Big Data and Computational Pathology

The Future of Pathology Informatics

Bruce Levy, MD, CPE
Associate Chief Health Information Officer
Associate Professor of Pathology
University of Illinois at Chicago

THE UNIVERSITY OF ILLINOIS AT CHICAGO
UIC

PATHOLOGY INFORMATICS SUMMIT 2016
May 23-26, 2016
Pittsburgh, PA
Brought to you by the Association for Pathology Informatics.

Big Data and Computational ~~Pathology~~ Medicine

The Future of Pathology Informatics

Bruce Levy, MD, CPE
Associate Chief Health Information Officer
Associate Professor of Pathology
University of Illinois at Chicago

THE UNIVERSITY OF ILLINOIS AT CHICAGO
UIC

PATHOLOGY INFORMATICS SUMMIT 2016
May 23-26, 2016
Pittsburgh, PA
Brought to you by the Association for Pathology Informatics.

Notice of Faculty Disclosure

In accordance with ACCME guidelines, any individual in a position to influence and/or control the content of this ASCP CME activity has disclosed all relevant financial relationships within the past 12 months with commercial interests that provide products and/or services related to the content of this CME activity.

The individual below has responded that he/she has no relevant financial relationship(s) with commercial interest(s) to disclose:

Bruce Levy, MD

Objectives

- Explain the Intelligence Cycle
- Define Big Data
- Discuss the emerging field of Computational Pathology and provide examples of how it is already changing pathology and medicine
- Explain why our leadership in Computational Pathology/Medicine is necessary and important

Where Are We Coming From?

The “great divide” between AP and CP

<u>Anatomic Pathology</u>

- Descriptive and interpretive
- Morphology-based
- Unstructured

<u>Clinical Pathology</u>

- Structured and computerized
- Quantitative
- Little interpretation

- More structure in AP
- More interpretation in CP
- Integrated reporting

?

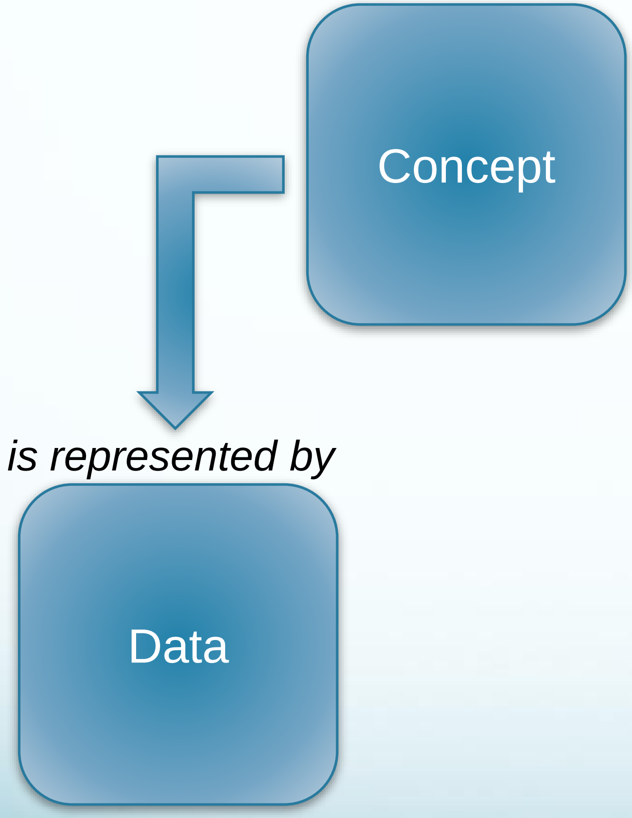
How do we achieve intelligence?

Got Intelligence?

A concept is the representation of an idea

Expression of ideas in healthcare requires detail

Got Intelligence?

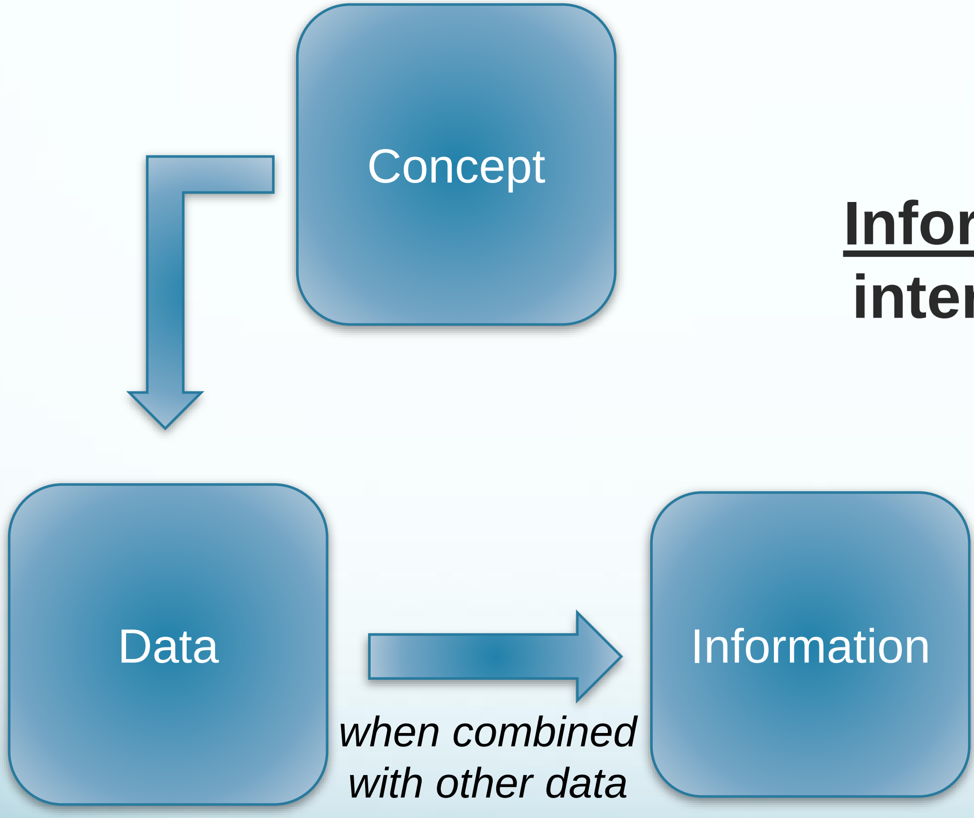


<u>Data</u> are facts used as a basis for reasoning, discussion or calculation

<u>Examples of health data:</u>

- Patient demographics
- Clinical signs and symptoms
- Medical, family and social histories
- Tests ordered and their results
- Problems and diagnoses
- Treatments provided
- Results of treatments
- Provider data

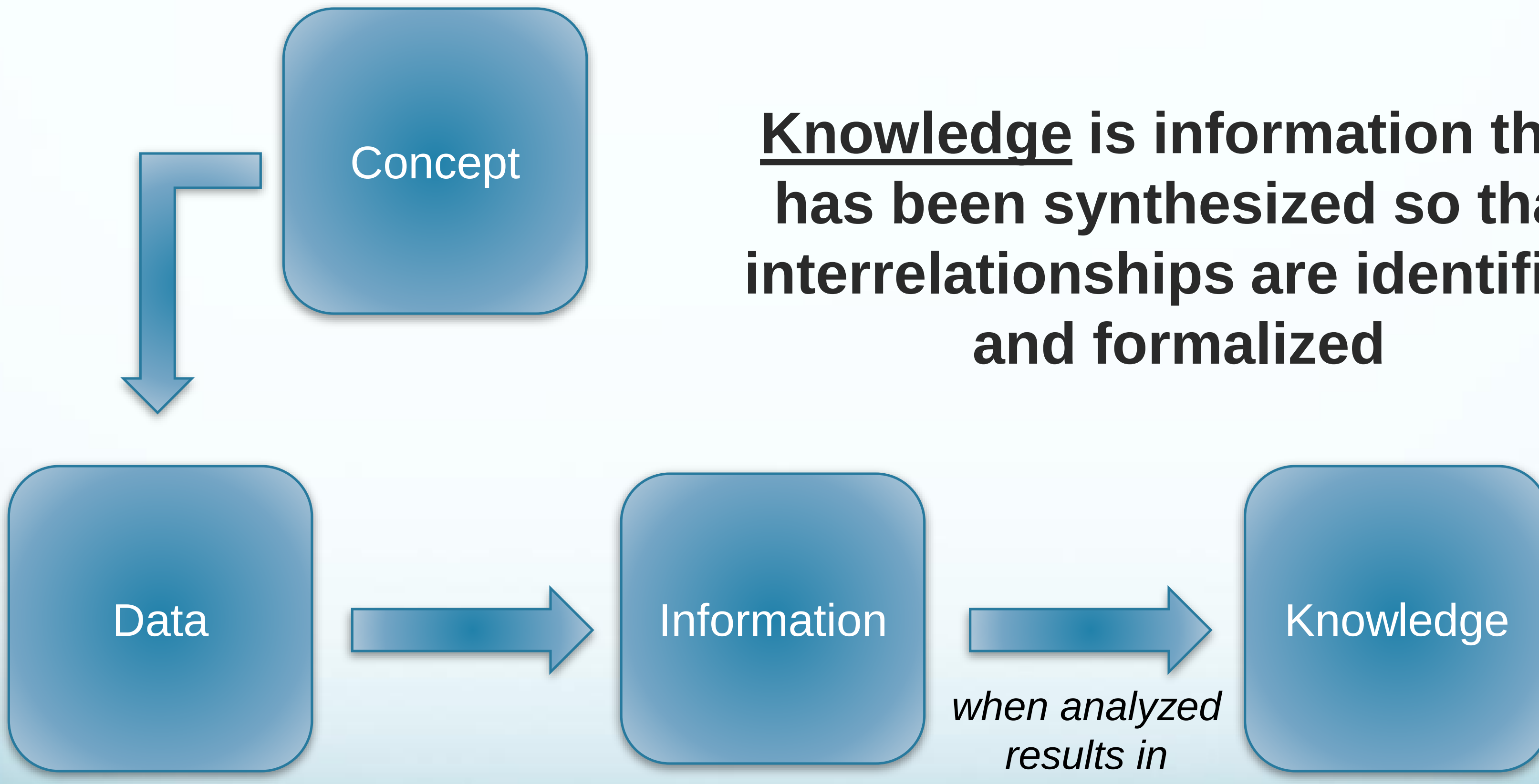
Got Intelligence?



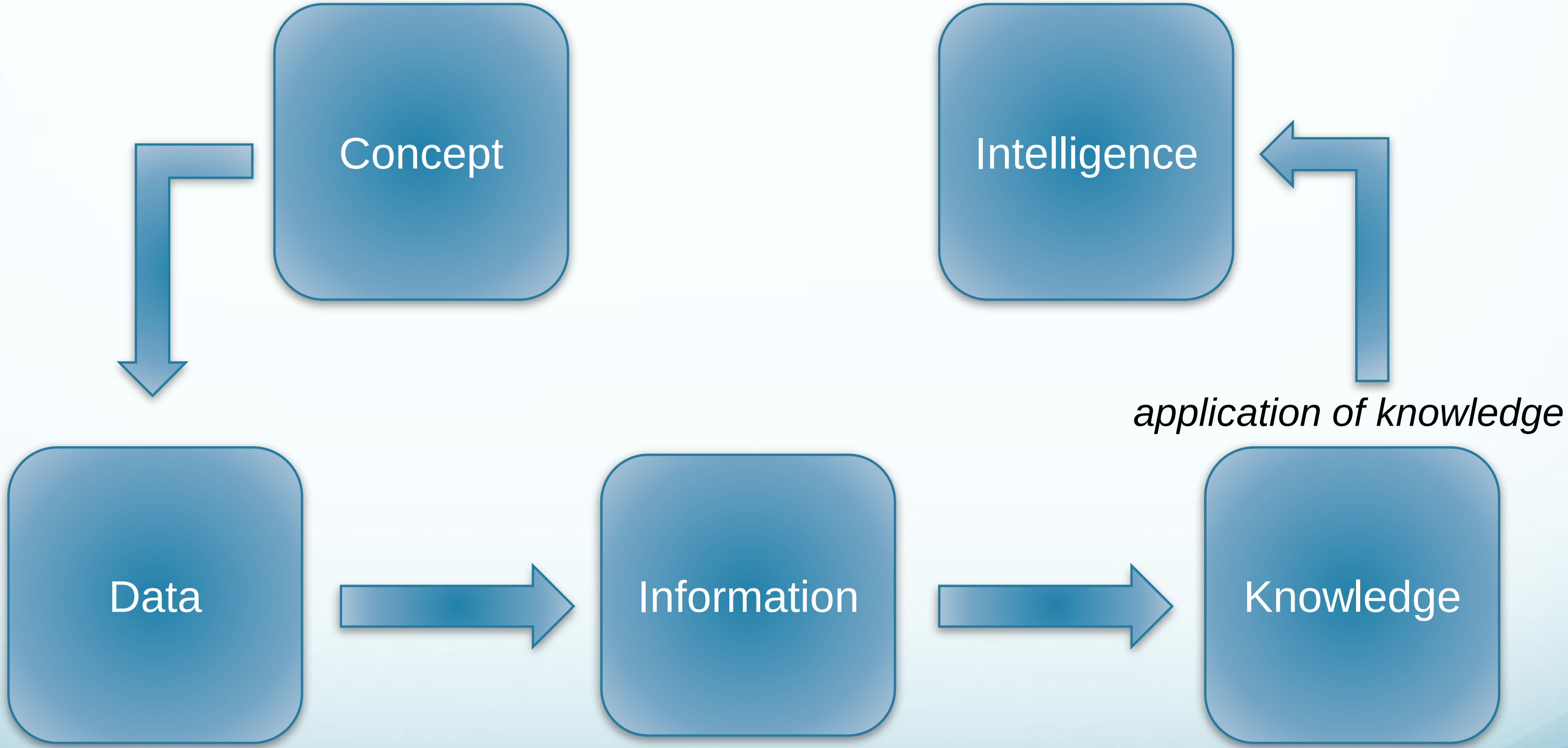
<u>Information</u> is data that are interpreted, organized and structured

Got Intelligence?

<u>Knowledge</u> is information that has been synthesized so that interrelationships are identified and formalized



Got Intelligence?




BIG DATA


I LOVE IT WHEN
YOU CALL ME BIG DATA
memegenerator.net


Semi-Structured
structured data
BIG
DATA
Volume
Useful
Metadata
AMP
Structured
VAST
Concept extraction
Text analytics
Semantic Metadata
database
Smart content
Framework
Zettabyte
Decision making
People driven
Unstructured
Data
analysis
content

What is ‘Big Data’ from the point of view of Pathology and Laboratory Medicine?

Papers in Press. Published October 20, 2015 as doi:10.1373/clinchem.2015.248591
The latest version is at http://hwmaint.clinchem.org/cgi/doi/10.1373/clinchem.2015.248591

Clinical Chemistry 61:12
000–000 (2015)

Q&A

"Big Data" in Laboratory Medicine

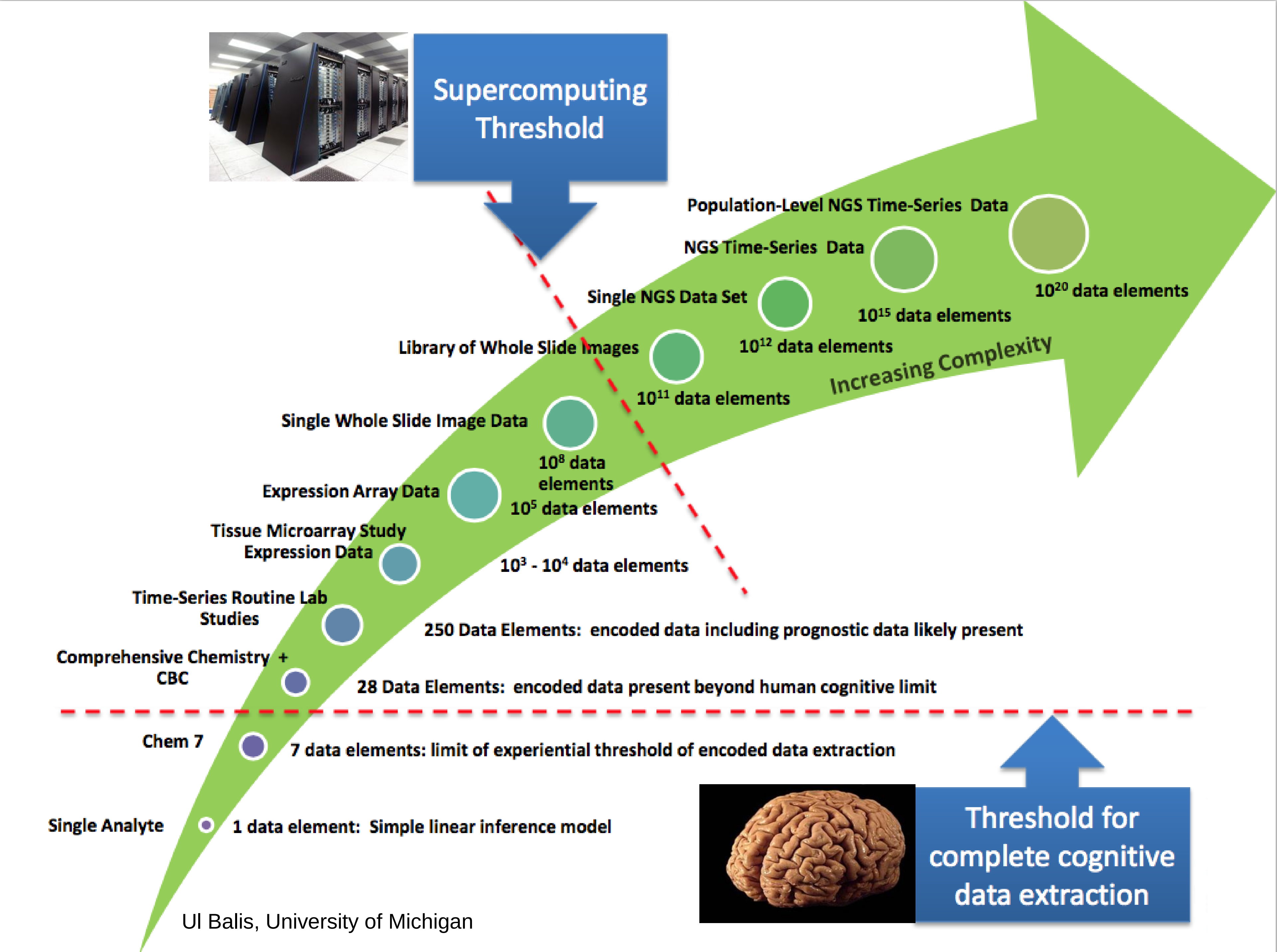
Moderators: Nicole V. Tolan[1*] and M. Laura Parnas[2]
Experts: Linnea M. Baudhuin,[3] Mark A. Cervinski,[4] Albert S. Chan,[5] Daniel T. Holmes,[6] Gary Horowitz,[7]
Eric W. Klee,[8] Rajiv B. Kumar,[9] and Stephen R. Master[10]

Beth Israel Deaconess Medical Center
Sutter Health Shared Laboratory
Mayo Clinic
Dartmouth-Hitchcock Medical Center
Sutter Health
St. Paul's Hospital, Vancouver
Stanford
New York Presbyterian

“Big Data” from our perspective

- “Anything that challenges an institution’s computational infrastructure”
  - “Too much information to process or store on a single computer”
- “I can’t do the analysis in Excel”
- “Any data set that challenges or exceeds an individual’s ability to manually evaluate all data points for clinical relevance”

Tolan and Parnes et al. *Clinical Chemistry* 2015;61:12


Supercomputing Threshold
Population-Level NGS Time-Series Data
NGS Time-Series Data
10^{20} data elements
Single NGS Data Set
10^{15} data elements
Library of Whole Slide Images
10^{12} data elements
Increasing Complexity
10^{11} data elements
Single Whole Slide Image Data
10^{8} data elements
Expression Array Data
10^{5} data elements
Tissue Microarray Study Expression Data
10^{3} - 10^{4} data elements
Time-Series Routine Lab Studies
250 Data Elements: encoded data including prognostic data likely present
Comprehensive Chemistry + CBC
28 Data Elements: encoded data present beyond human cognitive limit
Chem 7
7 data elements: limit of experiential threshold of encoded data extraction
Single Analyte
1 data element: Simple linear inference model
Threshold for complete cognitive data extraction
Ul Balis, University of Michigan

“Big Data” from our perspective

- “The kind of analytics that Google does”

Tolan and Parnes et al. *Clinical Chemistry* 2015;61:12

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]Google Inc. [2]Centers for Disease Control and Prevention

"One way to improve early detection is to monitor health-seeking behavior in the form of **online web search queries**, which are submitted by millions of users around the world each day. Here we present a method of analyzing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, **we can accurately estimate the current level of weekly influenza activity in each region of the United States**, with a reporting lag of about one day."

"Big Data" from our perspective

- "The kind of analytics that Google does"
- "Information we get from our large numbers of patients, samples and analytes in the lab"
  - "Includes all of the non-result data"
- "Assessment of massive amounts of information from multiple electronic sources in unison…to reveal otherwise unrecognized patterns"

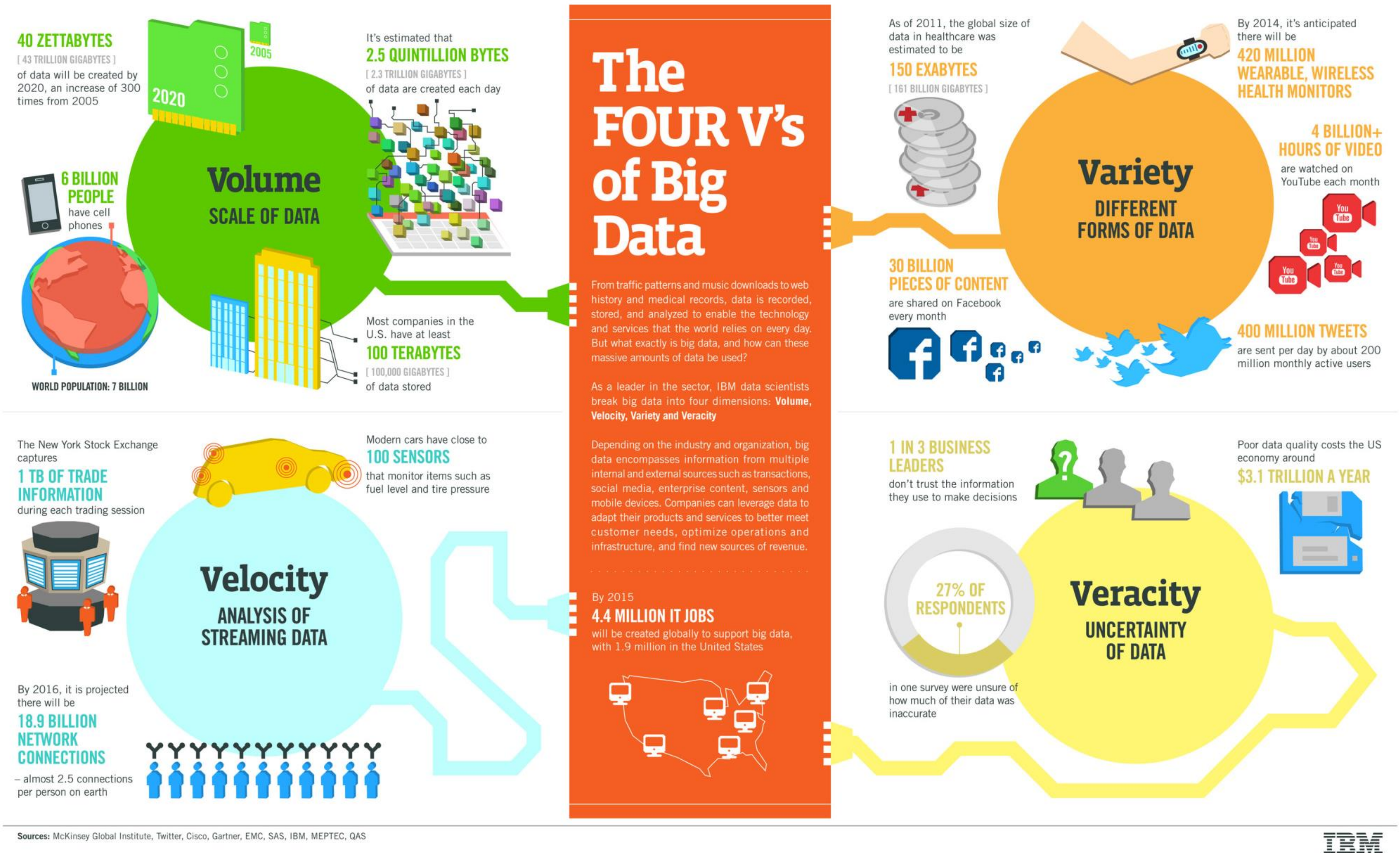
Tolan and Parnes et al. *Clinical Chemistry* 2015;61:12

Big Data Definition – 3 V’s

- “Big data” is high-Volume, -Velocity and –Variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making

  Doug Laney (Gartner) 2001

- The fourth ‘v’ is Veracity



The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
Volume
SCALE OF DATA
40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
Variety
DIFFERENT FORMS OF DATA
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

Other V's

- Variability – Not variety. Variability is when the meaning is changing (rapidly)
- Visualization – Challenging. How is data organized and analyzed in a manner that is easy to understand and read?
- Value – The data has to generate value. After all, the collection, storage and analysis of data costs a large amount of money. Data needs to be turned into intelligence

Big Data Definition

- “Big data” is high-Volume, -Velocity and –Variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making

Doug Laney (Gartner) 2001

Big Data Definition

- “Big data” is high-Volume, -Velocity and –Variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making

Doug Laney (Gartner) 2001

Computational Pathology

Computational Pathology

- Terms ‘computational’ and ‘pathology’ have been used together in literature as far back as 1980

- First defined by Fuchs and Buhmann in 2011

  “Computational Pathology investigates a complete probabilistic treatment of scientific and clinical workflows in general pathology.”

  “It combines experimental design, statistical pattern recognition and survival analysis within a unified framework to answer scientific and clinical questions in pathology.”

Fuchs and Buhmann. *Comput Med Imaging Graph* 2011;35:515-530

2014 Definition

"An approach to diagnosis that incorporates multiple sources of raw data; extracts biologically and clinically relevant information from these data; uses mathematic models at the molecular, individual and population levels to generate diagnostic inferences and predictions; and presents this clinically actionable knowledge to customers through dynamic and integrated reports and interfaces, enabling physicians, patients, laboratory personnel and other health care system stakeholders to make the best possible medical decisions."

Louis et al. *Archives of Pathology* 2014;138:1133-1138

2014 Definition

"An approach to diagnosis that incorporates multiple sources of raw data; extracts biologically and clinically relevant information from these data; uses mathematic models at the molecular, individual and population levels to generate diagnostic inferences and predictions; and presents this clinically actionable knowledge to customers through dynamic and integrated reports and interfaces, enabling physicians, patients, laboratory personnel and other health care system stakeholders to make the best possible medical decisions."

Louis et al. *Archives of Pathology* 2014;138:1133-1138

2014 Definition

"An approach to diagnosis that incorporates multiple sources of raw data; extracts biologically and clinically relevant information from these data; uses mathematic models at the molecular, individual and population levels to generate diagnostic inferences and predictions; and presents this clinically actionable knowledge to customers through dynamic and integrated reports and interfaces, enabling physicians, patients, laboratory personnel and other health care system stakeholders to make the best possible medical decisions."

Louis et al. *Archives of Pathology* 2014;138:1133-1138

2014 Definition

"An approach to diagnosis that incorporates multiple sources of raw data; extracts biologically and clinically relevant information from these data; uses mathematic models at the molecular, individual and population levels to generate diagnostic inferences and predictions; and presents this clinically actionable knowledge to customers through dynamic and integrated reports and interfaces, enabling physicians, patients, laboratory personnel and other health care system stakeholders to make the best possible medical decisions."

Louis et al. *Archives of Pathology* 2014;138:1133-1138

2014 Definition

"An approach to diagnosis that incorporates multiple sources of raw data; extracts biologically and clinically relevant information from these data; uses mathematic models at the molecular, individual and population levels to generate diagnostic inferences and predictions; and presents this clinically actionable knowledge to customers through dynamic and integrated reports and interfaces, enabling physicians, patients, laboratory personnel and other health care system stakeholders to make the best possible medical decisions."

Louis et al. *Archives of Pathology* 2014;138:1133-1138

2014 Definition

"An approach to diagnosis that incorporates multiple sources of raw data; extracts biologically and clinically relevant information from these data; uses mathematic models at the molecular, individual and population levels to generate diagnostic inferences and predictions; and presents this clinically actionable knowledge to customers through dynamic and integrated reports and interfaces, enabling physicians, patients, laboratory personnel and other health care system stakeholders to make the best possible medical decisions."

Louis et al. *Archives of Pathology* 2014;138:1133-1138

Potential Benefits

1. Minimize surprise of diseases
   a. Likelihood of disease before onset
   b. Disease trend before complications
2. Improve quality and efficiency of care
3. Hub for data-related research
4. Better selection of patients for clinical trials

Comp Path Use-Cases

- Molecular Pathology
- Test utilization
- Hematology
- Surgical Pathology
- Microbiology/Infectious disease
- Visualization of data/SAGE

Test Utilization

- Study of patients in Calgary over 1 year
- Six common lab tests (cholesterol, Hb A1C, TSH, vitamin B12, vitamin D and ferritin)
- Looked at repeat rates at 3, 6 and 12 months
- Found 16% inappropriately repeated tests
- Excess costs of $2.2 million (Canadian)
- This could be built as rules in CPOE systems

Morgan and Naugler. *AJCP* 2015;144:704-712

Hematology

- Analyzing populations of CBC data:

1. Predict the onset of anemia before patients have clinical disease

Higgins and Mahadevan. *Proc Natl Acad Sci* 2010 ;23;107(47):20587-92

1. Identify myelodysplastic syndrome in unselected outpatient populations

Raess et al. *Am J Hematol*. 2014;89(4):369–374

Surgical Pathology

- Computational image analysis to differentiate breast intraductal proliferative lesions
  - Differentiate UDH from DCIS
  - Stratify nuclear grade in DCIS
- 116 breast bxs to build classification models
- Applied to 51 breast biopsies
  - AUC of 0.86 for differentiation
  - AUC of 0.98 for low grade v. high grade

Dong et al. *PLOS One* 2014;9(12):e114855

Antibiogram

- Antibiograms
  - Microorganism specific
  - Single antibiotic
  - Not patient specific
- Use computational pathology and informatics to change frame of reference
  - from will this drug work for this bug
  - to will this regimen work for this patient

Weighted Incidence Syndromic Combination Antibiogram

WISCA

1. Reorganize from organism-antibiotic table to patient-infection-regimen

Organism	Antibiotic 1
E. coli	R
E. coli	S
E. coli	S
E. coli	S
E. coli	S
E. coli	S
E. coli	S
E. coli	S
E. coli	S

Patient	Syndrome	Regimen 1
1	ABI	1
2	ABI	0
3	ABI	0
4	ABI	0
5	ABI	0
6	ABI	1
7	ABI	0
8	ABI	1
9	ABI	0

Ari Robicsek, North Shore University Health, Chicago

Weighted Incidence Syndromic Combination Antibiogram

WISCA

1. Reorganize from organism-antibiotic table to patient-infection-regimen
2. Create a model that uses clinical and previous test data to predict the probability of coverage of each regimen
3. Input new patient's characteristics and WISCA engine outputs probabilities for regimens
4. Models automatically recalculated based on continuously collected data

Ari Robicsek, North Shore University Health, Chicago

WISCA for UTI

Characteristics	Base case	patient 1	patient 2	patient 3	patient 4	patient 5	patient 6	patient 7	patient 8
Age	45	45	75	75	75	75	75	75	75
Gender	F	M	F	F	F	F	F	F	F
Hospital	1	1	1	4	1	1	1	1	1
≥4 positive UCx in prior yr	no	no	no	no	yes	no	no	no	no
Albumin <2.5 g/dL	no	no	no	no	no	yes	no	no	no
≥2 recent hospitalizations[a]	no	no	no	no	no	no	yes	yes	yes
Cephalosporins in last 30 d	no	no	no	no	no	no	no	yes	no
FQ in last 30 d	no	no	no	no	no	no	no	no	yes
Antibiotic Regimens	Calculated probability of Coverage								
TMP-SMX	71%	61%	73%	63%	60%	71%	75%	66%	56%
Cefazolin	77%	60%	81%	78%	78%	75%	79%	66%	61%
Ertapenem	84%	66%	85%	90%	81%	80%	79%	69%	60%
Ciprofloxacin	85%	80%	85%	72%	77%	82%	88%	90%	56%
Ceftriaxone	85%	67%	86%	85%	84%	82%	83%	74%	65%
Ceftazidime	87%	75%	89%	87%	88%	85%	84%	80%	66%
Ampicillin + Gentamicin	93%	91%	93%	90%	89%	91%	94%	93%	87%
Ampicillin-Sulbactam	94%	85%	95%	95%	92%	91%	95%	90%	88%
Ceftazidime+ vancomycin	96%	94%	96%	95%	95%	94%	96%	94%	89%
Pip-Tazo	97%	95%	96%	97%	95%	94%	95%	93%	88%
Meropenem	98%	95%	97%	98%	96%	95%	96%	95%	90%

Ari Robicsek, North Shore University Health, Chicago

Visualizing BIG Data

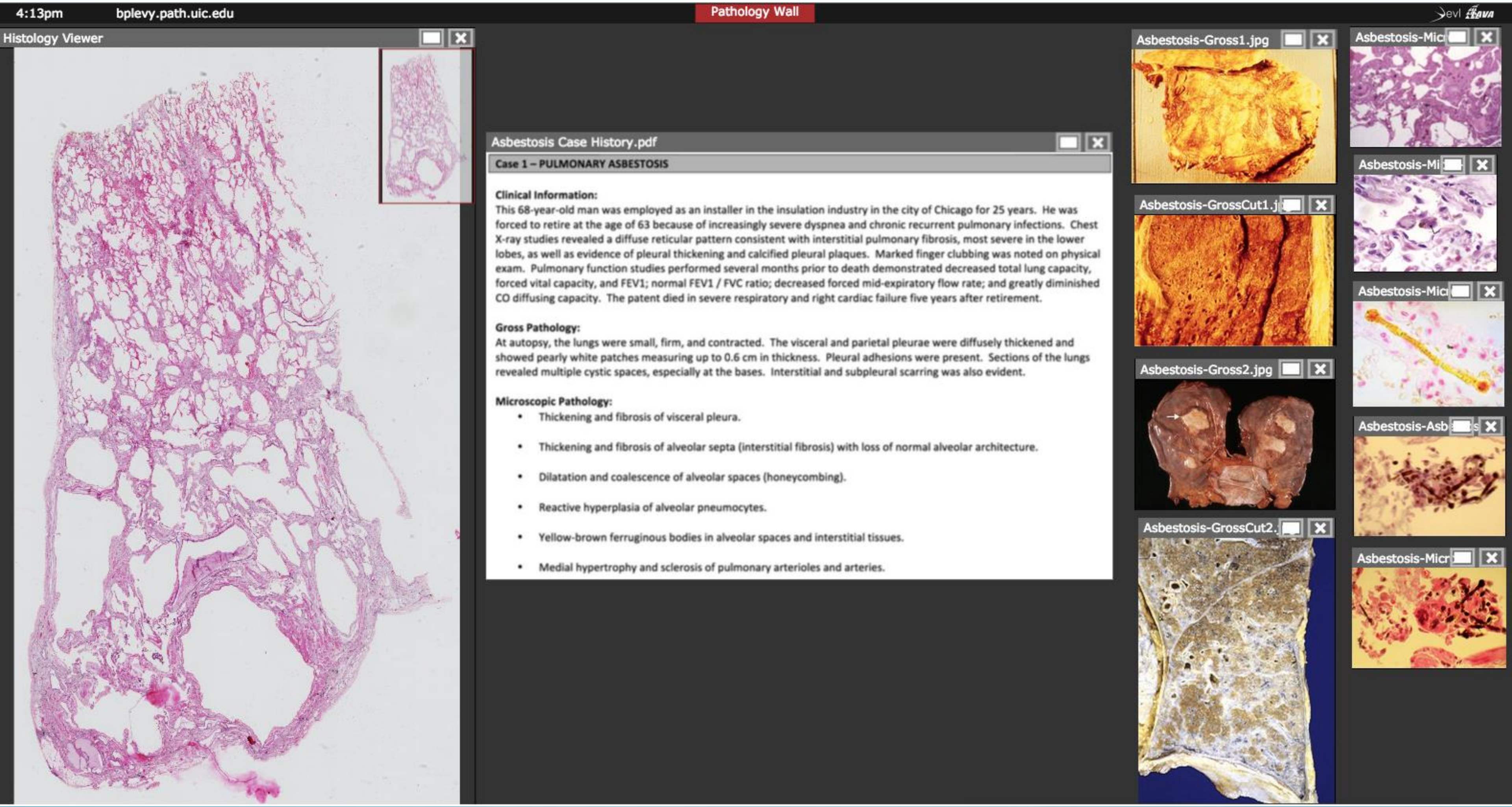
Including Whole-Slide Images

Scalable Adaptive Graphics Environment

- Access, display, share and collaborate in real time
  - Data-intensive information
  - Variety of resolutions and formats
  - Multiple simultaneous user input
  - Multiple drivers, no passengers
- Cloud based and web browser enabled
- Open source

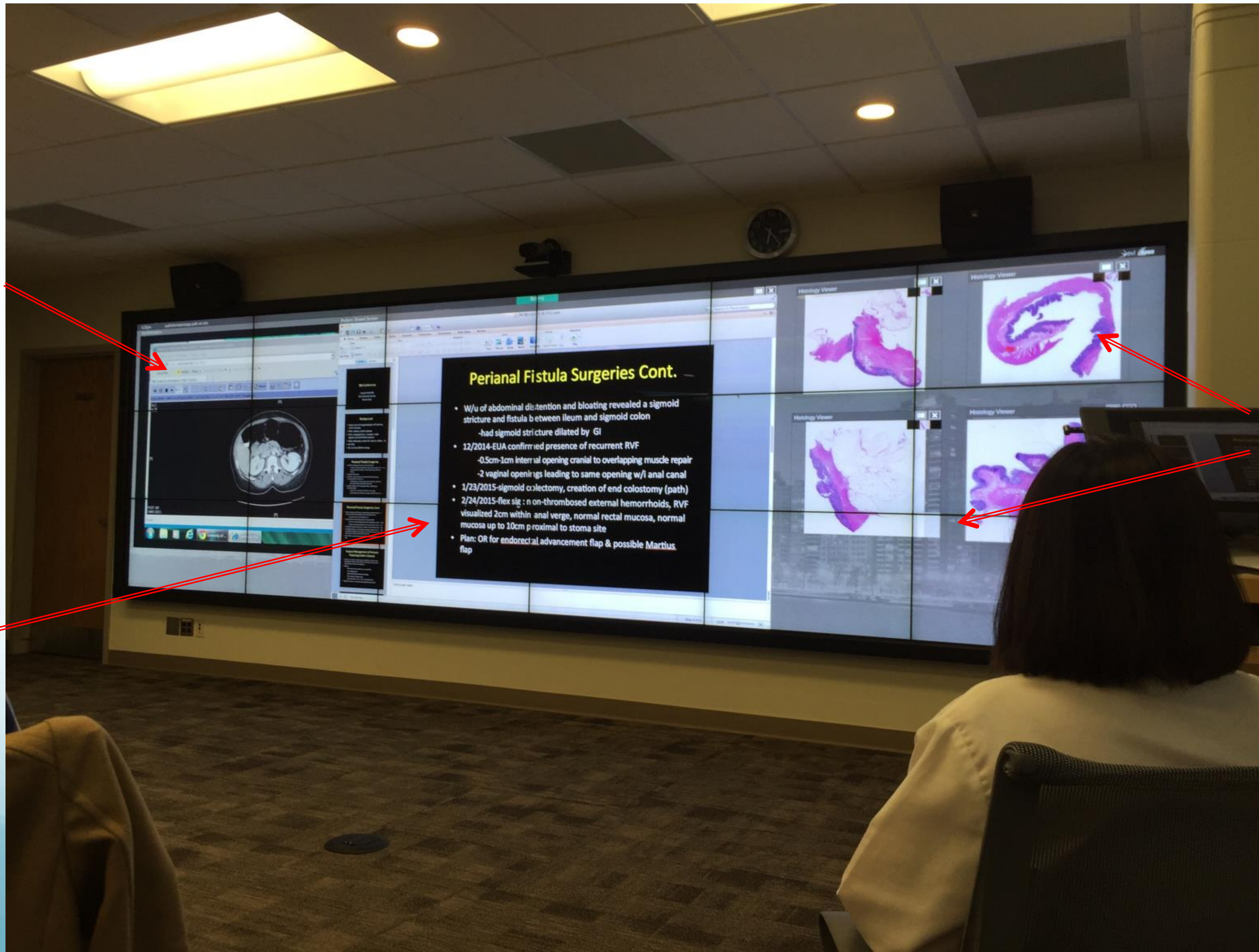
electronic visualization laboratory

Medical Education



Multidisciplinary Case Conference

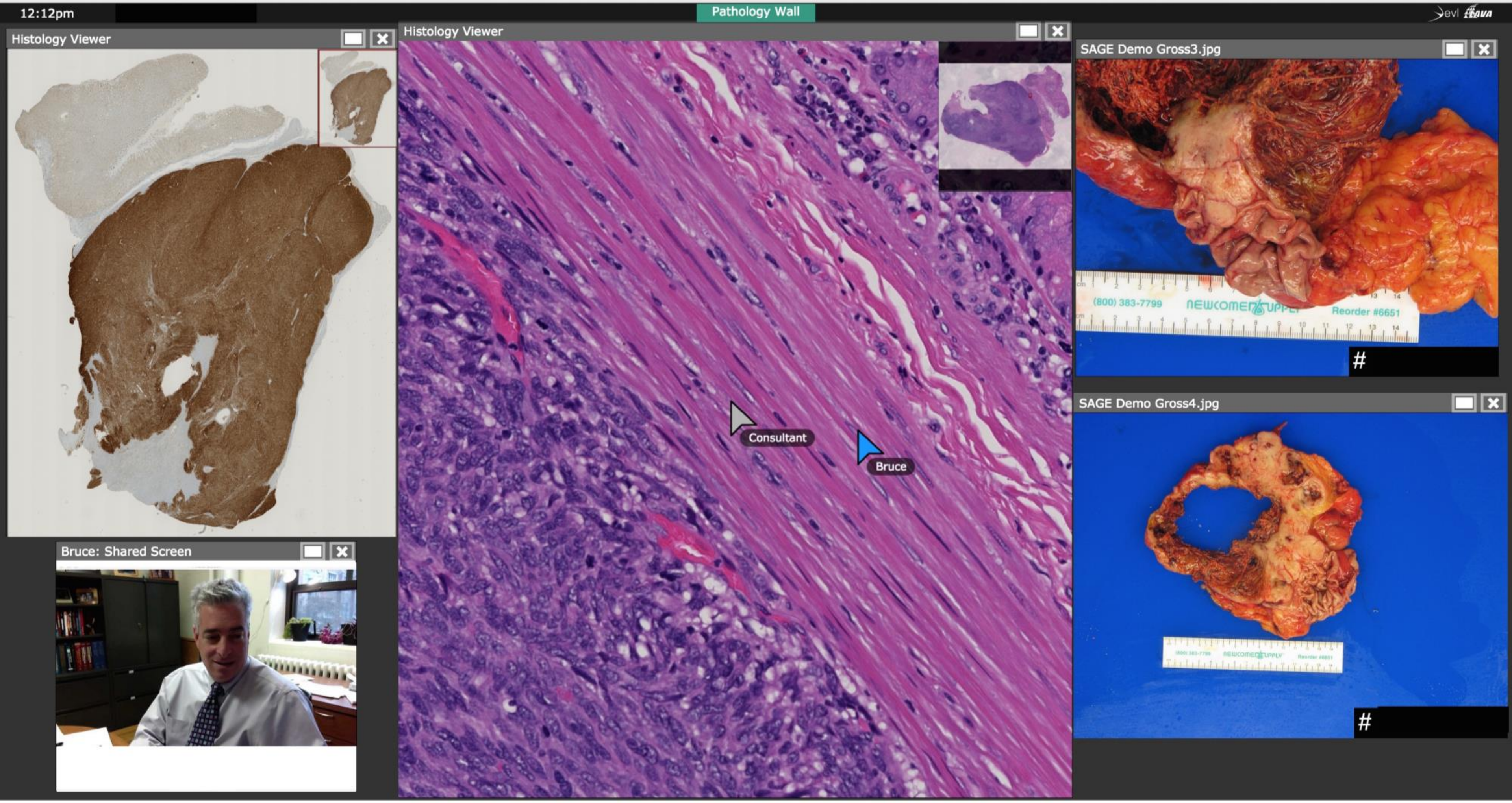
Radiology PACS

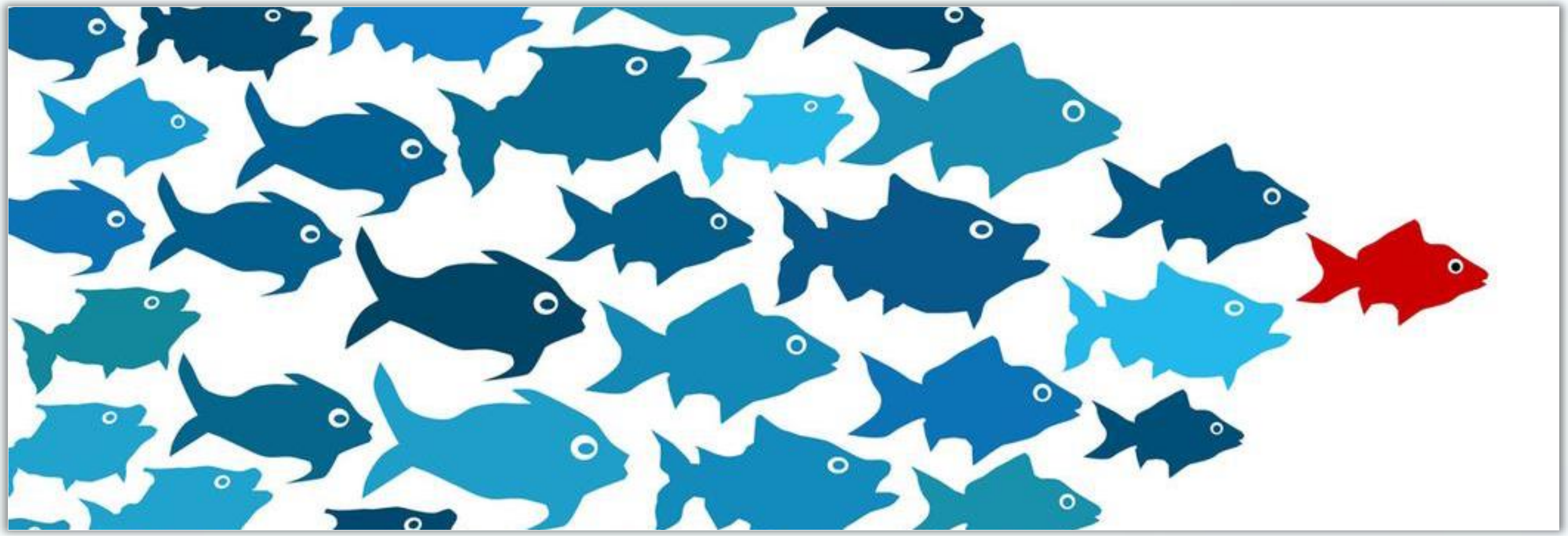


WSIs

Clinical Info

Medical Consultation





Leadership of Computational Medicine

We need to recognize and assume our proper place in the leadership of health care

"All Physicians are Leaders"

American Association for PHYSICIAN LEADERSHIP
Inspiring Change. Together.

We need to recognize and assume our proper place in the leadership of health care

70% Rule

“All Physicians are Leaders”

American Association for PHYSICIAN LEADERSHIP
Inspiring Change. Together.

We need to recognize and assume our proper place in the leadership of health care

70% Rule

"All Pathologists are Leaders"

Operational Leadership

- Electronic Health Record
  - Participate in EHR selection/implementation
    - Leverage experiences with LIS vendors
  - Be involved with EHR committees and groups
  - Develop EHR expertise among lab professionals
- Clinical Decision Support and test utilization
- Document and communicate lab's contributions and successes
- Institutional Leadership

Educational Leadership

Undergraduate and Graduate Medical Education

Undergraduate Med Ed

- Evolution of undergraduate medical education from the Flexner model is an opportunity for pathology, not a challenge
- Pathologists should lead sections of the systems-based model of medical school education
- Required lab medicine rotation
- Informatics education for medical students

PIER: Model for all Residents

PIER Scope and Sequence

E - 1

PIER Essentials 1

- Informatics in Pathology Practice
- Information Systems Fundamentals
- Importance of Databases
- Introduction to Data Standards
- Data Availability & Security

Entry-Level Proficiency
ACGME Milestone Level 1
Instructional Hours: 4-6

E - 2

PIER Essentials 2

- LIS Components & Functions
- Specialized LISs & Middleware
- Data & Communication Standards
- Digital Imaging
- Basics of the Health Care Information Ecosystem

Basic Proficiency
ACGME Milestone Level 2
Instructional Hours: 8-10

Pathology
Informatics
Essentials
for
Residents

E - 3

PIER Essentials 3

- Pathologist Role in LIS & EHR Projects
- LIS Installation & Configuration
- Information Systems & Laboratory Performance
- Data Security, Regulatory & Accreditation Requirements

Intermediate Proficiency
ACGME Milestone Level 3
Instructional Hours: 10-12

E - 4

PIER Essentials 4

- LIS Management & Oversight
- Order and Results Management
- Laboratory Data for Quality Improvement & Research
- Laboratory Data & Enterprise Health Care Analytics

Advanced Proficiency
ACGME Milestone Level 4
Instructional Hours: 10-14

Clinical Informatics Fellowships

Summary

- Intelligence in health care depends on big data
- Computational Medicine is heavily reliant on Pathology and Informatics
- Leadership of Computational Medicine is ours if we are willing to work for it
- Leadership requires us to be willing to look beyond the walls of our labs and understand our critical role in the care of patients and populations


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
BIYAN
SHUKRIA
TINGKI
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
MEHRBANI
GRAZIE
MAAKE
KOMAPSUMNIDA
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
PALDIES
BOLZIN
MERCI
SPASSIBO
SNACHALHUYA
NUHUN
CHALTU
WABEEJA
MAITEKA
SPASIBO
DENKAUJA
NENACHALHYA
SIKOMO
MINMONCHAR
MERASTAWHY
GAEJTHO
TAVTAPUCH
MEDAWAGSE
AGUYJE
FAKAAUE
LAH
ATTO
HUI
GUI
HATUR
UNALCHEESH
EKOJU
MAKETAI

Questions?