Enabling large scale data exploration and analysis for direct feedback to healthcare centers

Florian Guitton
Data Science Institute
Imperial College
London

Data Science Institute

Data Science Institute aims to enhance Imperial's excellence in data-driven research across its faculties

Multi disciplinary environment

The Institute’s staff contributes to more then 67 grants projects in various field and is enhanced by a large groups of Fellows and their groups running parallel high impact research

UK-Biobank

UK Biobank is a national and international health resource with unparalleled research opportunities, open to all bona fide health researchers.

Large publicly available cohort

More than 500 000 participants involved at different stage of the data collection process

Rich range of data modalities

More than 6000 individual variables collected longitudinally over a 5 years period in diverse modality format including genetic and detailed imaging data

Imperial College NHS Trust

One of the largest NHS trusts in England and together with Imperial College London forms one of the most powerful national academic health science centre.

Large secondary and tertiary care cohort

More than 1 200 000 patients attending the wards of the 6 Imperial College NHS Trust hospitals

Rich range of data modalities

Entire patient medical history jointed with biased, disease specific data modalities.

Many challenges ahead

Data Standards

Adopting and supporting data standards across the platform, i.e., CDISC compliant

Complex Variable Selection

Supporting advanced subset selection criteria based on run-time derived variables

Cross-Study Comparison

Supporting analysis and comparison of data from different studies and potentially different and heterogeneous data sources

Large High-dimensional Datasets

Managing and manipulating large high-dimensional datasets such as genotype or imaging data

Scalable Distributed Computation

Efficiently distributing and executing intensive analysis workflows

Architecture Considerations

Opportunity to start reconsidering the application stack from the ground up re-evaluating technologies and tools.

Modern web application

Based on the most modern and popular web framework, we are making Borderline extendible on many levels. The whole interface is pluggable and integrates with rich in-the-browser code editing

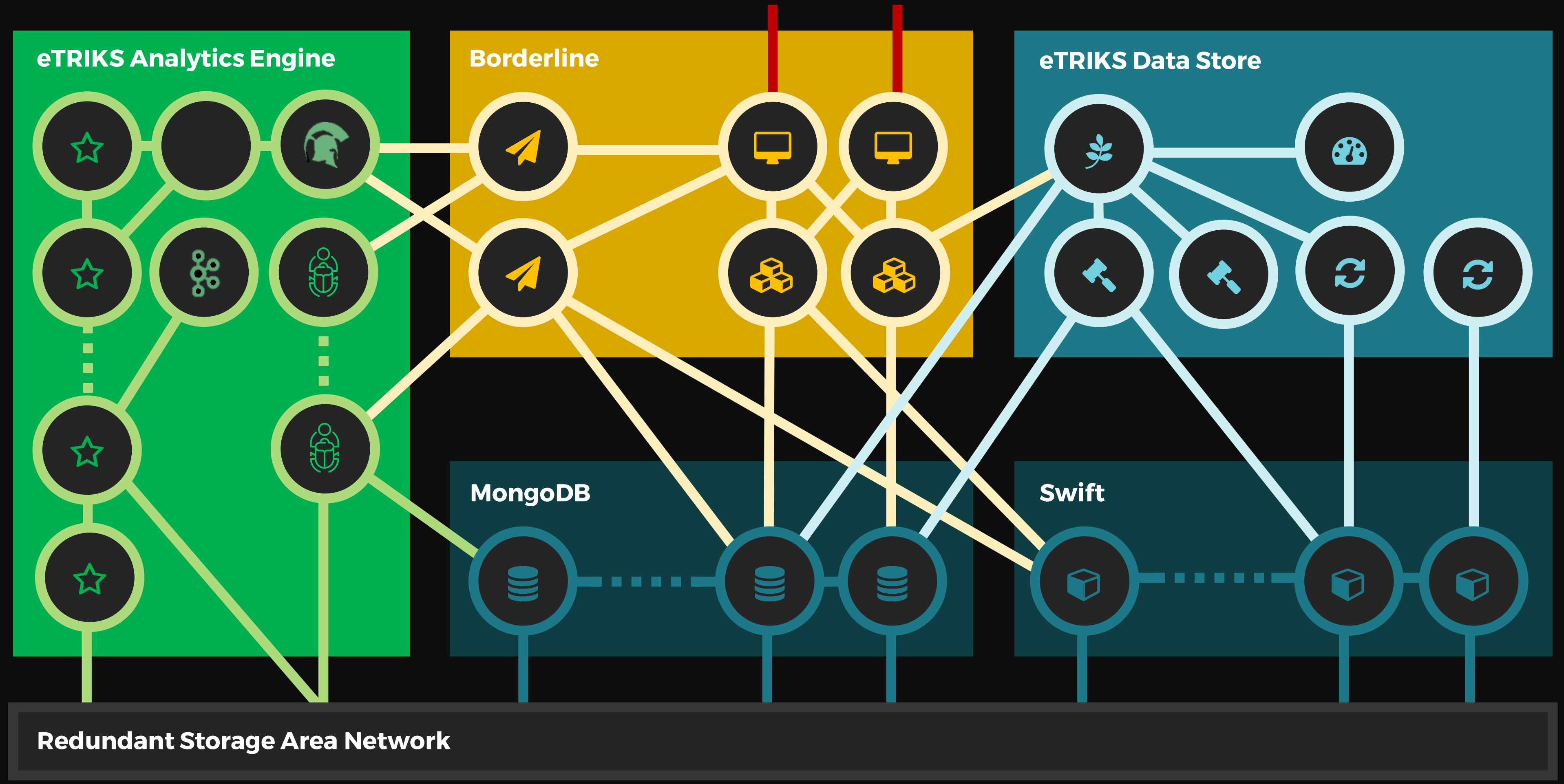
Microservice Architecture

The backend of Borderline and the components it relies on are constructed around the principle of microservices, dividing software in small executables

System Security

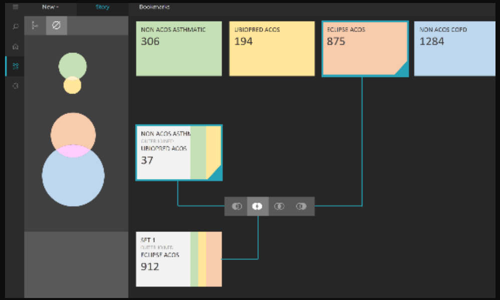
Data safety as a first citizen. Building on the work done with eTRIKS Horde and its bidirectional hybrid encryption system.

Microservice Architecture for Scalability

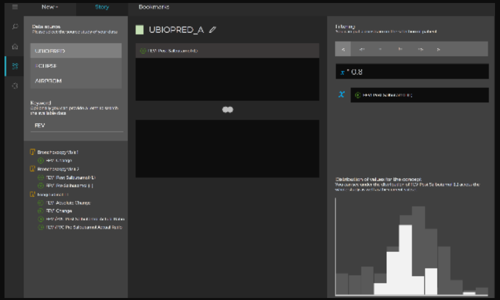


Borderline

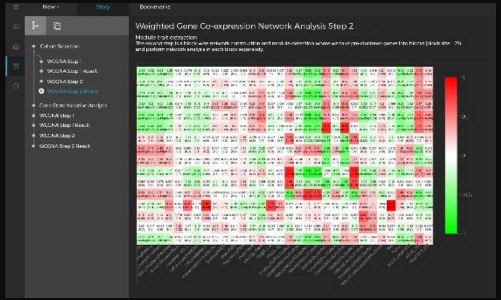
Developing user oriented tooling



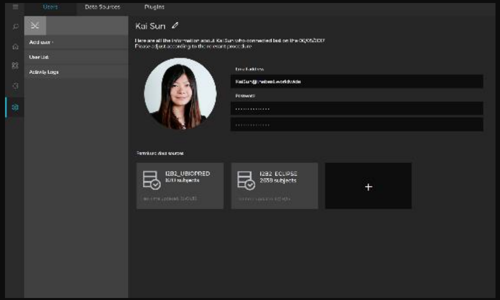
Logic operation on cohorts



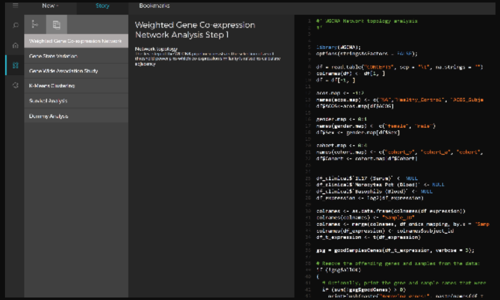
Familiar variable constraining



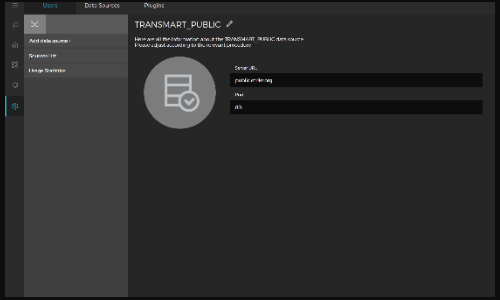
Data exploration in “stories”



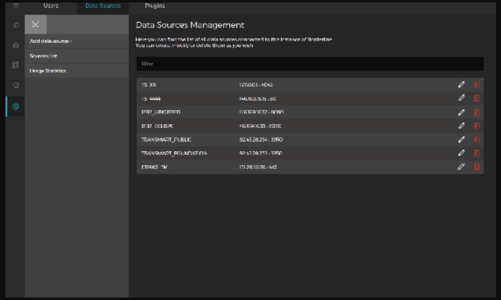
Centralise user delegation



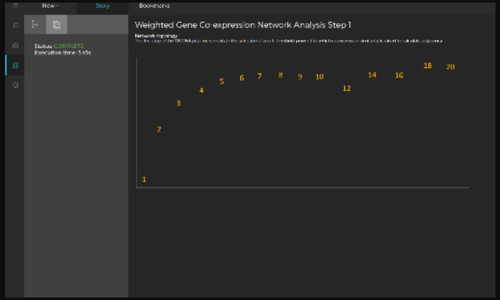
In-browser code edition



Multiple data sources



Rich extensibility



Audit and reporting

Infrastructure progress

Large investments planned over the next 5 years to allow 300 researchers to get access and analyze their data.

First “stones” already in place

More than 50 servers and 500TB of storage supported by a 80Gb/s network have already been deployed for the project following a £ 440.000 investment.

Scaling to ever-growing numbers

The project it currently dealing with 140TB of preliminary data and targets the handling of up to 1PB by March 2020

The team

Florian Guitton

Frontend development

General architecture and web from React to the coding of the RxJS backed event bus and built in IDE

Jean Grizet

Backend development

Scalability design and layer data oriented middleware abstraction propelled by NodeJS and top knowledge

Pierre-Marie Danieau

Backend development

Scalability design and layer data oriented middleware abstraction propelled by NodeJS and top knowledge

Axel Oehmichen

eTRIKS Analytical Engine

The art of large scale data distribution and computation from the Spark supported eTRIKS Analytical Engine

Ibrahim Emam

eTRIKS Data Store

Brining on board the expertise of data standard management and manipulation

Borderline

Resources

https://borderline.zone