SC13
Date
Time

Dan Crichton. M.S., NASA Jet Propulsion Laboratory

Focus Questions

- What is your lab doing in big data and AI?
- What do you see as the future of big data and AI in cancer?
- Where do you see science opportunities for big data and cancer biomarkers within EDRN?
- Are there projects that you would want to see EDRN pursue?

Panel Speakers

1. Zhen Zhang, Ph.D., Johns Hopkins University School of Medicine
2. Ziding Feng, Ph.D., Fred Hutchinson Cancer Research Center
3. Steven Skates, Ph.D., Massachusetts General Hospital
4. Matthew Schabath, Ph.D., H. Lee Moffitt Cancer Center and Research Institute, Inc.
5. Paul Boutros, Ph.D., University of California, Los Angeles
6. Dan Crichton M.S., NASA Jet Propulsion Laboratory

Where are the "Big Data" for biomarker research?

Zhen Zhang, Johns Hopkins Medicine

- Huge amount of data (e.g. genomic, proteomic data) ≠ big data.
- The role of transfer learning (we have a lot of unlabeled data)
- Integration of knowledge to reduce search "space" in learning.
- Methodology research in "small sample learning", e.g., tradeoff between fast convergence and asymptotic bias.
- Tagging quantitative information to facilitate extraction and evaluation of results from biomarker research literature.

Where do you see science opportunities for big data and cancer biomarkers within EDRN?

- Imaging combined with biomarkers
  - Imaging improves sensitivity, radiomics and biomarkers to reduce over diagnosis
- Use large EHR data
  - Use longitudinal data to improve detection (cirrhosis/NALFA patients)
  - Use readily available EHR data to enrich risk population (NOD)
- Integration of genomics, proteomics, and clinical data
  - Strengthen biomarker selection process
- Mobile device
  - Real time monitoring biomarkers and behaviors that provide clue for cancer

Z. Feng

Big Data + Basic Biology + Low CV Proteomics Biomarker Discovery and Validation for Multiple Cancers in Pre-diagnostic Longitudinal Samples

Leverage unique feature of cancer

- doubling time – uncontrolled cell division – exponential growth
- Other diseases have spikes, level shifts, Dx acute phase reactants but exponential growth over mnths/yrs is **unique to cancer**

Leverage individual plasma proteome phenotype

- each individual has their own set of plasma protein baselines
- Between patient variability >> within patient variability

Proximity Extension Assay PEA – power of NGS/PCR applied to proteins

- ~1,500 assays
- CVs 6 – 12%

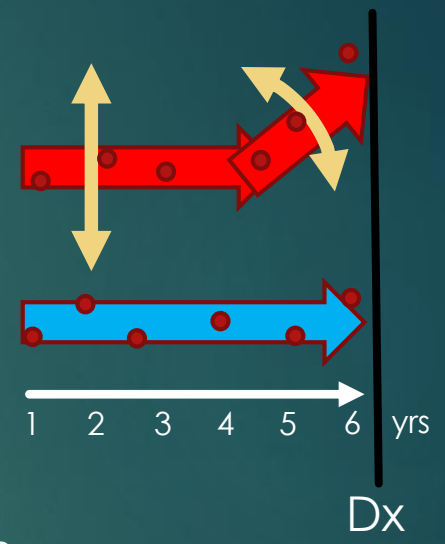
Preliminary Results: Ovarian Cancer: 25 markers (CA125, HE4, $FOLR_{\alpha}$, …)

S. Skates

Big Data + Basic Biology + Low CV Proteomics

Prediagnostic Longitudinal Plasmas (6 samples – 6 yrs)

- 20 cancers
- Assay ~1,500 proteins 100 cases + 100 controls
- Longitudinal change-point model: ~1,500 proteins
- Identify plasma proteins which:
  - rise exponentially from baseline in some cases
  - flat profile in most controls (high spec)



Big Data:

200 patients (cases+controls) x 6 samples/pt x 1500 proteins x 20 cancers

- 36 M data points

Aim:

- Identify earliest rising biomarkers for each cancer
- Panel of 5-10 biomarkers unique to each cancer
- Multi-cancer plasma 100-200 marker longitudinal algorithm
- Validation in large patient nested case/cntl with custom panel

S. Skates

Biomarkers and Big Data in Quantitative Imaging

1a. Applications in nodule malignancy discrimination

- Conventional machine learning (radiomics), deeply learned models, convolutional neural networks (for segmentation and classification), hybrid models and ensembles (for classification), and habitat imaging (clustering subregions)
- Classification models perform as well as (or better) Google's end-to-end results *with a fraction of the size**

1b. Advances in radiomic pipelines/analysis

- Deeply learned models and CNNs to predict radiologist-defined features: AUCs and accuracy: 0.82 to 0.84
- When scaling CT images in CNNs, nodule size is implicitly encoded into texture information (i.e., size/volume features are likely redundant in CNNs)

2. Future of AI in quantitative imaging

- Defining biological basis of image features
- Open access to [large] well annotated images and data
- Scalable end-to-end DL solutions from point-to-care to image processing to algorithm development/delivery
- Distributed learning

3/4. Opportunities within the EDRN

- Prospective observational and intervention trials to determine clinical utility and decision support systems for radiomic models: *end-to-end solutions and distributed learning could facilitate*
- Targeted biopsies for mapping image features to biology

**The "myth" of needing enormous datasets: you need a dataset large enough to train/tune, test, and validate*

Big Data Opportunities for the EDRN

Benchmarking

- Many biomarker discovery efforts underway simultaneously across EDRN
- A general understanding of best-practices for discovery data science exists across BDLs
- Consortium-wide, many large standard datasets exist that can be leveraged.

Proposal: regular consortium-wide, crowd-sourced quantitative benchmarking of biomarker discovery methods.

Centralized Bioinformatics

- EDRN does a lot of technology development
- But it also increasingly exploits standard methodologies like DNA- and RNA-sequencing
- Harmonized best-practice analysis would make all projects better and move faster.
- And centralizing would facilitate standardized data availability across the consortium.

Proposal: define centralized cloud-based bioinformatics pipelines to harmonize and accelerate consortium-wide analyses.

UCLA Health
Institute for Precision Health

P. Boutros

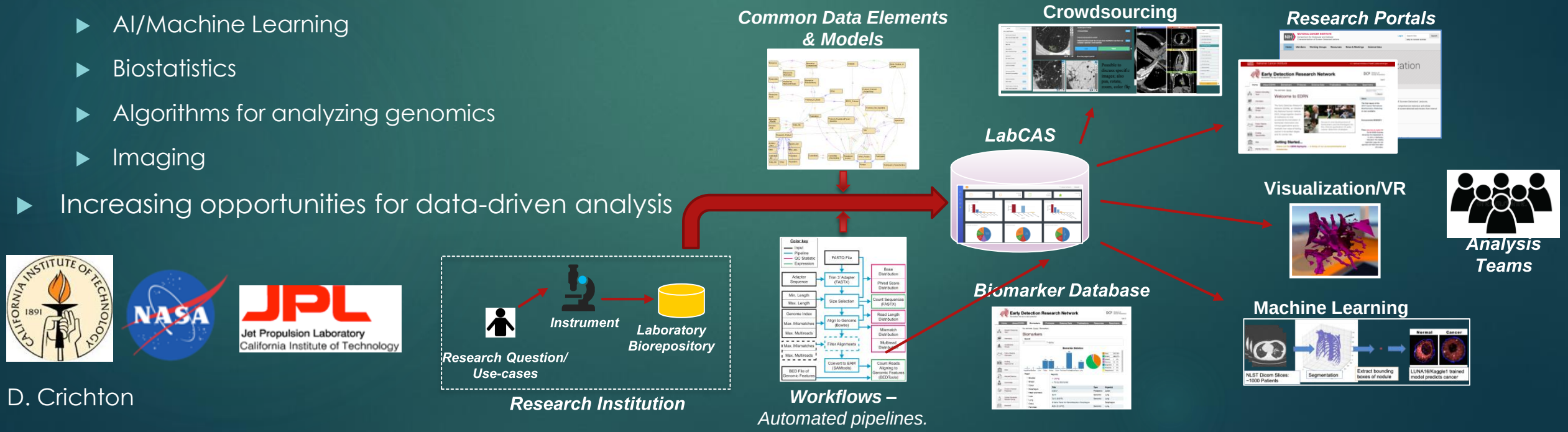
UCLA Jonsson Comprehensive Cancer Center

UNIVERSITY OF TORONTO

EDRN Knowledge Environment: A National Infrastructure

Enabling Data-Driven Research for Cancer Biomarkers

- Leverage cloud computing-based capabilities to support the capture, sharing, and analysis of data for EDRN
- Integrate within the EDRN portal, provides secure, multi-layer access to data for EDRN, NCI, research, and public communities, that link biomarkers, studies, and data together
- Provide biomarker common data elements to enable common annotation and linking across datasets
- Annotations of cancer biomarkers linked to terabytes of imaging, genomics data and analysis results, in the EDRN data commons, known as "LabCAS".
- Enabling analytical biomarker data pipeline and analysis supporting:
  - AI/Machine Learning
  - Biostatistics
  - Algorithms for analyzing genomics
  - Imaging
- Increasing opportunities for data-driven analysis



D. Crichton

Panel Discussion

- Zhen Zhang, Ph.D., Johns Hopkins University School of Medicine
- Ziding Feng, Ph.D., Fred Hutchinson Cancer Research Center
- Steven Skates, Ph.D., Massachusetts General Hospital
- Matthew Schabath, Ph.D., H. Lee Moffitt Cancer Center and Research Institute, Inc.
- Paul Boutros, Ph.D., University of California, Los Angeles
- Dan Crichton M.S., NASA Jet Propulsion Laboratory