OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

Demonstration on data curation and creation of pre-reasoned datasets in the OpenRiskNet framework

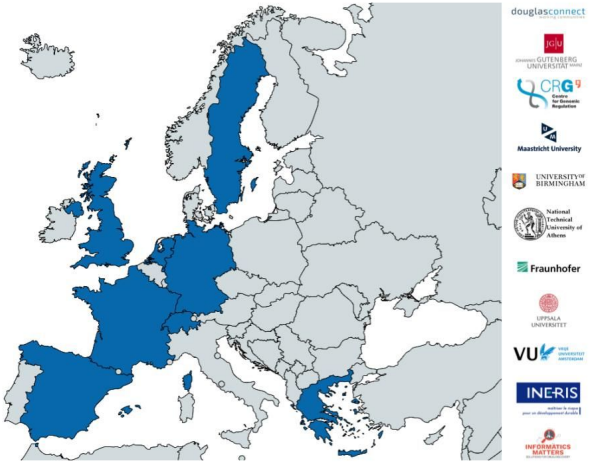
Noffisat Oki (Edelweiss Connect GmbH, Switzerland), **Danyel Jennen** (Maastricht University, The Netherlands), **Marc Jacobs** (Fraunhofer Institute, Germany), **Tim Dudgeon** (Informatics Matters Ltd, UK)

Webinar - 18 March 2019

OpenRiskNet: Open e-Infrastructure to Support Data Sharing, Knowledge Integration and *in silico* Analysis and Modelling in Risk Assessment
Project Number 731075

About the project

OpenRiskNet is a 3-year EU Horizon 2020 project with the main objective to develop an open e-infrastructure providing resources and services to a variety of communities requiring risk assessment, including chemicals, cosmetic ingredients, therapeutic agents and nanomaterials.



Main components:

- **Case-study-driven development** - examples of tools to be integrated are selected based on the case study needs. More information: https://openrisknet.org/e-infrastructure/development/case-studies/
- Solutions for all areas by **integrating existing tools** from consortium and associated partners (via the implementation challenge)
- **Integrated approach** combining experimental data (*in vivo, in vitro, in chemico*) with analysis, modelling and simulation tools into risk assessment workflows

Webinars series

Live demonstrations on the e-infrastructure deployment and the risk assessment case studies

	Topic	Date & Time
Past events	Introduction sessions to the OpenRiskNet e-infrastructure	See Webinar recordings: • Session 1 (24 Sep 2018) • Session 2 (27 Sept 2018) • Session 3 (4 Oct 2018) • Session 4 (30 Oct 2018)
	Learn how to deploy the OpenRiskNet virtual research environment	See Webinar recordings (25 Feb 2019)
Current event	Demonstration on data curation and creation of pre-reasoned datasets in the OpenRiskNet framework	Monday, **18 March 2019** 16:00 CET
Future events	Identification and linking of data related to AOPWiki (an OpenRiskNet case study)	Tuesday, **26 March 2019** 17:00 CET
	Semantic annotation	Monday, **1 April 2019** 16:00 CET (*to be confirmed*)
	The Adverse Outcome Pathway Database (AOP-DB)	Monday, **8 April 2019** 16:00 CET
	Additional demo on case studies	May - June 2019 (*to be announced*)

https://openrisknet.org/events/

DataCure Webinar Content

1.) What is DataCure and how is data curation on OpenRiskNet implemented?
2.) Demonstrations on various aspects of DataCure showing practical examples from the DataCure case study

   a.) Transcriptomics data extraction and metadata annotation

   b.) Text Mining workflow for metadata extraction for carcinogenicity prediction

   c.) Data extraction and curations for liver toxicity modeling

3.) Q&A/Discussion

DataCure case study

- DataCure establishes a process for data curation and annotation that makes use of APIs (eliminating the need for manual file sharing) and semantic annotations for a more systematic and reproducible data curation workflow.

- The aim is to deliver curated and annotated datasets for OpenRiskNet service users as well as preparation and development of tools and workflows that provide examples of useful toxicogenomic data analysis methods will also allow users perform their own data curation and analysis.

Implementation

1.) **Data sources** including Physchem, toxicological and omics databases (e.g. EdelweissData Explorer, PubChem, Liver Toxicology Knowledge Base (LTKB), etc), ontology/terminology/annotation tools (e.g. SCAIView), and literature databases (e.g. Pubmed and Toxplanet).
2.) **Data Extraction** from the EdelweissData Explorer and all other resources using API calls and text mining workflows.
3.) **Data Searching** using workflows that employ text mining capabilities for searching, refinement and curation of the extracted data.
4.) **Data curation and reasoning** using workflows provided in Jupyter or Squonk notebooks and which cover exercises such as extraction of specific data, merging of datasets, and annotation of data for downstream analysis purposes.
5.) **Resubmission to data source**

Use case: finding toxicological information

Scientist: Is this chemical carcinogenic? Or maybe hepatotoxic? Or …?
Is there an easy way to retrieve toxicological information?

Internet:

TOXNET
Toxicology Data Network

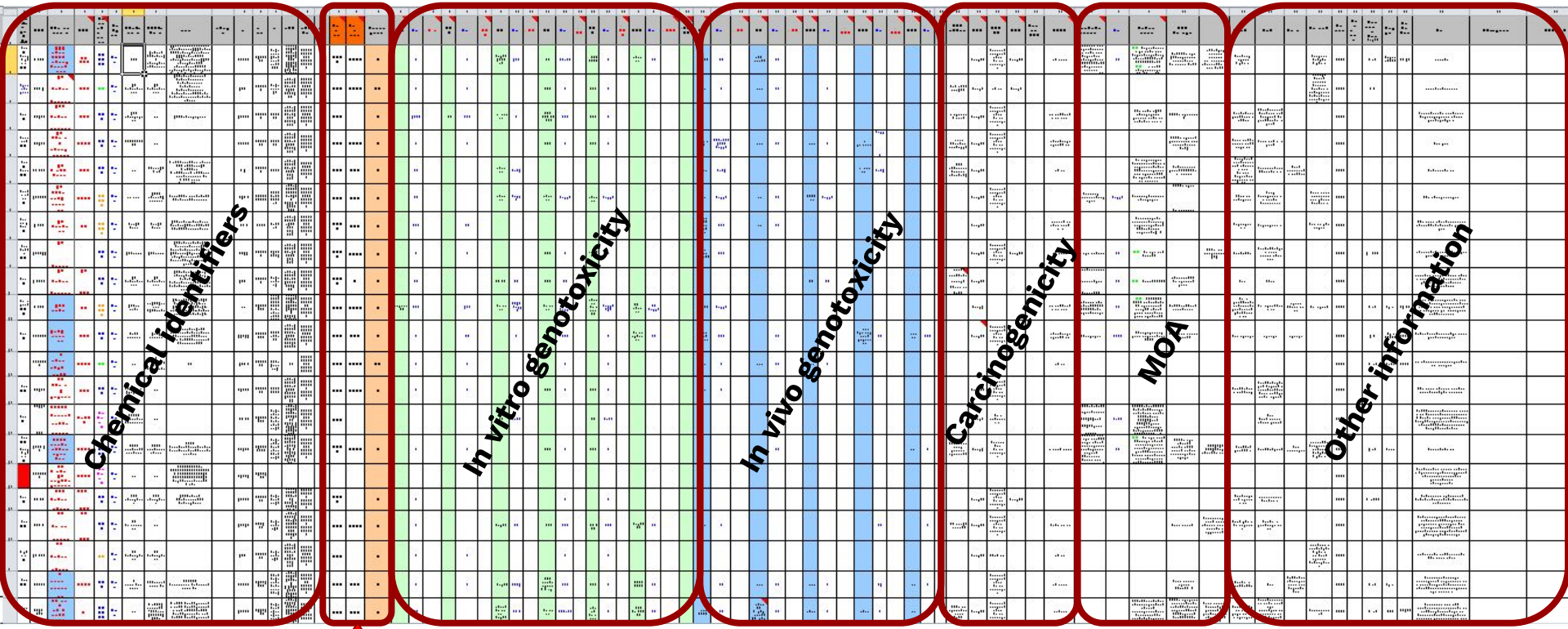
International Agency Research on Cancer
World Health Organization

ChemAgora

EPA

Click-click-click-read-click-read-click-read → answer?

Manual annotation >300 chemicals, >200 references



Data Curation Workflow

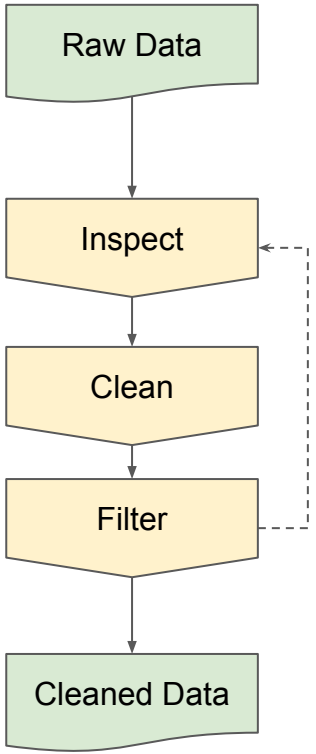
Presented by
Noffisat Oki (Edelweiss Connect Inc, Durham, NC, USA)

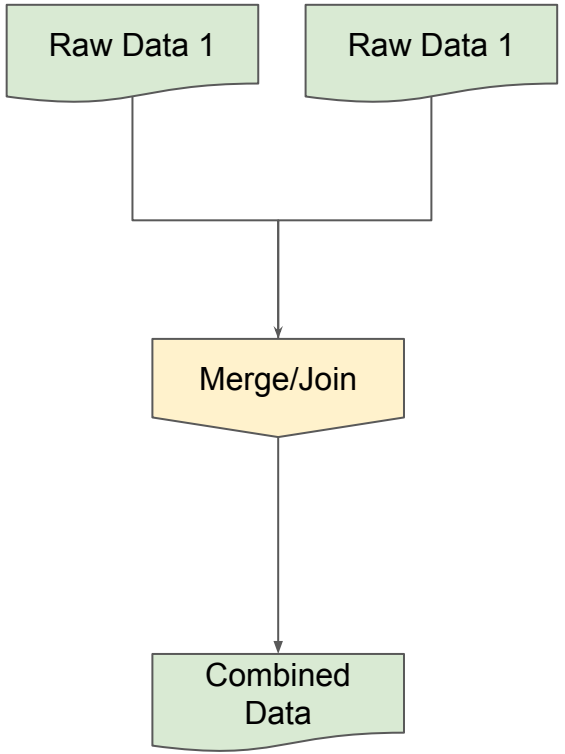
Data curation workflows

Data is messy.

It nearly always needs to be cleaned up and manipulated before being used.

We will illustrate two common needs being performed on the ORN infrastructure.

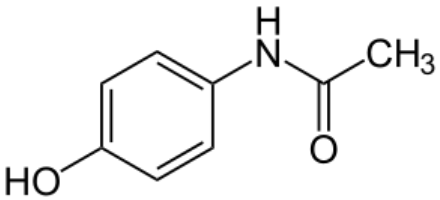




Text Mining Workflow

Presented by
Marc Jacobs (Fraunhofer Institute, Germany)

Text Mining Workflow



acetaminophen

???

OpenRiskNet

IARC classification	*Carcinogenity*
Group 1	*Carcinogenic to humans*
Group 2A	*Probably carcinogenic to humans*
Group 2B	*Possibly carcinogenic to humans*
Group 3	*Not classifiable as to its carcinogenicity to humans*

https://monographs.iarc.fr/agents-classified-by-the-iarc/

1) https://openrisknet.org/e-infrastructure/development/case-studies/case-study-datacure/

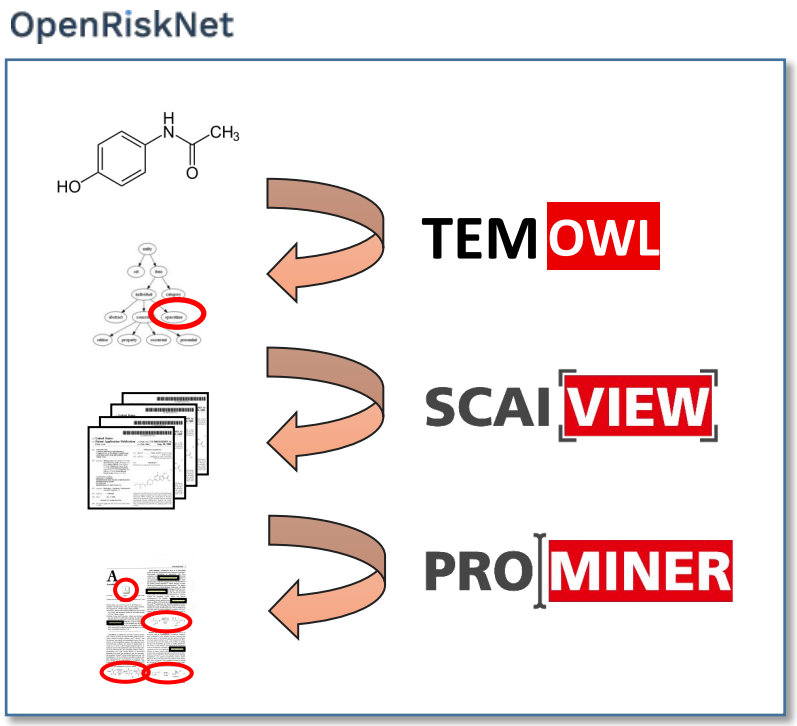
Text Mining workflow

Task:

- Identify the concept of acetaminophen (definition, identifiers, synonyms)
- Find all relevant documents in the context of acetaminophen and carcinogenity
- What are the most relevant statements?

Technology:

- Semantic index of PubMed/PMC (> 20 terminologies)
- Solr index + OLS index + UIMA pipeline



Data extraction and curations for liver toxicity modeling

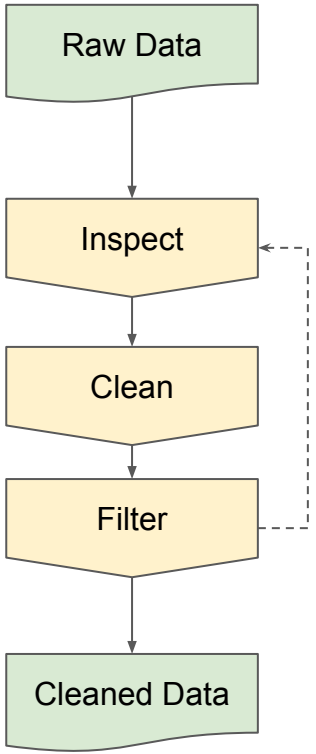
Presented by
Tim Dudgeon (Informatics Matters, UK)

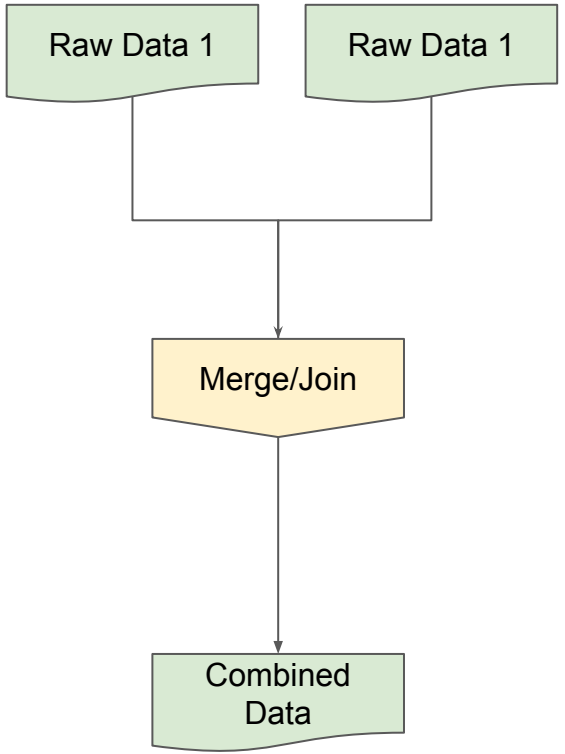
Data curation workflows

Data is messy.

It nearly always needs to be cleaned up and manipulated before being used.

We will illustrate two common needs being performed on the ORN infrastructure.





CPSign - LTKB modeling using Conformal Prediction and SVM

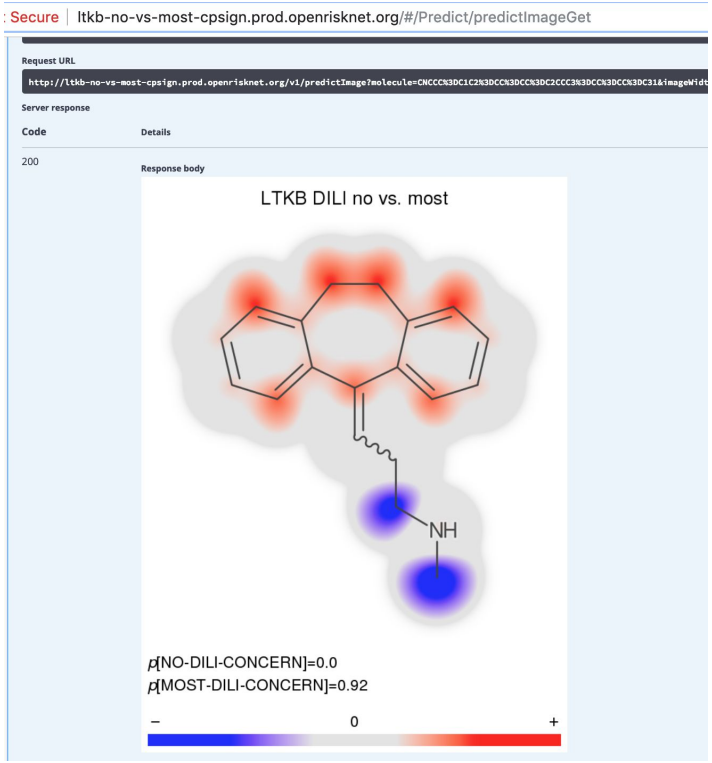
Model building

SVM with RBF kernel; exhaustive parameter sweep for C and γ
Evaluation using 10-fold cross validation
Finding the most **efficient** model
Final model packaged as JAR archive

Web service

OpenShift templating used for deploying the JAR as a web service
Automatically generates an OpenAPI definition for each model
e.g. http://ltkb-no-vs-most-cpsign.prod.openrisknet.org/

Example: nortriptyline



Jaqpot – Example modelling the LTKB data using a Multi-layer Perceptron classifier

Import JaqpotPy package

```
import pandas as pd
from jaqpotpy import Jaqpot
from sklearn import linear_model
from sklearn import neural_network
```

Train NN using your preferred algorithm

```
[ ]  nn=neural_network.MLPClassifier()
```

```
[ ]   neural_model=nn.fit(X,Y)
```

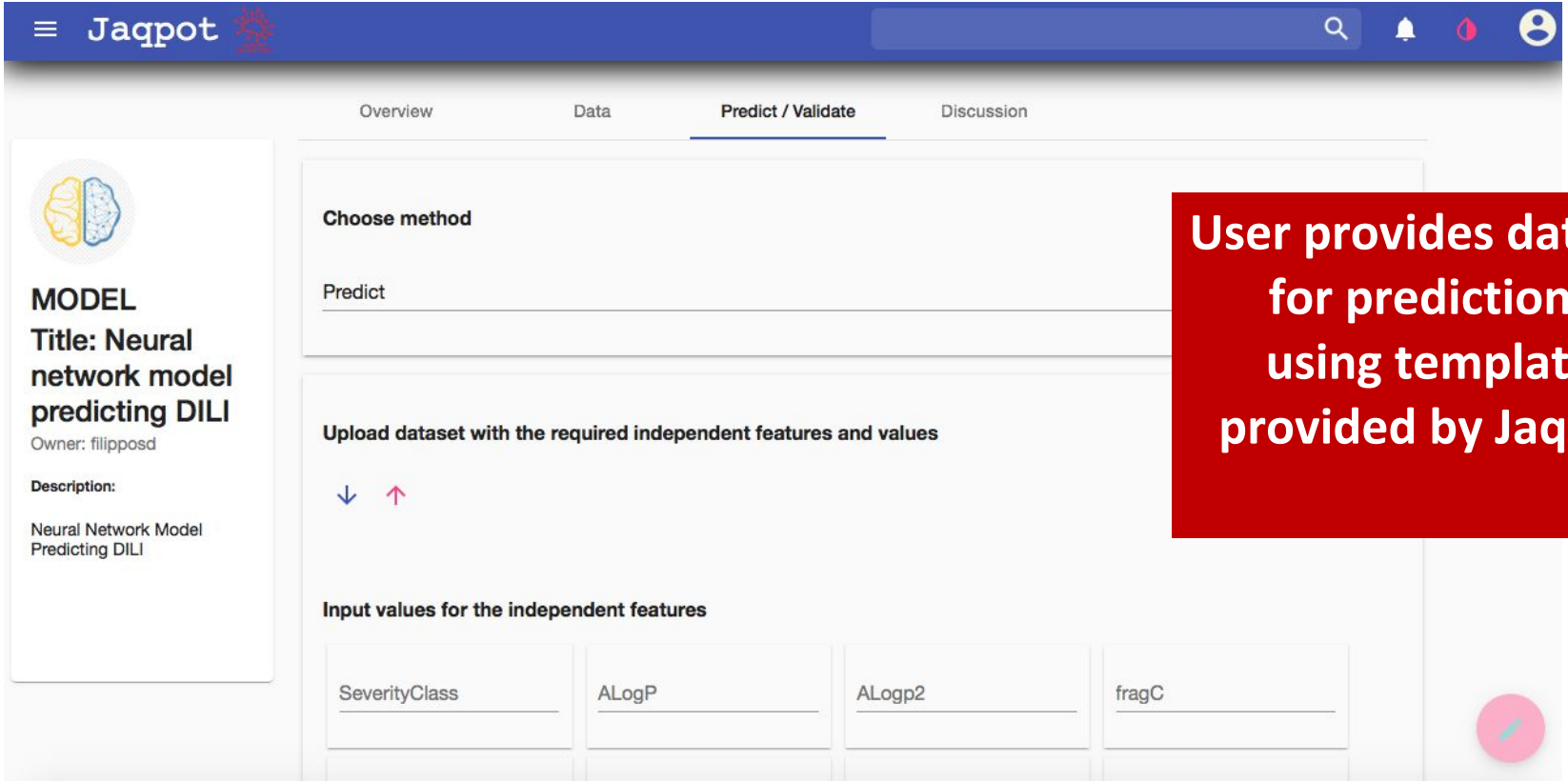
```
[ ]  nn.predict(X)
```

Turn the model into web service

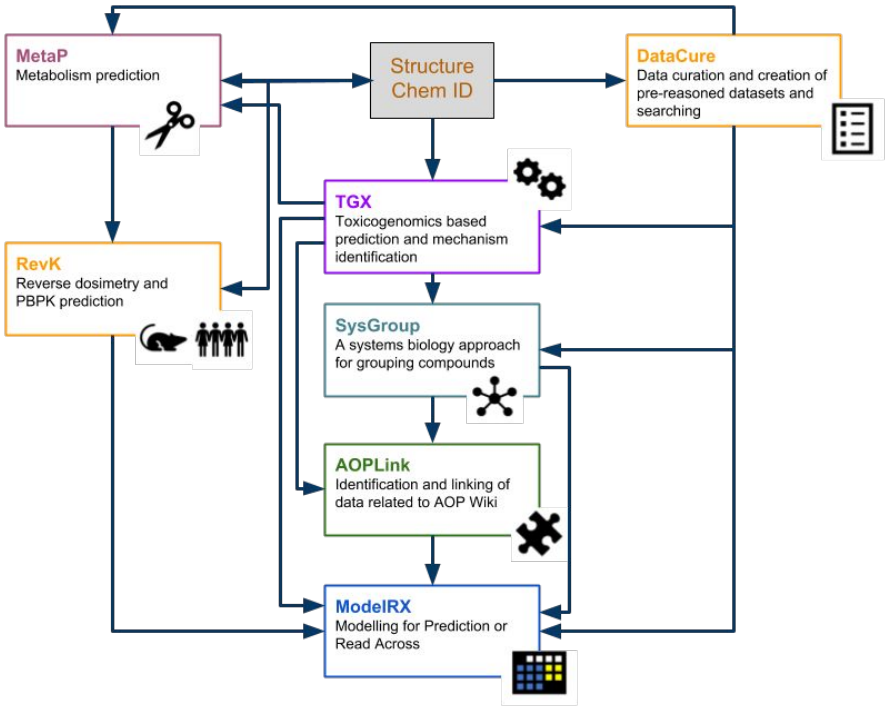
```
[ ]  jaqpot.deploy_neural_network(neural_model,X,Y,"Neural network
```

- **any** *scikit-learn* algorithm under Python
- 1 command gets model uploaded
- Instant
- Model becomes a **web service**

Jaqpot - predict using web service



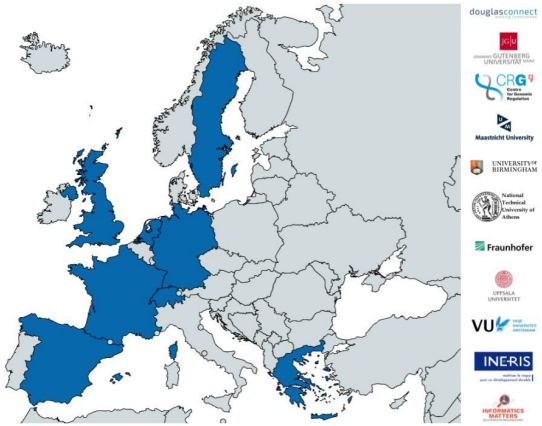
User provides dataset for predictions using template provided by Jaqpot


MetaP
Metabolism prediction
Structure
Chem ID
DataCure
Data curation and creation of pre-reasoned datasets and searching
TGX
Toxicogenomics based prediction and mechanism identification
RevK
Reverse dosimetry and PBPK prediction
SysGroup
A systems biology approach for grouping compounds
AOPLink
Identification and linking of data related to AOP Wiki
ModelRX
Modelling for Prediction or Read Across

Acknowledgements

OpenRiskNet (Grant Agreement 731075) is a project funded by the European Commission within Horizon 2020 Programme

Project partners:



P1 Douglas Connect GmbH, Switzerland (DC)
P2 Johannes Gutenberg-Universität Mainz, Germany (JGU)
P3 Fundacio Centre De Regulacio Genomica, Spain (CRG)
P4 Universiteit Maastricht, Netherlands (UM)
P5 The University Of Birmingham, United Kingdom (UoB)
P6 National Technical University Of Athens, Greece (NTUA)
P7 Fraunhofer Gesellschaft Zur Foerderung Der Angewandten Forschung E.V., Germany (Fraunhofer)
P8 Uppsala Universitet, Sweden (UU)
P9 Medizinische Universität Innsbruck, Austria (MUI)
P10 Informatics Matters Limited, United Kingdom (IM)
P11 Institut National De L'environnement Et Des Risques INERIS, France (INERIS)
P12 Vrije Universiteit Amsterdam, Netherlands (VU)

Webinars series

Live demonstrations on the e-infrastructure deployment and the risk assessment case studies

	Topic	Date & Time
Past events	Introduction sessions to the OpenRiskNet e-infrastructure	See Webinar recordings: • Session 1 (24 Sep 2018) • Session 2 (27 Sept 2018) • Session 3 (4 Oct 2018) • Session 4 (30 Oct 2018)
	Learn how to deploy the OpenRiskNet virtual research environment	See Webinar recordings (25 Feb 2019)
Current event	Demonstration on data curation and creation of pre-reasoned datasets in the OpenRiskNet framework	Monday, **18 March 2019** 16:00 CET
Future events	Identification and linking of data related to AOPWiki (an OpenRiskNet case study)	Tuesday, **26 March 2019** 17:00 CET
	Semantic annotation	Monday, **1 April 2019** 16:00 CET (*to be confirmed*)
	The Adverse Outcome Pathway Database (AOP-DB)	Monday, **8 April 2019** 16:00 CET
	Additional demo on case studies	May - June 2019 (*to be announced*)

https://openrisknet.org/events/