BALÁZS KÉGL

DR / CNRS

LAL & LRI

CNRS & University Paris-Sud

ARNAK DALALYAN

Pr / ENSAE

Laboratoire de Statistique

ENSAE / CREST

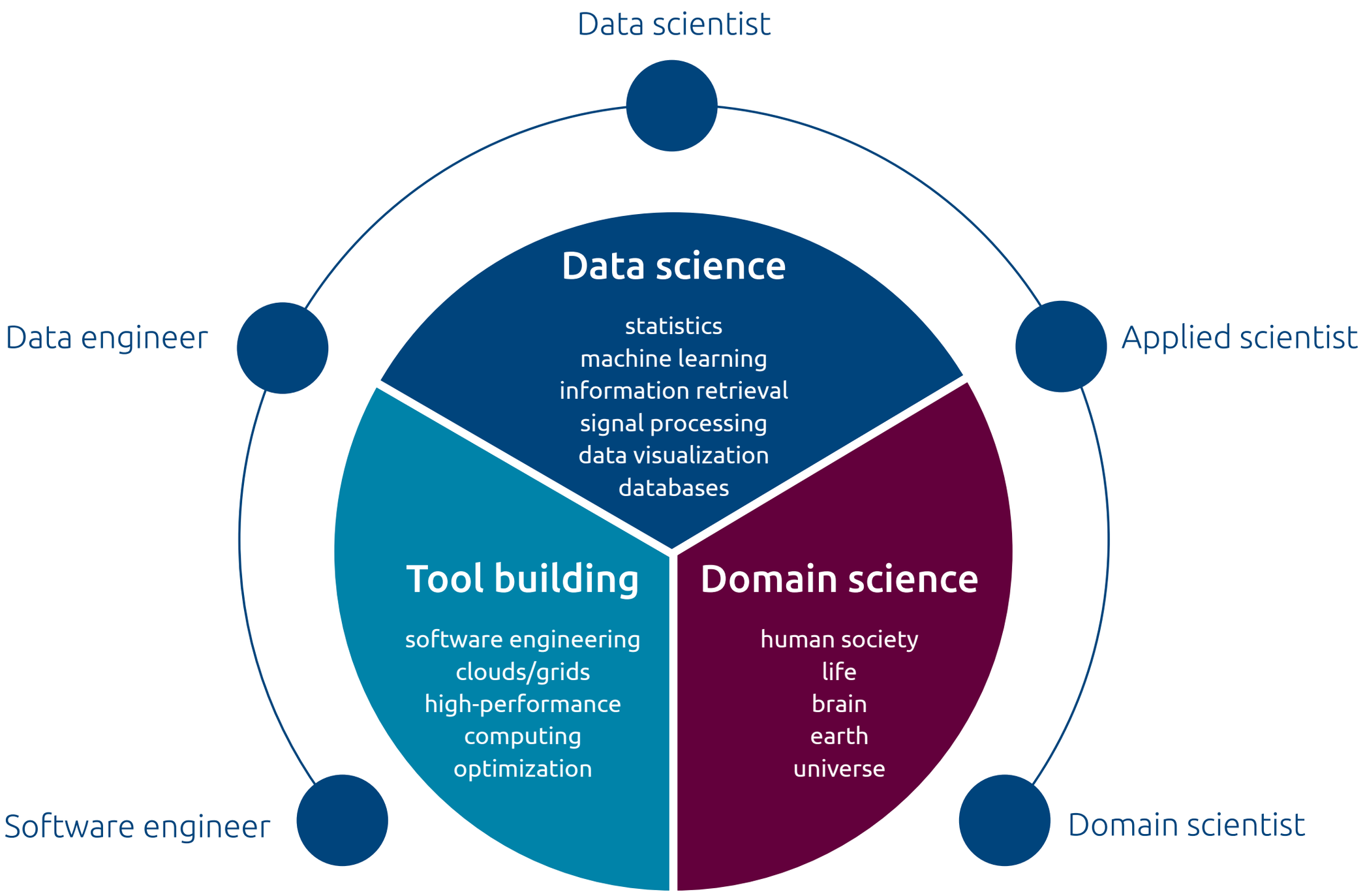
DATA SCIENCE

Design of automated methods

to analyze massive and complex data

to extract useful information

THE DATA SCIENCE LANDSCAPE



THE PARTICIPANTS

250 researchers in 35 laboratories

Biology & bioinformatics
IBISC/UEvry
LRI/UPSud
Hepatinov
CESP/UPSud-UVSQ-Inserm
IGM-I2BC/UPSud
MIA/Agro
MIAj-MIG/INRA
LMAS/Centrale

Chemistry
EA4041/UPSud

Earth sciences
LATMOS/UVSQ
GEOPS/UPSud
IPSL/UVSQ
LSCE/UVSQ
LMD/Polytechnique

Economy
LM/ENSAE
RITM/UPSud
LFA/ENSAE

Neuroscience
UNICOG/Inserm
U1000/Inserm
NeuroSpin/CEA

Particle physics astrophysics & cosmology
LPP/Polytechnique
DMPH/ONERA
CosmoStat/CEA
IAS/UPSud
AIM/CEA
LAL/UPSud

Machine learning
LRI/UPSud
LTCI/Telecom
CMLA/Cachan
LS/ENSAE
LIX/Polytechnique
MIA/Agro
CMA/Polytechnique
LSS/Supélec
CVN/Centrale
LMAS/Centrale
DTIM/ONERA
IBISC/UEvry

Visualization
INRIA
LIMSI

Signal processing
LTCI/Telecom
CMA/Polytechnique
CVN/Centrale
LSS/Supélec
CMLA/Cachan
LIMSI
DTIM/ONERA

Statistics
LMO/UPSud
LS/ENSAE
LSS/Supélec
CMA/Polytechnique
LMAS/Centrale
MIA/AgroParisTech

PARAMETERS

- 2 years: April 2014 - June 2016, 1.2M€
  - +1 year, conditional on evalution
- Light management
- executive committee of 17 members
- work groups
  - management (around objectives)
  - thematic (around scientific themes)

GOALS

- Build a community at Saclay around data science
- Get interdisciplinary collaborations off the ground
  - seeding postdocs, creativity workshops, thematic days, data challenges
- Support software tool building
  - coding sprints, bootcamps, engineering projects, open software initiative
- Data science IT platform (open data)
  - io.cds, open data, open software, reproducible research
- Making the CDS a contact point to big data industry

A UNIQUE OPPORTUNITY WITH UNIQUE CHALLENGES

- Unparallelled depth and breadth of talent: how to make them work together?

- Shortening the collaboration turnaround using coached workshops

- Fast-forward bootcamp-style data science training for a new crop of researchers

- Changing the carrier incentives: tool building, interdisciplinary research

- Inventing the institutional framework

GOALS AND TOOLS

- Build a community at Saclay around data science
  - an agora where researchers and engineers can meet and talk
  - a culture of crossing the disciplinary aisles
  - get to know each other's expertise and data analysis problems
- Tools
  - an interactive web portal: http://datascience-paris-saclay.fr (preliminary)
  - workshops, thematic days, and creativity workshops
  - summer school(s)

GOALS AND TOOLS

- Get interdisciplinary collaborations off the ground
  - seeding larger projects for outside funding (ANR, Europe)
  - embedding data scientists in domain science labs and vice verse
- Tools
  - financing postdoctoral projects, theses and (incoming) sabbatical/visiting stays
    - 5 postdocs in first call (July 2014)
    - 6 theses (within COFUND, launched)
    - 2-3 projects in second call (planned)
  - financing data challenges
    - http://higgsml.lal.in2p3.fr

GOALS AND TOOLS

- Support software tool building
  - primarily open source
  - development, maintenance, reusability across disciplines
  - leadership of Alexandre Gramfort (LTCI Telecom/CEA Neurospin) and Guillaume Wisniewski (LIMSI/UPSud)

- Tools
  - the Open Software Initiative for interns and doctoral students
    - 7 doctoral missions, very popular (17 candidates)
  - "code consolidator" and engineering projects
    - 4 engineering projects + 1 code consolidator in first call (July 2014)
  - coding sprints
    - scikit-learn coding sprint (July 2014)
  - data science bootcamps

GOALS AND TOOLS

- Data science IT platform
  - open data, open access, reproducible research (see IPOL)
  - disseminating and sharing data and software
  - connecting the CDS to data centers (e.g. Virtual Data)
  - leadership of Cécile Germain (LRI / UPSud)
- Tools
  - web portal: http://io.datascience-paris-saclay.fr (preliminary)
  - work group

GOALS AND TOOLS

- Design strategies
  - the institutional structure to stabilize the CDS
  - the ideal structures for data science research
  - projecting it onto existing structures (e.g., hotel à projets)
- Tools
  - work group
  - international workshop on organizing/managing data science research

GOALS AND TOOLS

- Making national and local decision-makers aware of the challenges
  - lack of incentives for interdisciplinary research and tool building
  - unprecedented brain drain into industrial research
- Tools
  - lobbying, explaining, thinking together

GOALS AND TOOLS

- Creating a map of Saclay data science masters
  - making the masters transparent
- Tools
  - web portal
  - work group
  - initiating an annual meeting on data science education

GOALS AND TOOLS

- Making the CDS a contact point to big data industry
- Tools
  - web portal
  - work group
  - building on existing tools (e.g. CIFRE theses, SystemX, Cap Digital)
  - supporting an ecosystem around start-ups (through SATT)