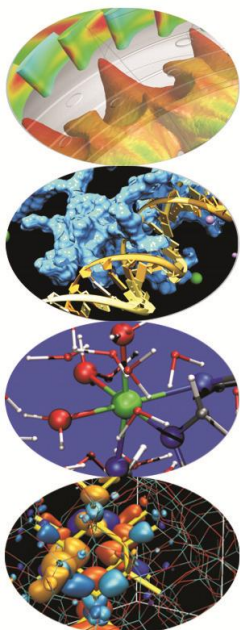


5 Aprile 2017

Soluzioni innovative per il 4.0

Big Data e Data Analytics

Gabriella Scipione, CINECA
Roberta Turra, CINECA



Connessione

Comunicazione Dispositivi e
Raccolta Dati

1

Internet of
Things



Cloud



Big Data



Analytics



Smart Factory



Augmented
Reality



Smart
Devices

2

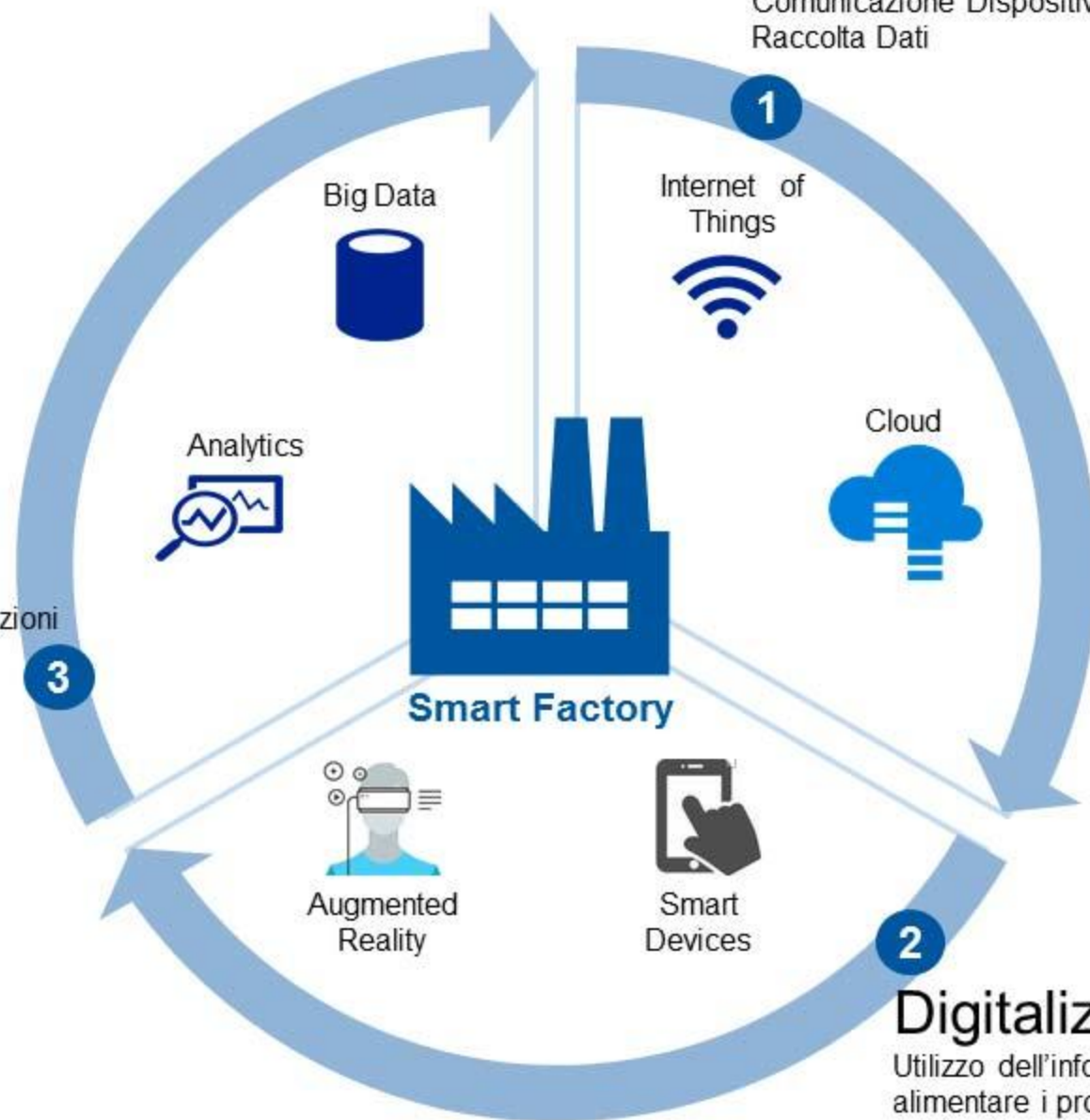
Digitalizzazione

Utilizzo dell'informazione per
alimentare i processi decisionali

Intelligence

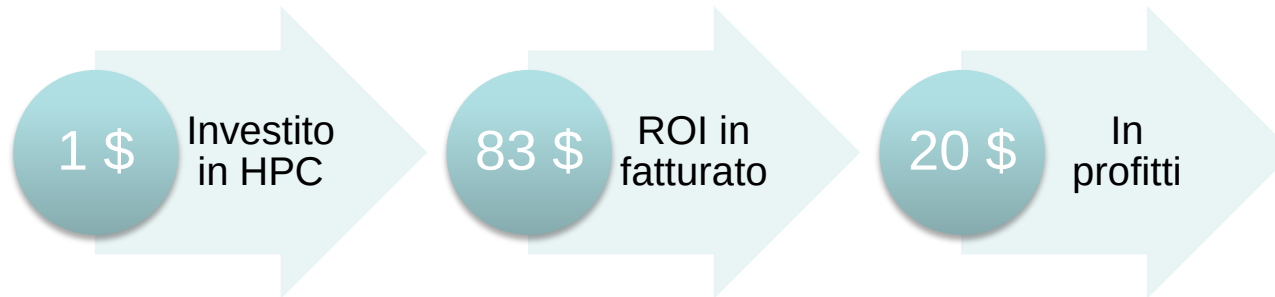
Gestione dei dati e
trasformazione in informazioni
a valore per il business

3

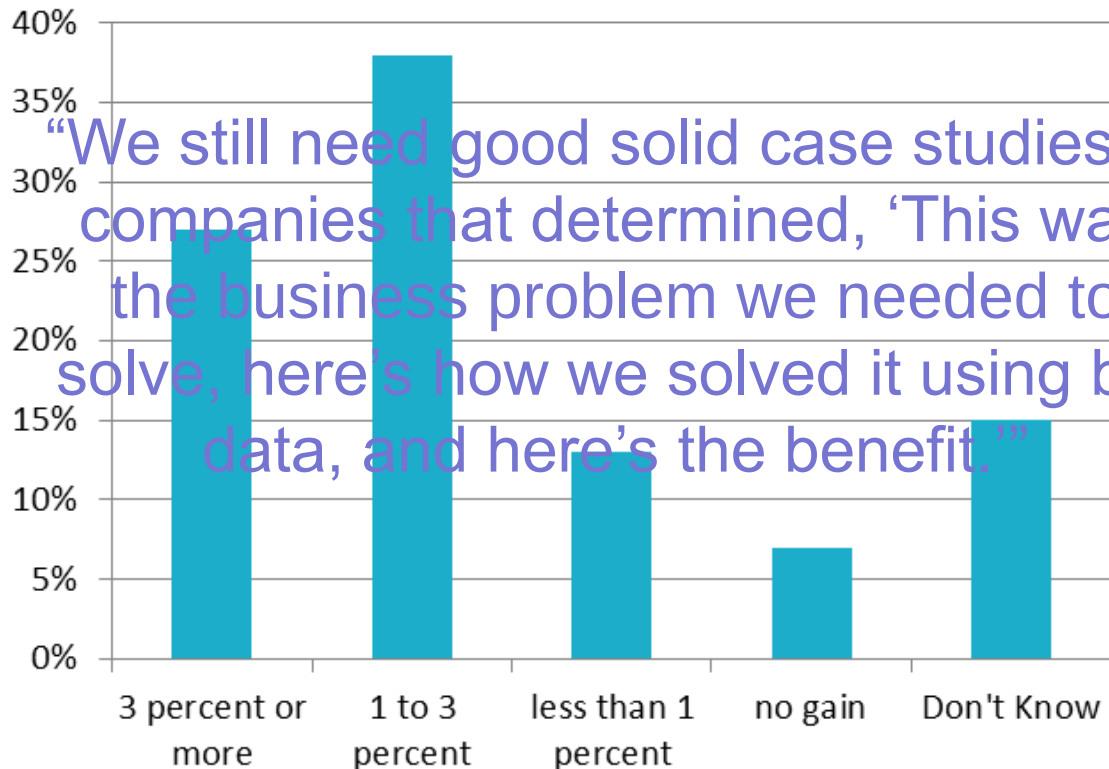


Big data analytic's impact on revenues according to Teradata's survey

HPC

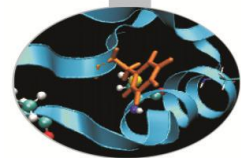


**Big
Data**



“We still need good solid case studies of companies that determined, ‘This was the business problem we needed to solve, here’s how we solved it using big data, and here’s the benefit.’”

Quali sono i problemi industriali che Big Data e Data Analytics permettono di affrontare

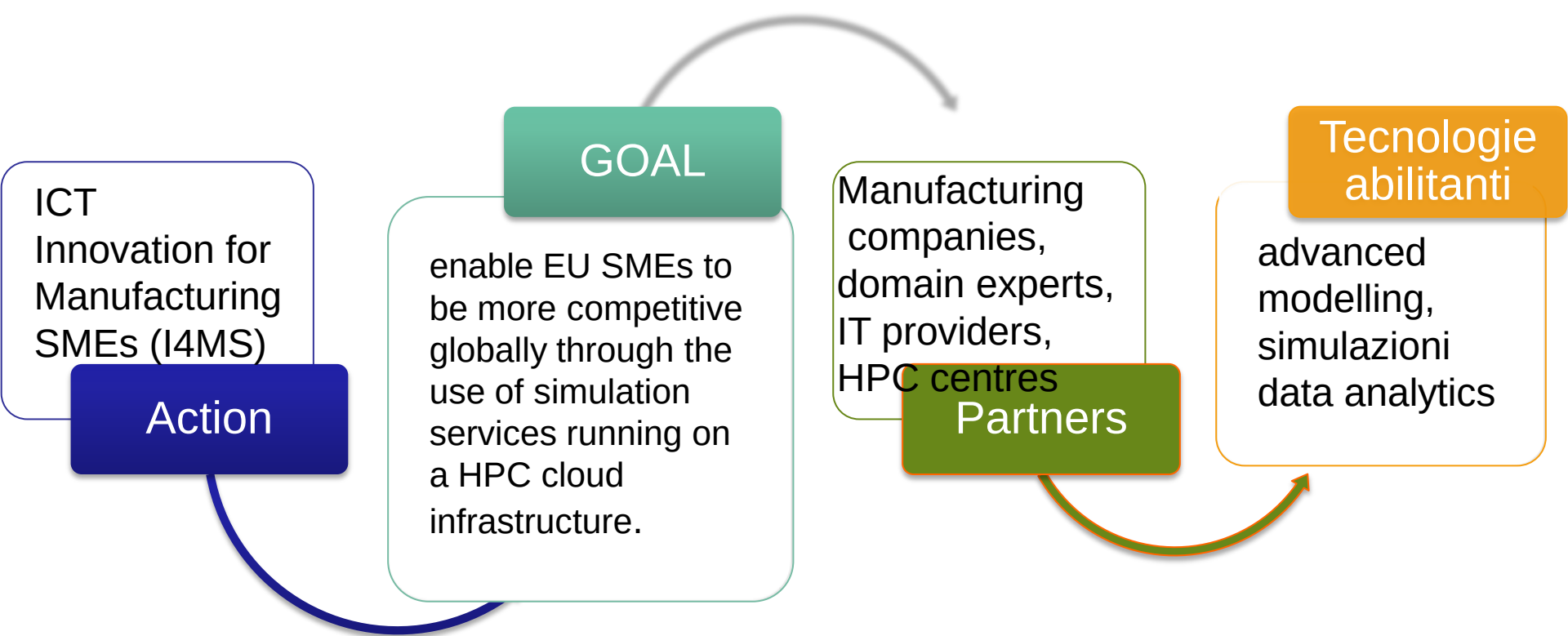


Assicurativo



**Automobilistico /
Manifatturiero**

Solutions and services for manufacturing SMEs to help grow your business.





FORTISSIMO



**PRESERVE: PREdictive diagnosis
SERvices for automotIVE industry**

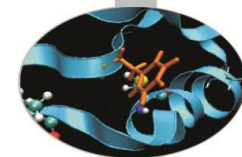
TEXA: specializzazione nella progettazione e costo delle realizzazioni e gestioni di filiali diagnostiche per i sistemi estensivi e i periodi di garanzia

CINECA: coinvolgimento di specialisti in HPC:

- per diminuire i tempi di elaborazione
- per miglioramento degli algoritmi per l'estrazione informazioni dai dati e trattamento dei dati stessi e per le analisi predittive



UNIPOL



Simulazione



Simulazione dell'evoluzione del
market risk



Solvency II, Art. 44 (1) stabilisce
che le imprese di assicurazione e
di riassicurazione abbiano in atto
un sistema di gestione dei rischi

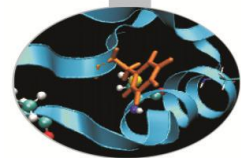
Big Data Lab



- Personalizzazioni delle Polizze assicurative per gli utenti
- Diminuzione delle frodi assicurative



Sviluppo di PoC
Presidio Tecnologico



Le PMI e le principali sfide

Partire può essere
costoso

Mancano le storie
di successo che
aiutini a convincere
i decision maker
dell'azienda

Il progetto viene a
collocarsi nella
voce di costo
dell'azienda (R&D)

Le competenze
sono difficili da
trovare

È difficile scegliere
ed accedere alle
giuste risorse di
calcolo e software

CINECA è un consorzio interuniversitario senza scopo di lucro al servizio della ricerca nazionale.

Nasce nel '69 con lo scopo di promuovere l'utilizzo dei più avanzati sistemi di elaborazione dell'informazione a favore della ricerca tecnologica e scientifica.

Punto di riferimento per il sistema accademico nazionale.
...ma anche per le imprese

CINECA con la sua infrastruttura si posiziona:

- MARCONI 12 posto Top500 per sistemi di calcolo dedicati alla ricerca e big data
- 6 posto dopo Cina, USA, ...



SUPERCOMPUTER MARCONI

New Tier-0

MARCONI

Model: Lenovo NeXtScale
Architecture: Intel OmniPath Cluster

Configuration 2016
Nodes: 1512 (BDW) + 3600 (KNL)
Processors: 2 x 18 cores Intel Broadwell @ 2.30 GHz, 54432 cores
1 x 68 cores Intel KnightsLanding @ 1.40 GHz, 244800 cores

Configuration 2017
Nodes: 3024 (SKL) + 3600 (KNL)
Processors: 1 x 68 cores Intel KnightsLanding @ 1.40 GHz, 244800 cores
2 x ≥20 cores Intel SkyLake @ ~2 GHz, ≥ 120960 cores

Internal Network: Intel OmniPath
Disk Space: >20PB (raw) of local storage
Peak Performance: about 20 PFlop/s

Disk Space: > 20 PB
Potenza di picco di 13Pflop/s

National Tier-1
CAHED - IBM NeXtScale Cluster

CERN LHC produce circa 15 PB di dati all'anno.
Luglio 2012 CERN ha prodotto 200 PB di data dagli 800 trillioni di collisioni per arrivare alla scoperta del bosone di Higgs

PICO - IBM NeXtScale Cluster

- 80 computing nodes
- thin/fat nodes 128/512 GB RAM
- hadoop and map reduce
- data insight
- remote visualization
- cloud computing

Data repository, curation and preservation

MULTI TIER STORAGE

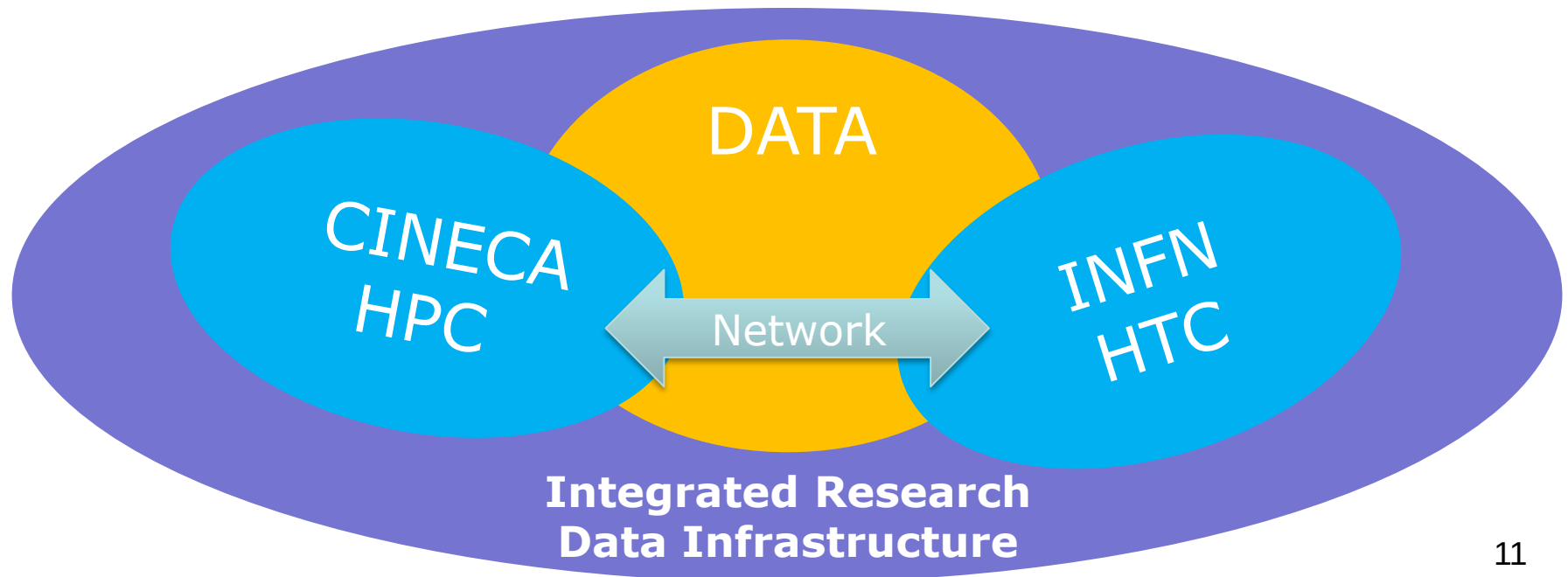
- 40 TB fast (SSD) storage
- 5 PB GSS storage
- 12 PB TAPE storage, integrated with GSS through LTFS

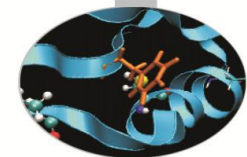
Infrastruttura Nazionale Big Data

Science Park Bologna

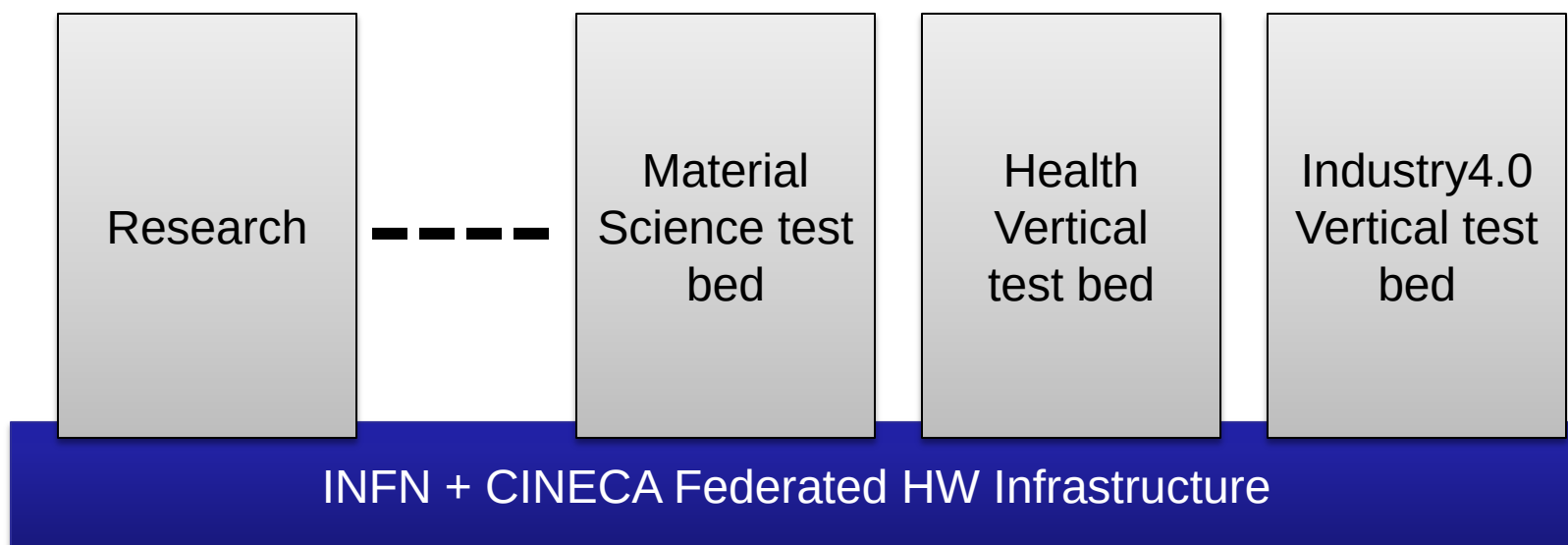
Integrazione dell'infrastruttura computazionale CINECA-HPC e INFN-HTC per fornire servizi per :

- Ricerca istituzionale di base e applicata
- P.A.
- Proof of Concept e innovazione per industrie e organizzazioni private





ER Big-data Framework Project



The European perspective: EuroHPC



DIGITAL DAY

Rome, 23 March 2017

Who

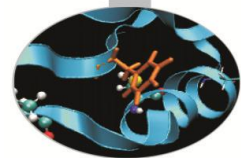
- 7 paesi europei (Italia, Francia, Germania, Lussemburgo, Paesi Bassi, Portogallo e Spagna)

What

- firmano la dichiarazione sul "supercalcolo",
- creazione EuroHPC, che sarà messa a disposizione delle comunità scientifiche, dell'industria e del settore pubblico in tutta l'UE

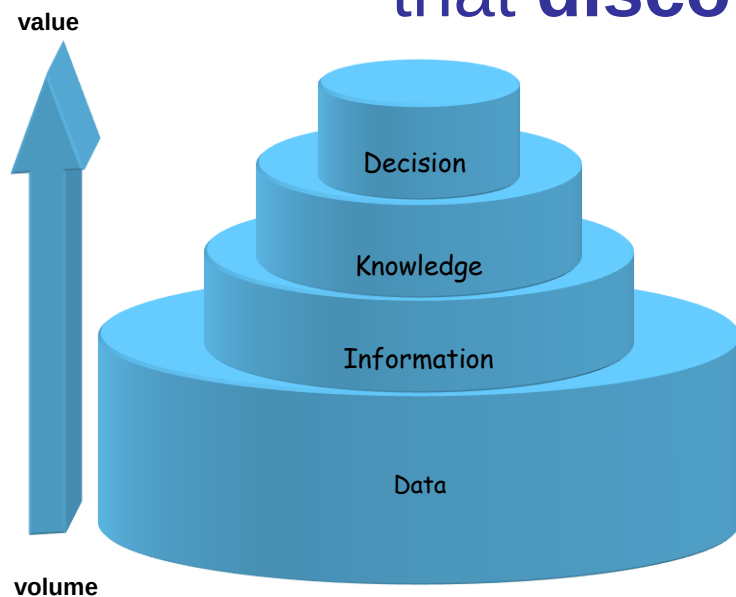
Why

- che potrà accrescere le capacità scientifiche e la competitività industriale dell'Europa.

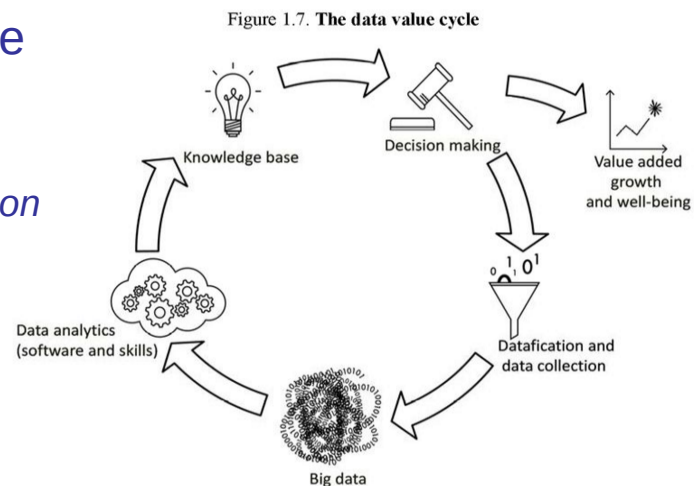


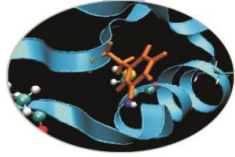
Data analytics

The **process** of extracting useful insights from raw data using algorithms that **discover** hidden patterns

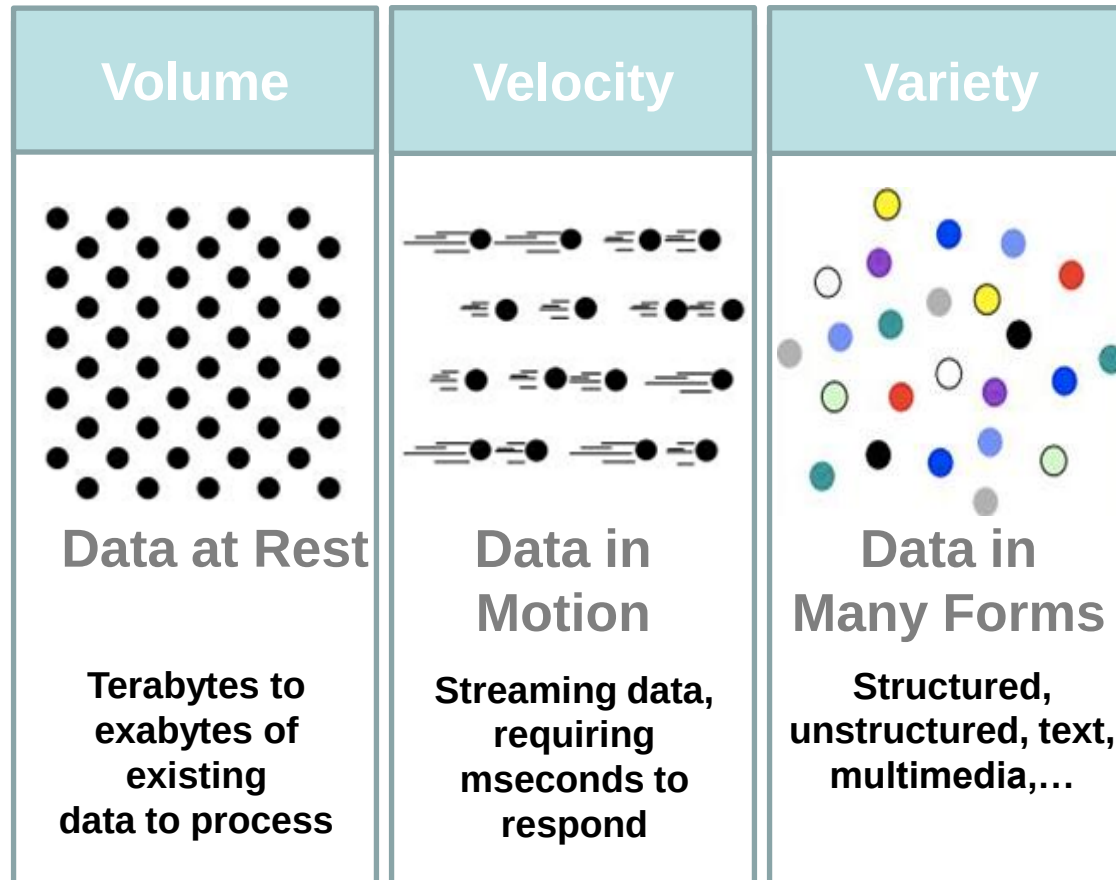


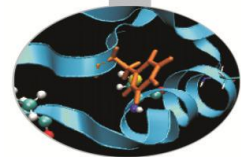
A step in the data value cycle
(OECD report on Data Driven Innovation)





Why is it challenging





Data typologies

structured data

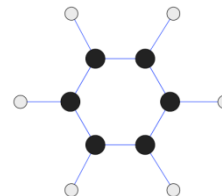
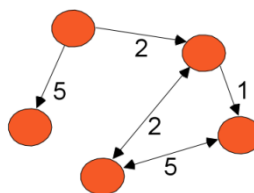
- data matrix
- transactional data

TID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

graph

- web and social networks
- molecular structures



ordinal data

spatial data

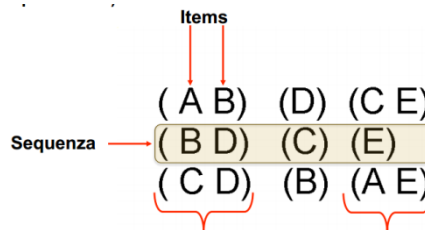
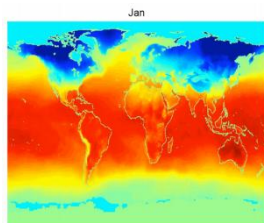
time series

sequences

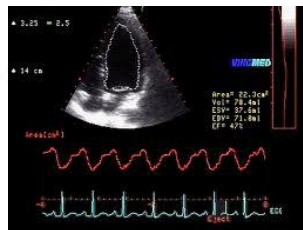
- genetic sequences

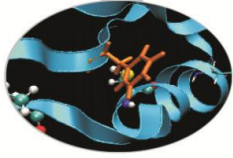
unstructured data

- textual documents
- images
- audio and videos (multimodal)



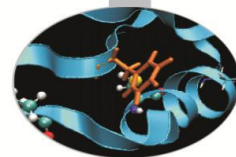
GGTTCGCGCTTCAGCCCCGCGCC
 CGCAGGGCCCCCGCCGCGCGTC
 GAGAAGGGCCCCGCTGGCGGGCG
 GGGGGAGGCGGGCGCCCGAGC
 CCAACCGAGTCCGACCAGGTGCC
 CCCTCTGCTCGGCCCTAGACCTGA
 GCTCATTAGGCGGCAGCGGACAG
 GCCAAGTAGAACACGCGAAGCGC
 TGGGCTGCCTGCTGCGACAGGG





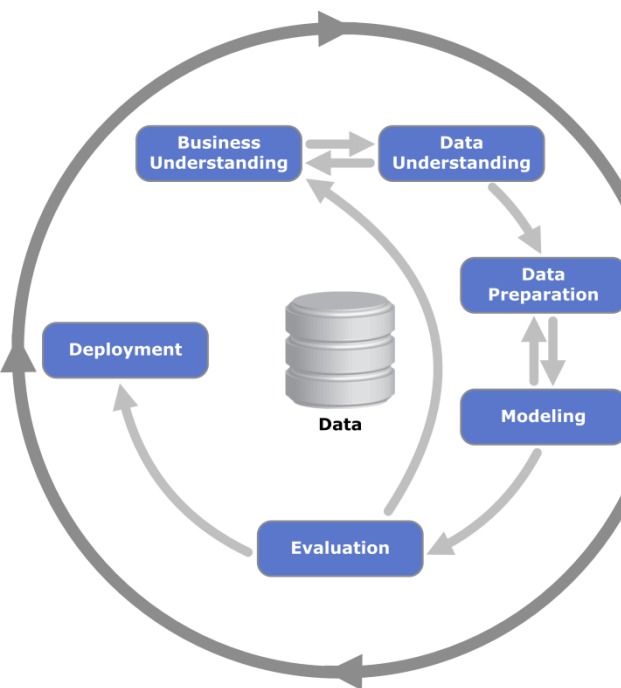
Data as an infrastructure

Data has become the key infrastructure for 21st century knowledge economies. Data are not the “new oil”, they are rather an infrastructure and capital good that can be used across society for a theoretically unlimited range of productive purposes, without being depleted.

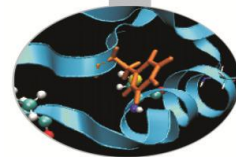


CRISP-DM reference model

Cross Industry Standard Process for Data Mining



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives <i>Background Business Objectives Business Success Criteria</i>	Collect Initial Data <i>Initial Data Collection Report</i>	Select Data <i>Rationale for Inclusion/ Exclusion</i>	Select Modeling Techniques <i>Modeling Technique Modeling Assumptions</i>	Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</i>	Plan Deployment <i>Deployment Plan</i>
Assess Situation <i>Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</i>	Describe Data <i>Data Description Report</i>	Clean Data <i>Data Cleaning Report</i>	Generate Test Design <i>Test Design</i>	Review Process <i>Review of Process</i>	Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i>
Determine Data Mining Goals <i>Data Mining Goals Data Mining Success Criteria</i>	Explore Data <i>Data Exploration Report</i>	Construct Data <i>Derived Attributes Generated Records</i>	Build Model <i>Parameter Settings Models Model Descriptions</i>	Determine Next Steps <i>List of Possible Actions Decision</i>	Produce Final Report <i>Final Report Final Presentation</i>
Produce Project Plan <i>Project Plan Initial Assessment of Tools and Techniques</i>	Verify Data Quality <i>Data Quality Report</i>	Integrate Data <i>Merged Data</i>	Assess Model <i>Model Assessment Revised Parameter Settings</i>	Review Project <i>Experience Documentation</i>	
		Format Data <i>Reformatted Data</i>			
		<i>Dataset Dataset Description</i>			

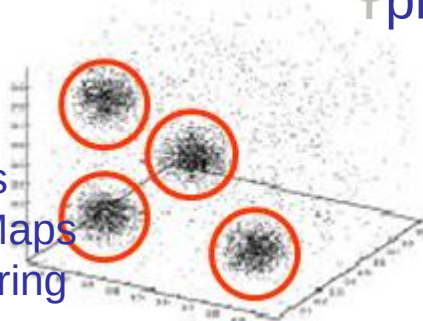


Tasks and techniques

descriptive

clustering

- k-means
- relational analysis
- Self Organizing Maps
- hierachical clustering
- mixture model
- ...



association rules

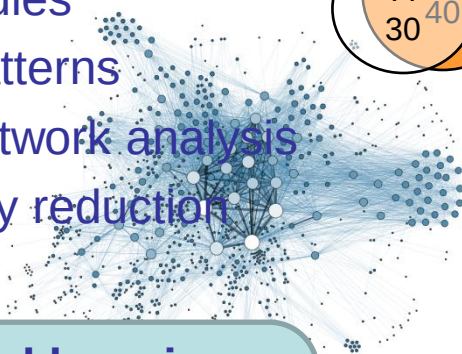
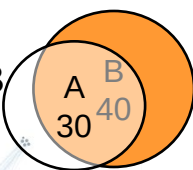
sequential patterns

graph and network analysis

dimensionality reduction

...

$$A \Rightarrow B$$



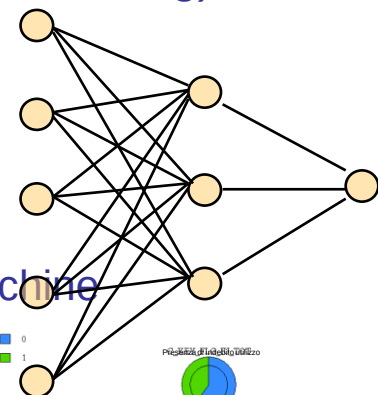
Unsupervised learning

training samples have no class information
guess classes or clusters in the data
 we are given inputs but no outputs
 (unlabeled data)
 we learn the "latent" labels

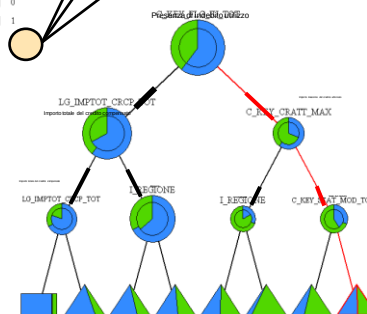
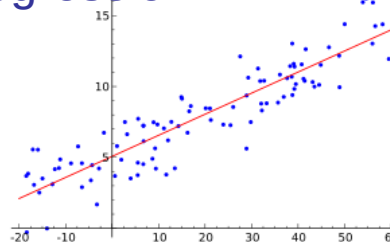
predictive

classification (machine learning)

- Naive Bayes
- Decision Trees
- Neural Networks
- KNN
- Rocchio
- Support Vectors Machine
- ...



regression



Supervised learning

use training samples with known classes
to classify new data
 we are given examples of inputs and
 associated outputs
 we learn the relationship between them

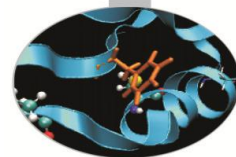
Caso Alstom

Obiettivo

Valutare in modo automatico se una segnalazione di guasto attivata dal sistema di monitoraggio, sia effettivamente da segnalare all'area di manutenzione

Tecniche di Machine Learning utilizzate

- Decision Tree
- Random Forest
- Neural Networks
- Gradient Boosting Machine



Variabili per ogni evento osservato: 300 var. delle 2.000 disponibili (eventi di diagnostica)

last_run_id	ISS 1	CAB_A_1	SERV_T1	EB_ANTENN	...	2	CAB_A_2	SERV_T2	EB_ANTENN	...	3	CAB_A_3	SERV_T3	EB_ANTENN	verofa
	UE	ON	ETTO	A_SWITCH			ON	ETTO	A_SWITCH			ON	ETTO	A_SWITCH	so
1559453	1	30	50	25		6	41	28			29	51	17		0
1561388	2	17	55	16		23	33	28			12	52	25		1
1561966	1	13	67	11		30	50	26			5	41	27		1
1593270	3	15	67	14		29	45	24			7	45	28		1
1656659	2	16	72	30		27	43	21			9	49	32		0
1656661	2	16	72	32		27	43	21			9	49	32		0
1656676	1	21	72	47		27	43	21			9	49	32		0
1699514	1	19	97	12		22	80	13			20	83	15		1
1704569	1	13	66	15		16	103	14			24	56	15		1
1748299	1	23	78	10		26	80	15			14	40	14		0
1783005	1	32	42	16		17	61	24			15	108	14		0
1817617	1	27	67	13		21	42	11			16	52	10		0
1653170	1	20	35	18		37	66	28			39	32	17		1
1658885	1	23	61	12		18	30	14			43	69	36		1

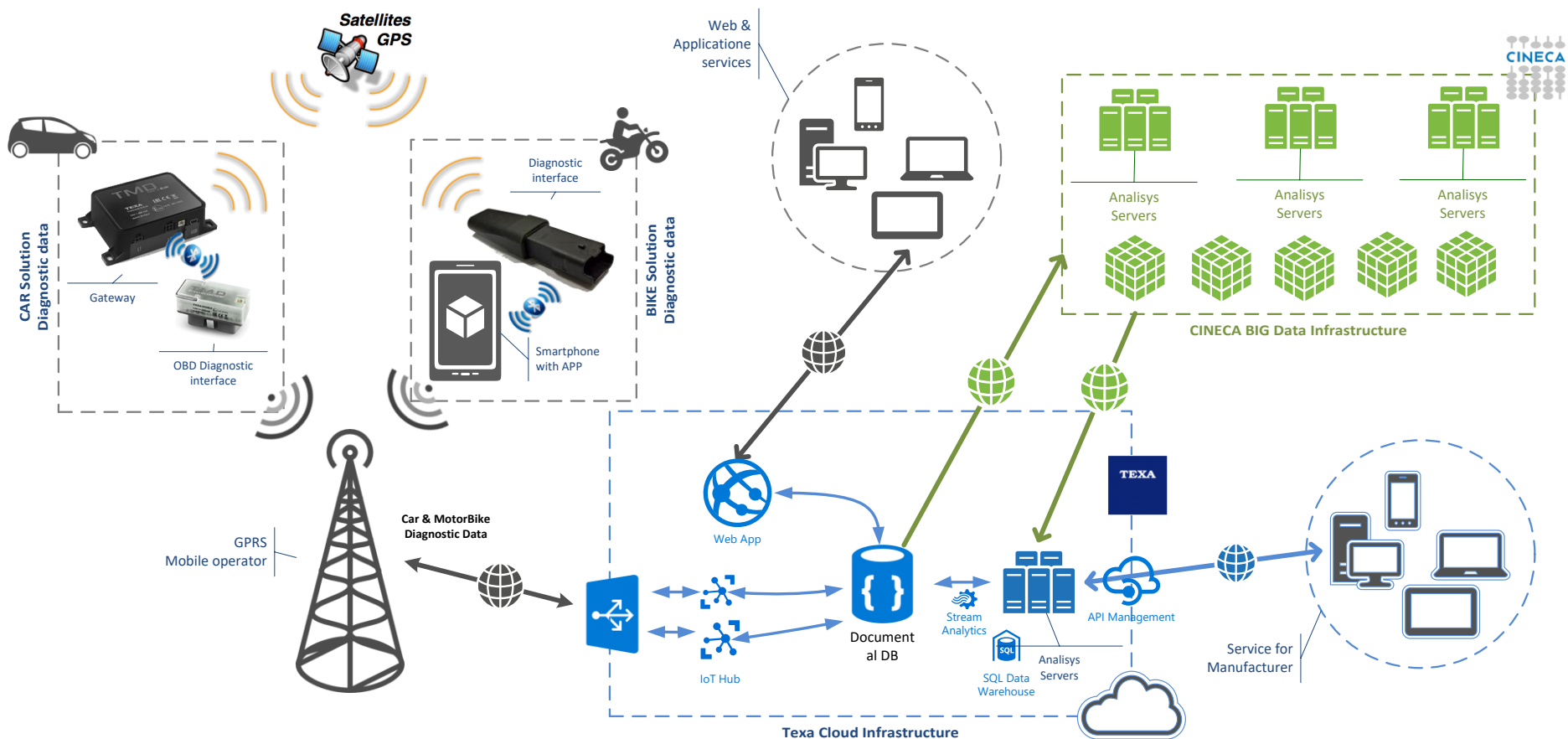
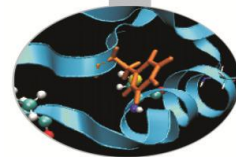
all'evento

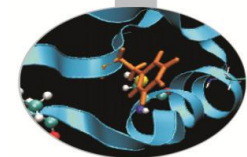
500 Km dall'evento

1000 Km dall'evento

Risultati riduzione fino al 25% delle false segnalazioni di guasti, che potrebbe tradursi in una riduzione dell'impiego del personale di manutenzione
 Sfruttando le medesime informazioni si è poi riusciti ad individuare possibili guasti in alcune apparati di terra (boe di segnalazione di transito del treno)

Caso Texa



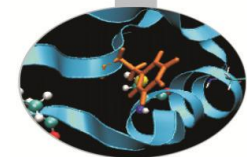


Modalità di accesso

Accesso alle risorse di calcolo
e a competenze di tipo
computazionale e/o di data
science (Machine Learning)

- Convenzioni / accordi quadro
- Finanziamento da progetti EU (Fortissimo)
- Open Innovation (progetto di ricerca congiunta co-finanziato)

Training in 2016



20 courses – 37 editions
including
4 schools – 5 editions
2 workshops
761 students

50
teachers
150
days of lectures
761
post-graduated
students

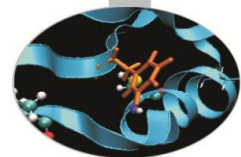
20
courses
4
schools
2
workshop
37
editions
5
editions

Alternating training

Stage and
traineeships



Training in 2017



Schools:

- 26th Summer School on Parallel Computing
- 13th Advanced School on Parallel Computing (*)
- 3rd School on Scientific Data Analytics and Visualization (*)
- 13th Advanced School on Computer Graphics for Cultural Heritage

Courses:



- Introduction to modern Fortran
- Introduction to Scientific and Technical Computing in C (*)
- Introduction to Scientific and Technical Computing in C++
- Python for computational science
- Debugging and Optimization of Scientific Applications (*)
- Introduction to Parallel Computing with MPI and OpenMP
- HPC Numerical Libraries
- Programming paradigms for GPU devices
- Tools and techniques for massive data analysis
- Parallel I/O and management of large scientific data
- High Performance Molecular Dynamics
- Scientific Visualization for Computational Chemistry
- Material Science codes on innovative HPC architectures: targeting exascale (*)
- High Performance Bioinformatics (*)
- Introduction to R for data analytics

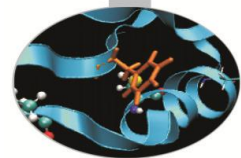


Workshop:



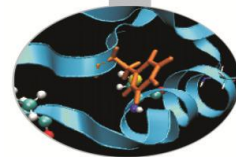
- HPC Methods for Engineering Applications (*)
- HPC methods for Computational Fluid Dynamics and Astrophysics (*)
- Introduction to Marconi KNL cluster, for users and developers (*)
- Introduction to PRACE-PCP pilot (*)

(*) are in the PATC (Prace Advanced Training Center)



Training in collaborazione

- BBS – Open Program Big Data Analytics
- BBS – Master in Data Science
- Course on Data Analytics and Visualization –
Università Modena e Reggio Emilia
- Dottorato in Data Science and Computation –
Università di Bologna e fondazione Golinelli



Grazie!

Contatti:

- Sanzio Bassini, Direttore dipartimento Supercalcolo s.bassini@cineca.it
- Claudio Arlandini, Project Manager HPC per le Industriec.arlandini@cineca.it
- Giuseppe Fiameni, Data Management e Middleware g.fiameni@cineca.it
- Roberta Turra, Data Analytics r.turra@cineca.it
- Gabriella Scipione, Metadata Managent e Integration g.scipione@cineca.it