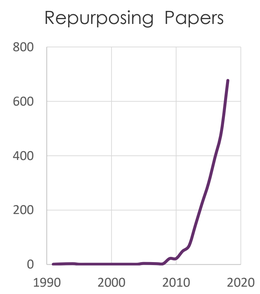
NCATS Stitcher:
Data Integration and Guided Curation Tool

Ivan Grishagin[1], Tyler Peryea[2], Daniel Katzel[1], Tongan Zhao[2], Trung Nguyen[2], Ajit Jadhav[2], Noel Southall[2]

[1]Rancho BioSciences, LLC, 16955 Via Del Campo #220, San Diego, CA
[2]National Center for Advancing Translational Sciences (NCATS), 6701 Democracy Boulevard, Bethesda, MD
Email: Ivan.Grishagin@RanchoBioSciences.com or Noel.Southall@NIH.gov

Background

- ❑ Untapped Big Data [1]
  - = Data accumulation is exponential
  - = Only 2% of potentially useful data is analyzed
- ❑ Aggregate to validate!
  - = Drug repurposing research up 35 times in 2018 vs. 2010 [2]
- ❑ Data merging is challenging
  - = Multiple formats
  - = "Dirty" data – need manual curation!
  - = Entity resolution/normalization is still unsolved
  - = ETL tools or custom scripts or expensive subscriptions



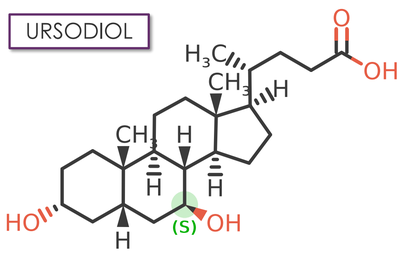
[1] https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
[2] https://www.ncbi.nlm.nih.gov/pubmed/?term=repurpose

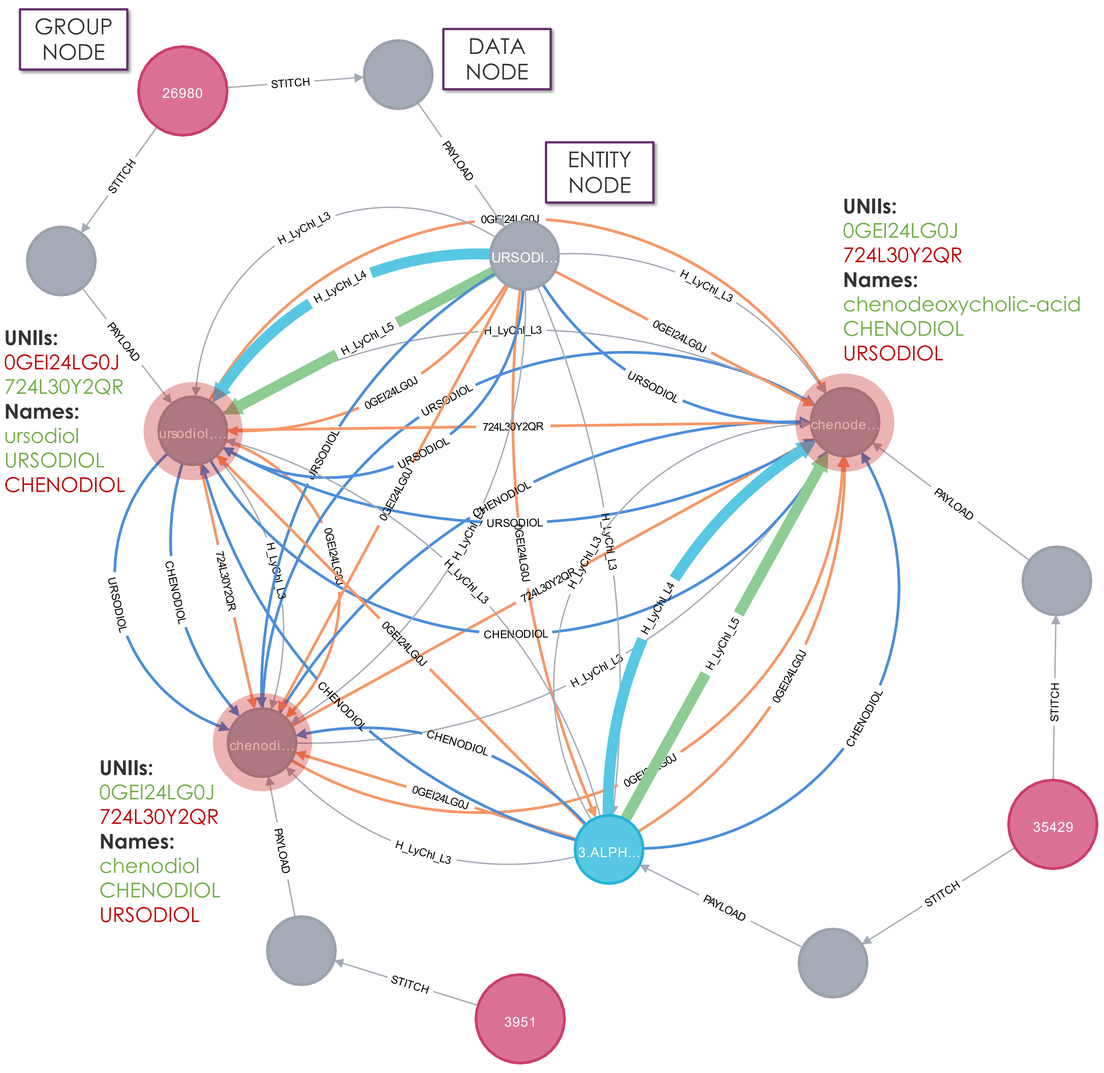
Implementation

- ❑ Stored as a Neo4j graph database
  - = computationally efficient
- ❑ Novel deterministic algorithm
  - = guarantees accuracy of all connections
  - = allows to detect data fidelity issues
- ❑ Java app => JSON response
- ❑ RESTful API:
  stitcher.ncats.io/api
  GET
  + /stitches/latest/:id
- ❑ Curation Web UI

Case Study

- ❑ Test via Stitcher API for shared UNIIs **=>**
- ❑ CHENODIOL is erroneously connected to URSODIOL **=>**
- ❑ Original data source has three entries with incorrect UNIIs / Names



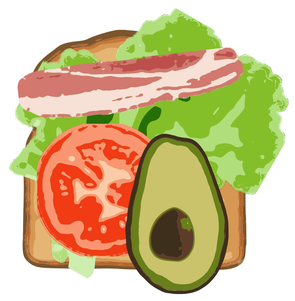


Goal and Concept

- ❑ **Combine** multiple incomplete data sources to **identify ground truth**
- ❑ Auto-detect inconsistencies and errors to **guide manual curation**

- ❑ **Build** a comprehensive knowledge **graph** for all substances
  - ➢ Use multiple reliable public sources
- ❑ **Locate** unique entities and **cluster** pertaining nodes
- ❑ **Test** the graph automatically
  - ➢ Shared UNII clashes
  - ➢ "Orphan" substances
  - ➢ Megaclusters
- ❑ **Annotate** where appropriate
  - ➢ Use test results to guide manual curation
  - ➢ Non-destructive edits
  - ➢ Maintain data provenance
  - ➢ Contribute to original data source

B.L.T.A.



Algorithm