COMAS
Compound Management and Screening Center

MAX-PLANCK-GESELLSCHAFT

Data Pipelines and Mol_Lists: RDkit Tools for the Jupyter Notebook

RDKit UGM 2016

26. - 28. Oct 2016

Axel Pahl

Dr. Axel Pahl
Medicinal Chemistry
COMAS – Compound Management and Screening Center
MPI of Molecular Physiology

whoami

- **Medicinal Chemist**
  - COMAS: COmpound MAnagement and Screening center
    - MPI of Molecular Physiology, Dortmund
  - still doing synthesis
- **Linux / OpenSource Enthusiast**
- **Programming Interests**
  - Python, Postgresql, RDKit, HTML
  - Nim (http://www.nim-lang.org)
- **Pipeline Pilot, KNIME**
- **Married, two children (12 & 15 yrs. old)**

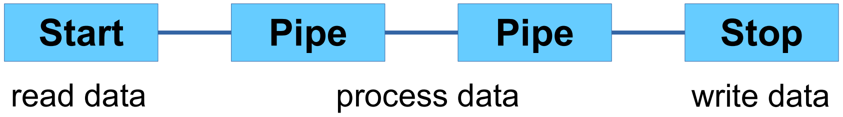
Project Introduction

Introduction

- **Two years ago: interactive SDF viewer with Qt interface**
- **More and more work shifted to the Juypter Notebook**
  - reproducibility and traceability of tasks
- **Wanted:**
  - to work with lists of molecules instead of dataframes
  - high quality (publication grade) plots with structure tooltips
  - workflows (read: pipelines) to search and filter large data sets and break them down to manageable sizes
- **This started the rdkit_ipynb_tools project**

(obviously I suck at naming projects)

rdkit_ipynb_tools



- **Module pipeline**
  - data workflows implemented as Python generators
  - low memory impact, high performance
- **Module tools**
  - Mol_List class
- **Module clustering**
  - tools for compound clustering and reporting (not covered today)
- **Module scaffolds *WIP***
  - tools for working with scaffolds, may be moved to a devel branch
- **Available on GitHub: https://github.com/apahl/rdkit_ipynb_tools (includes the tutorial notebook)**

Demo Time!