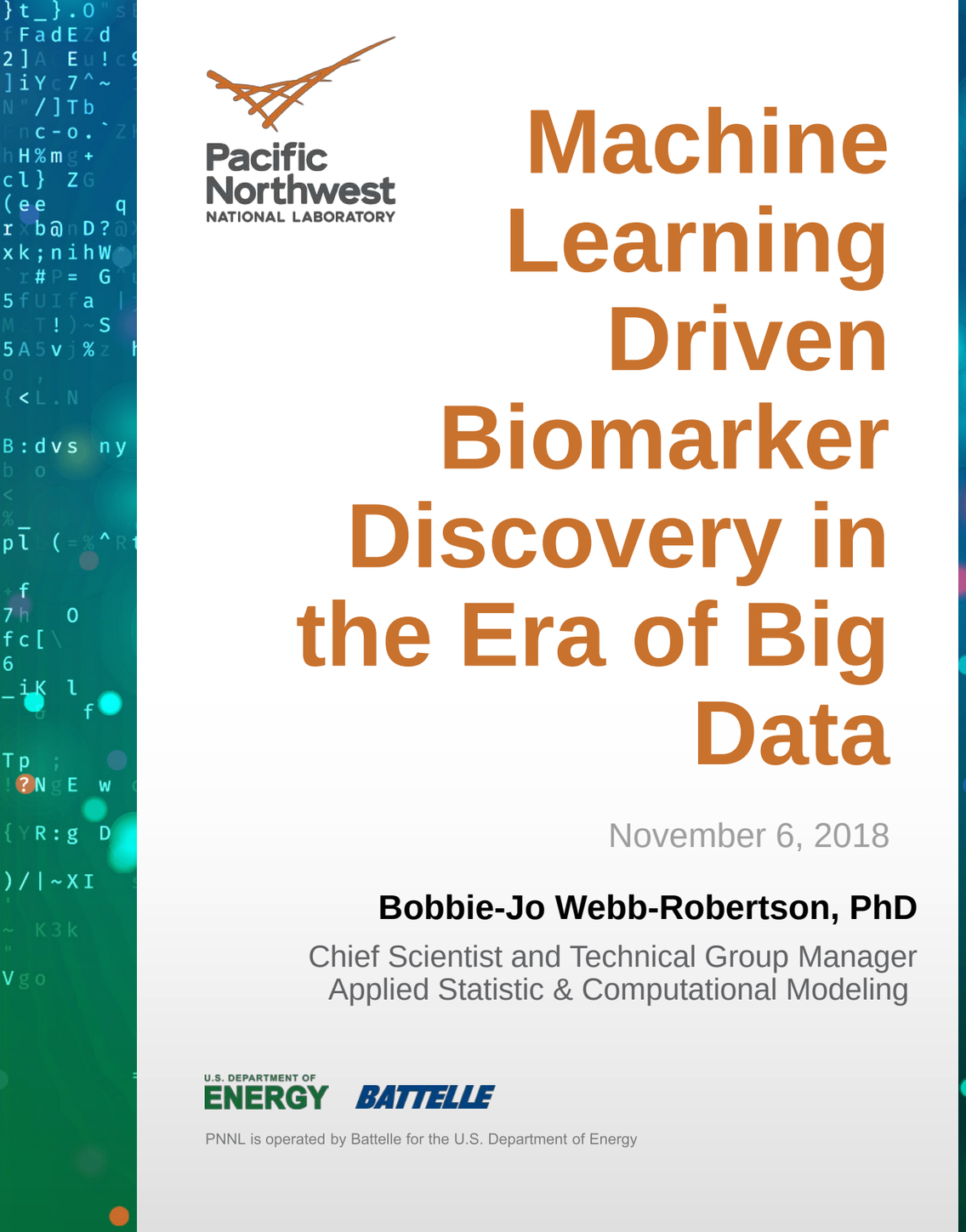
Pacific Northwest
NATIONAL LABORATORY

Machine Learning Driven Biomarker Discovery in the Era of Big Data

November 6, 2018

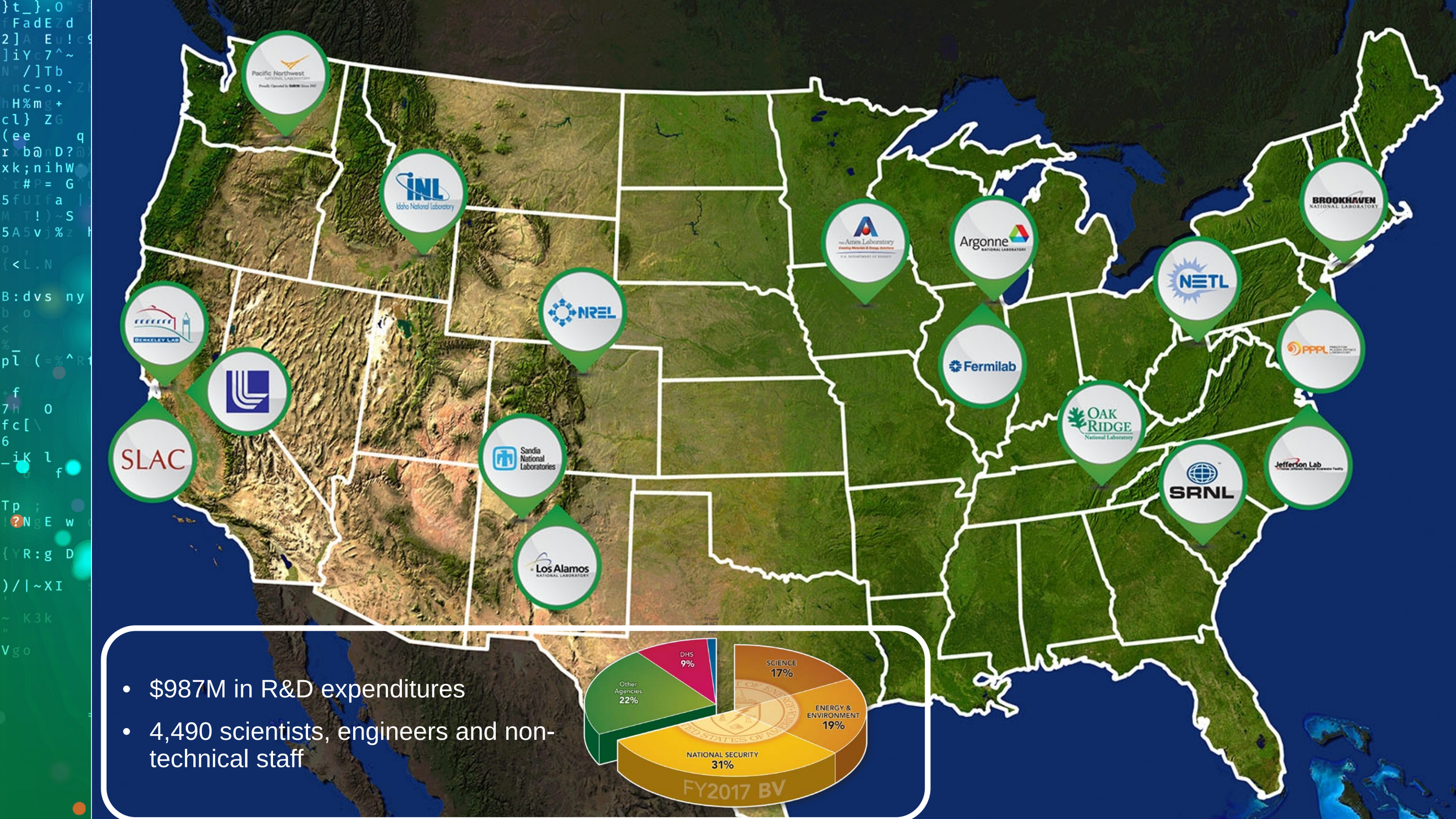
Bobbie-Jo Webb-Robertson, PhD

Chief Scientist and Technical Group Manager
Applied Statistic & Computational Modeling



U.S. DEPARTMENT OF ENERGY BATTELLE

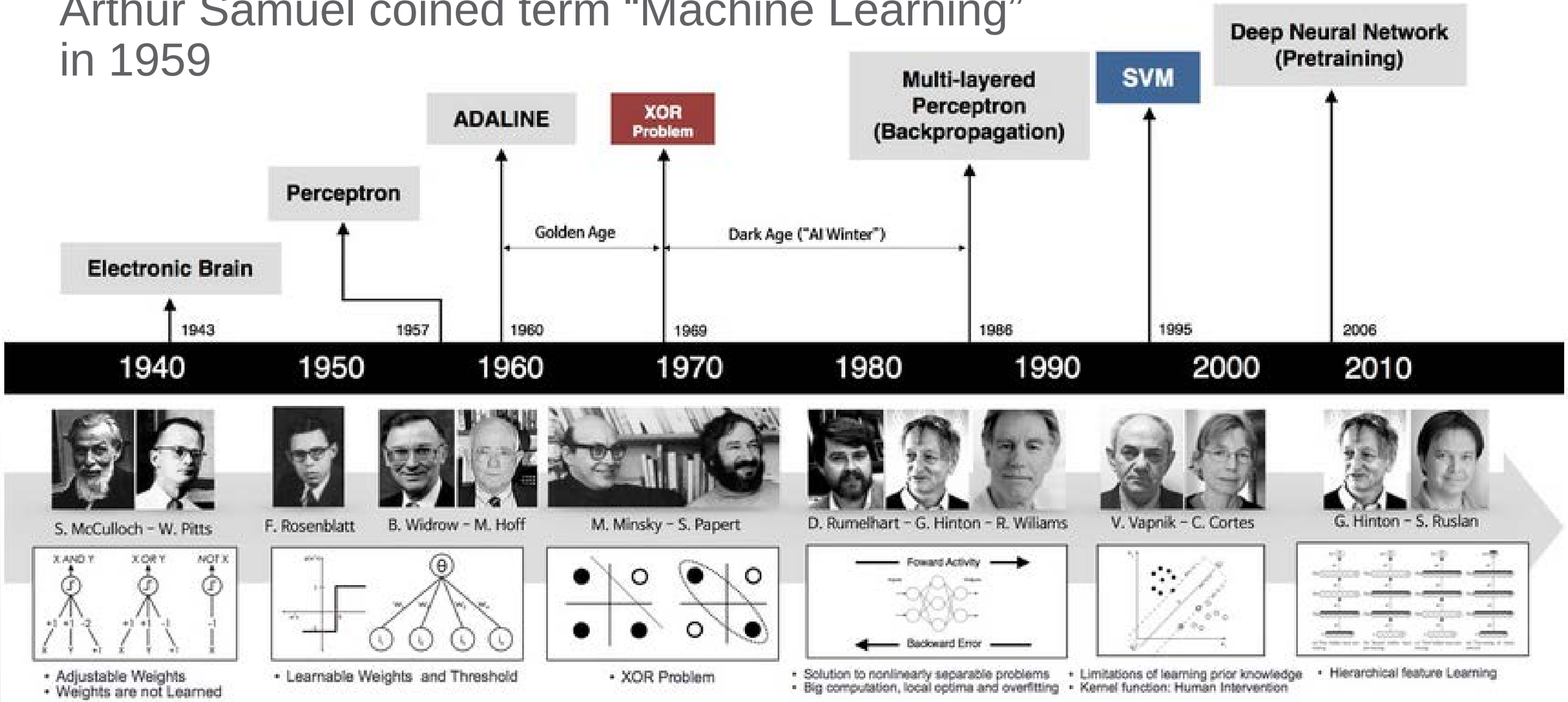
PNNL is operated by Battelle for the U.S. Department of Energy



- $987M in R&D expenditures
- 4,490 scientists, engineers and non-technical staff

Machine Learning

Arthur Samuel coined term “Machine Learning” in 1959



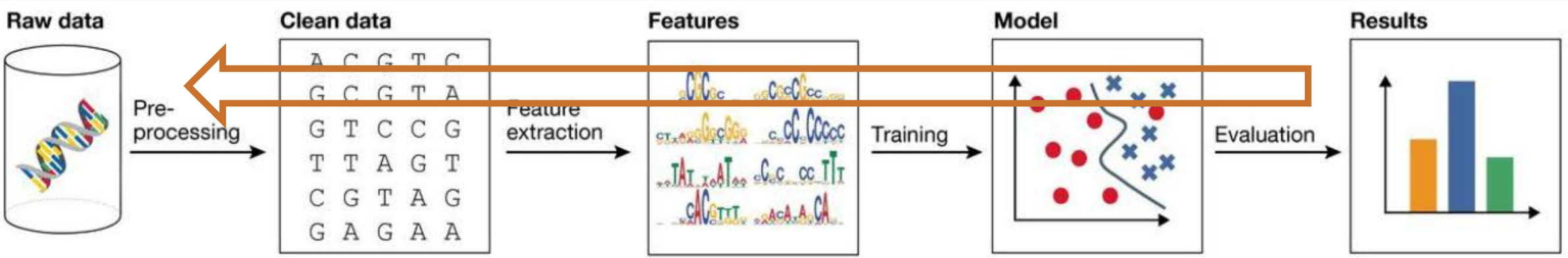
Machine Learning

Definition

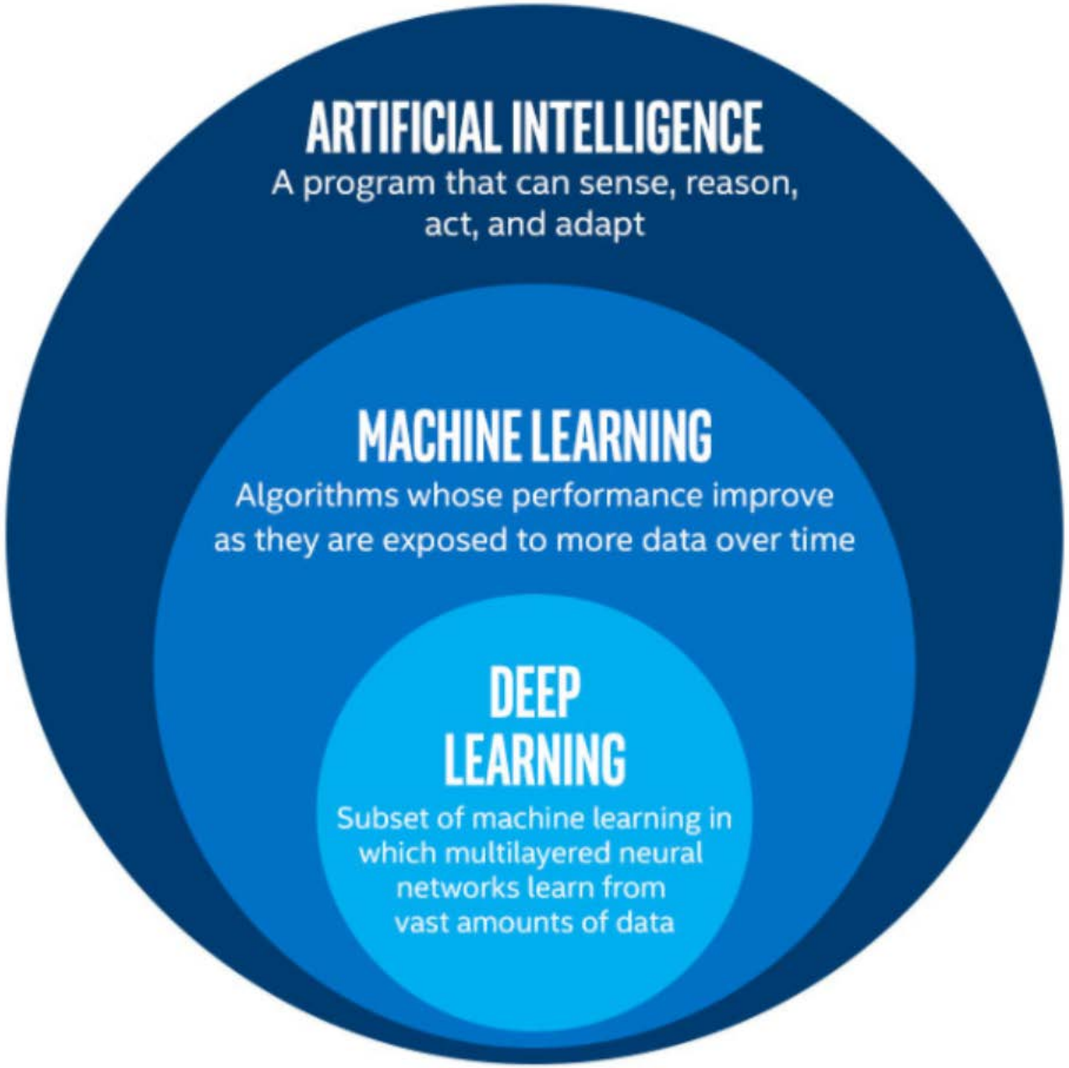
Machine learning (ML) is a field of artificial intelligence that uses statistical techniques to give computer systems the ability to "learn" from data, without being explicitly programmed. - Wikipedia

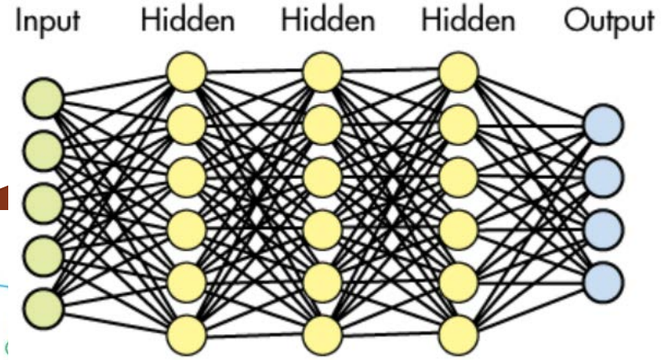
Challenges

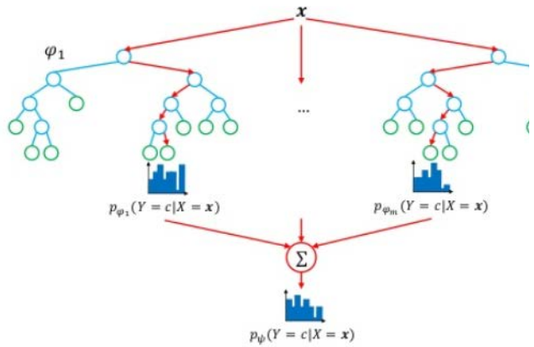
- Data is Discrete or continuous?
- Qualitative or Quantitative outcome?
- Low or High dimensionality in the Data?
- Non-linear relationships in the data?
- Complete data?
- Known dependencies in the data?
- Model interpretation needed?

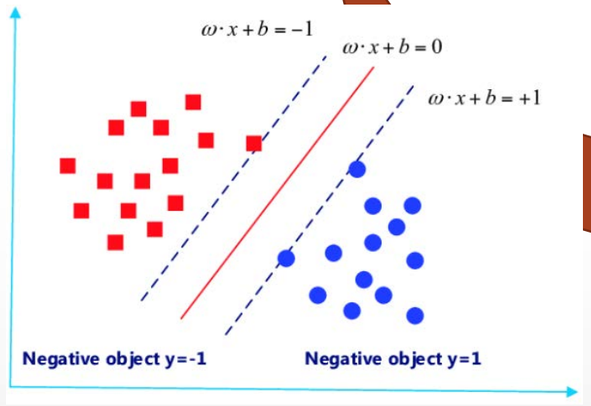


Machine Learning









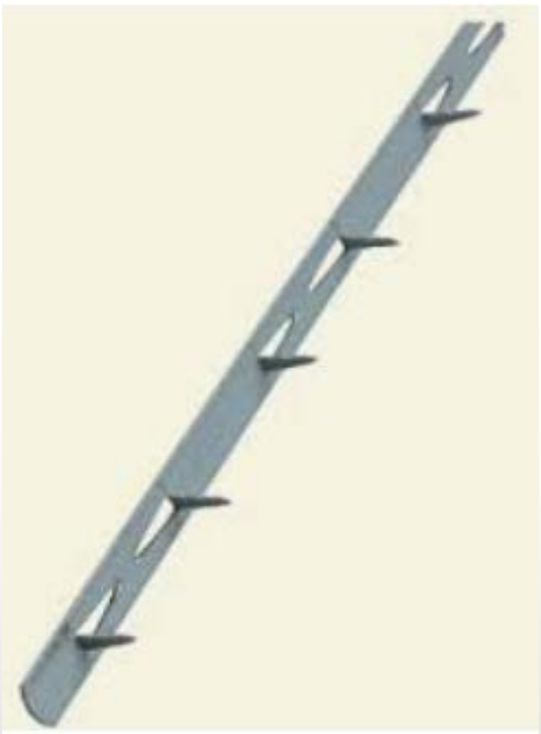
https://www.act-on.com/blog/3-things-about-machine-learning-every-marketer-needs-to-know/

Machine Learning for Big Data?

Same Methods

Improved algorithms in terms of code or processors

New type of method



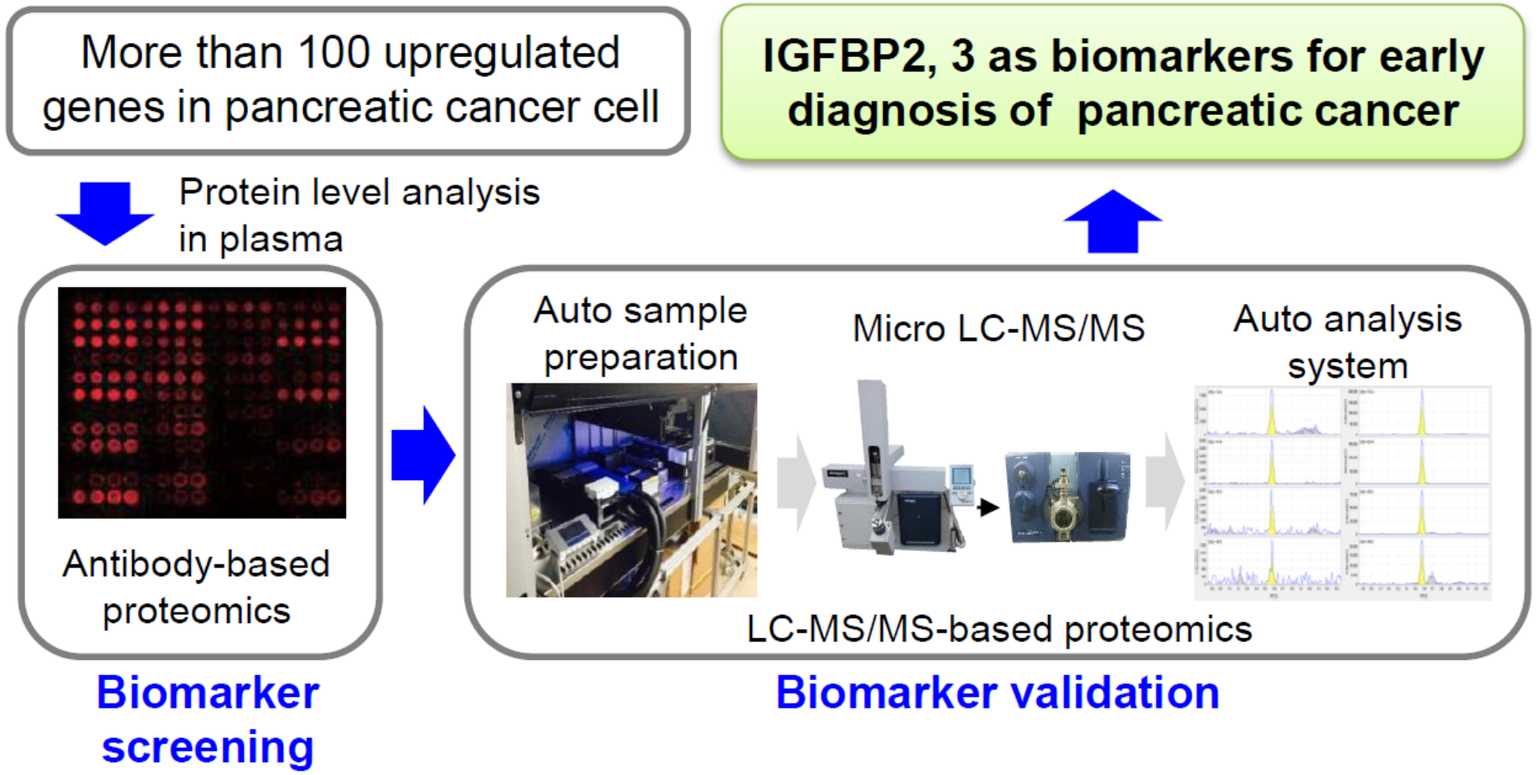
SVM

HPC SVM

CNN

What are Biomarkers?

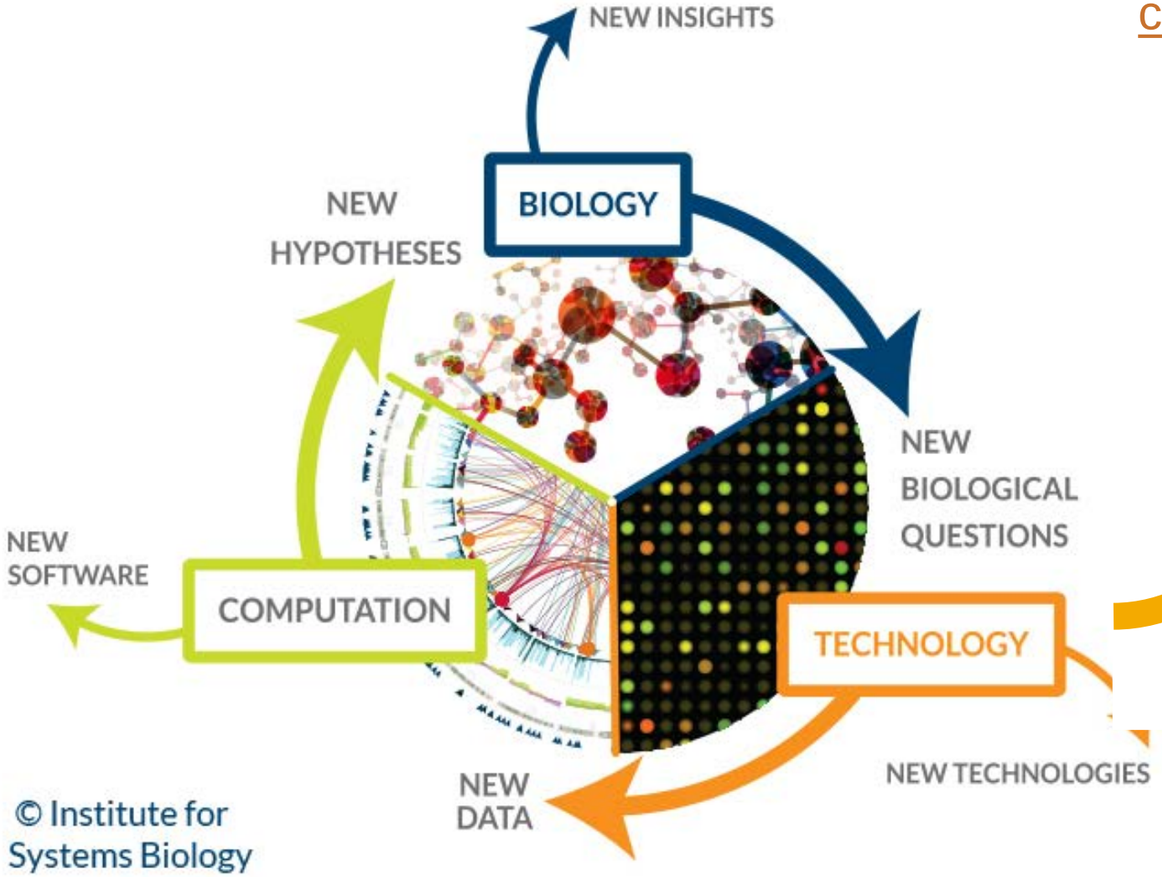
"The term *biomarker*, refers to a broad subcategory of medical signs – that is, objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly."



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161009

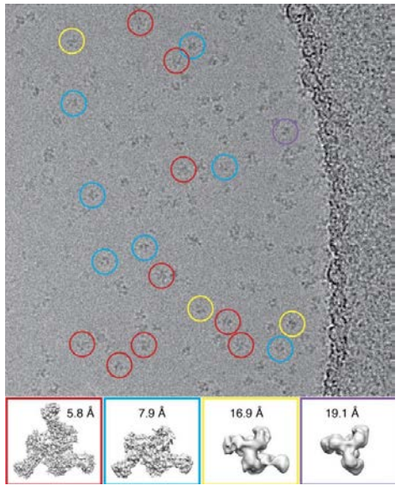
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3078627/pdf/nihms259967.pdf

Big Data Challenge for Biomarker Discovery - Data Scale and Collection Rates



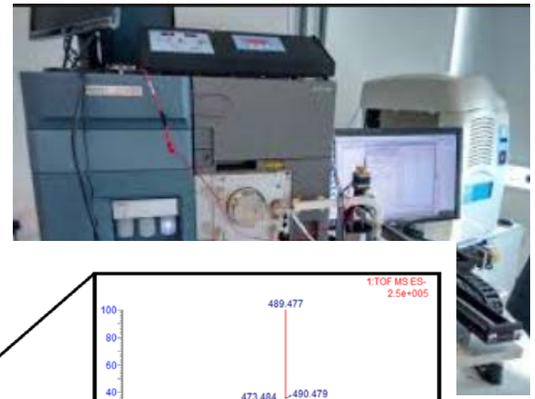
https://systemsbiology.org/about/what-is-systems-biology/

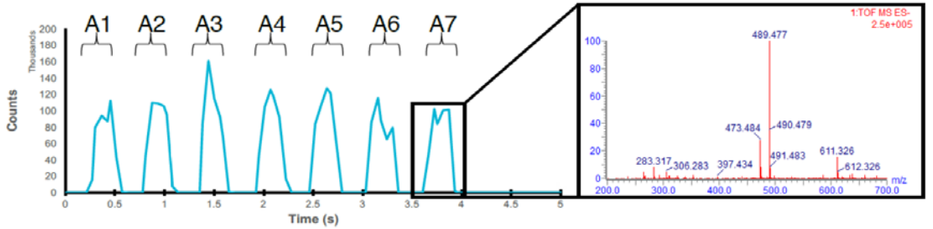
cyroEM is ~2000 movies per day



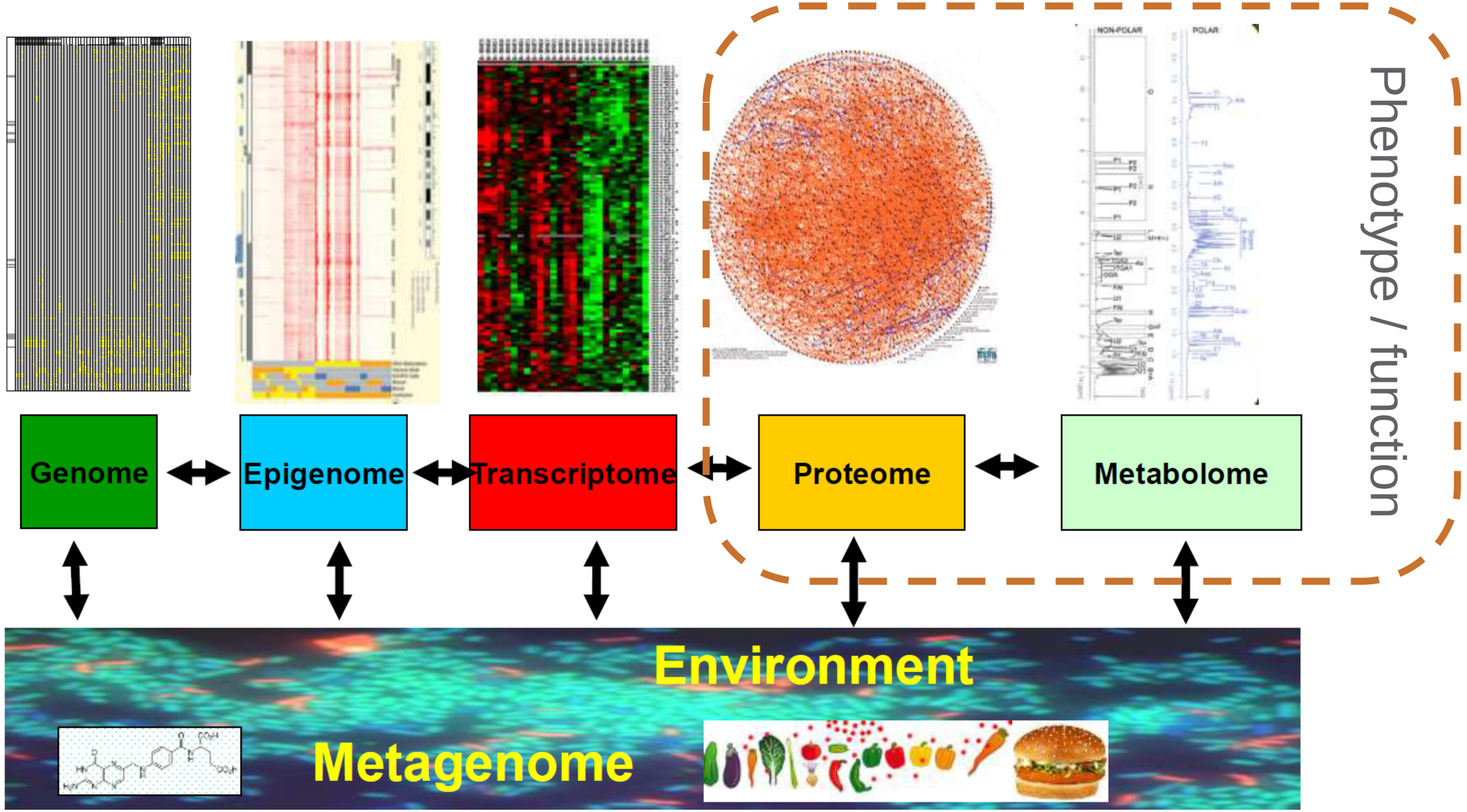
AMI Metabolomics is > 350 in 5 hours

https://www.labcyte.com/echo-technology/acoustic-mass-spec

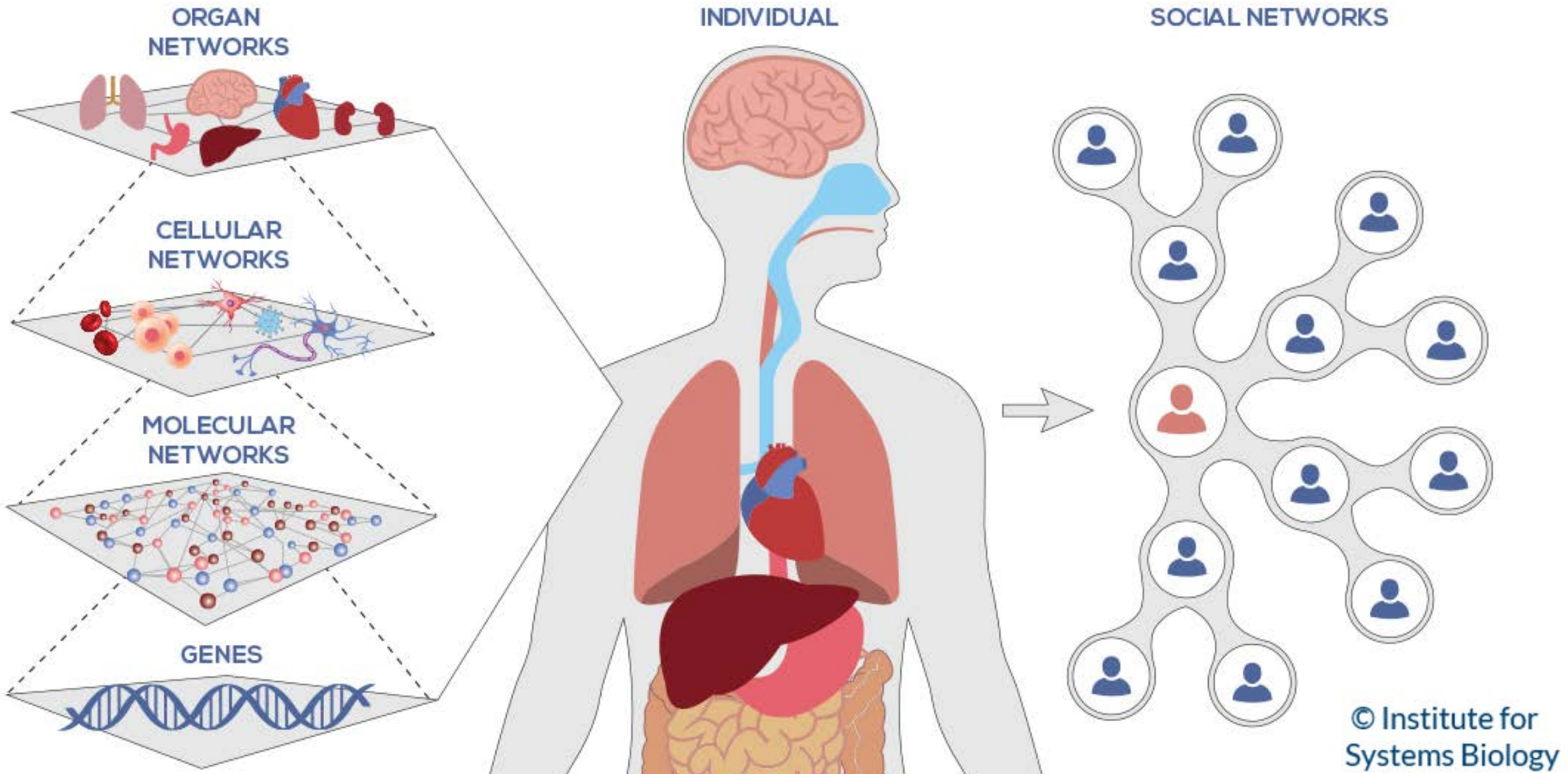




Big Data Challenge for Biomarker Discovery - Data Diversity and Complexity

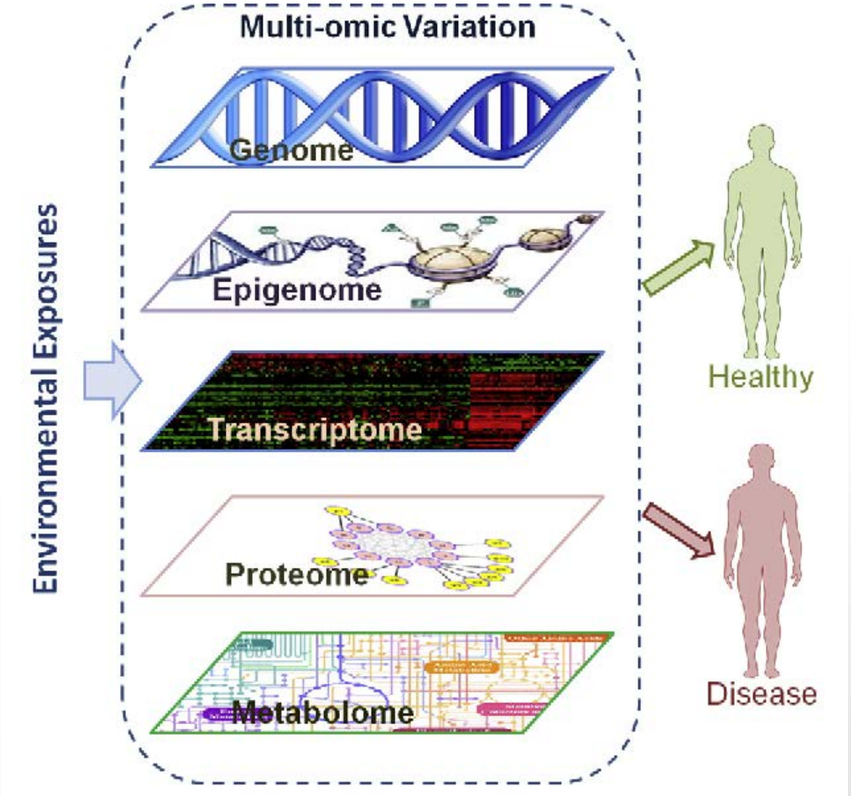


Big Data Challenge for Biomarker Discovery - Systems Complexity



Machine Learning Needs for Biomarker Discovery from Big Data - Interpretation

- Biologists need to understand the context that multiple markers work together to infer mechanism.

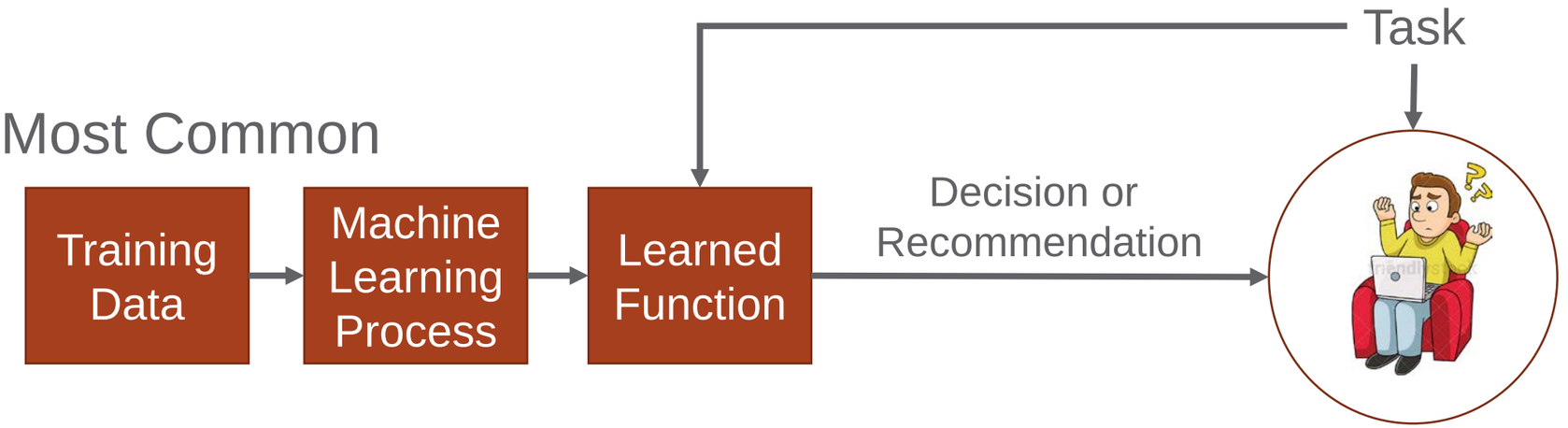


Sun and Hu, (2016) adv Gent

- The clinicians need to know which markers to build assays.

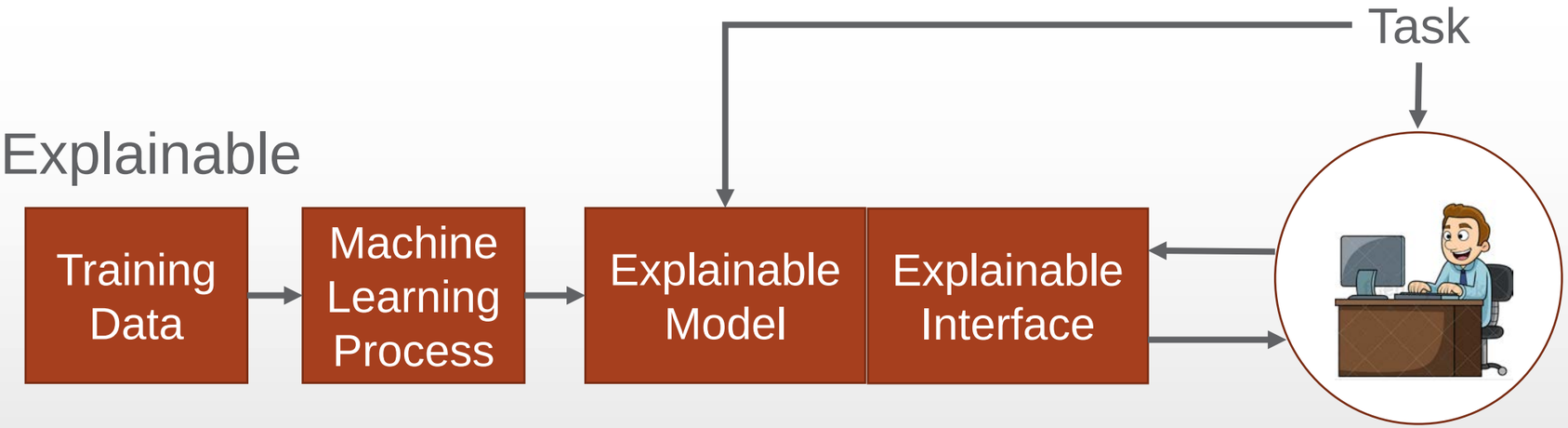
Machine Learning for Biomarker Discovery Needs Explainable Models

Most Common



- Why did you select that?
- What is your confidence in that decision?
- What causes you to fail?
- Can I correct an error?

Explainable



- I understand why you selected that.
- I understand why you did not select another.
- I understand where your successes and failures are.

https://www.darpa.mil/program/explainable-artificial-intelligence

Machine Learning Challenges – Opening the Black Box

PwC AI predictions for 2018

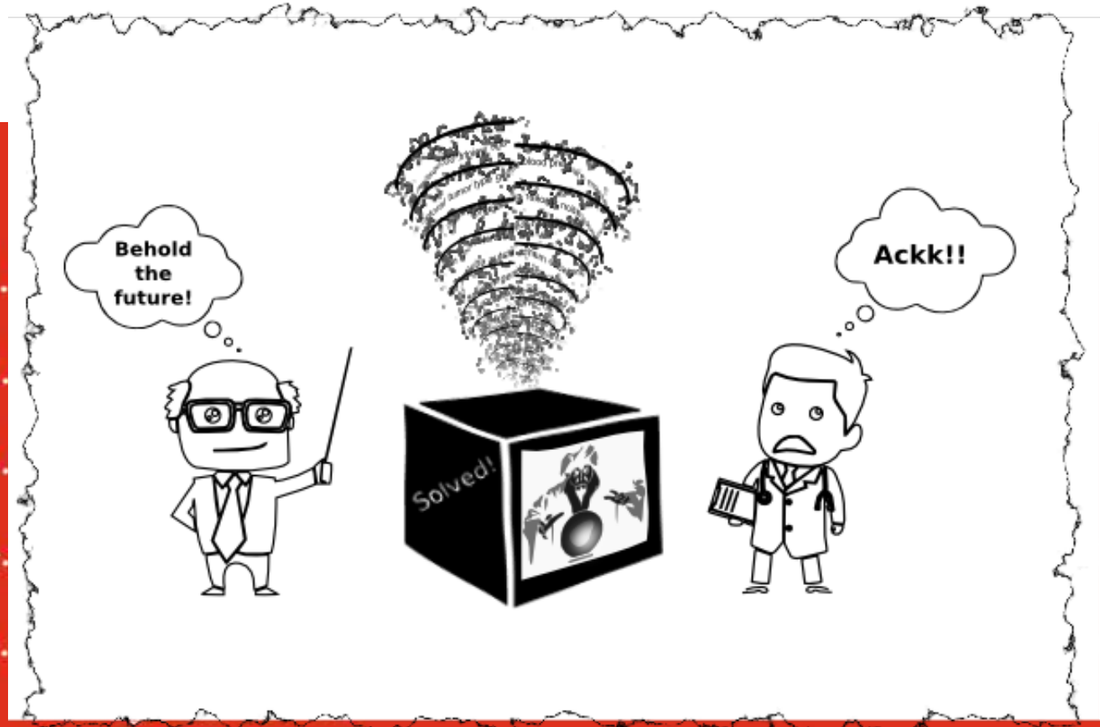
1. AI will impact employers before it impacts employment
2. AI will come down to earth—and get to work
3. AI will help answer the big question about data
4. Functional specialists, not techies, will decide the AI talent race
5. Cyberattacks will be more powerful because of AI—but so will cyberdefense
6. Opening AI's black box will become a priority
7. Nations will spar over AI
8. Pressure for responsible AI won't be on tech companies alone

2018 AI predictions
pwc.com/us/AI2018

2

https://www.pwc.com/us/en/advisory-services/assets/ai-predictions-2018-report.pdf

Machine Learning Challenges – Opening the Black Box



PwC AI predictions for 2018

1.
2.
3.
4.
5.

nt

talent race

o

6. Opening AI's black box will become a priority
7. Nations will spar over AI
8. Pressure for responsible AI won't be on tech companies alone

2018 AI predictions
pwc.com/us/AI2018

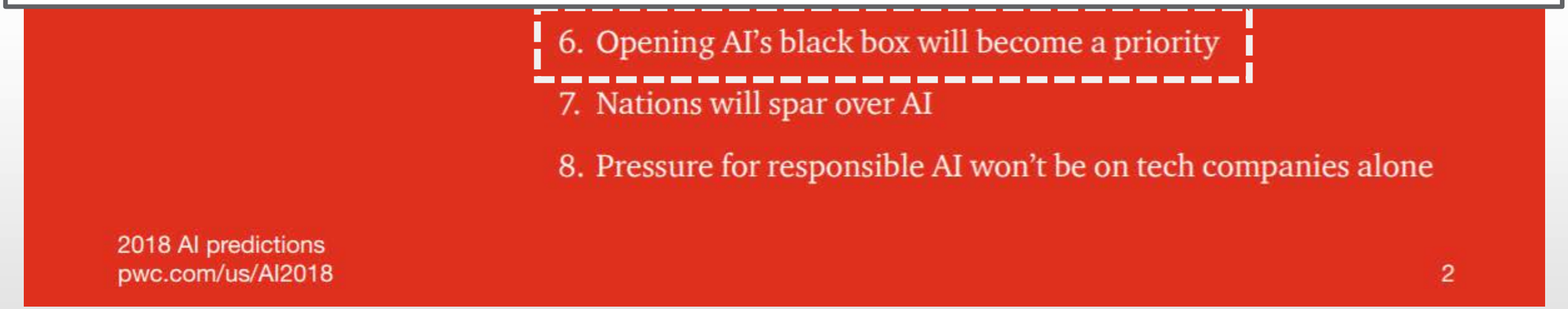
2

https://www.pwc.com/us/en/advisory-services/assets/ai-predictions-2018-report.pdf

Machine Learning Challenges – Opening the Black Box

As we let AI take over higher-risk tasks we will need to be able to answer why a decision is made or trust in the systems will be broken:

- **Why was my mortgage turned down?**
- **Why is this person being stopped at the airport?**
- **Why did the self-driving car move right?**
- **…**



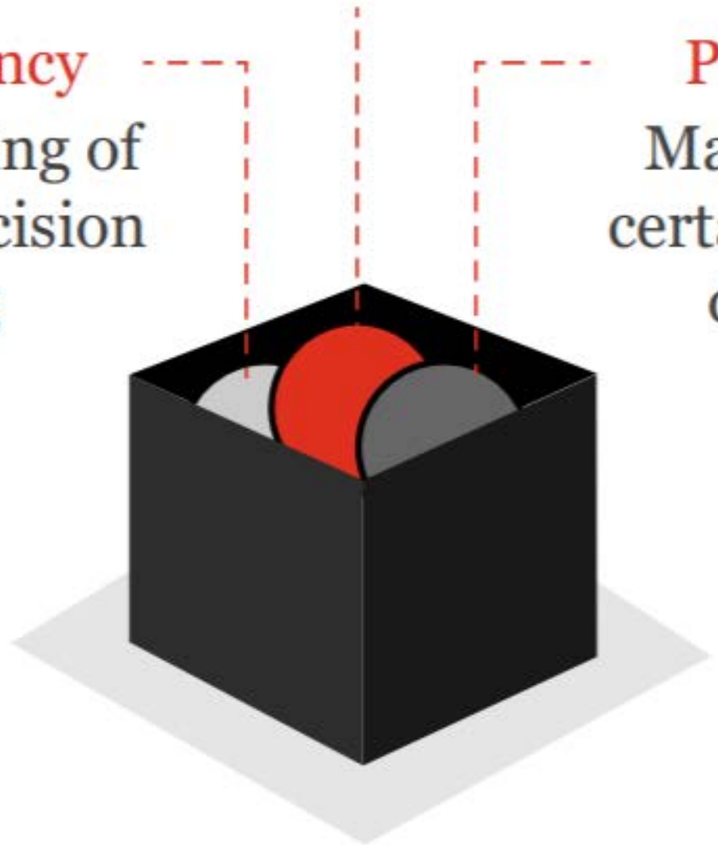
https://www.pwc.com/us/en/advisory-services/assets/ai-predictions-2018-report.pdf

What Does it Mean to Look Inside the Black Box?

Explainability
Understanding reasoning behind each decision

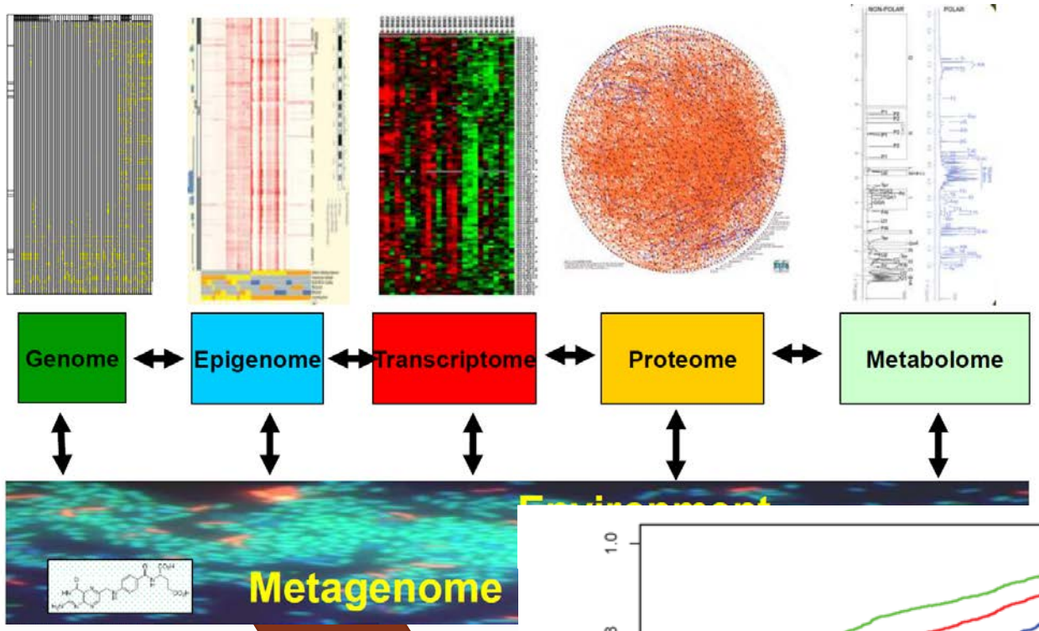
Transparency
Understanding of AI model decision making

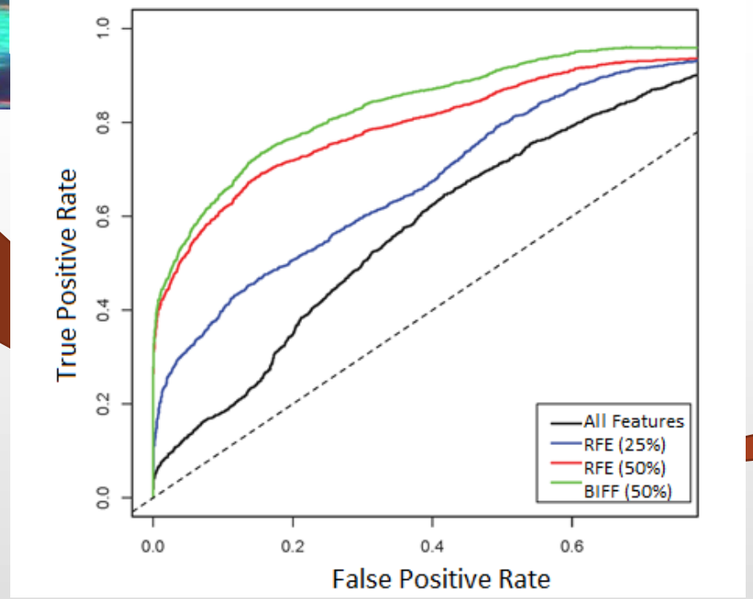
Provability
Mathematical certainty behind decisions



Source: PwC

What does interpretability Mean for Biomarkers?





Progression to T1D	
Feature	% Selected
Feature A	100
Feature B	100
Feature C	100
Feature D	99
Feature E	98

(http://www.daisycolorado.org/)
Example from my Work – DAISY

Clinical Research - DAISY

The Diabetes Auto Immunity Study in the Young

The **D**iabetes **A**uto **I**mmunity **S**tudy in the **Y**oung - started in July 1993 and has been continuously funded by the National Institutes of Health. The primary goal of DAISY is to learn how genes and the environment interact to cause childhood (type 1) diabetes. In order to do this, the study follows 2542 high-risk children with a diabetic relative (a sibling or parent) and children without a diabetic relative but found to have high genetic risk by screening of 30,000 Denver newborns. Infections, diets, genes and immunological markers are compared in children who have developed pancreatic inflammation and diabetes with those who remained healthy. Investigators led by Dr. Marian Rewers were able to map out the events leading to childhood diabetes. For instance, they developed immunological and genetic tests that can identify children who will develop diabetes in the next 5-10 years; they demonstrated that routine immunizations and baby milk formulas based on cow's milk do not increase th risk for diabetes; that omega free fatty acids may be protective, but certain viral infections increase the risk. On the foundations of DAISY, the National Institutes of Health funded an international consortium – The Environmental Determinants of Diabetes in the Young (TEDDY). TEDDY has screened 424,000 children in Europe and America and is following 8766 those at the highest risk. DAISY and TEDDY are likely to deliver definitive answers concerning the cause and prevention of childhood diabetes.

20th Anniversary DAISY

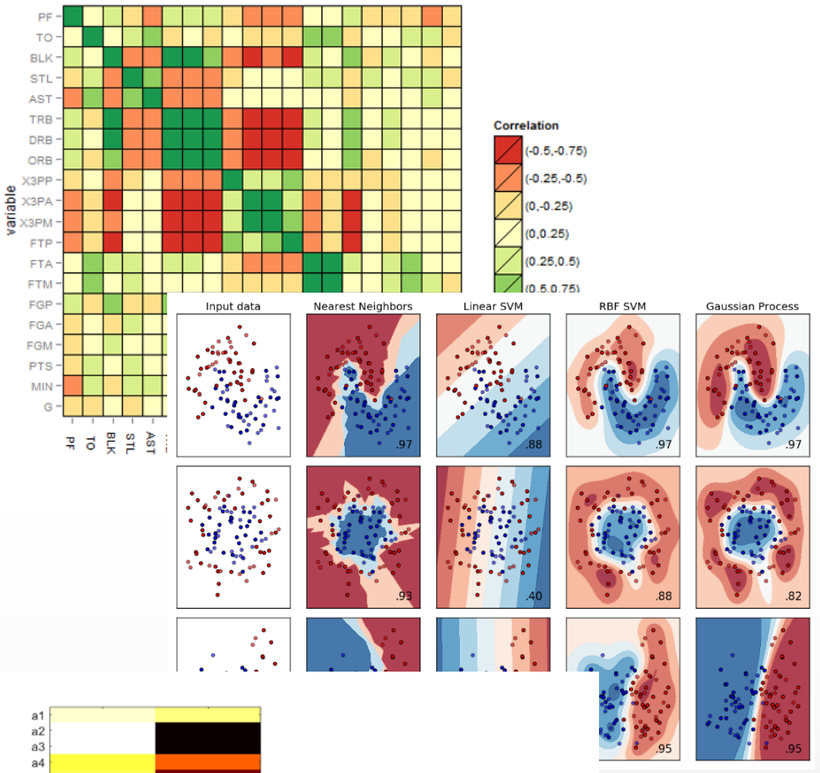
… follows 2542 high-risk children with a diabetic relative

Infections, diets, genes and immunological markers are compared in children who have developed pancreatic inflammation and diabetes with those who remained healthy.

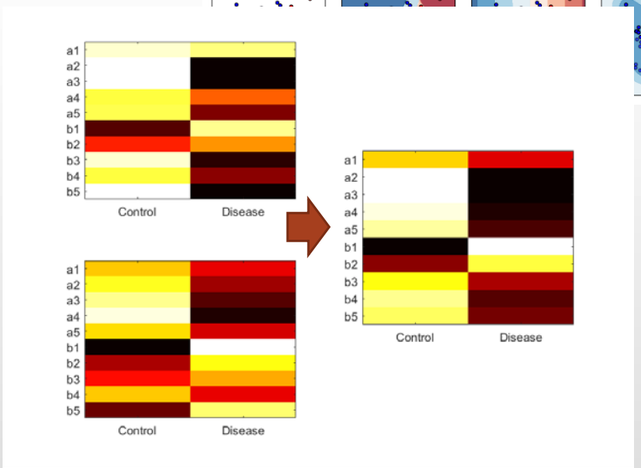
TEDDY has screened 424,000 children in Europe and America and is following 8766 those at highest risk.

Current Research in Integrative and Interpretable Machine Learning

Goal: Focus on identifying features that work in combination across multiple omics and meta-data that can predict a disease versus control state



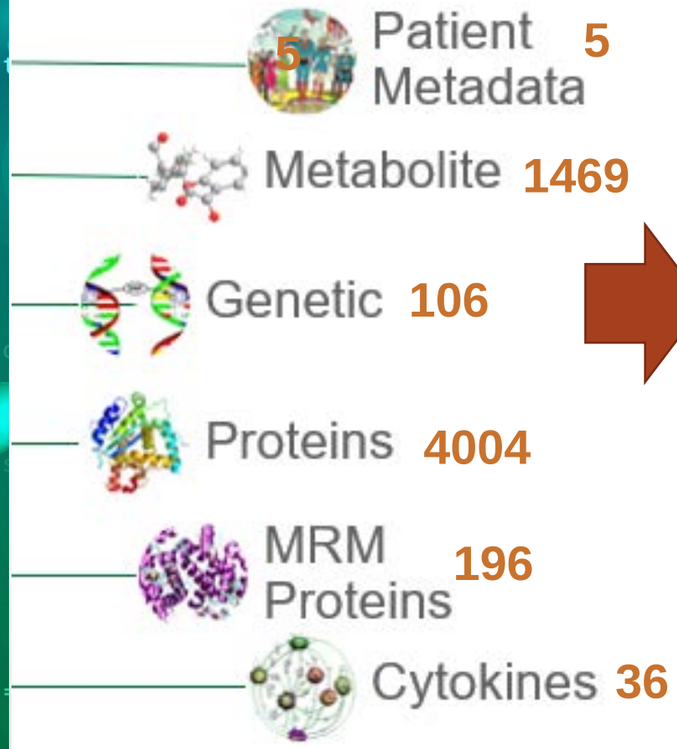
Approach: Integrative machine learning in combination with feature selection that models uncertainty in the solution



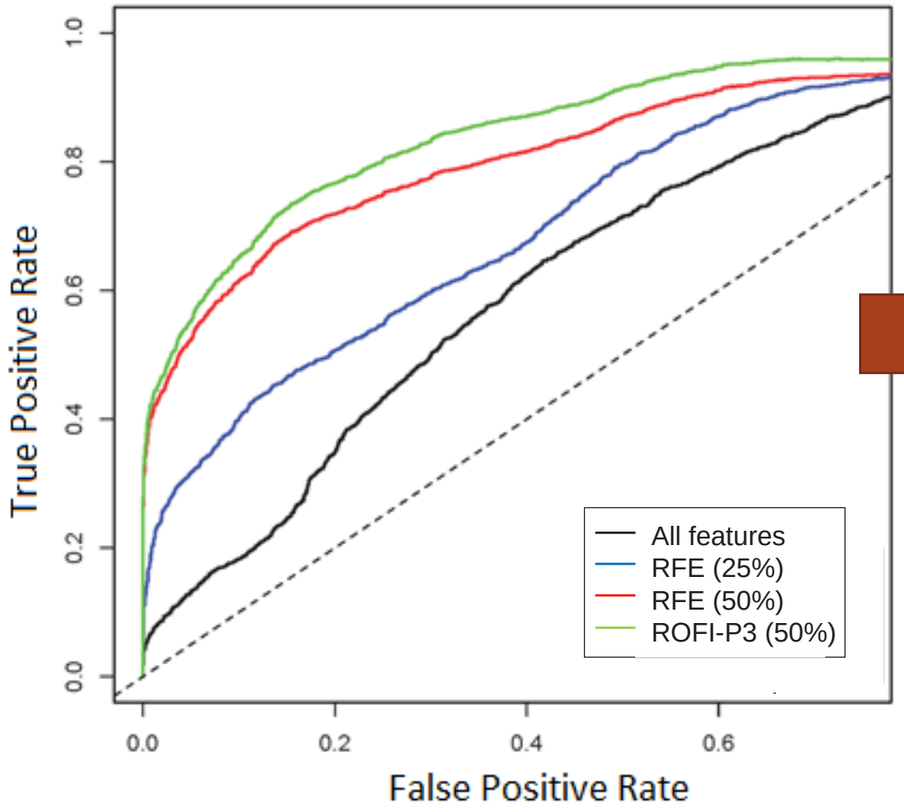
Webb-Robertson et al., 2009 Pac Symp Biocomput
Beagley et al., 2010, Bioinformatics
Webb-Robertson et al., 2012 J Biomed Biotechnol
Webb-Robertson et al., 2017 CSCI

Example – DAISY

Hypothesis: *Biomarker panels can differentiate the control group from the diabetic endpoints prior to clinical symptoms.*



5,816 potential markers



Good Performance

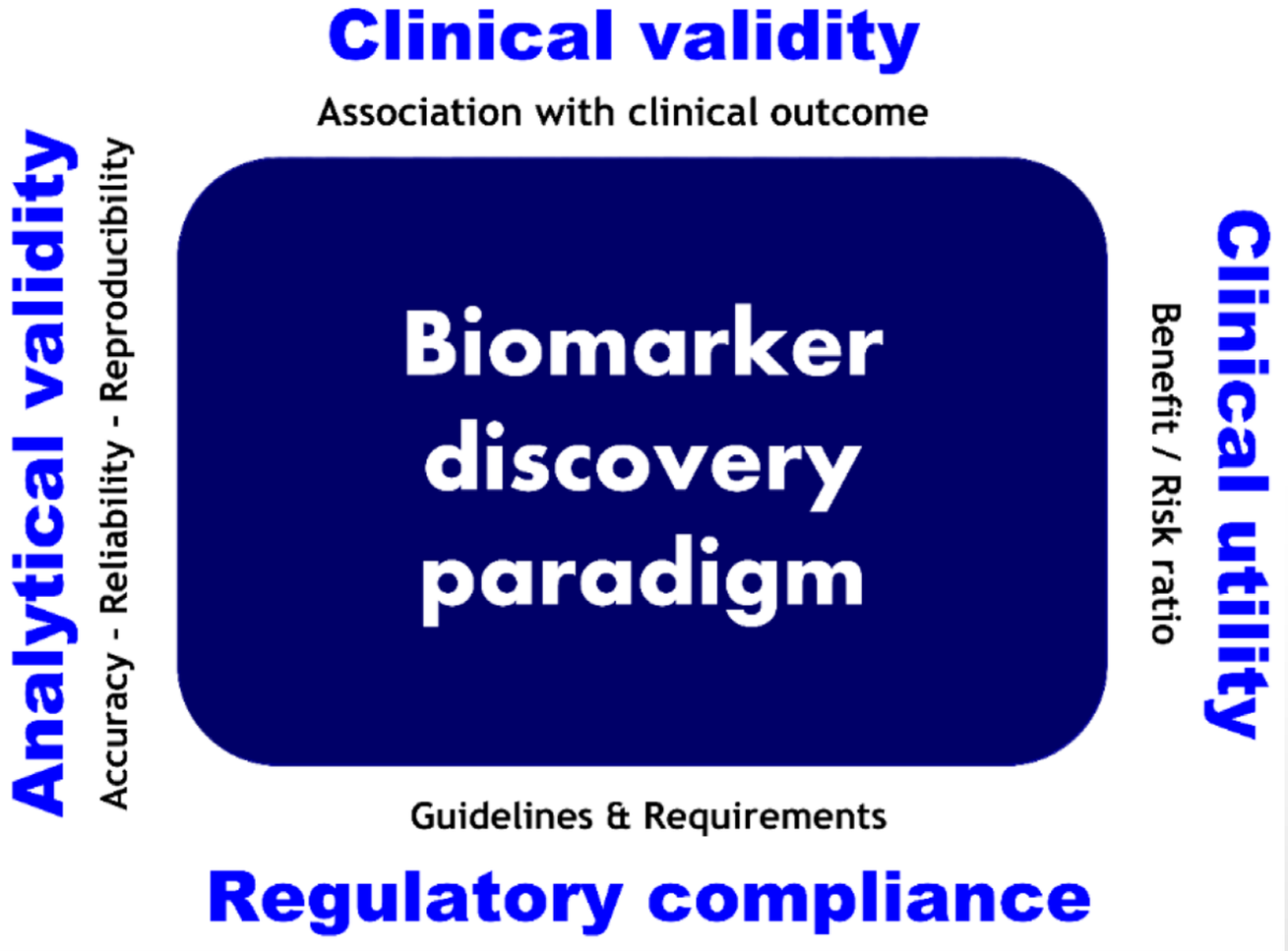
Progression to T1D	
Feature	% Selected
Feature A	100
Feature B	100
Feature C	100
Feature D	99
Feature E	98
Feature F	98
Feature G	97
Feature H	92
Feature I	91
Feature J	90
Feature K	89
Feature L	89

Clinical Markers

Machine Learning Biomarker Discovery

Biomarker Discovery is a complex task and machine learning plays one small role

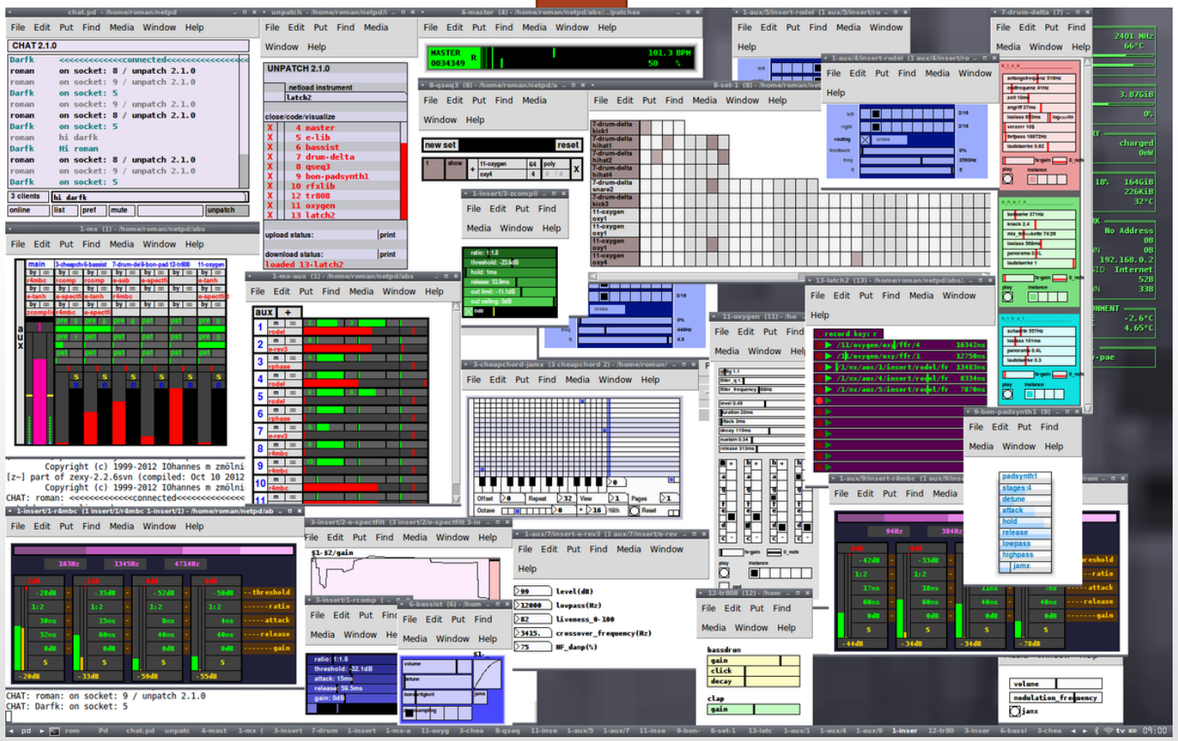
- Silver-bullets are unlikely and thus integration is becoming more important
- Understanding uncertainty is a necessity – validation is expensive



https://www.mdpi.com/1422-0067/17/9/1555/htm

Machine Learning Biomarker Discovery – Holds tremendous promise for Healthcare

Focus on how machine learning can improve or speed up the translation to clinic.



https://theanalyticalscientist.com/issues/1216/biomarkers-sweat-and-tears/

Acknowledgements

Core Team

PNNL: Applied Statistics & Computational Modeling	• Lisa Bramer • Sarah Reehl • Bryan Stanfill
PNNL: Integrative Omics	• Tom Metz • Charles Ansong • Ernesto Nakasuya
Barbara Davis Center for Diabetes	• Marian Rewers (Lead PI) • Brigitte Frohnert
University of North Carolina	• Qibin Zhang

Funding Agencies

- National Institutes of Health (NIH)
- Juvenile Diabetes Research Fund (JDRF)
- Helmsley Trust
- Centers for Disease Control and Prevention (CDC)

- Support in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida and the University of Colorado

Thank you

bj@pnnl.gov