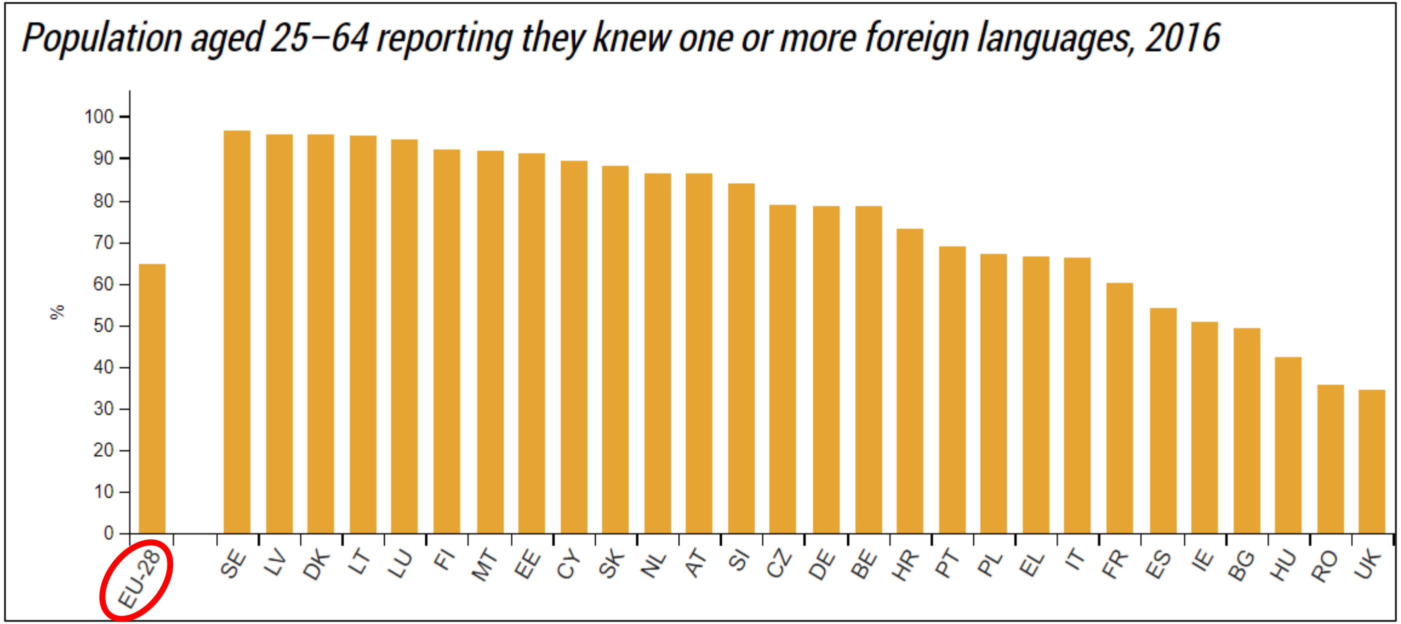
(% of pupils at upper secondary general education, 2019)

In 2019, **96 %** of pupils in upper secondary general education in the EU were **learning English as a foreign language**.

[Education and training in the EU - facts and figures]

[Which are the foreign languages studied most commonly]

DIGITAL Background Information:
Foreign Language

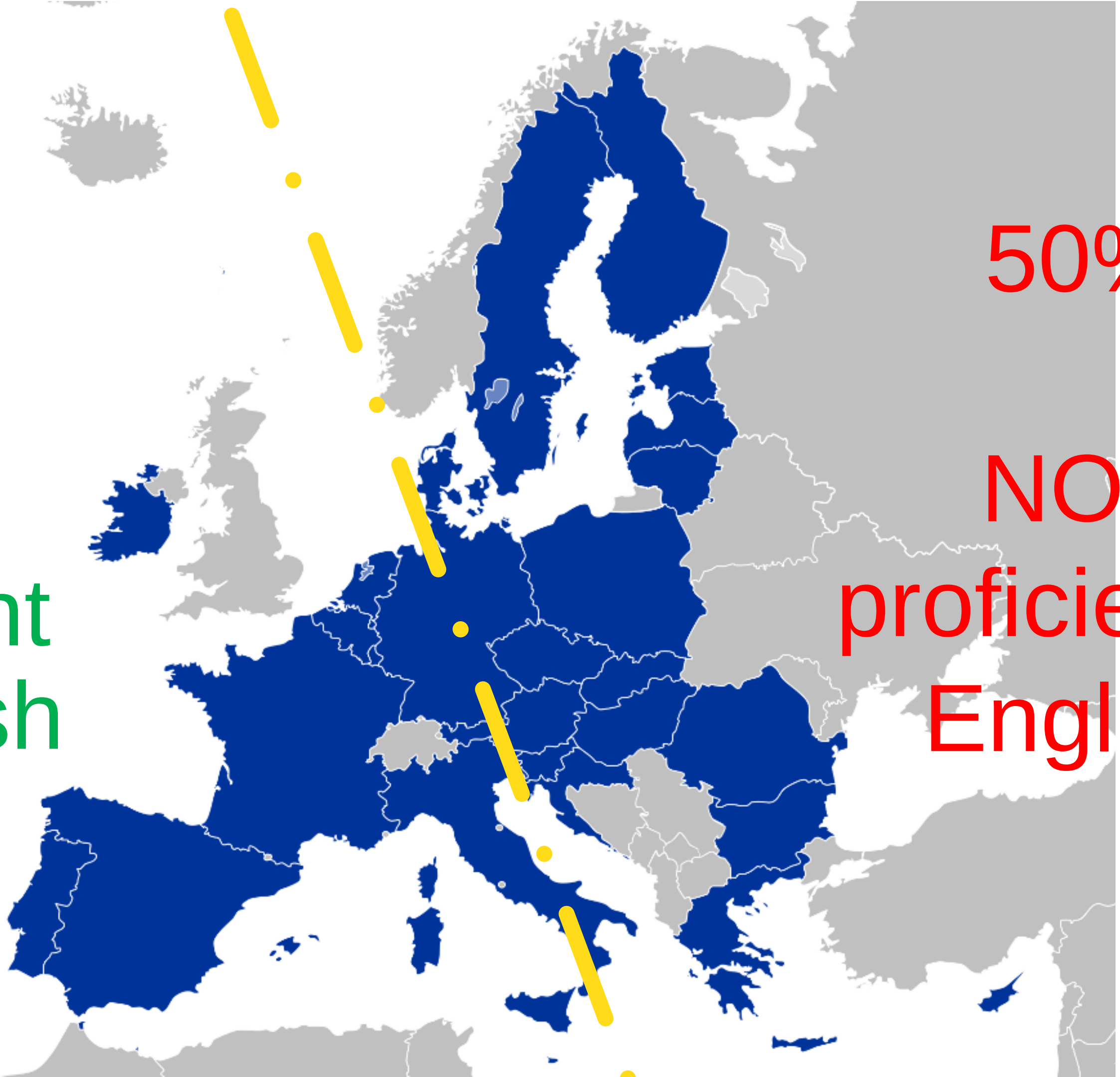


Population aged 25–64 reporting they knew one or more foreign languages, 2016

Foreign language skills statistics

50%

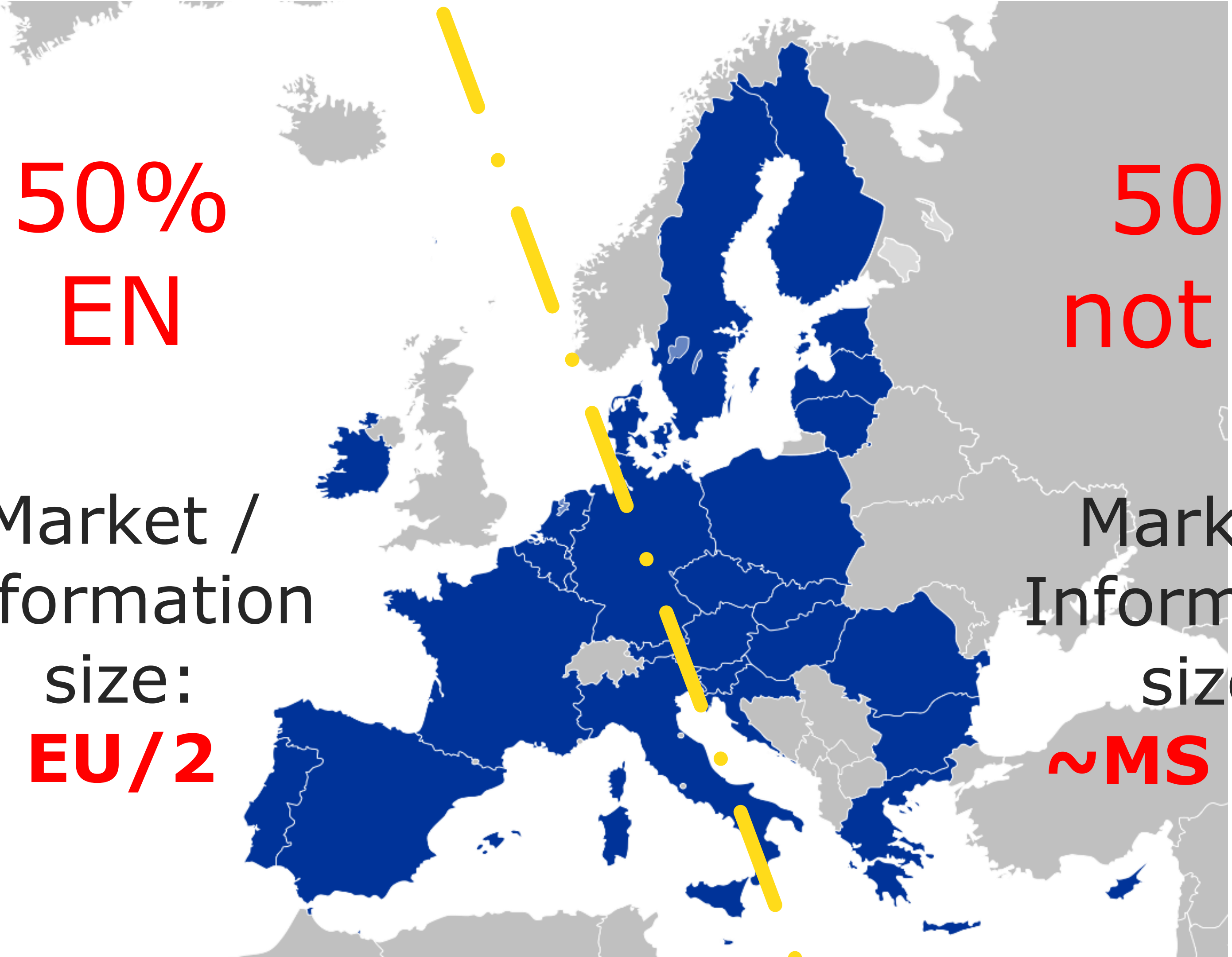
proficient in English



50%

NOT proficient in English

European Commission



50%
EN

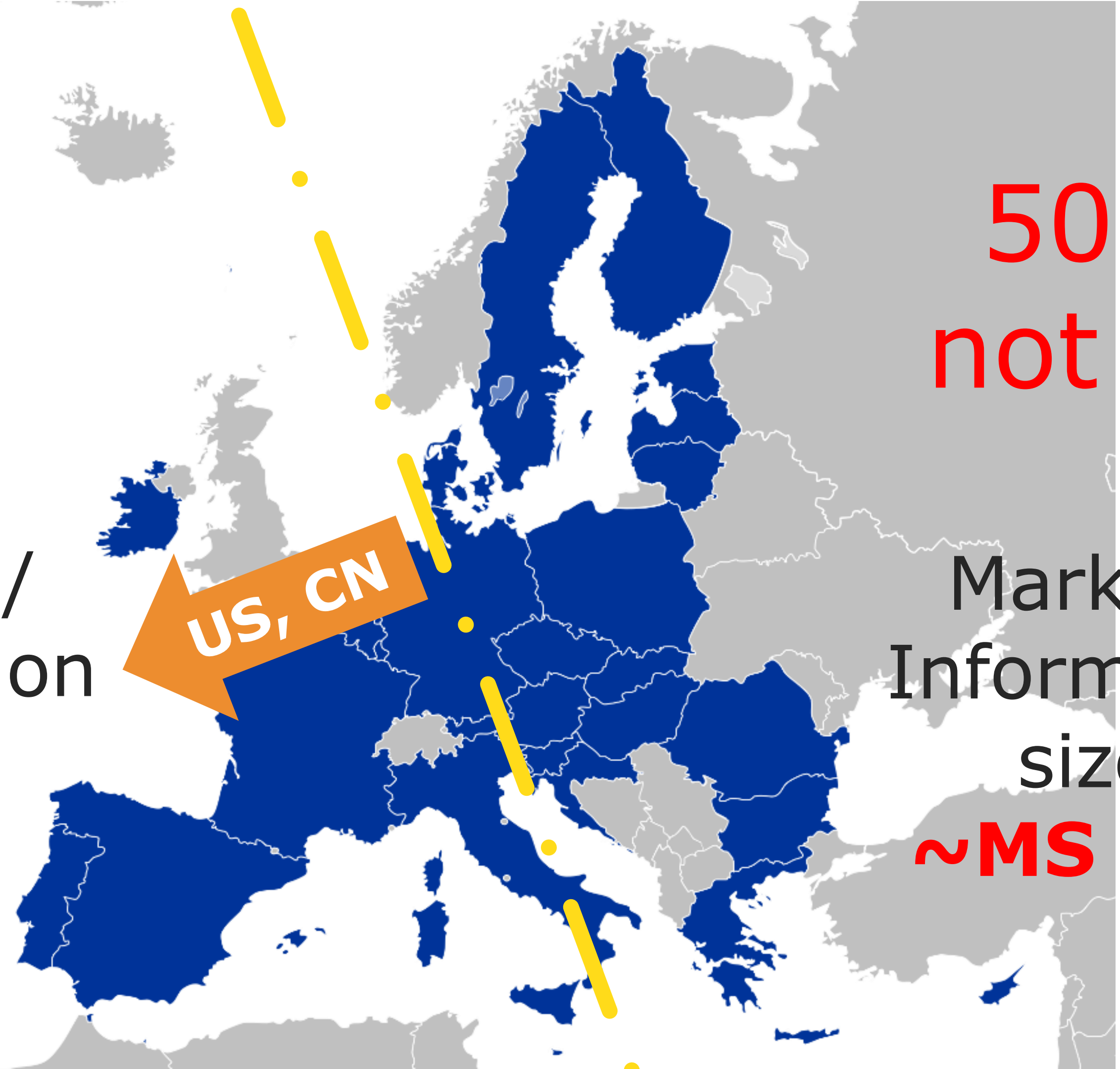
Market /
Information
size:
EU/2

50%
not EN

Market /
Information
size:
~MS only

50%
EN

Market /
Information
size:
EU/2



50%
not EN

Market /
Information
size:
~MS only

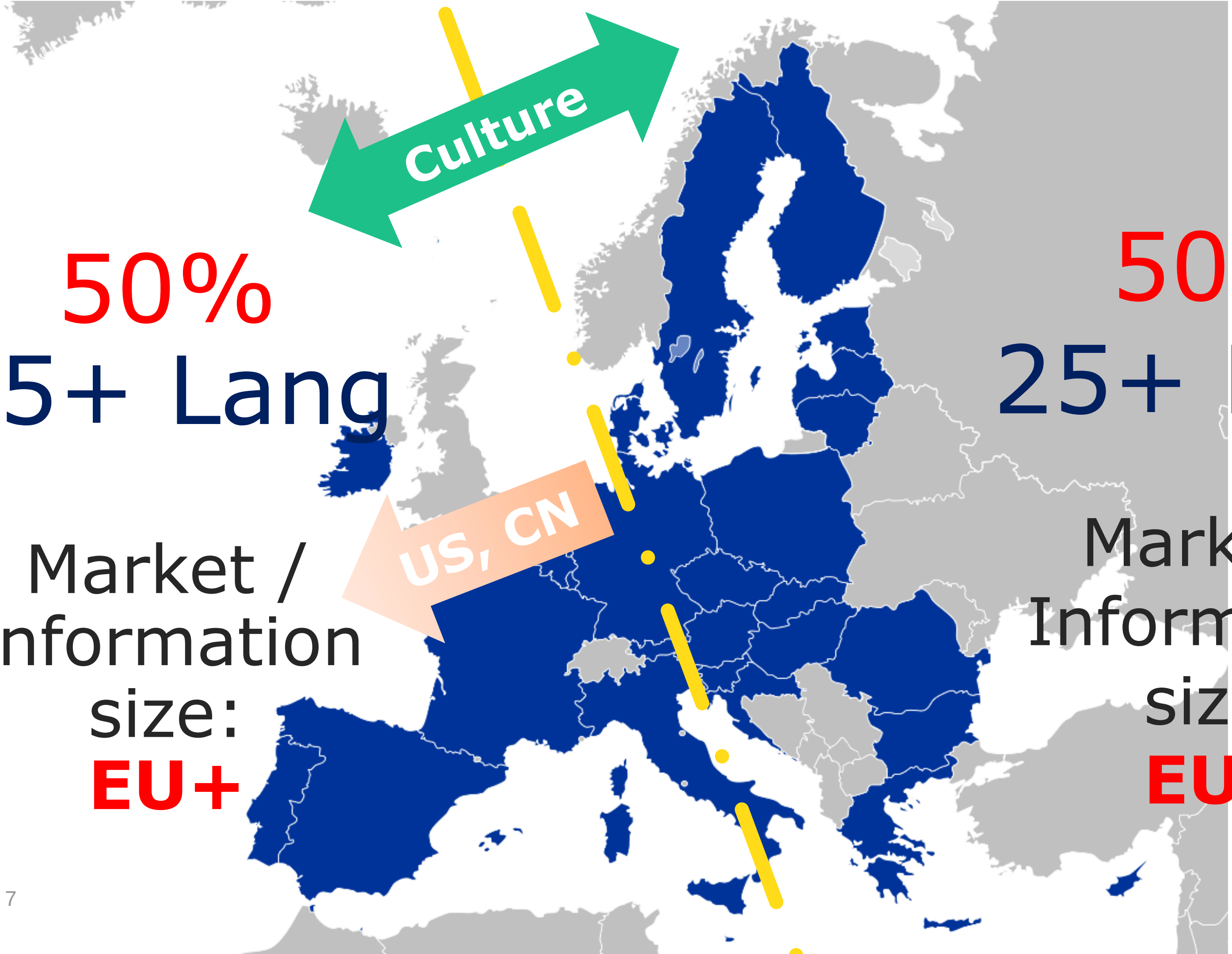
European
Commission

50%

25+ Lang

Market / Information size:

EU+



50%

25+ Lang

Market / Information size:

EU+

DIGITAL in a Nutshell:
The Programme *at a Glance*

Digital Europe Programme

VISION

Digital Decade and digital targets for 2030

FOCUS

Building the **strategic digital capacities** of the Union and facilitating the wide **deployment** of digital technologies

GOAL

Bridging the gap between digital technology research and market deployment

DIGITAL in a Nutshell:
The *Synergies*

DIGITAL in a Nutshell:
The Big Tickets

DIGITAL EUROPE PROGRAMME

HIGH PERFORMANCE COMPUTING

ADVANCED DIGITAL SKILLS

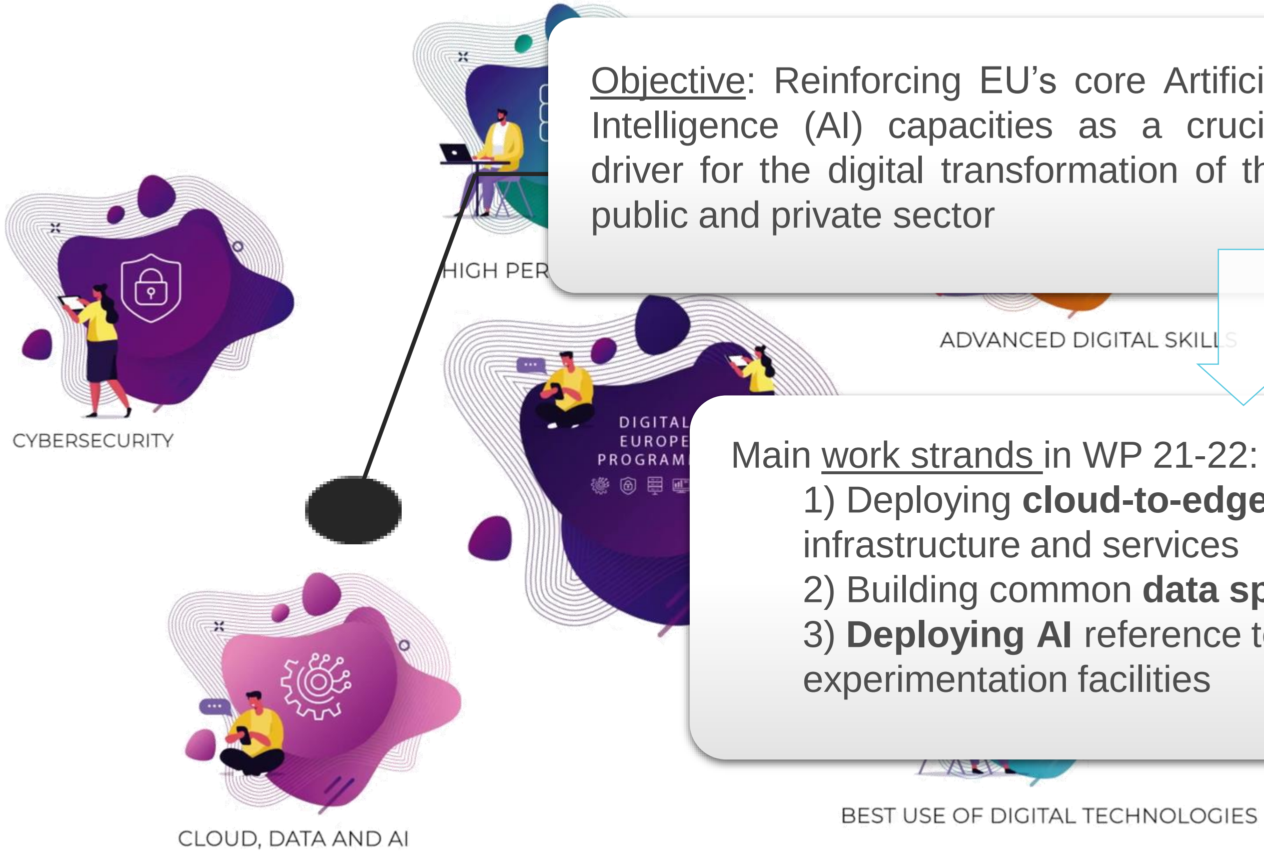
CYBERSECURITY

CLOUD, DATA AND AI

BEST USE OF DIGITAL TECHNOLOGIES

European Commission

DIGITAL in a Nutshell:
The Cloud, Data and AI



Objective: Reinforcing EU’s core Artificial Intelligence (AI) capacities as a crucial driver for the digital transformation of the public and private sector

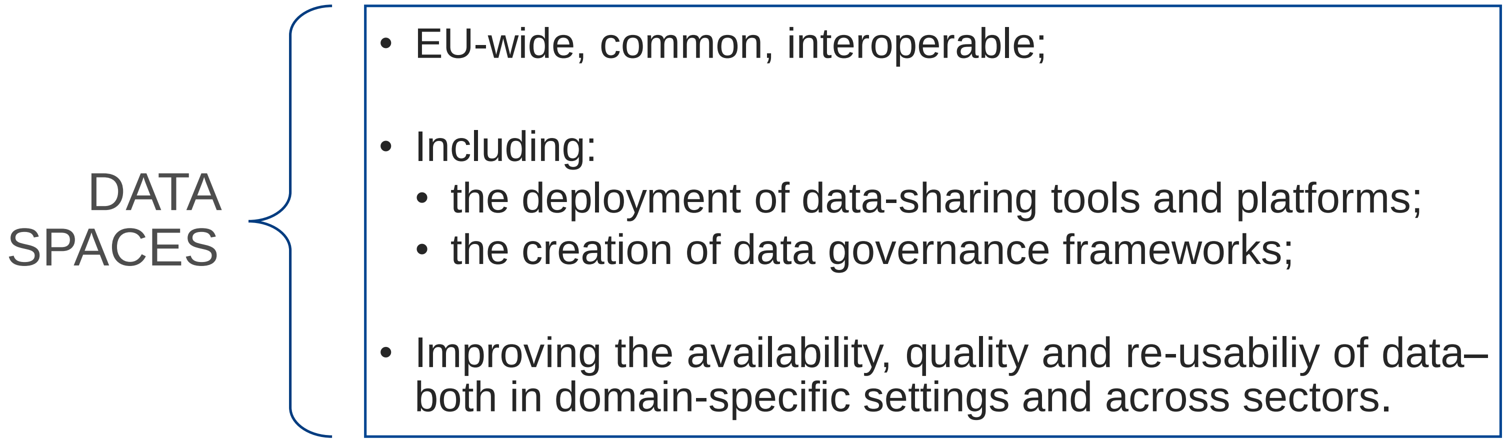
Main work strands in WP 21-22:

1) Deploying **cloud-to-edge** infrastructure and services
2) Building common **data spaces**
3) **Deploying AI** reference testing and experimentation facilities

Data Spaces

DATA SPACES Background Information:
European Data Strategy

High Impact Project: Data spaces should foster an ecosystem (of companies, civil society and individuals) creating new products and services based on more accessible data.

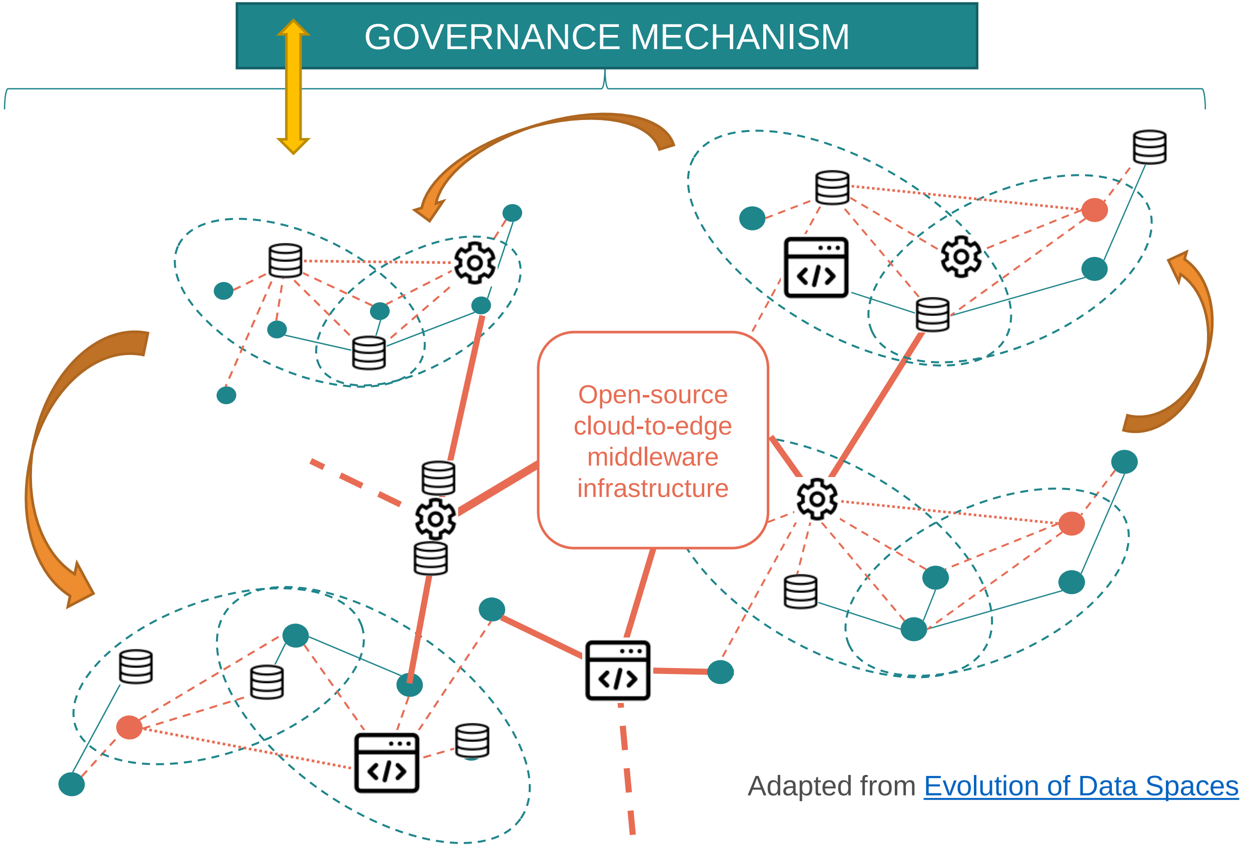


DATA SPACES

- EU-wide, common, interoperable;
- Including:
  - the deployment of data-sharing tools and platforms;
  - the creation of data governance frameworks;
- Improving the availability, quality and re-usabiliy of data– both in domain-specific settings and across sectors.

A European strategy for data

DATA SPACES in a Nutshell (1):



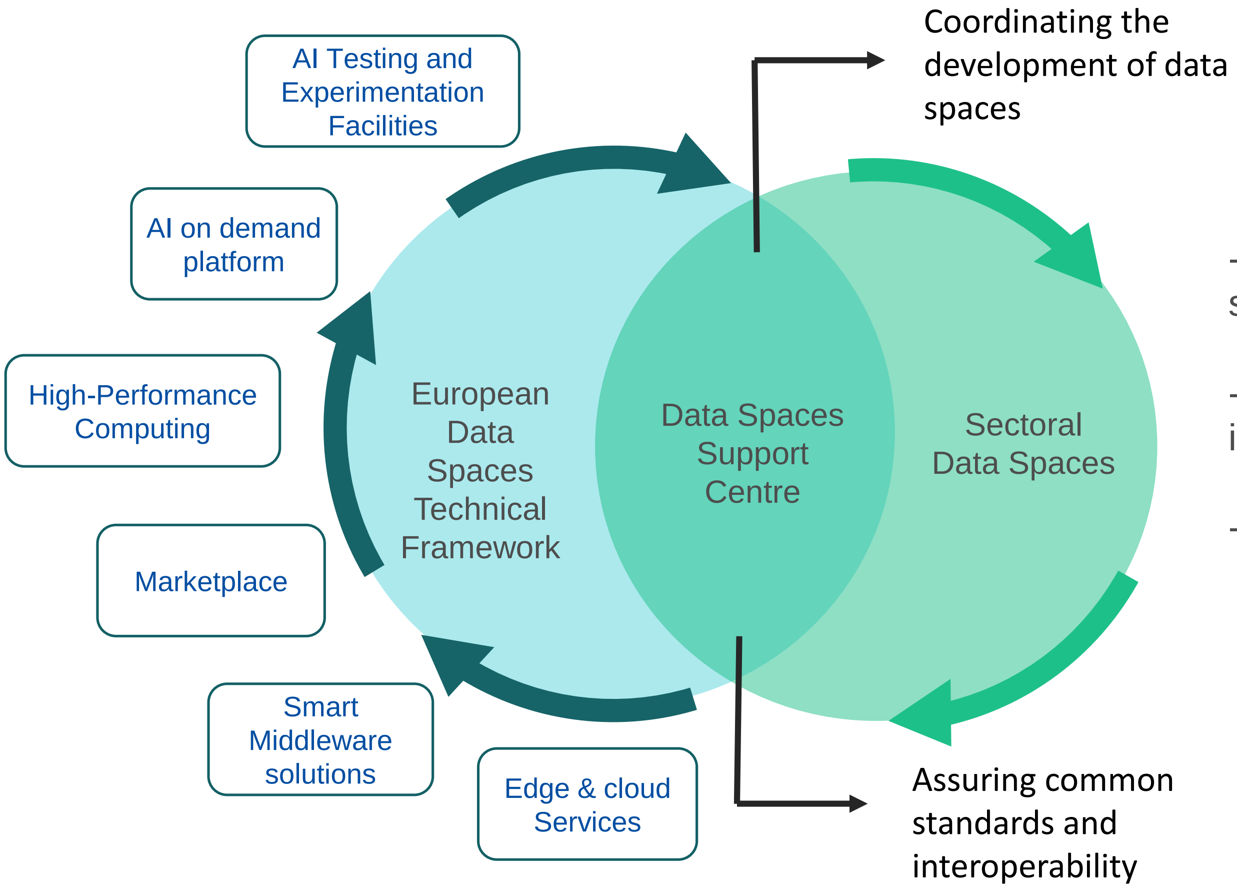
Data Spaces bring together

1. **data**
2. data **infra-structures**
3. **governance** structures

to facilitate pooling, accessing, processing, using and sharing data.

Adapted from Evolution of Data Spaces

DATA SPACES in a Nutshell (2):



AI Testing and Experimentation Facilities

AI on demand platform

High-Performance Computing

Marketplace

Smart Middleware solutions

Edge & cloud Services

European Data Spaces Technical Framework

Data Spaces Support Centre

Sectoral Data Spaces

Coordinating the development of data spaces

Assuring common standards and interoperability

- Bring together stakeholders.

- Identify the data the infrastructure needs.

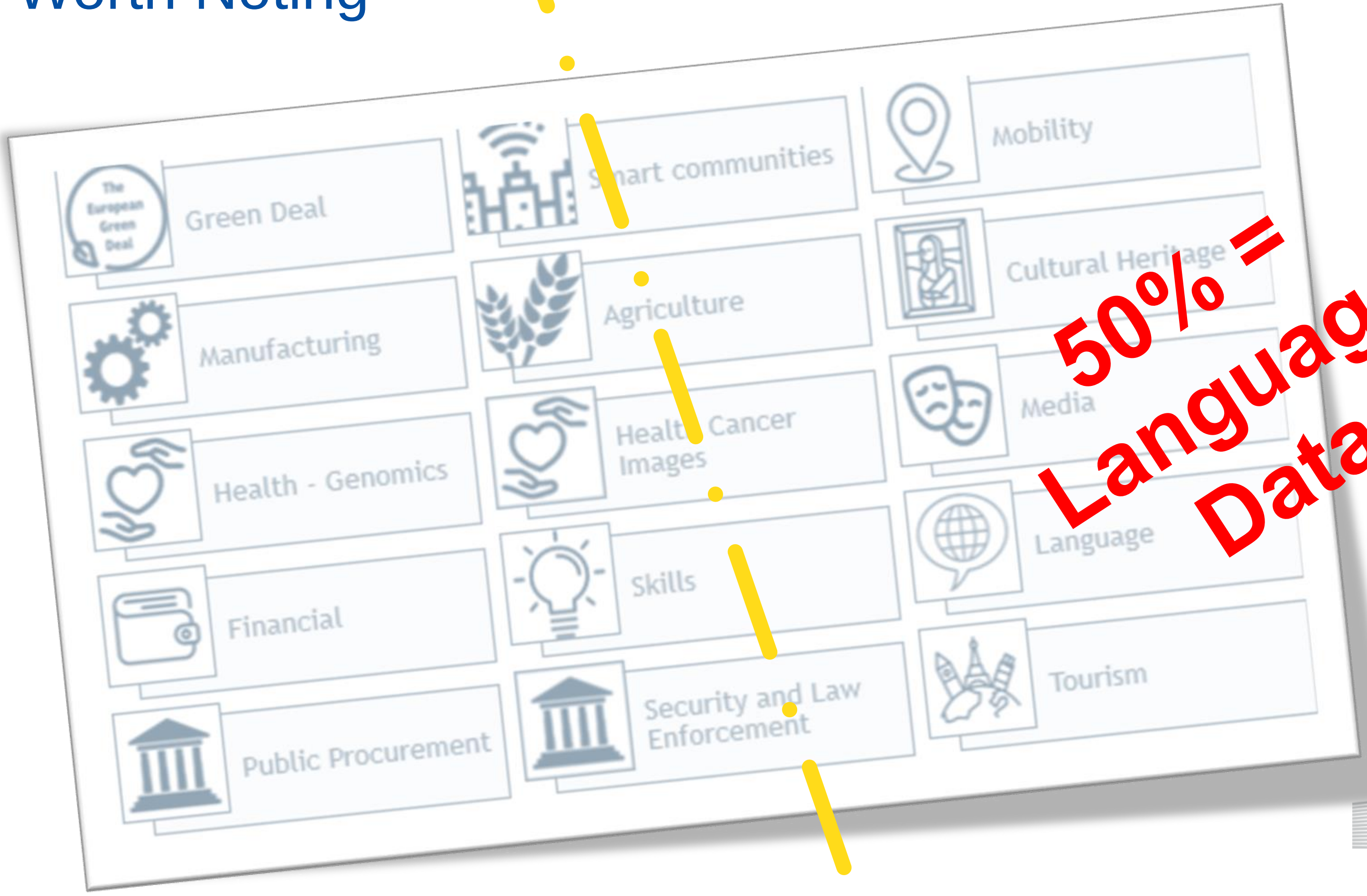
- Build the data spaces

Data Spaces in DIGITAL:
An Overview of all DSs in WP 21-22

Green Deal	Smart communities	Mobility
Manufacturing	Agriculture	Cultural Heritage
Health - Genomics	Health Cancer Images	Media
Financial	Skills	Language
Public Procurement	Security and Law Enforcement	Tourism

A Language Data Space for Europe

Worth Noting



The European Green Deal

Green Deal

Manufacturing

Health - Genomics

Financial

Public Procurement

Smart communities

Agriculture

Health Cancer Images

Skills

Security and Law Enforcement

Mobility

Cultural Heritage

Media

Language

Tourism

50% = Language Data

Reasoning

Multimodal Data

- Text, audio and video are a large part of the data produced every day. *However,* …

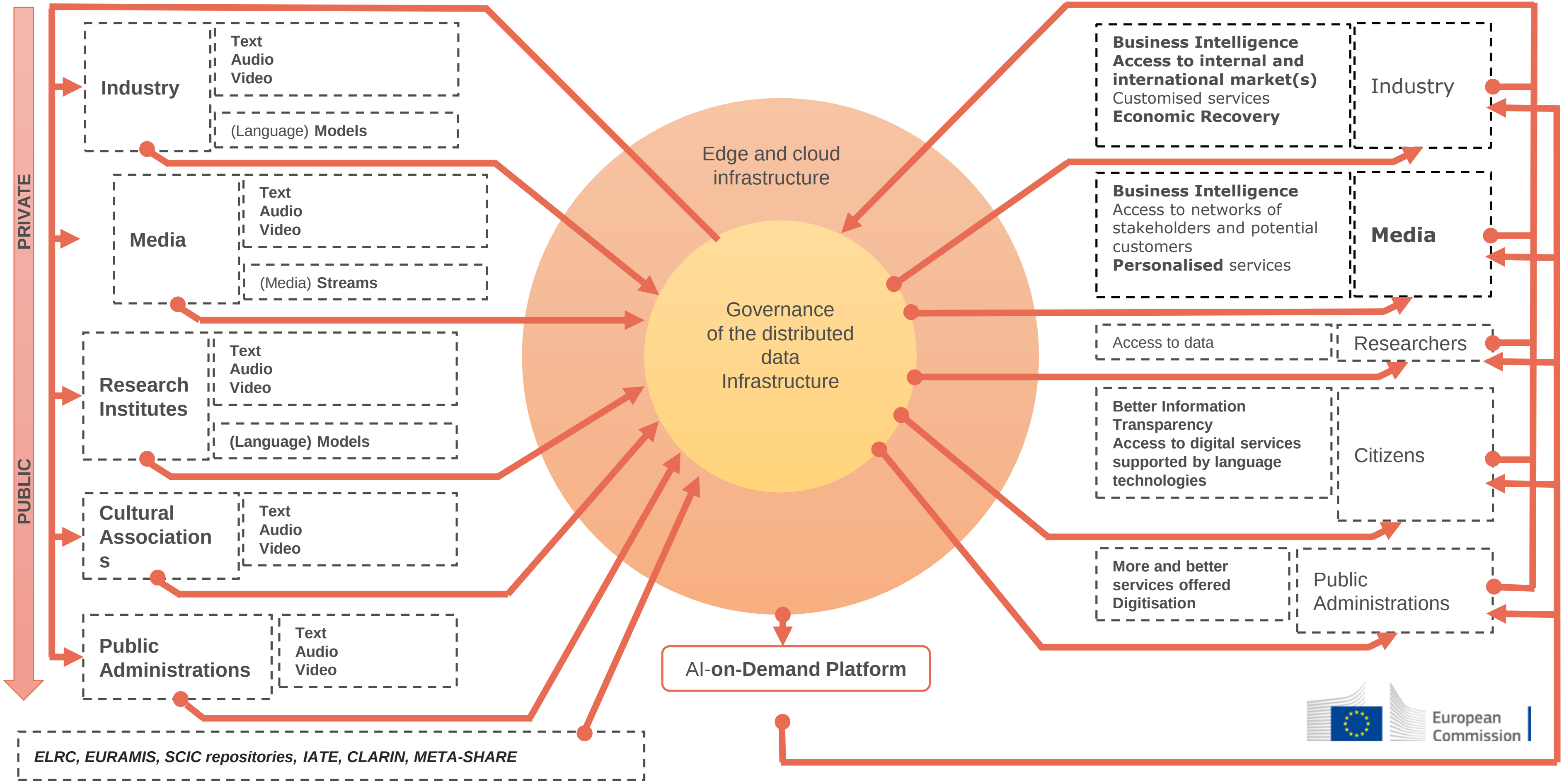
Data Ecosystem

- Such data need to be aggregated and organized for AI-based language services to process and extract relevant information. *Hence*, …

Language Data Space

- for the collection, creation, sharing and reuse of multimodal language data &
- as a support to the deployment of large multimodal language models + AI language technologies services through the AI platform.

LANGUAGE DATA SPACE in a Nutshell



PRIVATE

PUBLIC

Industry

Text
Audio
Video

(Language) Models

Media

Text
Audio
Video

(Media) Streams

Research Institutes

Text
Audio
Video

(Language) Models

Cultural Association
s

Text
Audio
Video

Public Administrations

Text
Audio
Video

ELRC, EURAMIS, SCIC repositories, IATE, CLARIN, META-SHARE

Edge and cloud infrastructure

Governance of the distributed data Infrastructure

AI-**on-Demand Platform**

Business Intelligence
Access to internal and international market(s)
Customised services
Economic Recovery

Industry

Business Intelligence
Access to networks of stakeholders and potential customers
Personalised services

Media

Access to data

Researchers

Better Information Transparency
Access to digital services supported by language technologies

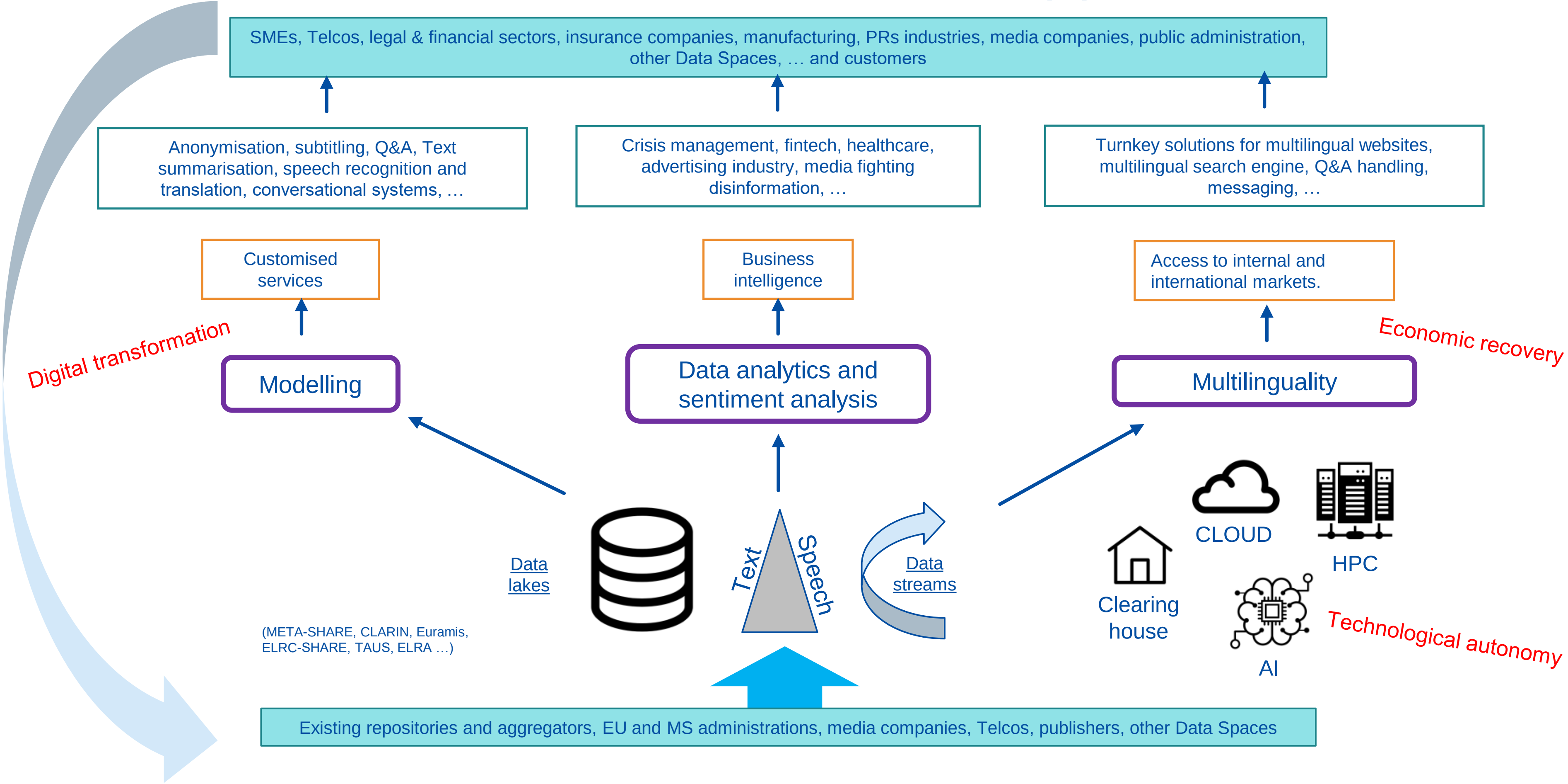
Citizens

More and better services offered
Digitisation

Public Administrations

European Commission

LANGUAGE DATA SPACE in a Nutshell (2)



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

Anonymisation, 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

Modelling

Data analytics and sentiment analysis

Multilinguality

Economic recovery

Data lakes

Text

Speech

Data streams

CLOUD

HPC

Clearing house

AI

Technological autonomy

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

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

Language Data Space

- Collection, creation, sharing and reuse of **multimodal language data**.
- This will support the deployment of **large multimodal language models** and a wide range of **AI language technologies services** to be offered through the AI platform.

Two Strands:

- Establish an institutional Centre of Excellence for Language Technologies (**CELT**)
- **Deployment** of the language data space

1st Work Strand

- Establish an institutional Centre of Excellence for Language Technologies (CELT) **to coordinate across the Member States** the creation and collection of multimodal language data and models.

This will make use of existing EU initiatives of language data collections such as ELRC, EURAMIS, SCIC repositories, IATE, CLARIN, META-SHARE.

1st Work Strand - Detail

- Develop in close collaboration with the Member States a multi-stakeholder **data and services governance scheme**, bringing together large industrial entities, public stakeholders and small-and-mid-size enterprise stakeholders.

- **Identify and bring together** existing European stakeholders and initiatives.

- Elaborate a blueprint for the **language data ecosystem** based on existing EU legislation and language data policies to build innovative business models across all stakeholders.

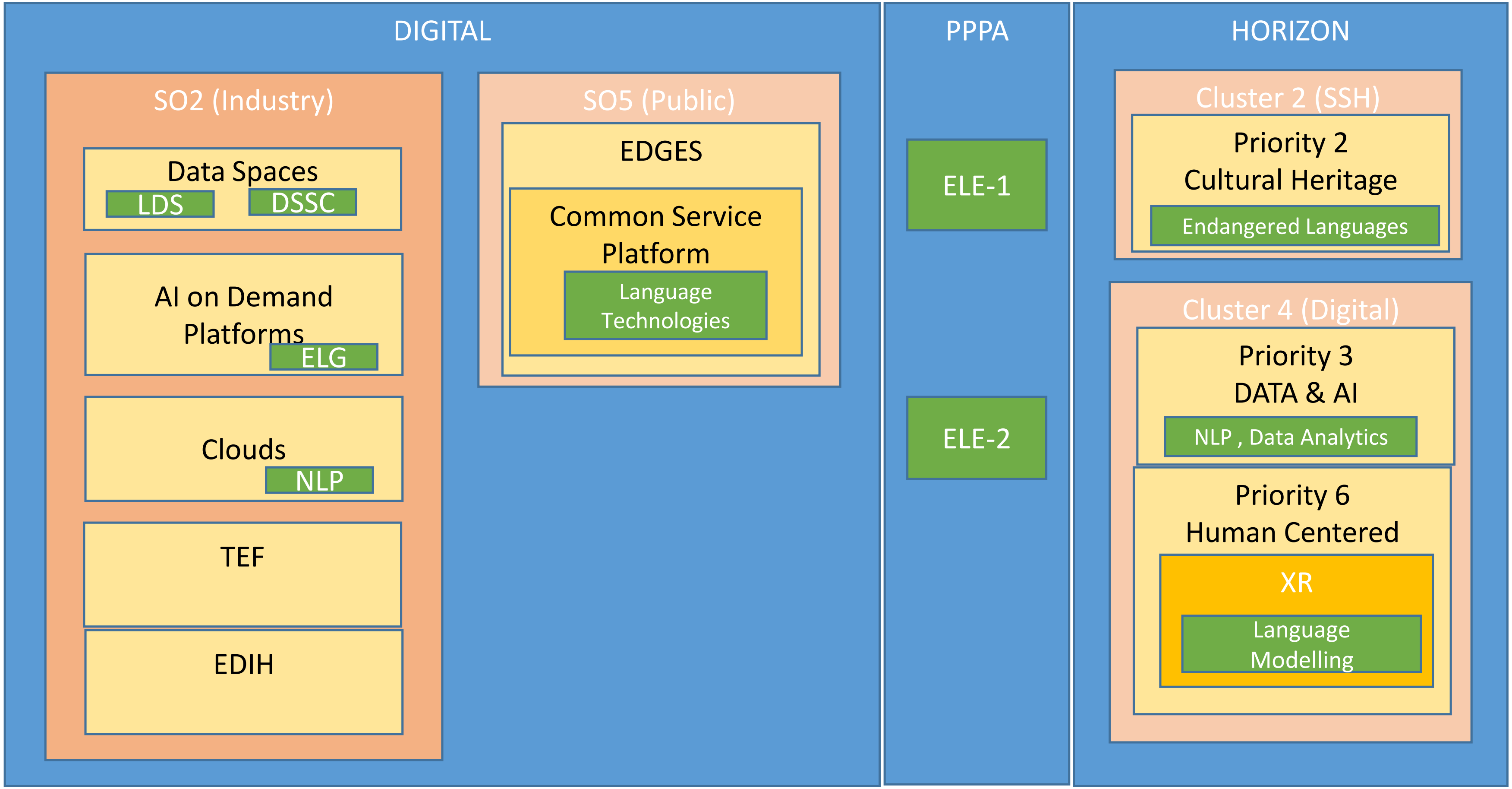
- Identify the **large multimodal language models** to be deployed.

- **Identify datasets and data streams** (public, private, citizen-collected….) relevant to the creation of large language models to be brought into conformity with the new blueprint standards and principles

- Establish a **detailed roadmap** on how to deploy the language data space.

2nd work strand - Deployment

- deploying the necessary **infrastructure** for the collection and sharing of multimodal language data and models;
- implementing the **data & services governance**, business models and strategies as defined by CELT;
- supporting the increased uptake and usage of the language data space and its derived language technologies services among European **private and public sectors**, including various DIGITAL platforms and endeavours/initiatives, also through **piloting and deployment projects**;
- establishing Member States and industrial collaboration to bring the set of identified language datasets and data streams into **conformity** with the new blueprint standards and principles;
- establishing Member States and industrial collaboration to **create and deploy the identified large multimodal language models**;
- deploying advanced **AI-based language processing services and tools** to be made available through the AI-on-demand platform.



DIGITAL

SO2 (Industry)

Data Spaces

LDS

DSSC

AI on Demand Platforms

ELG

Clouds

NLP

TEF

EDIH

SO5 (Public)

EDGES

Common Service Platform

Language Technologies

PPPA

ELE-1

ELE-2

HORIZON

Cluster 2 (SSH)

Priority 2
Cultural Heritage

Endangered Languages

Cluster 4 (Digital)

Priority 3
DATA & AI

NLP , Data Analytics

Priority 6
Human Centered

XR

Language Modelling

Links & contacts

PHILIPPE GELIN
European Commission

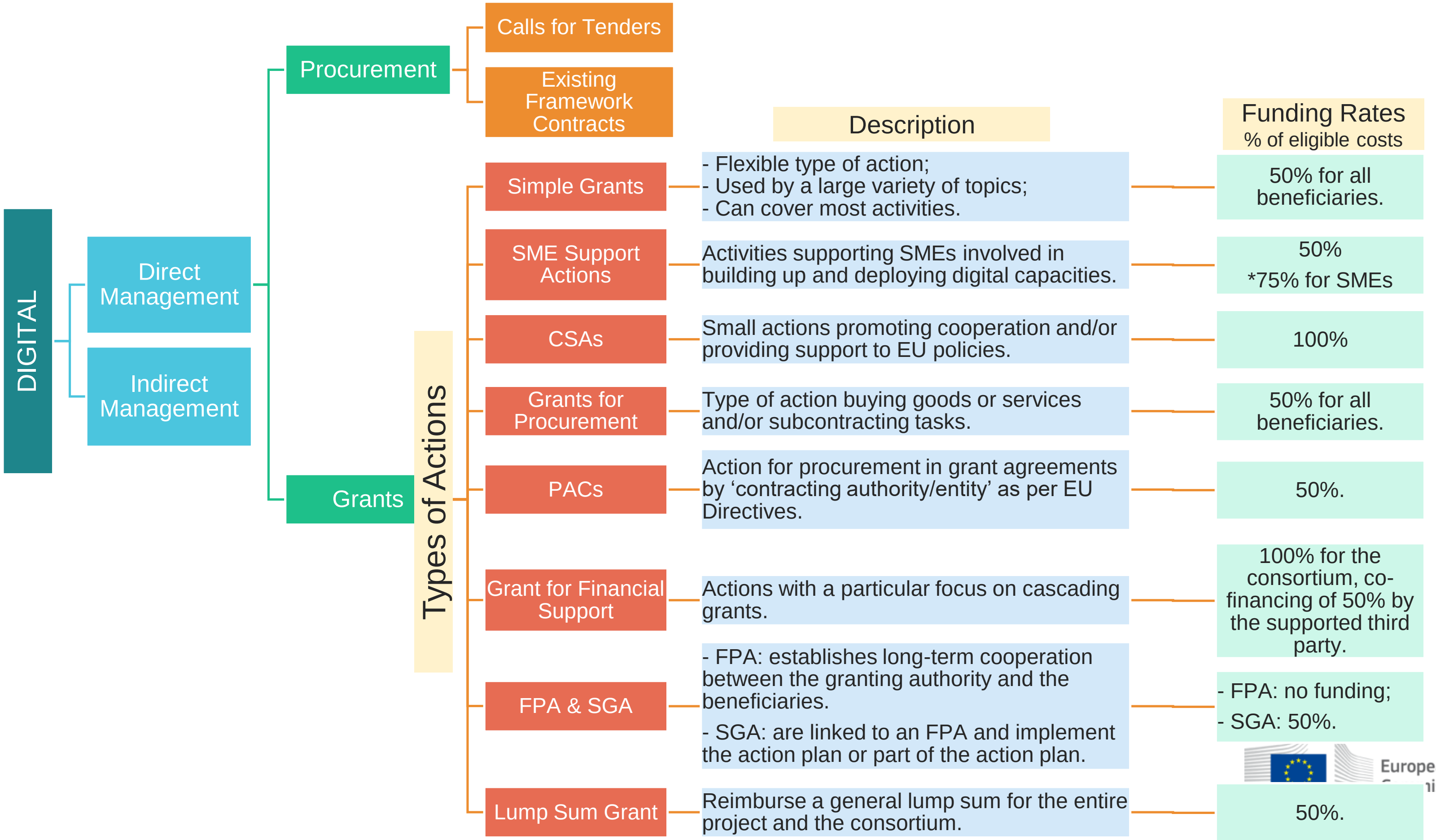
DG CONNECT - Communications Network, Content and Technology

Dir G: Data

Unit G3: Accessibility, **Multilingualism** and Safer Internet

Philippe.Gelin@ec.europa.eu

DIGITAL Implementation in WP 21-22

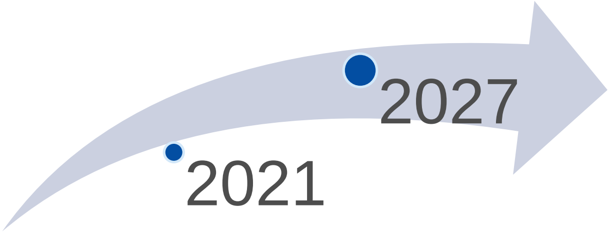


- DIGITAL
  - Direct Management
    - Procurement
      - Calls for Tenders
      - Existing Framework Contracts
    - Grants
  - Indirect Management

Types of Actions

Types of Actions	Description	Funding Rates % of eligible costs
Simple Grants	- Flexible type of action; - Used by a large variety of topics; - Can cover most activities.	50% for all beneficiaries.
SME Support Actions	Activities supporting SMEs involved in building up and deploying digital capacities.	50% *75% for SMEs
CSAs	Small actions promoting cooperation and/or providing support to EU policies.	100%
Grants for Procurement	Type of action buying goods or services and/or subcontracting tasks.	50% for all beneficiaries.
PACs	Action for procurement in grant agreements by ‘contracting authority/entity’ as per EU Directives.	50%.
Grant for Financial Support	Actions with a particular focus on cascading grants.	100% for the consortium, co-financing of 50% by the supported third party.
FPA & SGA	- FPA: establishes long-term cooperation between the granting authority and the beneficiaries. - SGA: are linked to an FPA and implement the action plan or part of the action plan.	- FPA: no funding; - SGA: 50%.
Lump Sum Grant	Reimburse a general lump sum for the entire project and the consortium.	50%.

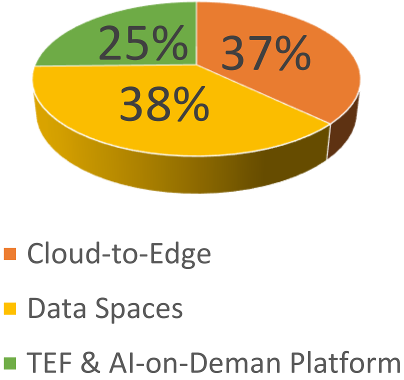
DIGITAL in a Nutshell:
The Programme *Budget*



DIGITAL in MFF **2021-2027**

Total financing **€7.6 Bio**	Total financing for *Cloud, data and Artificial Intelligence* **€2.1 Bio**

[Funding & tenders (europa.eu)]



DIGITAL WP **21-22**

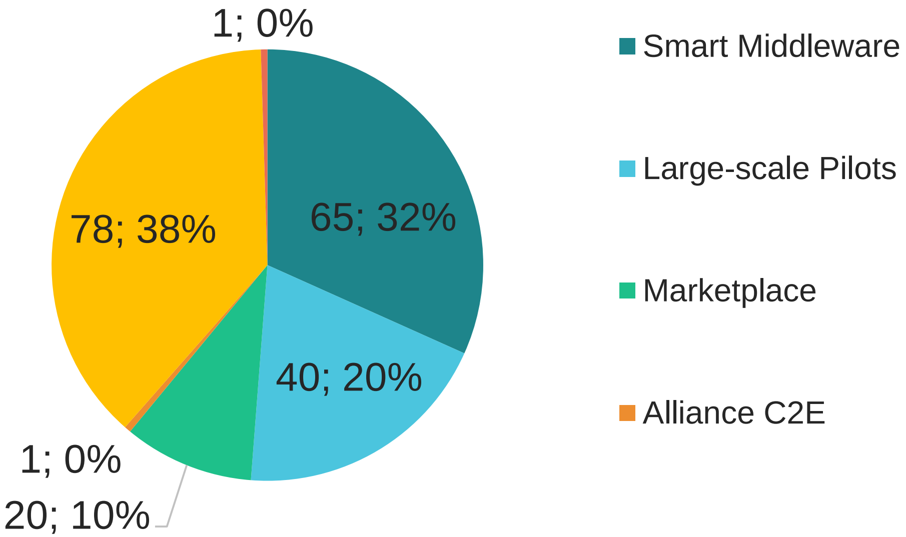
Total financing **€1.98 Bio**	Cloud to Edge: **€204 Mio** Data Spaces: **€206 Mio** TEF & AI-on-demand Platform: **€139 Mio** Total financing for *Cloud, data and AI*: **€549 Mio**

[DIGITAL Work Programme]

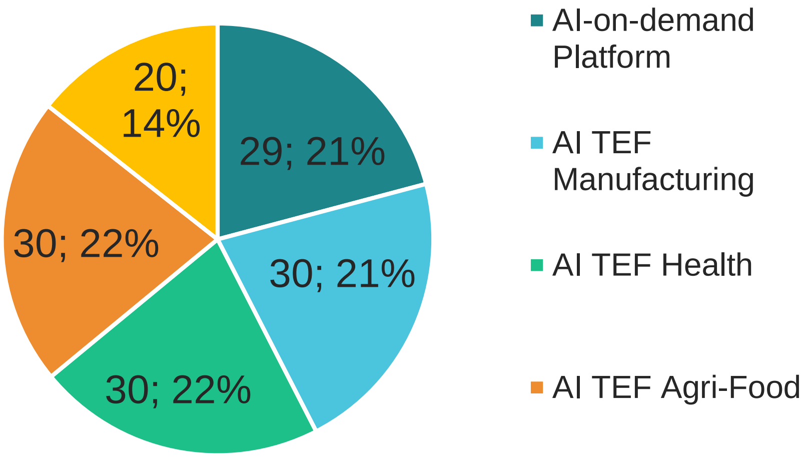
European Commission

DATA SPACES Implementation in WP 21-22

2. Cloud to Edge: **€204 Mio**

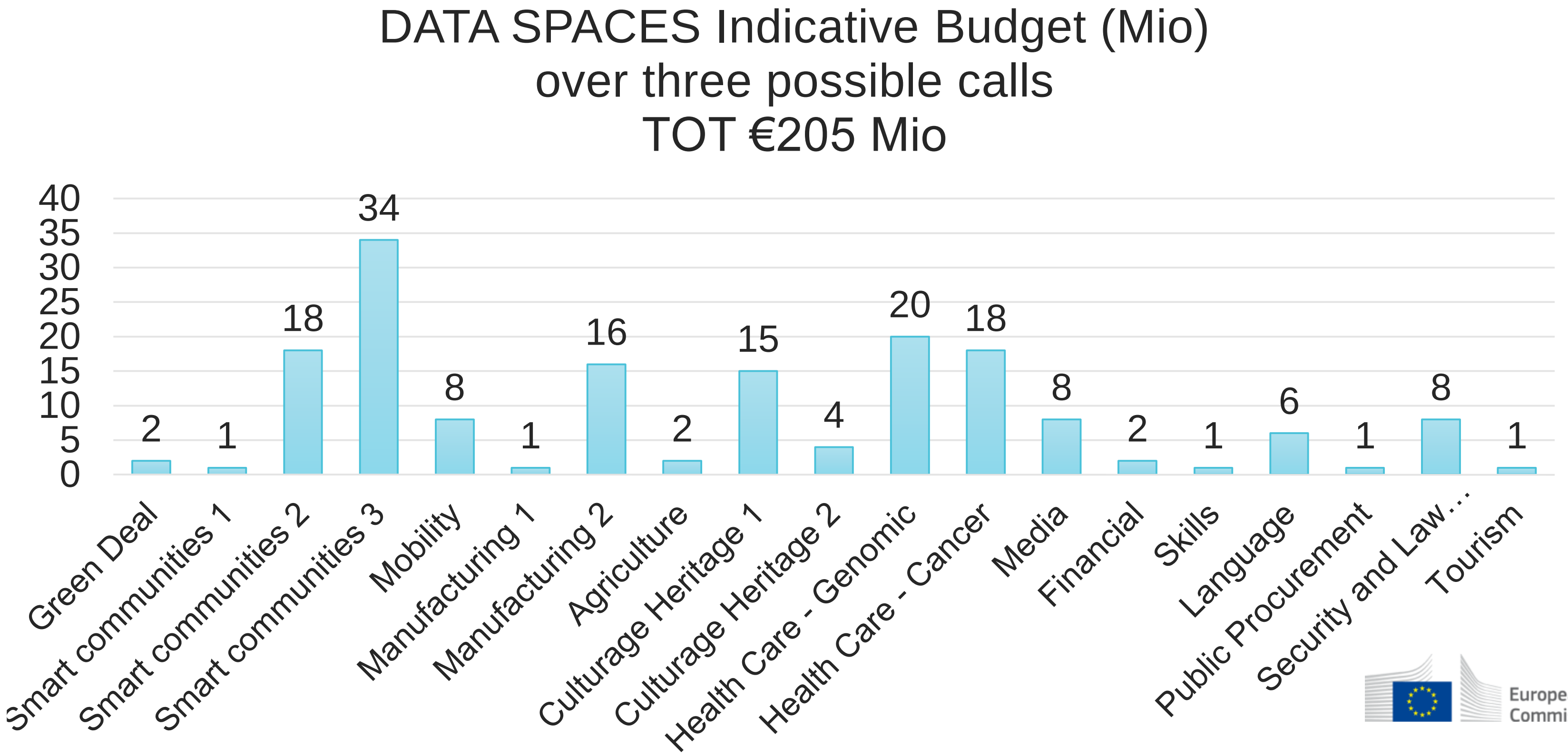


3. TEF & AI-on-demand Platform: **€139 Mio**



European Commission

Data Spaces in DIGITAL:
Indicative Budget according to WP 21-22 (1)



European
Commission