On monsters and tags...

Jeanette Prinz
Greg Landrum
KNIME
Fun with Tags

Blog post: https://www.knime.com/blog/fun-with-tags
Workflow:
knime://EXAMPLES/08_Other_Analytics_Types/02_Chemistry_and_Life_Sciences/03_Fun_with_Tags
We can’t rely on other people’s graffiti

Banksy Painting Self-Destructs After Fetching $1.4 Million at Sotheby’s

https://www.nytimes.com/2018/10/06/arts/design/uk-banksy-painting-sothebys.html

https://consequenceofsound.net/2018/10/banksy-painting-self-destructs/
Text Tagging

Text tagging is the process of adding a tag (language element descriptor) to unstructured data.
Motivation

- Automated access to disease information is an important goal of text-mining efforts
- Rapid growth of biomedical literature => impossible for humans alone to extract and exhaust all of the useful information
- Enables integration with other data types and the generation of new hypotheses
Dictionary: contains disease names from Ensembl Biomart assembled from different sources such as OMIM, Orphanet, and DDG2P.
Corpus Creation?

- In linguistics, a **corpus** is a large and structured set of texts (usually electronically stored and processed) ([https://en.wikipedia.org/wiki/Text_corpus](https://en.wikipedia.org/wiki/Text_corpus))
- KNIME node “Document Grabber” allows to extract abstracts from **PubMed** containing disease names from our dictionary. We only keep results with min. 20 hits in PubMed, and collect max. 100 documents per disease
The Workflow

Dictionary & Corpus Creation
- Create Corpus from PubMed
- String Manipulation
- Disease Names
  - read disease names
  - this node accesses PubMed several times
  - It needs internet connection and it might take long time to fully execute!

Model Training & Evaluation
- Model Training and Evaluation input dictionary
- Compare with
- Select diseases of interest to be inspected in next nodes/views
- Follow instructions provided in the view

Co-occurrence of Tagged Disease Names
- Co-Oc Network
- SubGraph
  - builds co-occurrence network
  - Select disease of interest
  - visualizes sub-graph
The Workflow
Demo
Ohdo Syndrome - KAT6B

Whole-Exome-Sequencing Identifies Mutations in Histone Acetyltransferase Gene KAT6B in Individuals with the Say-Barber-Biesecker Variant of Ohdo Syndrome

Jill Clayton-Smith, James O'Sullivan, Sarah Daly, Shehla Mohammed, Dian Donnai, Graeme Black

Open Archive • Published online: November 10, 2011 • DOI: https://doi.org/10.1016/j.ajhg.2011.10.008
Monster Model Factory

Video: https://bit.ly/2O8GIqs
Workflow: knime://EXAMPLES/50_Applications/37_Monster_Model_Factory
Build, validate, and deploy models for multiple data sets

- >1500 datasets from ChEMBL that I would like to build models for
- I want to actually use the models, so they need to be deployed
- The whole process needs to be automated and reproducible so that I can do it again when ChEMBL is updated
- Maybe we can learn something interesting from the models themselves
The model process
The model process, multiple models
The model process, multiple models
AUTOMATE

ALL THE THINGS!

https://www.publicdomainpictures.net/view-image.php?image=155188
Make each step a separate workflow. Use KNIME to orchestrate calling those workflows.

KNIME blog post: [https://goo.gl/LvESqB](https://goo.gl/LvESqB)
White paper: [https://goo.gl/d6UpUu](https://goo.gl/d6UpUu)
Monster Model Factory workflow:

knime://EXAMPLES/50_Applications/37_Monster_Model_Factory
The heart of the factory: Call Local Workflow$^1$

- Executes another workflow in the same local repository

$^1$ Call Remote Workflow when run on the KNIME Server$^2$

$^2$ If we were doing this today, we’d use the Call Workflow node.

Monster Model Factory workflow:

knime://EXAMPLES/50_Applications/37_Monster_Model_Factory
Monster Model Factory workflow:

knime://EXAMPLES/50_Applications/37_Monster_Model_Factory
Monster Model Factory workflow:

knime://EXAMPLES/50_Applications/37_Monster_Model_Factory
Monster Model Factory workflow:

knime://EXAMPLES/50_Applications/37_Monster_Model_Factory
Monster Model Factory workflow:

knime://EXAMPLES/50_Applications/37_Monster_Model_Factory
Monster Model Factory workflow:

knime://EXAMPLES/50_Applications/37_Monster_Model_Factory
Monster Model Factory

Abstract Model Factory

Parameter Optimization Loop Start
Compute additional method parameters
Random Forest Learner

Brute Force Fingerprint Choice Number of Trees

Partitioning
Random Forest Predictor
Compute Kappa & AUC

stratified 80-20

Random Forest
H2O Gradient Boosting

Load

Math Formula (Variable)

Counting Loop Start seed with iterations
10 times random

Defining path to server

Init Access Preprocess Train Score Deploy
Monster Model Factory


Workflow: knime://EXAMPLES/04_Analytics/11_Optimization/08_Model_Optimization_and_Selection
Task Scale

- 1500 sets
  With up to 40k rows
- 4 ML
- Parameter optimization
  with 5 features
- 10 iterations with a random split to Training and Test sets
- 300k models

Flowchart:

- Access
- Preprocess
- Train
- Score
- Deploy
Scaling Execution
Scale with KNIME Server Distributed Executors

Build/test workflows → KNIME Server → Run model factory → Run individual data sets

65-70 load-balanced distributed executors
Model Factory

Abstract Model Factory

Prepare Modelling Configuration

© 2018 KNIME AG. All rights reserved.
Summary

• Automated the Model Management Process using KNIME Analytics Platform
• In total 300K models were built\(^1\)
• Scaled it for our task with KNIME Server distributed executors
• Less than a week\(^2\)

1 ~1550 assays * 4 methods * 5 FPs * 10 repeats
2 ~4400 hours computing time on single CPU

Workflow: knime://EXAMPLES/50_Applications/37_Monster_Model_Factory