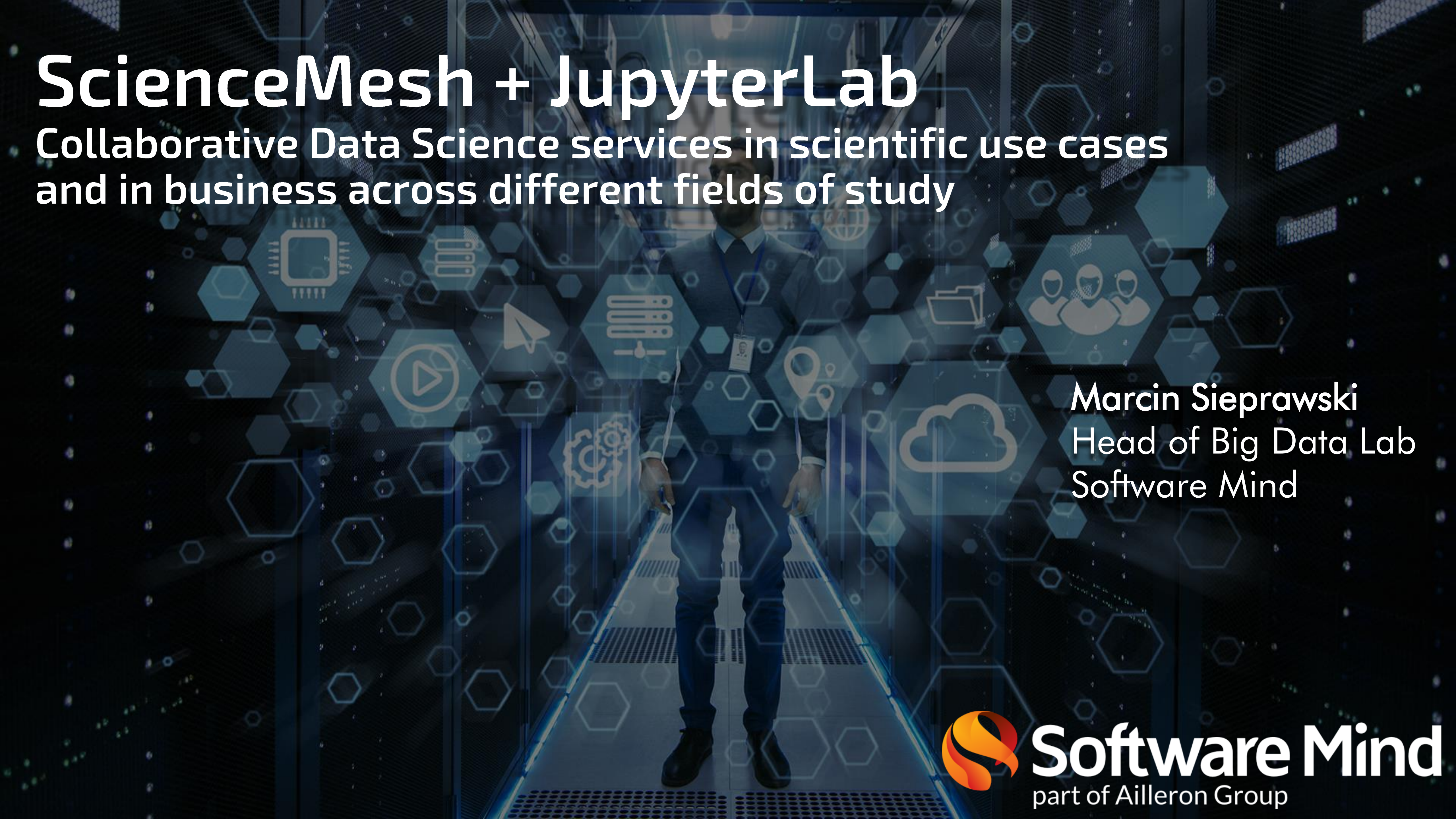


ScienceMesh + JupyterLab

Collaborative Data Science services in scientific use cases and in business across different fields of study

Marcin Sieprawski
Head of Big Data Lab
Software Mind

Software Mind
part of Ailleron Group

Software Mind

Software house focused on building **dedicated teams**
to extend **product engineering**
and **digital transformation** capabilities

4
R&D Labs in Poland

HQ in Cracow
Branches: Warsaw,
Rzeszow Bielsko-Biala

20
Years of experience
established in 1999

Branches
and representations

USA, Australia, Singapore,

Big enough to scale, small enough to care

ScienceMesh

CS3MESH4EOSC project

Reva — Cloud inter-operability platform

cloud-native

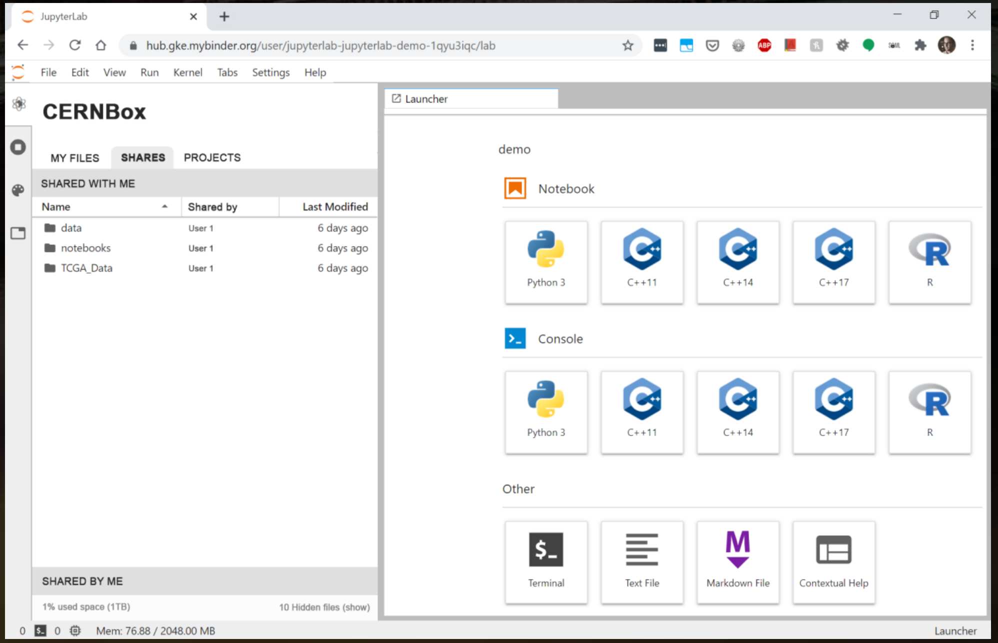
Distributed Data Science environments

JupyterLab extension (CS3 APIS)

- **Leading tasks on**
  - Reference inter-operability platform
  - Distributed Data Science environments

- ScienceMesh Inter-operability platform
  - make cloud storage and application providers inter-operable, via the CS3 APIS

- JupyterLab extension (Cs3Api4Lab)
  - Integration with ScienceMesh IOP (CS3 APIS)
  - replaces the default file manager
  - new UI elements for share functionalities



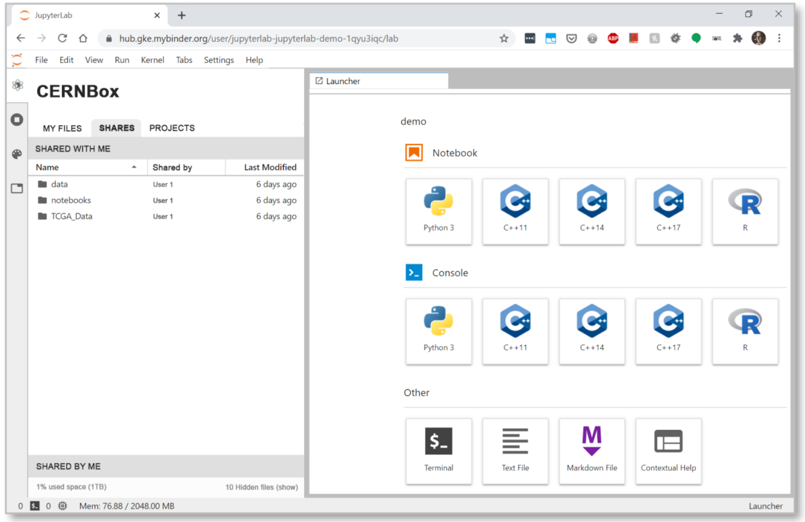
Name	Shared by	Last Modified
data	User 1	6 days ago
notebooks	User 1	6 days ago
TCGA_Data	User 1	6 days ago

1% used space (1TB)   10 Hidden files (show)

0   0   Mem: 76.88 / 2048.00 MB

JupyterLab extension (**Cs3Api4Lab**)

- Integration with ScienceMesh IOP (CS3 APIS)



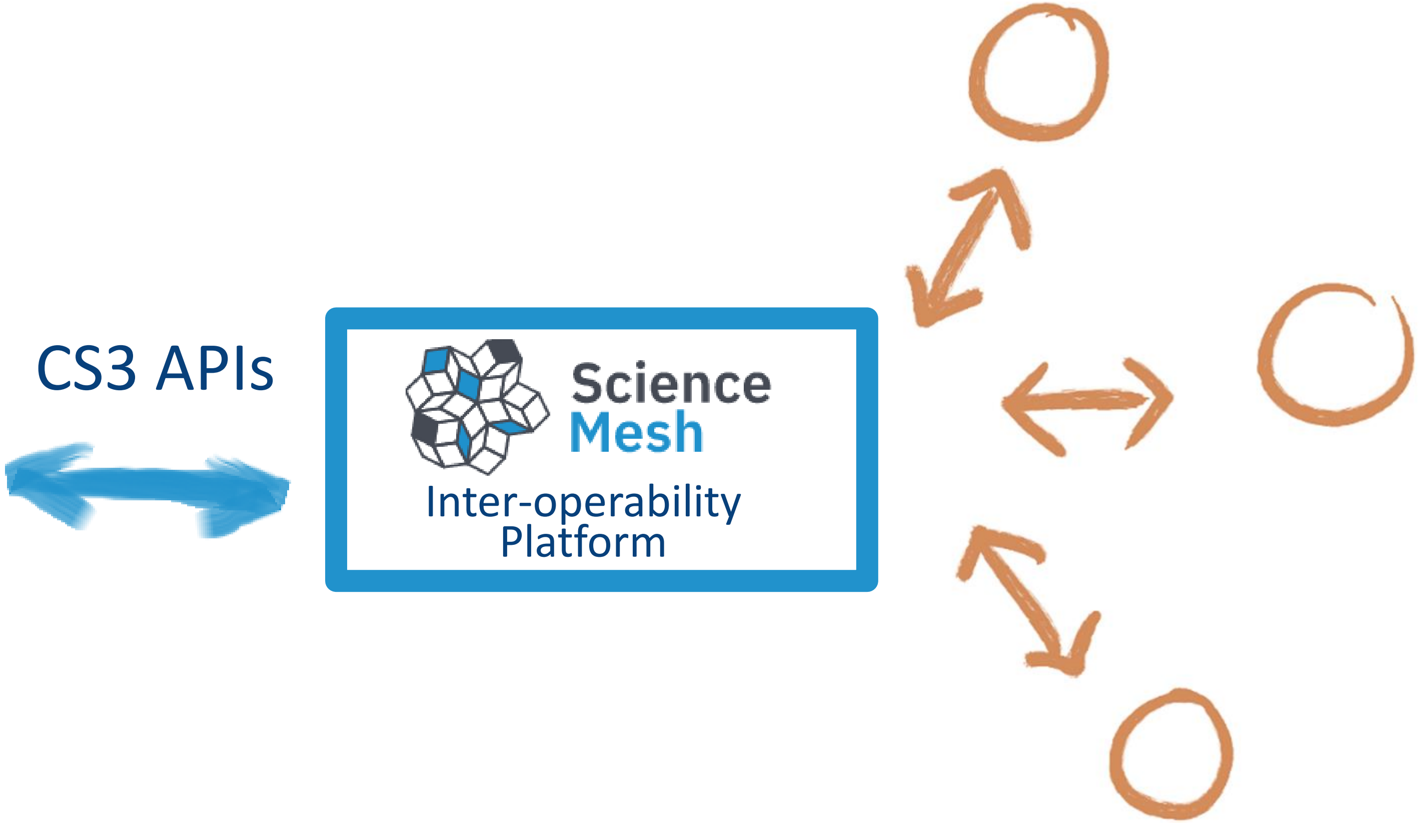
SHARED WITH ME

Name	Shared by	Last Modified
data	User 1	6 days ago
notebooks	User 1	6 days ago
TCGA_Data	User 1	6 days ago

SHARED BY ME

1% used space (1TB) 10 Hidden files (show)

jupyterlab

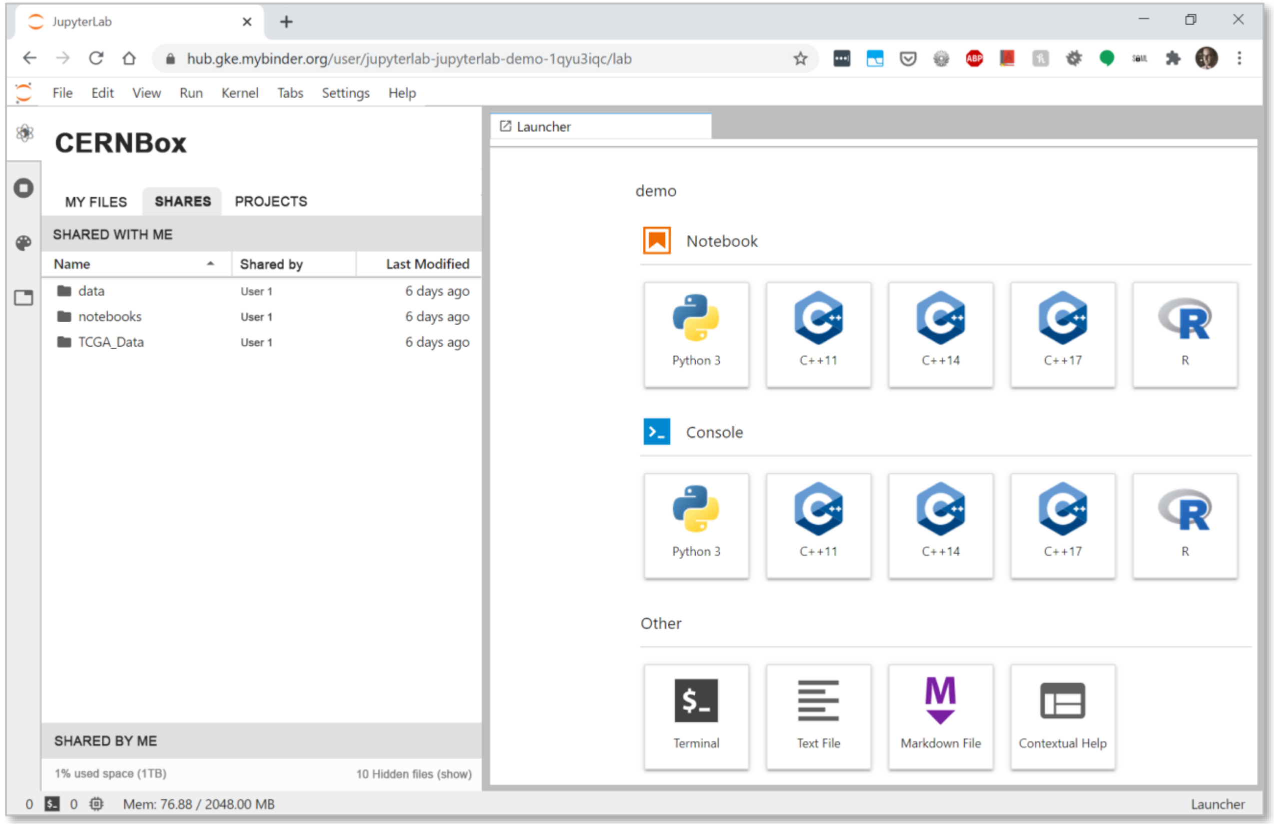

CS3 APIs

Science Mesh

Inter-operability Platform

JupyterLab extension (**Cs3Api4Lab**): Frontend

- Full client in Lab
- File browser – share functionalities
  - Shared by/with tab
  - Sharing buttons
  - Entries in the context menus
  - Pop-up windows: file information and sharing status
  - Account info
- File browsing



SHARED WITH ME

Name	Shared by	Last Modified
data	User 1	6 days ago
notebooks	User 1	6 days ago
TCGA_Data	User 1	6 days ago

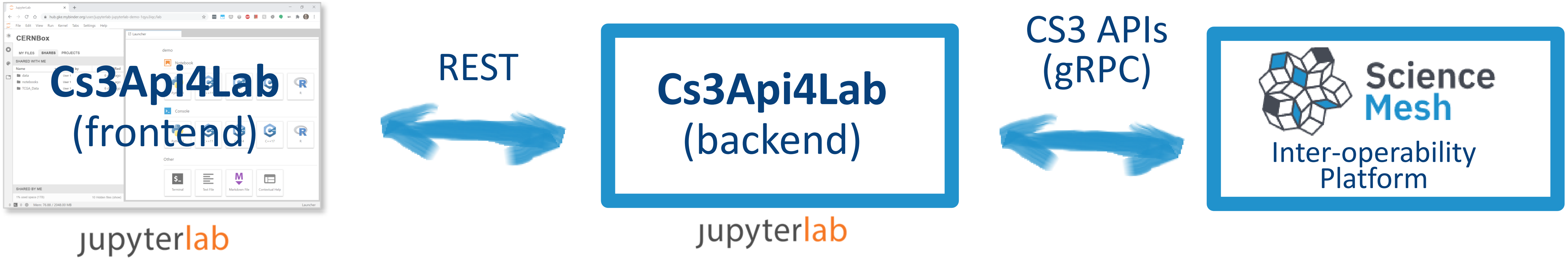
SHARED BY ME

1% used space (1TB) 10 Hidden files (show)

0 0 Mem: 76.88 / 2048.00 MB Launcher

JupyterLab extension (**Cs3Api4Lab**): Backend

- Replaces ContentsManager and Checkpoints
- REST endpoints for integration with the frontend:
  - API for content operations
  - API for checkpoints operations (todo)
  - API for share operations
- Connecting IOP: gRPC (CS3 APIs)



Cs3Api4Lab (frontend)

jupyterlab

REST

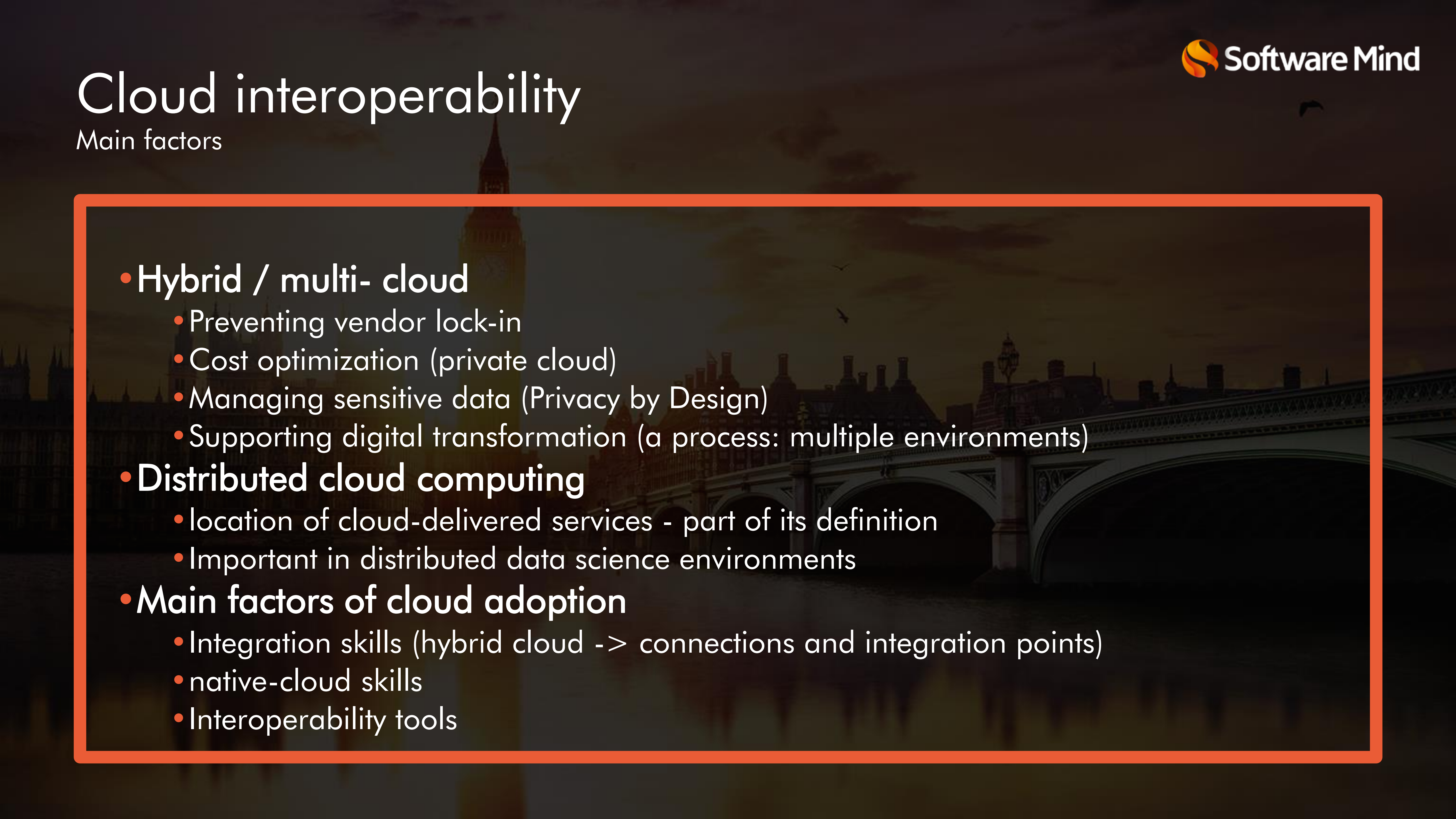
Cs3Api4Lab (backend)

jupyterlab

CS3 APIs (gRPC)

Science Mesh

Inter-operability Platform

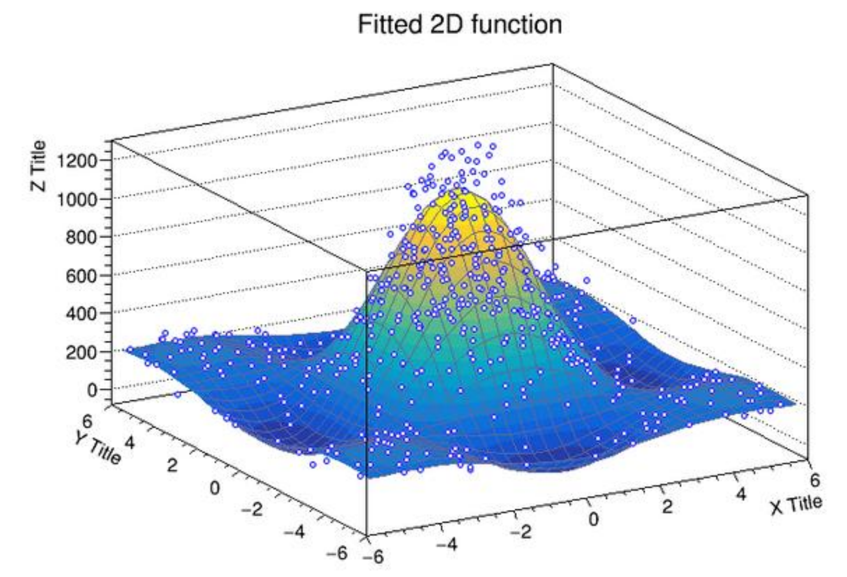


Cloud interoperability

Main factors

- **Hybrid / multi- cloud**
  - Preventing vendor lock-in
  - Cost optimization (private cloud)
  - Managing sensitive data (Privacy by Design)
  - Supporting digital transformation (a process: multiple environments)
- **Distributed cloud computing**
  - location of cloud-delivered services - part of its definition
  - Important in distributed data science environments
- **Main factors of cloud adoption**
  - Integration skills (hybrid cloud -> connections and integration points)
  - native-cloud skills
  - Interoperability tools

Collaborative Data Science: Beyond High Energy Physics



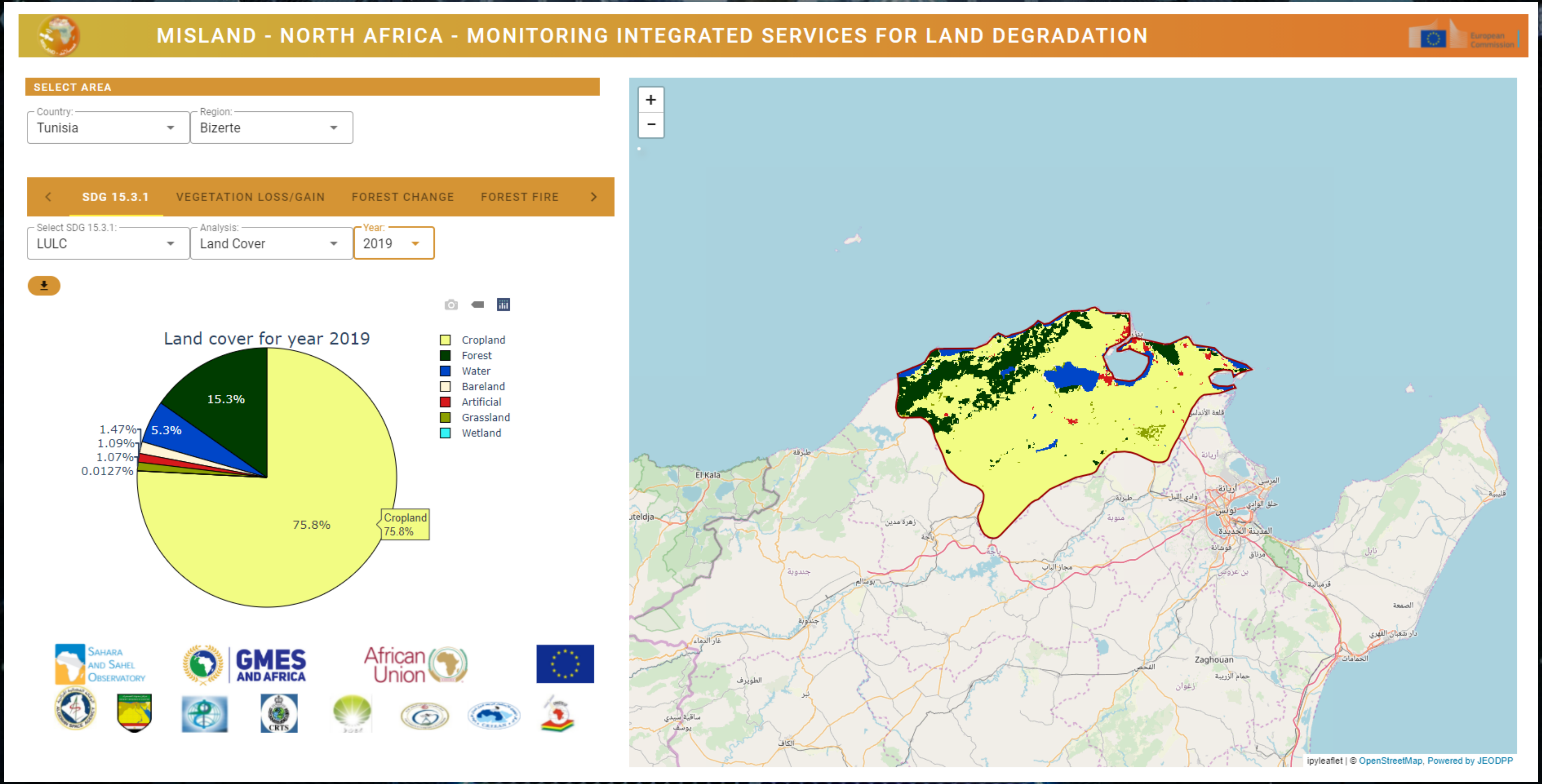
- All scientific disciplines nowadays are data-driven
  - Data analytics play an increasing role in all types of research
  - Distributed data science environments => all fields of study
  - A more effective collaboration between scientific institutions
- Business: develop new products in all sectors
  - Finance, IoT, SmartCities, energy and many others
- Gartner - Critical Capabilities for Data Science and Machine Learning Platforms
  - (13 March 2021)
  - **By 2023, 30% of organizations will harness the collective intelligence** of their analytics communities, outperforming competitors that rely solely on centralized analytics or self-service.
  - **By 2024, 70% of enterprises will use cloud and cloud-based AI infrastructure** to operationalize AI, thereby significantly alleviating concerns about integration and upscaling.

ScienceMesh

CS3MESH4EOSC project

Collaborative Data Science

Earth Observation use case



MISLAND - NORTH AFRICA - MONITORING INTEGRATED SERVICES FOR LAND DEGRADATION

SELECT AREA

Country: Tunisia

Region: Bizerte

SDG 15.3.1 | VEGETATION LOSS/GAIN | FOREST CHANGE | FOREST FIRE

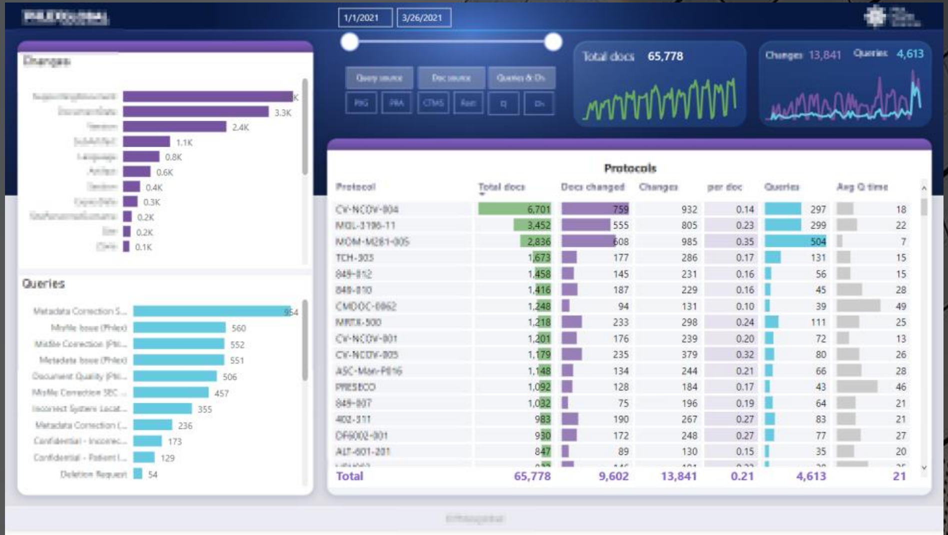
Select SDG 15.3.1: LULC

Analysis: Land Cover

Year: 2019

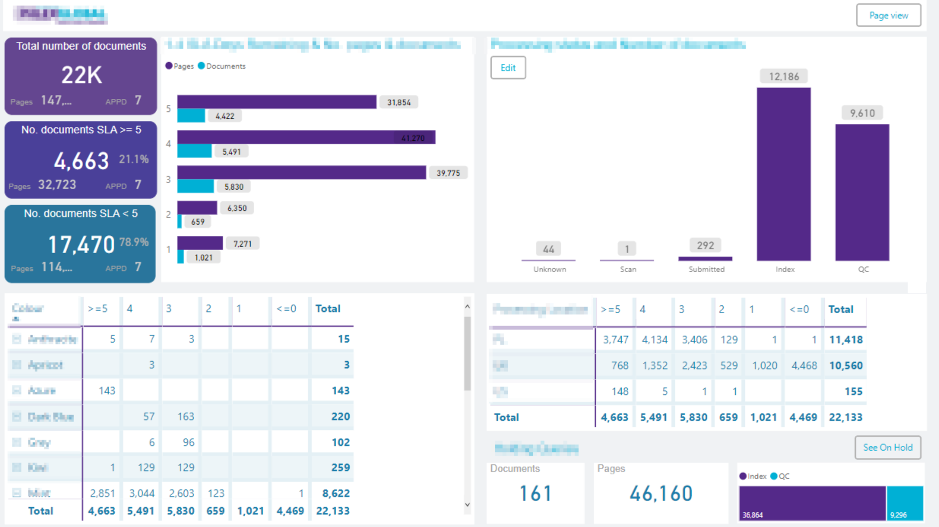
Software Mind

PowerBI


[redacted]
1/1/2021
3/26/2021
Total docs 65,778
Changes 13,841
Queries 4,613
Changes
3.3K
2.4K
1.1K
0.8K
0.6K
0.4K
0.3K
0.2K
0.2K
0.1K
Queries
Metadata Correction S... 954
Misfile Issue (Pfilex) 560
Misfile Correction (Pfil... 552
Metadata Issue (Pfilex) 551
Document Quality (Pfil... 506
Misfile Correction SEC ... 457
Incorrect System Locat... 355
Metadata Correction (... 236
Confidential - Incorrec... 173
Confidential - Patient I... 129
Deletion Request 54

Protocols

Protocol	Total docs	Docs changed	Changes	per doc	Queries	Avg Q time
CV-NCOV-004	6,701	759	932	0.14	297	18
MGL-3196-11	3,452	555	805	0.23	299	22
MOM-M281-005	2,836	608	985	0.35	504	7
TCH-303	1,673	177	286	0.17	131	15
849-012	1,458	145	231	0.16	56	15
848-010	1,416	187	229	0.16	45	28
CMDOC-0062	1,248	94	131	0.10	39	49
MRTX-500	1,218	233	298	0.24	111	25
CV-NCOV-001	1,201	176	239	0.20	72	13
CV-NCOV-005	1,179	235	379	0.32	80	26
ASC-Man-P016	1,148	134	244	0.21	66	28
PRESBCO	1,092	128	184	0.17	43	46
849-007	1,032	75	196	0.19	64	21
402-311	983	190	267	0.27	83	21
OF6002-001	930	172	248	0.27	77	27
ALT-601-201	847	89	130	0.15	35	20
Total	65,778	9,602	13,841	0.21	4,613	21



Total number of documents

22K

Pages 147,... APPD 7

No. documents SLA >= 5

4,663 21.1%

Pages 32,723 APPD 7

No. documents SLA < 5

17,470 78.9%

Pages 114,... APPD 7

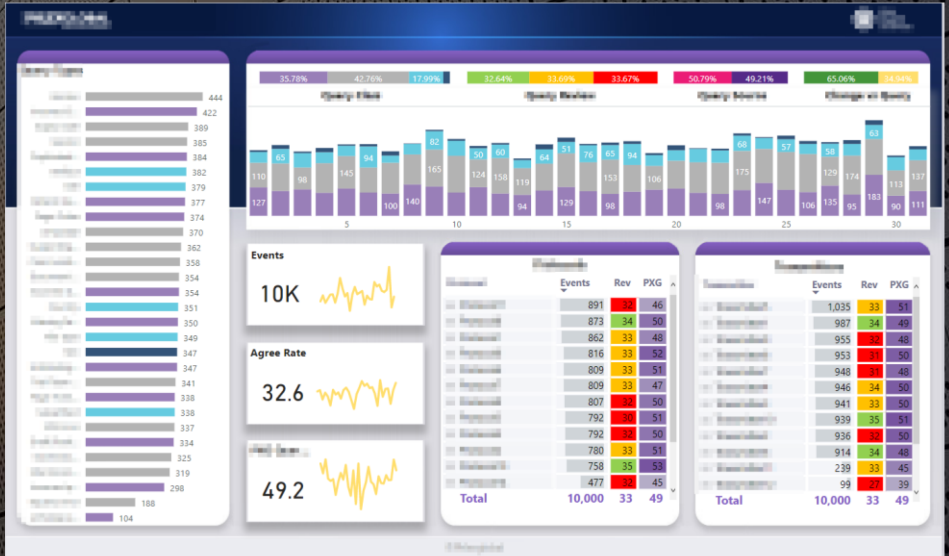
[redacted]	>=5	4	3	2	1	<=0	Total
[redacted]	5	7	3				15
[redacted]		3					3
[redacted]	143						143
[redacted]		57	163				220
[redacted]		6	96				102
[redacted]	1	129	129				259
[redacted]	2,851	3,044	2,603	123		1	8,622
Total	4,663	5,491	5,830	659	1,021	4,469	22,133

[redacted]	>=5	4	3	2	1	<=0	Total
[redacted]	3,747	4,134	3,406	129	1	1	11,418
[redacted]	768	1,352	2,423	529	1,020	4,468	10,560
[redacted]	148	5	1	1			155
Total	4,663	5,491	5,830	659	1,021	4,469	22,133

[redacted]

Documents 161

Pages 46,160



Events

10K

Agree Rate

32.6

49.2

[redacted]	Events	Rev	PXG
[redacted]	891	32	46
[redacted]	873	34	50
[redacted]	862	33	48
[redacted]	816	33	52
[redacted]	809	33	51
[redacted]	809	33	47
[redacted]	807	32	50
[redacted]	792	30	51
[redacted]	792	32	50
[redacted]	780	33	51
[redacted]	758	35	53
[redacted]	477	32	45
Total	10,000	33	49

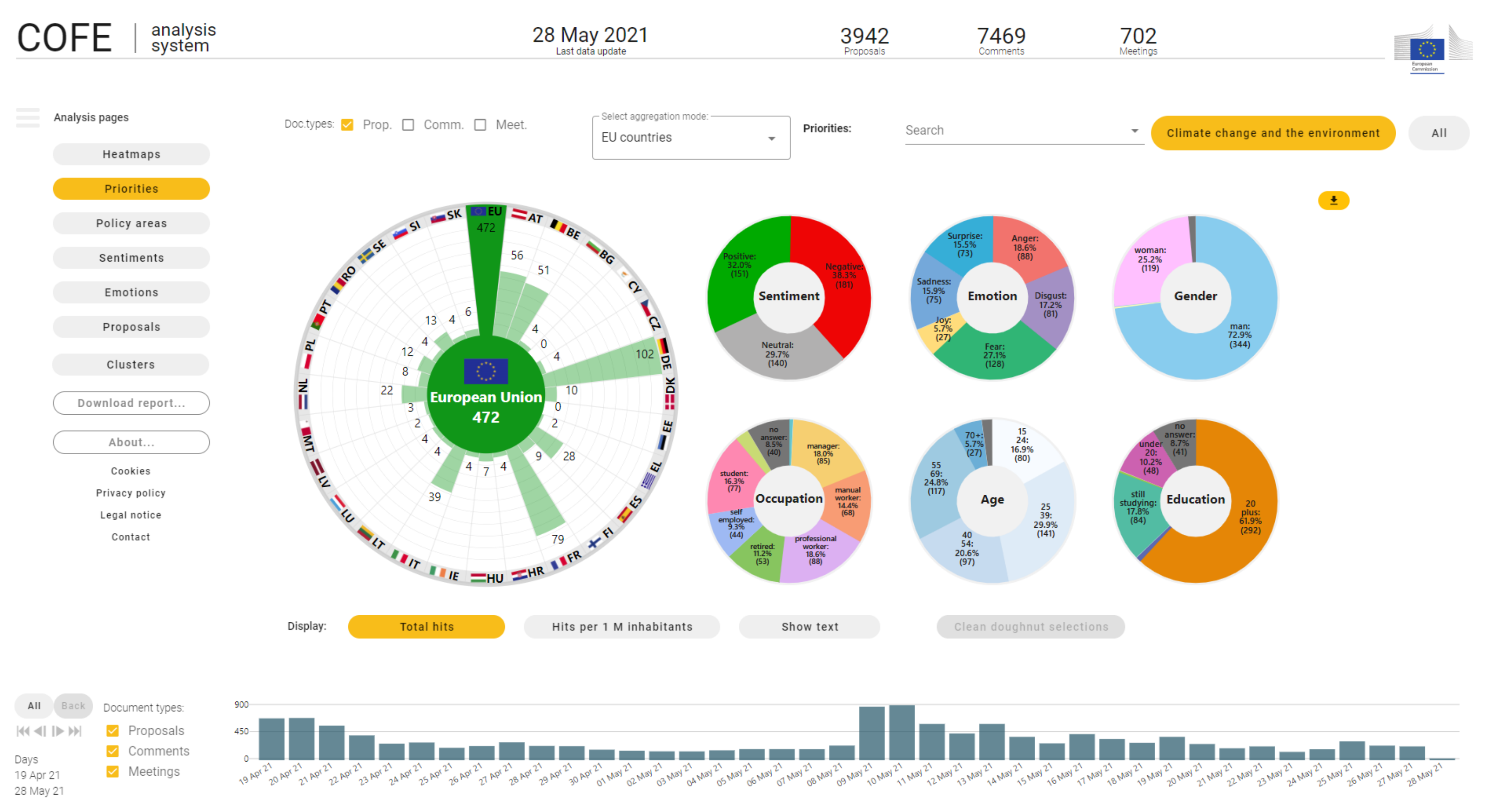
[redacted]	Events	Rev	PXG
[redacted]	1,035	33	51
[redacted]	987	34	49
[redacted]	955	32	48
[redacted]	953	31	50
[redacted]	948	31	48
[redacted]	946	34	50
[redacted]	941	33	50
[redacted]	939	35	51
[redacted]	936	32	50
[redacted]	914	34	48
[redacted]	239	33	45
[redacted]	99	27	39
Total	10,000	33	49

Replacing BI tools with Jupyter

Analysys of the proposals and comments the Conference on the Future of Europe

Dashboard to analyse the proposals and comments inserted by european citizens in the Conference on the Future of Europe (https://futureu.europa.eu/).

Vue/Vuetify Material Design



COFE | analysis system

28 May 2021 Last data update

3942 Proposals

7469 Comments

702 Meetings

Doc.types: ☑ Prop. ☐ Comm. ☐ Meet.

Select aggregation mode: EU countries

Priorities: Search | Climate change and the environment | All

Display: Total hits

Days 19 Apr 21 28 May 21

Document types: ☑ Proposals ☑ Comments ☑ Meetings

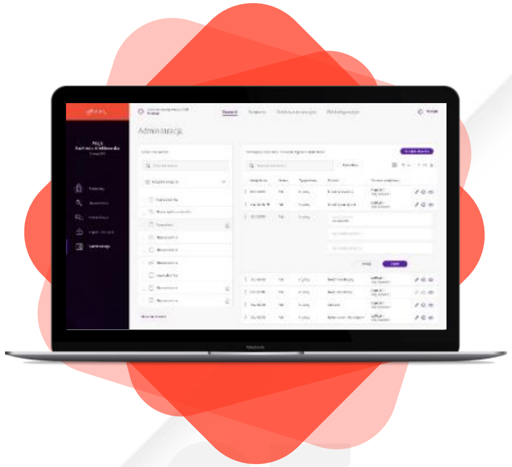
Ailleron offers a comprehensive
DIGITAL BANKING PLATFORM

01

Omni-channel platform providing flawless user experience

02

Ready to go capabilities – value added services



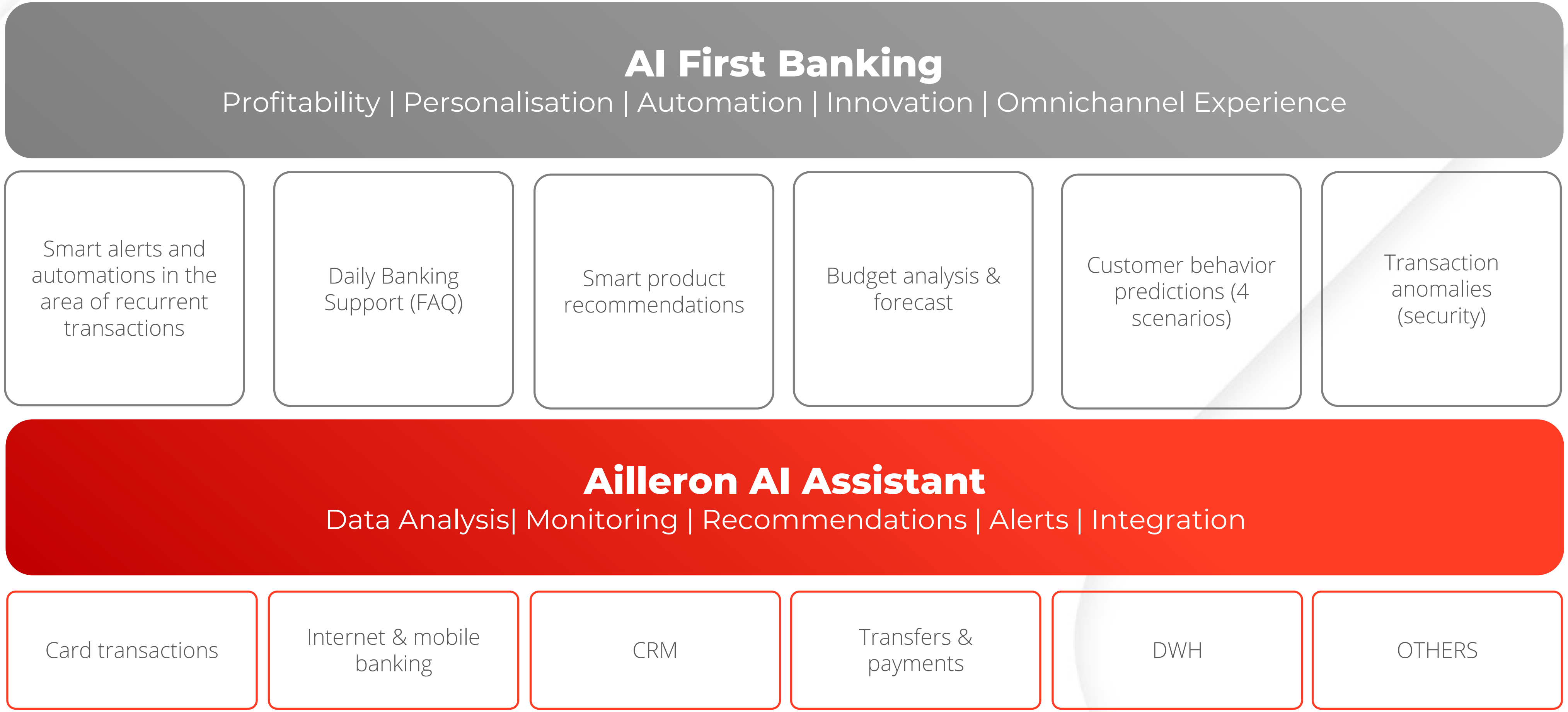
03

Back-office modules which support retail services administration

with Value Added Modules

01 Digital Onboarding (eKYC) > 02 Meeting Scheduler > 03 Subscription Manager > 04 Notifications & Campaign Manager > 05 Robo Assistant

AI Bank
SOLUTION



AI First Banking

Profitability | Personalisation | Automation | Innovation | Omnichannel Experience

Smart alerts and automations in the area of recurrent transactions

Daily Banking Support (FAQ)

Smart product recommendations

Budget analysis & forecast

Customer behavior predictions (4 scenarios)

Transaction anomalies (security)

Ailleron AI Assistant

Data Analysis| Monitoring | Recommendations | Alerts | Integration

Card transactions

Internet & mobile banking

CRM

Transfers & payments

DWH

OTHERS

Selected case studies

Startup: Electronic Identity Protection Platform

Web-scale Semantic Web startup (2005-2008)
commercializing new technology

Bare-metal cluster

Hadoop

Natural Language Processing

Main Facts:

Customer
UK Startup

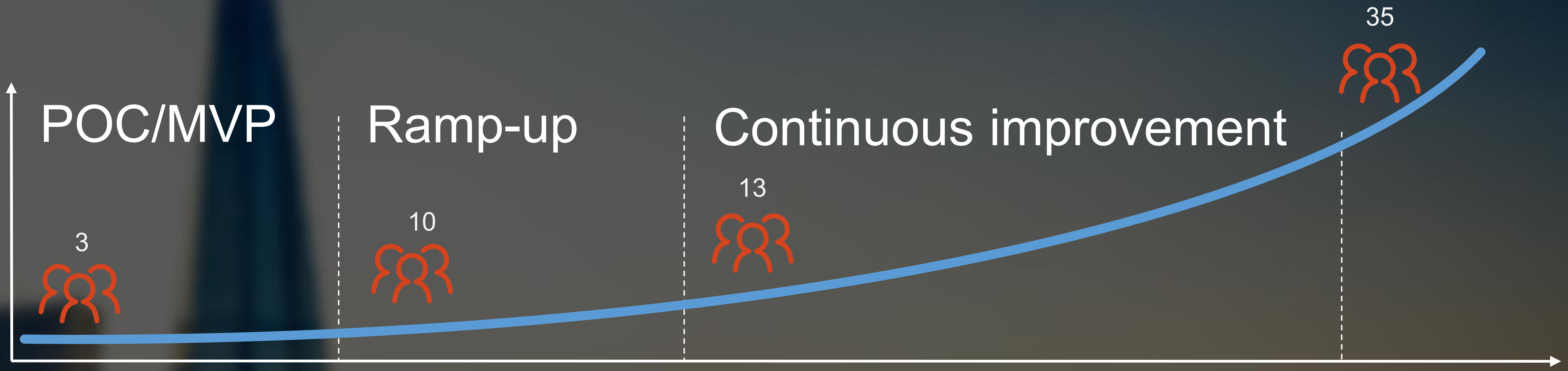
Project
Building new product from scratch – electronic identity protection platform

Operating Model
Dedicated Teams, MVP approach, R&D with universities (University of Southampton, University of Sheffield)

Challenges:

- Transfer of innovative technologies (Semantic Web, NLP,Graph Databases) from universities to real business
- Scaling the niche technology to fully blown commercial solution (e.g. browsing 4 billion of web pages)
- Building Big Data solution even before this term was defined

Software Mind Digital Transformation Services



POC/MVP

Introduced set of practices on top of SCRUM framework to reduce time-to-market

- Agile workshops to set product boundaries easily

Ramp-up

Implementing processes focused on quality step by step

- TDD
- Continuous Integration & Continuous Delivery

Automated testing

- Getting rid of manual testing and Waterfall, implementing automated tests and Scrum, while keeping in mind legal regulations of drug processing

Continuous improvement

Introducing Infrastructure as Code (IoC)

- Automatically creating coherent, fast and firm environments (70% done)

Breaking monolith into microservices

- Minimizing dependency on tightly coupled code
- Extracting microservices from monolith structure
- Introduction of asynchronus message queues

Growing our teams up to 40

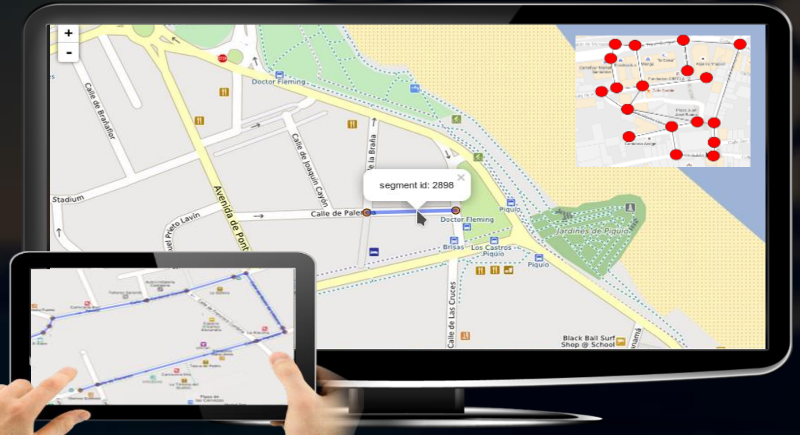
- True DevOps approach
- Fully automated Continoues Delivery pipeline
- Reducing time to market from 3 months to 2 weeks

SETA

Big Data - large scale smart mobility managent platform (2017-2019)

Private Cloud | Cloud - native | Apache Spark | Data Science environment

- Big Data technologies for monitoring and managing mobility in large metropolitan areas.
- Solution based on data from millions of citizens, thousands of connected cars, thousands of city sensors and hundreds of distributed databases

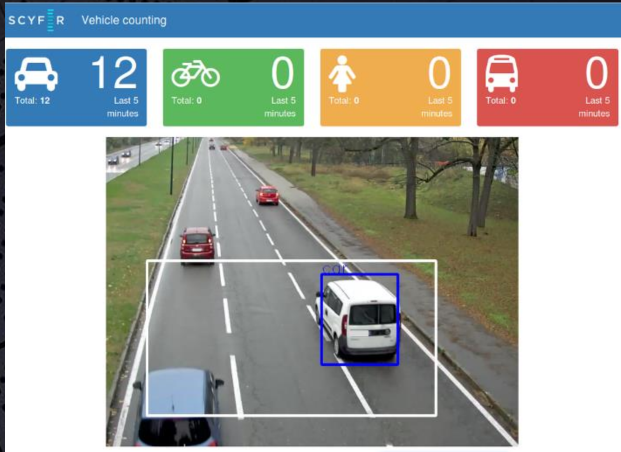


Main Facts

- Management of huge geospatial and spaciotemporal data
- GPU-based acceleration of geospatial indexes
- Privacy by Design
- High performance geo-located event processing engine
- Scalable backend for mobile aps
- Integrated machine learning components
- Data Science environment built into Agile software development process

Seta

Data Science Environment - examples

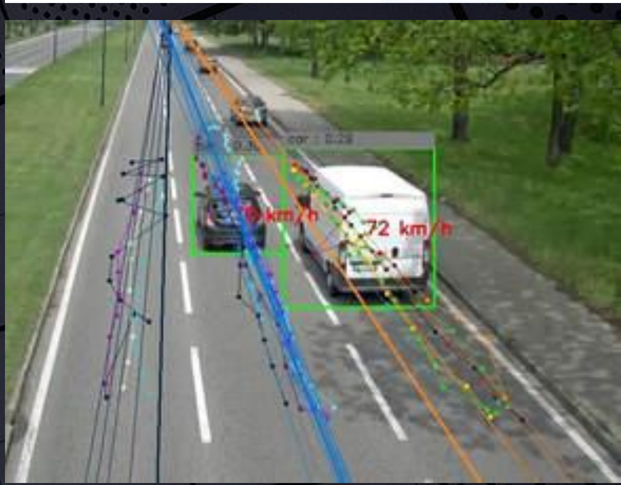


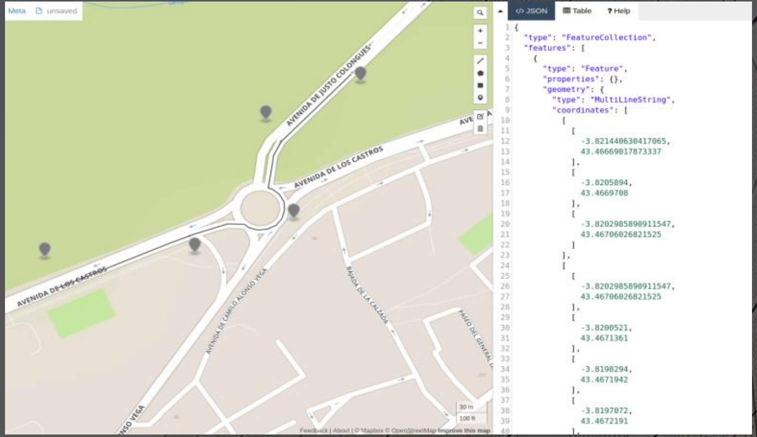
12 Total: 12 Last 5 minutes

0 Total: 0 Last 5 minutes

0 Total: 0 Last 5 minutes

0 Total: 0 Last 5 minutes





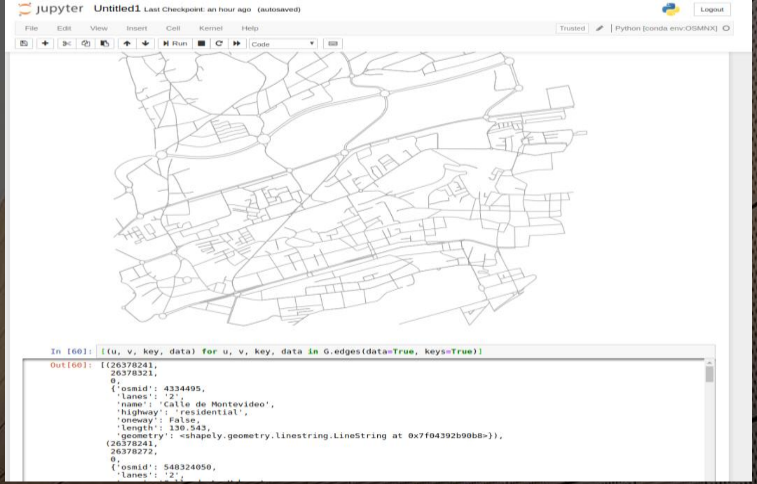
```
{
  "type": "FeatureCollection",
  "features": [
    {
      "type": "Feature",
      "properties": {},
      "geometry": {
        "type": "MultiLineString",
        "coordinates": [
          [
            [
              -3.821440630417065,
              43.46669017873337
            ],
            [
              -3.8205894,
              43.4669708
            ],
            [
              -3.8202985890911547,
              43.46706026821525
            ]
          ],
          [
            [
              -3.8202985890911547,
              43.46706026821525
            ],
            [
              -3.8200521,
              43.4671361
            ],
            [
              -3.8198294,
              43.4671942
            ],
            [
              -3.8197072,
              43.4672191
            ],
```

Geospatial visualisation + map matching

- low veracity of GPS tracks
- map matching: Hidden Markov models

Graph analysis/algorithms

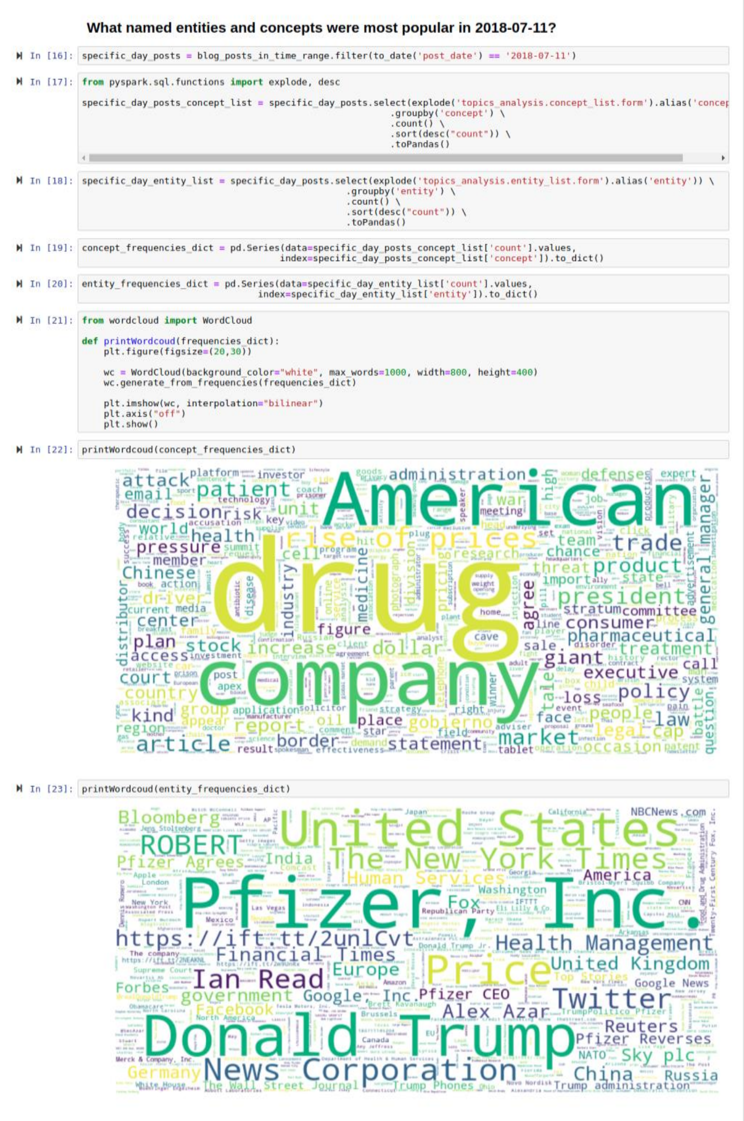
- Analysis and visualisation in the context of road infrastructure graph
- Jupyter Notebook + Spark/GraphX



```
In [60]: [(u, v, key, data) for u, v, key, data in G.edges(data=True, keys=True)]

Out[60]: [(26378241,
           26378321,
           0,
           {'osmid': 4334495,
            'lanes': '2',
            'name': 'Calle de Montevideo',
            'highway': 'residential',
            'oneway': False,
            'length': 130.543,
            'geometry': <shapely.geometry.linestring.LineString at 0x7f04392b90b8>}),
          (26378241,
           26378272,
           0,
           {'osmid': 548324050,
            'lanes': '2',
```

Data Science environment - FinTech



What named entities and concepts were most popular in 2018-07-11?

```
In [16]: specific_day_posts = blog_posts_in_time_range.filter(to_date('post_date') == '2018-07-11')
```

```
In [17]: from pyspark.sql.functions import explode, desc

         specific_day_posts_concept_list = specific_day_posts.select(explode('topics_analysis.concept_list.form').alias('concep
                                                                  .groupby('concept') \
                                                                  .count() \
                                                                  .sort(desc("count")) \
                                                                  .toPandas()
```

```
In [18]: specific_day_entity_list = specific_day_posts.select(explode('topics_analysis.entity_list.form').alias('entity')) \
                                                          .groupby('entity') \
                                                          .count() \
                                                          .sort(desc("count")) \
                                                          .toPandas()
```

```
In [19]: concept_frequencies_dict = pd.Series(data=specific_day_posts_concept_list['count'].values,
                                              index=specific_day_posts_concept_list['concept']).to_dict()
```

```
In [20]: entity_frequencies_dict = pd.Series(data=specific_day_entity_list['count'].values,
                                             index=specific_day_entity_list['entity']).to_dict()
```

```
In [21]: from wordcloud import WordCloud

         def printWordcoud(frequencies_dict):
             plt.figure(figsize=(20,30))

             wc = WordCloud(background_color="white", max_words=1000, width=800, height=400)
             wc.generate_from_frequencies(frequencies_dict)

             plt.imshow(wc, interpolation="bilinear")
             plt.axis("off")
             plt.show()
```

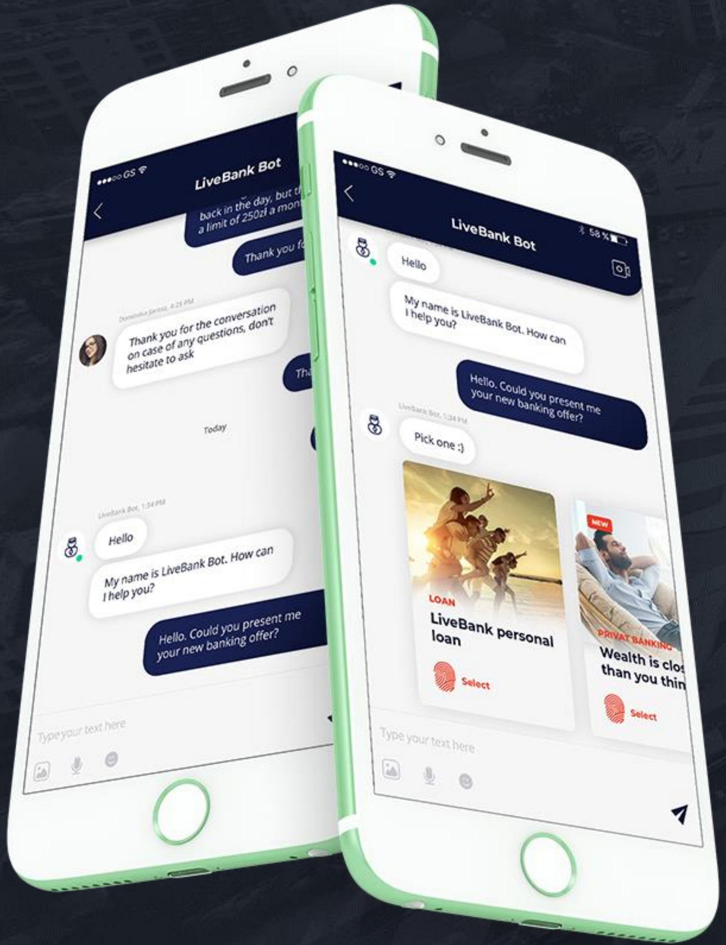
```
In [22]: printWordcoud(concept_frequencies_dict)
```

```
In [23]: printWordcoud(entity_frequencies_dict)
```

- Collecting blog posts
- Natural Language Processing
  - Named Entity regocnition
  - Concepts (topic extraction)
  - Insight Engine,
    - eg. patterns of named entities: distance between entities, entities in the same sentences, statistical patterns, etc.)
- Combining data
- Visualisation and data mining
- Text Classification
  - Binary, Multiple class, Multilabel, Complex taxonomy
  - common use cases
    - IPTC Subject Codes
    - EuroVoc
    - Business Reputation
    - IAB Taxonomy
    - Social Media

Software Mind

Conversational AI



back in the day, but t
a limit of 250zł a mon

Thank you f

Thank you for the conversation on case of any questions, don't hesitate to ask

Hello

My name is LiveBank Bot. How can I help you?

Hello. Could you present me your new banking offer?

Pick one :)

Wealth is clos
than you thin

Chatbot for financial institutions

- Ailleron product
- Dialogue-based machine learning (advanced conversations not just question and answer)
- Intent classification, Named Entity Recognition state-of-the-art NLP methods implemented and used
- An Advanced administration panel means adding and managing bot knowledge with ease

Software Mind

Thank you for your attention!

Marcin Sieprawski
Head of Big Data Lab
marcin.sieprawski@softwaremind.com

Headquarters
ul. Zyczkowskiego 20
31-864 Krakow

Tel.: + 48 12 252 34 00
web: www.softwaremind.com

Branches in Poland
Warsaw, Rzeszow, Bielsko-Biala

Worldwide
Australia / Sydney, Brisbane
Singapore / Singapore
USA / Boise, Dallas

Krakow, Poland