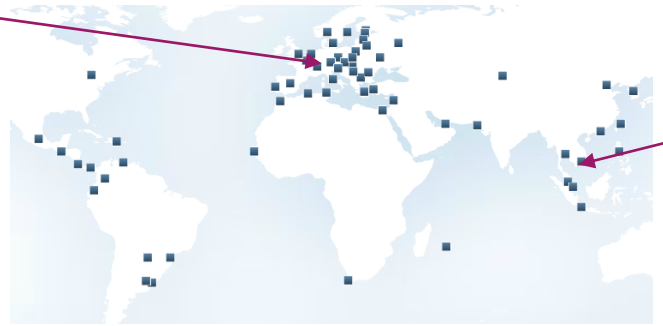
PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

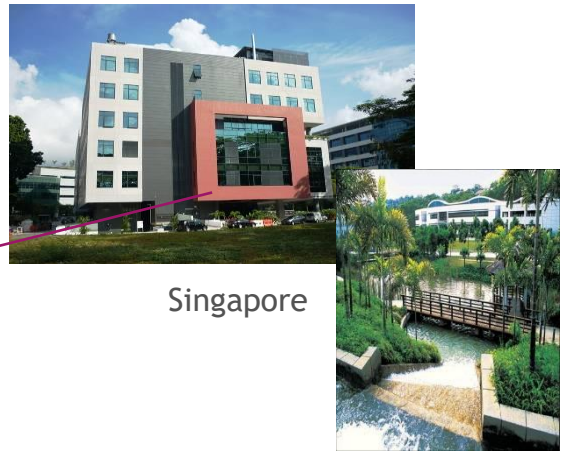
Building an Efficient R&D Chemo-Centric Data Repository System that Leads to Knowledge Discovery

***Pavel Pospisil*, Elyette Martin, Antonio Castellon, Mark Bentley**

Philip Morris R&D – large scientific company

Neuchatel, Switzerland



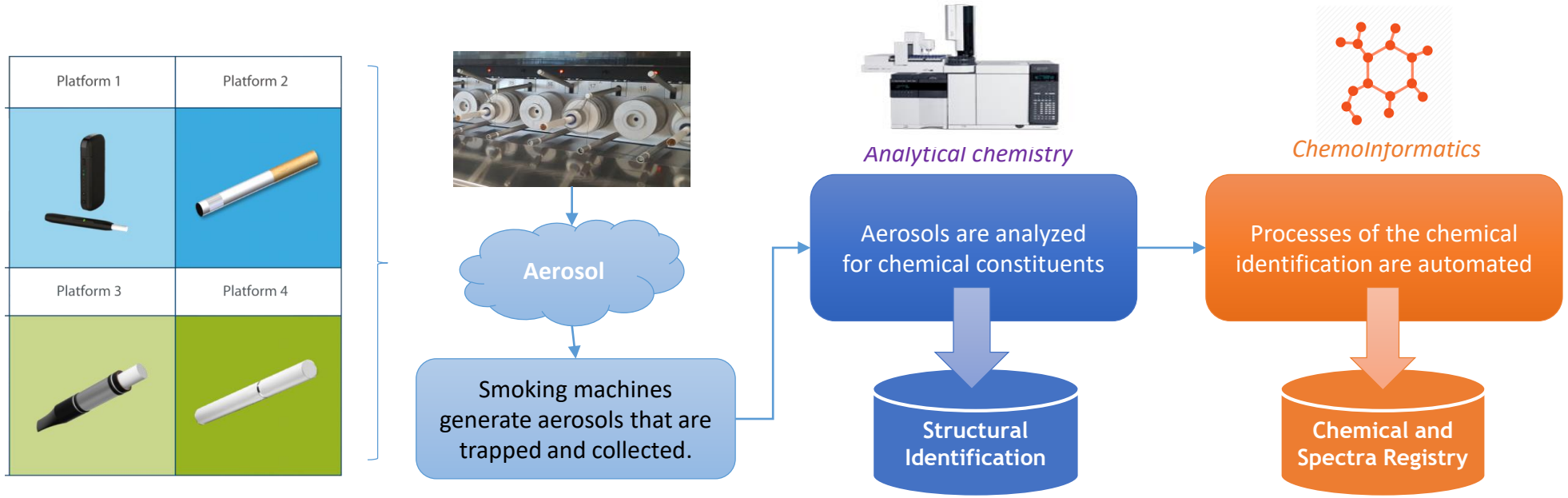


Singapore

- More than 300 scientists and engineers work at PMI R&D in the fields of biology, physics, chemistry, engineering, computational science, etc.
- As in many other R&D companies, PMI R&D deals with extremely high quantities of scientific data.
- Primary scientific data, including results from chemical, biological, and toxicological assays, are recorded in different formats, derived from scientific and laboratory management systems, data warehouses, and documents.

Example of large sets of chemical data

- At PMI, we analyze complex chemical matrices related to aerosols of our Reduced-Risk Products.

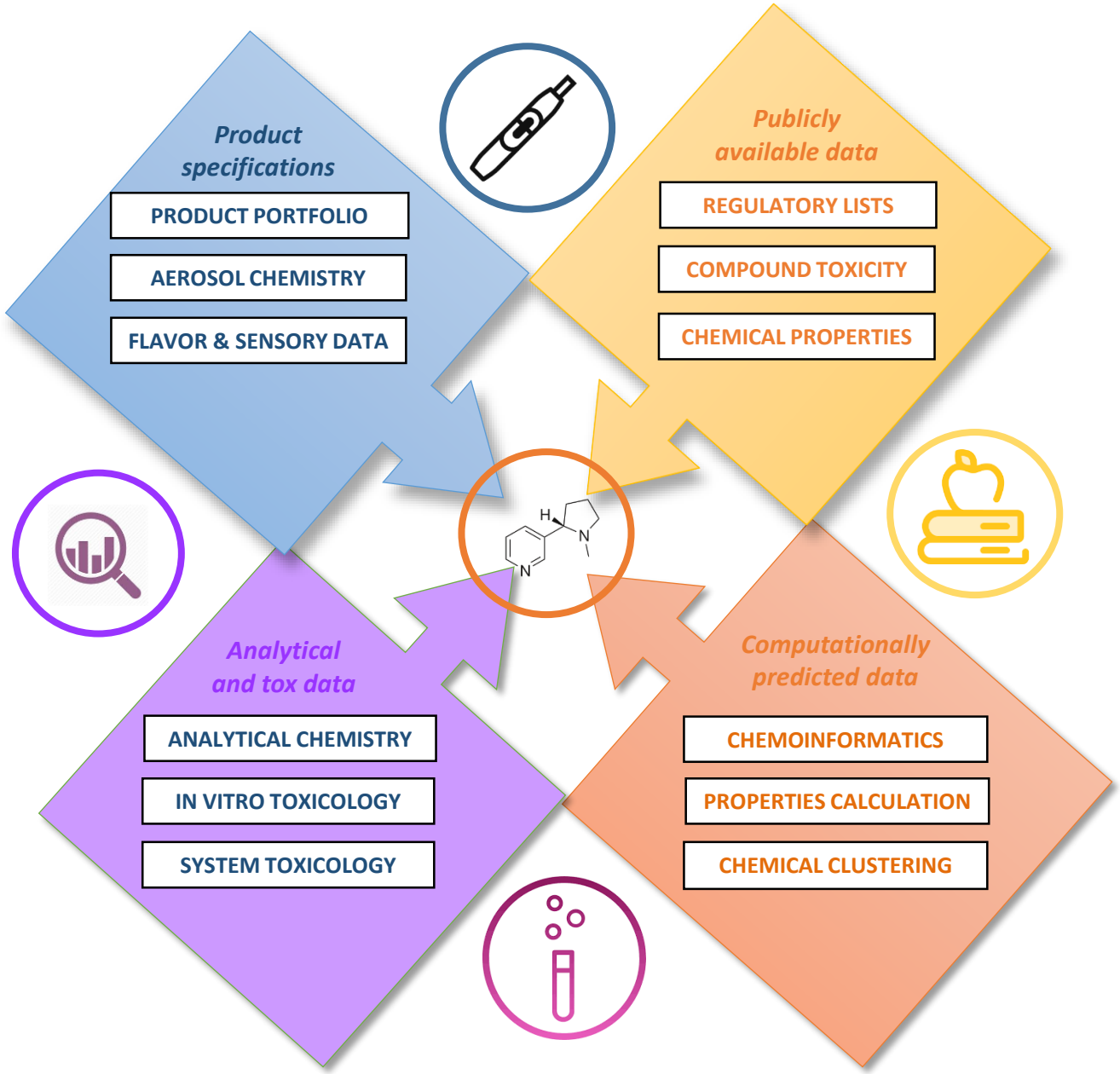


- Aerosol characterization generates large sets of chemical data, which are recorded in the LIMS, SDMS, and DWH fit-for-purpose systems as well as documents.

- In similar industries, trends have been to centralize data in a smart way. However, the centralization involves a significant effort to reformat the data and redesign the schemes.

→How can we provide the users with a convenient platform quickly that **overviews the data without these efforts of centralization?**

Type of data and its central node – the molecule



High diversity of data

- Product specifications, chemical structures
- Tox and biological endpoints
- Links to raw data lab systems
- Formats: tables, texts, files, and images
- Security: confidential vs. public

Our concept

- Integrate but not centralize
- Common denominator - chemical substance
- Dynamic: future modification-ready

Integrated ChemoInformatics Platform (ICIP)

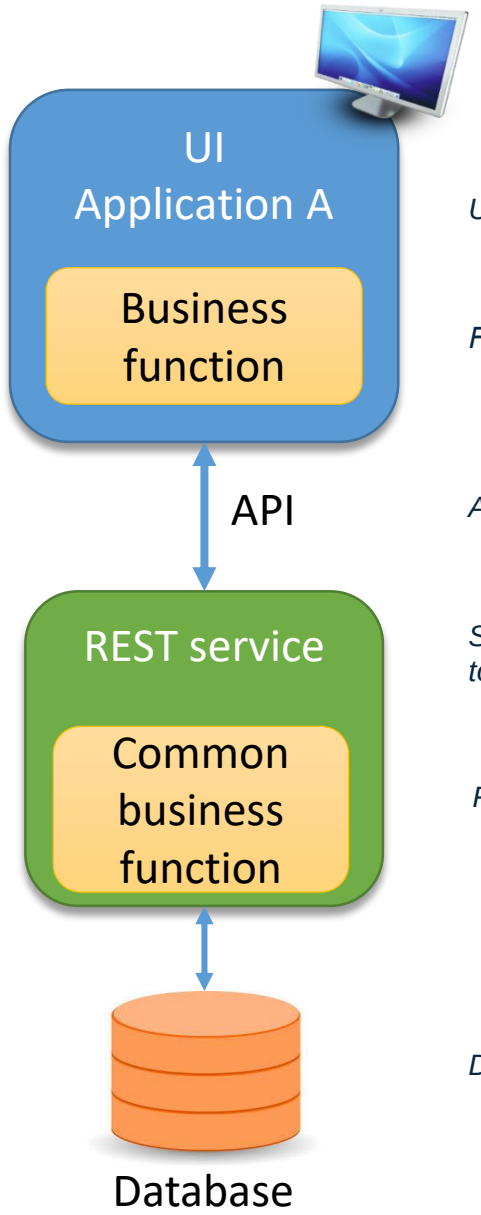
ICIP architecture

Approach

- Build applications as modules
- Build the core REST service for every application
- Incorporate microservices to connect apps
- Each service is self-contained (no dependencies)
- Use the best tools suited for each type of data

The architecture comprises three logic layers:

1. **User interface (UI):** data graphical visualization
2. **Business logic layer:** controls processes
3. **Data layer:** processed data stored in separate databases



User interface of the application

Functions as search, filter…

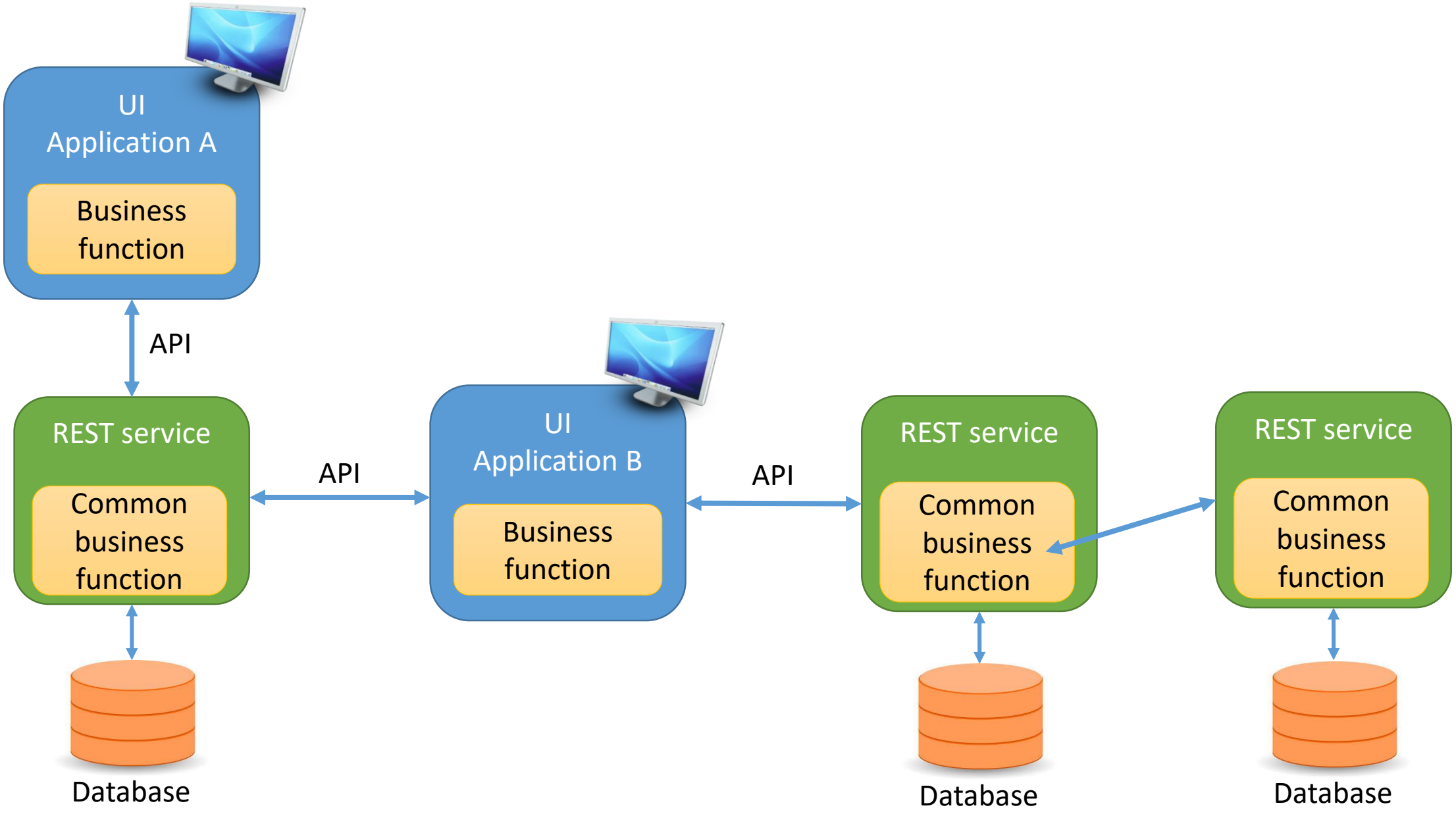
Application programming interface

Style that defines the set of constraints to be used for creating web service

Functions applicable to the domain

Data repository (databases, files, pictures)

Microservices – the principle



Microservices – example and demo

UI graphs
UI controller

UI business function

UI

API

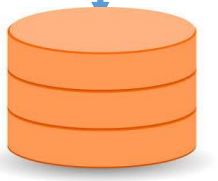
On client computer

On Server

REST web service

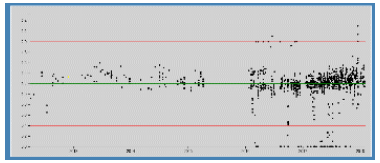
Common business function

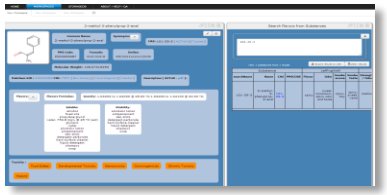
Functional service domain



Database

Persistent data layer
(database, file system)

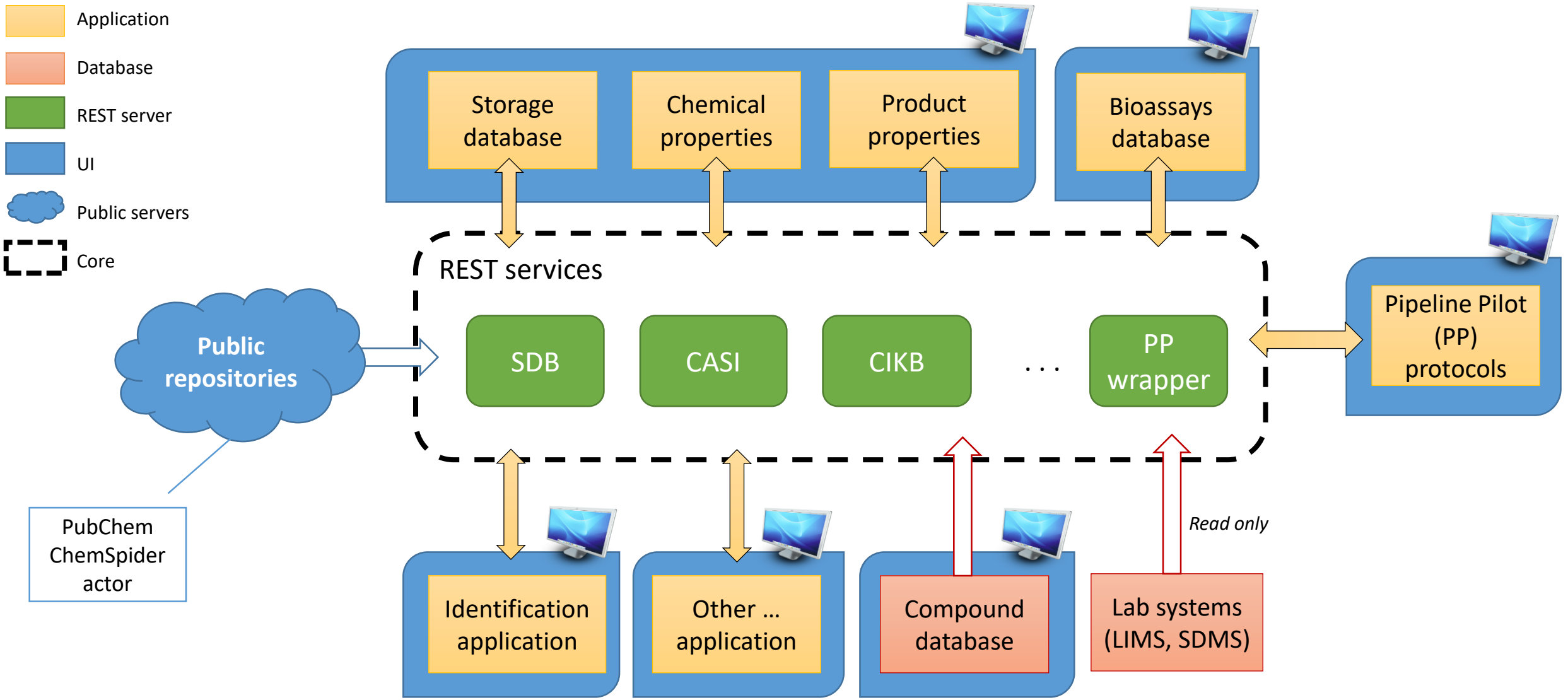




ANGULARJS
by Google

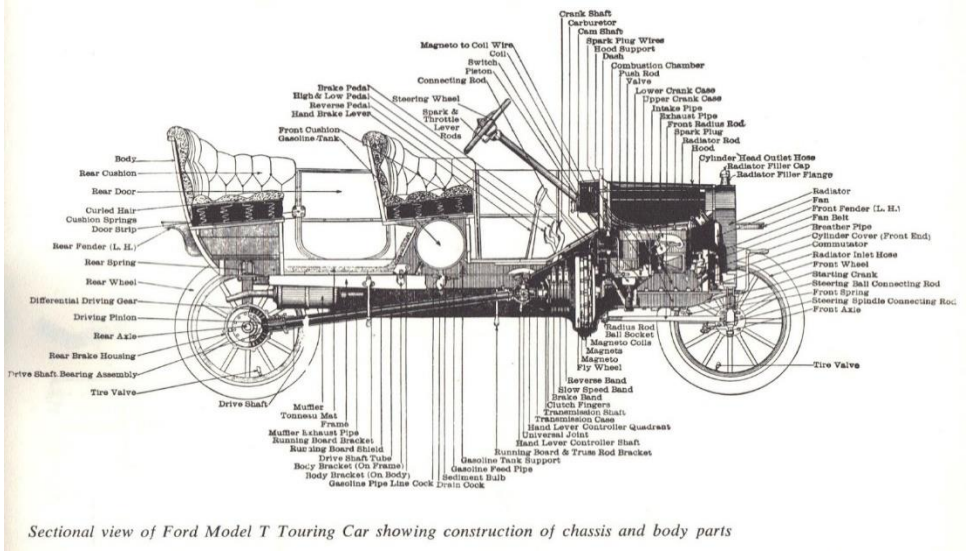
neo4j

ICIP architecture

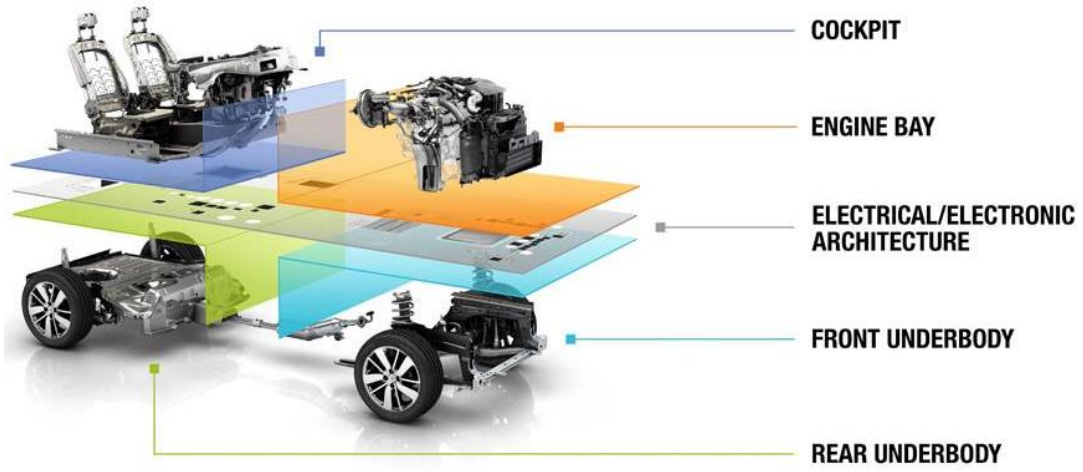


ICIP analogy

- Rather then building a monolithic system made from specifically tailored body parts, we assemble systems by modules.



Sectional view of Ford Model T Touring Car showing construction of chassis and body parts

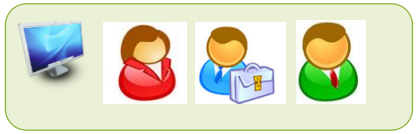


Nissan Renault modular concept

We go from **platform strategy**
- synergies only within the platform

to **modular strategy**
- synergies between more than one platform

Why scientists and developers love it



Scientists

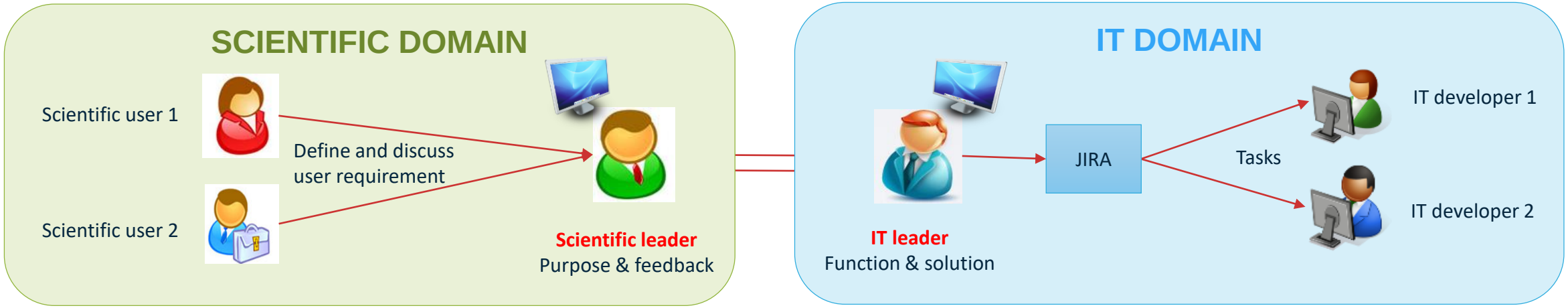
- Single web interface
- No need to install yet another software
- Fast delivery
- Services “under construction” will not obstruct the working services
- Can express new needs
- Data-mining capabilities that promote knowledge discovery

Developers

- Robust architecture as each service is self-contained
- Reduction of complexity
- Flexibility to modify features in response to new needs
- Rapid development of new features
- Easy deployment - only requires installation of new services and not the full application
- Easy to redistribute the services to several servers
- Dependency upon other services exist only in the case of security (login access features)

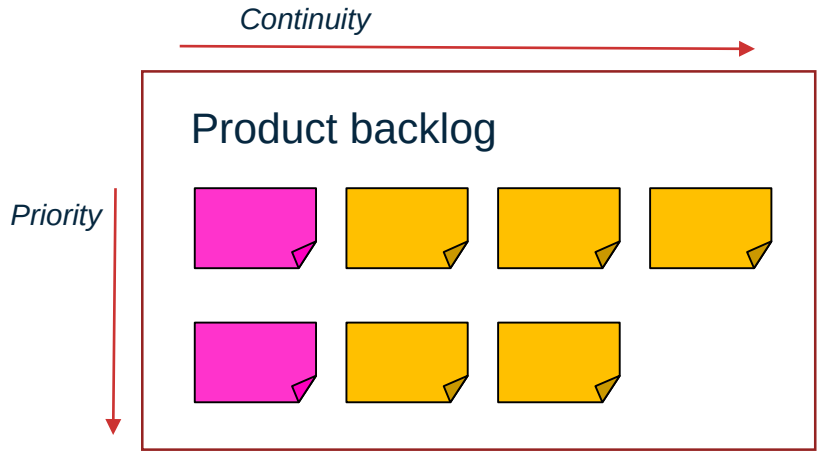
Agile collaboration through leaders' teams

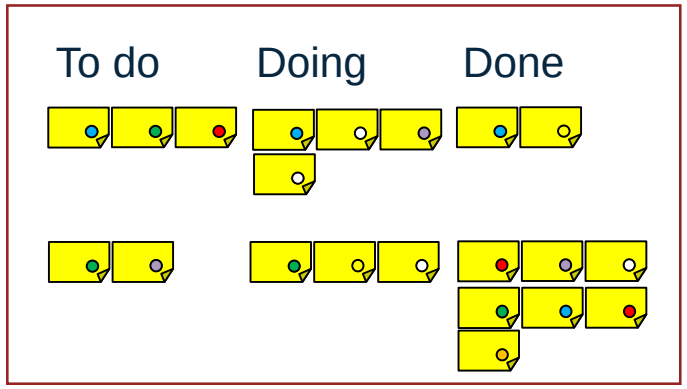
- Systems of ICIP contain different functions required by different departments.
- An effective communication between the scientists and IT experts is a must.
- Building modular systems requires agile collaboration.

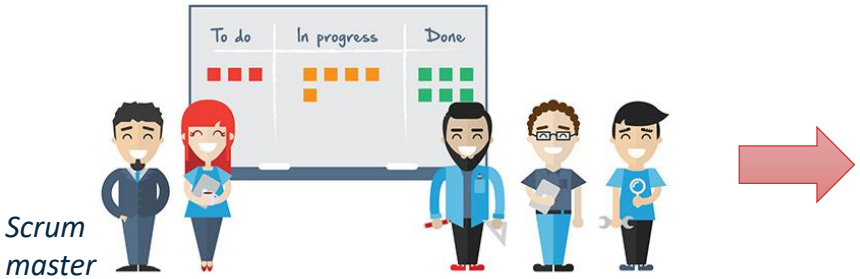


SCRUM

- To deliver the **MOST** valuable work **RAPIDLY** in a free innovative mindset







Lessons learned

- Requires a multidisciplinary developer(s) with different programming skills (server, languages, frameworks, etc.)
- Maintain regular discussions with scientists
- Ongoing learning process, don't plan too much, execute

Conclusion

- At PMI, we analyze complex chemical matrices of product aerosols that generate high quantities of data.
- ChemoInformatics team developed several databases and applications that automate scientists' work.
- Processes accelerate product assessment and assemble the resulting data in one noncentralized but integrated platform: **ICIP**.
- Architecture is modular, with modules interconnected using microservices:
  - Chemocentricity – the substance is the central element.
  - All fit-for-purpose laboratory systems, data formats, and technologies are kept.
  - Public repositories are connected via web services, and data are always updated.
  - Browsers can be customized.
  - Data can be visualized using state-of-the-art 3D visualization.
- **ICIP** becomes a true **knowledge base** that allows storing, exploring ,and understanding. In other words:

Store large sets,
include the latest data,
explore from any angle,
& discover the unexpected!

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Thank you for your attention

Building an Efficient R&D Chemo-Centric Data Repository System that Leads to Knowledge Discovery

Pavel Pospisil, Elyette Martin, Antonio Castellon, and Mark Bentley
Philip Morris International R&D