THE DATA SCIENCE ECOSYSTEM

BALÁZS KÉGL

DR / CNRS

Laboratoire de l'Accélérateur Linéaire &
Laboratoire de la Recherche en Informatique

CNRS & University Paris-Sud

universite PARIS-SACLAY

Paris-Saclay Center for Data Science

Outline

- Who are we?
  - Université Paris-Saclay
  - Center for Data Science
- The data science ecosystem
- What do we design?
  - In experimental physics
  - In data science

19 founding partners

Université Paris-Saclay

19 *fondateurs*

60 000 *étudiants*

6 000 *doctorants*

15 000 *étudiants en master*

8 *Schools*

11 000 *chercheurs et enseignants-chercheurs*

300 *laboratoires*

8 000 *publications /an*

15 % *de la recherche publique française*

10 *départements*

+ horizontal multi-disciplinary and multi-partner initiatives ("lidexes") to create cohesion

A multi-disciplinary initiative to define, structure, and manage the data science ecosystem at the Université Paris-Saclay

http://www.datascience-paris-saclay.fr/

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

Data Science

Design of automated methods

to analyze massive and complex data

to extract useful information

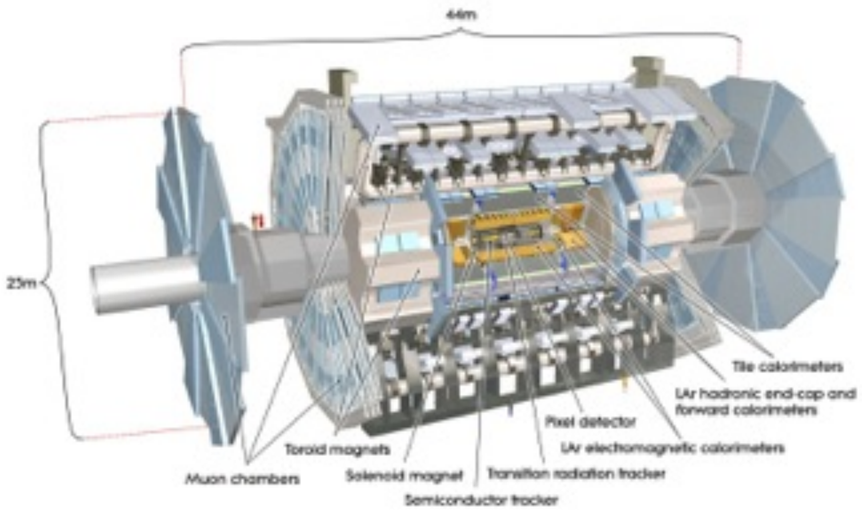
Data Science

Design of automated methods
to analyze massive and complex data
to extract useful information

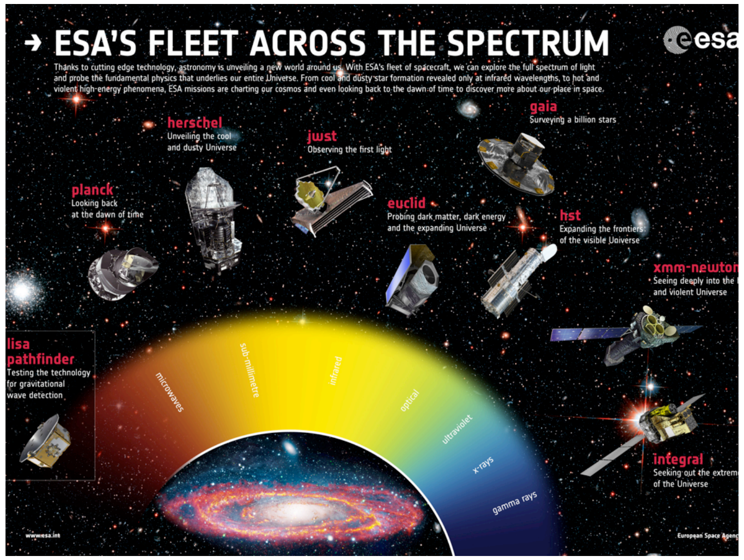
Focusing on inference:
data → knowledge

DATA IN SCIENCE: THE FOURTH PARADIGM

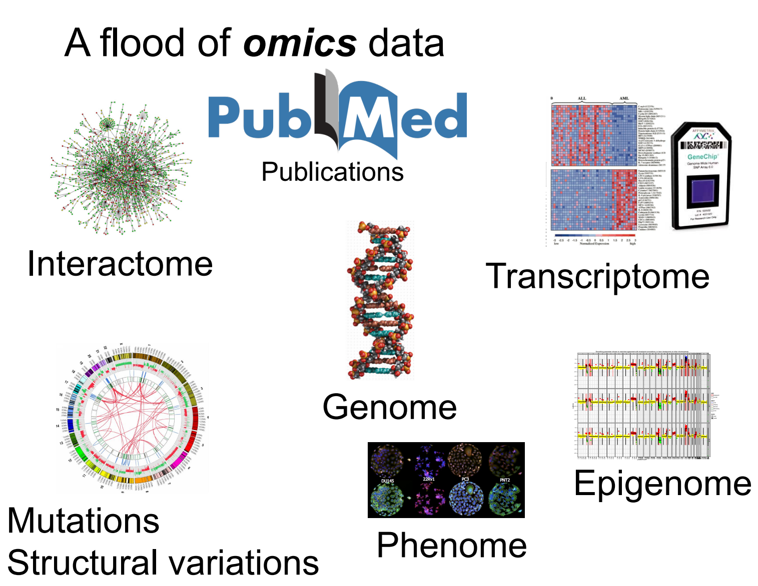
High-energy physics



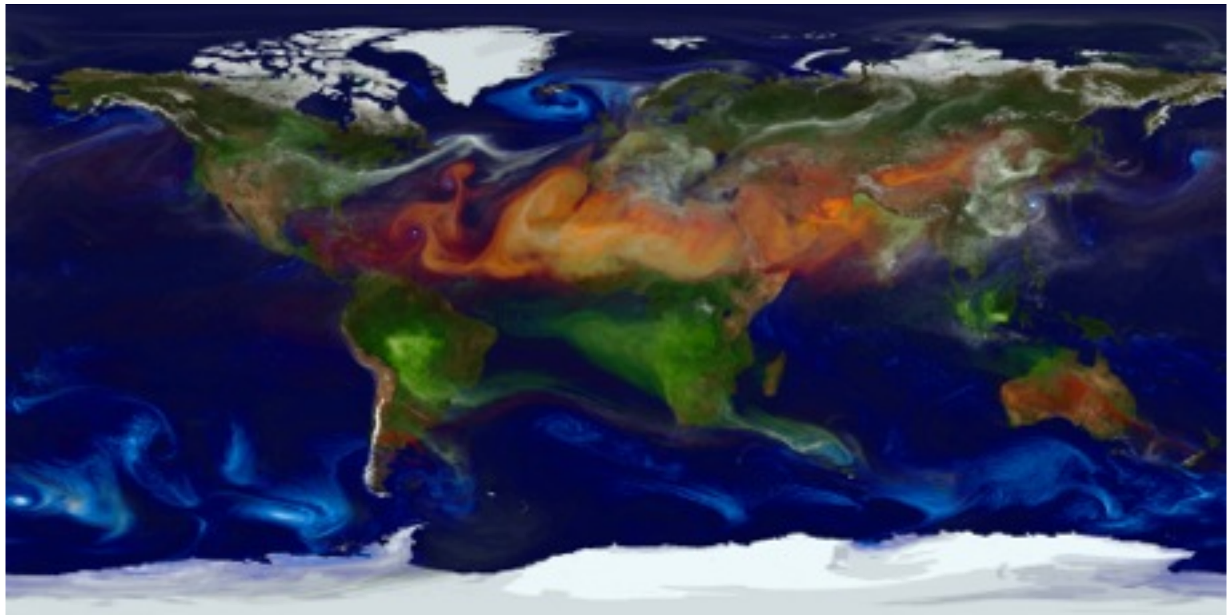
Astrophysics



Biology/health

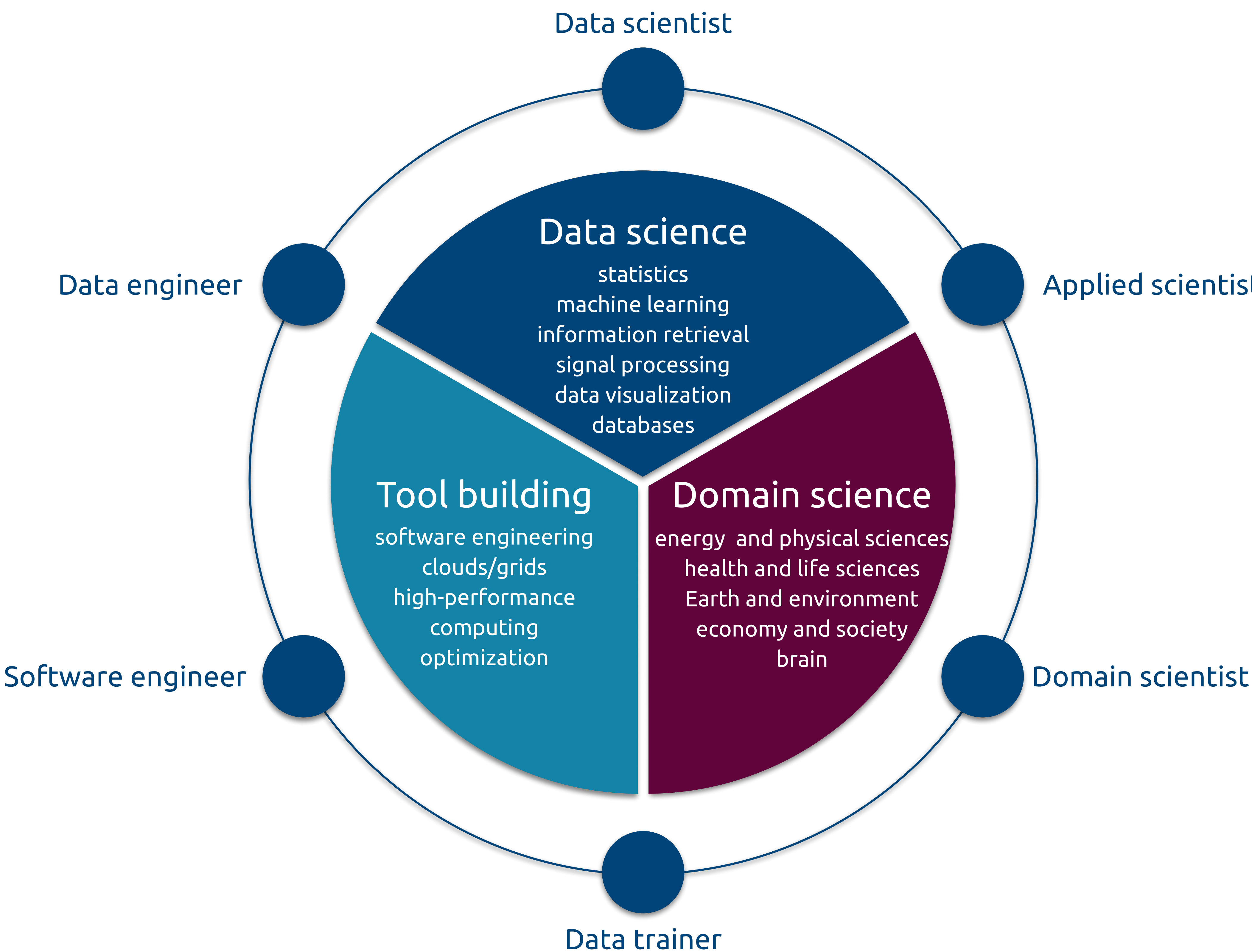


Environmental sciences

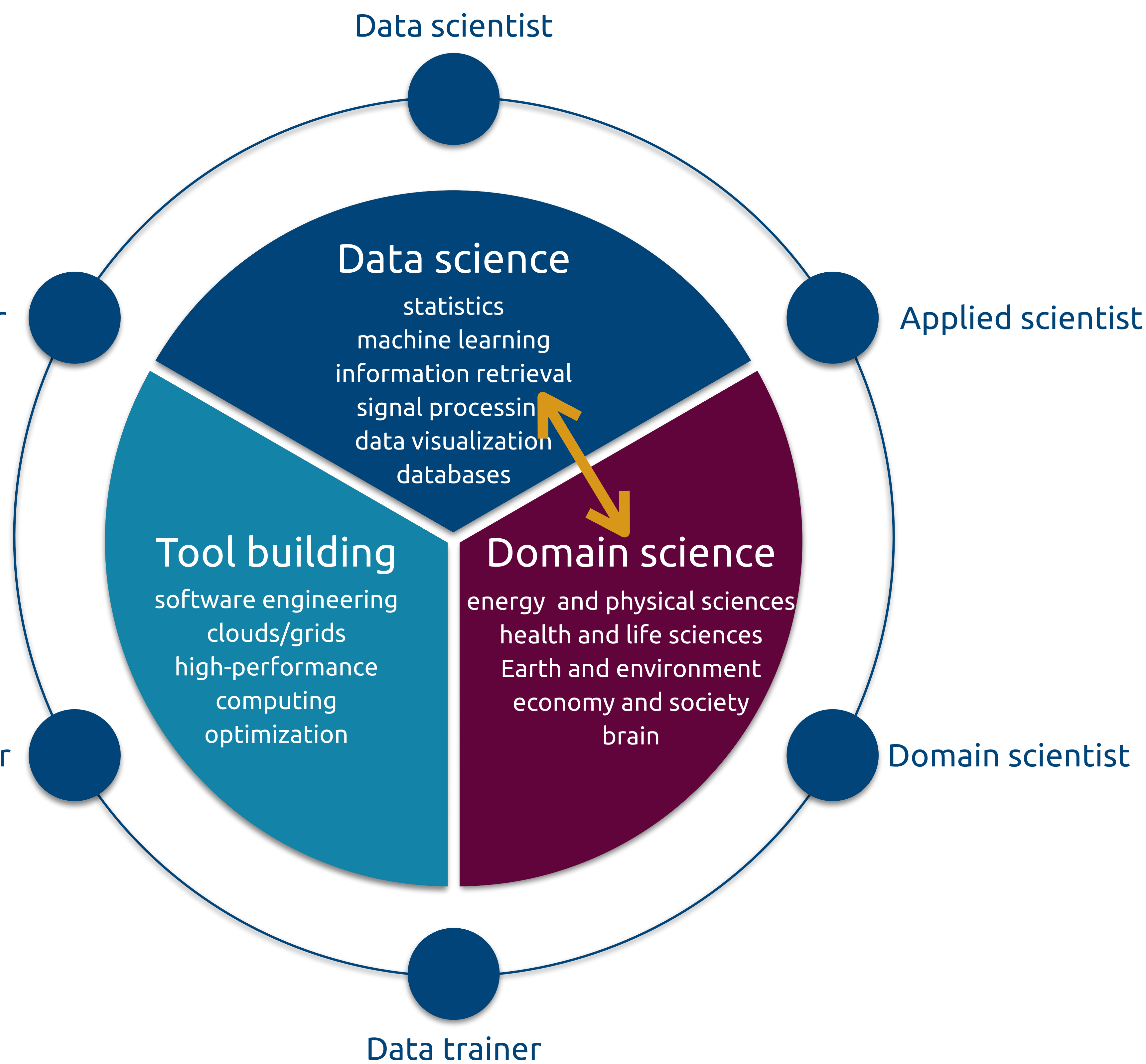


université PARIS-SACLAY
Paris-Saclay Center for Data Science

THE DATA SCIENCE LANDSCAPE



The data science landscape



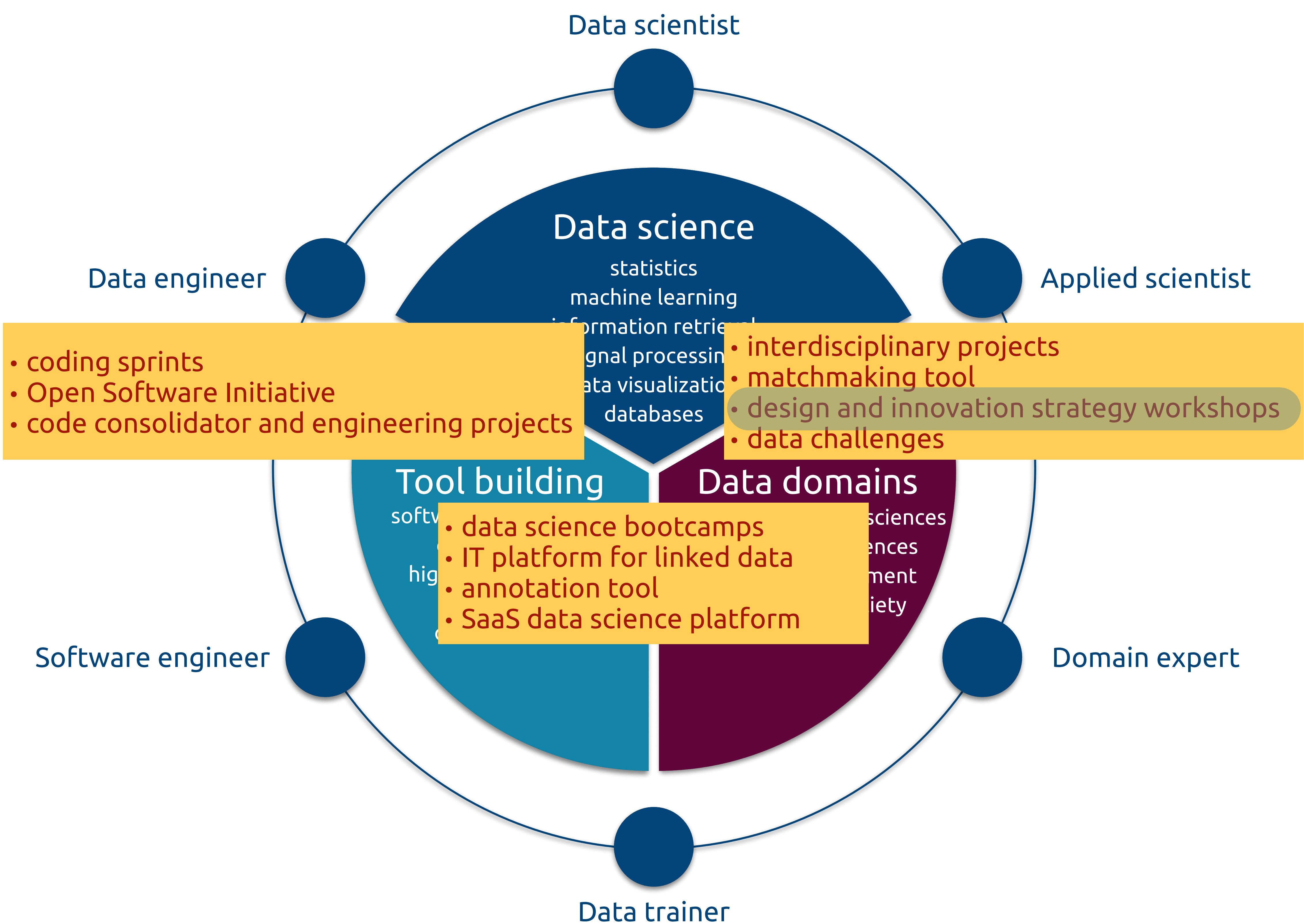
TOOLS

We are designing and learning to manage tools to accompany data science projects with different needs

CHALLENGES

Few tools exist that can help domain scientists and data scientists to collaborate efficiently

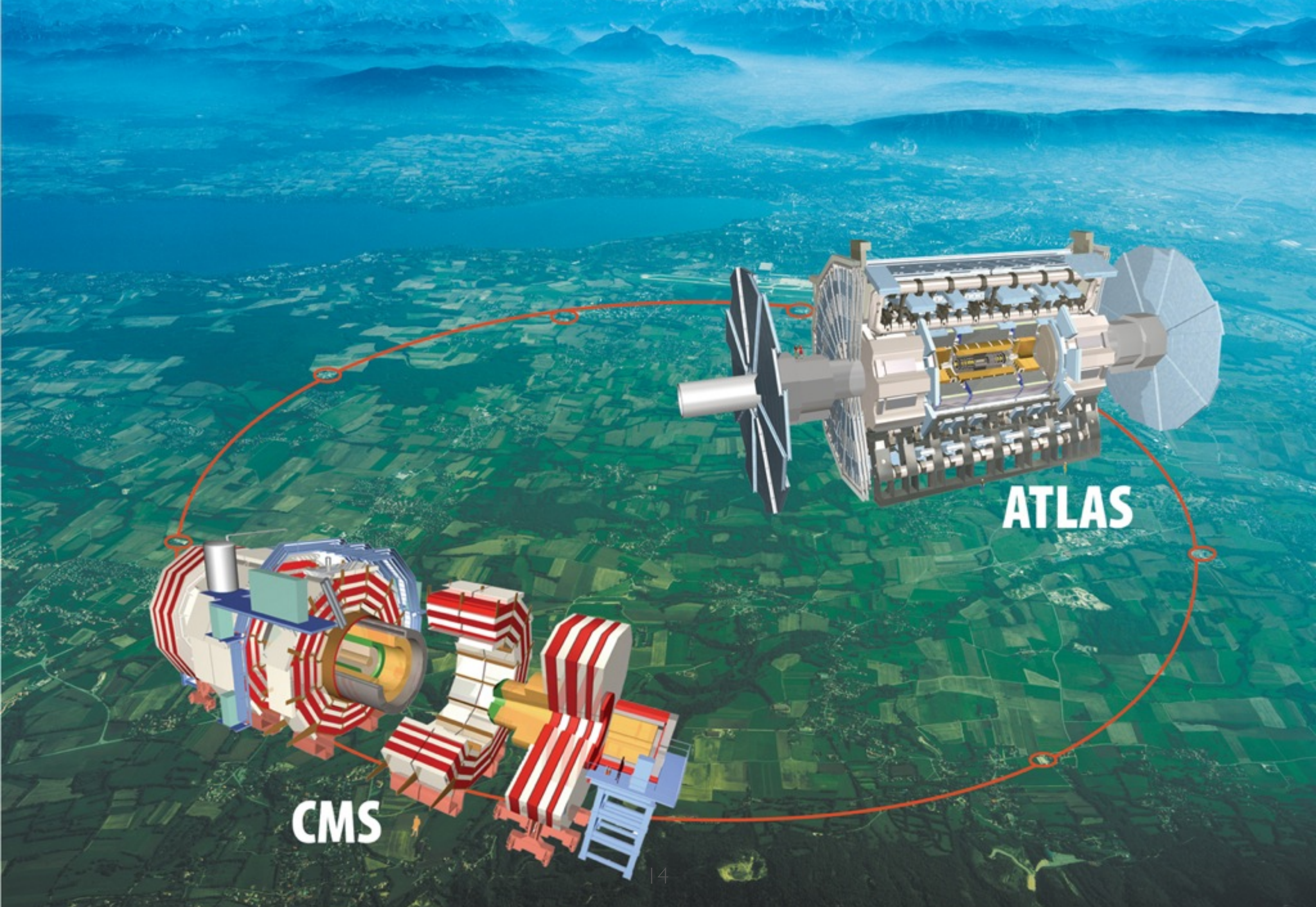
TOOLS: LANDSCAPE TO ECOSYSTEM



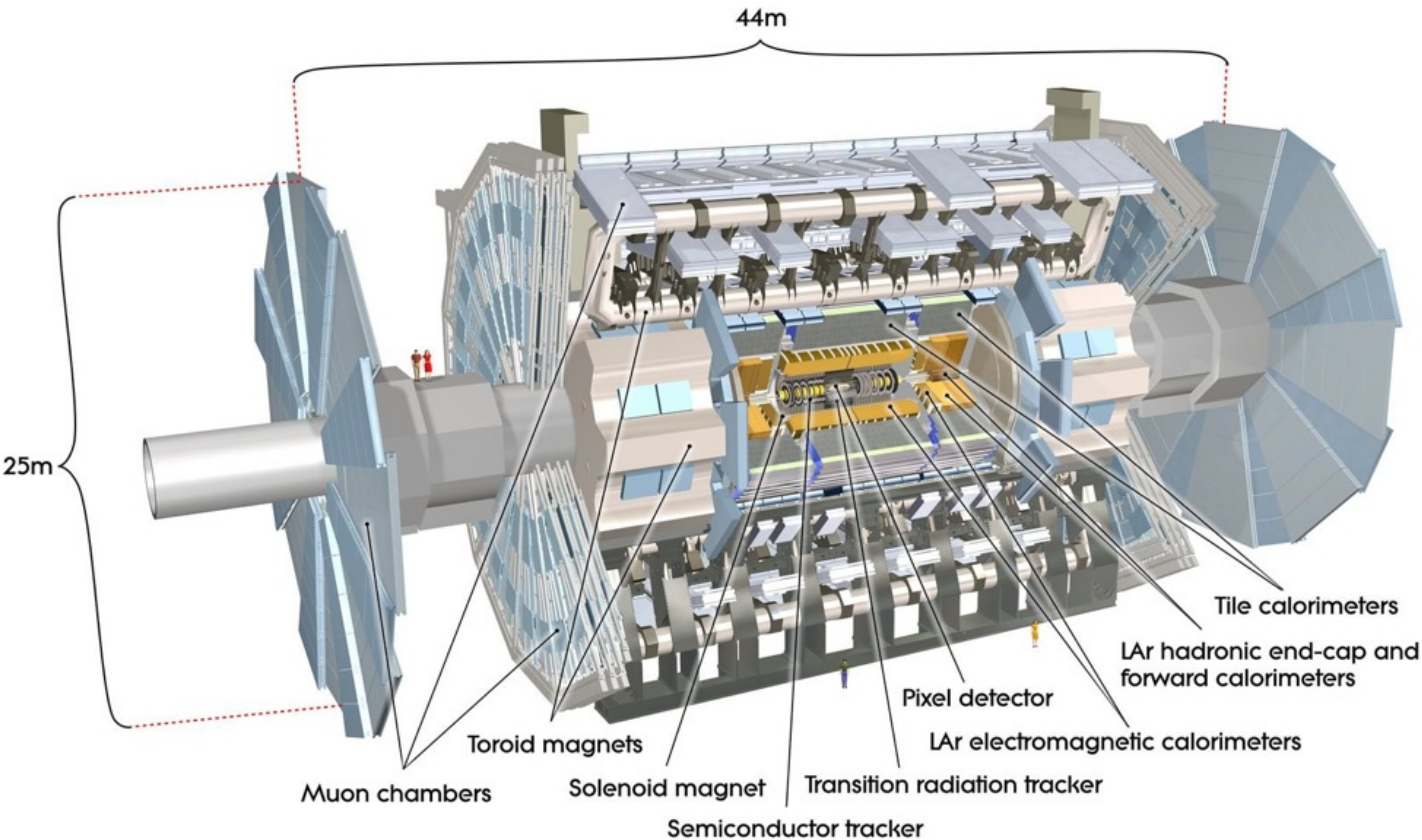
What do we design?

WHAT DO EXPERIMENTAL PHYSICISTS DESIGN?

EXPERIMENTS



DETECTORS

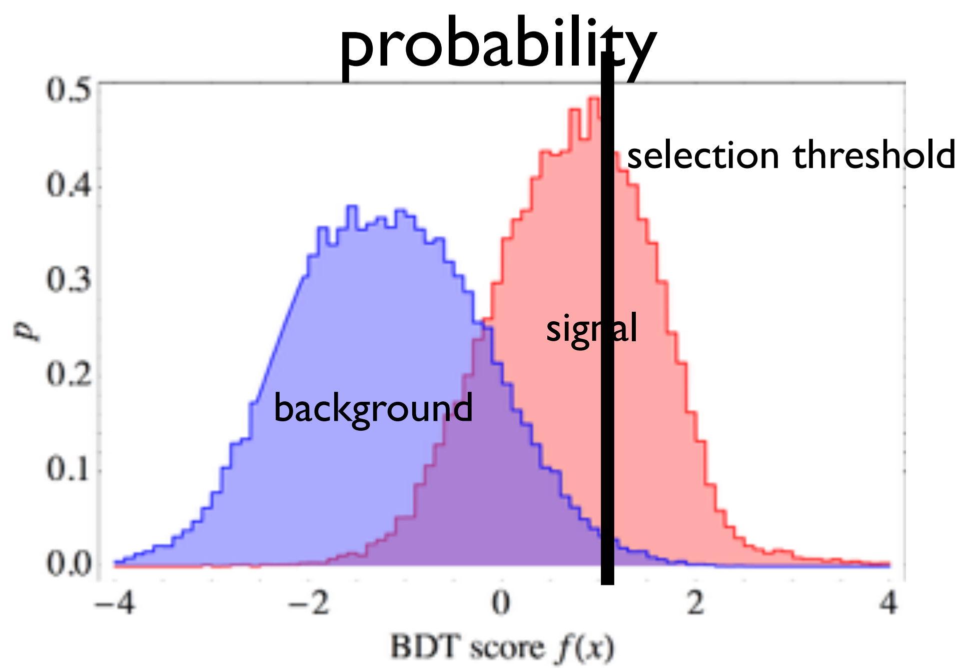

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

DATA COLLECTION PIPELINES

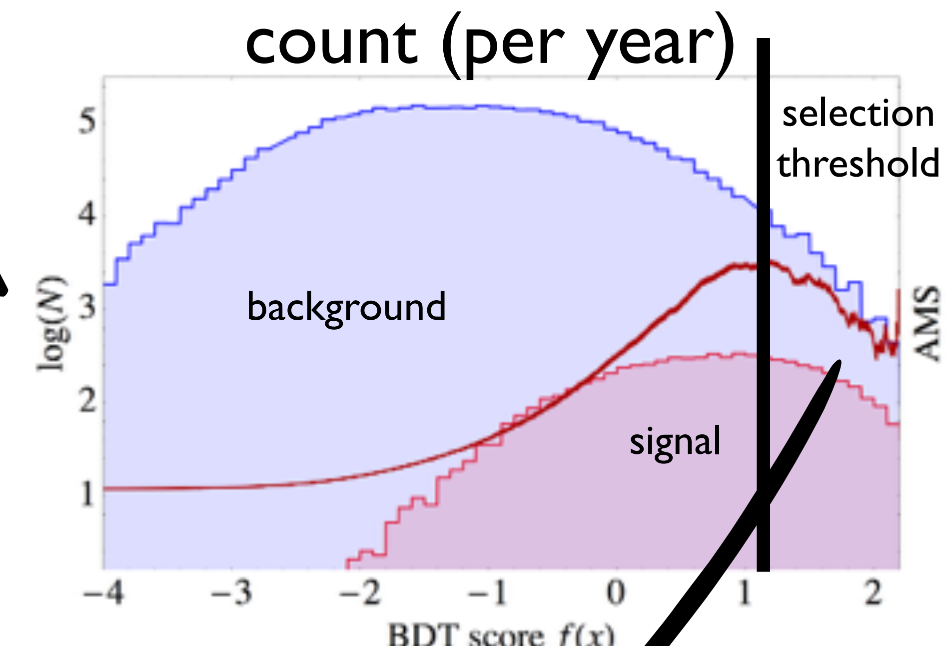
- **Hundreds of millions** of proton-proton collisions **per second**
- Filtered down to **400 events per second**
  - still **petabytes per year**
  - **real-time** (budgeted) classification: trigger

ANALYSIS PIPELINES

Goal: optimize the expected discovery significance



flux × time



selection

total count,
say, 150 events

excess is s = 50 events

expected background
say b = 100 events

$$\text{AMS} = \frac{s}{\sqrt{b}} = 5 \text{ sigma}$$

Experimental physicists design discovery processes following the scientific method

Experimental physicists design

discovery processes

Experimental physicists design discovery processes

On the way they use data science techniques, even motivate the development of new techniques, but they don't care about methodological improvements as long as the job gets done reasonably efficiently

EXPERIMENTAL PHYSICISTS DESIGN DISCOVERY PROCESSES

On the way they use data science techniques, even motivate the development of new techniques, but they don't care about methodological improvements as long as the job gets done reasonably efficiently

Innovation in data science might be hindered in such collaborations

WHAT DO DATA SCIENTISTS DESIGN?

THE CYNICAL VIEW:
WE DESIGN RESEARCH PAPERS

The cynical view: we design research papers

The cynical view: we design research papers

- In a prefect world, research papers are a means to the end of communicating scientific results

THE CYNICAL VIEW: WE DESIGN RESEARCH PAPERS

- In a prefect world, research papers are a means to the end of communicating scientific results
- Of course, good scientific results are not uncorrelated to good research papers

THE CYNICAL VIEW: WE DESIGN RESEARCH PAPERS

- In a prefect world, research papers are a means to the end of communicating scientific results
- Of course, good scientific results are not uncorrelated to good research papers
- But: the dominant design of research papers frames not only how we tackle problems, but also what problems we work on

What do data scientists design?

What do data scientists design?

- Methods to solve data science problems (data to knowledge)

What do data scientists design?

- Methods to solve data science problems (data to knowledge)
- Problems to work on

What do data scientists design?

- Methods to solve data science problems (data to knowledge)
- Problems to work on
- Theoretical (mathematical) frameworks to analyze data science methods

What do data scientists design?

- Methods to solve data science problems (data to knowledge)
- Problems to work on
- Theoretical (mathematical) frameworks to analyze data science methods
- Experimental techniques to validate data science methods

Designing data science methods

Designing data science methods

- A messy mixture of principles

Designing data science methods

- A messy mixture of principles
  - The engineering approach: does it work for solving a problem?

DESIGNING DATA SCIENCE METHODS

- A messy mixture of principles
  - The engineering approach: does it work for solving a problem?
  - The mathematical approach: can we prove that it works for solving a problem?

DESIGNING DATA SCIENCE METHODS

- A messy mixture of principles
  - The engineering approach: does it work for solving a problem?
  - The mathematical approach: can we prove that it works for solving a problem?
  - The scientific approach: can we motivate it by our (spotty) knowledge on how the brain works? Does it tell us something about the brain (simulator)?

DESIGNING DATA SCIENCE METHODS

- A messy mixture of principles
  - The engineering approach: does it work for solving a problem?
  - The mathematical approach: can we prove that it works for solving a problem?
  - The scientific approach: can we motivate it by our (spotty) knowledge on how the brain works? Does it tell us something about the brain (simulator)?
- Fierce fighting on what the right principle is

The engineering approach

THE ENGINEERING APPROACH

- No first principles (electric engineering without Maxwell's laws), trial and error

THE ENGINEERING APPROACH

- No first principles (electric engineering without Maxwell's laws), trial and error
- Needs rigorous experimental design setup, benchmark instantiations of a problem, quantitative quality measurements

THE ENGINEERING APPROACH

- No first principles (electric engineering without Maxwell's laws), trial and error
- Needs rigorous experimental design setup, benchmark instantiations of a problem, quantitative quality measurements
- Benchmark problems are often "abstracted away" from real problems

THE ENGINEERING APPROACH

- No first principles (electric engineering without Maxwell's laws), trial and error
- Needs rigorous experimental design setup, benchmark instantiations of a problem, quantitative quality measurements
- Benchmark problems are often "abstracted away" from real problems
- Data scientists usually don't care if a "real" problem is solved, as long as his/her method can be shown to improve results on benchmarks

DATA SCIENCE PROBLEMS

DATA SCIENCE PROBLEMS

- Usually come from outside of data science, we call them "real problems"

Data science problems

- Usually come from outside of data science, we call them "real problems"
- We turn them into an abstract problems by formalizing them (i.e, input, output, objective or merit function)

DATA SCIENCE PROBLEMS

- Usually come from outside of data science, we call them "real problems"
- We turn them into an abstract problems by formalizing them (i.e, input, output, objective or merit function)
- Introducing a new problem into a community is harder than it looks, needs marketing

DATA SCIENCE PROBLEMS

- Usually come from outside of data science, we call them "real problems"
- We turn them into an abstract problems by formalizing them (i.e, input, output, objective or merit function)
- Introducing a new problem into a community is harder than it looks, needs marketing
  - no benchmark in the beginning: paper cannot be formatted in the right way

DATA SCIENCE PROBLEMS

- Usually come from outside of data science, we call them "real problems"
- We turn them into an abstract problems by formalizing them (i.e, input, output, objective or merit function)
- Introducing a new problem into a community is harder than it looks, needs marketing
  - no benchmark in the beginning: paper cannot be formatted in the right way
- Once it's done, everybody tries his/her favorite hammer

Data scientists design methods

Data scientists design methods

Their goal is to improve methods on established benchmarks, and they don't care if a real problem is solved or if the improvement matters

DATA SCIENTISTS DESIGN METHODS

They don't care if a real problem is solved

PHYSICISTS DESIGN DISCOVERY PROCESSES

They don't care about methodological improvements

Thank you!

The mathematical approach

The mathematical approach

- Design a mathematical framework in which data science methods can be analyzed

THE MATHEMATICAL APPROACH

- Design a mathematical framework in which data science methods can be analyzed
- Often several steps further on abstraction than even the benchmark problems

THE MATHEMATICAL APPROACH

- Design a mathematical framework in which data science methods can be analyzed
- Often several steps further on abstraction than even the benchmark problems
- Often the results are "loose", vacuous for the practical problems (e.g. worst case, infinite sample size)

The mathematical approach

- Design a mathematical framework in which data science methods can be analyzed
- Often several steps further on abstraction than even the benchmark problems
- Often the results are "loose", vacuous for the practical problems (e.g. worst case, infinite sample size)
- A new analysis technique can make your carrier: all methods can be reanalyzed

The mathematical approach

- Design a mathematical framework in which data science methods can be analyzed
- Often several steps further on abstraction than even the benchmark problems
- Often the results are "loose", vacuous for the practical problems (e.g. worst case, infinite sample size)
- A new analysis technique can make your carrier: all methods can be reanalyzed
- The coincidence of practical success of a method and its successful (but loose) mathematical analysis is often mistaken for causality, justification of the mathematical approach

THE SCIENTIFIC METHOD

THE SCIENTIFIC METHOD

- Model1 is well established, we are testing Model2

THE SCIENTIFIC METHOD

- Model1 is well established, we are testing Model2
- Model2 predicts <u>tufas</u>, Model1 doesn't

THE SCIENTIFIC METHOD

- Model1 is well established, we are testing Model2
- Model2 predicts tufas, Model1 doesn't
- We design an experiment/detector/analysis pipeline to generate and see tufas (if they exist)

THE SCIENTIFIC METHOD

- Model1 is well established, we are testing Model2
- Model2 predicts <u>tufas</u>, Model1 doesn't
- We design an experiment/detector/analysis pipeline to generate and see tufas (if they exist)
- If we see tufas, Model1 is invalidated

THE SCIENTIFIC METHOD

- Model1 is well established, we are testing Model2
- Model2 predicts tufas, Model1 doesn't
- We design an experiment/detector/analysis pipeline to generate and see tufas (if they exist)
- If we see tufas, Model1 is invalidated
- If Model2 has no competitors, it is accepted

THE SCIENTIFIC METHOD

- Model1 is well established, we are testing Model2
- Model2 predicts <u>tufas</u>, Model1 doesn't
- We design an experiment/detector/analysis pipeline to generate and see tufas (if they exist)
- If we see tufas, Model1 is invalidated
- If Model2 has no competitors, it is accepted
- Today's tufas are hard to generate, and our observation is noisy, so tufas + noise can look like known objects (say birds). All we can say that if Model1 is valid, the number of tufa-looking birds is significantly smaller than the number of tufas we see.

A LESS CYNICAL VIEW:
WE DESIGN DISCOVERY PROCESSES