SoBigData Research Infrastructure

Social Mining & Big Data Analytics
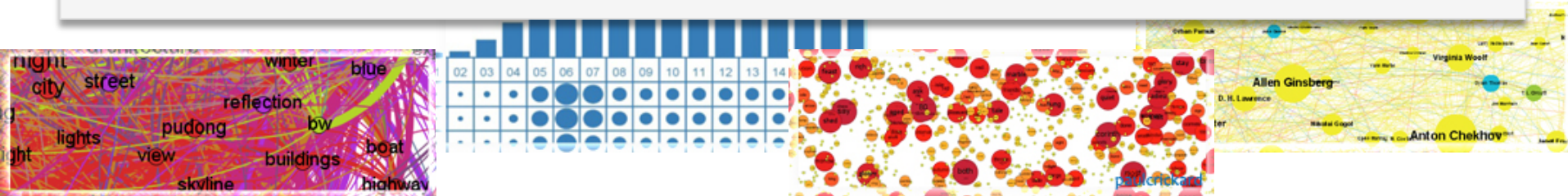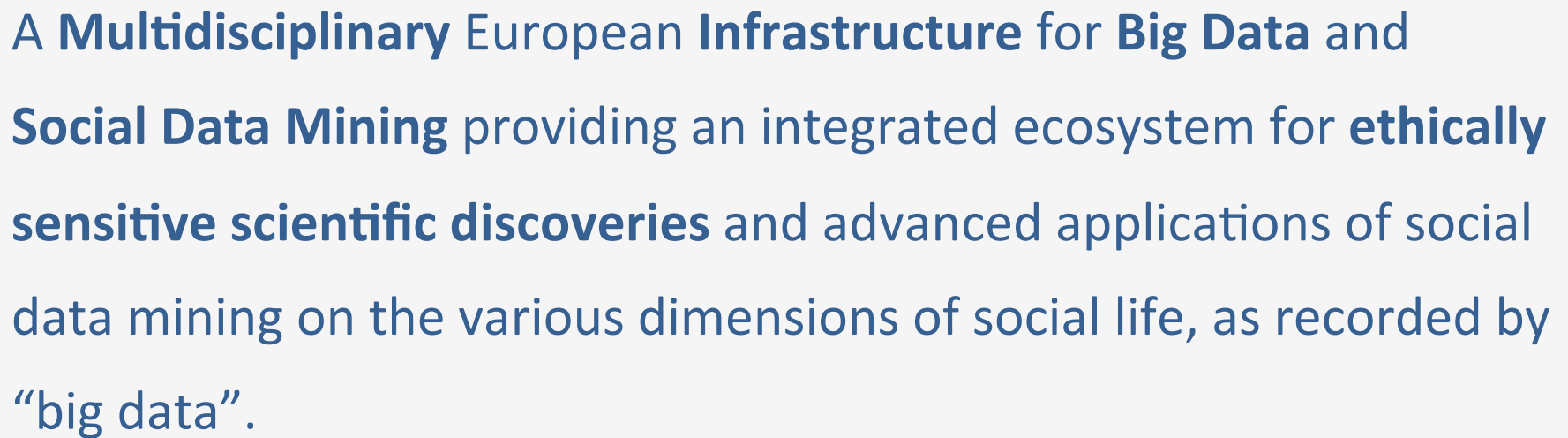
Social Mining
& Big Data Ecosystem

www.sobigdata.eu

HORIZON 2020

Speaker: Massimo Coppola, CNR - ISTI

SoBigData is…

A **Multidisciplinary** European **Infrastructure** for **Big Data** and **Social Data Mining** providing an integrated ecosystem for **ethically sensitive scientific discoveries** and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data".

The Consortium

Italy United Kingdom Germany Estonia

Finland Switzerland Nederlands

Consiglio Nazionale delle Ricerche

The University Of Sheffield.

Università di Pisa

Fraunhofer FHR

TARTU ÜLIKOOL

IMT INSTITUTE FOR ADVANCED STUDIES LUCCA

Leibniz Universität Hannover

KING'S College LONDON

SCUOLA NORMALE SUPERIORE

A! Aalto University

ETH Zürich

TU Delft Delft University of Technology

Existing national RI's to be integrated

SoBigData Euro Lab on Big Data Analytics & Social Mining

GATE general architecture for text engineering

Fraunhofer IGD

L3S Research Center

LIVINGARCHIVE QLectives

The **GOAL** of the Research Infrastructure

The **SoBigData Research Infrastructure** has been designed:

- to **integrate key national infrastructures** and centres of excellence at European level in social mining and big data analytics for creating a networked and virtual ecosystem

- to enable cutting-edge, **multi-disciplinary social mining experiments** with the Research Infrastructure assets: **big data sets**, analytical **tools**, **services** and **skills**

- to **grant access** (both virtual and trans-national on-site) to the SoBigData Research Infrastructure to multidisciplinary *scientists, innovators, public bodies, citizen organizations, SMEs, as well as data science students at any level of education.*

The **pillars** for reaching the goal

- a **distributed data ecosystem** for procurement, access and curation of big social data

- **distributed platform of interoperable, social data mining methods** and associated skills: tools, methodologies and services for mining, analysing, and visualising complex and massive datasets

- **a community of multidisciplinary scientists, innovators, public bodies, citizen organizations, SMEs**, as well as data science students at any level of education scientific, brought together by extensive networking and innovation actions

Thematic Clusters

Visual Analytics

Semantics visualization and real-time solutions for the simulation, visualization and rendering of data

Web Analytics

New models and tools for understanding user behaviour for improving the users' web experience.

Human Mobility Analytics

Methods for mobility analytics, pre-processing tools, integration of diverse mobility observations, advanced mobility data mining

Methods and algorithms covering the major research topics in social networks, community discovery, evolutionary analysis

Social Network Analysis

Intersection of natural, social and engineering sciences to address the challenges that social data generate

Social Data

Natural processing language, mining social media, information extraction, ontology-based semantic annotation

Text and Social Media Mining

Building the e-infra & boosting joint research: exploratories

- Our first exploratories: i.e. the Research Environments tailored to specific multidisciplinary domains
  - different **resources** will made be available (and discoverable) by exploratories: **data, methods and results/publications**

  - We will formulate different **driving scenarios**:
    - societal debates
    - societal well-being and economic performance
    - city of Citizens
    - migration studies

Big Data for City of Citizens
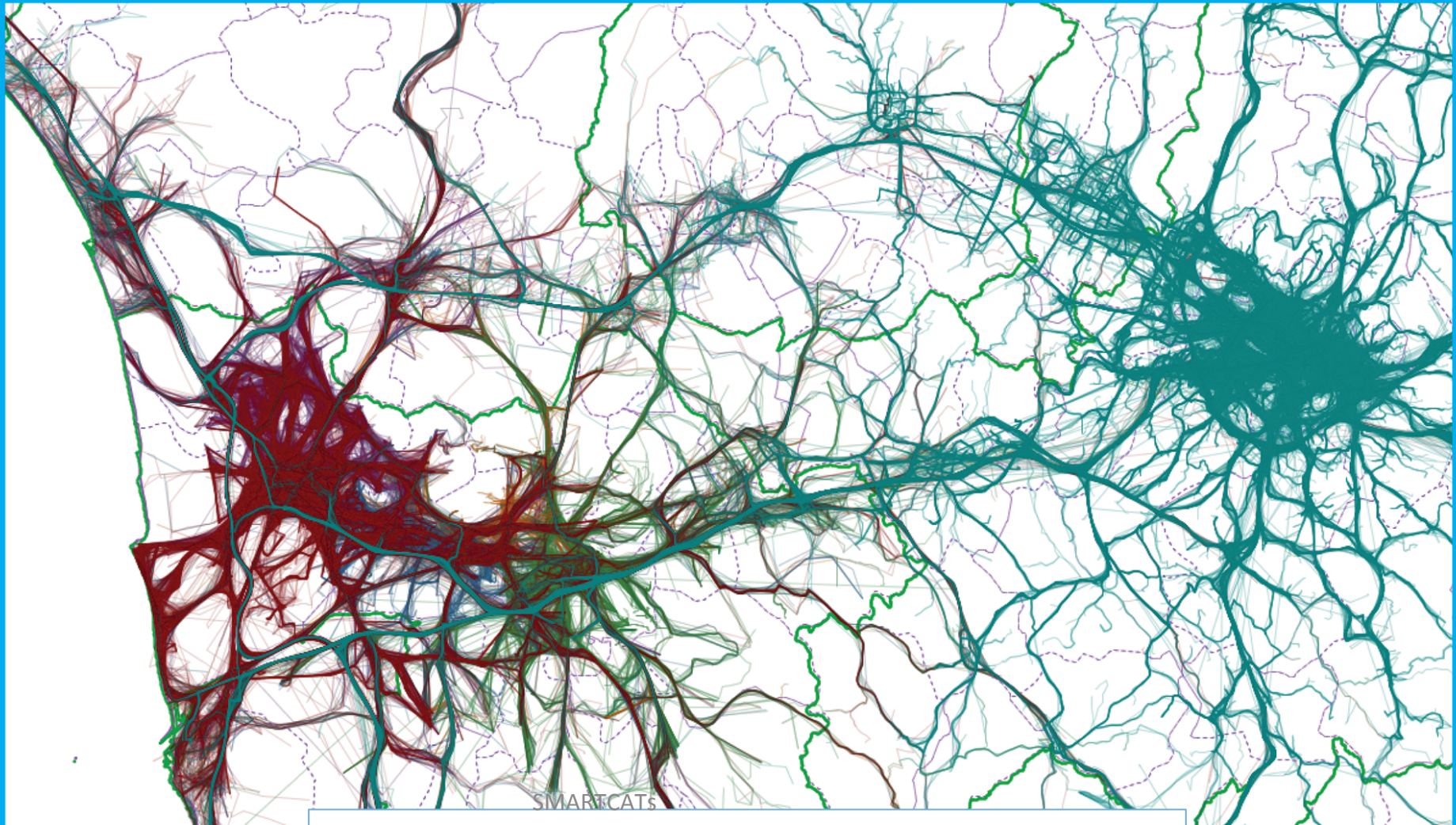
Personal Mobility, Social + Mobility, Personal Sensing

SoBigData.eu E-Infra

Data

- *Publication and validation*
- *Policy definition*
- *Anonymization*
- *Encryption*
- *Embargo definition*
- *Accounting monitoring*

Methods

- Publication and validation
- Policy definition
- Linking to data
- Contextualization
- Provisioning
- Accounting monitoring

SoBigData.eu E-Infra (2)

Data and Methods entities will become infrastructure resources
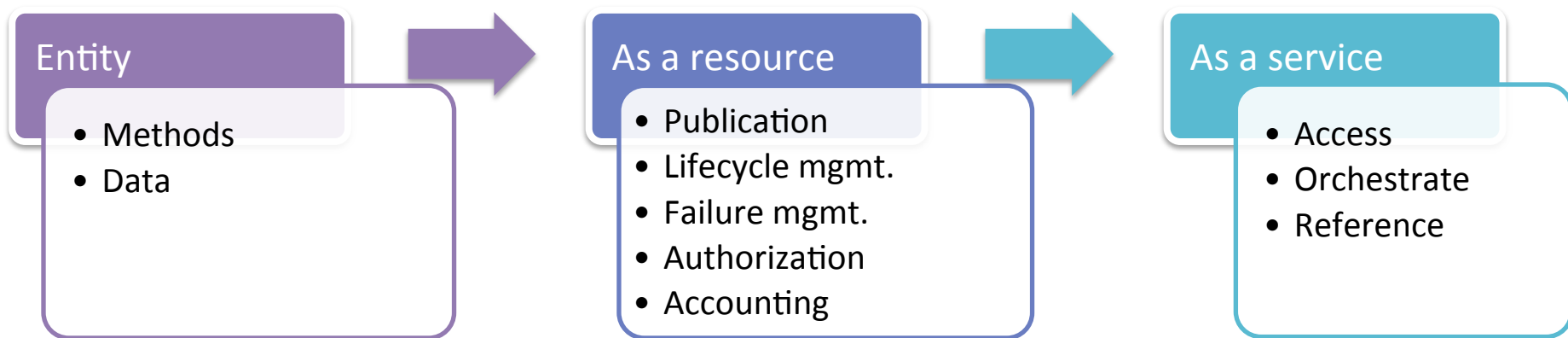
Entity
- Methods
- Data

As a resource
- Publication
- Lifecycle mgmt.
- Failure mgmt.
- Authorization
- Accounting

As a service
- Access
- Orchestrate
- Reference

☐ Data

- ☐ Publication and validation
- ☐ Policy definition
- ☐ Anonymization
- ☐ Encryption
- ☐ Embargo definition
- ☐ Accounting monitoring

☐ Methods

- ☐ Publication and validation
- ☐ Policy definition
- ☐ Linking to data
- ☐ Contextualization
- ☐ Provisioning
- ☐ Accounting monitoring

Managing Data: what this means?

Managing Data does not simply mean

<p style="text-align:center">Support discover
Provide access</p>

It must include support for

- legal interoperability
  - copyright management,
  - licensing of single and derivative products
  - terms of use
- fine-grained policies
  - attribution,
  - citation policy,
  - provenance management

There are three broad types of data:

- **Primary/raw** data: data coming directly from the source;

- **Derivative** data: data processed from primary/raw data;

- **Metadata**: reference data describing either the primary/ raw data or derivative data.

Primary/raw data and derivative data may be licensed under different conditions and by different stakeholders

Managing data: dealing with complexity

The e-infrastructure offering **discover** and **access** may be operated by a different actor and offered through specific terms of use.

In this case three types of licenses may be involved:

1. the one agreed between the system operator and the **primary** data owner,

2. the one selected for **derivative** product that may differ from the one associated with primary data, and

3. the one agreed between the **system operator** and the **data consumer**.

All these licenses have to be captured by the "**terms of use**" of the e-infrastructure, i.e., **they are part of the rules a consumer must agree to accept when using the system**.

Managing data:

VRE as an instrument to manage the complexity

A re-use license (to specify in the ToU of the e-infrastructure) concerns at least attribution, copyleft requirement, and control on commercial exploitation of the dataset. Moreover it is needed to manage and apply some forms of control on access.

Virtual Research Environments (VREs) offer flexible and secure web-based, community-centric platforms, so researchers can work together on common challenges.

VREs terms of use are automatically composed according to the combined data and services selected at the time of VRE definition.

- **Raw data are then licensed according to the license expressed by the data owner/** custodian and expressed at time of registration of the data content to the e-infrastructure.

- **Derivative data products instead are licensed with a license compatible** and legally interoperable **with the one associated with the primary data**.

- It remains under the responsibility of a **single user**, as expressed in the VRE terms of use, to confirm the license to associate with any produced derivative data.

The SoBigData **Stakeholders**

- **Big data analysts and social informatics researchers**, who want to enhance their algorithms to deal with social data, gain multi-disciplinary research skills, harmonise existing data and analytics infrastructures, and engage other research communities in the development of these key enabling technologies for the future digital economy and society;

- **Economists, social science and humanities researchers, journalists, policy and law makers**, who have to analyse the avalanche of (big) social data, in order to gain insight and actionable knowledge.

- **Researchers in related communities,** who would like to use the algorithms, the analytical competences and data infrastructure;

- **Industrial innovators & startuppers**, who would like to create rapid proof-of-concepts of data-driven innovative ideas and services;

- **The public as a whole**, who would like to understand their role in the production, consumption and value-creating of social data.

Where we are

- Project starting date: 01 September 2015

  – 81 people involved, 23 tasks started

  – 3 events scheduled in 2016

  - workshops

    – **Hannover: 5 – 7 Sept in collaboration with Workshop TPDL 2016**

  - summer schools

    – **Lipari 10 – 17 July:**

      http://lipari.cs.unict.it/LipariSchool/ComplexSocialSystems/

  - datathons

    – **ETHZ: http://www.nervousnet.ethz.ch/hackathon - 22 April**

Thank You!

www.sobigdata.eu

info@sobigdata.eu

@SoBigData (https://twitter.com/SoBigData)

https://www.facebook.com/SoBigData

Communication channels

www.sobigdata.eu   info@sobigdata.eu

https://www.facebook.com/SoBigData

@SoBigData (https://twitter.com/SoBigData)