A Data Driven Approach to Diagnosing and Treating Disease

August 27, 2014

Icahn School of Medicine at Mount Sinai
An orthodoxy is a long-held core belief that is no longer true or applicable in the current landscape.

Orthodoxies are biases, in an industry, company, academic research, and among leaders, that stifle new approaches – and ultimately – innovation.
Who wants to hear actors talk?
– H.M. Warner
Warner Brothers, 1927

There is no reason anyone would want a computer in their home
– Ken Olsen
Founder and president of Digital Equipment Corp., 1977

Anyone who thinks the ANC is going to run South Africa is living in cloud-cuckoo-land
– Margaret Thatcher
British Prime Minister, 1987

I think there is a world market for maybe five computers
– Thomas Watson
Chairman of IBM, 1943

There will never be a bigger plane built
– A Boeing engineer after first flight of the 247, a 10-seater

Biological processes are organized into simple pathways
– Life and Biomedical Sciences
Is medicine poised for a fundamental transformation?
Team Oracle and Big Data at America’s Cup

300 Sensors:
- Strain on mast, hull and wing
- Weather information
- Load on all components (jib, winches, etc.)

3000 variables measure 10 times a sec

300gb of data generated a day
Team Honda and Big Data at Indy

200 Sensors:
- Engine, clutch, gearbox
- Differential, fuel system, oil
- Steering, tires, drag reduction system
- Driver health

Roughly 1gb per lap is generated

Via thousands of simulations, the performance of each component can be predicted with 90% accuracy
What about new born screening in the state of New York?

Heal stick

- Blood spotted on paper
- Assays run to test for 40 or so disorders
- Generates about 10kb of data
We are on the crest of a tsunami in consumer sensor technologies
Printable tattoo biosensor
At Theranos, we’re working to shape the future of lab testing. Now, for the first time, our high-complexity CLIA-certified laboratory can perform your tests quickly and accurately on samples as small as a single drop.
Fake Rocks
Transmitters hidden inside rocks and other objects can receive information from NSA taps implanted in nearby computers even if they’re “air gapped” machines or networks that aren’t hooked up to the Internet—among the hardest of all digital targets.
Even the great city of New York is in on the game

New York City’s government creates a terabyte of raw information every day about everything from parking tickets to electricity. Urban informatics can give structure and new meaning to that sea of data.

How Does Urban Informatics Work? To see how urban informatics works, take this example of how CUSP might help New York City solve issues related to traffic congestion.

See the NYU Center for Urban Science and Progress for More Details
Get answers about your health without the hassle.

Feel better, have more energy & find out what's going on inside your body - quickly and conveniently. Spend more time doing what you love, less time at the doctor.

Create FREE Account  Compare all packages
Considering the digital universe of data to better diagnose and treat patients.
Multiscale measures of patients now available through efforts like Mount Sinai’s Biobank (>25,000 identified patients and growing fast)

- Germ line
- Somatic
- Epigenome
- DNA
- RNA
- Microbiome
- Immune
- Gut
- Mouth
- Environment
- EMR
- Diagnoses
- Mobile Apps (eg, GINGER.io)
- Drugs
- Labs
- Hospital Based
- Consumer Based
- Procedures
That promise to enable the construction of molecular networks that define the biological processes that comprise living systems.
Thinking outside of the box: The Candle Problem

Group 1

Group 2
Thinking outside of the box: The Candle Problem

Solution
The more expert you are on a topic the harder it is for your brain to think differently.

The brain is extremely complex but also has a fixed energy budget... it is lazy.

To make sense of what it sees it takes short cuts all the time.

To break free, you must consciously jolt your brain regularly.
Many examples of others having successfully broken “Orthodoxies”

No one believed consumers would pay extra to have equipment installed in their homes.

No one believed people who shopped in discount stores would pay extra for designer products.

No one believed that BMW drivers were more concerned about parking damage than high performance.

No one believed consumers would buy electronics in stores after Dell had changed the model.
These technologies are enabling scoring of very large-scale, high-dimensional data on individuals for low cost.
Integrating data to build predictive models of living systems

( - DNA,  - RNA,  - Protein,  - Metabolite)
Computational Infrastructure

Our two newest systems rank among the largest in academic medicine in the U.S. and are capable of more than 100 million hours of computation per year.

Our Minerva high-performance computing system provides a robust computational and data-intensive infrastructure geared specifically for genomics. Minerva was recently expanded with new Intel processors to over 400 servers, comprising over 15,000 cores, over 50TB of RAM, and 7.3 PB of raw storage capacity.

Demeter is the core of our new Data Science Ecosystem - a Hadoop cluster currently composed of 250 servers, 10,000 logical cores, 32 TB of RAM, and 3.6 PB of raw storage set up for integrated mining across all data assets at Mount Sinai.

Demeter runs the same software that is used by companies such as Facebook and Walmart to improve business outcomes through the analysis of petabytes of data.
Building networks from high-dimensional data scored in populations

Monitor disease and molecular traits in populations

Putative causal gene

Disease trait
Establishing causality:

Classic Approach: Direct, hard hitting, artificial perturbations of genes

Insulin Causal Gene

Insulin Levels
Mendelian Randomization as a Path to Causal Inference

An integrative genomics approach to the reconstruction of gene networks in segregating populations

J. Zhu,¹ P.Y. Lum,¹ J. Lamb,¹ D. GuhaThakurta,¹ S.W. Edwards,¹ R. Thieringer,² J.P. Berger,³ M.S. Wu,⁴ J. Thompson,⁵ A.B. Sachs,¹ and E.E. Schadt¹
Integrating all data to predictive network models of living systems
Markov Equivalent Structures
(Statistically Indistinguishable Candidate Relationships)

Integration with Bayesian Network Reconstruction Methods

Final Network Model

True Relationship

Inferred Relationship

Genotype Locus L1
Genotype Locus L2
Genotype Locus L3
Genotype Locus L4

Trait Values

Time Lag

Green Node

Blue Node
Toward whole-cell models for science and engineering

Jonathan Karr
A better *understanding* of the network of molecular and cellular processes has the potential to lead to network enabled wisdom in medicine.
Organizing 163 genetic loci for IBD

Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease

Problem: How do you make sense of 163 loci to understand a complex disease like IBD?
Constructing predictive network models for IBD
From these causal network structures we can identify points of therapeutic intervention.
Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer’s Disease

Connections between diseases and tissues: IBD network driving Alzheimer’s

Building networks from 500 prefrontal cortex samples
Constructing the co-expression networks

“Normal” versus LOAD Networks
Causal probabilistic network relating to a PFC module correlating with multiple LOAD clinical covariates, enriched for immune function/pathways related to microglia activity

We identified TYROBP as a key regulator of this network.

CD33, MS4A4A, MS4A6A (from LOAD GWAS)

Two papers in NEJM today reporting on rare variants in TREM2 associate with LOAD.
The microglia pathogen phagocytosis pathway

TREM2 signals directly via TYROBP
Systems analysis of eleven rodent disease models reveals an inflammatome signature and key drivers

<table>
<thead>
<tr>
<th>Similar set: upregulated</th>
<th>Enrichment P</th>
<th>Overlap</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inflammatory response</td>
<td>4.76E-61</td>
<td>208</td>
<td>704</td>
</tr>
<tr>
<td>Leukocyte activation</td>
<td>2.13E-32</td>
<td>164</td>
<td>704</td>
</tr>
<tr>
<td>Regulation of immune response</td>
<td>1.44E-25</td>
<td>84</td>
<td>260</td>
</tr>
<tr>
<td>Cytokine production</td>
<td>6.10E-18</td>
<td>85</td>
<td>335</td>
</tr>
<tr>
<td>Chemotaxis</td>
<td>4.97E-16</td>
<td>74</td>
<td>284</td>
</tr>
<tr>
<td>Humoral immune response</td>
<td>3.25E-14</td>
<td>69</td>
<td>271</td>
</tr>
<tr>
<td>Mitotic cell cycle</td>
<td>7.64E-13</td>
<td>87</td>
<td>414</td>
</tr>
<tr>
<td>Induction of apoptosis</td>
<td>1.74E-12</td>
<td>86</td>
<td>412</td>
</tr>
<tr>
<td>TLR signaling pathway</td>
<td>4.66E-12</td>
<td>21</td>
<td>47</td>
</tr>
<tr>
<td>Phagocytosis</td>
<td>2.74E-11</td>
<td>38</td>
<td>111</td>
</tr>
<tr>
<td>Innate immune response</td>
<td>9.29E-11</td>
<td>48</td>
<td>173</td>
</tr>
<tr>
<td>ECM remodeling</td>
<td>9.67E-11</td>
<td>22</td>
<td>59</td>
</tr>
<tr>
<td>Osteoclast differentiation</td>
<td>3.61E-10</td>
<td>31</td>
<td>82</td>
</tr>
<tr>
<td>Regulation of cell proliferation</td>
<td>4.19E-10</td>
<td>112</td>
<td>662</td>
</tr>
<tr>
<td>Antigen processing and presentation</td>
<td>7.44E-10</td>
<td>32</td>
<td>89</td>
</tr>
<tr>
<td>Positive regulation of translation</td>
<td>3.64E-09</td>
<td>45</td>
<td>170</td>
</tr>
<tr>
<td>Cytokine production by Th17 cells</td>
<td>6.33E-09</td>
<td>17</td>
<td>40</td>
</tr>
<tr>
<td>Angiogenesis</td>
<td>9.41E-09</td>
<td>60</td>
<td>277</td>
</tr>
<tr>
<td>Cell-cycle process</td>
<td>2.79E-08</td>
<td>104</td>
<td>636</td>
</tr>
<tr>
<td>Wound healing</td>
<td>1.79E-07</td>
<td>57</td>
<td>274</td>
</tr>
<tr>
<td>Regulation of translation</td>
<td>4.44E-07</td>
<td>66</td>
<td>349</td>
</tr>
<tr>
<td>Macrophage activation</td>
<td>1.61E-06</td>
<td>20</td>
<td>49</td>
</tr>
<tr>
<td>Interleukin-12 production</td>
<td>1.70E-06</td>
<td>18</td>
<td>40</td>
</tr>
</tbody>
</table>
Core disease modules harbor pluripotent drug targets
The predictive network models we will construct will enable stratification of patient populations.
Integrating diverse data for psychiatric disease to get at predictive models of these diseases
Identify regions of interest from cases and controls

A significant cluster, based on both univariate test and WGCNA.
DLPFC Co-expression Network

DLPFC Bayesian Sub-Network

Neuroimaging Sub-Network

Imaging QTL

DNA Methylation

DNA-Protein

Literature

Cis eQTL

Protein-Protein

SNP1

CNV2

SNP3

CNV4

SNP5

SNP6

SNP7

Imaging-based Brain Connectivity Network

Subtype 1

Subtype 2
Making the link between imaging and molecular data
Constructing a new map of the Allen Brain Atlas
To link imaging data we start with a map of 115 regions of interest constructed using Ayasdi’s topological data analysis platform.
Regions identified as differential between cases and controls

brain stem
parahippocampal_gyrus_posterior_division
thalamus
postcentral_gyrus
Now project the imaging graph into the gene expression graph

- **go_process activation of Rap GTPase activity**
- **go_process positive regulation of Rap GTPase activity**
- **go_process regulation of hippo signaling cascade**
- **go_process regulation of macrophage fusion**

**parahippocampal_gyrus**

**Brain Stem**

- **go_process** The regulation of cysteine-type endopeptidase activity
Now project the imaging graph into the gene expression graph

Post central gyrus

- go_process base conversion or substitution editing
- go_process T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell
- go_process leucine catabolic process
- go_process positive regulation of histone methylation
- go_process positive regulation of centriole replication
- go_process regulation of histone H3-K36 methylation
- go_process positive regulation of histone H3-K36 methylation
- go_process histone H3-K36 methylation

thalamus

- go_process cellular response to prostaglandin stimulus
- go_process nickel ion transport
- go_process iron ion transmembrane transport
- go_process ferrous iron transport
- go_process nickel ion transmembrane transport
- go_process lead ion transport
- go_process ferrous iron import
- go_process vanadium ion transport
Telencephalon (Lateral ventricle is near telencephalon and thalamus)

- go_process cellular response to prostaglandin stimulus
- go_process response to prostaglandin stimulus
- go_process positive regulation of cholesterol storage
- go_process triglyceride mobilization
- go_process regulation of lipid biosynthetic process
- go_process macrophage derived foam cell differentiation
- go_process foam cell differentiation
- go_process negative regulation of fatty acid metabolic process

From these graphs we can project the genes comprising the pathways enriched in these regions to predictive network models.
A network informed view of schizophrenia
From this projection we identify a causal, probabilistic network structure that is 40-fold enriched for genes that harbor variants in conserved regulatory regions that associate with brain gene expression and schizophrenia.
Not Perturbed

Perturbed

Profile SCZ Lines

Project Signature Onto Network

Component of Normal or Disease Network

Perturbagen Target or Perturbagen Node
What is the temporal and spatial identity of hiPSC neural cells relative to the human brain?
12 weeks = electrophysiologically active

induced N\textsuperscript{a\textsubscript{+}}/K\textsuperscript{+} currents

induced action potentials
12 weeks = electrophysiologically active

induced Na⁺/K⁺ currents

spontaneous EPSCs

induced action potentials

spontaneous IPSCs
12 weeks = electrophysiologically active

**induced Na⁺/K⁺ currents**

Control

SCZD

**spontaneous EPSCs**

Control

SCZD

**induced action potentials**

Control

SCZD

**spontaneous IPSCs**

Control

SCZD

Fluo-4AM, 3 month old neurons, real time
Screening networks for novel drug discovery

Integrative Biology Platform

- RNA
- Protein
- Metabolite
- Clinical Data
- Literature
- Image Data

DNA

SZ Patients and Controls

hiPSCs

NPCs

Dopaminergic

GABAergic

Glutamatergic

Neurons

Healthy Network

Toxicity Network

HZ Subtype 1

HZ Subtype 2

SZ Gene Reporter Set

HTS Assays

hiPSC derived neurons

Screening networks for novel drug discovery

Integrative Biology Platform

- RNA
- Protein
- Metabolite
- Clinical Data
- Literature
- Image Data

DNA

SZ Patients and Controls

hiPSCs

NPCs

Dopaminergic

GABAergic

Glutamatergic

Neurons

Healthy Network

Toxicity Network

HZ Subtype 1

HZ Subtype 2

SZ Gene Reporter Set

HTS Assays
## Acknowledgements

<table>
<thead>
<tr>
<th>Mount Sinai</th>
<th>Mount Sinai</th>
<th>Cornell</th>
<th>Harvard</th>
</tr>
</thead>
<tbody>
<tr>
<td>----------------------------------------------------------------------------</td>
<td>--------------------------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Sage Bionetworks               University of Bonn                  McKinsey &amp; Company</td>
<td>Stay connected with us!</td>
<td>@multiscalebio</td>
<td></td>
</tr>
</tbody>
</table>