BRIEF BIO

Nuno Bandeira, Ph.D., bandeira@ucsd.edu

Associate Professor, University of California, San Diego

- Computer Science and Engineering
- Skaggs School of Pharmacy and Pharmaceutical Sciences
- Founding faculty, Halıcıoğlu Data Science Institute (HDSI)
- Research interests: algorithms for interpretation of mass spectrometry data from metabolomics, proteomics, natural products and microbiome samples

Executive Director, Center for Computational Mass Spectrometry (CCMS)

- NIGMS P41 Biomedical Technology Research Resource, since 2008
- Multi-PI with Vineet Bafna, Pavel Pevzner
- Research algorithms developed for proteomics, metabolomics, proteogenomics, natural products drug discovery, protein-protein interactions, etc.
  - Including algorithms for global-scale integration of discoveries into reusable knowledge bases
- Major service platforms for sharing data, tools and community-curated knowledge

CENTER FOR COMPUTATIONAL MASS SPECTROMETRY

Big Data

MassIVE

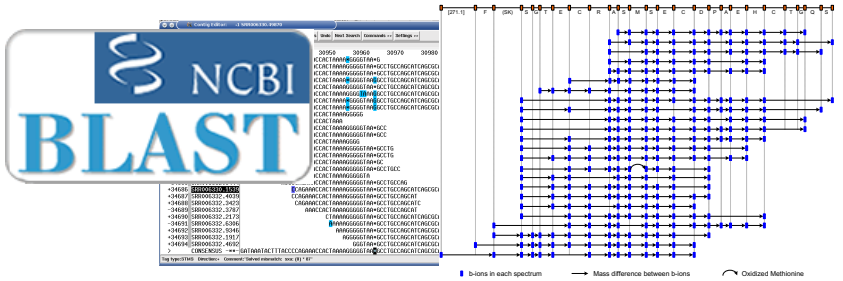
Mass Spectrometry
Interactive **V**irtual **E**nvironment

proteome CHANGE

>11,500 public + private datasets, hundreds of terabytes

http://massive.ucsd.edu

Algorithms



Designed to build on rather than just scale to big data

http://proteomics.ucsd.edu/software

Compute

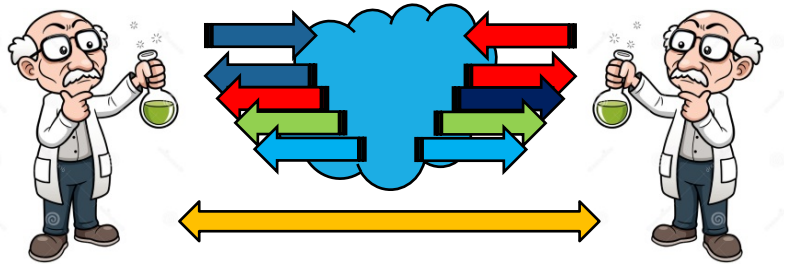
***Proteo*mics *S*calable, *A*ccessible and *F*lexible *e*nvironment**

ProteoSAFe

80+ data analysis workflows
>200k jobs on >2500 cores, >30 billion spectra

http://proteomics.ucsd.edu/ProteoSAFe

Community



Community-wide sharing of data+knowledge, >45,000 users from >140 countries

http://gnps.ucsd.edu

INITIAL THOUGHTS

→ *Assess + Increase value*
→ *Engage researchers*

Tools and practices that NLM could use to help researchers and funders better integrate risk management practices and considerations into data preservation, archiving, and accessing decisions

- Develop (and support) domain-specific community standards for what constitutes high-quality reusable data
- Assess "storage value" of datasets for future research based on i) uniqueness and ii) potential for additional discoveries (e.g., quantify "undiscovered knowledge")
- Repositories for each data type should function like "data journals": i) enforcing publication standards, ii) evaluating datsets and iii) increasing+assessing dataset value over life cycle
  - Repositories need to track direct AND indirect need for datasets; should report to funders *and* data generators

Methods to encourage NIH-funded researchers to consider, update, and track lifetime data costs; and

- Proposals asking for new data should assess "generation cost" of dataset by requiring dataset deposition with metadata *and publicly-accessible itemized costs*
  - Cover full reproducibility cost of cohort + data acquisition + analysis + organization/release, etc.
  - Justify why existing data is insufficient → this should inform a) funders supporting the data and b) repositories (if inadequacy relates to data quality) to update the assessed dataset value

Challenges for the academic researchers and industry staff to implement these tools, methods, and practices.

- Research methods to assess "Cost" and "Value" of a dataset – constantly evolving and highly domain-specific
  - Also complicated by diverse data types and levels of knowledge representation
- "Data recommendation" platforms for on-demand integration of tools+datasets into existing research projects
- Public/Private data models, including pay-wall/subscription models