Leverage the Wealth of Internal and External Information to Drive Collaboration and Project-Centricity into Your Research Informatics Platform for Drug Discovery and Development; A Strategic Imperative

BioIT World Conference 2016

April 5-7, 2016

James Connelly

Head of Global Research Data Management,
Sanofi

SANOFI

Synopsis

Every pharmaceutical company holds a wealth of information collected for numerous drug candidates during research, pre-clinical and clinical development and for marketed drugs.

Much of this information is trapped, undiscoverable and not optimally useable for drug discovery and development or effectively combined with the massive amount of externally available data.

Two pilot studies with IBM Watson successfully extracted high quality information from toxicology reports and enabled researchers to discover critical insight for drug-repurposing proposals.

Cloud-based SAR data platforms will further stimulate the use of broad data sources for research and create a project-centric and collaborative environment.

Increasing reliance of Pharma on a diverse range of drug discovery collaborations has caused severe challenges to adapt Big Pharma data warehouse-centric SAR data platforms.

Evolution and innovation of cloud-based SAR data platforms has revolutionized our approach to SAR data integration, sharing and analysis.

We will describe our current approach to collaboration data and the evolutive roadmap towards a fully capable SAR data platform in the cloud that will utilize cloud-based services and big data technology.

This will reduce the cost of internally supported systems and create a scalable external SAR informatics system.

These are major trends in the pharma industry and are driven by the acceptance, evolution and scalability of cloud-based and computational services and for collaborative SAR data sharing and analysis.

There is a strong value-added in utilization of this wealth of information for R&D and strategic imperative to build informatics systems using Big Data technology to access and utilize all relevant information sources.

Outline

- **IBM Watson Drug Phenotype "Information Cloud"**
  - **Drug Safety and Toxicology "Cloud" Pilot**
    - Concept and Execution
    - Data Access and Visualization
- **Project-Centric, SAR Data Platforms "in the Cloud"**
  - **Traditional approaches to harmonize SAR Data**
    - Compounds, Bioassays and ISIS DB's
    - Integrated SAR Data Platform and DataWarehouse-centricity
    - High Cost of big RED DP integration and support
  - **Back to the future with project-centric DB's**
    - Drug Discovery Collaboration Data Exchange "in the cloud"
    - Evolution of full single-vendor SAR data platforms
  - **Leverage "Big Data" Technologies to build a distributed SAR data environment**

IBM Watson

Drug Phenotype Information "Cloud"

April 5, 2016

Sanofi
James Connelly, Research Director
David Aldous, Senior Vice President
Richard Brennan, Director Scientific Advisor

IBM Watson
Louisa Roberts, Scott Spangler and Ying Chen

SANOFI

IBM **Watson**

SANOFI

Create an advanced, next generation R&D Informatics capability that revolutionizes the use of internal and external data content for an information-driven, iterative approach to Drug Discovery and Development

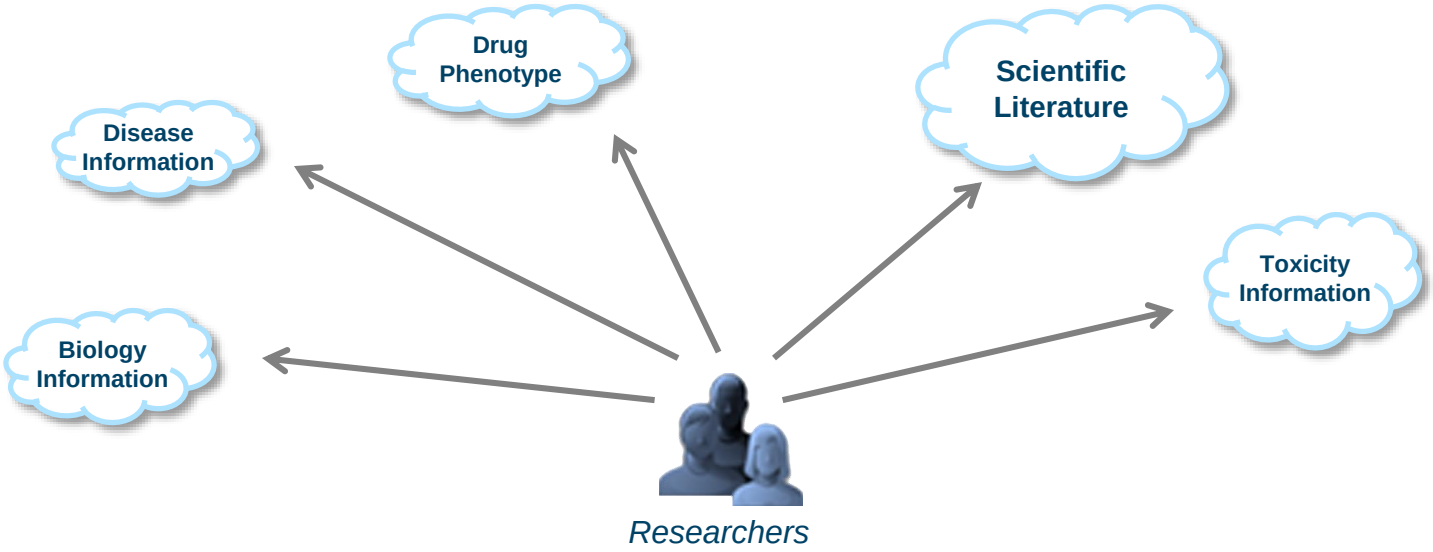
Information-Driven Approach to

1. Learn from legacy projects & information?
2. Detect liabilities for early resolution or fast fail?
3. Uncover and analyze information in many different silos?
4. Make the right/best decisions around projects?
5. Accelerate programs, decrease attrition and lower costs?
6. Find new indications for existing Sanofi compounds?

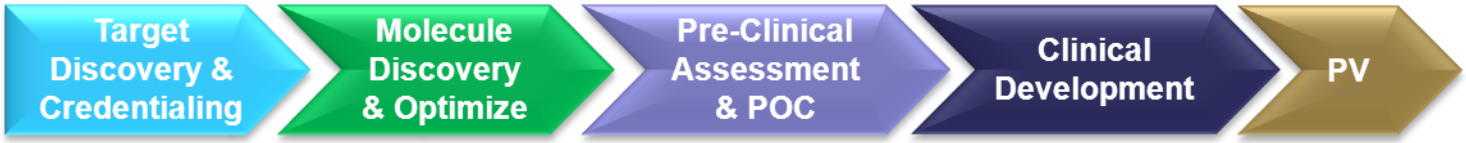
To Create the Anticipated Impact

- ✓ Shape the future Drug R&D paradigm and profiling strategies to advance superior candidates to the clinic
- ✓ Identify new therapeutic targets, anti-targets, biomarkers, MOA, liabilities and correlation with responders/non-responders in patient populations
- ✓ Reduce R&D cost and accelerate the delivery of innovative therapeutic options for patients.

Sanofi has a wealth of information trapped in reports, presentations and siloed databases that is unavailable or undiscoverable for scientific queries, analysis or prospective insight and hypothesis generation/testing



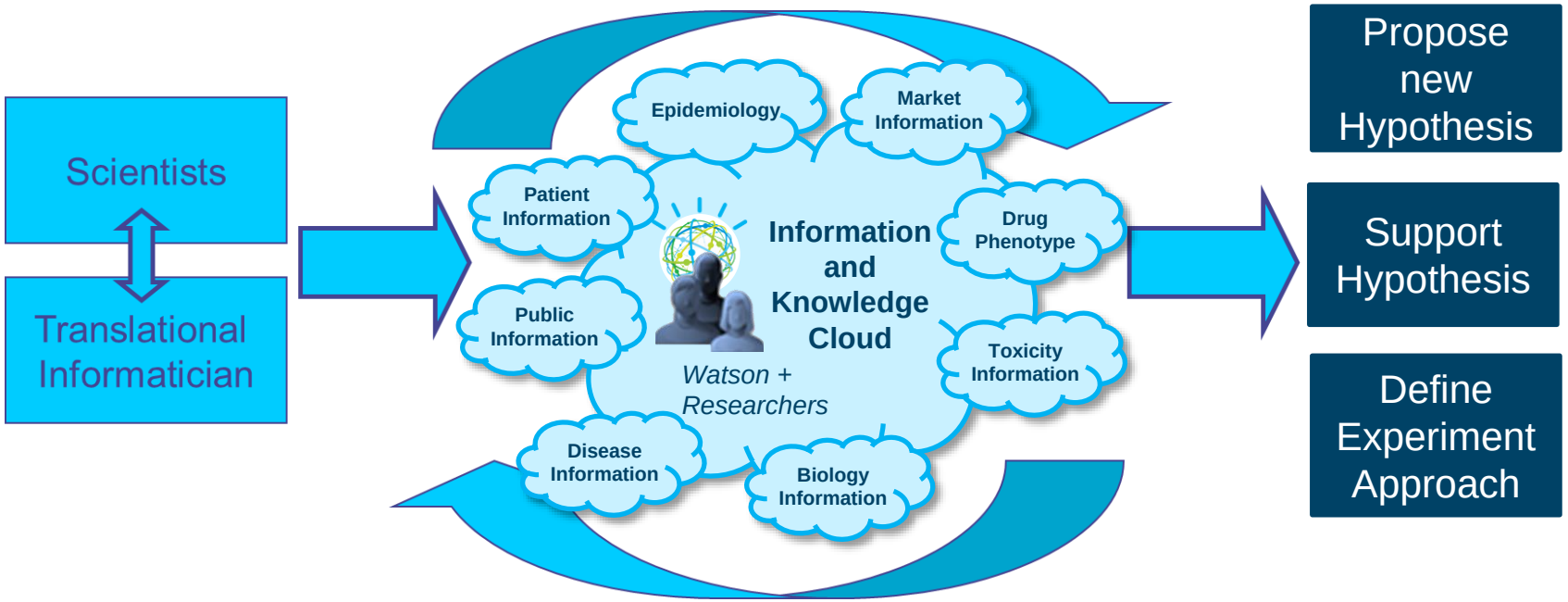
Drug Discovery has historically worked in a linear model (Shots on Goal)



- **Programs end somewhere on this continuum and new projects are initiated**
- **Relatively little of the information collected is re-used for learning and insight**

Next-generation R&D Informatics for an Information-driven Era of Drug Discovery Research

The future is an iterative information-driven model



The key is information availability, interoperability and usability

Background

- Information critical to drug discovery research is typically siloed, fragmented, and poorly accessible
- Sanofi partnered with IBM Watson to demonstrate that text analytics, data integration, and machine-learning could enable the extraction of scientific findings from unstructured data
- A pilot study was undertaken to extract and structure preclinical safety data from historical repeat dose toxicology study reports
- And to integrate and visualize these data alongside discovery pharmacology data to facilitate the resolution of compound → target → toxicity relationships

The Information and Knowledge Cloud will drive value across three key areas

Efficiency

- Speed up the deep exploration of scientific literature, less reliant on experts
- De-risking programs before significant investment to allow early targeted resolution or fast-fail. Information driven = reduce redundant/expensive experimentation

Effectiveness

- Strategic advantage through immediate and effective leveraging of pre-competitive information or new information from collaborations
- Increase quality of decision making at Target, Lead and Development Candidate, Choose the best programs.
- Shape the future profiling and disease area strategies to advance superior candidates to the next stage of clinical trials

Enablement

- Enable every researcher to be as good as the best with access to accumulated knowledge

Potential to decrease attrition, reduce R&D costs, and increase success rate for selected candidates to accelerate the delivery of new therapeutic options to patients

Pilot Exercise Concept Illustration

Information Source	Research Questions	Information Silo's	Cross Silo Aggregation and Analysis	Expected Outcome
Lead Discovery and Optimization	What is the activity of my Lead Compounds on Targets? What are the eADMET results for my Lead?	Target Activity Databases (RED DW)	Are there pre-clinical toxicology effects for compounds in a Lead series?	De-risk Program for early issue resolution or No Go
			What are the target activity profiles for compounds with a common toxicology effect?	Testable Hypothesis for toxicity mechanism or new biomarker
Pre-Clinical Research	What are the animal toxicity effects observed for Candidates? What are the efficacy model results?	Many Pre-Clinical Report Types	Are there any toxicology effects that would stop an NIH/NCATS Proposal?	Support for or No Go on Proposal or a New re-purposing proposa
External Research	What biological effects are observed with Sanofi Compounds? What new therapeutic indications Are proposed?	Public Journals And Patents	Do biological effects observed support an NIH/NCATS drug Re-purposing proposal?	Validation of proposed Therapeutic Indication for NIH/NCATS proposal
			IBM/Sanofi DPIC and Analysis Engine	

Safety & Toxicology Use Case – Pilot Overview

- Focus on a prototype Toxicology Information Cloud that will extract information from a Rat Toxicology report set used in the IMI eTOX initiative and known toxicity test-cases for benchmarking.

Key Questions:

1. Were there any pre-clinical rat toxicology effects observed for compounds similar to my lead series?
2. Can we observe a pattern of biological target activity among a series of compounds that show a common toxicological effect?
3. Are there any toxicity issues that impact decisions on NIH/NVCATS initiative?

Evaluation Criteria 1:

Queries return at least 80% of known effects in the eTOX data in our VITIC database benchmark

Evaluation Criteria 2:

Detect a pattern of target-based activity for compounds with a common toxicity effect and leads to an experimentally testable hypothesis

Evaluation Criteria 3:

Successful detection of known toxicity effects on the NIH/NCATS compounds

Safety & Toxicology Benchmarking

1. Find compounds with specific toxicity outcomes
2. Identify toxic effects caused by similar compounds
3. Link adverse effects to pharmacology
4. Demonstrate the breadth/richness of the data extracted and more exploratory uses

- The above searches could be linked to additional information such as:
  - **Which of these compounds also caused ALT increase?**
  - **What was the NOAEL for these compounds?**
  - **What were their indications?**
  - **What published compounds cause liver toxicity and also have structural similarity to Sanofi compounds withdrawn from development?**

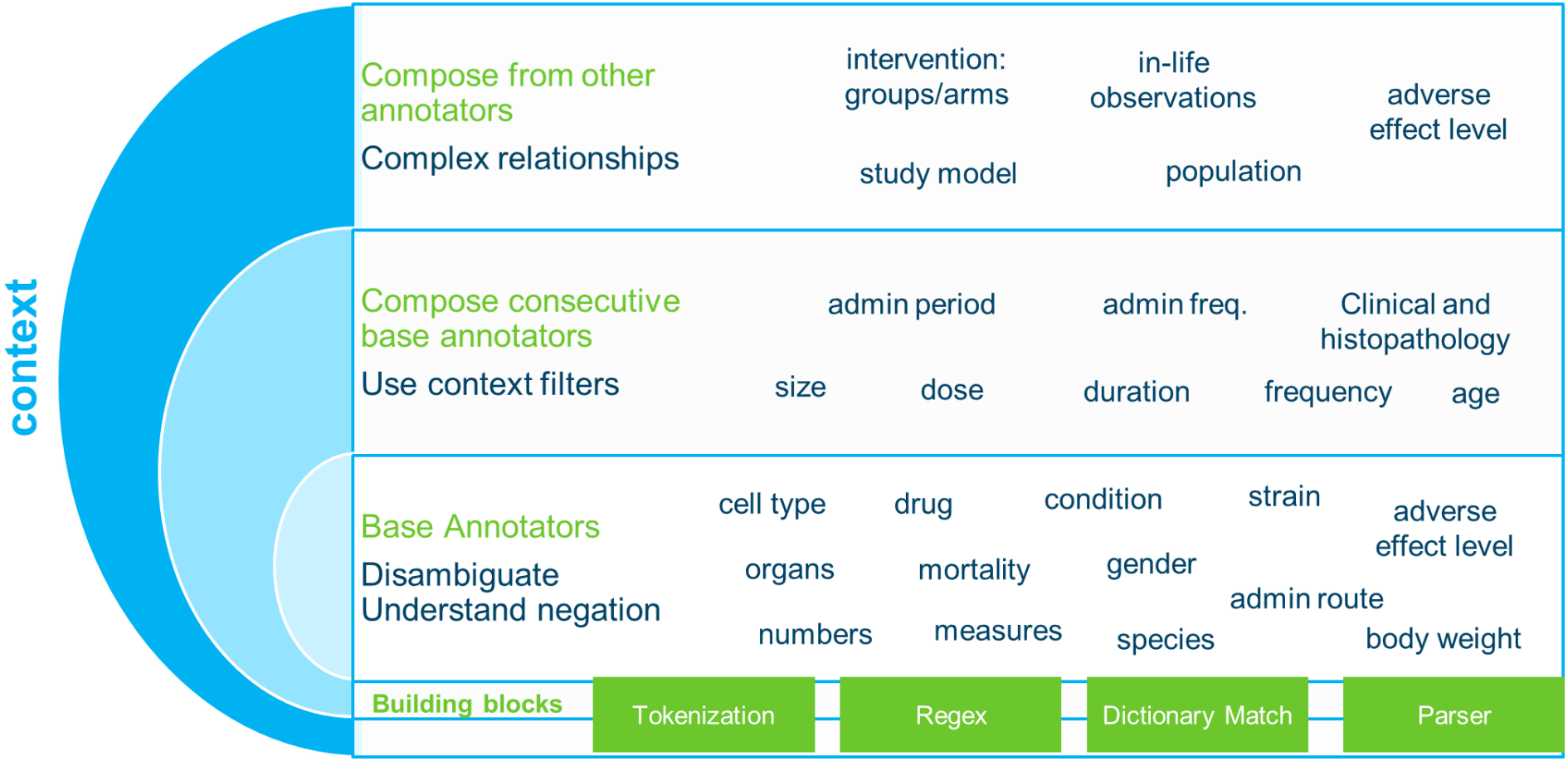
Execution

The configuration of complex text annotators was required to automatically extract highly contextual information, and accurately recognize experimental design concepts such as test-article, dose levels, and administration frequency and period.

Standardized terminologies were adapted or built for multiple concepts such as organ, cell type, histopathological finding, and severity.

Intelligent recognition and reconstruction of experimental design concepts from free text enabled the association of outcomes to specific study groups

Configuration of complex annotators required to extract highly contextual information



In the liver a minimal diffuse hepatocellular hypertrophy was seen in 9 of 10 males and all 10 females at 0.03 mg/kg/day, minimal to mild diffuse hepatocellular hypertrophy occurred in all 10 males and 10 females at 0.3 mg/kg/day, and mild to moderate diffuse hepatocellular hypertrophy was noted in all 10 males and females at 3 mg/kg/day.

1. Identify structure of the document: e.g. sentence belongs to “Results” section

2. Identify sub-entities and sub-elements. May require use of specifically-developed dictionaries and/or pattern based discovery
   - Detect cell type via a mix of dictionary and discovery
   - Imprecise severity term “mild to moderate” is mapped to “moderate”
   - Detect modifier term “diffuse” for histological outcome “hypertrophy”
   - Organ type may be recognized or inferred: “hepatocellular” maps to cell type (hepatocyte) AND organ (liver)

3. Map to canonical terms for compatibility across studies/pathologists/authors/CROs etc.

4. Relate entities within sentence to form facts (”tuples”)

Organ	Cell Type	Severity	Modifier	Outcome	Section
LIVER	HEPATOCYTE	MINIMAL	DIFFUSE	HYPERTROPHY	Results

Challenges that have been overcome

- Build relevant dictionaries for entities, including information for mapping synonyms to canonical version (required iterations with SME)
- Configure text extractors to understand:
  - Contextual disambiguation
  - Negation
  - Variations
- Identify relationships between entities to surface complex facts
- Produce generic text extractors while using only a limited training set

Intervention Group (Canonicals)

StudyIntervention

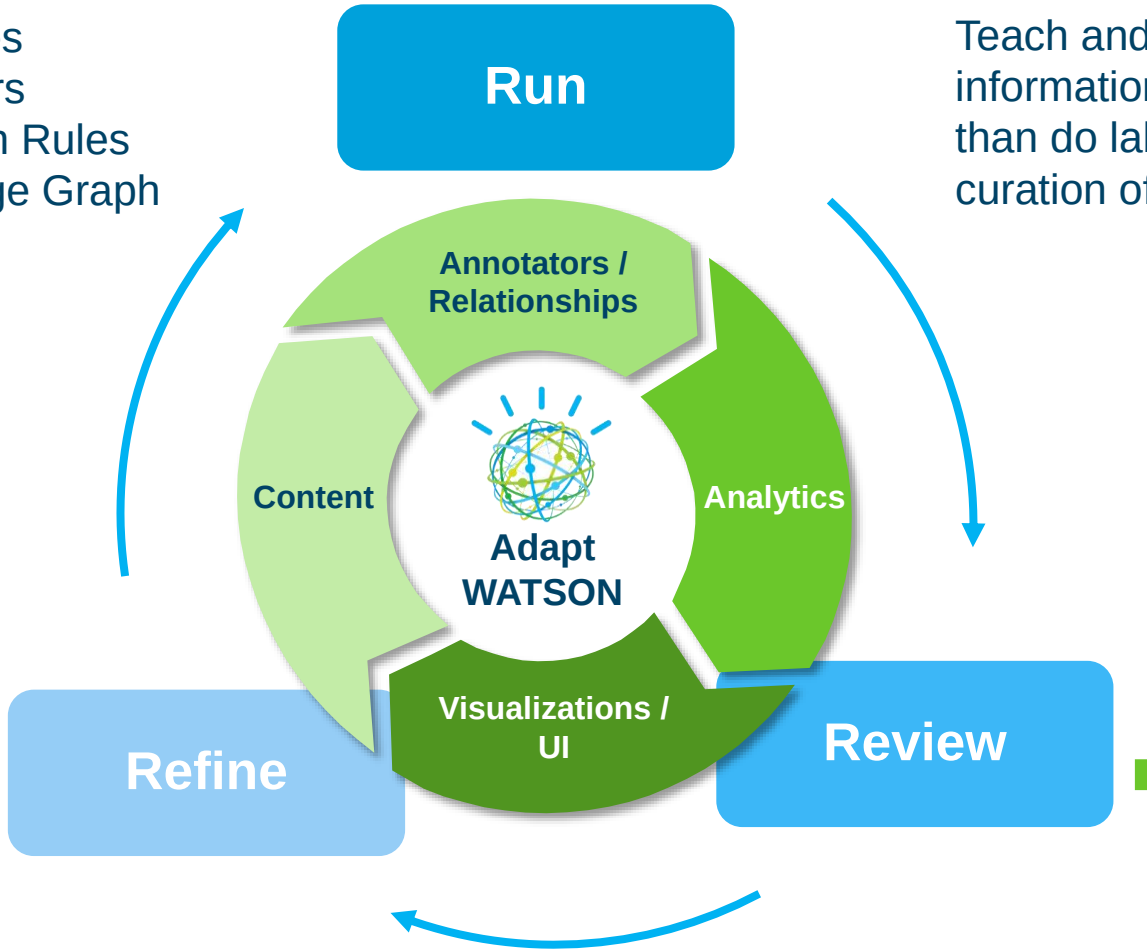
Text analytics result, Number of rows: 53 Showing page 1 of 1

gro...	gender (S...	drug (TEXT)	adminPeriod (...	adminFrequen...	adminRoute (TEXT)	dose (TEXT)	spans (STRING)	headers (STRI...	Input Document
0	BOTH	Cpd1	1 month	1 dose per 1 day	Administration, Oral	0 mg/kg	[68582-68583];...	FIGURES	Sanofi_
1	BOTH	Cpd1	1 month	1 dose per 1 day	Administration, Oral	0.03 mg/kg	[16856-16860];...	SUMMARY;M...	Sanofi_
2	BOTH	Cpd1	1 month	1 dose per 1 day	Administration, Oral	0.3 mg/kg	[16862-16865];...	SUMMARY;M...	Sanofi_
3	BOTH	Cpd1	1 month	1 dose per 1 day	Administration, Oral	10 mg/kg	[16870-16872];...	FIGURES;SUM...	Sanofi_
0	BOTH	Cpd2	1 month	1 dose per 1 day	Administration, Oral	0 ppm	[19153-19154];...	REFERENCES;...	Sanofi_
1	BOTH	Cpd2	1 month	1 dose per 1 day	Administration, Oral	10000 ppm	[11195-11201];...	REFERENCES;S...	Sanofi_
2	BOTH	Cpd2	1 month	1 dose per 1 day	Administration, Oral	50000 ppm	[11206-11212];...	REFERENCES;S...	Sanofi_
1	BOTH	Cpd3	1 month	1 dose per 1 day	Administration, Oral	1 mg/kg	[30025-30026];...	FIGURES;MAT...	Sanofi_
2	BOTH	Cpd3	1 month	1 dose per 1 day	Administration, Oral	15 mg/kg	[30086-30088];...	FIGURES;MAT...	Sanofi_
3	BOTH	Cpd3	1 month	1 dose per 1 day	Administration, Oral	35 mg/kg	[16230-16232];...	FIGURES;SUM...	Sanofi_
0	BOTH	Cpd3	1 month	1 dose per 1 day	Administration, Oral	0 mg/kg	[56815-56816];...	FIGURES	Sanofi_
1	BOTH	Cpd3	1 month	1 dose per 1 day	Administration, Oral	0.2 mg/kg	[16323-16326];...	FIGURES;SUM...	Sanofi_
2	BOTH	Cpd3	1 month	1 dose per 1 day	Administration, Oral	2 mg/kg	[16328-16329];...	FIGURES;SUM...	Sanofi_
3	BOTH	Cpd3	1 month	1 dose per 1 day	Administration, Oral	5 mg/kg	[16335-16336];...	FIGURES;SUM...	Sanofi_
0	BOTH	Cpd4	2 weeks	1 dose per 1 day	Administration, Oral	0 mg/kg	[41966-41969];...	EXPERIMENTA...	Sanofi_
1	BOTH	Cpd4	2 weeks	1 dose per 1 day	Administration, Oral	10 mg/kg	[25205-25207]	INTRODUCTI...	Sanofi_
2	BOTH	Cpd4	2 weeks	1 dose per 1 day	Administration, Oral	150 mg/kg	[16538-16541];...	LIST OF ABBR...	Sanofi_
3	BOTH	Cpd4	2 weeks	1 dose per 1 day	Administration, Oral	25 mg/kg	[16563-16565];...	LIST OF ABBR...	Sanofi_
0	BOTH	Cpd4	1 month	1 dose per 1 day	Administration, Oral	0 mg/kg	[29811-29812]	LIST OF ABBR...	Sanofi_
1	BOTH	Cpd4	1 month	1 dose per 1 day	Administration, Oral	10 mg/kg	[29114-29116];...	LIST OF ABBR...	Sanofi_
2	BOTH	Cpd4	1 month	1 dose per 1 day	Administration, Oral	100 mg/kg	[30798-30801];...	LIST OF ABBR...	Sanofi_
3	BOTH	Cpd4	1 month	1 dose per 1 day	Administration, Oral	30 mg/kg	[29121-29123];...	LIST OF ABBR...	Sanofi_

Gender | Drug | Admin Period | Admin Frequency | Admin Route | Dose

Watson Training and Evaluation Process

1. Ontologies
2. Annotators
3. Extraction Rules
4. Knowledge Graph



Teach and improve the information extractor rather than do laborious manual curation of extracted content

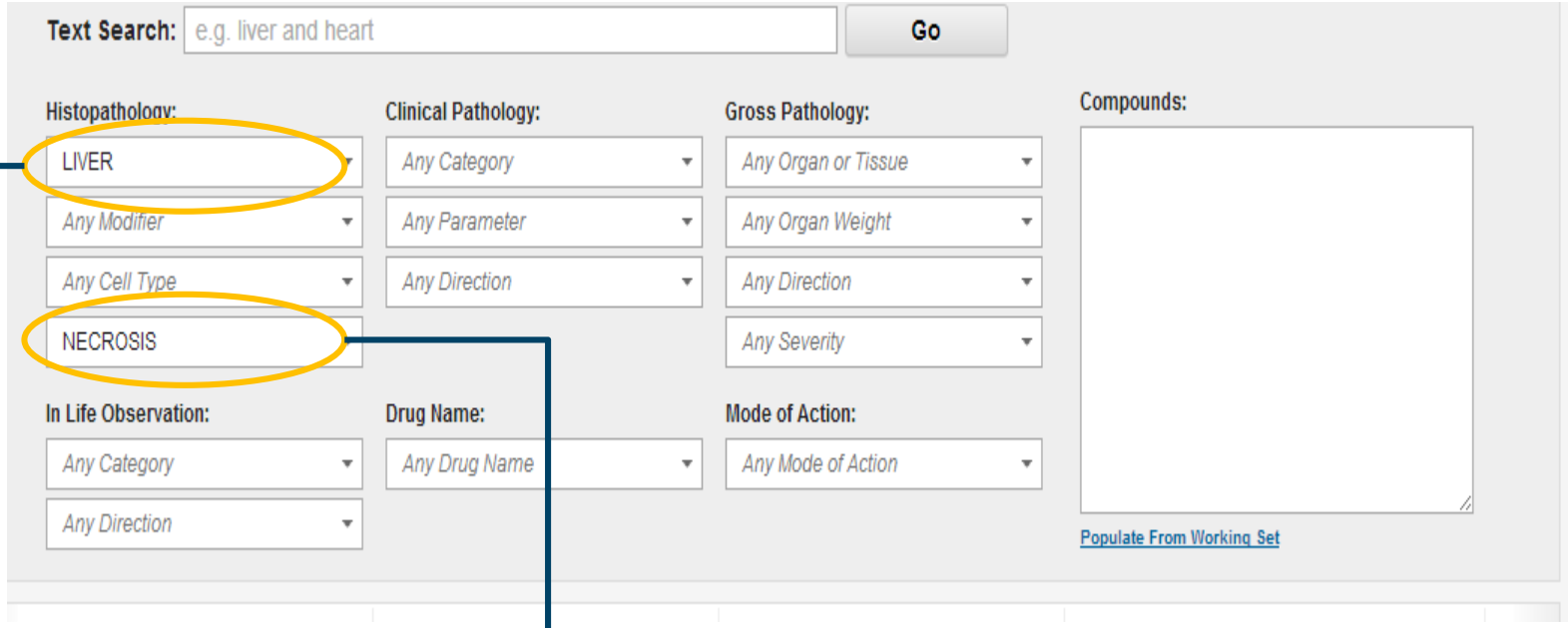
Pilot Goal
Precision > 80%
Recall > 80%

Data Access & Visualization

- A flexible, web-based search and retrieval interface, designed by and for toxicologists, provides access to detailed study report data from multiple query points (compound, outcome, target organ, MOA…).
- Real-time filtering of results on multiple criteria, and complex multi-parameter searches are possible.
- Extracted findings are associated to the text from which they were extracted and access to the full study report is provided if required.
- Incorporation of target activity data from an internal SAR database and matching to study via unique compound ID, coupled with innovative visualization tools facilitates the identification of novel relationships between different data domains, such as target-toxicity relationships, co-occurrence of different pathologies.

Find compounds with specific toxicity outcomes

Example: Liver Necrosis

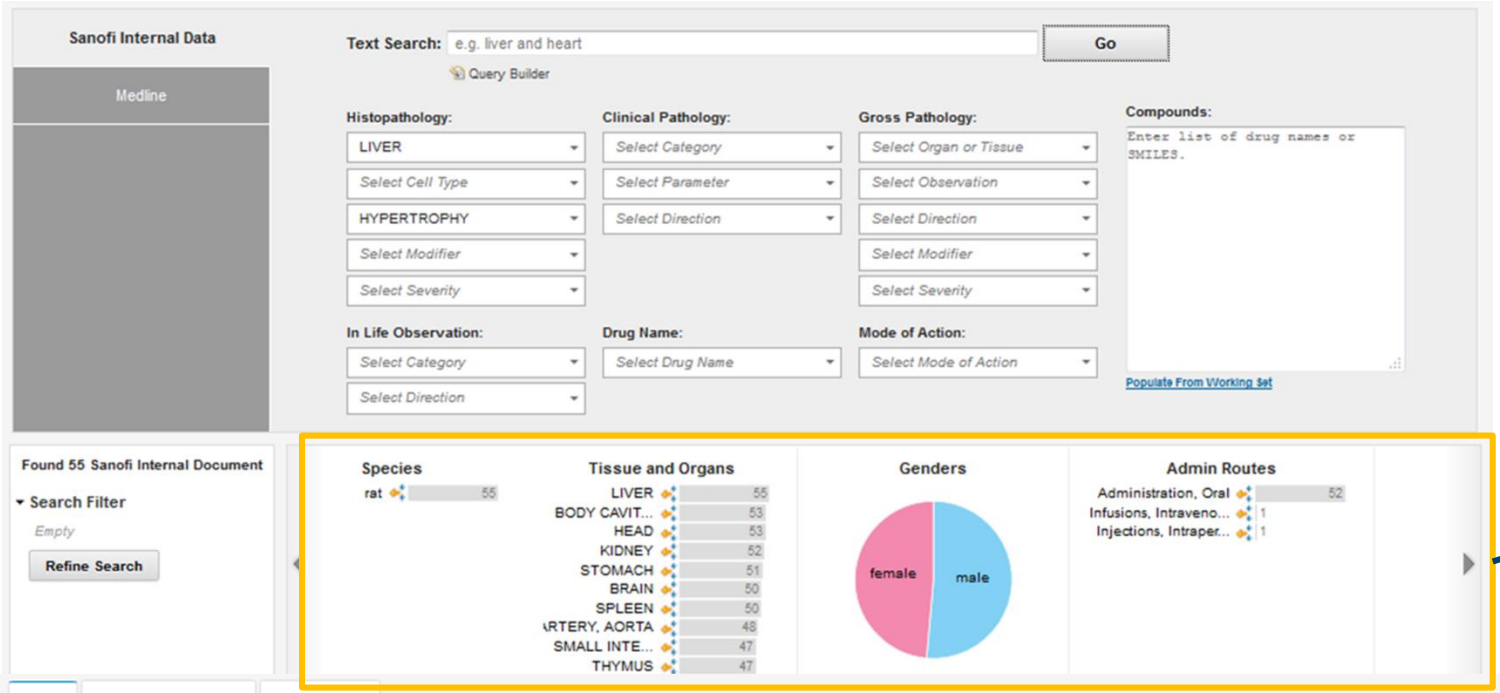


Hepatocellular necrosis was observed in one male (4M37, minimal) and one female (4F36, mild) at 200 mg/kg/day, and minimal to moderate karyorrhexic or apoptotic cells were observed in 4/6 males (2M10, 2M11, 2M12 and 2M13) at 600 mg/kg/day. [52178 - 52418]

IBM Watson

SANOFI

Watson extracts a list of drugs associated with liver necrosis by reading through its corpus of information



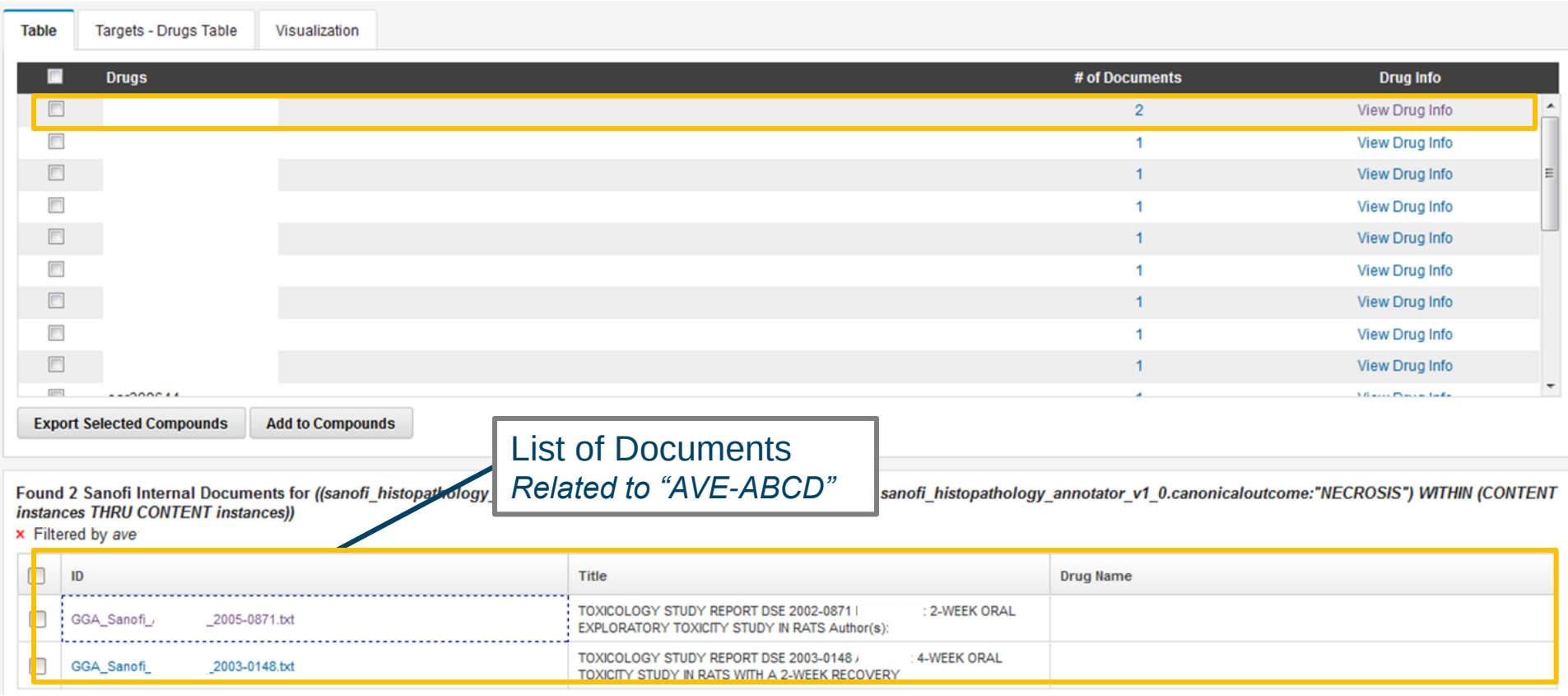
Identified Sanofi Tox Reports related to Liver Hypertrophy

Dashboard
Categorizes documents by species, tissue, gender, etc.

List of Drugs
Associated with Liver Necrosis

Watson enables the scientist to drill down into the source documents to investigate further

The toxicologist can view the related documents and drug information for AVE-ABCD…



Watson understands, extracts and organizes toxicological information from a source document

Histopathology information extracted and organized from a Sanofi Tox Report

TOXICOLOGY STUDY REPORT DSE 2002-0871 : 2-WEEK ORAL EXPLORATORY TOXICITY STUDY IN RATS Author(s):
GGA_Sanofi_ _2005-0871.txt

Report | Study Intervention | Study Population | In Life Observation | **Histopathology** | Gross Pathology

Organ/Tissue	Modifier	Cell Type	Outcome	Outcome Mention	Sentence
HAIR	N/A	N/A	ULCER	ulceration	Minimal skin hair follicle atrophy or moderate epidermal ulceration were noted in 1/5 and 2/5 female rats treated with 600 and 1200 mg/kg/day, respectively.
HAIR	N/A	N/A	ULCER	ulceration	Minimal skin hair follicle atrophy or moderate epidermal ulceration were noted in 1/5 and 2/5 female rats treated with 600 and 1200 mg/kg/day, respectively.
	/A	HEPATOCYTE	HYPERTROPHY	hypertrophy	Compound—related microscopic findings consisted of centrilobular hepatocellular hypertrophy and centrilobular hepatocellular Vacuolation at all doses with the highest incidence and/or severity at 300 and 600 mg/kg/day.
	A	HEPATOCYTE	ENLARGE	enlarged	Minimal to mild centrilobular hypertrophy was characterized by enlarged hepatocytes with ground glass cytoplasm.
	/A	HEPATOCYTE	ENLARGED	enlarged	Minimal to mild centrilobular hypertrophy was characterized by enlarged hepatocytes with ground glass cytoplasm.
LIVER	CENTRILOBULAR	HEPATOCYTE	VACUOLES	Vacuoles	Centrilobular Vacuolation was characterized by intra—cytoplasmic round Vacuoles in the centrilobular hepatocytes; this finding was minimal to moderate, more pronounced in males than in females and occurred at all doses with the highest incidence at 300 and 600 mg/kg/day.
LIVER	N/A	HEPATOCYTE	NECROSIS	necrotic	Mononuclear cell infiltrates were characterized by small aggregates of lymphocytes and/or macrophages occasionally centered upon single necrotic hepatocytes.
LIVER	N/A	N/A	CONTENT	Contents	3.4.3 Microscopic observations (Table 14 and 23- see Table of Contents) Liver
LIVER	N/A	N/A	CONTENT(S)	Contents	3.4.3 Microscopic observations (Table 14 and 23- see Table of Contents) Liver
LIVER	N/A	N/A	CONTENTS	Contents	3.4.3 Microscopic observations (Table 14 and 23- see Table of Contents) Liver
LIVER	N/A	N/A	NECROSIS	necrosis	1074) dosed at 1200 mg/kg/day had microscopic focal subcapsular necrosis of the liver.

Histopathology Table
Specific entry related to Liver Necrosis and sentence from the source document that supports it

The scientist can view the full tox report, which highlights toxicological information

The toxicologist can use this information to:

- Obtain more context around liver necrosis within the source document
- Decide what to explore next

Minimal to mild centrilobular hypertrophy was characterized by enlarged hepatocytes with ground glass cytoplasm. This change was present at all doses and was more pronounced in females. This microscopic finding was the histological correlate of the increases in liVer weights.

Centrilobular Vacuolation was characterized by intra—cytoplasmic round Vacuoles in the centrilobular hepatocytes; this finding was minimal to moderate, more pronounced in males than in females and occurred at all doses with the highest incidence at 300 and 600 mg/kg/day. The content of the Vacuoles was consistent with neutral lipids (Oil Red 0 stain positive).

Mononuclear cell infiltrates were characterized by small aggregates of lymphocytes and/or macrophages occasionally centered upon single necrotic hepatocytes. This change increased in incidence and severity in treated groups as compared with controls.

Aventis Pharma — Confidential

Toxicology Study Report

DSE 2002-0871

One female animal (no. 1074) dosed at 1200 mg/kg/day had microscopic focal subcapsular necrosis of the liver. The relationship of this finding to the treatment is considered doubtful based on its focal and isolated nature and the occasional spontaneous occurrence of such change in untreated animals.

Stomach

Treatment—related microscopic changes consisted of submucosal edema and epithelial erosion/ulcer of the forestomach that were noted in all groups with an increase in incidence in animals treated with when compared to controls.

Non— glandular stomach microscopic changes are presented in the table below.

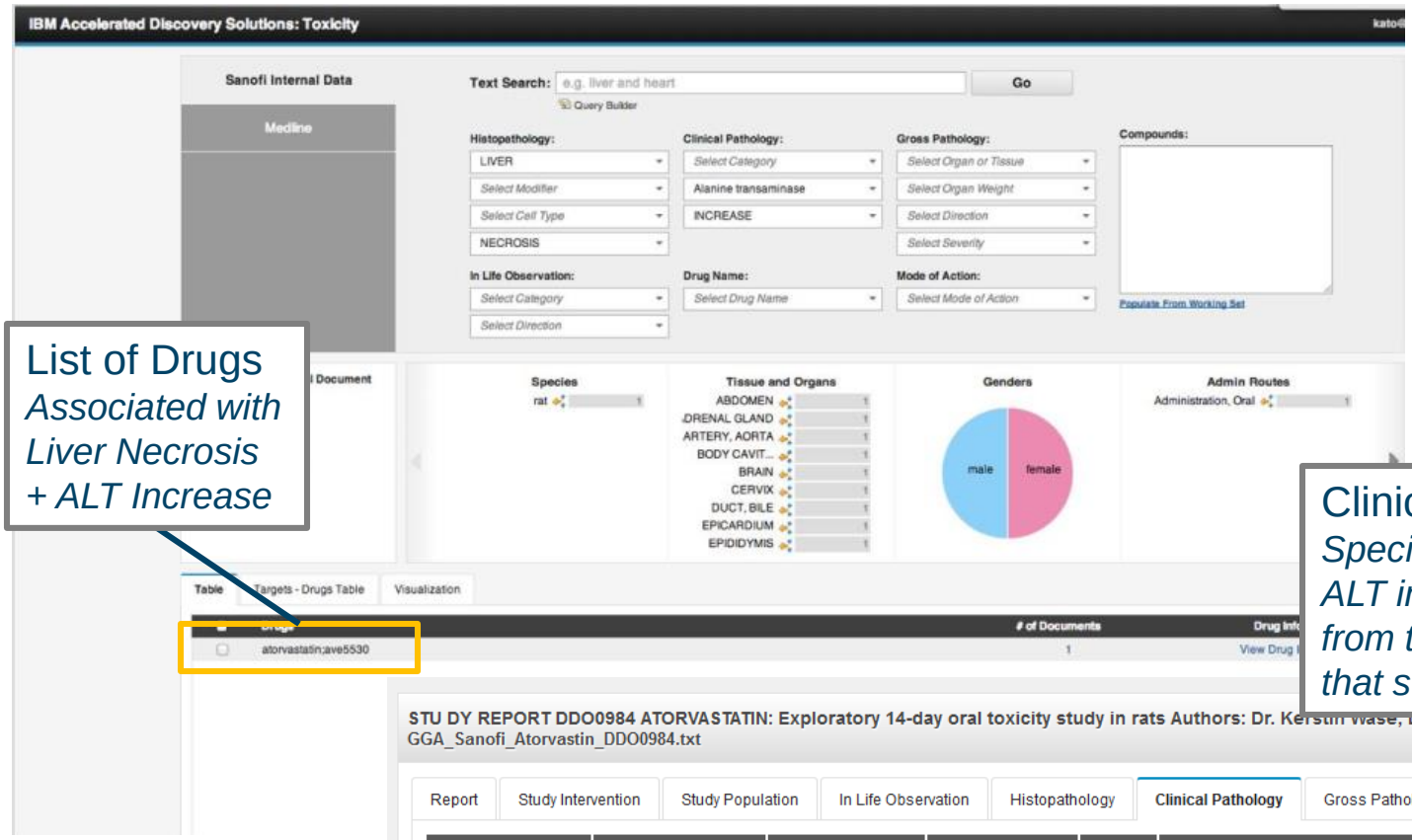
Full Tox Report
Paragraph describing an instance of liver necrosis

IBM Watson

SANOFI

Demonstrate the breadth/richness of the data extracted and more exploratory uses

Example: Liver necrosis in rats + ALT increase



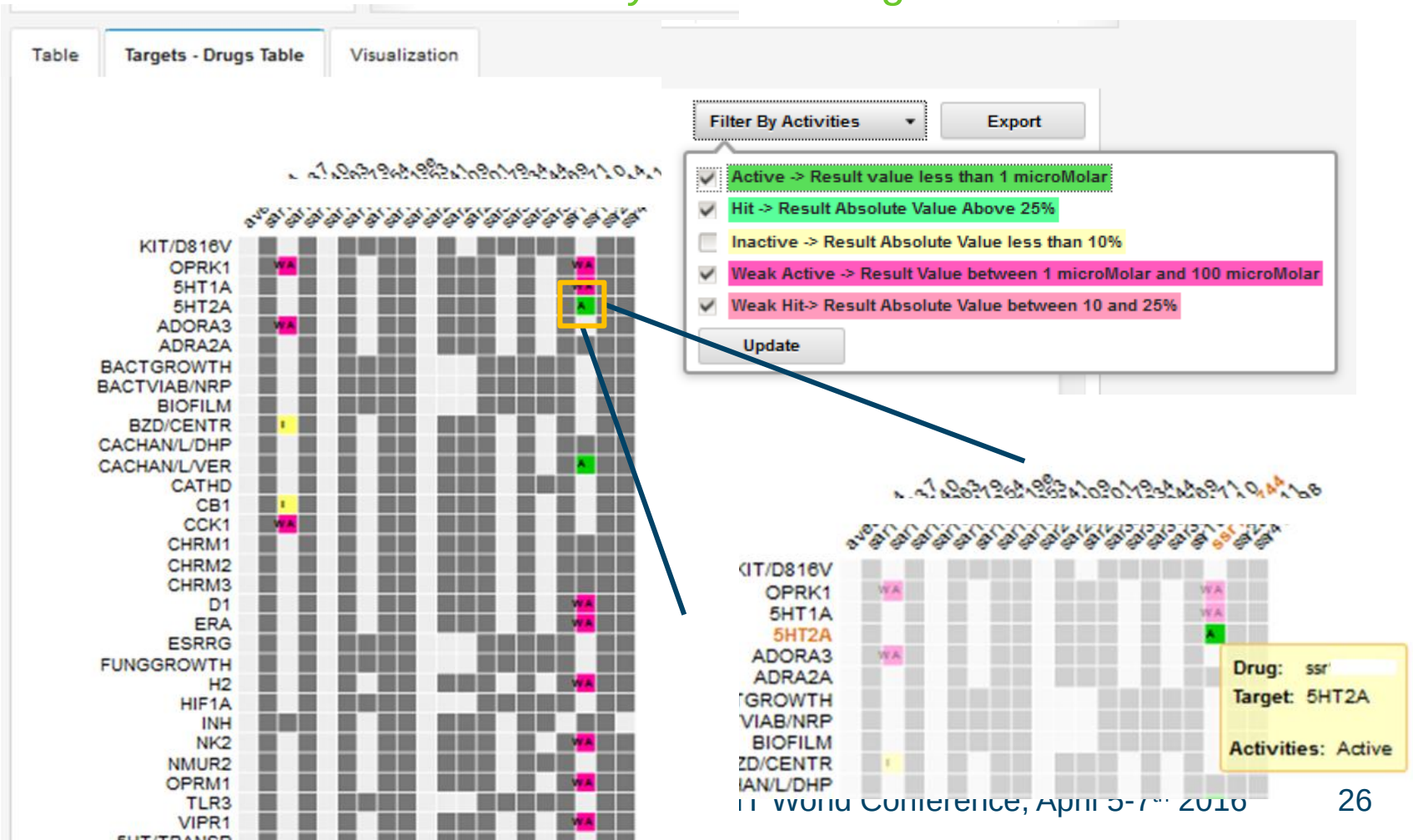
List of Drugs *Associated with Liver Necrosis + ALT Increase*

Found 1 compound (atorvastatin) that includes both attributes

Clinical Pathology Table *Specific entry related to ALT increase and sentence from the source document that supports it*

Watson understands, extracts and organizes target-drug relationships using Sanofi's target activity data

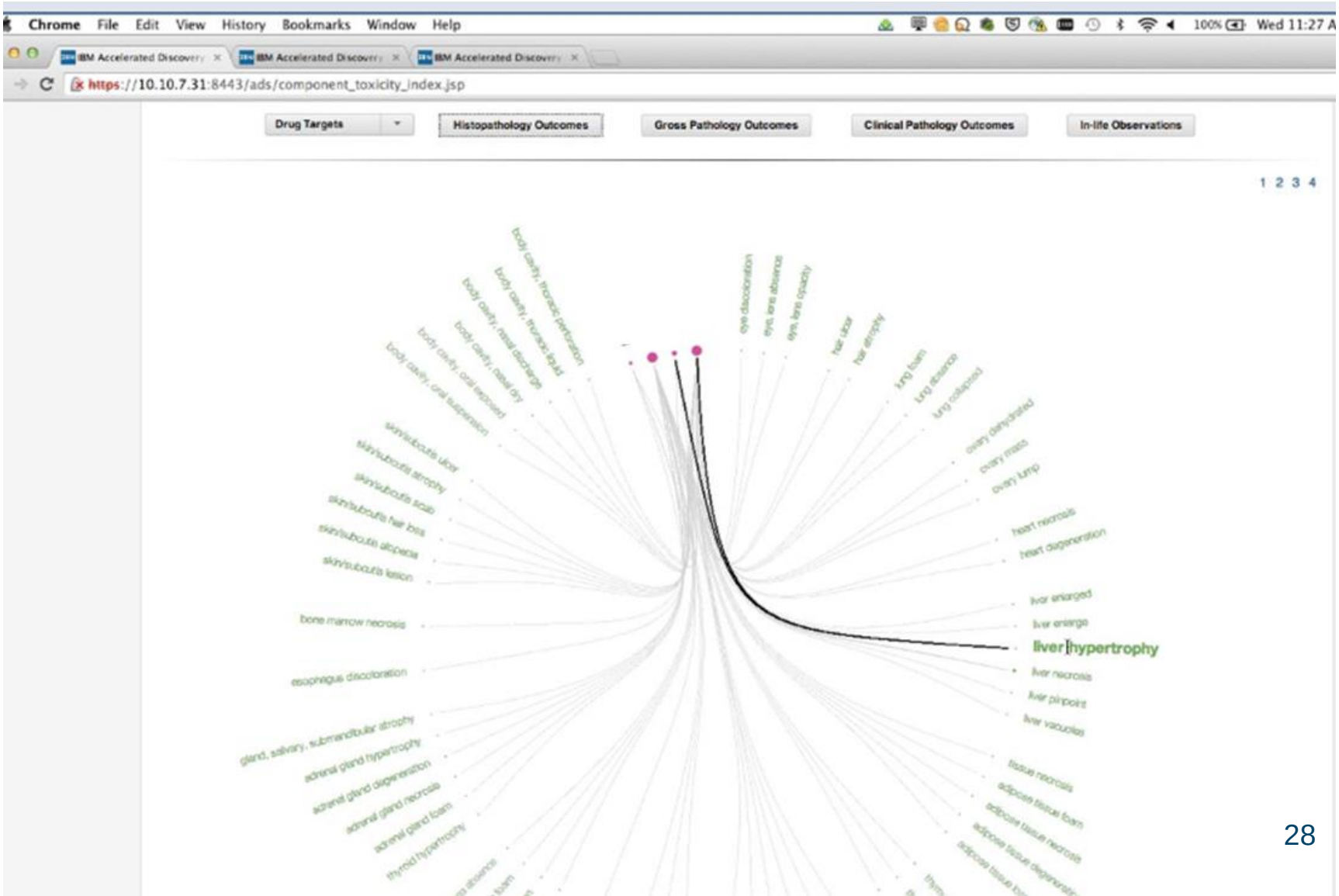
SSR-ABCDEF shows some activity towards target 5HT2A



Using visualizations, the toxicologist can find compounds with specific toxicity outcomes and visa versa

AVEABC is linked to liver necrosis and about 30 other toxicity effects such as liver hypertrophy, skin lesions, etc. (**bold cords**)

Identify toxic effects caused by similar compounds

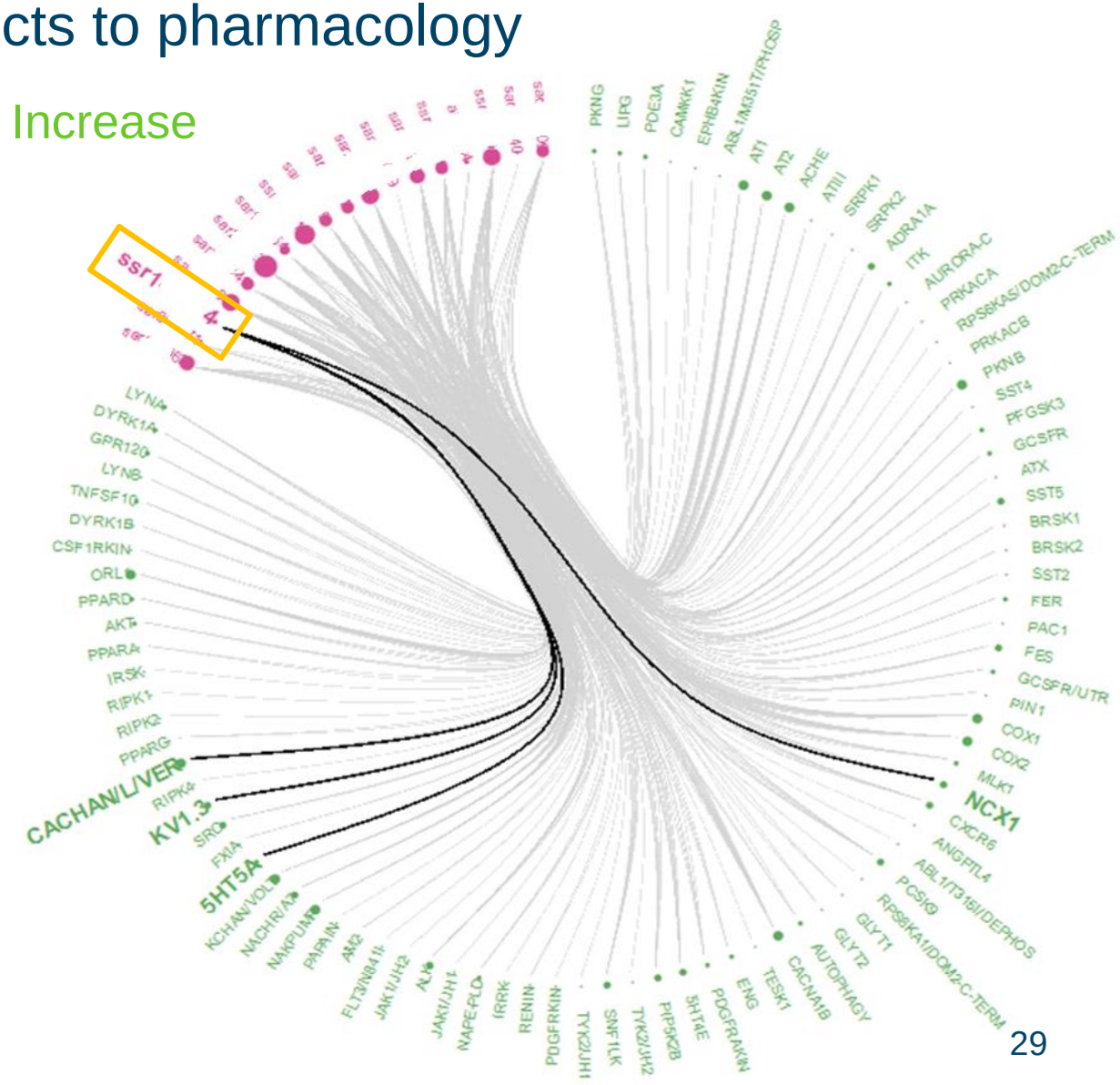


3. Link adverse effects to pharmacology

Example: Liver Weight Increase

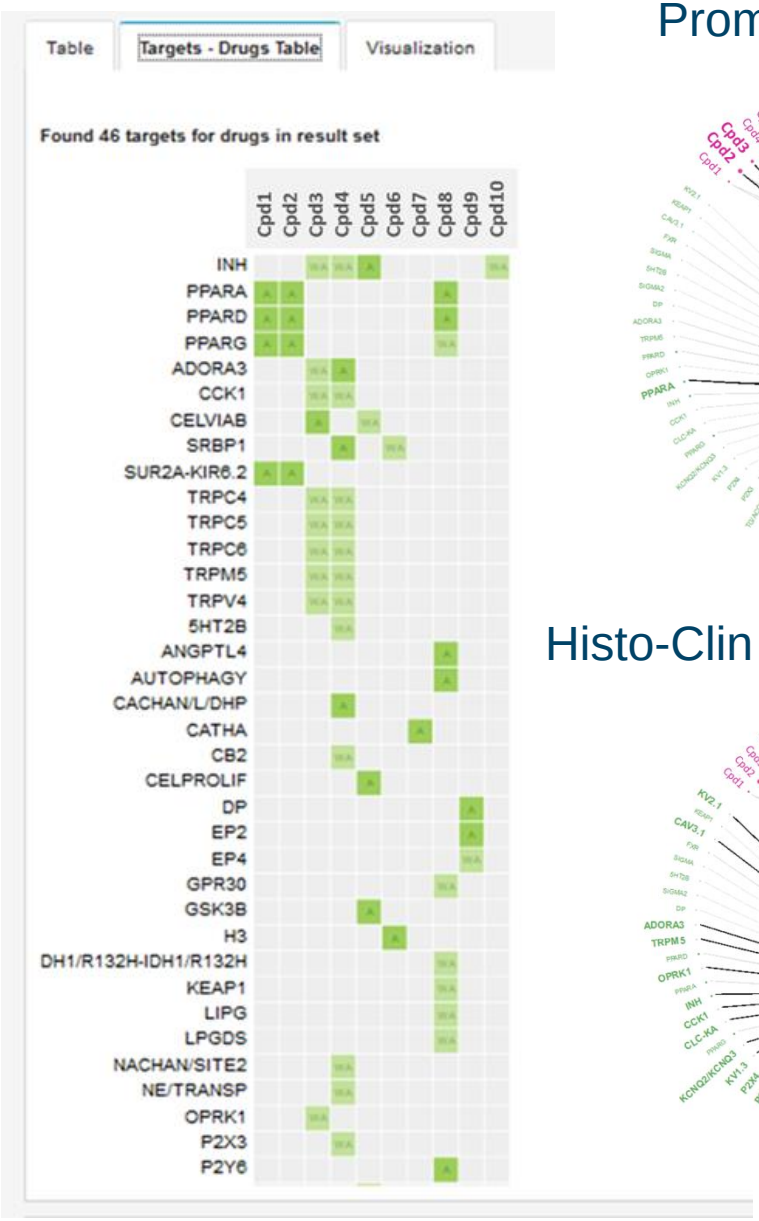
"SSR-ABCDEF" is associated with liver weight increase.

Visually, the toxicologist can see the drug-target relationships

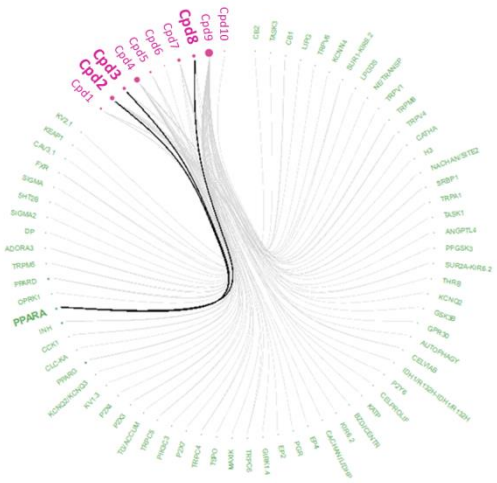


IBM **Watson**

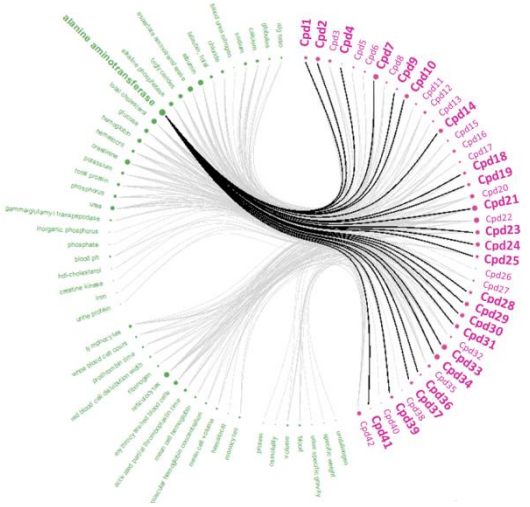
SANOFI



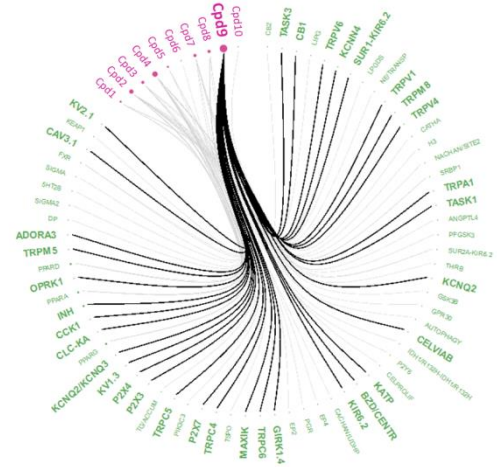
Promiscuous Target



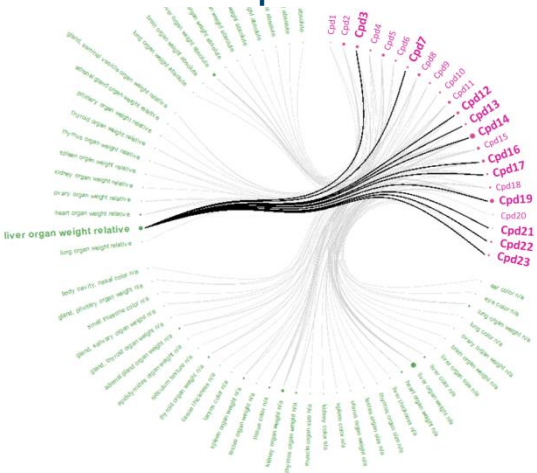
Promiscuous Compound



Histo-Clin path Relationships



Histo-Gross path Relationships



BioIT World Conference, April 5-7th 2016

Safety & Toxicology Results Executive Summary – *Unlocking scientific information from unstructured data*

Watson was able to exceed all pilot objectives and metrics

- ✓ Build Toxicology relevant ontology, annotators and Knowledge Graph
- ✓ Extract High Quality and detailed information from Toxicology reports
- ✓ Find compounds with specific toxicity outcomes
- ✓ Identify toxic effects caused by similar compounds
- ✓ Link toxicology outcomes with pharmacology
- ✓ Extract rich and deep information that can be further explored
- ✓ Precision and recall of information extraction ranged from 80-95%

WATSON has demonstrated considerable capabilities to help identify safety and toxicity signals, and we plan to expand the Watson corpus with additional internal toxicity and safety reports to further drive value

Sanofi Pilot Team

- Business Development Teams

Project Steering Committee

- Provides overall direction
- Provides client specific insights
- Resolve escalated issues

- David Aldous (Sanofi Sponsor)
- Jim Connelly (Sanofi Sponsor)
- Ying Chen (IBM Sponsor)
- Roxana Stanoi (IBM Project Lead)

Sanofi Program Mngmt Team

- Shiv Krishnan
- Andy Keesler
- Mark Nelligan
- Brian Ellerman

- Provide process/program leadership
- Provide industry insights
- Contract management
- Contribute to steering committee meetings

Project Leadership

- Lead the overall effort
- Provide thought and process leadership
- Lead communication with key participants
- Ensure quality

Jim Connelly (Sanofi)
Roxana Stanoi (IBM Project Lead)

IBM Watson Life Sciences

- Jamie Morris – Watson Healthcare & Life Sciences (*)
- Louisa Roberts - Watson Life Science SME (*)

IBM Sanofi Account Team

- Nicolas Forien – Global Client Executive (*)
- Pascal Sempe – Bus. Dev.
- Peter Maluso – LS Executive

(*) Additional Steering Committee members (IBM)

Joint Project Team

Sanofi Contributors

- Brian Ellerman
- Michael Chin
- David Borcherding
- Diana Brassard
- Richard Brennan
- Zaid Jayyosi
- Jacquelyn Dwyer
- Alexander Amberg
- Susan Schiavi
- Sherry Cao
- Heinz Haenel
- Hahn Chang
- Ram Dharanapragada
- Hans-Peter Spirkl

- Provide subject matter expertise, knowledge and historical data
- Participate in Workshops and Interviews
- Perform reviews and validation

IBM Research Team

- Griff Weber – Research
- Stephen Boyer - Research
- Scott Spangler – Research
- Ana Lelescu - Reserach
- Meenakshi Nagarajan – Research
- Anshu Jain – Research
- Susanne Glissmann – Research
- Su Yan – Research
- Nidhi Rajshree – Research
- Nikhil Ingole – Research
- Brad Wade – Research
- Cassidy Kelly – Research
- Eric Louie - Research
- Cartic Ramakrishnan – Lead Software Architect
- Brian Caufield – Software Engineer
- Linda Kato – Software Engineer
- Marco Patricio Crasso – Software Engineer
- Ru Fang – Software Engineer
- Shilpi Ahuja – Architect
- Jacques Labrie – Data Mgmt
- Patrick Mcneillie – Research
- John Colino – Research

IBM Extended Team

- Jon Richter – WDA Product Manager
- Christine Bonini – Watson Analyst
- Jeffrey Kreulen – Watson Engineer
- Isaac Cheng – WDA Engineer
- Rajesh Ramachandran – Watson Solutions
- Mark Marrara – Design Team Lead
- Claire Kohler – Design team
- Ashwini Kamath – Design team
- Anne Aldous – Watson Implementation
- Jaya Kaja – Watson Implementation

- Manage the day-to-day aspects of the project
- Perform workshops and interviews
- Perform business requirements analysis
- Execute required tasks
- Perfrom reviews and validation of architecture

Single Vendor SAR Data Platforms "in the Cloud"

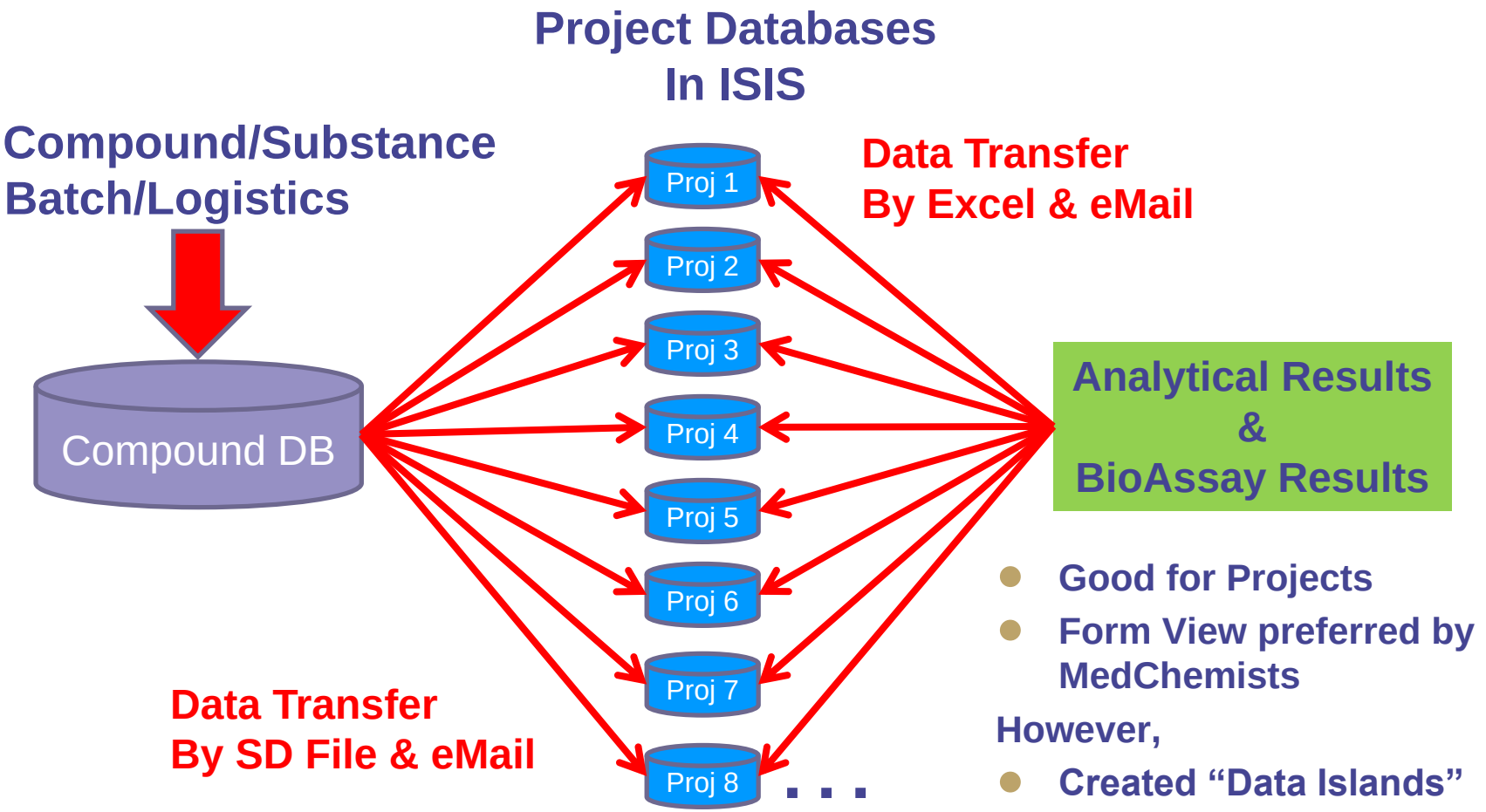
James Connelly
Head of Global Research Data Management,
Sanofi

SANOFI

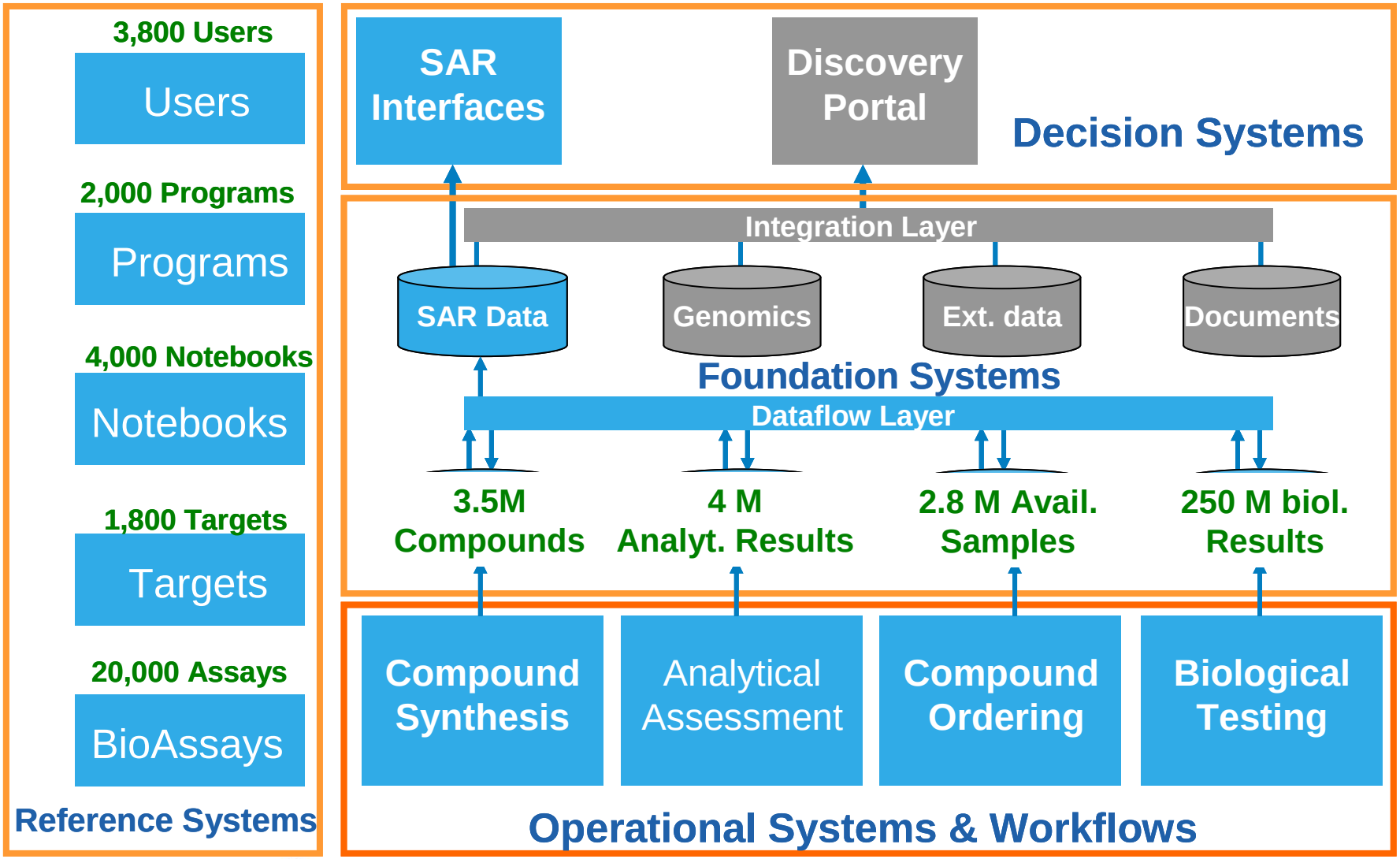
Outline

- Traditional approaches to harmonize SAR Data
  - **Compounds, Bioassays and ISIS DB's**
  - **Integrated SAR Data Platform and DataWarehouse-centricity**
  - **Evolution of RED DP and challenges/gaps**
  - **High Cost of big RED DP integration and support**
- Back to the future with project-centric DB's
  - **Drug Discovery Collaboration "in the cloud"**
  - **Local project-centric DM's**
  - **Evolution of full single-vendor SAR data platforms**
- Leverage new "Big Data" Technologies to build a distributed SAR data environment

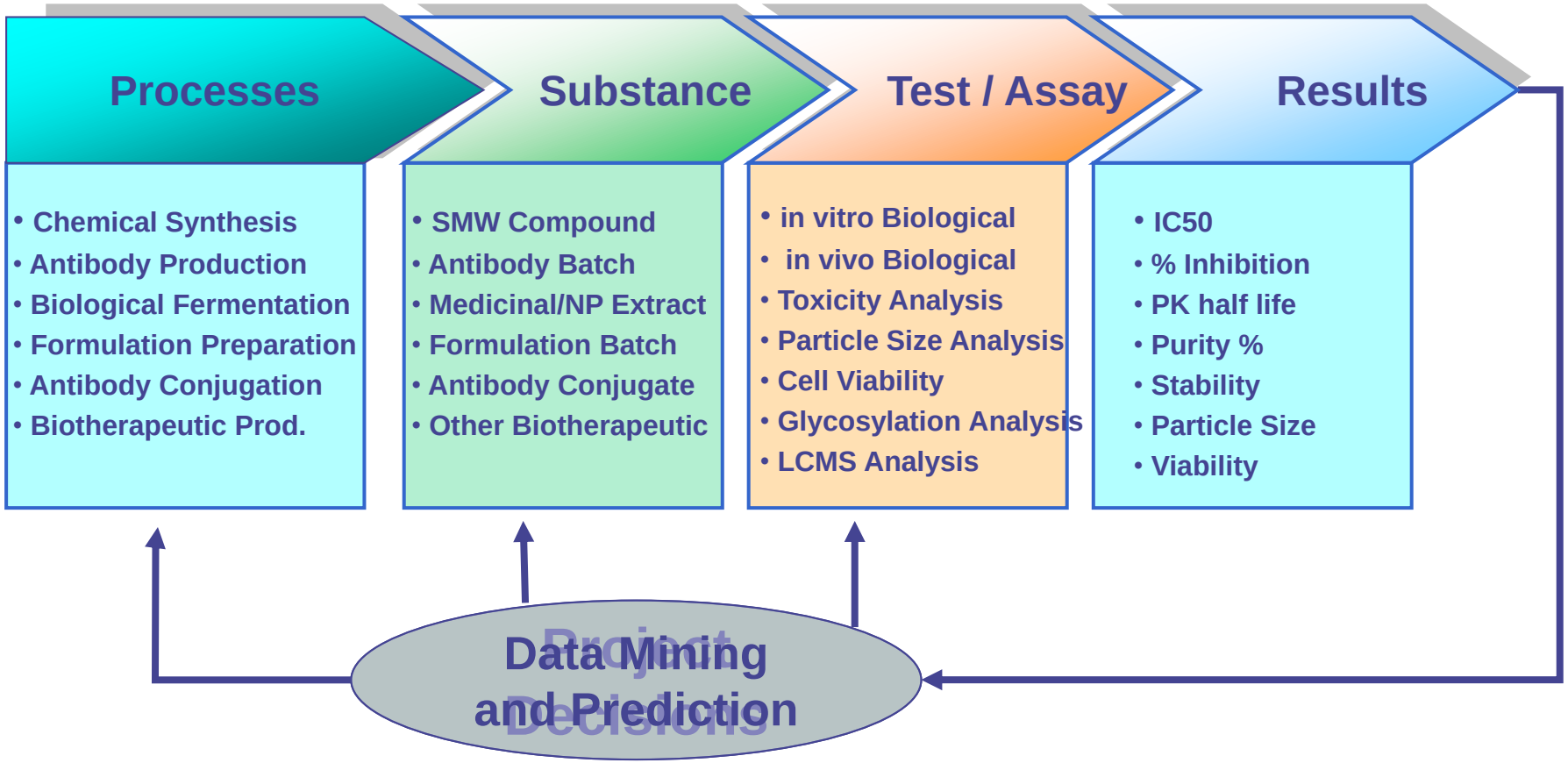
Original “Project Databases”



Discovery SAR Data System



From Structure to Substance Activity Relationship



Need to view results in context of processes, substance metadata and test/method

Evolution of Discovery SAR Data Systems

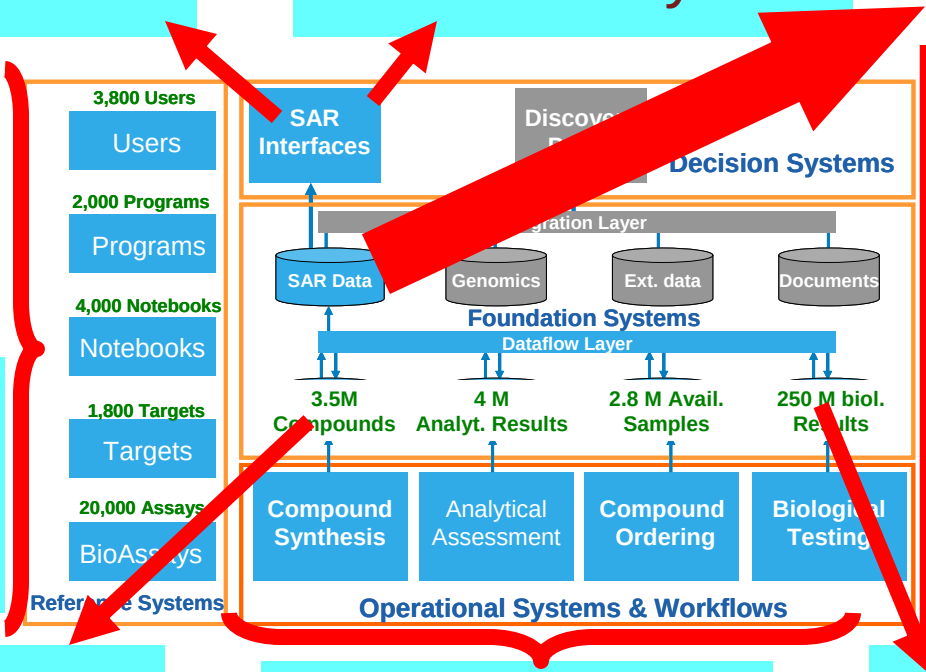
Improve Expert Query Tools

Consolidate Data Access/Analysis

Evolve Data Warehouse and Marts

Update SARUsers PRObase

Update Referentials TAR & BAR



Integrate New Data Across Value Chain

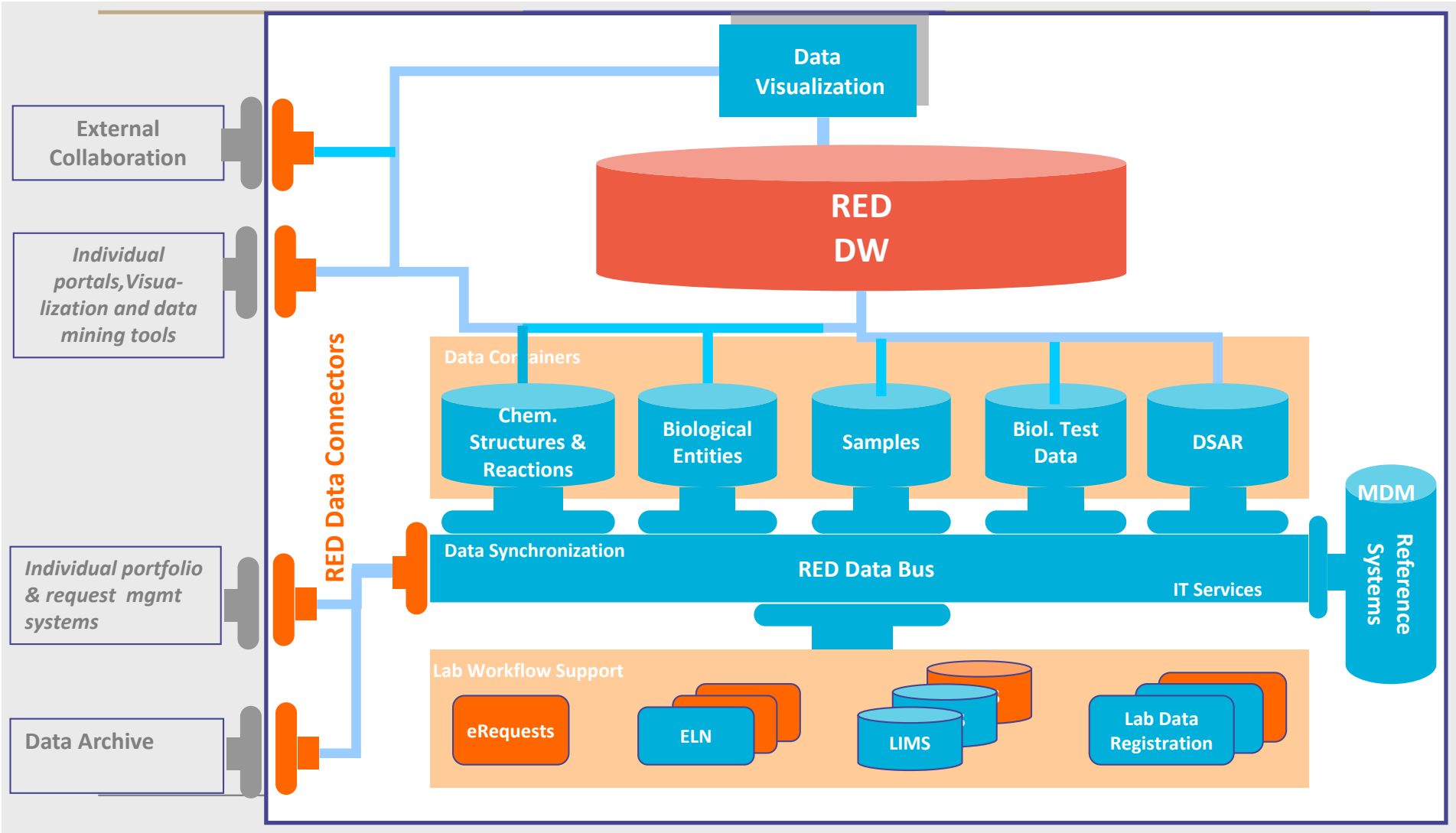
Capture External Collaboration SAR Data

Substance Registration and Repositories

ELN Integration

Enhance Biol. Data Capture

Data Warehouse-Centric Platform



Large, Integrated DataWarehouse Platforms

- Chose “Best in Class” applications and systems
- Enabled data-mining, modeling and drug design
- Provided harmonized, high quality data

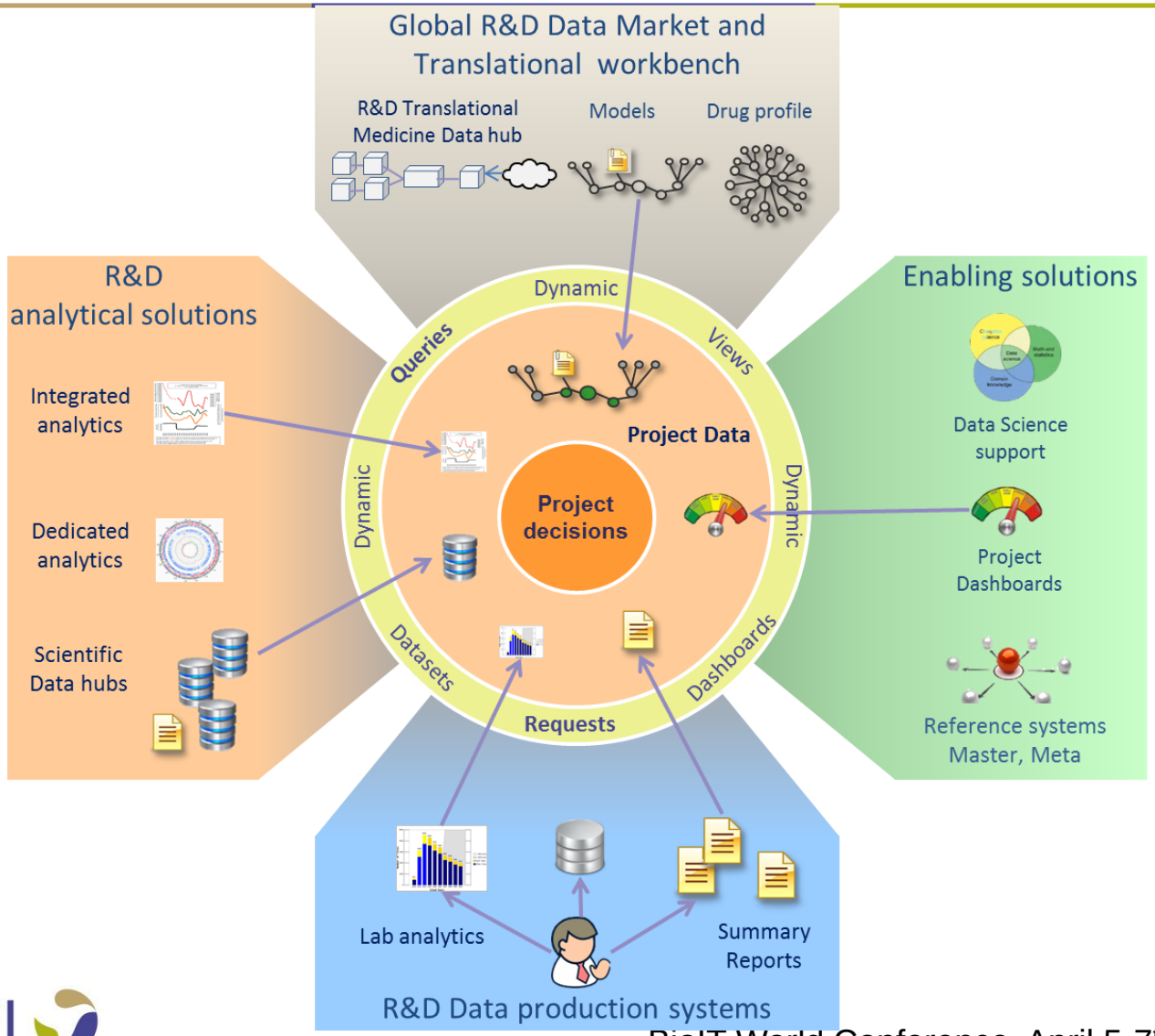
However,

- Difficult to introduce new data types and complex assay results
- Performance and capability gap issues that delayed project progression
  - **Data availability lag, query performance, “one size fits all”**
- Complex data integration required high level of support, high cost
- Platform was rigid, relatively static and did not evolve with availability of new technologies and capabilities.
- Vendors raised prices because we were locked in

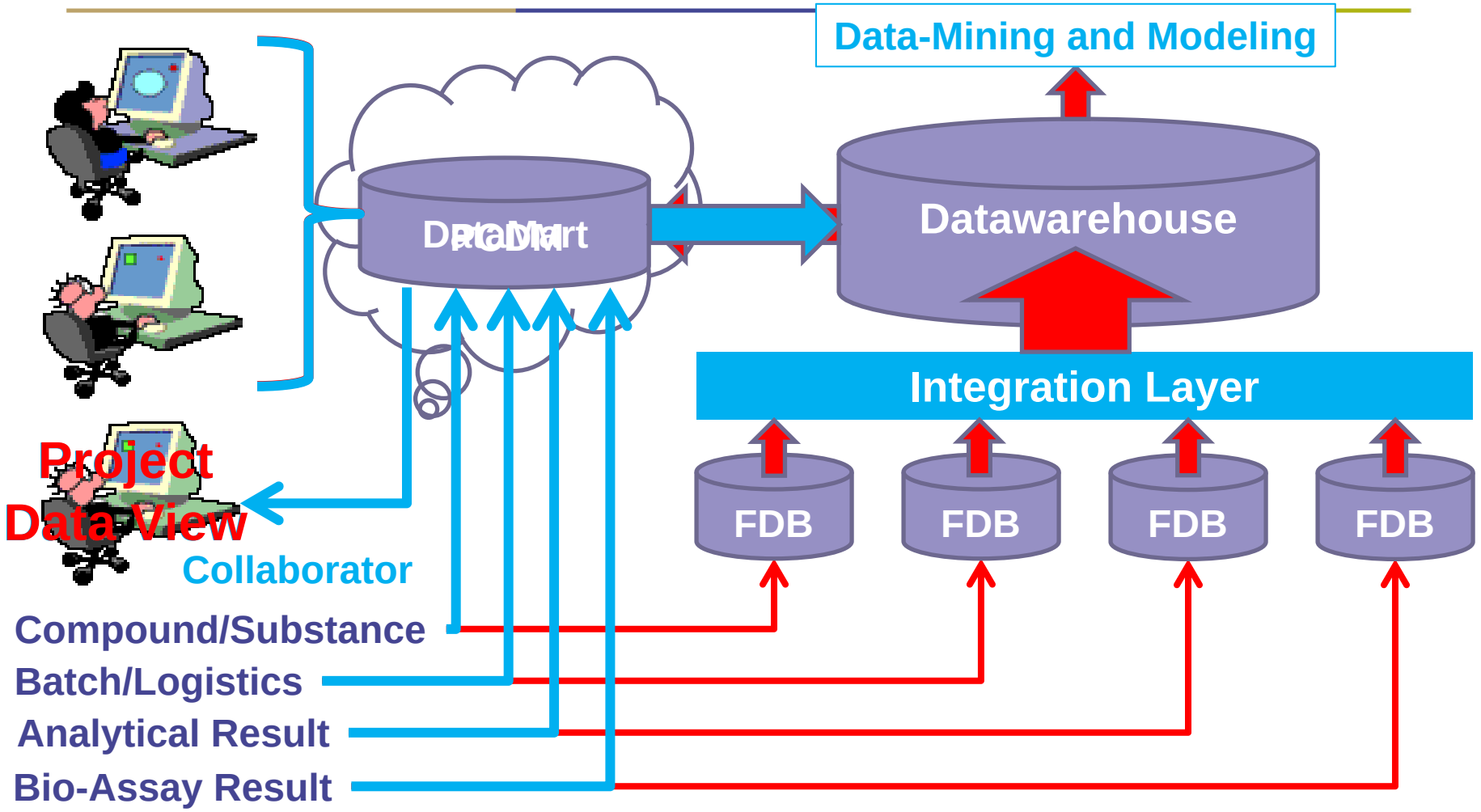
Related to collaborations

- IP tracking related issues and data access control issues
- Difficult to remove data from DW at end of collaboration
- Registration without structures and subsequent data visualization issues

The Vision
A Project-Centric Platform



Reverse the Flow of Data



Prioritize project over data-mining

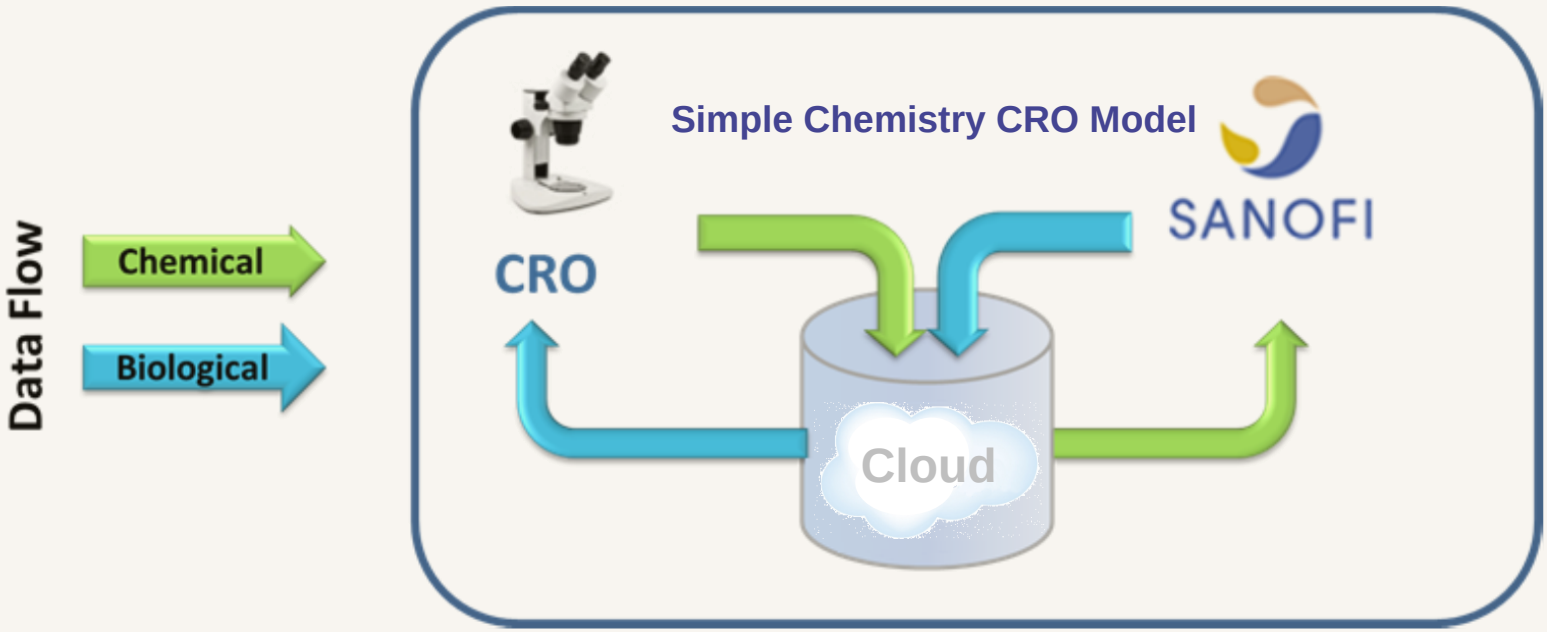
"Set up a **service** to support drug discovery **data exchange** with **external partners**"

Previous Data Exchange Practices and Issues

- Excel Sheet data exchange by email or "SharePoint Dumps"
- Confusion on registration in corporate databases
- Difficulty to integrate and aggregate with internal data
- Limited data visualization and analysis options
- Intellectual property and contractual considerations
- Diversity of data standards and conventions
- Diversity of partner types
  - **(program purchase, non-profit venture, CRO and true collaboration)**
- Large workload on scientists

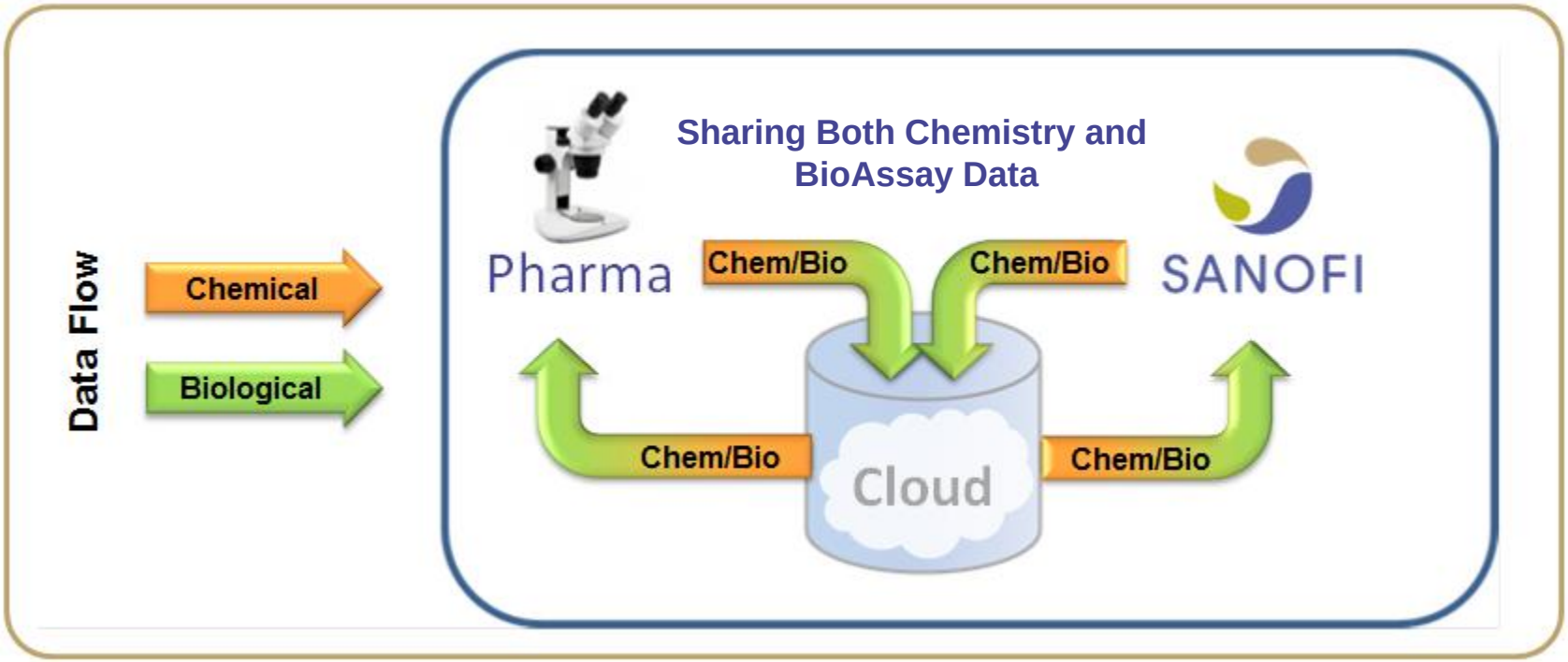
Basic CRO Model

- To share, exchange and register drug-discovery data with external partners (CRO, academics…)
- To collaborate in an external cloud



Partnership Model

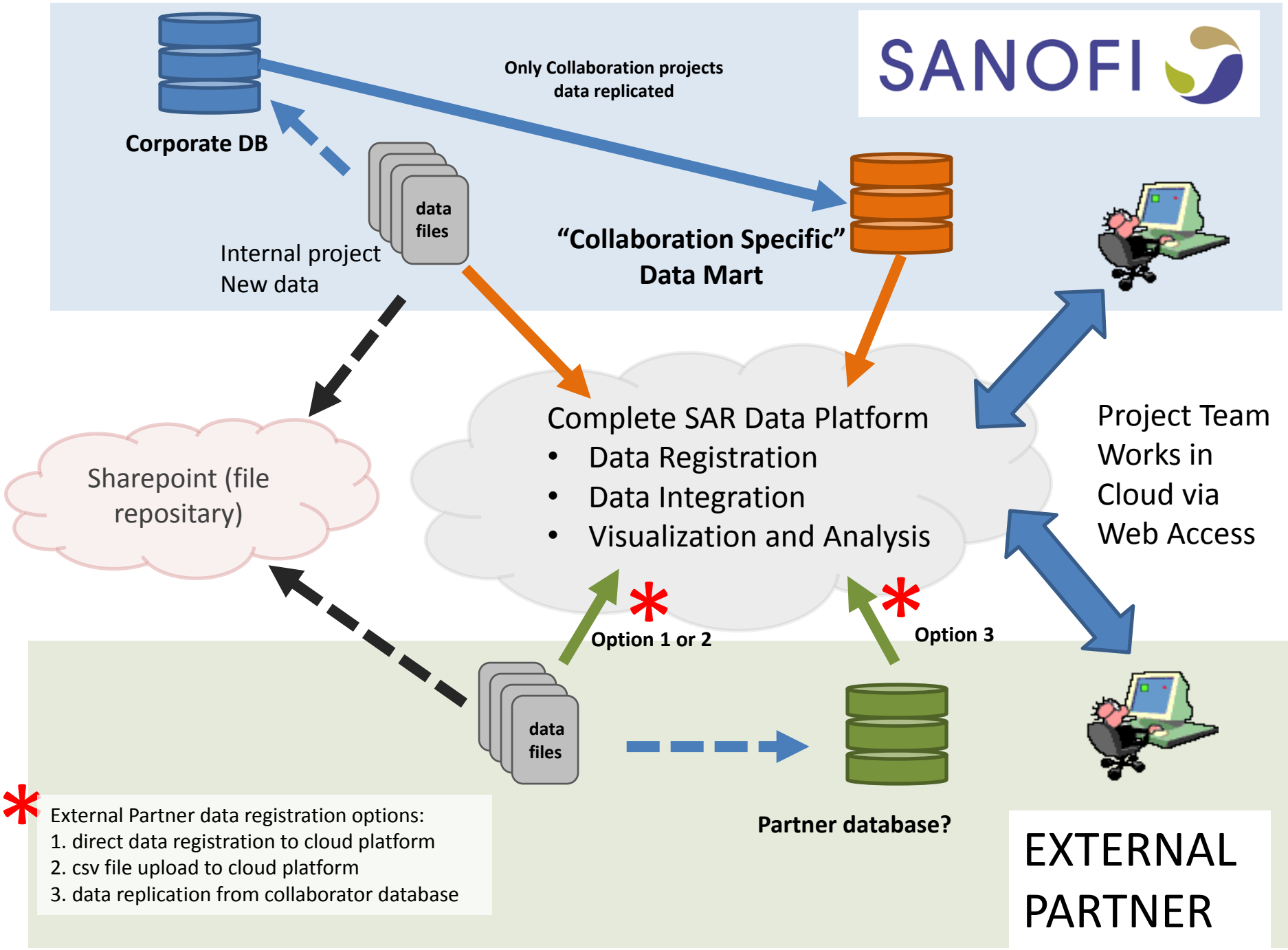
- **In the context of a partnership with a big Pharma or Biotech Partner, the requirements are more complex…**



Enterprise Solution for External Collaborations

Flexibility to allow configuration for **many Collaboration Models**

- Outsourcing to **Chemistry CROs** (smw, peptides and soon biologics)
- Outsourcing to **Bio-profiling assay providers**
  - **Integration with internal request system**
  - **automated results upload by providers**
- Collaboration with other **Pharmas, Biotechs and Academics,** with or without **Data Ownership**
  - **Harmonization and conversion of chemical representations and biological data terminologies and models**
  - **Automatic import from DW with mapping of partner model**
  - **Applications provided to partner and used by all scientists**
  - **Fast and clean project closure and IP tracking**
- **Consortiums, PPPs**
  - **Adjust automatic transfer with on-premise systems to comply consortiums business rules**
- **Evolution to Fully Capable SAR Data Platform "in the Cloud"**


SANOFI
Only Collaboration projects data replicated
Corporate DB
data files
Internal project New data
"Collaboration Specific" Data Mart
Complete SAR Data Platform
• Data Registration
• Data Integration
• Visualization and Analysis
Sharepoint (file repositary)
Project Team Works in Cloud via Web Access
Option 1 or 2
Option 3
data files
Partner database?
External Partner data registration options:
1. direct data registration to cloud platform
2. csv file upload to cloud platform
3. data replication from collaborator database
EXTERNAL PARTNER

Single Vendor SAR Data Platforms "in the Cloud"; a Sampling

- Vendor 1 and 2; Platform designed for collaborative SAR data exchange, evolving into powerful full SAR Data Platform in the Cloud
- Vendor 3; Hosted Complete SAR Data Platform designed for rapid, inexpensive deployment to Biotech companies. (Also for collaboration)
- Vendor 4; Complete, fully integrated SAR Data Platform moved into the Cloud
- Vendor 5; Easily configurable SAR data platform that can be adapted to support collaborations
- Vendor 6; Powerful analysis and modeling platform available "in the Cloud" that can evolve into a full SAR data platform (with big data)
- Vendor 7; Data capture platform offering web access "in the Cloud" and integration with existing on-premise systems

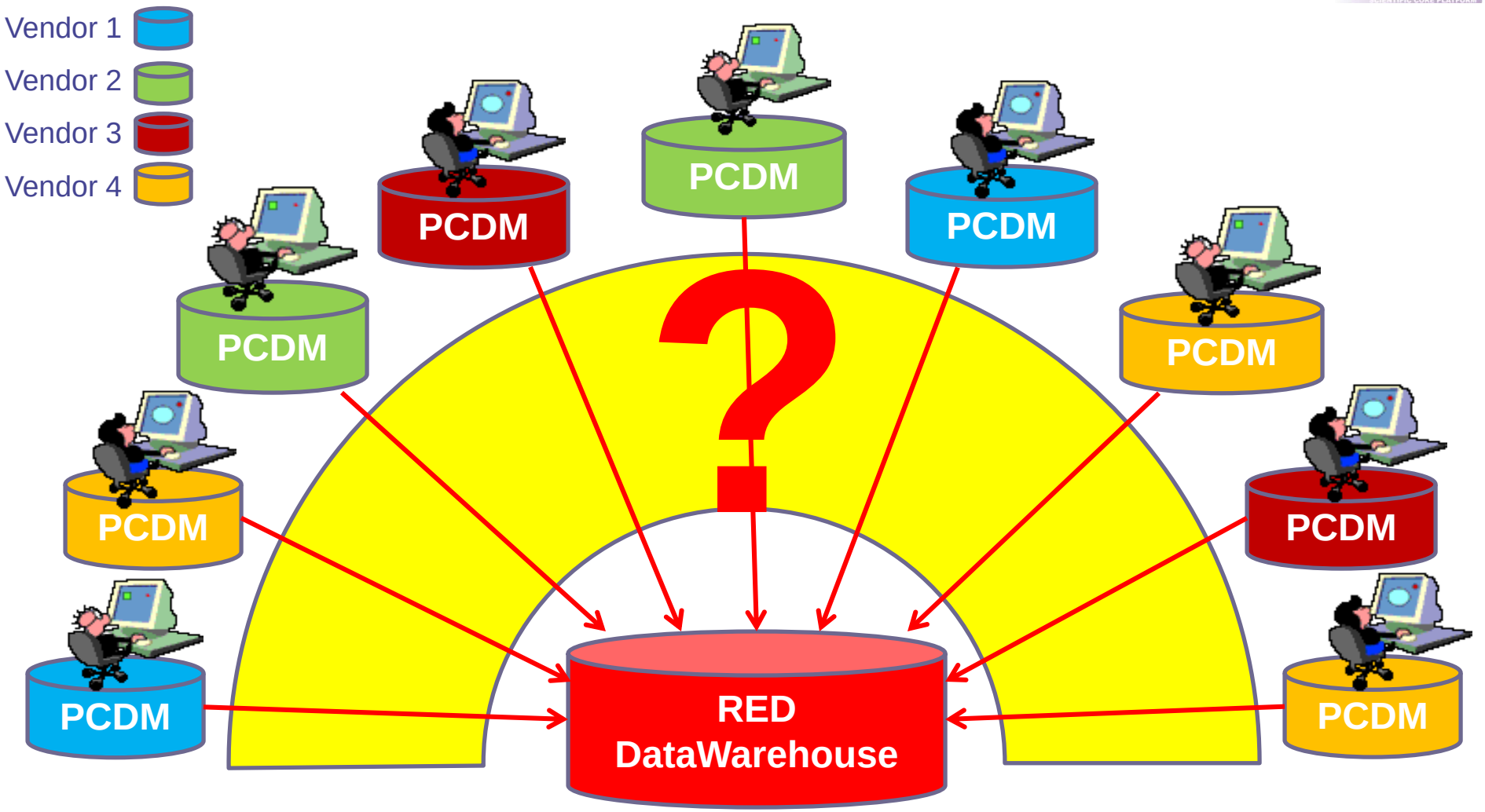
Commonalities

- **Web services access**
- **Easy configuration and access control**
- **Off-prem or using Cloud resources**

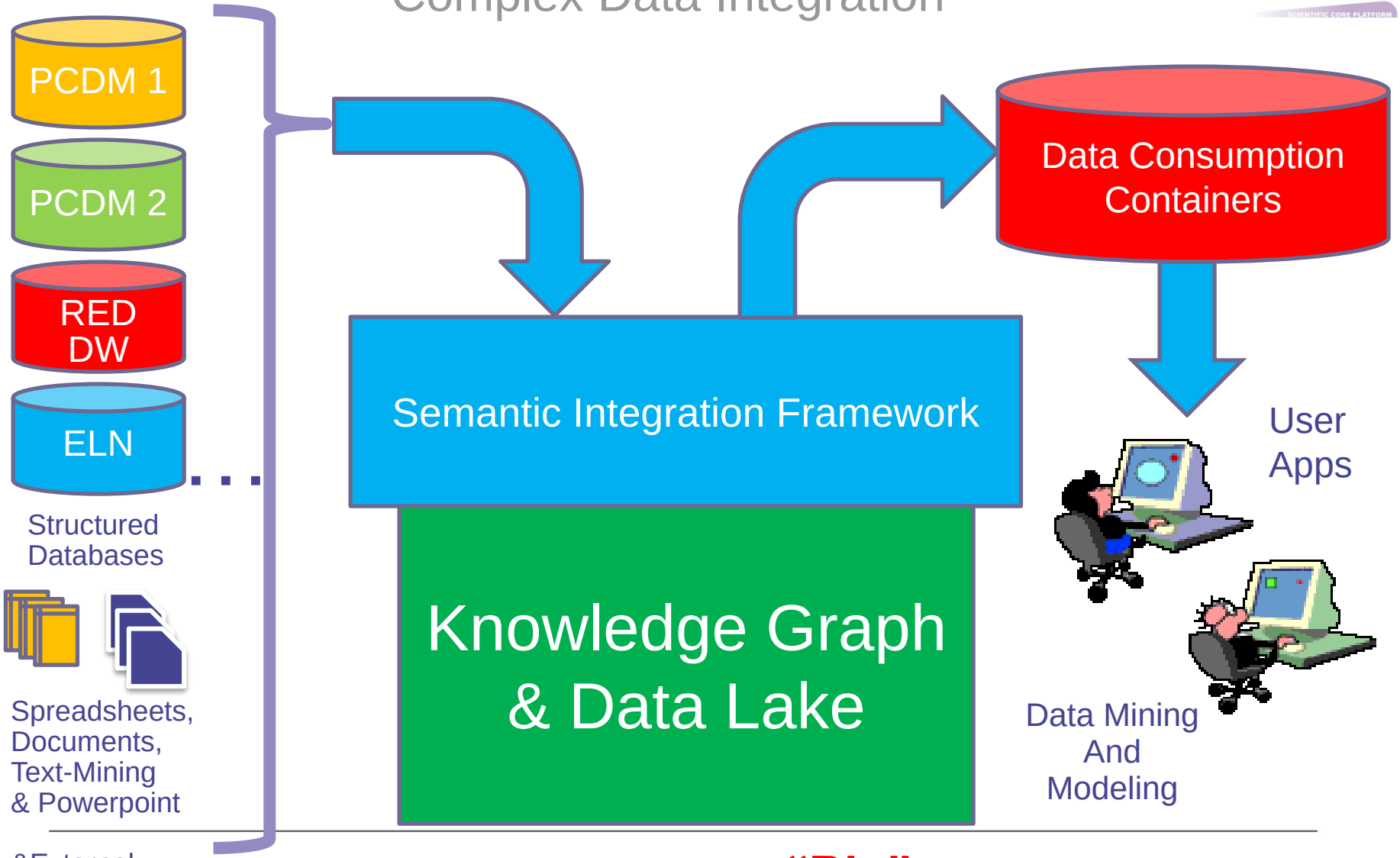
Vendor 8; Creating data integration platform that is application agnostic

These platforms should drive data standards

How do we Integrate Data Islands?



"Big Data" Technology used for Complex Data Integration



Data does not have to be "Big" to benefit from "Big Data"

Advantages

- Scalable to Fit Demand
- Changes Funding Model
- Flexible Configuration
- Rapid Evolution of Capabilities and Technology
- Efficient for Project Teamwork and Progression
- Rapid Deployment
- Configurable to Needs of Projects
- Vendors Compete to Deliver Best Platform/Cost

- If these systems, designed for external collaboration, work well, why would we not adopt them for internal projects?

- Vendors should Drive Data Standardization

THANKS FOR YOUR ATTENTION!