SmartData@PoliTo

3rd SmartData@PoliTO Workshop:

Status and Future Directions

24-25 September 2018

Thanks to...

External Advisory Board

Sandra Di Rocco

External Advisory Board
Professor, department head KTH, Mathematics

Alessandro Laio

External Advisory Board
Statistical and Biological Physics, Complex Heterogeneous Systems

Alberto Pisoni

External Advisory Board
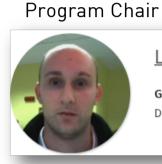Automotive, Electronic and Control System

Fosca Giannotti

External Advisory Board
Contraint-Based Frequent Pattern Mining, Mobility Data Mining, Complex Network Analysis

Pietro Michiardi

External Advisory Board
Probabilistic Machine Learning, Parallel and Distributed Learning Algorithms

Mirko Pieropan

PhD Student
Statistical Physics, Inverse Problems, Approximate Inference, Graphical Models, Image Reconstruction

Program Chair

Luca Cagliero

Group of Expert
Data Science, Text Mining

Logistic Chair

Annita Dei Tos

Administrative Manager
Taking care of all admin work.

Workshop agenda

Monday, September 24th

- 12:30 – Getting together – Welcome reception

- 14:00 - **Welcome** - *Luca Settineri* – Vice Rector of Policy Planning and Implementations

- 14:15 - Presentation of the **SmartData@Polito** center – *Marco Mellia*

- 14:45 - Keynote 1: **Practical and Scalable Probabilistic Machine Learning** – *Pietro Michiardi*

- 15:45 – **Deep learning applications in healthcare** – *Lia Morra*

- 16:15 - **Machine Learning for Fluxes Regression in Discrete Fracture Networks** – *Francesco Della Santa*

- 16:45 - *Coffee Break*

- 17:00 - **Poster session** (PhD Students)

Tuesday, September 25th

- 09:00 - Keynote 2: **Variety sampling** – *Sandra Di Rocco*

- 09:45 – **Automatic Longitudinal Exploration of Network Traffic Analysis and applications to cybersecurity** – *Andrea Morichetta*

- 10:15 – **Computational framework for inference in epidemic processes on livestock movement networks** – *Indaco Biazzo*

- 10:45 – 11:00 *Coffee Break*

- 11:00 - Keynote 3: **Synthetic maps for navigating high-dimensional data spaces** – *Alessandro Laio*

- 11:45 – **Extractive multilingual document sumarization** – *Luca Cagliero*

- 12:15 - **Learning Localized Generative Models for 3D Point Clouds via Graph Convolution** – *Giulia Fracastoro*
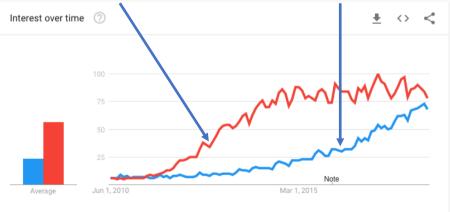
SmartData@PoliTO

Big Data & Data Science

"Extracting **meaning** from very large quantities of **data**"

The challenge of SmartData@PoliTO

- We spent a significant effort to understand
  - **What** we want to do
  - **Who** should we include
  - **What** topics we want to cover
- And to get to a shared working model
- Result:
  - A centre where experts **work and closely**
  - **Share** experience and
  - **Contribute** to the growth of the center

  - *Cooperate in the same spaces*

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

GOAL of SmartData@PoliTO

- Born on the experiences and competencies
  - EE – Internet monitoring
  - CS – Data mining & machine learning
  - MAT – Statistics and modelling
  - Phy – Modelling and inverse problems
  - MNG – Economics and business impact
  - ...

- And on the BigData@PoliTO centre
  - Fully dedicated Cluster hardware

- **Offer access and support** for these technologies
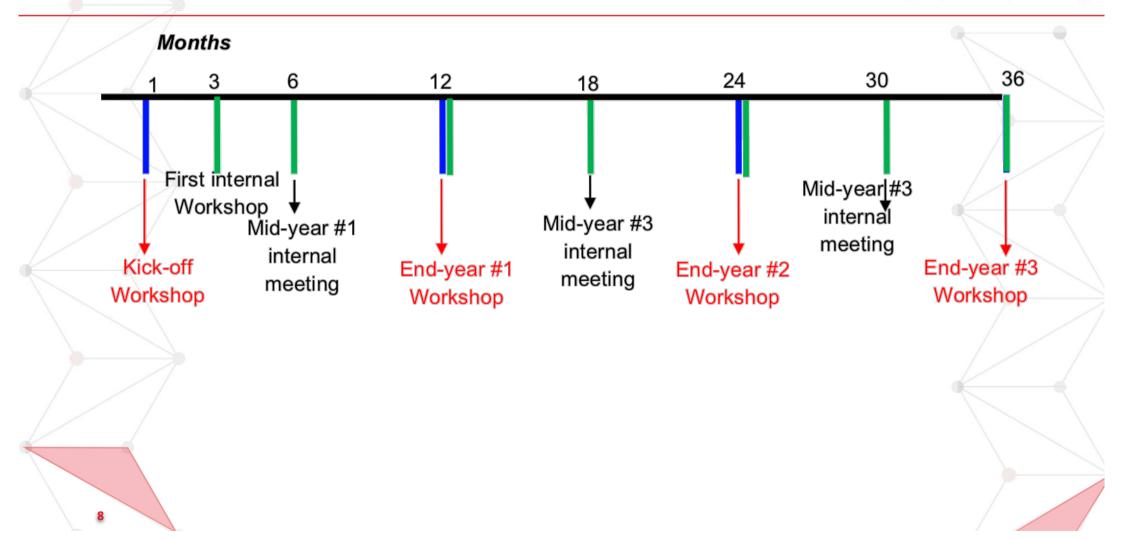
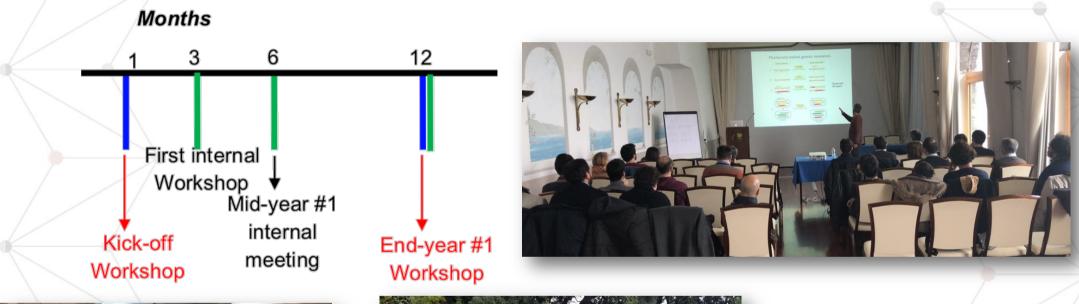- **Technology transfer:** toward industries

Where we are

Months

1	3	6	12	18	24	30	36

First internal Workshop

Mid-year #1 internal meeting

Kick-off Workshop

End-year #1 Workshop

Mid-year #3 internal meeting

End-year #2 Workshop

Mid-year #3 internal meeting

End-year #3 Workshop

Where we are

Months

1 3 6 12

First internal
Workshop

Mid-year #1
internal
meeting

Kick-off
Workshop

End-year #1
Workshop

24

	Initiative	Where	Core Competences	Personnel
1	Osservatorio Big Data Analytics & Business Intelligence	Politecnico di Milano, IT	Management	**2** contact persons
2	Despina Big Data Lab	University of Torino, IT	Economics and statistics	**13** researchers all from the same department
3	BIG DATA LAB	LUISS Business School, Roma, IT	Business	No research team is mentioned
4	CRISP, Interuniversity Research Centre on Public Services	University of Milano-Bicocca, IT	Business	**3** administrative staff, **18** researchers, all from the same department
5	ISI Foundation	Torino, IT	Data science	Around **50** scientists with different backgrounds, 15 fellows, and 25 other staff
6	Laboratorio CINI Big Data	Consortium of more than 30 Italian Universities	Computer science	No permanent researchers. Partners of different universities share competences
7	SoBigData (Social Mining & Big Data Ecosystem)	EU H2020 Consortium	Computer science	H2020 consortium composed of 9 institutions, including the CNR and UNIPI
8	**Manchester's Data Science Institute**	**University of Manchester, UK**	**Data science with bridges to "problem holders"**	**7 professors form the advisory board; 12 staff and researchers coordinate multidisciplinary projects**
9	**Berlin Big Data Center**	**Berlin, DE**	**Data science**	**5 managers, 11 principal investigators, around 20 researchers**
10	**National Center for Data Science**	**Lausanne and Zürich, CH**	**Data and computer sciences**	**40 data and computer scientists and experts in selected domains**
11	**Cambridge Big Data Strategic Research Initiative**	**University of Cambridge, UK**	**mathematics, computer science, bioinformatics, medicine, etc.**	**More than 100 researchers acting as principal investigators**
12	**MIT Big Data Initiative**	**Boston, USA**	**Data and computer sciences**	**37 faculties, 28 students, complemented by several external advisers**
13	**Stanford Data Science Initiative (SDSI)**	**Palo Alto, USA**	**medicine, computer science, biology, energy, etc.**	**16 researchers organize multi-disciplinary teams and the research agenda across the university**
14	**Berkeley Institute for Data**	**Berkeley, USA**	**psychology, statistics, mathematics, biology,**	**14 staff and 72 fellow researchers**

Positioning of Smartdata@polito

- We position ourselves as both
  - Research center for **fundamentals** of big data
  - Technology transfer to **application** domains

- We want to be **inclusive**
  - But limit participation to people really working on Big Data and Data Science topics
  - Involving domain experts ("**Group of Experts**")
  - Via open calls, co-advised PhD positions, co-funded PostDoc positions, …

| SmartData@PoliTO | Politecnico di Torino, IT | Statistics, Math, Physic, Computer Science, Communication, Domain Experts, etc.. | 30+ staff and 21+ PhDs and PostDocs |

SmartData@PoliTO in one slide

Data updated September 2018

7+ Industrial Project private [2017]

1 Large project [2017]

30+ Professors

Focus: AI ML BIG DATA INDUSTRY 4.0

14+ PhD Students

1 Cluster 2.5TB RAM, 324 Cores, 773TB

3Mln€/y Funding [2017]

Information

POLITECNICO DI TORINO

SmartData

Complete Name
SmartData@PoliTO

Parent organization
Politecnico di Torino

Sponsors
Both Private and public

Location
Torino

Born in
2017

Website
smartdata.polito.it

Coordinator
Marco Mellia

Address
Corso Duca degli Abruzzi 24 - 10129 Torino

Contacts
contact.smartdata@polito.it

People

Project Coordinator

Marco Mellia
Project Coordinator
Network Monitoring, Cybersecurity, Data Science

Steering Board

Elena Baralis
Scientific Board and Main Proponent
Data Science, Big Data Analytics

Andrea Pagnani
Scientific Board and Main Proponent
Statistical physics, combinatorial optimization, bioinformatics, biological modeling, statistical inference

Paolo Neirotti
Scientific Board and Main Proponent
Innovation Management, MIS, Economic and Industrial Impact of Big Data

Francesco Vaccarino
Scientific Board and Main Proponent
Topological data analysis, geometrical and topological aspects of data analysis, non commutative geometry and invariant theory

Participating Member

Stefano Berrone
Participating Member
Numerical Methods

Silvia Chiusano
Participating Member
Data Science, Data Mining

Alfredo Braunstein
Participating Member
Machine Learning, Statistical Physics, Combinatorial Optimization

Luca Dall'Asta
Participating Member
Theoretical Physicist

30+
features

Idilio Drago
Participating Member
Internet Measurements, Big Data Analysis, Network Security

Mauro Gasparini
Participating Member
Statistics

Fabrizio Lamberti
Participating Member
Data Visualization, Visual Analytics, Virtual and Augmented Reality, Human-Machine Interaction

Maurizio Munafò
Participating Member
Network Monitoring

Group of Experts

Andrea Acquaviva
Group of Experts
IoT, smart energy, smart manufacturing

Luca Cagliero
Group of Expert
Data Science, Text Mining

Paolo Garza
Participating Member
Data Science, Big Data Analytics

Danilo Giordano
Participating Member
Internet Traffic, Smart Mobility, Data Analysis

Enrico Magli
Participating Member
Image and Video Processing, Machine, Deep Learning

Luca Vassio
Participating Member
Data Science, Behaviour Modeling, Analytics, Analytical and Numerical

Ettore Bompard
Group of Experts
Energy Center, Lab, Power Systems, Electro-energetics, Energy Infrastructures

Claudio Canuto
Group of Experts
Numerical Methods, Signal Processing

Mario Casu
Group of Experts
VLSI Design, Hardware Accelerators for Computing, System of Chip (SoC), FPGA, Deep Learning

Juan Carlos De Martin
Group of Experts
Multimedia, Digital Revolution

Alessandro Mantelero
Group of Experts
Legal Expert on Data Protection and Big Data Regulation

Elio Piccolo
Group of Experts
Artificial Intelligence

Luca Ridolfi
Group of Experts
Hydraulics

Valentina Socco
Group of Experts
Applied Geophysics

Fernando Corinto
Group of Experts
Electrical Engineering

Antonio Lioy
Group of Experts
Cybersecurity

Maurizio Morisio
Group of Experts
Data Mining of Software Product Data, Software Analytics

Cristina Pronello
Group of Experts
ITS (Intelligent Transport Systems), Sustainable Mobility, Travel analysis

Riccardo Sisto
Group of Experts
Networks, Distributed Systems

Marco Torchiano
Group of Experts
Data Mining of Software Product Data, Software Analytics, Data Visualization

People

Andrea Bordone Molini

PhD Student
Unsupervised machine learning, Time Series Analysis, Anomaly Detection

Alessandro De Gregorio

PhD Student
Topological Data Analysis, Algebraic Topology

Marilisa Montemurro

PhD Student
Artificial Intelligence, Bioinformatics, Data Mining

Andrea Pasini

PhD Student
Data Science, Big Data, Deep Learning

Francesco Della Santa

PhD Student
Machine Learning, Numerical Analysis

Alessandro Destefanis

PhD Student
Big Data, Management, Strategic Decision Making, Statistical Modelling

Eliana Pastor

PhD Student
Data Science, Big Data Analytics

Mirko Pieropan

PhD Student
Statistical Physics, Inverse Problems, Approximate Inference, Graphical Models, Image Reconstruction

Marco Guerra

PhD Student
Topological methods, dimensionality reduction, privacy

Pinky Kumawat

PhD Student
Mobility as a Service, Sustainability

Nicola Prette

PhD Student
Deep learning models, next-generation video codec

Francesca Soro

PhD Student
Cybersecurity, Big Data Analytics

Moreno La Quatra

PhD Student
Deep learning , multimedia

Machine Learning, Embedded Systems

14+
PhD Students

- Each student has 2 Advisors and a well defined topic
- Grants are 75% co-funded by the lab with 2 grants 100% covered by companies

Competences

- **Algorithms** and **methodologies** for data analysis
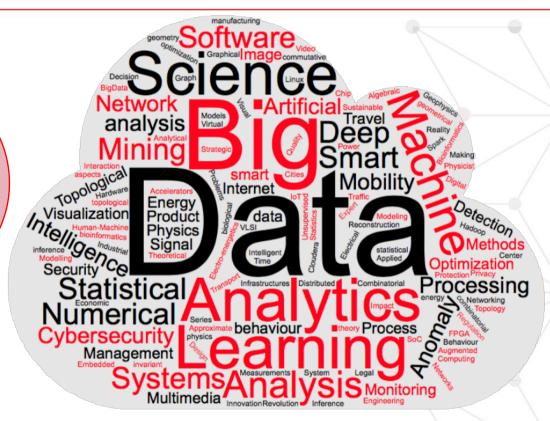 - Big data processing, supervised and un... learning, deep le... classification, pr... detection, cluste...

- **Methodologies** for da...
  - Statistics, compu... geometry with app... analysis, graph mod... information mining from graphs, ...

- Study the **transformations** in **industrial structures**
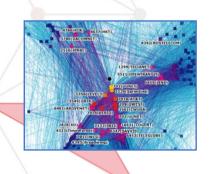 - Value chains, business models, managerial practices, ...

Focus:

AI
ML
BIG DATA
INDUSTRY 4.0

Generation ▸ Acquisition ▸ Storage ▸ Analysis

Applications and domains
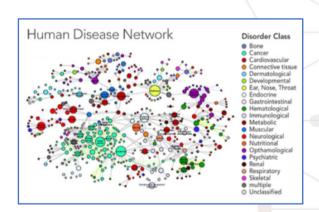
- Predictive maintenance for complex systems
- Anomaly detection with applications to
    time series analysis
- Cybersecurity
- Mobility in smart cities
- Data diffusion in complex systems

Cluster SmartData@PoliTO

- Heterogeneous solution

  - 3 Master Nodes DELL PowerEdge R620

    - 1U, 128GB, 2CPU 6Cores, 600GBx3 SAS

  - 18 Workers Dell R720

    - 2U, 96GB, 2CPU 6Cores, 3TBx12 SATA

  - 12 Workers Supermicro

    - 2U, 32GB, 1 CPU 6Cores, 2TBx5 SATA

  - cluster
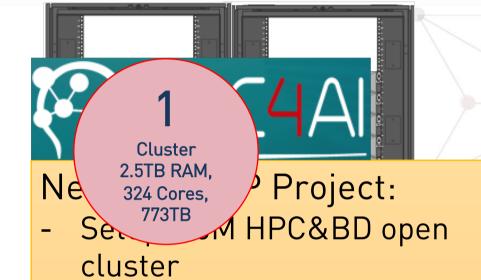
    - 2Rack, 2.5TB RAM, 324 Cores, 773TB

  - DELL N3048

    - 48Porte 1Gb/s, 2 porte 10Gb/s

  - version

    - 4 racks, more than 60 nodes, 10x bigger, fatter, faster,…
    - March 2019 – 1.5MEuros (50% funded by Regione Piemonte)

1

Large project [2017]

1

Cluster 2.5TB RAM, 324 Cores, 773TB

Ne___ P Project:

- Set___M HPC&BD open cluster

- To foster collaboration with SME and R&D center

Roberto Marturano

IT Manager
Big Data, Cloudera, Hadoop, Spark, Linux

7+

Industrial Project
private
[2017]

Selected Projects
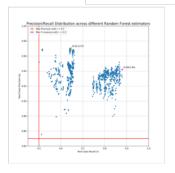
Collaboration with Private Companies
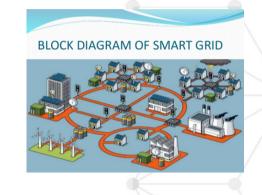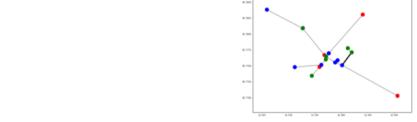
- **Predictive Maintenance**
  - Analysis of data from **electric grid** (SCADA) to predict **failures**/malfunctioning of vital elements

- **Advanced Diagnostic/Predictive Maintenance**
  - Analysis of data from **powertrain systems** to predict **failures**/malfunctioning

Collaboration with Private Companies

EDISON

- **Smart City**
  - Harvesting of **open data** for the characterization of energy consumption of residential and commercial **buildings**

- **Mobility prediction**
  - Study of **traffic patterns** from car sharing systems
  - Design of **electric car sharing system**

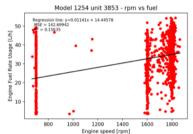# Collaboration with Private Companies
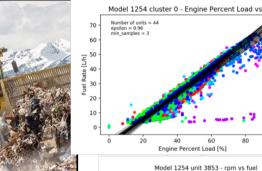
- **Service characterization, profiling, and shaping**
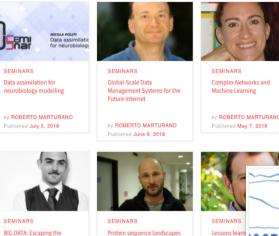 - Analysis of **IoT CAN bus** data related to on-road and off-road **vehicle** usage

- **Internet traffic monitoring and security**
  - Analysis of internet traces looking for **malware** and other **anomalies**

What we have done

- 3 SmartWorkshops
- 11 SmartSeminars
- 2 Assistant Professors RTD-A
- 14+ PhD grants
- PhD and master courses
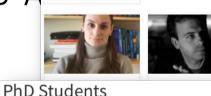- Industrial collaborations
- Conference organization
- ...
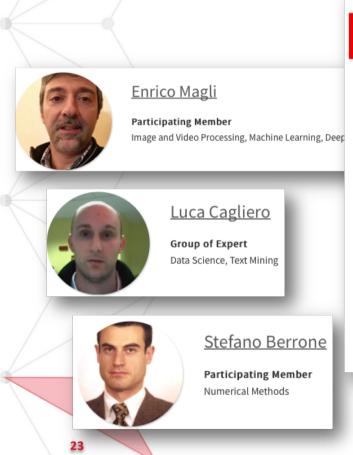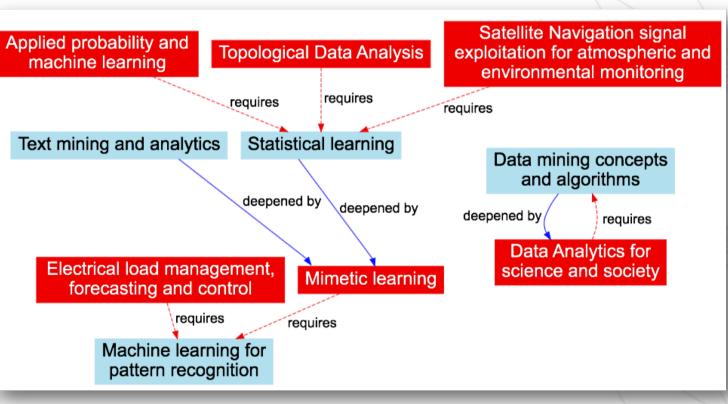
Working towards a PhD curriculum

Enrico Magli

Participating Member
Image and Video Processing, Machine Learning, Deep

Luca Cagliero

Group of Expert
Data Science, Text Mining

Stefano Berrone

Participating Member
Numerical Methods
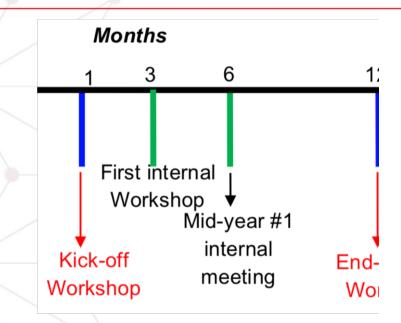
Tell us about your courses!

Months

| 1 | 3 | 6 | 12 |

Kick-off Workshop

First internal Workshop

Mid-year #1 internal meeting

End- Wor

Seed grant ends in 2020

Future directions
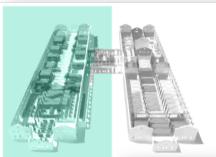
- In the TODO list
  - Get a real lab!

Officine Sud: 9.000 metri quadrati di futuro

Un hub per la creazione collaborativa e incubatore di idee focalizzato su start up, industrie creative e smart data.

25

Future directions

- In the TODO list
  - Get a real lab!

- Keep funding
  - ~~2 more RTD-A~~
  - PhD grants
    - 6+ more positions
  - PostDoc positions
    - 5+ more positions
  - Visiting professors
    - 3 positions

- Keep the **co-funding** model
  - 75% SmartData+ 25% proponent
  - With open calls

	Year 1	Year 2	Year 3
RTD-A	1	1	0
ADR - B1			4
	3	2	1
			cofunding
PhD students		cofunding	
	7	14	20
Development Manager	0	1	1
IT Manager	1	1	1
Administrative	0,5	0,5	0,5
Total	12,5	21,5	27,5

Future directions

- No more pictures here ☺
- Start focussing activities
  - Select research areas to focus on
  - Increase the involvement in funded EU projects (FET, RIA, ITN, …)
- Prefer long-lived collaborations
  - Collaborations with companies require a long start-up time
  - Topics are sometimes too practical

Future directions

- Start a **consolidations process**
  - There are members that did not even sent their profile picture...
- Open participations to other colleagues
  - There are colleagues that are already actively collaborating with the centre...
- Open the Steering Board
  - Formally 5 members – but practically open to participation
- **Receive suggestions**
  - From the External Advisory Board
  - From colleagues

- Open workshop winter 2019
  - February 25th, 26th, or February 28th- March 1st, or March 4th,5th

- With the participation of renowned experts
  - Prof Carlo Ratti – SENSEable City Lab – MIT
  - Dr. Edo Liberty - Head of Amazon AI Labs – Palo Alto
  - Dr. Jerome Pesenti – VP of AI at Facebook (TBC)
  - ~~Tomaso Poggio - McGovern Institute – MIT~~

- Organized as an open workshop
  - «The present and future of AI»

- Followed by a 1,5 days PhD school

- **Looking for volunteers to help**

Workshop agenda

Monday, September 24th

- 12:30 – Getting together – Welcome reception

- 14:00 - **Welcome** - *Luca Settineri* – Vice Rector of Policy Planning and Implementations

- 14:15 - Presentation of the **SmartData@Polito** center – *Marco Mellia*

- 14:45 - Keynote 1: **Practical and Scalable Probabilistic Machine Learning** – *Pietro Michiardi*

- 15:45 – **Deep learning applications in healthcare** – *Lia Morra*

- 16:15 - **Machine Learning for Fluxes Regression in Discrete Fracture Networks** – *Francesco Della Santa*

- 16:45 - *Coffee Break*

- 17:00 - **Poster session** (PhD Students)

Tuesday, September 25th

- 09:00 - Keynote 2: **Variety sampling** – *Sandra Di Rocco*

- 09:45 – **Automatic Longitudinal Exploration of Network Traffic Analysis and applications to cybersecurity** – *Andrea Morichetta*

- 10:15 – **Computational framework for inference in epidemic processes on livestock movement networks** – *Indaco Biazzo*

- 10:45 – 11:00 *Coffee Break*

- 11:00 - Keynote 3: **Synthetic maps for navigating high-dimensional data spaces** – *Alessandro Laio*

- 11:45 – **Extractive multilingual document sumarization** – *Luca Cagliero*

- 12:15 - **Learning Localized Generative Models for 3D Point Clouds via Graph Convolution** – *Giulia Fracastoro*