BIG

Big Data Public Private Forum

TOWARDS A BIG DATA ROADMAP FOR EUROPE

Martin Strohbach, AGT International

Tilman Becker, Edward Curry, John Domnique, Ricard Munné, Sebnem Rusitschka, Sonja Zillner

OVERVIEW

BIG
Big Data Public Private Forum

- **Big Data Public Private Forum and Partnership**
- **3 Data Sharing Use Cases**
- **Methodology for creating a community based roadmap**
- **Findings from the Use Cases, Sectors and Data Value Chain Analysis**
- **Impact of the findings towards a European roadmap**
- **Conclusions**

THE EU PROJECT BIG
BIG DATA PUBLIC PRIVATE FORUM

Trigger

Europe needs a clear strategy for leveraging Big Data Economy in Europe

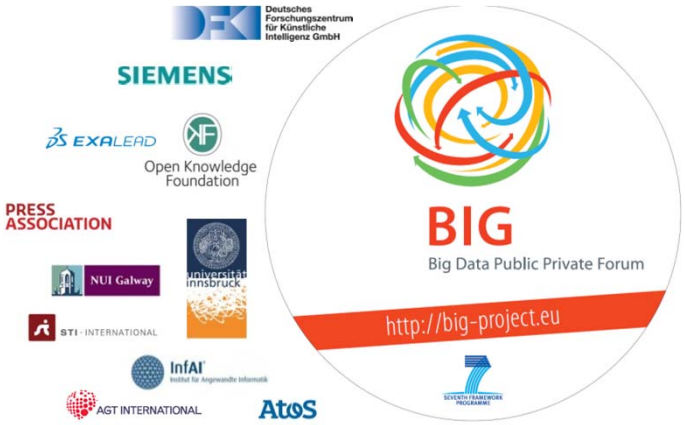
Objectives

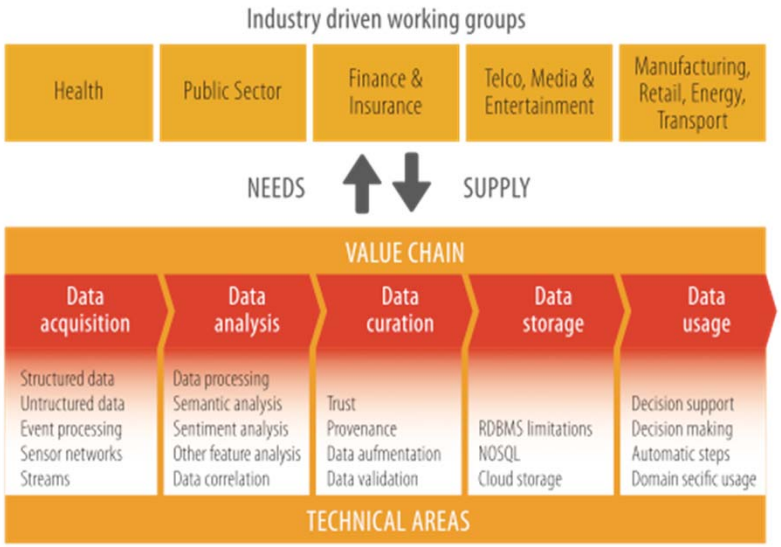
Work at technical, business and policy levels, shaping the future through the positioning of Big Data in Horizon 2020.

Bringing the necessary stakeholders into a **sustainable industry-led initiative**, which will greatly contribute to enhance the EU competitiveness taking full advantage of Big Data technologies.

Facts

Type of project: **Coordination & Support Action**
Project start date: **September 2012**
Duration: **26 months**
Call: **FP7-ICT-2011-8**
Budget: **3,038 M€**
Consortium: **11 partners**





BIG DATA PRIVATE PUBLIC PARTNERSHIP

BIG
Big Data Public Private Forum

- Objectives: *„The goal […] is to increase the amount of productive European economic activities and the number of European jobs that depend on the* ***availability of high quality data assets*** *and the technologies needed to derive value from them."*
  (Source. Strategic Research and Innovation Agenda, SRIA)

- Neelie Kroes, EU Commissioner for the Digital Agenda: **„This is a revolution and I want the EU to be right at the front of it"**

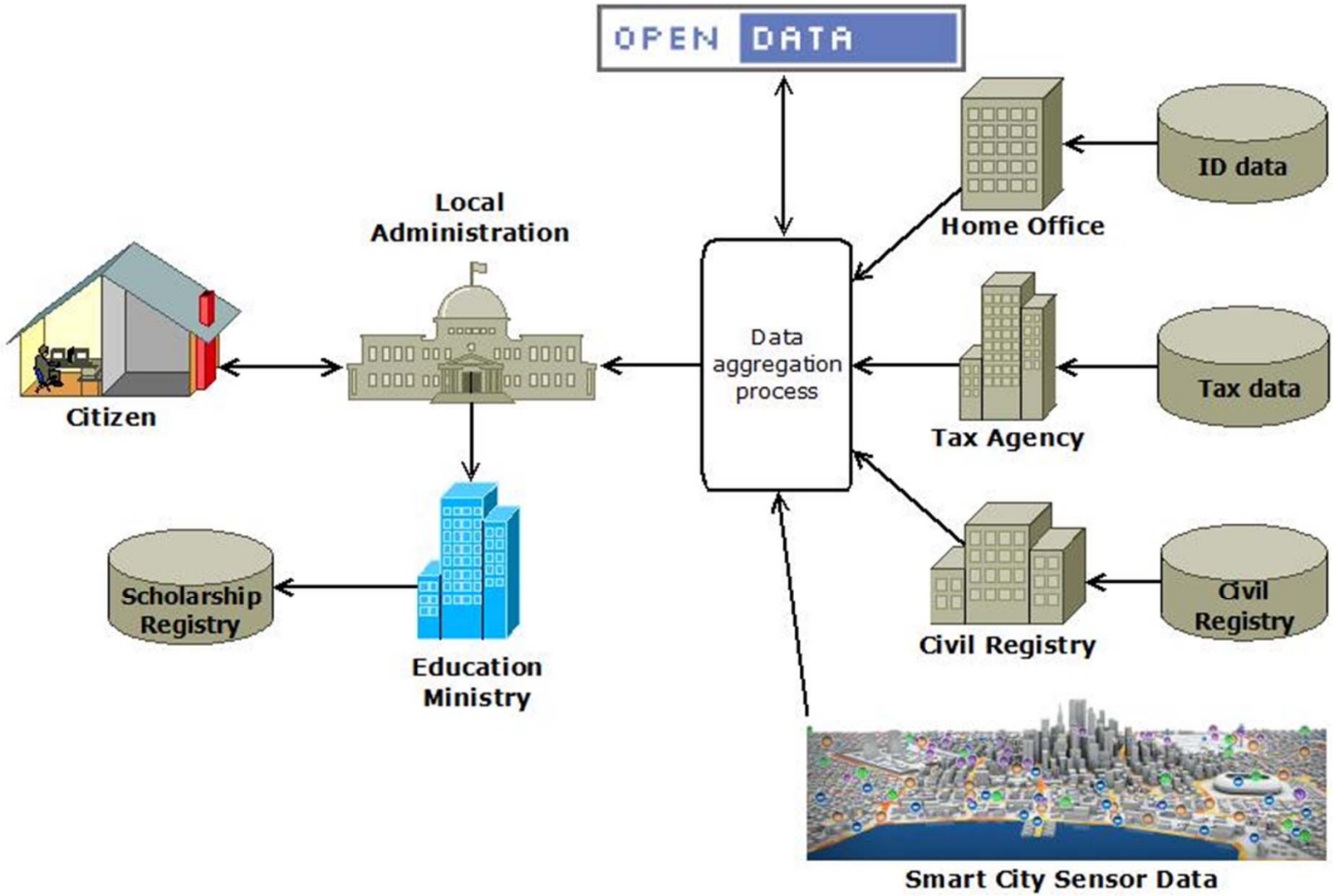
- 29 European organisations from industry and research

- Results of public consultattion to be presented at NESSI Summit Brussels, 27th May

SELECTED USE CASES

OVERCOMING THE DATA SILO CULTURE IN PUBLIC SECTOR

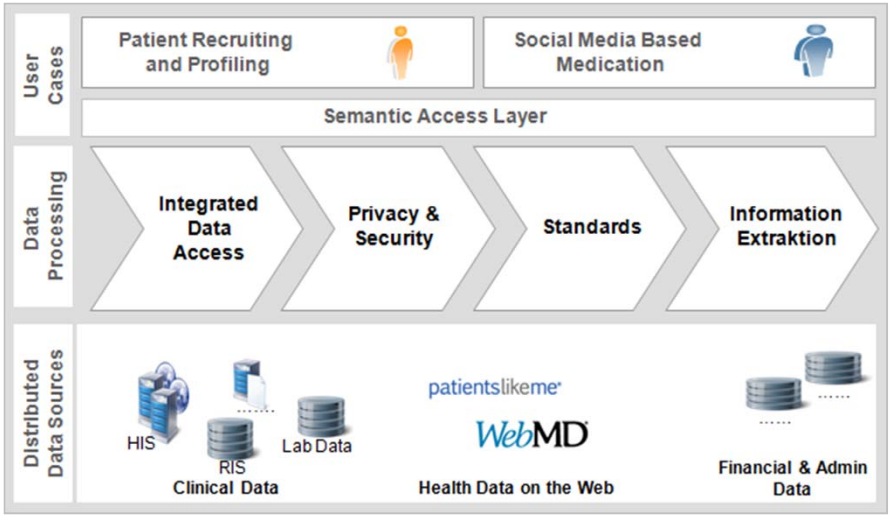
BIG
Big Data Public Private Forum



USE CASE HEALTHCARE
SECONDARY USAGE OF HEALTH DATA

Description

- Secondary usage of health data is defined as the aggregation, analysis and concise presentation of clinical, financial, administrative as well as other related health data
- in order to discover new valuable knowledge, for instance to identify trends, predict outcomes or influence patient care, drug development, or therapy choices.



Example Use Cases

- *Example 1: Patient recruiting and profiling* suitable for conducting clinical studies. Today, often clinical studies, in particular studies investigating rare diseases, fail due to the fact that not enough patients are available for conducting clinical studies (Berliner Forschungsplattform Gesundheit (BFG), Astra Zeneca)

- *Example 2: Social Media Based Medication Intelligence* Mining of health data on the discover insights about side effects of medications *(e.g. Treato)*

PEER ENERGY CLOUD

BIG
Big Data Public Private Forum

Smart grid pilot in Saarlouis
100 households

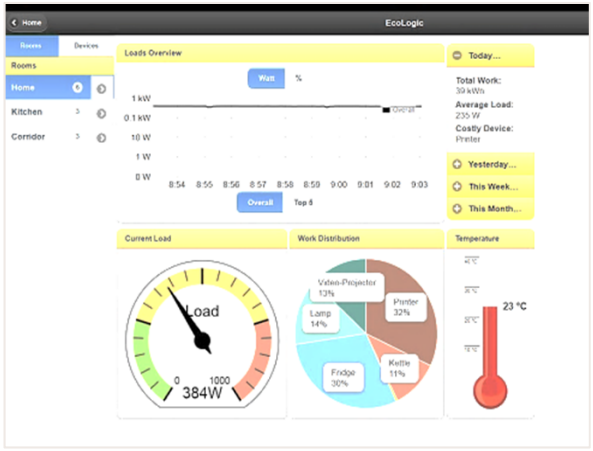
Supported by:

Federal Ministry
for Economic Affairs
and Energy

on the basis of a decision
by the German Bundestag

Innovation award
Germany
Land of Ideas
Selected Landmark 2012

peer energy cloud
AGT INTERNATIONAL
Stadtwerke Saarlouis
Mit Energie. Mit Sicherheit. Mit uns.
KIT Karlsruher Institut für Technologie
DFKI Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
SEEBURGER Business Integration



Engage consumers to optimally use local solar energy

- Understand consumption and save
- Trade solar energy in the neighborhood to balance the grid

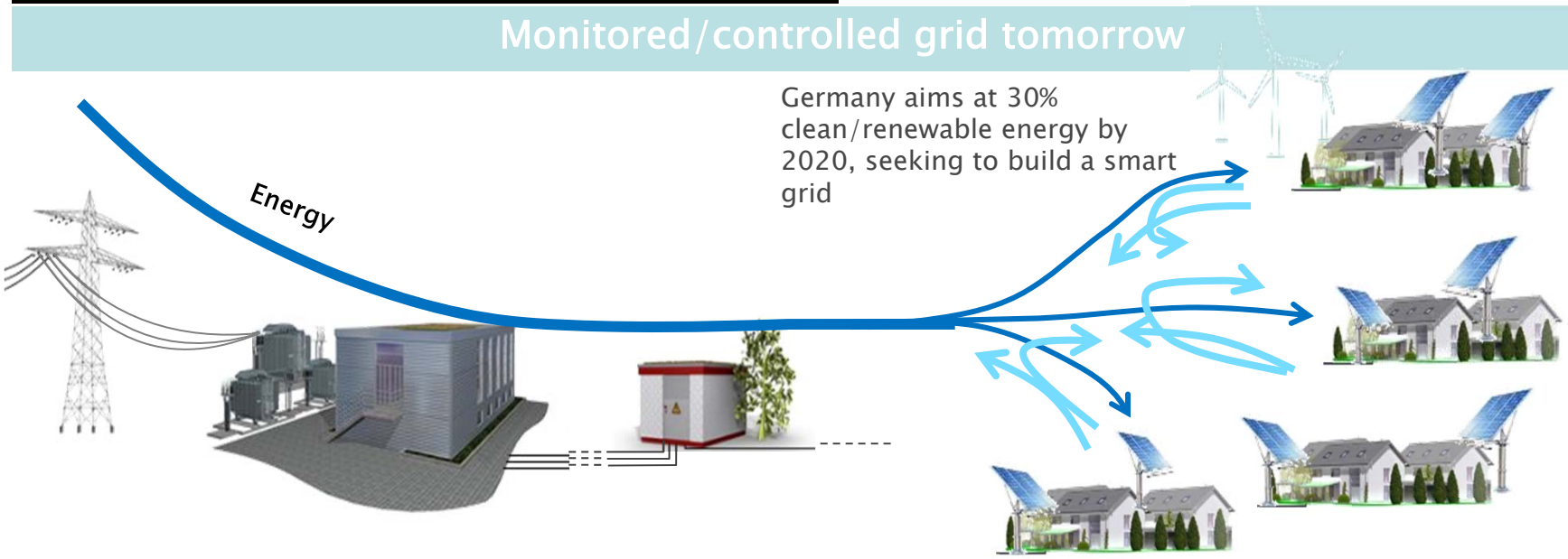
DEVICE LEVEL ENERGY MONITORING

peer energy cloud

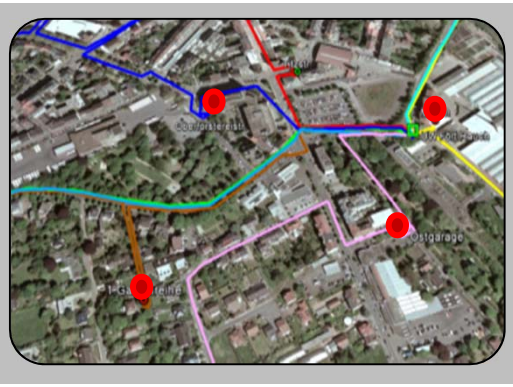
BIG
Big Data Public Private Forum

Monitored/controlled grid today

Monitored/controlled grid tomorrow



Sensors today



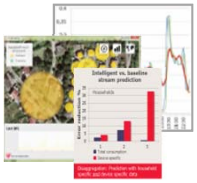
Sensors tomorrow *(consumer level)*

GETTING READY FOR DATA VOLUMES IN FUTURE GRIDS

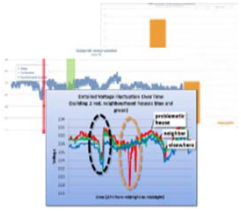
BIG
Big Data Public Private Forum

PeerEnergyCloud Pilots allows us to get ready for future data volumes today

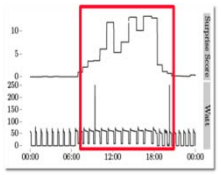
How much data is really needed for what?



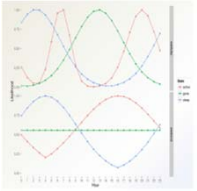
real-time analytics on mass data (grouped aggregation)



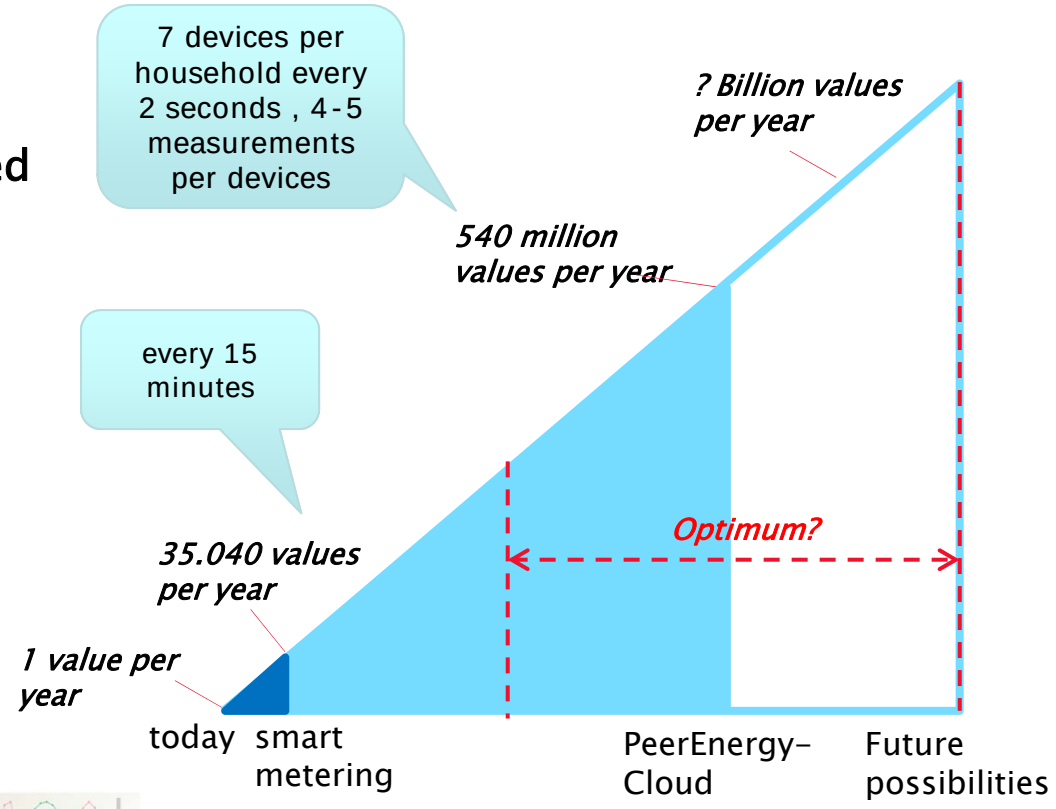
Scalable statistics over hundreds of millions of measurements



Automatic detection of load anomalies (spotting inefficiencies and defects)



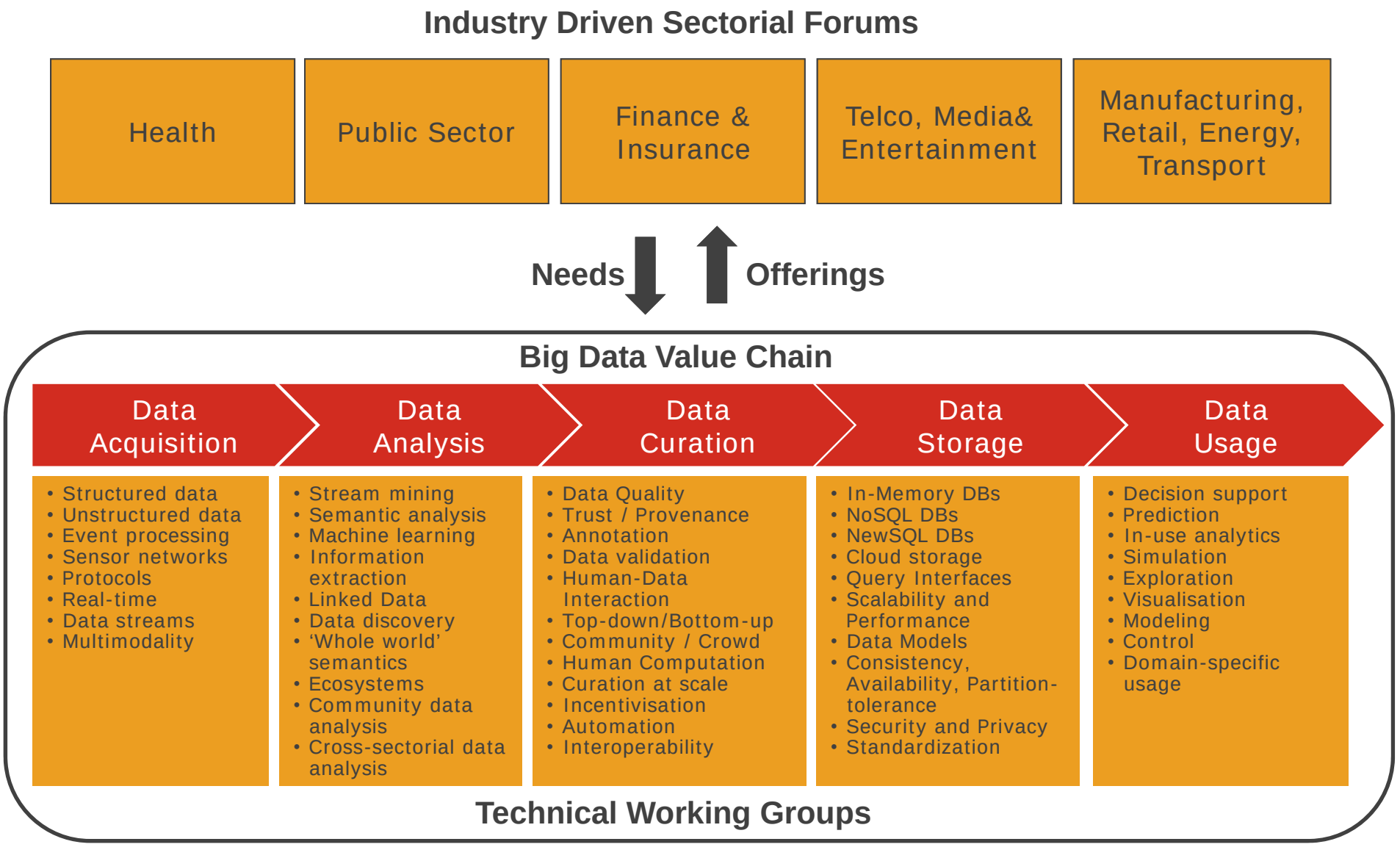
Household activity state inference and prediction



BIG METHODOLOGY

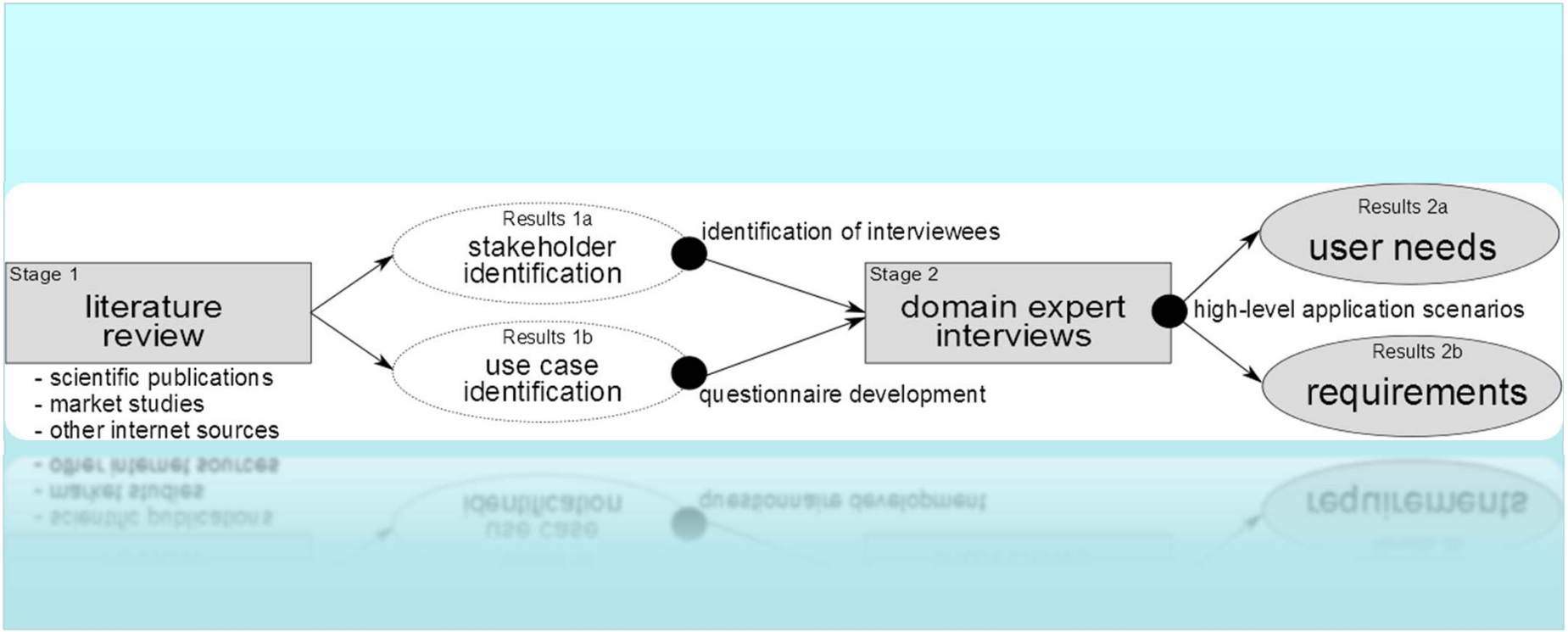
SECTORIAL FORUMS AND TECHNICAL WORKING GROUPS

BIG
Big Data Public Private Forum



SECTOR ANALYSIS METHODOLOGY

BIG
Big Data Public Private Forum



TECHNICAL WORKGROUP APPROACH

BIG
Big Data Public Private Forum

Methodology

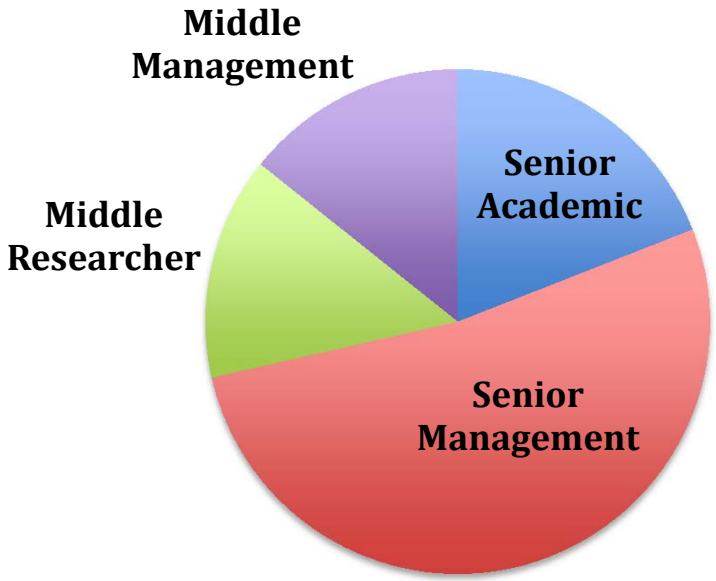
1. Literature & Technical Survey
2. Subject Matter Expert Interviews
3. Stakeholder Workshops
4. Online Questionnaire (with NESSI)

BIGDATAVALUeEU

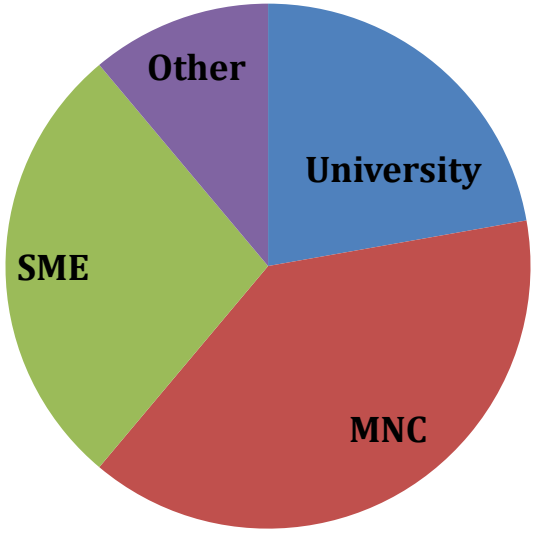
Target Interviewee

- Early adopters
- Business enablement
- Technical maturity
- Key Opinion Leaders

Interviewee Breakdown



Position in Organisation



Types of Organisations

TWG FINDINGS

KEY INSIGHTS

BIG
Big Data Public Private Forum

The Data Landscape

- Much of (Big Data) **technology is evolving evolutionary**
- But business processes **change must be revolutionary**
- ***Data variety and verifiability*** are key opportunities
- ***Long tail of data variety*** is a major shift in the data landscape

Biggest Blockers

- **Lack of Business-driven** Big Data strategies
- Need for format and data storage technology **standards**
- **Data exchange** between companies, institutions, individuals, etc.
- Regulations & markets for **data access**
- Human resources: Lack of skilled data scientists **and data engineers**

Key Trends

- Lower ***usability barrier for data tools***
- ***Blended human and algorithmic data processing*** for coping with for data quality
- Leveraging ***large communities (crowds)***
- Need for semantic ***standardized data representation***
- Significant increase in use of ***new data models*** (i.e. graph) (expressivity and flexibility)

DATA VALUE CHAIN - ANALYSIS

BIG
Big Data Public Private Forum

Data Acquisition | **Data Analysis** | Data Curation | Data Storage | Data Usage

Key Insights

- **Old technologies applied in a new context (Volume, Variety, Velocity)**
- **Need for**
  - Stream data mining
  - 'Good' data discovery
  - Techniques to deal with dealing with both very broad and very specific data
- **Features to Increase Take-up**
  - Simplicity including the 'democratisation of semantic technologies'
  - Ecosystems of tools
- **Communities and Big Data will be involved in new and interesting relationships**
  - In collection, improving data accuracy, analysis and usage
  - Improves community engagement
- **Cross-sectorial uses of Big Data will open up new business opportunities**

Social & Economic Impacts

- **Cross-sectorial businesses**
- **Data-cleaning today requires a lot of work – area for new business**
- **eHealth transformed from push (patient visits doctor) to pull (continuous monitoring)**
  - Also large scale trend analysis
  - Conduit between research and individual patient care
    - E.g. creation of patient avatars based on combination of specific data and generic research data
- **Proactive fraud detection even before it happens**
- **Engaged citizens create and have access to all data related to local, regional and national social and policy issues**

DATA VALUE CHAIN - CURATION

BIG
Big Data Public Private Forum

State of the Art

- **Master Data Management (MDM)**
  - Centralized, single point of reference for data representation
- **Collaboration platforms & Web 2.0 tools**
  - Wikis, Content Management Systems (CMS)
- **Early-stage crowdsourcing services**
  - E.g. Amazon Mechanical Turk

Use Case

- **Health and Life Sciences**
  - ChemSpider: Collaborative platform for data curation of chemical structures
  - Protein Data Bank: Data curation platform for 3D protein structures
  - FoldIt: Crowdsourcing game platform for protein folding
- **Telco, Media, Entertainment**
  - Press Association, Thomson Reuters, The New York Times
  - Use of semantic technologies to structure and categorize unstructured data (texts, images and videos), improving content accessibility and reuse
- **Retail**
  - Ebay: use of crowdsourcing services to improve product categorization
  - Unilever: use of crowdsourcing services for product feedback and sentiment analysis

DATA VALUE CHAIN - CURATION

BIG
Big Data Public Private Forum

Future Requirements

- Creation of incentives mechanisms for the maintenance and publication of curated datasets
- Definition of models for the data economy
- Understanding of social engagement mechanisms
- Reduction of the cost associated with the data curation task (scalability)
- Improvement of the Human-Data interaction aspects. Enabling domain experts and casual users to query, explore, transform and curate data
- Inclusion of trustworthiness mechanisms in data curation
- Integration and interoperability between data curation platforms / Standardization
- Investigation of theoretical and domain specific models for data curation
- Better integration between unstructured and structured data and tools

Emerging Trends

Incentives and social engagement
- Better recognition of the data curation role
- Understanding of social engagement mechanisms

Economic Models
- Pre-competitive and public-private partnerships

Curation at Scale
- Evolution human computation and crowdsourcing
- Instrumenting popular apps for data curation
- General-purpose data curation pipelines
- Human-data interaction

Trust
- Capture of data curation decisions & provenance management
- Fine-grained permission management models and tools

Standardization & interoperability
- Standardized data model and vocabularies
- Better integration between data curation tools

Data Curation Models
- Minimum information models
- Nanopulications
- Theoretical principles and domain-specific model

DATA VALUE CHAIN - STORAGE

BIG
Big Data Public Private Forum

Data Acquisition | Data Analysis | Data Curation | **Data Storage** | Data Usage

Key Insights

From state-of-the art

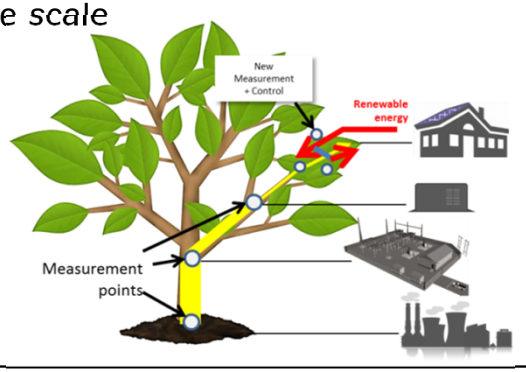
- Capability to **store and manage virtually unbounded volumes of data** has huge potential to transform society and business
- Better Scalability at Lower Operational Complexity and Costs
- Big Data Storage Has Become a Commodity Business.
- Maturity Has Reached Enterprise-Grade Level

Open Challenges

- Unclear Adoption Paths for Non-IT Based Sectors
- Lack of standards and best practices is major barrier for adoption
- Open Scalability Challenges
- Privacy and Security is Lacking Behind

Social & Economic Impacts

- **Enables a truly data-driven economy** that is able to manage a variety of data sets in an integrated way
- **Privacy and Security** becomes more important
- **Energy:** high resolution smart meter data contributes to the stable integration of renewable energies
- **Health Care:** storage is key enabling technology to solve integration challenges
- **Media:** enables using the voice of the crowd
- **Transport:** facilitates personalized and multi-modal on large scale



DATA VALUE CHAIN - STORAGE

BIG
Big Data Public Private Forum

Data Acquisition | Data Analysis | Data Curation | **Data Storage** | Data Usage

Future Requirements

- Standardization of Query Interfaces

Generic Graph API

- Legal Frameworks for data management
- Data Tracing and provenance in order to assess trust in data and ensure compliance
- Better support and scalability for storing semantic data models, e.g. in health care

Emerging Trends

- **Graph databases** will become more important
- **Columnar stores** experience wider adoptions as they are often faster in most practical applications
- **Convergence with analytical frameworks** Analytical databases for better performance and lower development complexity (Mahout, Spark, Hadoop/R, rasdaman, SciDB)
- **Data Hubs and Markets:** Hadoop-based solutions tend to become a central integration point for all enterprise, sectorial and cross-sectorial data
- **Smart Data:** only use relevant data in network and at the edge processessing

DATA VALUE CHAIN - USAGE

BIG
Big Data Public Private Forum

Data Acquisition → Data Analysis → Data Curation → Data Storage → **Data Usage**

Key Insights

- Key task of Data Usage is to **support business decisions**
- There is **great variety** of Big Data Applications. Some key areas are:
  - **Industry 4.0** (industrial internet)
  - **Predictive maintenance**
  - **Smart data** and **service integration**
- **Interactive exploration**: Big Data generates insights beyond existing models, new analysis interfaces must support browsing and modeling (visual analytics)
- Opportunities for **data markets:** integration with customer and supplier data, services from infrastructure (Iaas) to software (SaaS) to business processes (BPaaS) to knowledge (KaaS)

Social & Economic Impacts

- **Regulatory issues:**
  - **Data protection** and **privacy**
  - Ownership of **derived data**
- **Integration** of business processes and data
- Beyond single company; **data market places**
- **Transparency** through Data Usage impacts:
  - Economy
  - Society
  - Privacy
- **Current drivers** are big companies
- Opportunities for **SMEs** though services (integration) and data market places

USE CASE ANALYSIS

PUBLIC SECTOR

BIG
Big Data Public Private Forum

- What the Public sector is facing…
  - **Internal data silos** → no integration of systems across public bodies
  - **Raising pressure on productivity** → lacks productivity compared to the private sector
  - **Older workforce** → compared to the private sector, relies on a far older workforce
  - **Aging population** → increasing demand for medical and social services

- …and in what can Big data help:
  - Improving many areas in public sector services:
    - Strengthen collaboration among PS bodies
    - Social welfare
    - Citizen services
    - Improve healthcare
    - Government transparency and accountability
    - Public safety
  - **Open Data initiatives** → act as catalysts in the development of a data ecosystem through the opening of their own datasets, and actively managing their dissemination and use
  - **Analysis of sensor data**: Real time data processing to search for patterns and relationships and present real-time views on Smart Cities for better urban management and citizen engagement

PUBLIC SECTOR REQUIREMENTS
OVERCOMING THE DATA SILO CULTURE

BIG
Big Data Public Private Forum

Data ownership

Fragmentation of data ownership that leads to the data silo and interoperability issues

Common strategy at all levels of administration

So much energy is lost and will remain so until a common strategy is realized for the reuse of cross technology platforms

Data Privacy and Security

Legal framework supporting data access and usage of citizen's information

Legislative and political willingness

To promote legislation that allows the reuse of data for other purposes than those for what it was originally collected

Openness of Data

Leadership to promote common standards for Government Open Data: APIs, format and schemas, as well as for open data licensing

Data Sharing

Overcome data silos because of legacy technologies

Not-Technology-related	Regulation & Technology	Technology-related

HEALTHCARE REQUIREMENTS

BIG
Big Data Public Private Forum

High Investment

Long-term investments require conjoint engagement of several partners

Data Digitalization

only small percentage of data is documented (lack of time) with low quality

Data Security

Legal processes for data sharing & communication are needed

Value-based system incentives

Current incentives enforce “high number” instead of “high quality” of care services

Semantic Annotation

transform unstructured data into structured format

Data Quality

Reliable insights for health-related decisions require high data quality

Business Cases

Undiscovered und unclaimed potential business values

Data Sharing

Overcome data silos and inflexible interfaces

Not-Technology-related	Regulation & Technology	Technology-related

DATA POOLS IN HEALTHCARE

MAIN IMPACT BY INTEGRATING VARIOUS AND HETEROGENEOUS DATA SOURCES

BIG
Big Data Public Private Forum

Patient Behaviour & Sentiment Data

- Owned by consumers or monitoring device producer
- Encompass any information related to the patient behaviours and preferences

Pharmaceutical & R&D Data

- Owned by the pharmaceutical companies, research labs/academia, government
- Encompass clinical trials, clinical studies, population and disease data, etc.

Health data on the web

- Mainly open source
- Examples are websites such as PatientLikeMe, Linked Open Data, etc.

Highest Impact on integrated data sets

Clinical Data

- Owned by providers (such as hospitals, care centers, physicians, etc.)
- Encompass any information stored within the classical hospital information systems or EHR, such as medical records, medical images, lab results, genetic data, etc.

Claims, Cost & Administrative Data

- Owned by providers and payors
- Encompass any data sets relevant for reimbursement issues, such as utilization of care, cost estimates, claims, etc.

ENERGY SECTOR REQUISITES
STAKEHOLDERS, USE CASES, DATA SOURCES

BIG
Big Data Public Private Forum

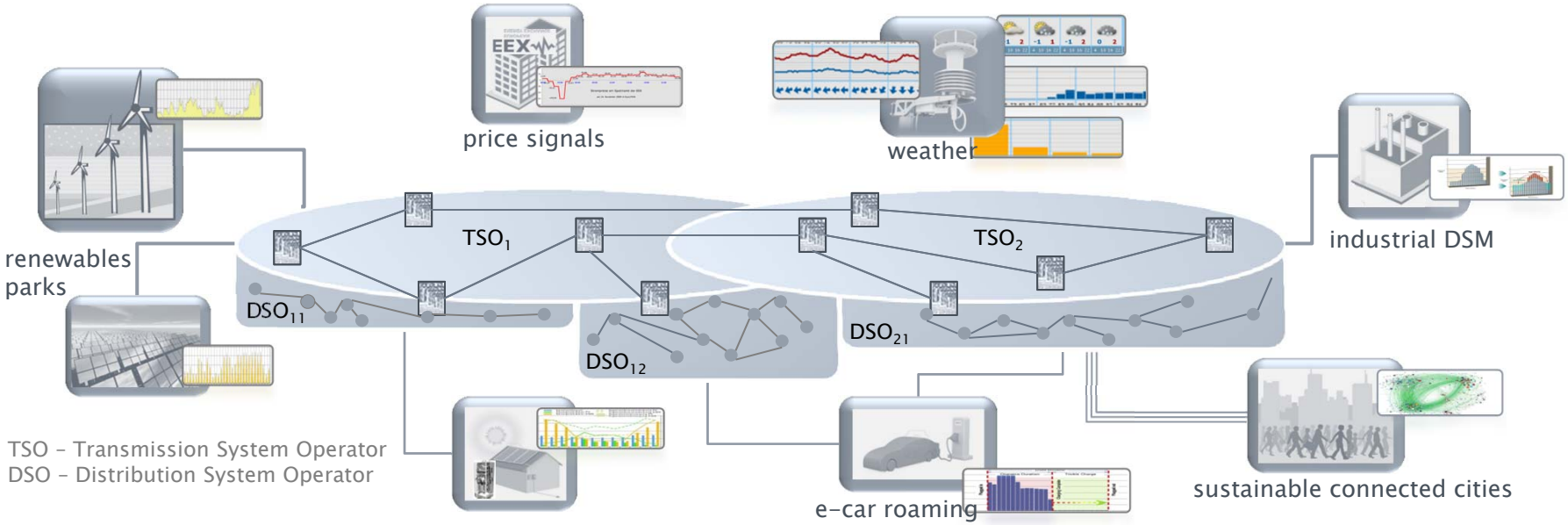
New business requires the combined value creation on mass data coming from a variety of data sources, once the volume and velocity of energy data is mastered

NETWORK OPERATORS

Increased situational awareness
- intermittent, decentralized supply
- new responsive demand
- network asset and weather conditions

RETAIL/WHOLESALE

Efficient portfolio management
- detailed data on market prices,
- actual energy usage, feed-in
- weather conditions



END USER ENERGY MANAGEMENT

Flexible energy tariffs for prosumers and eCars
- Actual energy usage instead standardized profiles

ENERGY SECTOR REQUIREMENTS

BIG
Big Data Public Private Forum

Investment in communication and connectedness

Broadband communication or ICT in general, needs to be widely available alongside energy and transportation infrastructure such that real-time data access is given

Standardization & Open Data

Open data is a great opportunity, but, standardization is required. Regarding data models, representation, as well as protocols practical migration paths

Skilled people

programming, statistical tools for example needs to be part of *engineering education* until big data technology becomes more business user-friendly or in case this becomes the “new normal”

Data Access

Dynamic configurability of data access of which data can be collected for what purpose in what granularity and time span and location

Privacy and confidentiality preserving data analytics are required to enable the service provider to retrieve the knowledge without violating the agreed upon granularity

A digitally united European Union

European stakeholders require reliable *minimally consistent* rules and regulation regarding digital rights and regulations, whilst making use of all technological enhancements

Abstraction

from the actual big data infrastructure is required to enable (a) *ease of use* and (b) extensibility and flexibility

Adaptive models of data and system model

new knowledge extracted from domain analytics or the ever changing circumstances of the system can be *redeployed* into the data analytics framework

Data interpretability & analytics

Expert and domain know-how must be blended into the data management and analytics. *Data analytics* is required as *part of every step* from data acquisition to data management to data usage. *Fast and even real-time* analytics is required to support decisions, which need to be made in ever shorter time spans

Not-Technology-related | **Regulation & Technology** | **Technology-related**

CONCLUSIONS

(SOME) COMMON REQUIREMENTS

BIG
Big Data Public Private Forum

High Investment

Long-term investments require conjoint engagement of several partners

Data ownership

Fragmentation of data ownership that leads to the data silo and interoperability issues

Business Cases

Undiscovered und unclaimed potential business values

Data Privacy and Security

Legal frameowrks for data sharing & communication are needed

Data Quality

Reliable insights for health-related decisions require high data quality

Openness of Data

Leadership to promote common standards for Open Data: APIs, format and schemas, as well as covering licensing and legal aspects

Data Digitalization

only small percentage of data is documented (lack of time) with low quality

Semantic Annotation

transform unstructured data into structured format

Data Sharing

Overcome data silos and inflexible interfaces

Not-Technology-related | **Regulation & Technology** | **Technology-related**

TECHNOLOGY ROADMAP DEVELOPMENT

BIG
Big Data Public Private Forum

Observations & Learnings from sector and technical investigations

- **Large Gap** between needs mentioned by the stakeholder and users of the (health) sector and the technological opportunities envisioned by the technical groups
  - **Technical requirements** mentioned within sectorial interviews mainly relate to efficient data management approaches
  - **Technical opportunities** mentioned by the technical groups highlight opportunities within a world of open data access, such a the web
- **Challenge**: How to align the two perspectives?
  - First Step "*Focus on big data readiness*": Any technological requirements, such as efficient data management, need to be addressed /solved before big data capabilities can be implemented (enabling technologies)
  - Second Step: "*Elaborate big data opportunities*": Develop transitional scenarios that could be realized assuming that the sector has achieved big data readiness (big data technologies), scenarios should generate value.

We need to distinguish between

Enabling Technologies

- Sector-specific data management technologies that need to be in place before any big data scenario can be realized
- Ensure that the relevant data of the sector is available (e.g. EHR, IED)
- Driven by **user / need driven (sector) analysis**
- Focus on domain-specific requirements of data management and related research questions

Big data opportunities

- advanced/big IT capabilities that help to improve healthcare delivery
- **Data-driven**: Investigate in public available health data sources as basis for use case brainstorming
- **Technology-Driven**: Investigate to which extent applications /technologies from other domains can be transferred to the healthcare sector

TOWARDS A BIG DATA ROADMAP

BIG
Big Data Public Private Forum

Scalable Storage

Distributed Data Acquistion

Secure Data Sharing

Encryption

Anonymisation

Scalable Storage

Semantic enrichment

Analytical Databases

Complex Event Processing

Analytical Platforms

Semantic Processing

Distributed Processing

Explorative Analysis

Media & Entertainment

Public Sector

Telco

Energy

Domain Specific Services

Manufacturing

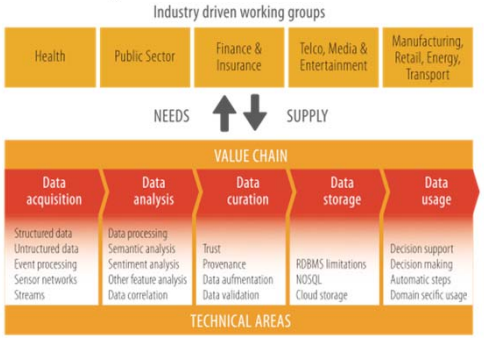
Health

Retail

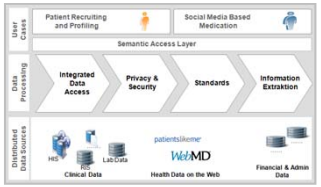
SUMMARY

BIG
Big Data Public Private Forum

- BIG is contributing to the **creation of an industry-lead Big Data community in Europe**
  - Sector and Data Value Chain Analysis
  - Initial results in SRIA
  - Next steps: roadmap development, stakeholder platform



- **3 Data Sharing Use Cases** Introduced



- **Data Sharing is a key requirement** for driving data driven-business on a larger scale and evolve technologies
  - Requires addressing issues across **business** (value), **regulatory** (society) and **technical** (principled engineering and development of new capabilities)
  - Initiatives such as PPP in combination with national will help address these issues

BIGDATAVALUeEU

http://www.bigdatavalue.eu

Thank you

Dr. Martin Strohbach
Senior Researcher, AGT International
Technical Lead Storage WG, BIG
mstrohbach@agtinternaitonal.com

Tilman Becker (DFKI, Data Usage), **Edward Curry** (NUI Galway, Data Curation), **John Domnique** (STI, Data Analysis), **Ricard Munné** (ATOS, Public Sector), **Sebnem Rusitschka** (Siemens, Energy and Transport), **Holger Ziekow** (AGT, PEC), **Sonja Zillner** (Siemens, Health)

BIGDATAVALUeEU

http://www.bigdatavalue.eu

BIG
Big Data Public Private Forum

http://www.big-project.eu