Towards a Digital Pathology Commons
Frontiers of Predictive Oncology and Computing, October 18th 2017

Michael J. Becich, MD PhD, FACMI

Chair, Department of Biomedical Informatics
University of Pittsburgh School of Medicine

University Distinguished Professor,
Associate Vice Chancellor for Informatics
Associate Director, UPCI and CTSI
Co-Director, Center for Commercial Applications of Healthcare Data of the Pittsburgh Healthcare Data Alliance
NCI Board of Scientific Advisors

UPMC LIFECHANGINGMEDICINE

DBMI

ACMI
INFORMATICS FELLOWS. LEADING WITH DISTINCTION.

Pittsburgh Health Data Alliance
Carnegie Mellon University   University of Pittsburgh   UPMC

Center for Commercial Applications of Healthcare Data
University of Pittsburgh

# Becich Conflicts of Interest (Disclaimer)

- **Spatial Diagnostics, Inc. or SPDx** (founder and stock) – computational pathology company
- **Nexi** – Newco by Rebecca Jacobson and TIES/TCRN team (royalties to my Department)
- **Cancer Center Consultancies and EABs – Baylor, University of Colorado, University of Michigan and Wake Forest**
- **CTSA Consultancies and EABs – numerous (not a conflict for this talk except possibly for U of Chicago Institute for Translational Medicine)**

Disclaimer – I am a member of NCI's Board of Scientific Advisors

Outline

- Introduction to Whole Slide Imaging and Computational Pathology

- Introduction to Digital Pathology Commons

- Data Commons - Precision Medicine, Data Science and Human Cell Atlas (CZI) programs are drivers

- Data to Knowledge to Wisdom – Machine Learning and Causal Modeling & Discovery in Comp Path

- **Key Impact Areas** – Predictive Analytics for better health and discovery science thru Comp Path!

Traditional vs. Digital Pathology Whole Slide Imaging (WSI)
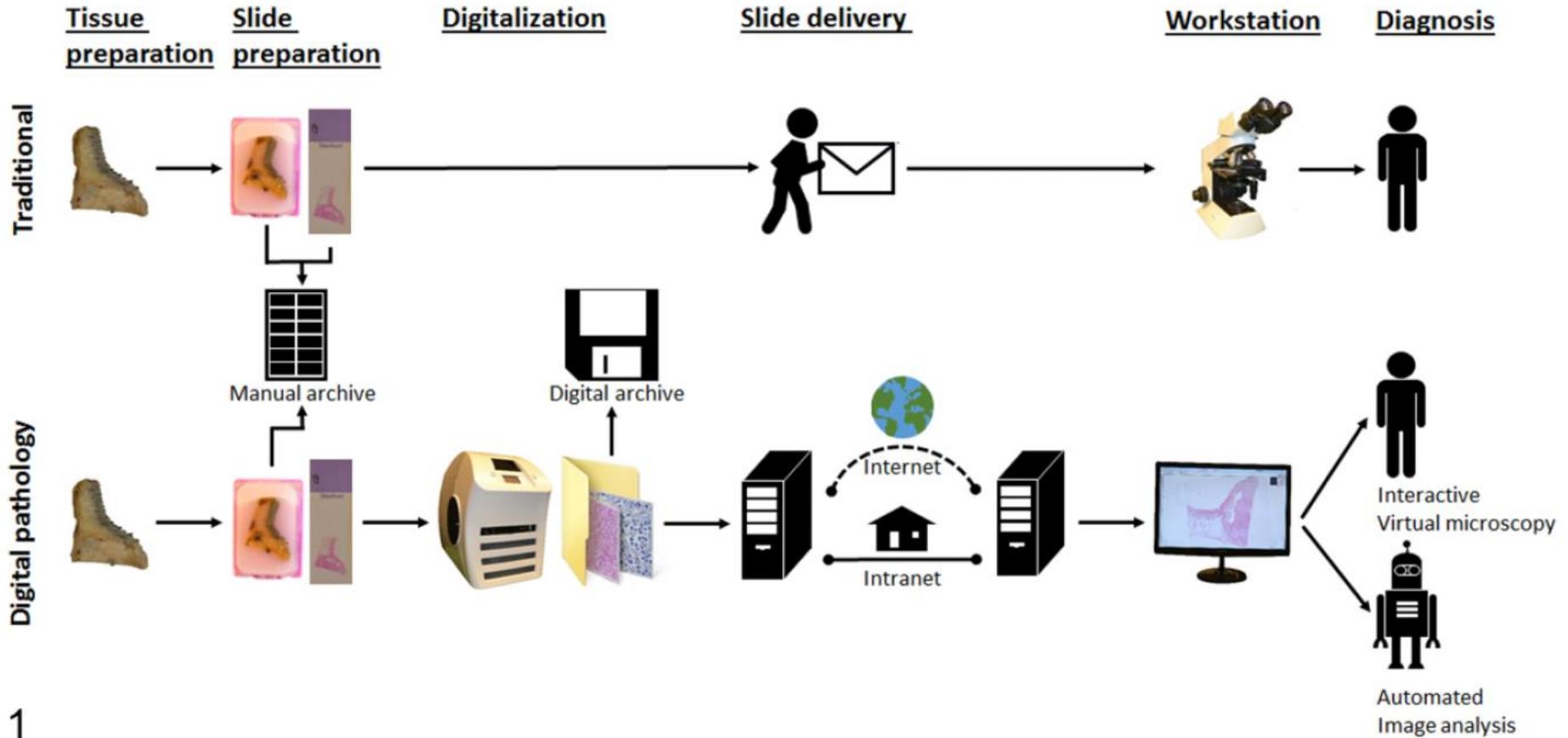
From Bertram and Kopfleish, Vet Path 2017

Introduction to Whole Slide Imaging

Pipeline for Whole Slide Feature Characterization

- 10^{10} pixels for each whole slide image: $10^5 \times 10^5$

- 10 whole slide images per patient

- 10^8 image features per whole slide image

- 10^{15} pixels

- 10^{13} features

From Saltz, circa 2011, to NLM

CAMELYON17

Bai du 百度

1000 scanned slides

3.03 Tbyte

444 registrants

FDA News Release

FDA allows marketing of first whole slide imaging system for digital pathology

f SHARE   TWEET   in LINKEDIN   PIN IT   EMAIL   PRINT

For Immediate Release   April 12, 2017

Histopatholomics and Computational Pathology

B Genomics · Epigenomics · Proteomics · Metabolomics

A

C

D ④ ③ ② ①
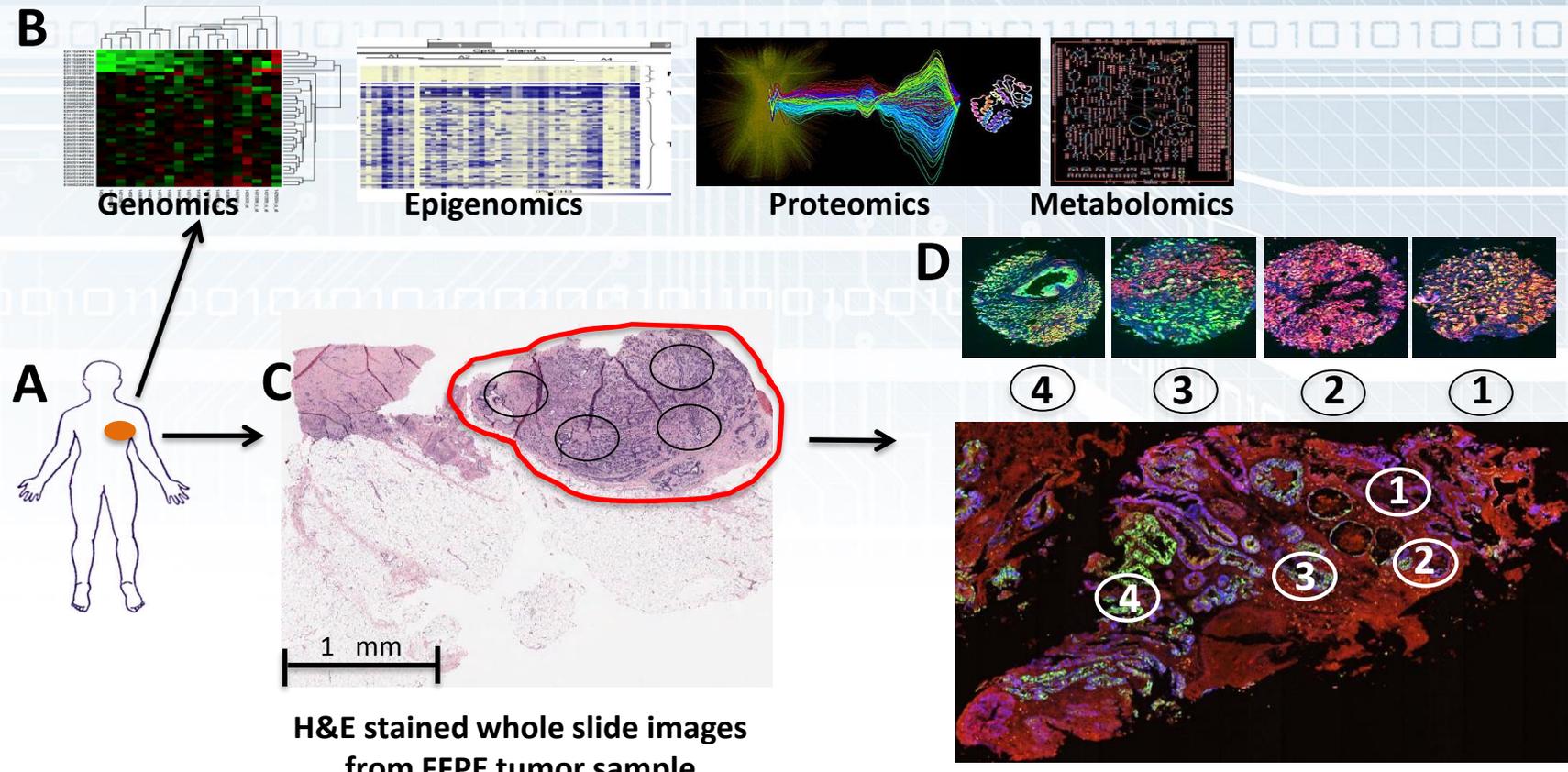
H&E stained whole slide images from FFPE tumor sample

Multi to hyperplexed fluorescence imaging of whole slide images for higher spatial resolution and tissue context

Thanks to Drs. Joe Ayoob and Chakra Chennubhotla for this illustration, 2017

Unique opportunities in WSI and Computational Pathology

- Computational pathology
  - Hyperplex Immunofluorescence (9 to > 50 Ab)
  - Machine learning + spatial statistics
  - Network Systems biology
  - Diagnostics/prognostics

- Iterative experimental–computational human tumor micro-environment studies
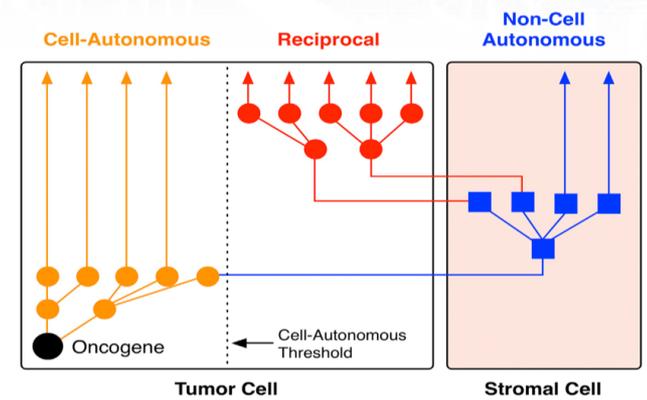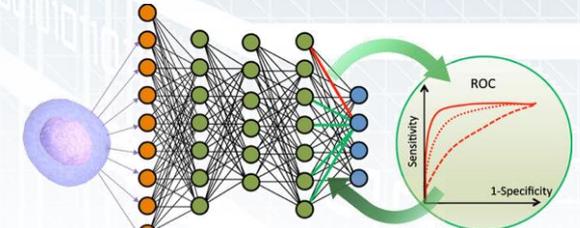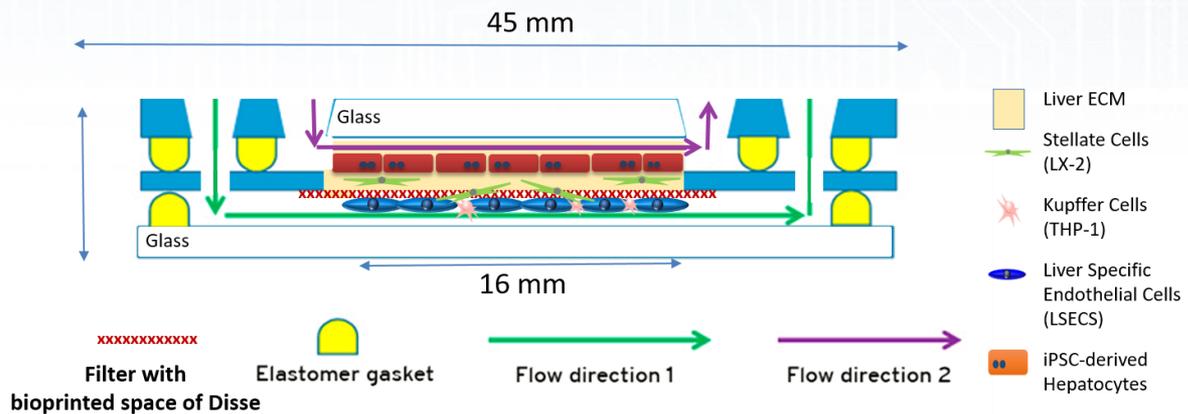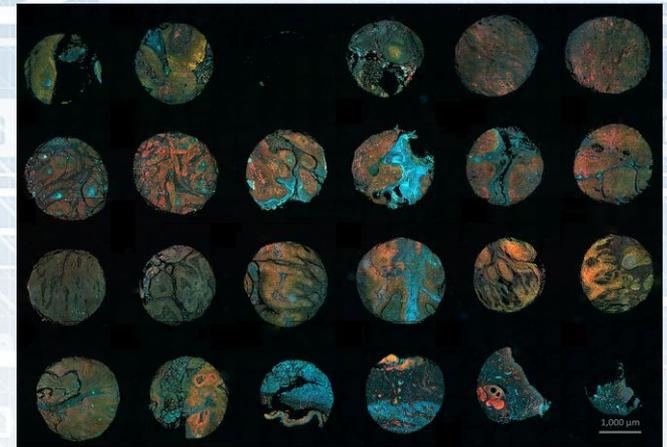
Courtesy Chakra Chennubhotla, 2017

A Case for Data Commons:
Toward Data Science as a Service

From Computing In Science & Engineering IEEE 2016

Robert L. Grossman, Allison Heath, Mark Murphy, and Maria Patterson | University of Chicago
Walt Wells | Center for Computational Science Research

"...Data commons collocate data, storage, and computing infrastructure with core services and commonly used tools and applications for managing, analyzing, and sharing data to create an interoperable resource for the research community..."

FAIR Data Principles

- Findable
- Accessible
- Interoperable
- Reusable

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*[#]

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

- Data Use Case for Data Commons
  - See Data Med v3.0 – http://datamed.org

Six Requirements of a Data Commons

- Permanent digital IDs (data and knowledge)
- Permanent metadata (data describing data)
- API (interface)-based access (interoperability)
- Data portability (standard containers)
- Data Peering (commons 1 can access commons 2)
- Pay for compute (allocate computing/charging)
  - Demand higher than computing resources available

From Grossman et al 2016 CISE IEEE

Towards Computational Pathology Imaging Commons for Cancer – Data Types

- Imaging – Rich source of computable information
  - Need to De-ID WSI and then "deeply" annotate
- Phenotype – From Anatomic Pathology Lab Info Sys
  - Structured data from synoptic reports
  - Unstructured data via NLP (Text Info Extract Sys – TIES)
- Computational Pathology Annotation – more later
- Outcomes Data – From Cancer Registry Systems
- Data to Integrate – Biobanks, Clinical Pathology & Molecular Pathology

The Text Information Extraction System or TIES?
NCI ITCR funded effort

- An **NLP and Information Retrieval system** for de-identifying, annotating, storing and retrieving pathology (and radiology) reports

- A system for **indexing research resources** (clinical data, biospecimens & images) with document annotations

- An **GUI for querying** large repository of annotated documents and obtaining resources locally, using an honest broker model

- A platform to support **phenotype, images and biospecimen sharing** among networks of cancer centers and other institutions

http://ties.dbmi.pitt.edu/

TIES Cancer Research Network (TCRN)

UPMC Hillman Cancer Center (lead)
- **Augusta University Cancer Center**
- **Abramson Cancer Center (Penn)**
- **Stonybrook University (new partner)**
- **Roswell Park Cancer Institute**

Network Trust Agreements
- IRBs agree that use of data for investigators is NHSR, no need for an additional IRB protocol even to access record level de-id data
- Governance
- Agreement to abide by SOPs
- Instrument of Adherence

Soliciting new WSI "ready" partners!

http://ties.dbmi.pitt.edu/tcrn/
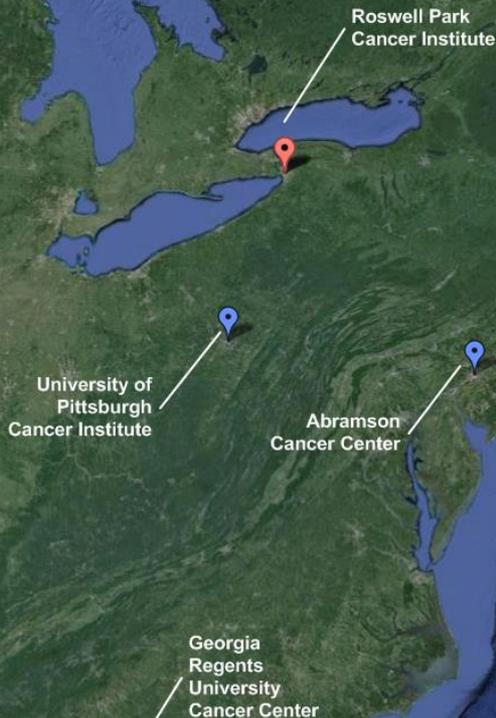
Table 2. TCRN case statistics for numbers of patients and cases (A) and the number of cases of rare tumors (B) and common cancer categories (C) based on final diagnosis

	GRU	RPCI	ACC	UPCI	Total
A. Case statistics					
Patients	76,404	72,376	465,717	1,840,156	2,454,653
Pathology cases	157,316	156,555	857,681	4,588,017	5,759,569
B. Rare tumors					
Adenoid cystic carcinoma	41	88	404	509	1,042
Adrenocortical carcinoma	5	20	59	63	147
Alveolar soft part sarcoma	3	15	10	25	53
Angioimmunoblastic lymphadenopathy	12	35	58	84	189
Chordoma	5	14	124	245	388
Follicular dendritic cell sarcoma	2	2	8	13	25
Merkel cell carcinoma	9	72	165	196	442
Ovarian granulosa cell tumor	4	10	23	34	71
Phaeochromocytoma	15	38	272	164	489
Pleomorphic xanthoastrocytoma	2	5	12	53	72
Pseudomyxoma peritonei	6	36	46	129	217
Rhabdomyosarcoma	34	70	86	270	460
Sebaceous adenocarcinoma	13	33	26	94	166
Sinonasal undifferentiated carcinoma	2	6	31	27	66
Thymoma	13	45	433	210	701
C. Common cancer categories					
Bladder carcinoma	345	1,618	3,873	6,711	12,547
Breast carcinoma	1,143	9,605	28,262	37,691	76,701
Colorectal carcinoma	465	2,530	6,898	11,608	21,501
Endometrial carcinoma	394	1,815	3,707	7,706	13,622
Esophageal carcinoma	63	1,477	2,452	3,514	7,506
Hepatic carcinoma	153	633	2,912	5,720	9,418
Lung carcinoma	820	4,264	10,208	17,955	33,247
Lymphoma	1,387	6,795	10,605	15,689	34,476
Malignant glial neoplasm	242	292	2,198	4,943	7,675
Malignant melanoma	335	2,675	5,180	7,068	15,258
Ovarian carcinoma	503	2,872	4,659	6,446	14,480
Pancreatic carcinoma	162	740	1,866	3,622	6,390
Prostate carcinoma	903	3,612	18,867	19,445	42,827
Renal cell carcinoma	364	1,319	3,183	10,950	15,816
Thyroid carcinoma	474	1,236	7,681	12,387	21,778

Jacobson et al, in press Cancer Research 2015

TIES Cancer Research Network Pubs

http://ties.dbmi.pitt.edu/

Adding Cancer Registry Data to TIES Outcomes Annotation

- **Identified as a high value development target from users**
- **We have secured additional funding from or Institute for Precision Medicine in Pittsburgh**
- **Senior Developer Mike Davis leads this effort.**
- **Starting with Breast Cancer first**
- **Work that we do here can immediately be leveraged by all of you to similarly add Cancer Registry data to your TIES instances**
- **Result = deeper patient annotation and outcomes data**

TIES http://ties.dbmi.pitt.edu/

Cancer Registry Data Elements

Demographics	Primary	Treatment	Outcome
Race	Primary Site	Surgery	Vital Status
Gender	Histology	Chemotherapy	Cancer Status
Age @ Diagnosis	Grade	BRM	Recurrence
Smoking	Path TNM	Hormonal	Cause of Death
Alcohol	Clinical TNM	Immunotherapy	
	Prognostic Factors (including site specific)	Rad Onc	

http://ties.dbmi.pitt.edu/

Enabling Research on the Cancer Registry

Enabling Research on the Cancer Registry

TIES

tranSMART

TIES and the TIES Cancer Research Network

TIES Team

Girish Chavan
Eugene Tseytlin
Kevin Mitchell
Julia Corrigan
Liz Legowski

Adi Nemlekar
Yining Zhao
Vanessa Benkovich
Liron Pantanowitz
Rajiv Dhir

Roswell Park
Carmelo Gaudioso
Monica Murphy
Mayurapriyan Sakthivel
Amanda Rundell

GRU
Roni Bollag
Samir Khleif
Jennifer Carrick
Nita Maihle
And more.....

Penn
Michael Feldman
Nate DiGiorgio
Tara McSherry
Joellen Weaver

Funding

Led by
Rebecca Jacobson, MD MSIS

http://ties.dbmi.pitt.edu/tcrn/

I have some pretty big shoes to fill...

Why Computationally Annotated WSI: To Quantify Tumor Heterogeneity

Inter-tumour heterogeneity

Intra-tumour heterogeneity

Dominance of clone 1

Dominance of clone 2

Mixed dominance

Marusyk A. Nat Rev Cancer. 2012 Apr 19;12(5):323-34
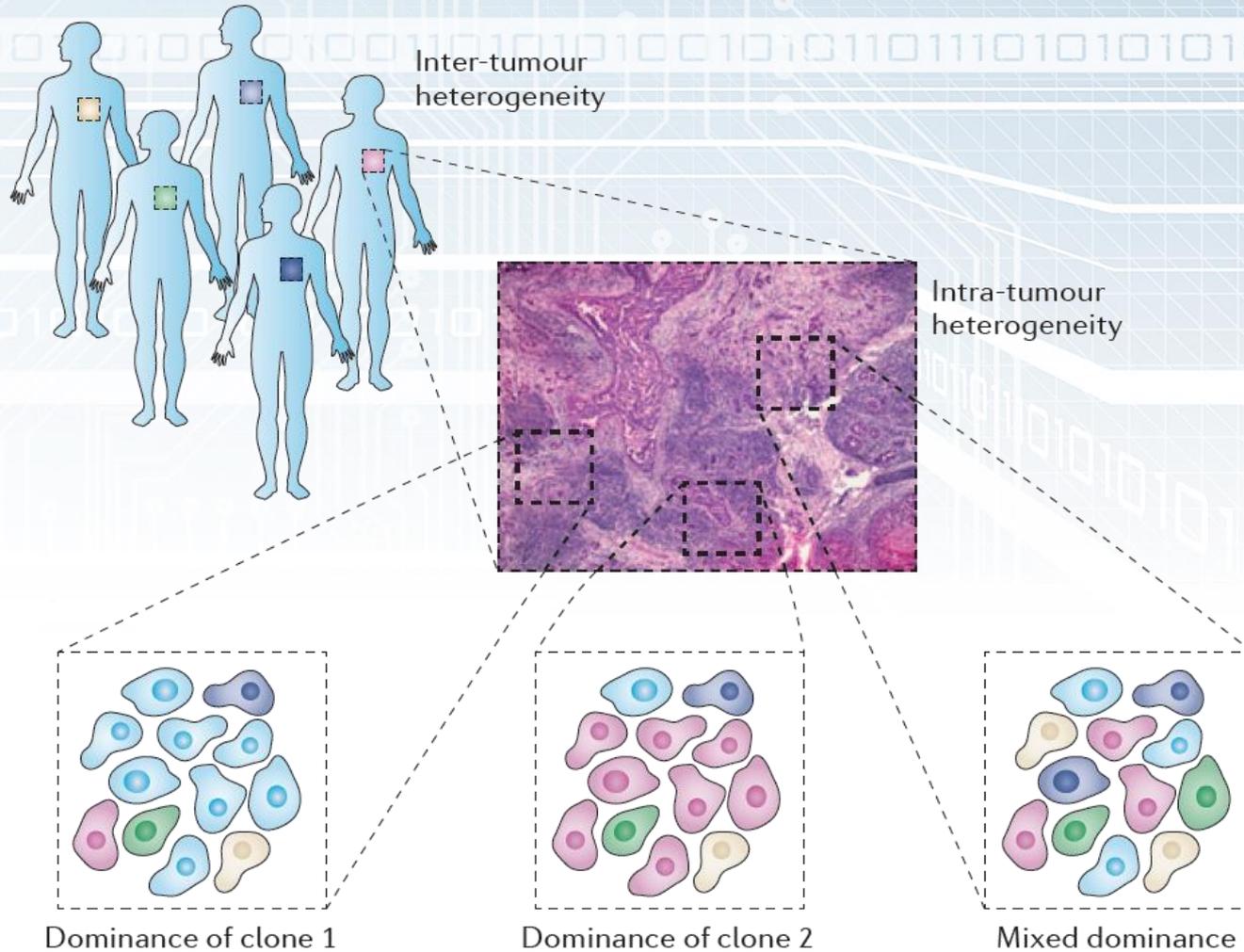
Intra-tumoral spatial heterogeneity complicates accurate diagnosis & prognosis

Normal epithelial cell — Invasive tumor cell — Mast cell — Monocyte — Lymphocyte — Cancer-associated fibroblast — Neutrophil — Macrophage — Lymphatic endothelial cell — Pericyte — Endothelial cell — Bone marrow-derived cells

Big Data to Knowledge (BD2K)

Wisdom	Applied	• I better stop the car!
Knowledge	Context	• The traffic light I am driving towards has turned red
Information	Meaning	• South facing traffic light on corner of Pitt and George Streets has turned red
Data	Raw	• Red, 192.234.235.245.678, v2.0

Pittsburgh, your Self-Driving Uber is arriving now!!!

Pitt's Department of Biomedical Informatics is a Center of Excellence in Big Data to Knowledge

Building Blocks – BD2K - Center for Causal Discovery (Greg Cooper - ACMI)

Center for Causal Discovery

CENTER FOR CAUSAL DISCOVERY

The center will develop the algorithms, software, and system architecture needed by biomedical scientists seeking to discover and represent causality using their large and diverse data sets.

Biomedical Science

We selected 3 very different biomedical problems to use as test beds for our algorithms and to drive the development of new algorithms that meet the needs of biomedical researchers.

Data Science

We are implementing an integrated set of methods that support the graphical representation, discovery, and application of causal knowledge from large and complex biomedical data (see samples of structural causal

http://www.ccd.pitt.edu/
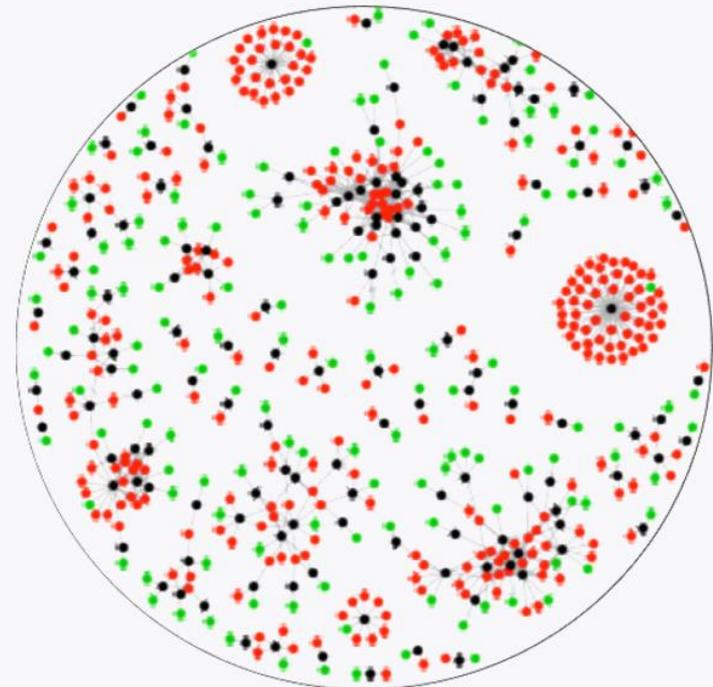
To help discover valid, novel, and significant causal relationships in big biomedical data that lead to new insights in health and disease

Driving Biomedical Problems (DBPs)
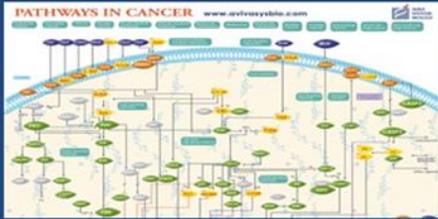
- Discovery of cell signaling networks in breast cancer and head and neck cancer
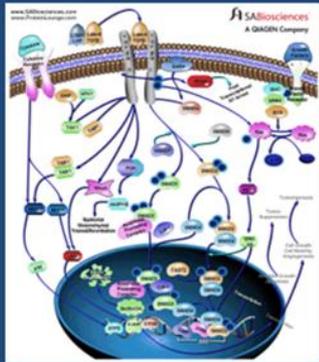
- Discovery of the mechanisms of disease onset and progression in chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis

- Discovery of the functional (causal) connectivity of regions of the human brain from fMRI data

NIH Data Commons Pilots - $55M

https://commonfund.nih.gov/bd2k/commons - Posted 5/19/17

- **Establishing guidelines for Data Commons** operations and what it means for digital objects in the Commons to be FAIR
- Developing and testing **cloud-based platforms** to store, manage and interact with biomedical data and tools
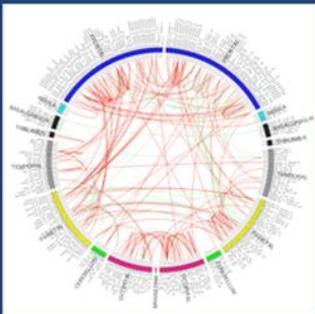- Setting up the ability to access data through appropriate **authorization and authentication** protocols
- **Support interoperability** between existing biomedical data and tool repositories and portability between cloud service providers
- **Creating portals** where users with all levels of expertise can access and interact with data and tools
- **Learning by doing**, which involves developing agile, iterative Data Commons pilots testing its utility, troubleshooting, and retesting
- **Analyzing and evaluating** Data Commons Pilots for cost, utility, efficiency, usability, and adherence to FAIR data principles

Pitt Harvard Data Commons Pilot

Cloud-based Integration of Causal Modeling and Discovery Tools with a Unified Patient Research Database

Jeremy Espino[1], Paul Avillach[2], Michael Davis[1], Jeremy Easton-Marks[2], Michael McDuffie[2], Gabor Korodi[2], Gregory Cooper[1], Isaac Kohane[2], Michael Becich[1]
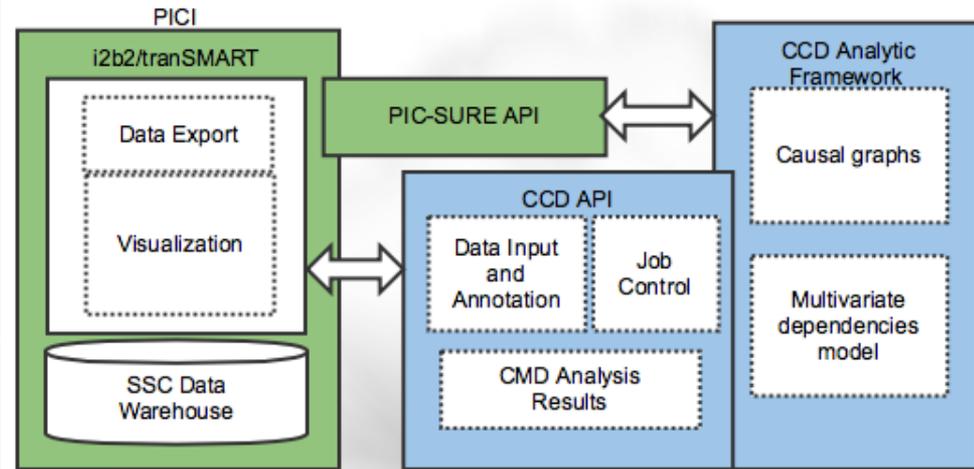
Findable – cloud hosted plus HPC
Accessible – SFARI Autism Data
Interoperable – CCD and PIC-I APIs
Reusable – federated ID management
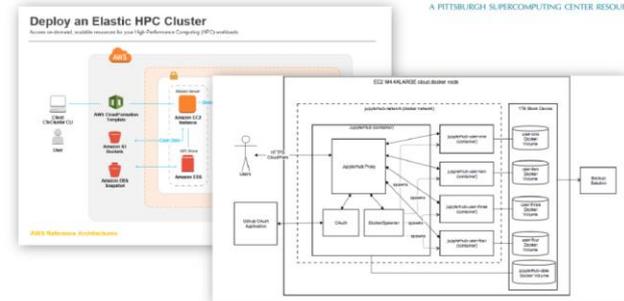
We have deployed our software at **Amazon** using existing **EC2** services as well as an on demand high performance computing cluster, **CfnCluster**. We have also deployed the causal discovery API service on **Bridges** at the **Pittsburgh Supercomputing Center**.

SFARI — SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

Informatics Fuels Data to Knowledge

DIKW paradigm today

Adding understandable data to the stack

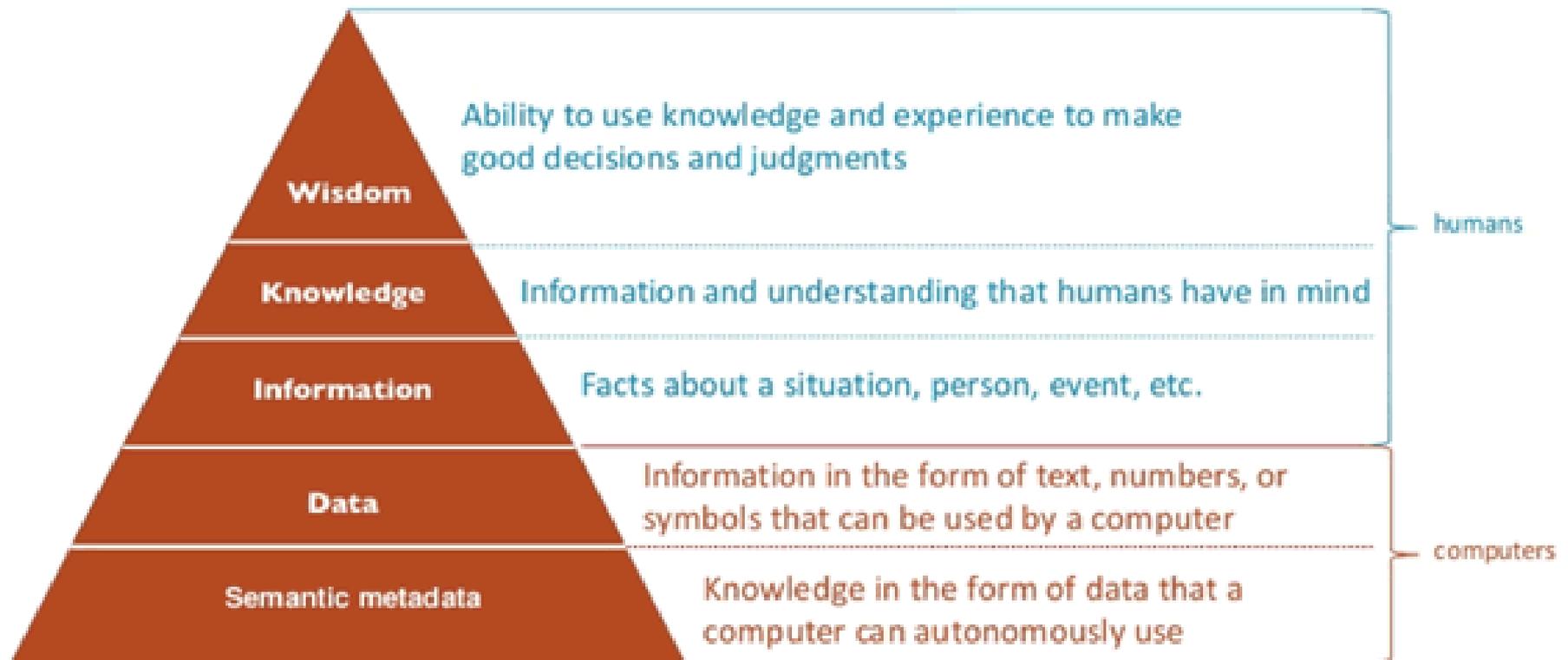| Level	Description
Wisdom	Ability to use knowledge and experience to make good decisions and judgments
Knowledge	Information and understanding that humans have in mind
Information	Facts about a situation, person, event, etc.
Data	Information in the form of text, numbers, or symbols that can be used by a computer
Semantic metadata	Knowledge in the form of data that a computer can autonomously use

humans

computers

Conclusions

- **Informatics will drive a FAIR Data Science Program with major new Pitt Computer Scientists Drs. Paul Cohen (Dean of SCI) and Rob Rutenbar (SVC Research) who will greatly benefit by the vision supported by both UPMC and Pitt.**

- **We are linking our current >$180M Big Data research portfolio (BD2K/NHGRI, CDC, CTSA/NCATS, NIGMS, NCI, PCORI, PMI) to new NIH Initiatives via Pitt Data Commons**

- **Key innovations in FAIR data and research (knowledge) objects, integrative metadata development, causal analytics and novel research computing environments (supercomputing/cloud computing/storage) are key!**

- **We are looking for partners! (Harvard & U Penn Linkages)**

*Please join in this effort by e-mailing me – becich@pitt.edu*
Provide Interest/Skills/Research Goals – I will send you Pitt's RoadMap

Computational Pathology @ Pitt
Led by Chakra Chennubhotla, PhD

- Organized as a Interest Group and Lecture Series

- Currently 109 members from Pitt, CMU, UPMC and regional companies/startups (*need sponsor!!!*)

- Pittsburgh Computational Pathology Interest Group - http://www.csb.pitt.edu/comppath/

- Computational Pathology Lecture Archive - https://www.youtube.com/channel/UCWfBS3PLWHTIACeccm2CIog

- Supported by Akif Burak Tosun, PhD (post-doc)

Computational Pathology @ Pitt
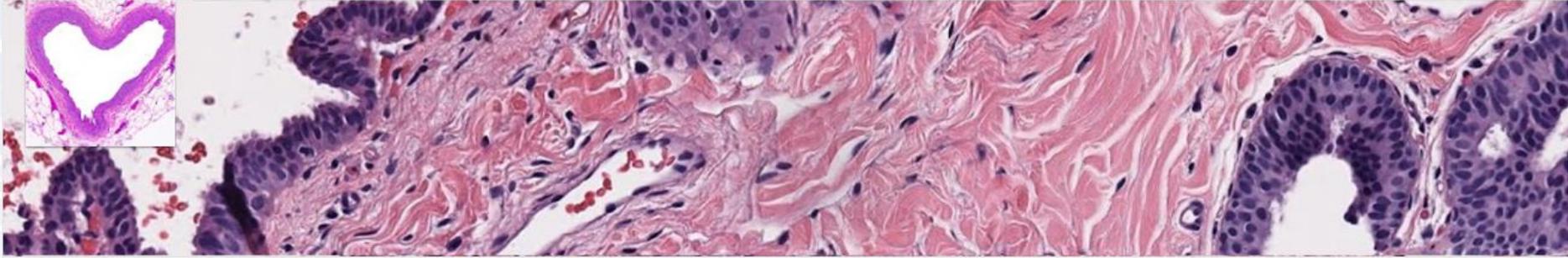
Pittsburgh Computational Pathology Lecture Series

▶ Subscribe 32

Pittsburgh Computational Pathology Interest Group and Lecture Series aims to bring focus on developing basic computer vision and and machine learning algorithms for transmitted light and fluorescence microscopy imaging data from histopathology and on highlighting the translation role of computational pathology research in interfacing with genomics, personalized medicine and microbiomics.

Our goal is to promote a vibrant community of computational pathology research and education practices here in Pittsburgh by engaging researchers and clinicians across Pitt, CMU and Duquesne campuses, and the local industrial partners.

http://www.csb.pitt.edu/comppath/
Show less

Uploads

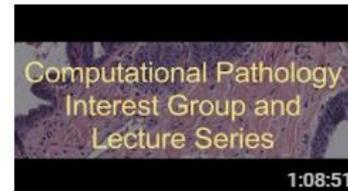| 52:07 | 51:03 | 1:37:28 | 1:08:51 | 48:44 |

CompPath Dr. Shikhar Uttam, Ph.D. Lecture (April 14th 2017)
21 views · 1 month ago
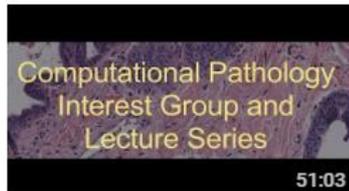
CompPath Dr. Shandong Wu, Ph.D. Lecture (March 14th 2017)
24 views · 1 month ago

CompPath Dr. Jeffrey L. Fine, M.D. Lecture (January 19th 2017)
56 views · 4 months ago

CompPath Dr. Kayhan Batmanghelich, PhD Lecture
48 views · 4 months ago

CompPath Dr. Joshua T Mattila, PhD Lecture (November 14th
41 views · 6 months ago

Computational Pathology Lecture Archive -
https://www.youtube.com/channel/UCWfBS3PLWHTIACeccm2CIog