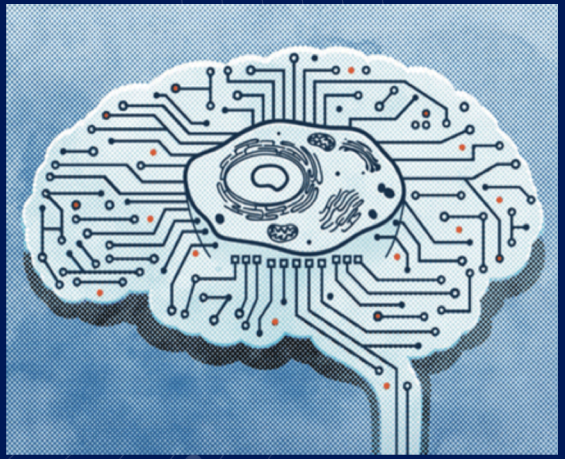
NHGRI Genomic Data Science Working Group (GDSWG)



Trey Ideker, PhD

Division of Genetics
Department of Medicine
University of California San Diego

February 2019

National Human Genome Research Institute

The Forefront of Genomics®

Data & Information Sciences are Integral to Genomics

G

(+e)

Genome reference databases
Common & rare variants

Translation of G to P

Statistical models
& functional knowledgebases

P

Clinical databases
Molecular phenotypes

Data Access & Privacy, Data Visualization

What is the Data Science Working Group?

- Subcommittee of the National Advisory Council for Human Genome Research.
- Newest of the subcommittees of NHGRI Council, formed Spring 2017.
- Meetings are approximately every two months by WebEx.

http://www.genome.gov/27569732/nhgri-genomic-data-science-working-group/

Working Group Functions

- Advises NHGRI on plans related to data sciences outside of Council meetings.
- Deeper engagement of Council in issues of genomic data science, advises the internal Data Sciences Focus Group
- Provides input to NHGRI Director and other Institute leaders about trans-NIH issues related to data science.
- Addresses broad challenges, from data management to data usage to data policy, as they relate to all areas of genomics, from basic science to genomic medicine.

Working Group Members

Michael Boehnke
University of Michigan

Eric Boerwinkle
UT School of Public Health
Baylor College of Medicine

Lon Cardon
BioMarin

Nancy Cox
Vanderbilt University

George Hripcsak
Columbia University

Trey Ideker
UC San Diego

Gail Jarvik
University of Washington

Mark Johnston
University of Colorado School of Medicine

Anthony Philippakis
The Broad Institute of MIT and Harvard

NHGRI Representatives

Valentina Di Francesco, Eric Green, Carolyn Hutter, Allison Mandich

NIH NHGRI

Workgroup Topics in 2017-2018: A Recap

Release of informatics PARs in 2018:

PAR-18-843 (R21), PAR-18-844 (R01), PAR-19-061 (SBIR).

- Based on recommendations from the last NHGRI Informatics Workshop (Sept. 2016).
- To encourage investigator-initiated algorithm and methods development research.
- Focus on algorithm and methods development and hardening.

Award of the NHGRI Genomic Cloud Platform “AnVIL”:

Analysis, Visualization and Informatics Labspace.

https://www.genome.gov/27569268/genomic-analysis-visualization-and-informatics-labspace-anvil/

Workgroup Topics in 2017-2018: A Recap

Restructuring of model organism databases (MODs) *via* the Alliance of Genome Resources (https://www.alliancegenome.org/)

Reviewing and enhancing the Genomic Data Sharing Policy implementation.

Plans for the modernization of the NIH Data Resources Ecosystem delineated in the NIH Strategic Plan for Data Science

Upcoming Workgroup Topics:
Role of Data Sciences in NHGRI Vision2020

Challenging cross-cutting nature of data sciences **across NIH institutes**

- Relation of NHGRI data sciences to that of other ICs
- Boundaries: what this space is, and what it is not.

Challenging cross-cutting nature of data sciences **across NHGRI programs**

- Informatics analysis is supported across multiple NHGRI focus areas and initiatives.
- E.g. stand-alone informatics program, also in ENCODE, Ctrs for Common Disease Gnmcs (CCDGs)
- How to build common analysis pipelines coordinated among data generation centers?

Challenging cross-cutting nature of informatics **across Council working groups**

- Integrated nature of data generation, storage and analysis.
- What data are missing for key analyses of the future?
- What can and cannot be interpolated?

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

Ronald Fisher 1938

Upcoming Workgroup Topics: Final Points

Leveraging advances in AI and machine learning in genomic data analysis

- Techniques like deep learning are revolutionizing many science and engineering disciplines

Education & Training

- Need additional mechanisms to encourage training of data scientists for genome research

Organizing focused data science workshop(s) in NHGRI 2020 Vision Planning

- Visualization
- Community engagement (data sharing & privacy, enabling specific groups with data & tools)

Questions?

The Forefront of Genomics®

Some Genome Analysis Goals for the Future

Predictive performance

Advances in deep learning provide major opportunity to improve current statistical models for polygenic risk, QTLs

Mechanistic interpretability

Models must be interpretable to a granularity that is appropriate depending on goals (multiple problems, multiple solutions)

Transferability across contexts

Evolutionary organisms, cell-line screens, patients, development & aging

Privacy-awareness

Strong practical need to compute on private (e.g. encrypted) data