Cancer Research: Computing and Data

Warren A. Kibbe, Ph.D.
Professor, Biostats & Bioinformatics
Chief Data Officer, Duke Cancer Institute
warren.kibbe@duke.edu

@wakibbe

#HECBDML

Take homes

- Cancer is a grand challenge
- Data generation is no longer the bottleneck in biomedical research – data management, analysis, reasoning are
- Highlight two technologies enabling a much more dynamic view of biology
- Two vignettes highlighting computational and big data challenges in biomedical research

Data access is pervasive

Cancer is a grand challenge

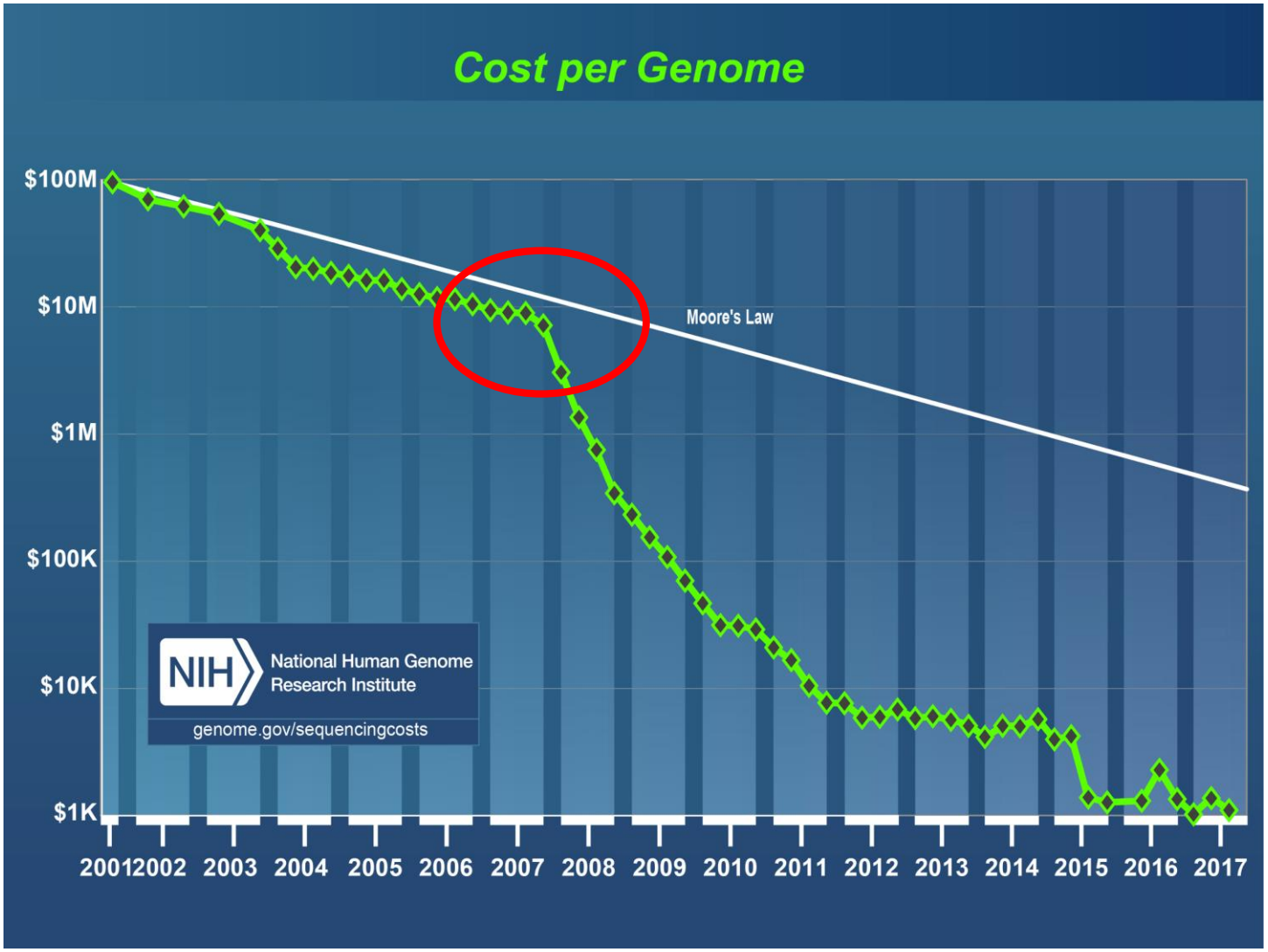
Courtesy of P. Kuhn (USC)

Requires:

- Deep biological understanding
- Advances in scientific methods
- Advances in instrumentation
- Advances in technology
- Data and computation
- Mathematical models

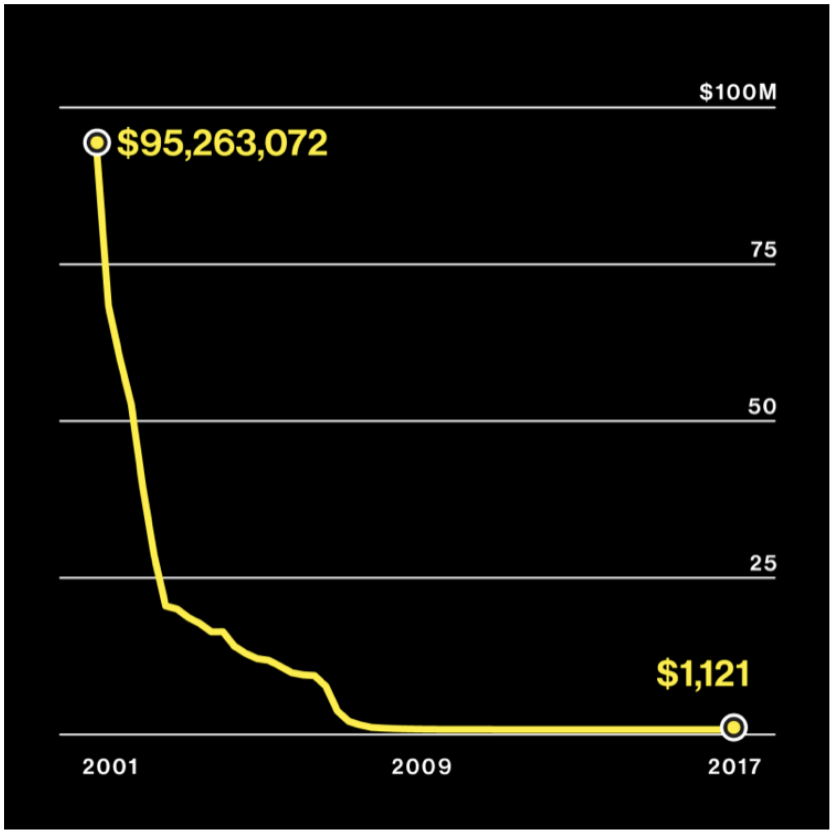
*Cancer Research and Care generate detailed **data** that is critical to create a learning health system for cancer*

Genomics – the poster child

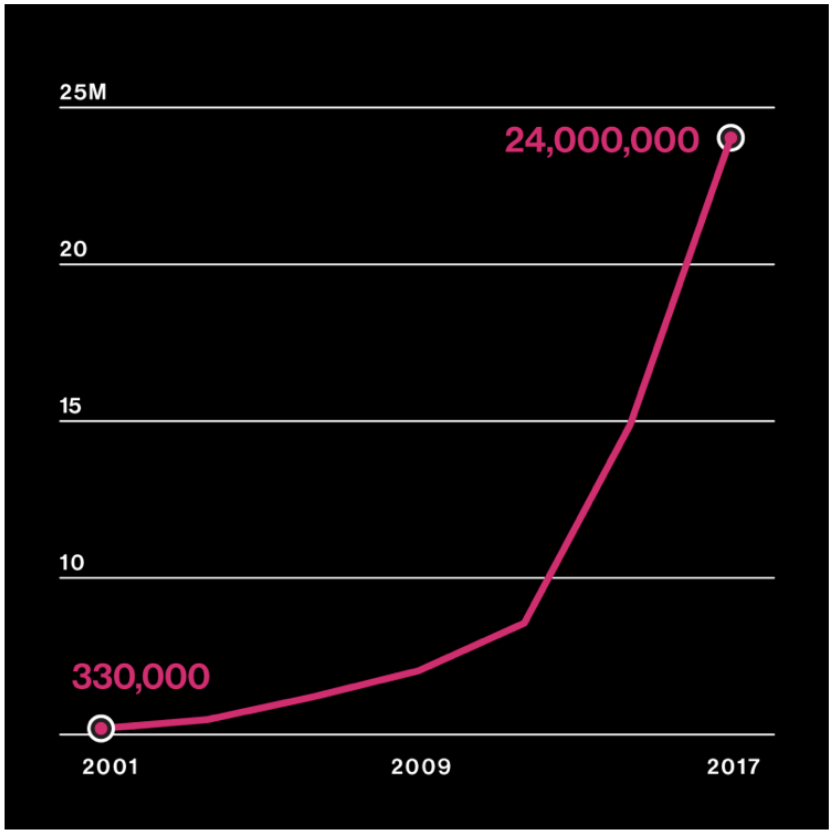


Genomics – the poster child

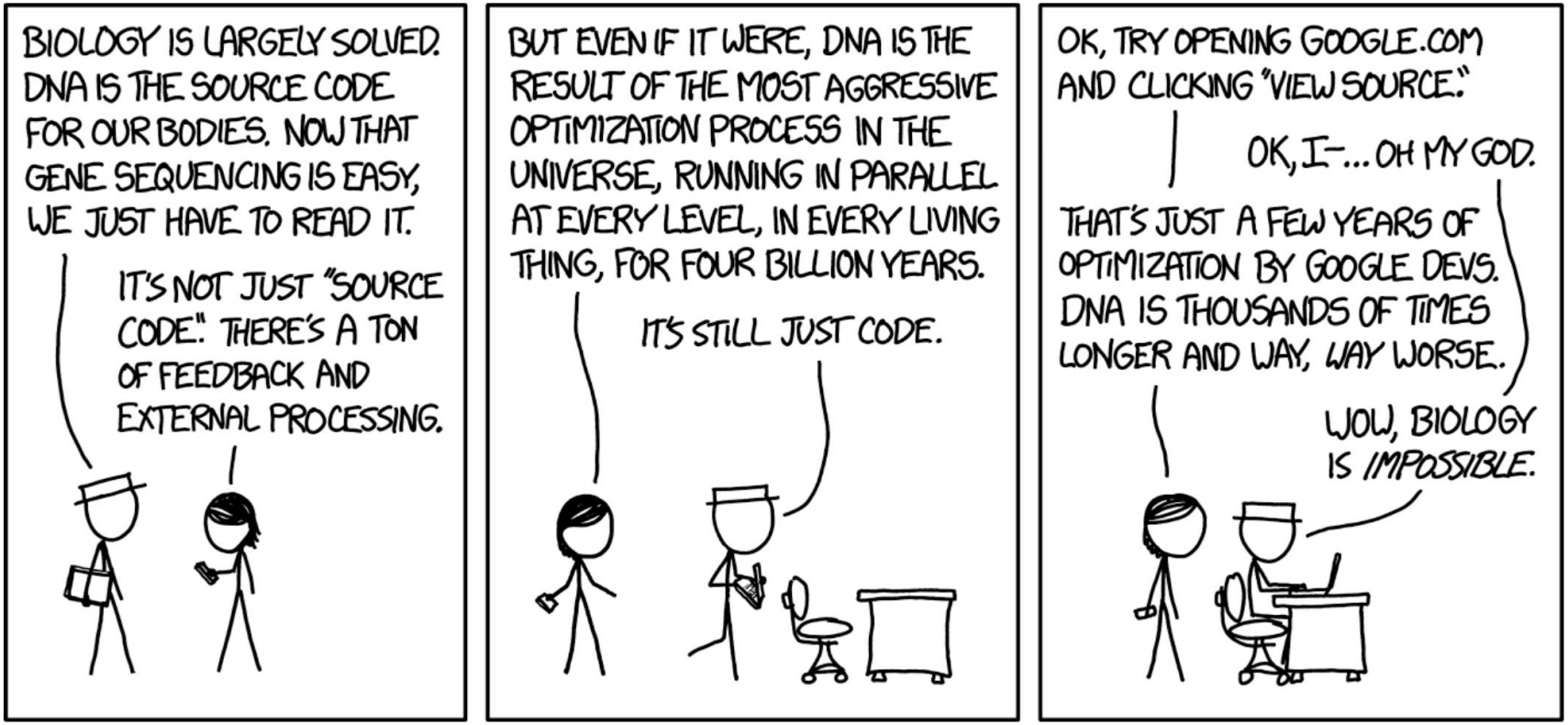
Cost per genome



Consumer genetic testing



Biology is so much more than DNA



https://xkcd.com/1605/

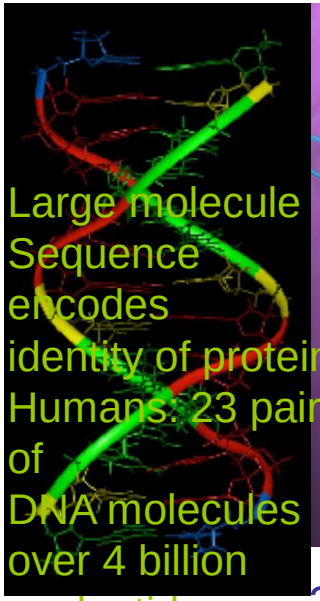
Biology in Four Dimensions

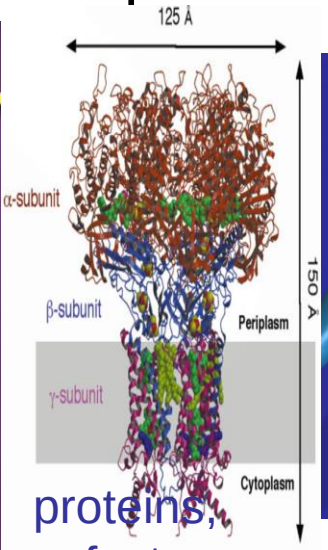
Biological Scales

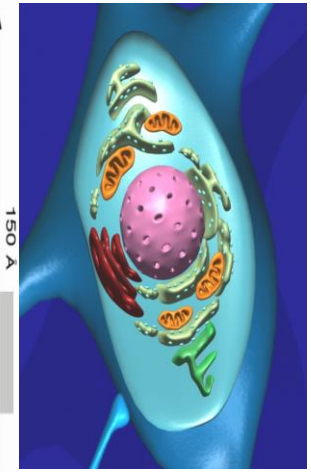
Molecular to Systems Biology

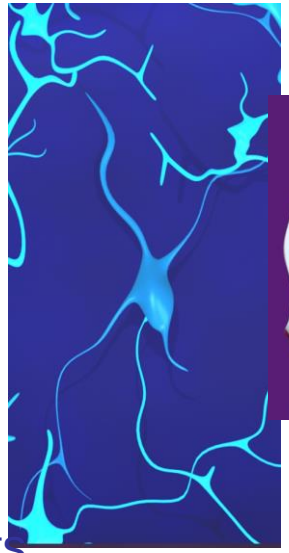
Size scale (meters)

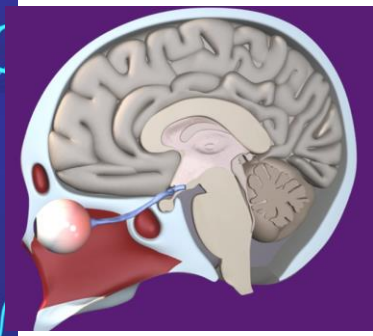
10^{-9}- 10^{-4}	10^{-9} -10^{-8}	10^{-8} -10^{-7}	10^{-7} -10^{-5}	10^{-6} -10^{-2}	10^{-3} -1
DNA	Protein	Complexes	Organelles cells	Tissues	Organs
Large molecule Sequence encodes identity of protein Humans: 23 pairs of DNA molecules over 4 billion nucleotides per pair	30,000 proteins with variants	proteins, cofactors, metabolites, 2nd msgrs	compartments structures function	signaling, networks	emergent properties
10^{9}	10^{-12} -10^{-4}	10^{-6} -10^{3}	10^{-3} -10^{4}	1 -10^{8}	1 -10^{9}









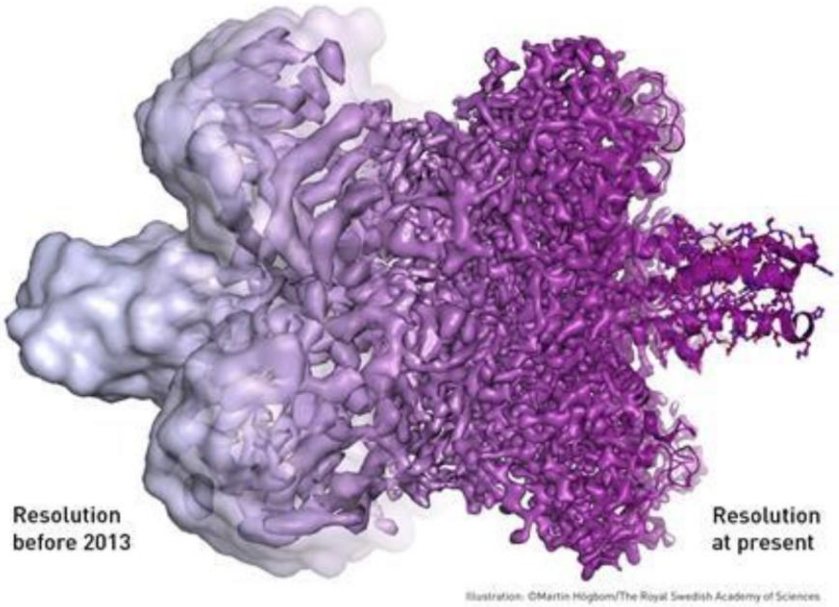


Time scale (seconds)

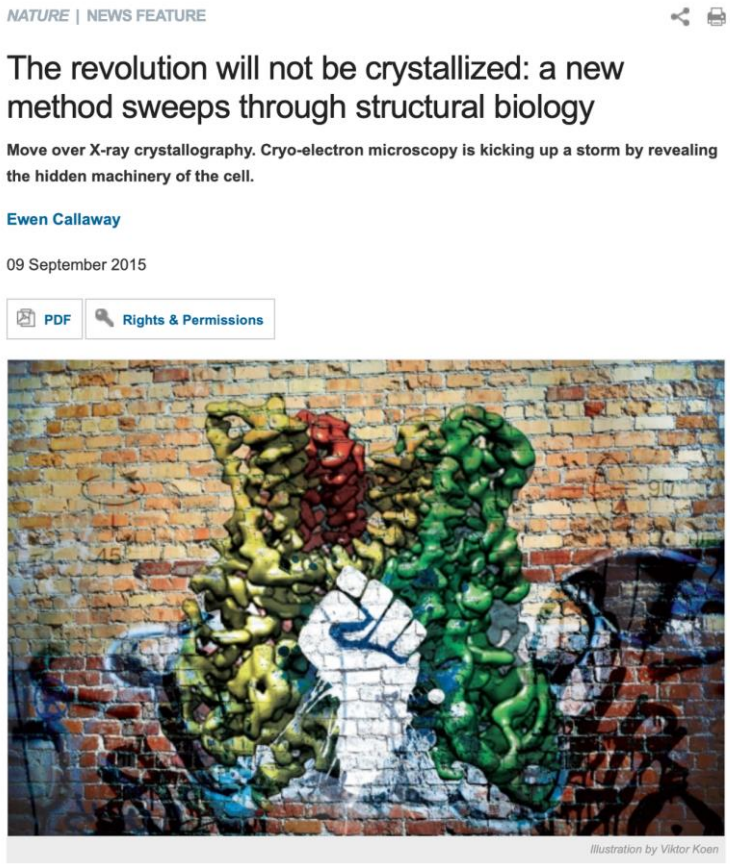
© 2004 WAKibbe, Northwestern University

Cryo-EM

- Able to get atomic resolution of flexible molecules, like membrane-bound proteins



Source: © Martin Högbom, Stockholm University. After an image by V. Falconieri.



NATURE | NEWS FEATURE

The revolution will not be crystallized: a new method sweeps through structural biology

Move over X-ray crystallography. Cryo-electron microscopy is kicking up a storm by revealing the hidden machinery of the cell.

Ewen Callaway

09 September 2015

PDF Rights & Permissions

Illustration by Viktor Koen

Single cell techniques

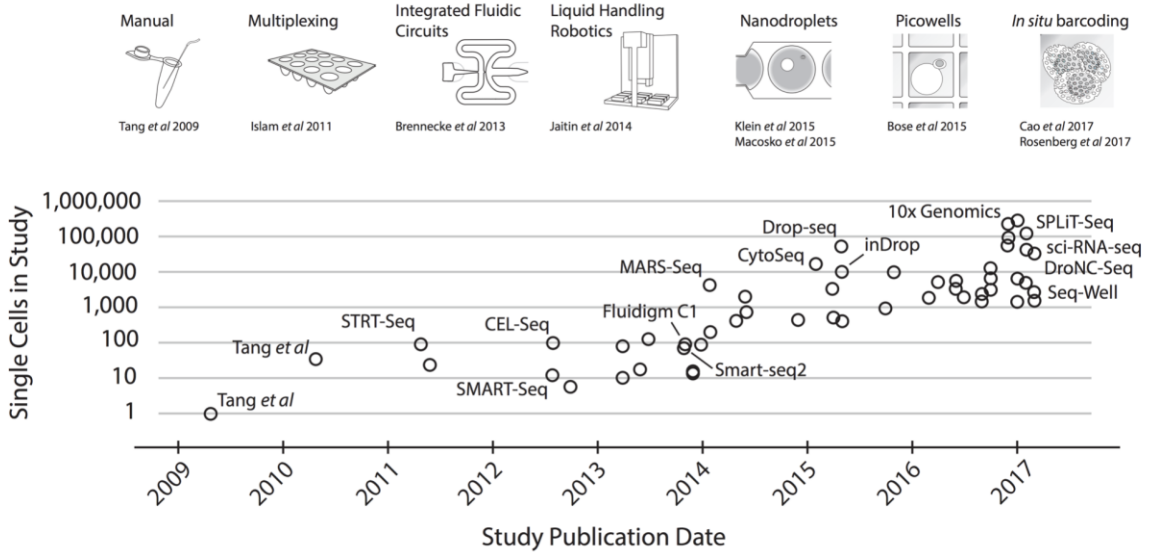


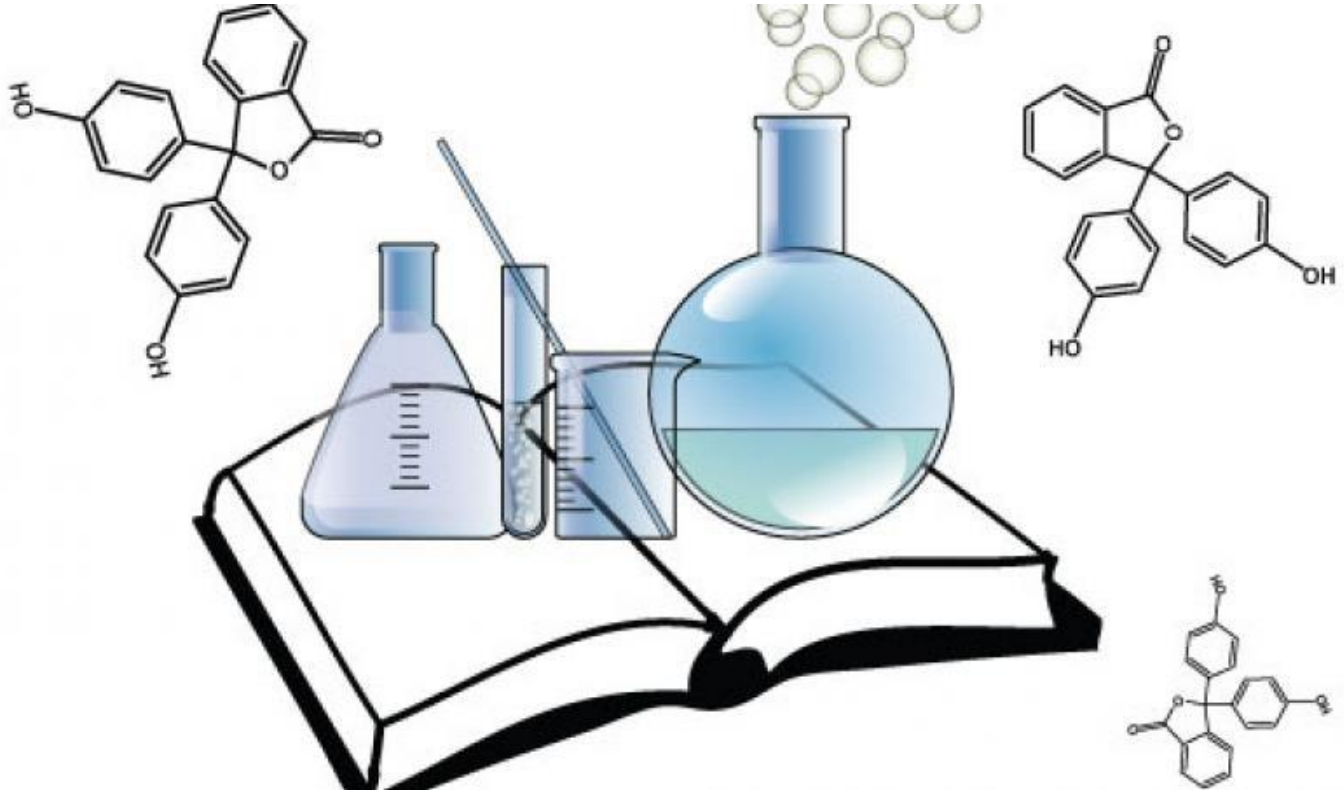
Figure 2.3: Moore's law in single cell transcriptomics (image taken from Svensson et al)

- Sequencing
- Proteomics
- Metabolomics
- Microenvironment

Growing ability to focus on dynamics!

https://arxiv.org/abs/1704.01379

Example Basic Science Problem



NCI RAS Initiative + NCI-DOE Joint Initiative

Oncogenic KRAS is responsible for many human cancers

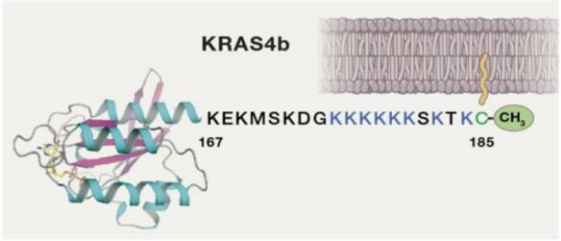
93% of all pancreatic

42% of all colorectal

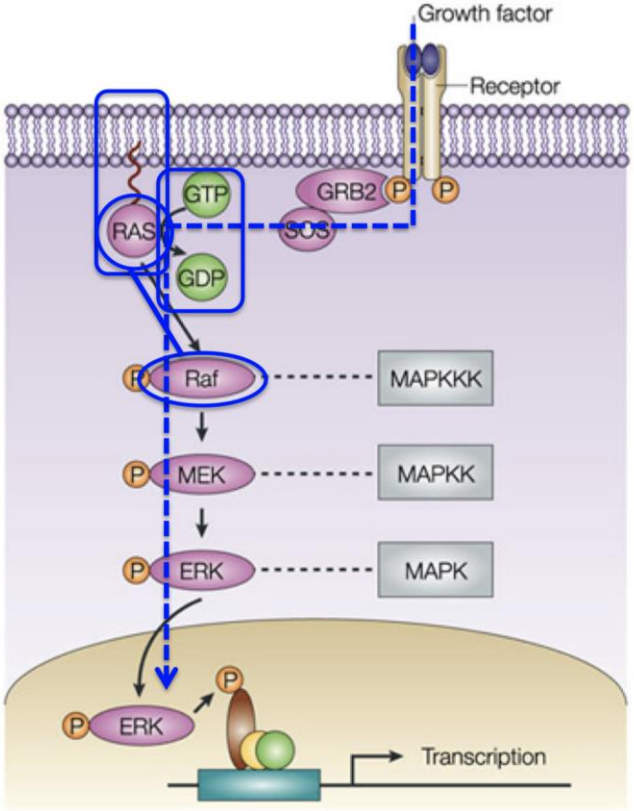
33% of all lung cancers

1 million deaths/year world-wide

No effective inhibitors



Simanshu,Cell 170, 2017



Nature Reviews | Molecular Cell Biology

Pathway transmits signals

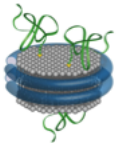
RAS is a switch
oncogenic RAS is "on"

RAS localizes to the plasma membrane

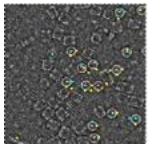
RAS binds effectors (RAF) to activate growth

Cancer Moonshot Pilot 2

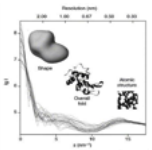
RAS activation experiments (FNLCR)



Experiments on nanodisc



CryoEM imaging

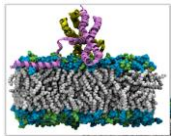


X-ray/neutron scattering

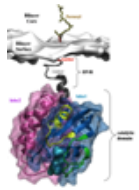
Multi-modal experimental data, image reconstruction, analytics

Protein structure databases

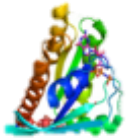
Predictive simulation and analysis of RAS



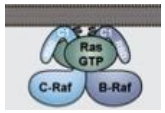
Phase Field model of lipid membrane



Granular RAS membrane interaction simulations



Atomic resolution RAS-RAF interaction

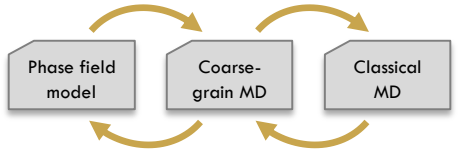


RAS Activation

Adaptive sampling molecular dynamics simulation codes

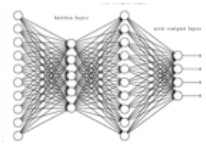
Adaptive time stepping

Adaptive spatial resolution

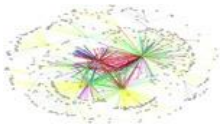


High-fidelity subgrid modeling

Machine learning guided dynamic validation



Unsupervised deep feature learning



Mechanistic network models

Uncertainty quantification

Close collaboration of experimentalists and theorists to build predictive model

Simulations to build model	Experimental data to inform model

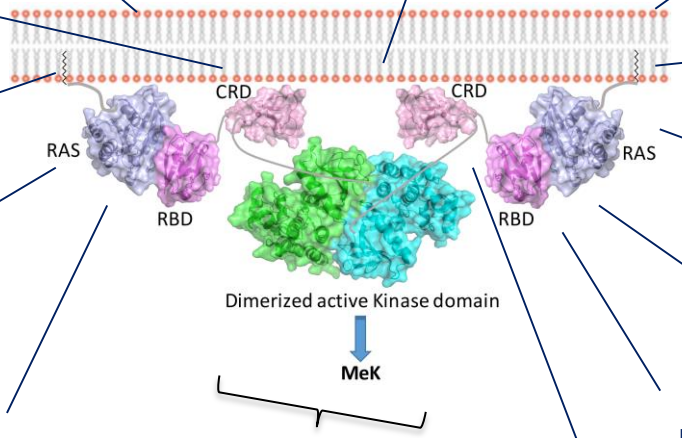
Lipid content: RAS/HVR binding by SPR, alpha assays in nanodiscs, liposomes, imaging in GVUs, lipidomics, SANS (possibly with contrast variation)

Lipid domains: Confocal microscopy RAS/HVR localization in GUVs, Calibrate coarse-grained (CG) simulations with all-atom (AA) Simulations, Calculate free energies of domains

RAS/HVR-membrane binding: SPR in liposomes, biophysical measurements, SANS (with contrast variation), AA and free-energy calculations of RAS/HVR binding to constrain CG parameters, free energies to inform phase field

RAF-membrane affinity: SPR in liposomes, biophysical measurements, MD simulations to identify regions of interest that interact with membrane

HVR structure/dynamics: crystallization, CD, MD of HVR in multi-component lipid platform to inform mobility in phase field model



Farnesyl dynamics: *solid state NMR*, AA and CG simulations of farnesyl in membranes and lipid bilayers informs phase field model

RAS/HVR multimeric state: BRET, step photobleaching, PALM, AA and CG MD of KRAS/HV R on nanodisc and multi-component lipid platform

RAS/HVR mobility & dynamics: single particle tracking, FCS, CG simulations of farnesylated HVR and RAS on nanodiscs and membranes, use to constrain phase field coupling

RAS-RBD binding: SPR, ITC, alpha assays in nanodiscs, TIRF, SANS (possibly with contrast variation), compare with AA simulations and constrain CG simualtions

RAS activity & structure: GTPase, GTP off-rate, crystallization, NMR, cryo-EM?, SANS, AA MD simulations constrain CG parameters

RAS-RBD structure: crystallization, NMR, AA simulations to constrain CG parameters

RAF activation: *dimerization, phosphorylation state(s),* long time-scale CG simulations and kinetic estimation, multi-scale simulations multi-scale simulations of RAS/RAF dynamics on membrane

RBD-CRD and CRAF structure: crystallization, NMR, cryo-EM, CG simulations validated against AA simulations

Team Science is critical

- Analytics and Visualization
- Bioinformatics
- Biostatists
- Clinical Care
- Clinical Research
- Clinical Trials
- Data Science, ML, BD, HPC
- EHRs, Imaging, Lab Systems

Open Data enhances collaboration and team science!

NCI RAS Initiative + NCI-DOE Joint Initiative

Multiscale Model of Lipid Bilayer

To bridge the particle and continuum scales, the relevant degrees of freedom can be described through the framework of a free energy functional.

$$\mathscr{F}\left[\{n_i(\mathbf{r},t)\}\right] = \int_{\mathbb{R}^2} \left(f_{\mathrm{mm}}(\{n_i\}) + \sum_{i=1}^{P} \left[\sum_{j=1}^{N} u_{\mathrm{pm}}^{(i)}(\mathbf{r}-\mathbf{R}_i) n_j(\mathbf{r}) + \frac{1}{2} \sum_{i'=1}^{P} u_{\mathrm{pp}}^{(i)}(\mathbf{r}-\mathbf{R}_i)\delta(\mathbf{r}-\mathbf{R}_{i'}) \right] \right) d\mathbf{r}$$

$$f_{\mathrm{mm}}(\{n_i\}) = \sum_{i=1}^{N} \left(T n_i(\mathbf{r},t) \log\left(\Lambda^2 n_i(\mathbf{r},t)\right) + \frac{1}{2} T \sum_{i'=1}^{N} \int_{\mathbb{R}^2} \Delta n_i(\mathbf{r},t) c_{i,i'}(\mathbf{r}-\mathbf{r}') \Delta n_{i'}(\mathbf{r}',t)\, d\mathbf{r}' + \dots \right)$$

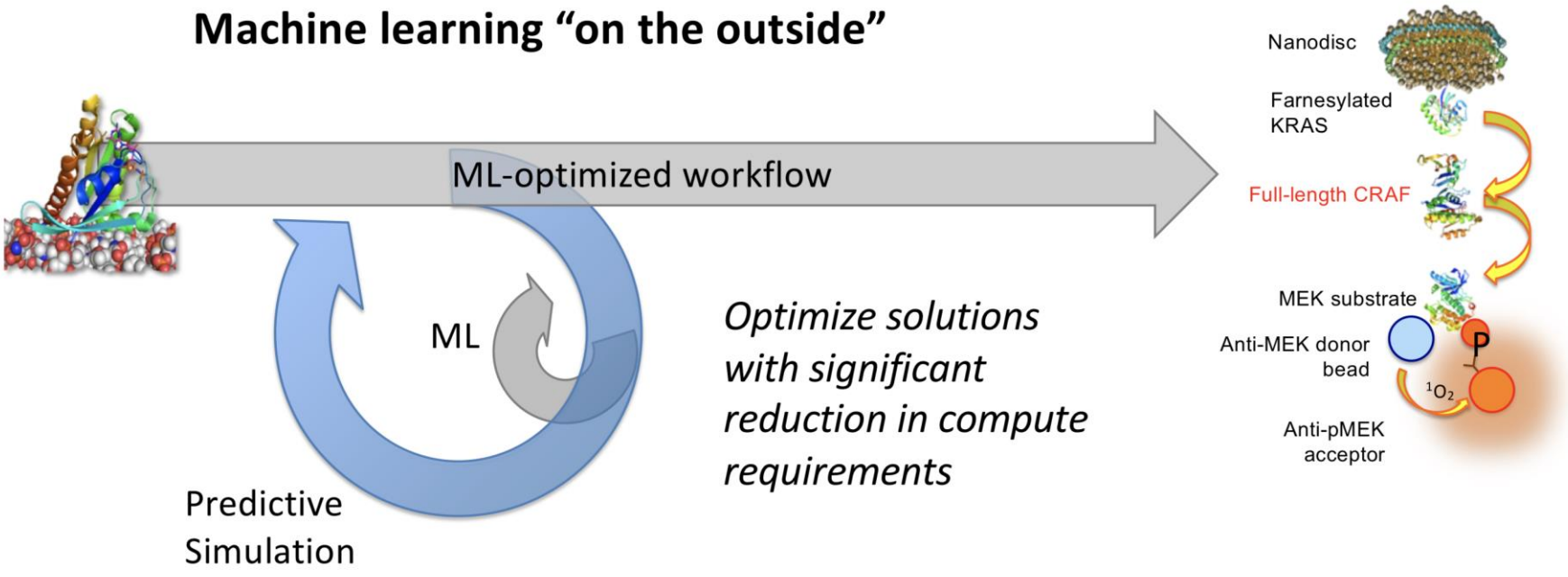
Protein-membrane interaction

Membrane-membrane interaction

Protein-protein interaction

NCI RAS Initiative + NCI-DOE Joint Initiative

Two ways we envision using machine learning techniques along with predictive simulation



Integrated workflows develop insight faster

NCI RAS Initiative + NCI-DOE Joint Initiative

NCI-DOE Pilots: Multi-institution/multi-disciplinary teams

FNLCR / NCI: Debanjan, Goswami, Gulcin Gulten, Rebika Shrestha, **Andrew Stephen**, Tommy Turbyville, Que Van

Oak Ridge National Lab: Debsindhu Bhowmik, **Arvind Ramanathan**

Los Alamos National Lab: Boian Alexandrov, **Angel Garcia**, Nick Hengartner, Jeevapani Hettige, Christoph Jungans, Cesar Lopez, Chris Neale, Sandrasegaram Gnanakaran, Tim Travers, Art Voter

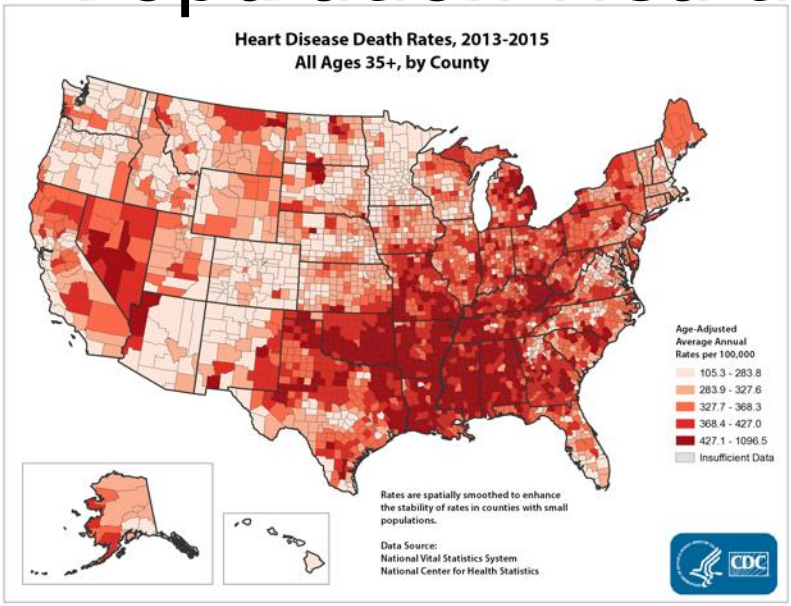
Lawrence Livermore National Lab: Ryan Berg, Harsh Bhatia, Timo Bremer, Tim Carpenter, Gautham Dharuman, Francesco Di Natale, **Jim Glosli**, Helgi Ingolfsson, Piyush Karande, **Felice Lightstone**, Tomas Oppelstrup, Liam Stanton, Shiv Sundram, Michael Surh, Brian Van Essen, Xiaohua Zhang

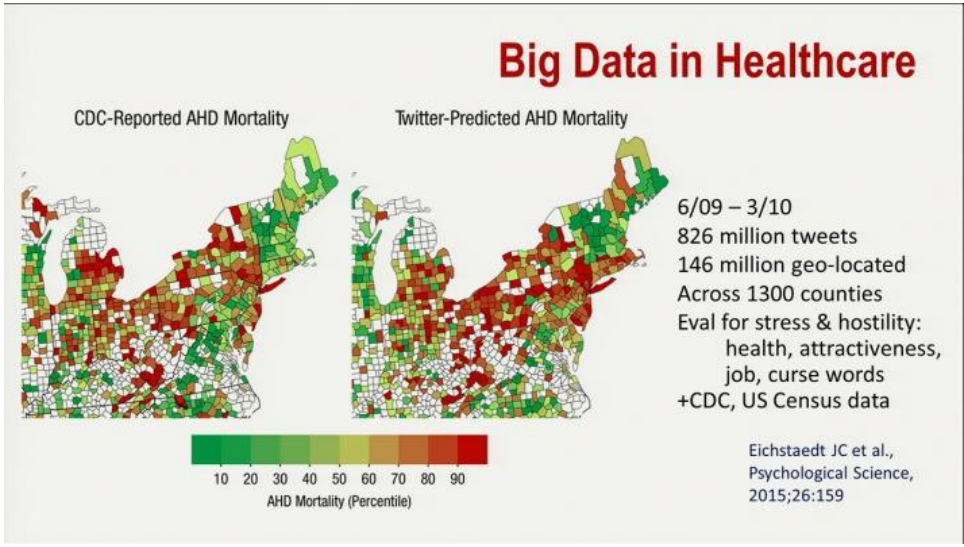
Example from Precision Medicine

Population vs Individual vs Clinic

Health vs Disease

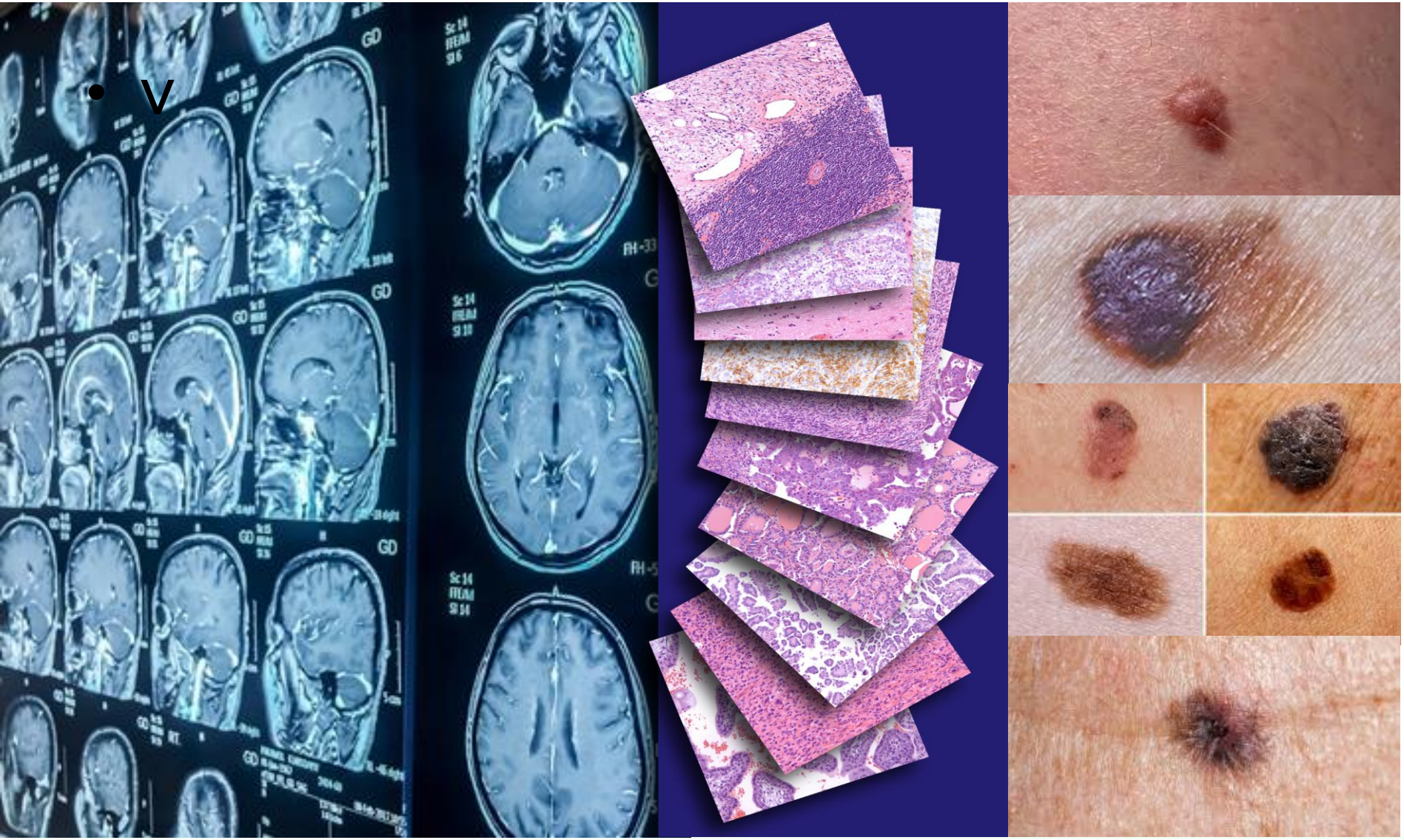
- What is 'normal'?
- Systematic and measurement error
- Biological heterogeneity
- Population Health



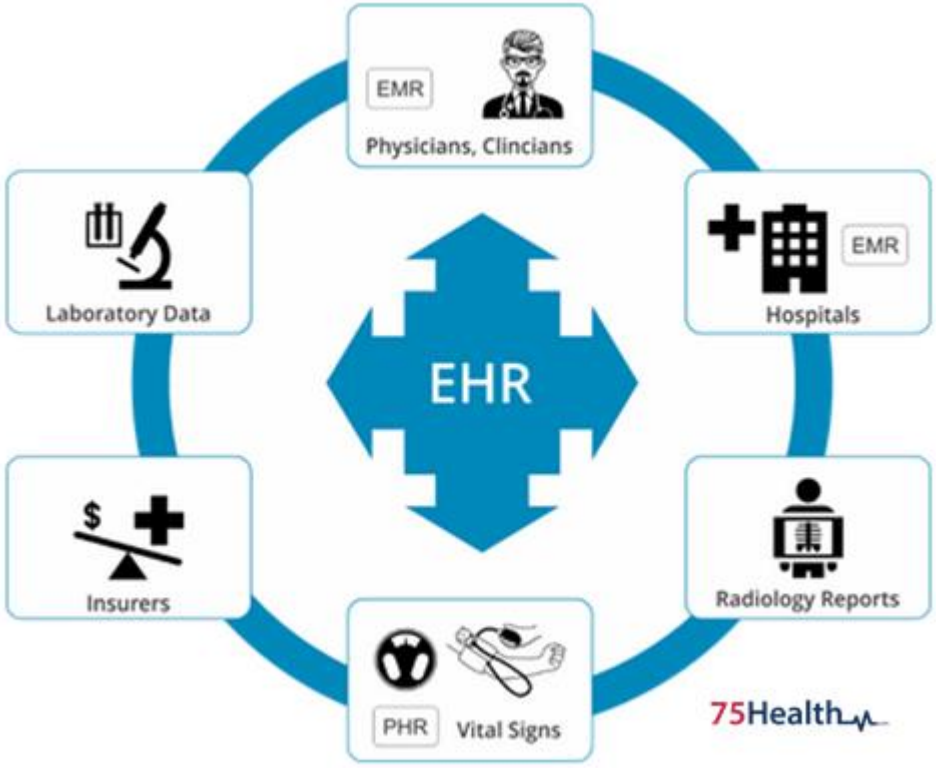


Machine Learning

- V



Clinical, Lab, Molecular data



http://blog.75health.com/what-components-constitute-an-electronic-health-record/

Access to data has changed-Epic

Details | Past Results | Graph of Past Results

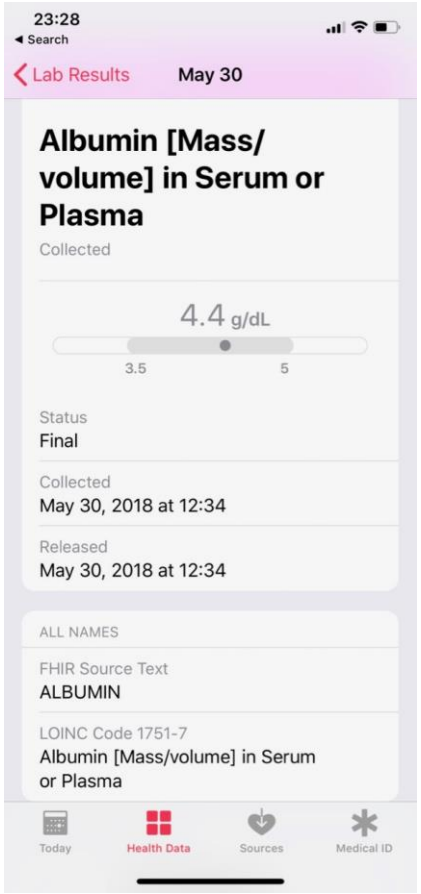
Comments from the Doctor's Office

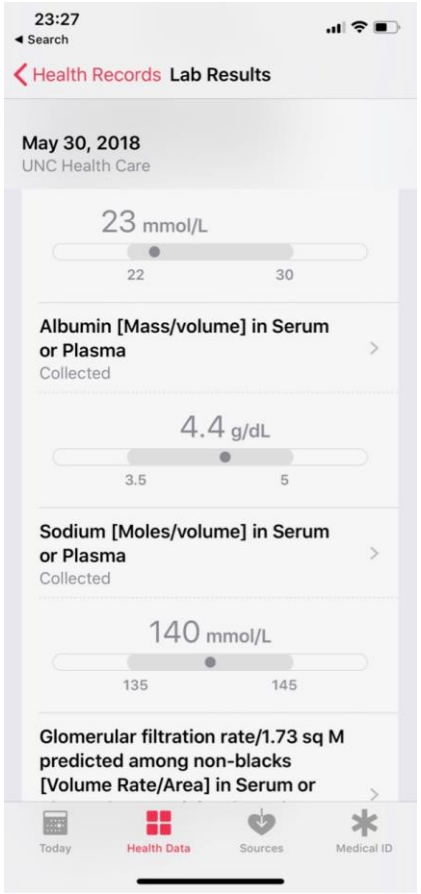
Liver and kidney function are normal. Electrolytes are normal

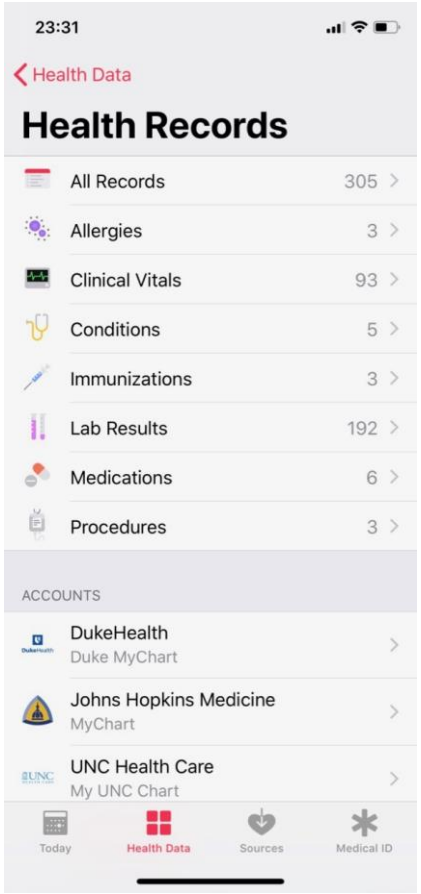
Component Results

Component	Your Value	Standard Range
Sodium	140 mmol/L	135 - 145 mmol/L
Potassium	4.3 mmol/L	3.5 - 5.0 mmol/L
Chloride	104 mmol/L	98 - 107 mmol/L
CO2	23.0 mmol/L	22.0 - 30.0 mmol/L
BUN	16 mg/dL	7 - 21 mg/dL
Creatinine	0.85 mg/dL	0.70 - 1.30 mg/dL
BUN/Creatinine Ratio	19	
EGFR MDRD Non Af Amer	>=60 mL/min/1.73m2	>=60 mL/min/1.73m2
EGFR MDRD Af Amer	>=60 mL/min/1.73m2	>=60 mL/min/1.73m2
Anion Gap	13 mmol/L	9 - 15 mmol/L
Glucose	126 mg/dL	65 - 99 mg/dL
Calcium	8.9 mg/dL	8.5 - 10.2 mg/dL
Albumin	4.4 g/dL	3.5 - 5.0 g/dL
Total Protein	7.2 g/dL	6.6 - 8.0 g/dL
Total Bilirubin	0.5 mg/dL	0.0 - 1.2 mg/dL
AST	36 U/L	19 - 55 U/L
ALT	36 U/L	19 - 72 U/L
Alkaline Phosphatase	45 U/L	38 - 126 U/L

From Apple Health App

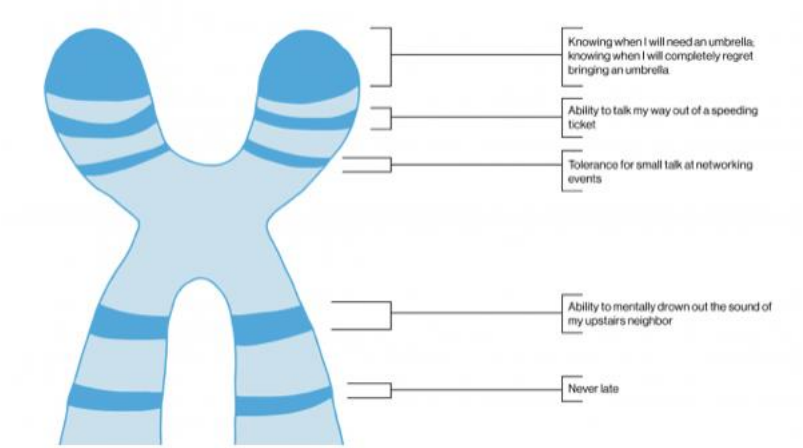






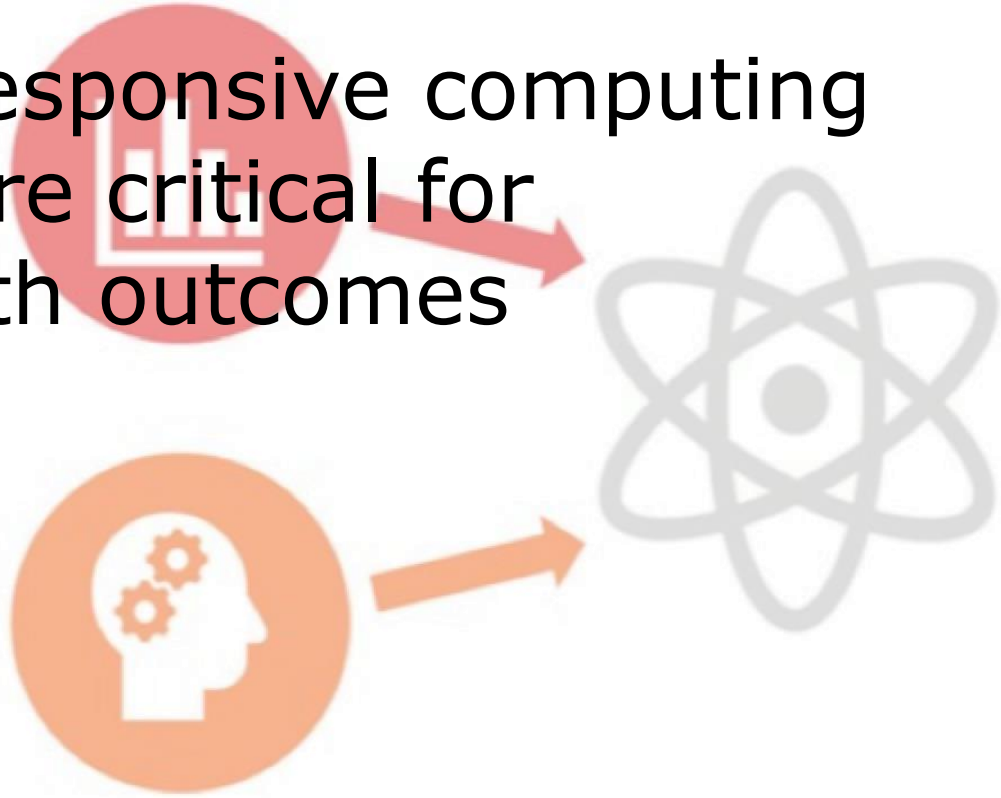
Healthcare

- Evidence is not consistently accessible and structured
- Outcomes are not connected to care
- Patient trajectories are not calculated or accessible



Healthcare

- More data is ‘digital first’ every day
- Decision aids are needed
- Good UX and responsive computing and analytics are critical for improving health outcomes



Understanding Cancer

- **Precision medicine** will lead to **fundamental understanding** of the complex interplay between genetics, epigenetics, nutrition, environment and clinical presentation and **direct effective, evidence-based prevention and treatment**.

Ramifications across many aspects of health care

Questions?

Warren Kibbe, Ph.D.

warren.kibbe@duke.edu

@wakibbe

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov