

Health



Food & Agriculture



Energy



Transport



Climate



Social Sciences



Security

**BIG DATA EUROPE**

Empowering Communities  
with Data Technologies

# **BIG DATA EUROPE H2020 CSA (2015-17)**

**22.9.2016**



SC4 Workshop

*Integrating Big Data, Software & Communities for Addressing  
Europe's Societal Challenges*



# Big Data in Marketing







# Big Data in Intelligence

TOP SECRET//SI//ORCON//NOFORN

Hotmail<sup>®</sup> Google<sup>®</sup> Skype<sup>®</sup> paltalk.com<sup>®</sup> YouTube<sup>®</sup>

Gmail<sup>™</sup> facebook<sup>™</sup> YAHOO!<sup>™</sup> AOL<sup>™</sup> mail<sup>™</sup>

 (TS//SI//NF) **PRISM Collection Details** 

**Current Providers**

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!
- Facebook
- PalTalk
- YouTube
- Skype
- AOL
- Apple

**What Will You Receive in Collection (Surveillance and Stored Comms)?**  
It varies by provider. In general:

- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target activity – logins, etc.
- Online Social Networking details
- **Special Requests**

Complete list and details on PRISM web page:  
Go PRISMFAA

TOP SECRET//SI//ORCON//NOFORN



# Big Data Europe (CSA: 2015-17)

- ⊙ **Show** societal value of Big Data: 7 Domains



Health



Food & Agriculture



Energy



Transport



Climate



Social Sciences



Security

- ⊙ **Lower** barrier for using big data technologies

- Required effort and resources
- Limited data science skills

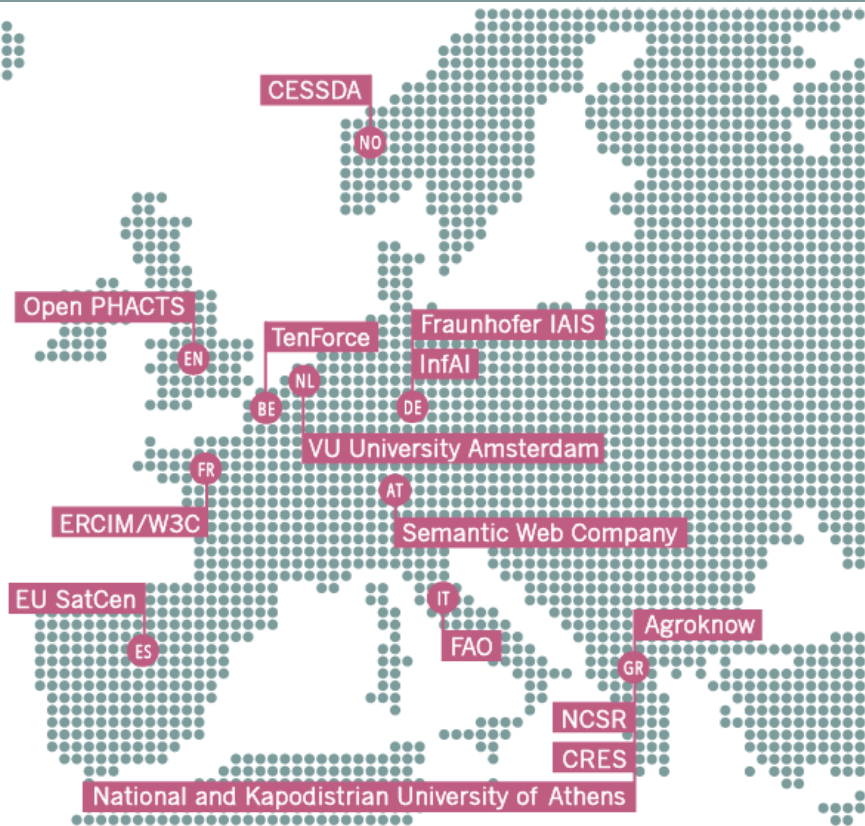
- ⊙ **Help** establishing cross-lingual/organizational/domain Data Value Chains



# Consortium

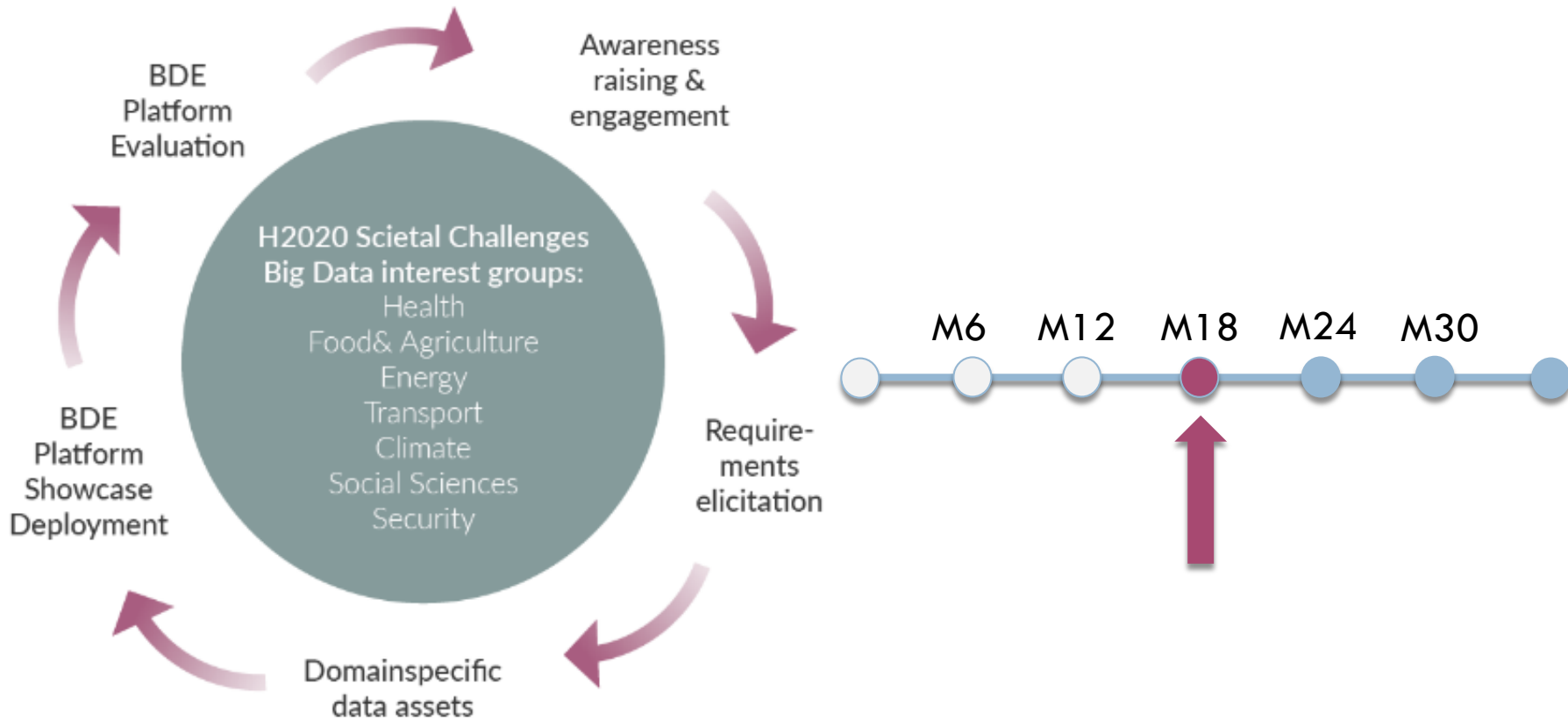


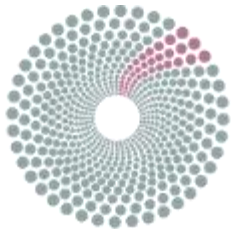
cessda





# Stakeholder Engagement & Activities





Health



Food & Agriculture



Energy



Transport



Climate



Social Sciences



Security

**BIG DATA EUROPE**

Empowering Communities  
with Data Technologies

# ARCHITECTURE & COMPONENTS



*Integrating Big Data, Software & Communities for Addressing  
Europe's Societal Challenges*

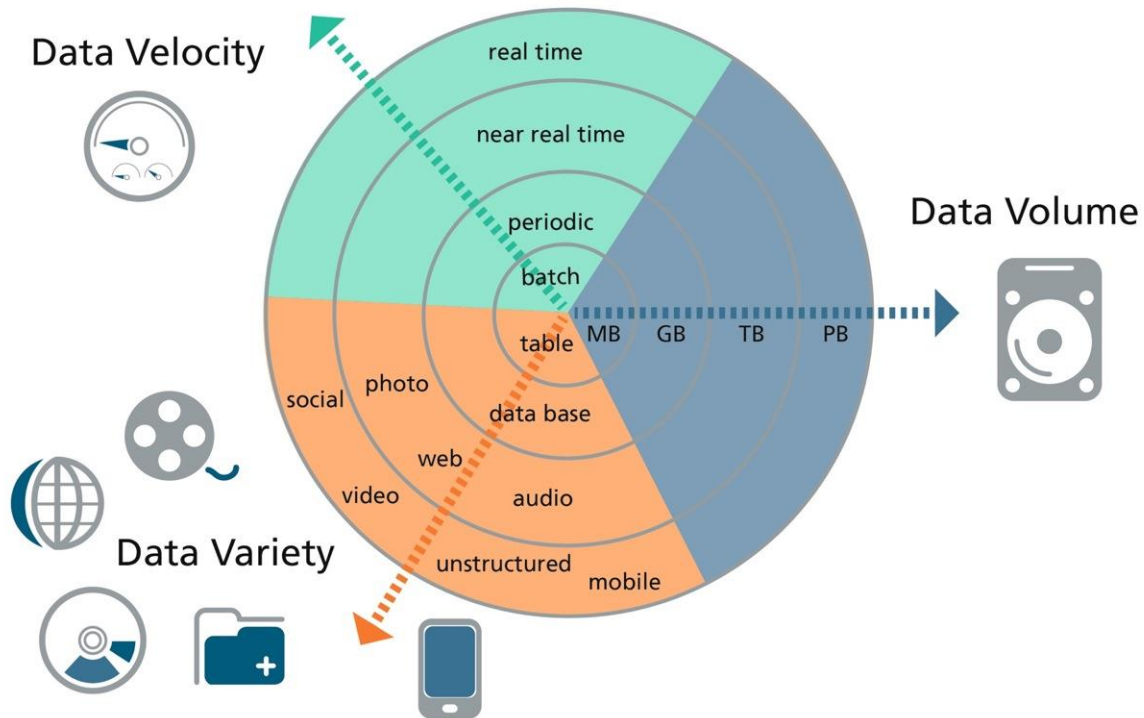


# The three Big Data „V“

## Variety is often neglected



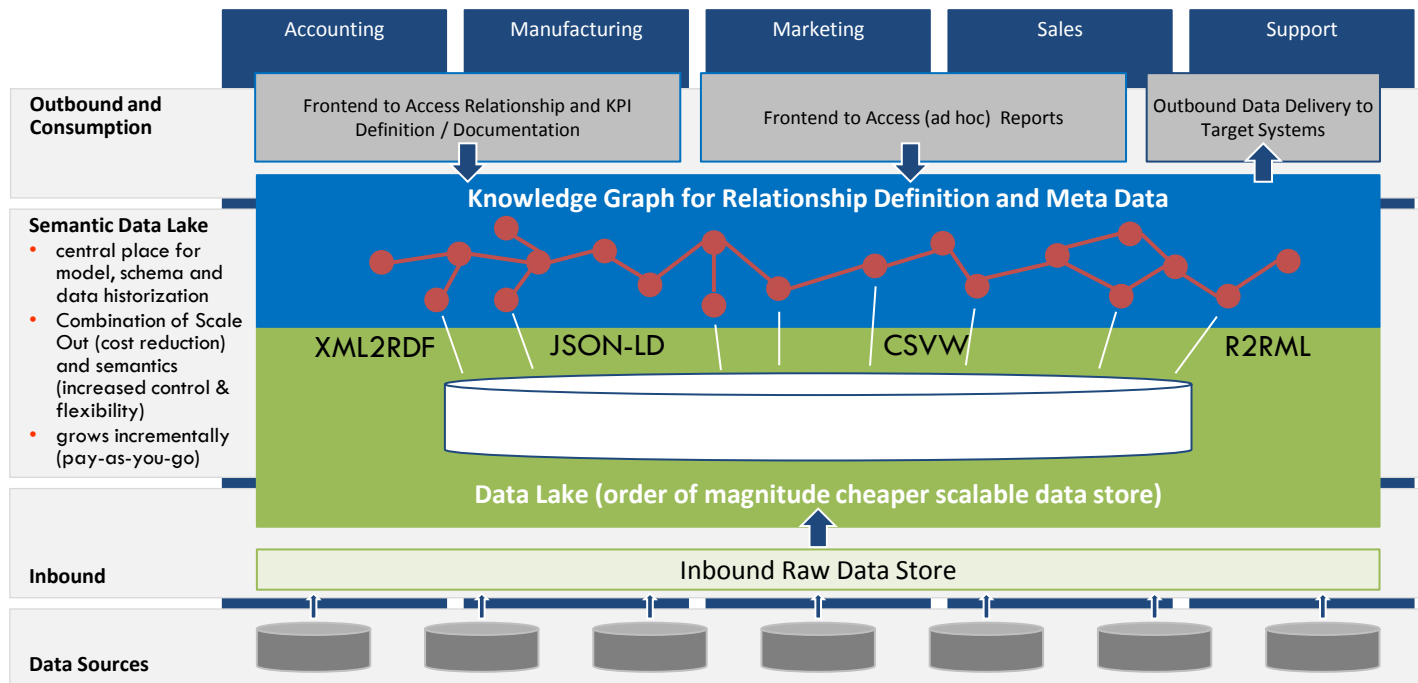
**BIG DATA EUROPE**  
Empowering Communities  
with Data Technologies





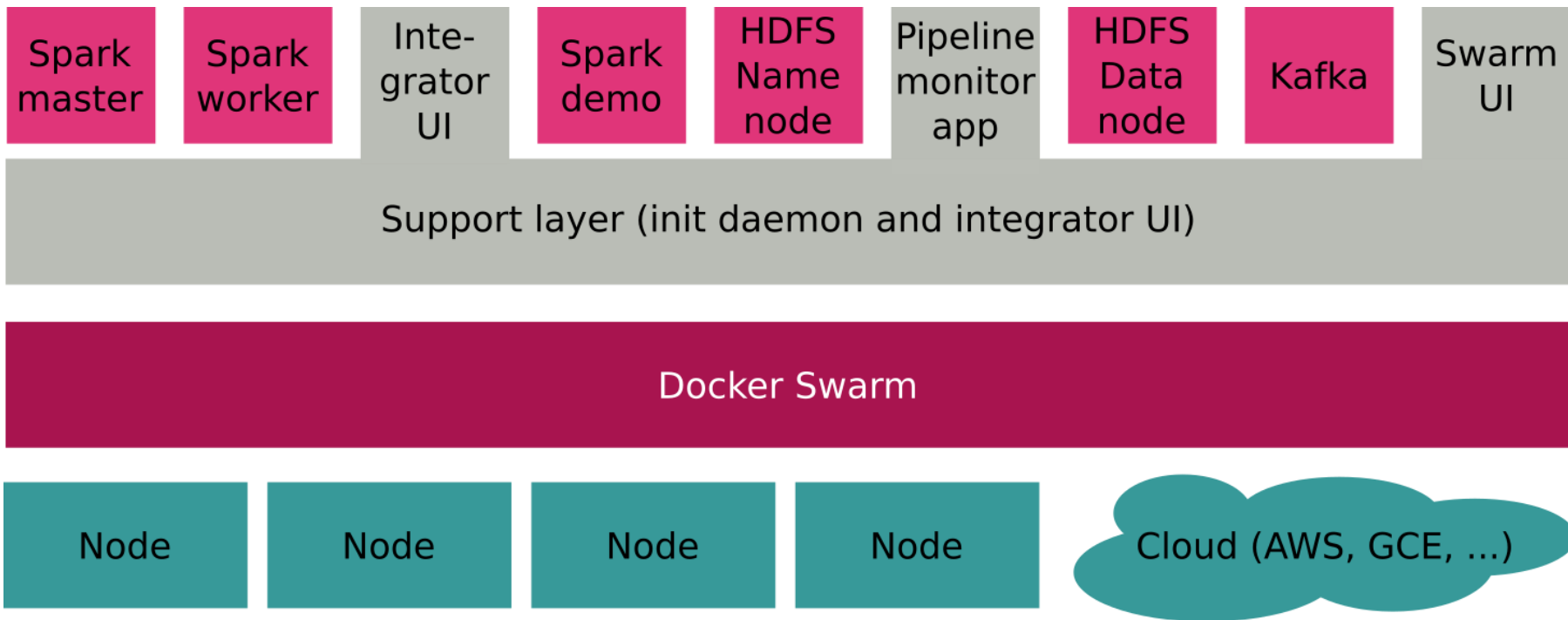


# Adding a Semantic Layer to Data Lakes





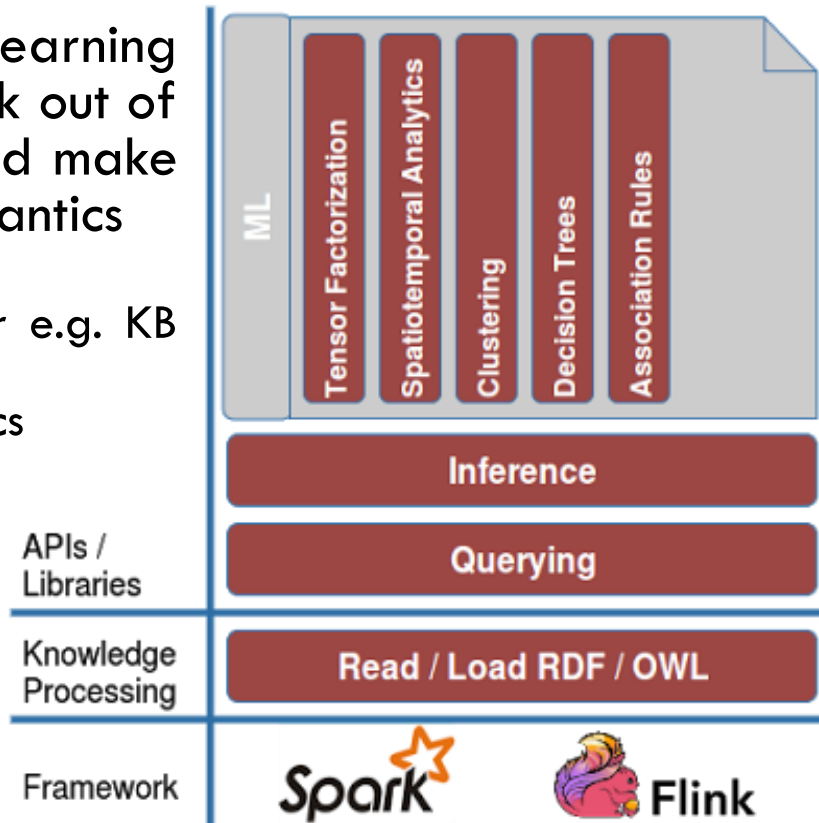
# Current State of Platform Architecture





# BDE SANSA Stack

- ⊙ Distributed Machine Learning (ML) algorithms that work out of the box on RDF data and make use of its structure / semantics
- ⊙ Examples:
  - Tensor Factorization for e.g. KB completion
  - Spatiotemporal analytics
  - Anomaly prediction
  - Clustering
  - Association rules
  - Decision trees





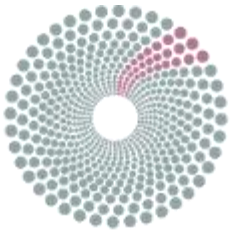
# Platform components

|                         |                            |
|-------------------------|----------------------------|
| <b>Search/indexing</b>  | <b>Data processing</b>     |
| Apache Solr             | Apache Spark               |
| <b>Data acquisition</b> | Apache Flink               |
| Apache Flume            | <b>Semantic Components</b> |
| <b>Message passing</b>  | Strabon                    |
| Apache Kafka            | Sextant                    |
| <b>Data storage</b>     | GeoTriples                 |
| Hue                     | Silk                       |
| Apache Cassandra        | SEMAGROW                   |
| ScyllaDB                | LIMES                      |
| Apache Hive             | 4Store                     |
| Postgis                 | OpenLink Virtuoso          |



# BDE vs Hadoop distributions

|   | Hortonworks                    | Cloudera                       | MapR                             | Bigtop                         | BDE                               |
|---|--------------------------------|--------------------------------|----------------------------------|--------------------------------|-----------------------------------|
| <i>File System</i>                                  | HDFS                           | HDFS                           | NFS                              | HDFS                           | <b>HDFS</b>                       |
| <i>Installation</i>                                 | Native                         | Native                         | Native                           | Native                         | <b>lightweight virtualization</b> |
| <i>Plug &amp; play components (no rigid schema)</i> | no                             | no                             | no                               | no                             | <b>yes</b>                        |
| <i>High Availability</i>                            | Single failure recovery (yarn) | Single failure recovery (yarn) | Self healing, mult. failure rec. | Single failure recovery (yarn) | <b>Multiple Failure recovery</b>  |
| <i>Cost</i>   | Commercial                     | Commercial                     | Commercial                       | Free                           | <b>Free</b>                       |
| <i>Scaling</i>                                      | Freemium                       | Freemium                       | Freemium                         | Free                           | <b>Free</b>                       |
| <i>Addition of custom components</i>                | Not easy                       | No                             | No                               | No                             | <b>Yes</b>                        |
| <i>Integration testing</i>                          | yes                            | yes                            | yes                              | yes                            | --                                |
| <i>Operating systems</i>                            | Linux                          | Linux                          | Linux                            | Linux                          | <b>All</b>                        |
| <i>Management tool</i>                              | Ambari                         | Cloudera manager               | MapR Control system              | -                              | <b>Docker swarm UI+ Custom</b>    |



Health



Food & Agriculture



Energy



Transport



Climate



Social Sciences



Security

**BIG DATA EUROPE**

Empowering Communities  
with Data Technologies

# 7 SC-PILOT INSTANTIATIONS



*Integrating Big Data, Software & Communities for Addressing  
Europe's Societal Challenges*



# Pilots: Overview

## ⊙ SC1: Health & Pharm.



## ⊙ SC2: Food & Agr.



## ⊙ SC3: Energy



## ⊙ SC4: Transport



## ⊙ SC5: Climate



## ⊙ SC6: Social Sciences



## ⊙ SC7: Security





# SC1: Life Sciences & Health



SC1: Life Sciences & Health

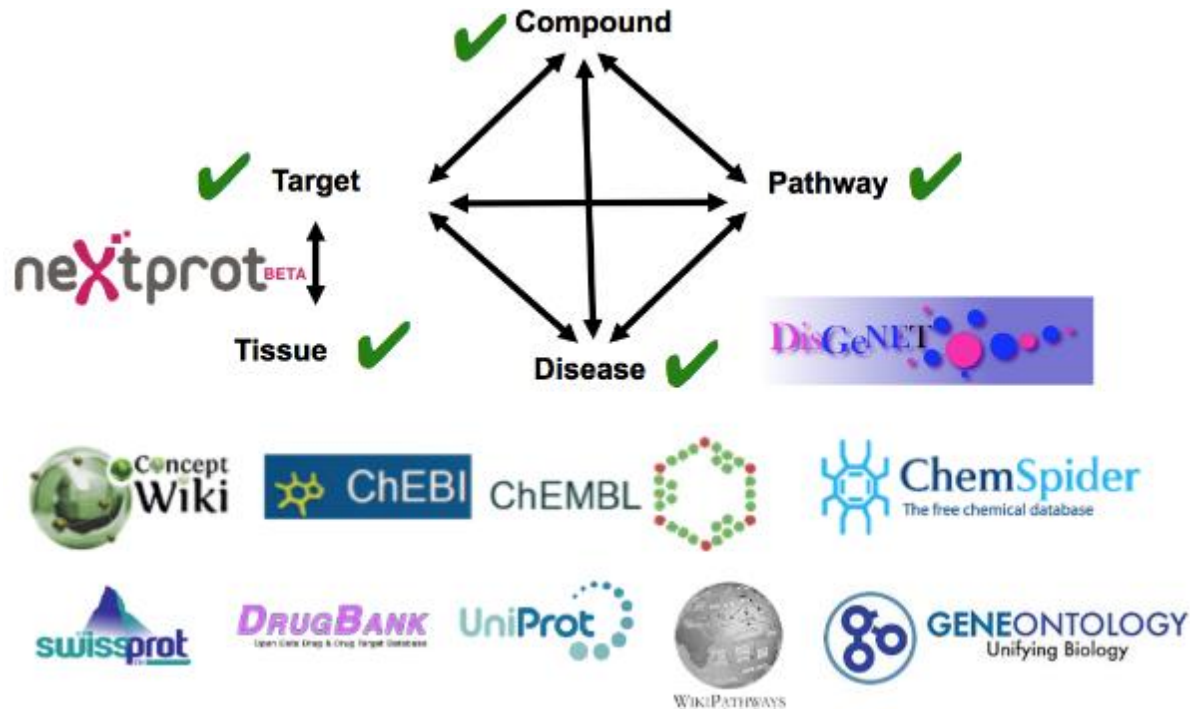




# SC1: Life Sciences & Health

**Big Data Focus area:** Large-scale heterogeneous pharmaceutical research data linking & integration

**Selected Key Data assets:** ACD Labs / ChemSpider, ChEBI, ChEMBL, ConceptWiki, DrugBank, ENZYME, Gene Ontology, GO Annotation, SwissProt, WikiPathways





# SC1: Life Sciences & Health

Pilot 1: Replicate Open PHACTS functionality on the BDE infrastructure using Open Source solutions

Reasons:

- Deployment possible in-house
- Apply to other domains (e.g. Agriculture)
- Using extra BDE functionalities (e.g. logging, analysis)

[www.big-data-europe.eu](http://www.big-data-europe.eu)

Open PHACTS pgroth

Developer Home | Want help? | Documentation | Get my API keys | Featured Apps

Search

### OpenPHACTS v1.4 API Active Docs

The response template for each operation colour coded as follows:

- Required elements that always return a single value
- Required elements that return either a single value or an array
- Optional elements that always return a single value
- Optional elements that return either a single value or an array

#### Operations

OpenPHACTS API v1.4

|  |                                     |     |
|--|-------------------------------------|-----|
| Compound Information                   | /compound                           | GET |
| Compound Information: Batch            | /compound/batch                     | GET |
| Compound Class Members: Count          | /compound/members/count             | GET |
| Compound Class Members: List           | /compound/members/pages             | GET |
| Compound Classifications               | /compound/classifications           | GET |
| Classification of Targets for Compound | /compound/classificationsForTargets | GET |
| Compound Pharmacology: Count           | /compound/pharmacology/count        | GET |
| Compound Pharmacology: List            | /compound/pharmacology/pages        | GET |
| Compound Class Pharmacology: Count     | /compound/tree/pharmacology/count   | GET |
| Compound Class Pharmacology: List      | /compound/tree/pharmacology/pages   | GET |



# SC2: Food & Agriculture



## SC2: Food & Agriculture



# SC2: Food & Agriculture

## AGINFRA

**Big Data Focus area:**  
**Large-scale distributed**  
**agricultural data**  
**integration**

**Selected Key Data assets:**

INFOODS, AQUASTAT

Green Learning Network

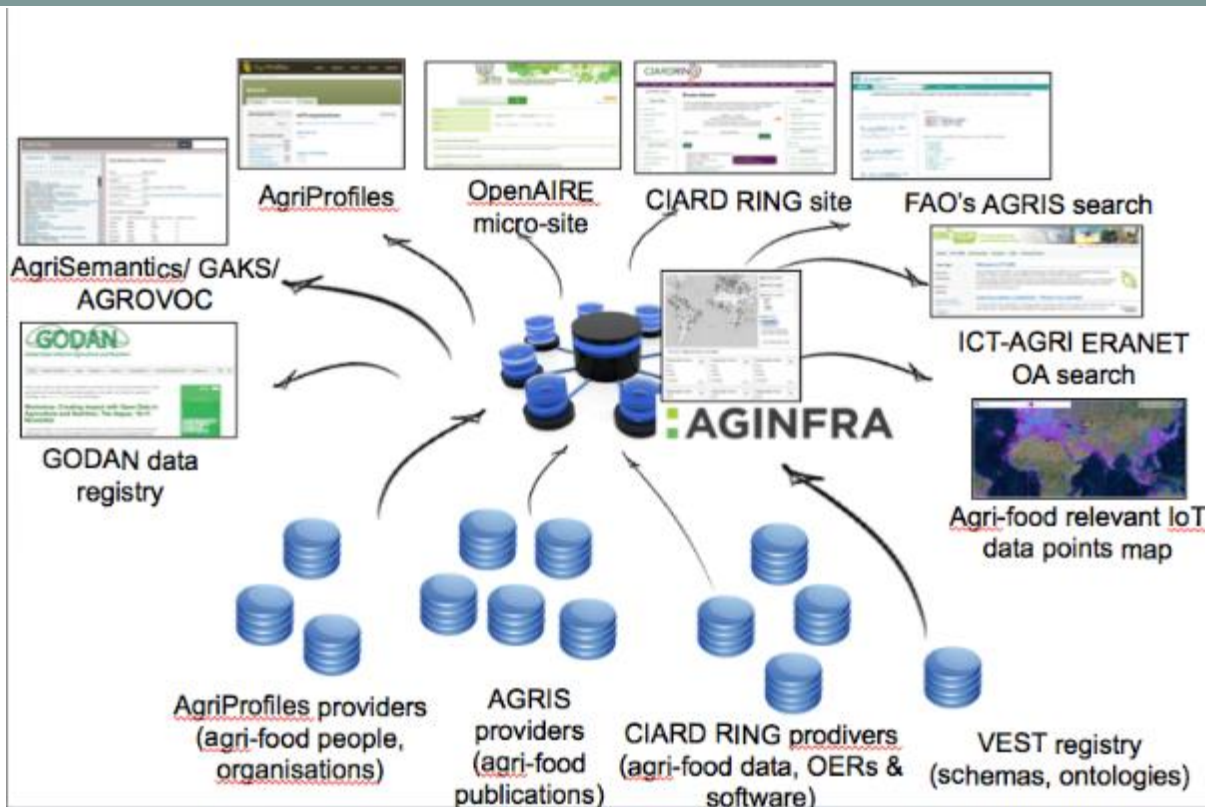
(GLN), Agricultural

Bibliography Network

(ABN), AgroVoc, AquaMaps,

Fishbase

[www.big-data-europe.eu](http://www.big-data-europe.eu)





# SC2: Food & Agriculture



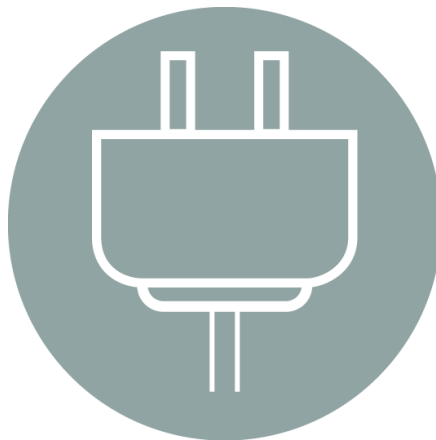
## Pilot focus area: **Viticulture**

*(from the Latin word for vine)  
is the science, production,  
and study of grapes.  
It deals with the series of  
events that occur in the vineyard.*





# SC3: Energy



SC3: Energy



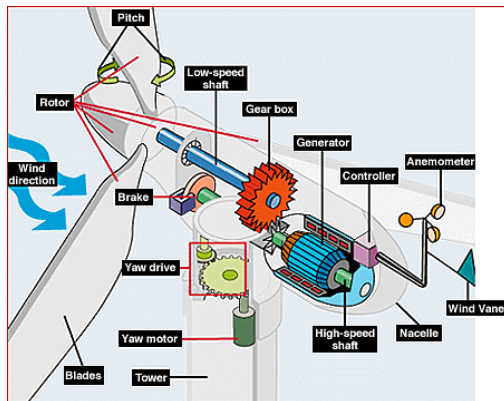
# SC3: Energy

**Big Data Focus area:** Real-time turbine monitoring stream processing and analytics

**Selected Key Data assets:** European Energy Exchange Data, smart meter sensor data, gas/fuels market/price data, consumption statistics, stratigraphic model data (geology, geophysics)

Pilot focus area:

**System monitoring in energy production units.**



[www.big-data-europe.eu](http://www.big-data-europe.eu)





# SC4: Transport



The Fraunhofer Society is a German research organization with 67 institutes spread throughout Germany, each focusing on different fields of applied science.



The **Centre for Research and Technology-Hellas (CERTH)** founded in 2000 is one of the leading research centres in Greece. CERTH includes the **Hellenic Institute of Transport (HIT):** Land, Sea and Air Transportation as well as Sustainable Mobility services



ERTICO - ITS Europe is a partnership of around 100 companies and institutions involved in the production of Intelligent Transport Systems (ITS).



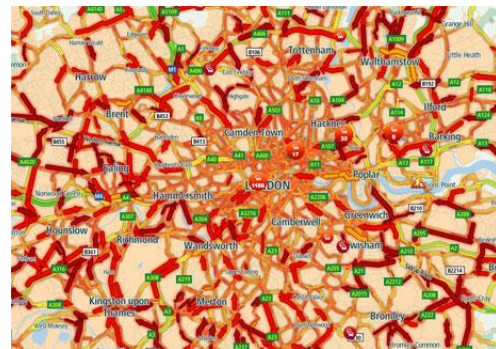


# SC4 Pilot Focus Area

## Info mobility based on *Mobility Pattern* Identification

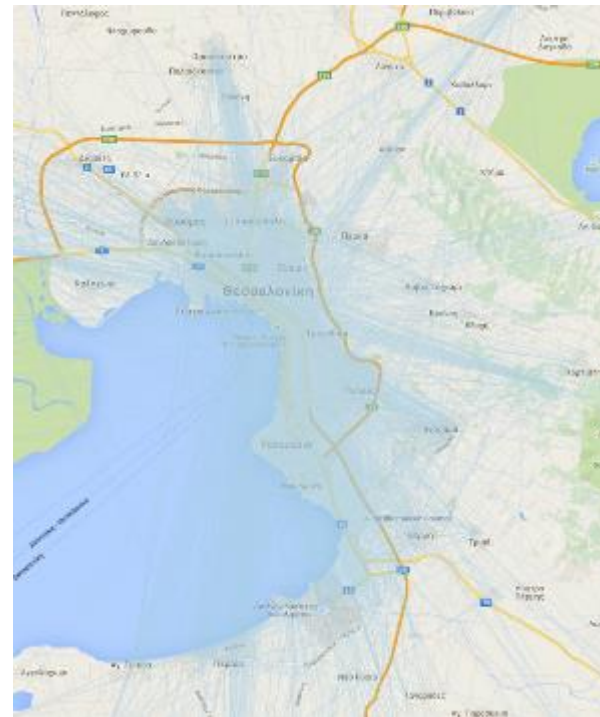
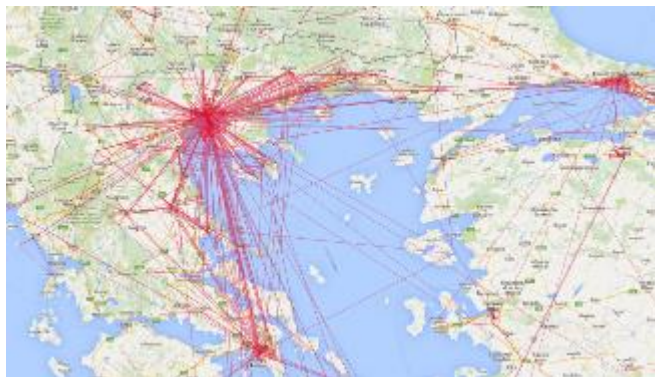
Pilot 4: Multisource data collection  
for the provision of accurate info-  
mobility and advanced transport  
planning service in **Thessaloniki,**  
**Greece**

[www.big-data-europe.eu](http://www.big-data-europe.eu)





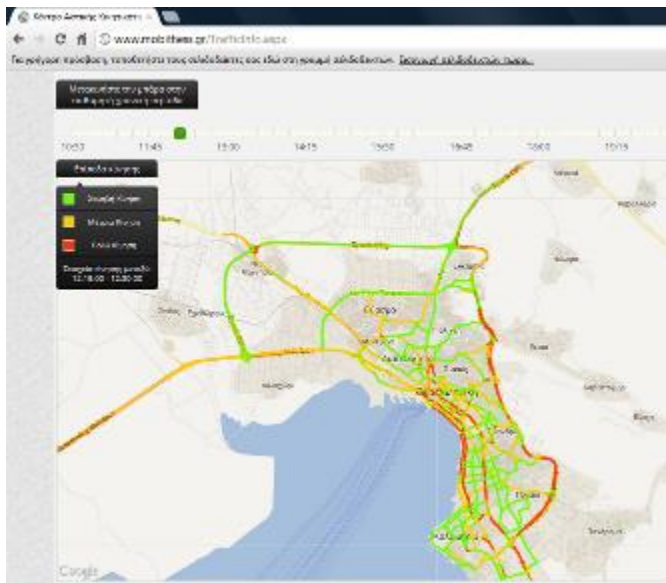
# SC4: Twitter data in Thessaloniki



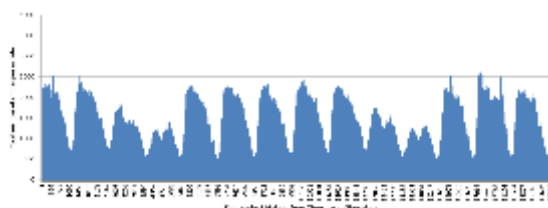


# SC4: Floating Car Data

Real time traffic conditions information based on a combination of traffic modeling and real time measurements (traffic flow and speed)

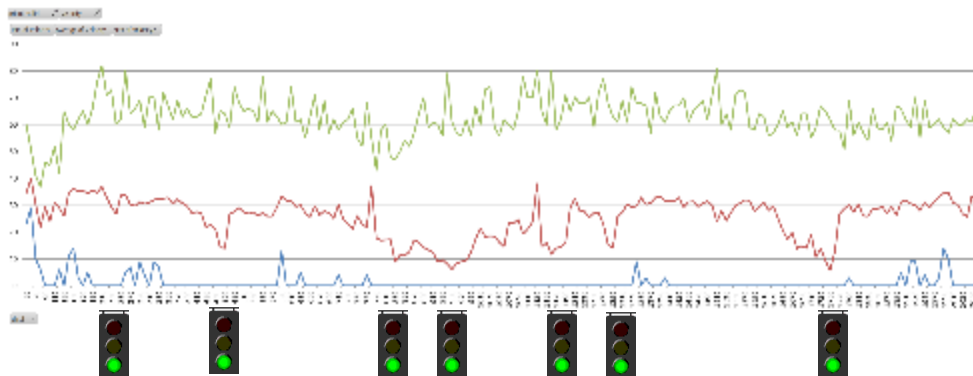


[www.big-data-europe.eu](http://www.big-data-europe.eu)



>1.200 vehicles (one taxi fleet)

- Circulating 16-24 hours/day
- Pulse each 100m or 10s
- 500-2.500 pulses /minute



Speeds along a 2km stretch



# SC5: Climate



SC5: Climate



# SC5: Climate

Pilot focus area:

## Supporting data-intensive climate research

**Big Data Focus area:** Enormous simulation time. Extremely complicated computing model.

**Selected Key Data assets:** European Grid Infrastructure (EGI). Access to several data centres hosted at CNRS-Lyon, NCSR-D Athens, INFN-Milan, NIKhEF-Amsterdam.



[www.big-data-europe.eu](http://www.big-data-europe.eu)





# SC6: Social Sciences

---



## SC6: Social Sciences



# SC6: Social Sciences

## Pilot focus area: Citizens budget spending on municipal level

**Big Data Focus area: Statistical and research data linking & integration**  
**Selected Key Data assets:** Federated social sciences data catalogs, statistical data from public data portals and statistical offices (e.g. EuroStats, UNESCO, WorldBank)



[www.big-data-europe.eu](http://www.big-data-europe.eu)





# SC7: Security



SC7: Security





# SC7: Security

## Pilot focus area: Getting insight in man-made surface changes triggered by automatic detection, news, or social media information

**Big Data Focus area: Image data analysis**

**Selected Key Data assets:** Earth Observation data (e.g. Very High Resolution Satellite Imagery acquired from commercial providers and governmental systems) and collateral data for supporting CFSP/CSDP missions and operations

[www.big-data-europe.eu](http://www.big-data-europe.eu)



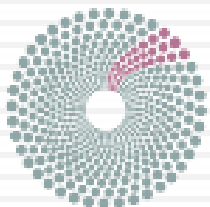


# SC7: Security

**Pilot 7: Ingestion of remote sensing images and social sensing data to detect and verify man-made changes on the Earth surface for security applications**

## Reasons:

- © **Evacuation** route planning
- © **Monitoring** of critical infrastructures
- © **Border security**
- © Satellite image data is HUGE and computational intensive to compare
- © Smart 'focus' algorithms are needed to prioritize the analysis jobs



[www.big-data-europe.eu](http://www.big-data-europe.eu)



#BigDataEurope



# Questions & Contacts



**WEB:** [www.big-data-europe.eu](http://www.big-data-europe.eu)

**EMAIL:** [info@big-data-europe.eu](mailto:info@big-data-europe.eu)

## PROJECT COORDINATION

**Prof. Sören Auer**, auer © cs.uni-bonn · de ([Fraunhofer IAIS](#))

> **Dr. Simon Scerri (Deputy)**, scerri © cs.uni-bonn · de ([Fraunhofer IAIS](#))



EIS Department/Group,  
Fraunhofer IAIS & CS Department Uni-Bonn,  
Bonn, Germany

**Fraunhofer IAIS: Leads Fraunhofer Big Data Alliance**