Grand Challenges of Measurement Science with Big Data

Presenter: Peter Bajcsy

Software and Systems Division

Information Technology Laboratory

National Institute of Standards and Technology

Goal & Objective

- **Goal:** to articulate "Grand Challenges" in data-intensive research
- **Objective:** to identify differences between measurement science for Little Data and Big Data

Concrete Application Domain

- **Analysis of cell biology microscopy images**

- **NIST Project: Computational Science in Biological Metrology**
  - **CS-BIO-MET: nist.gov/itl/ssd/is/computational-science-in-biometrology.cfm**

Basic Biological Questions

- What are the quantitative dynamic characteristics of cell changes as a function of their surrounding environment, cell signaling, phenotype and genotype?
  - Parameter estimation
  - Bayesian inference and learning methods
  - Development of mathematical models

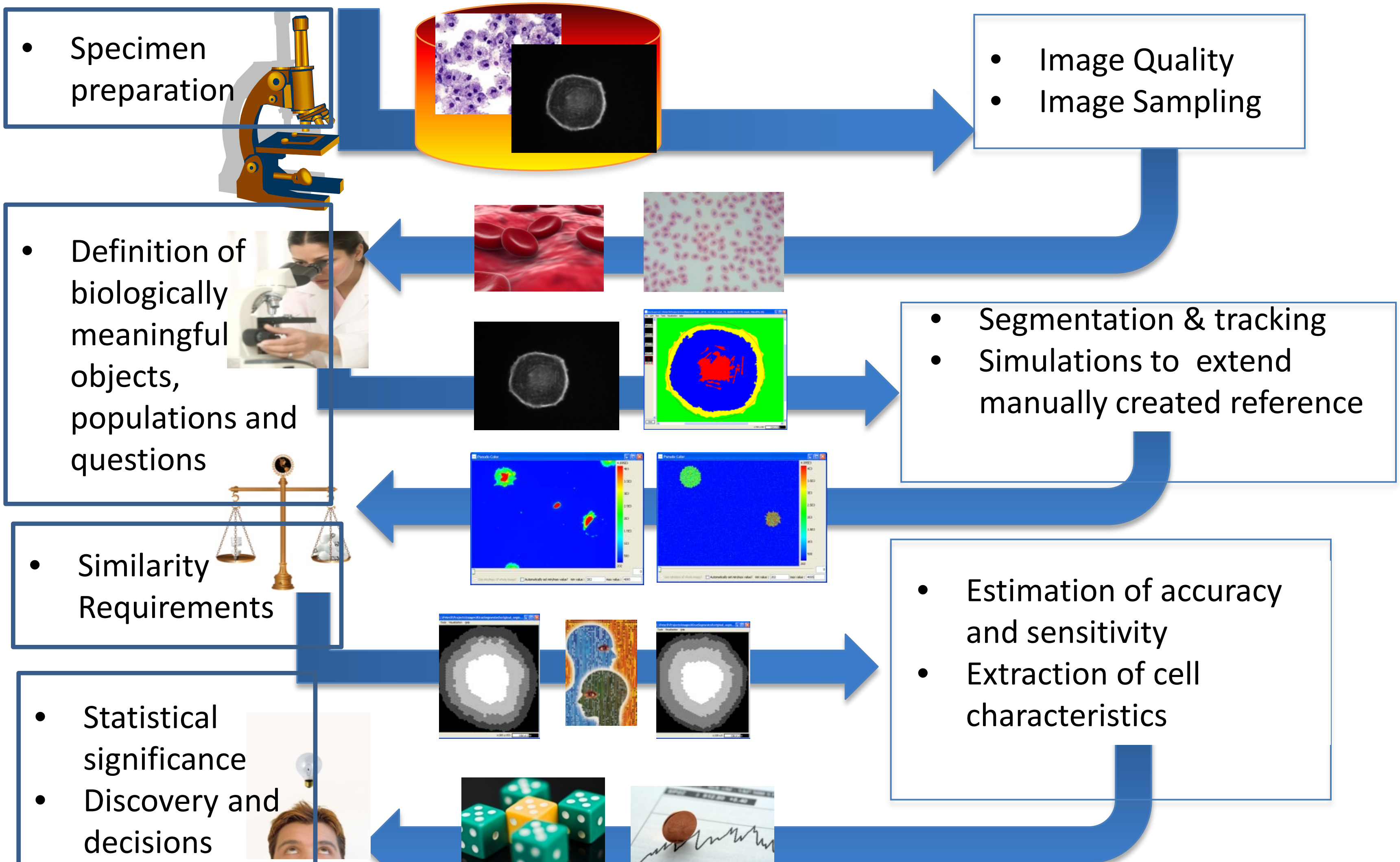
Mission Oriented Results

- Big Data → **Limited human input** → Need for automation and measurements of automation accuracy
- Automation Outcome
  - → Quantitative cell measurements
  - → Cost reduction
  - → New scientific discovery
  - → Avoid missed opportunities

Computational Science in Biological Metrology

Biological Metrology

Computational Science



- Specimen preparation

- Image Quality
- Image Sampling

- Definition of biologically meaningful objects, populations and questions

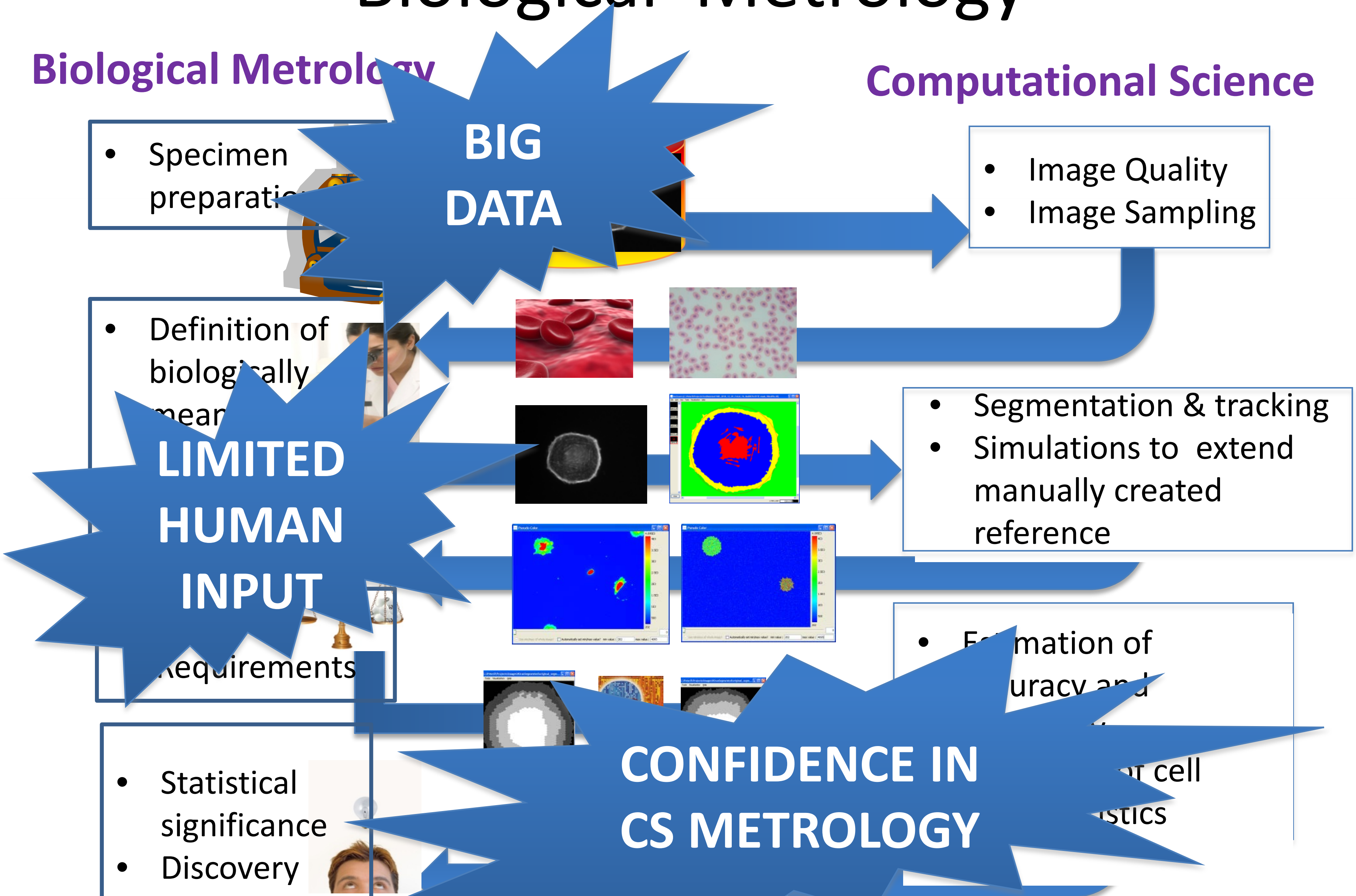
- Segmentation & tracking
- Simulations to extend manually created reference

- Similarity Requirements

- Estimation of accuracy and sensitivity
- Extraction of cell characteristics

- Statistical significance
- Discovery and decisions

Computational Science in Biological Metrology



Biological Metrology

Computational Science

- Specimen preparati

- Definition of biologi ally mean

Requirements

- Statistical significance
- Discovery

- Image Quality
- Image Sampling

- Segmentation & tracking
- Simulations to extend manually created reference

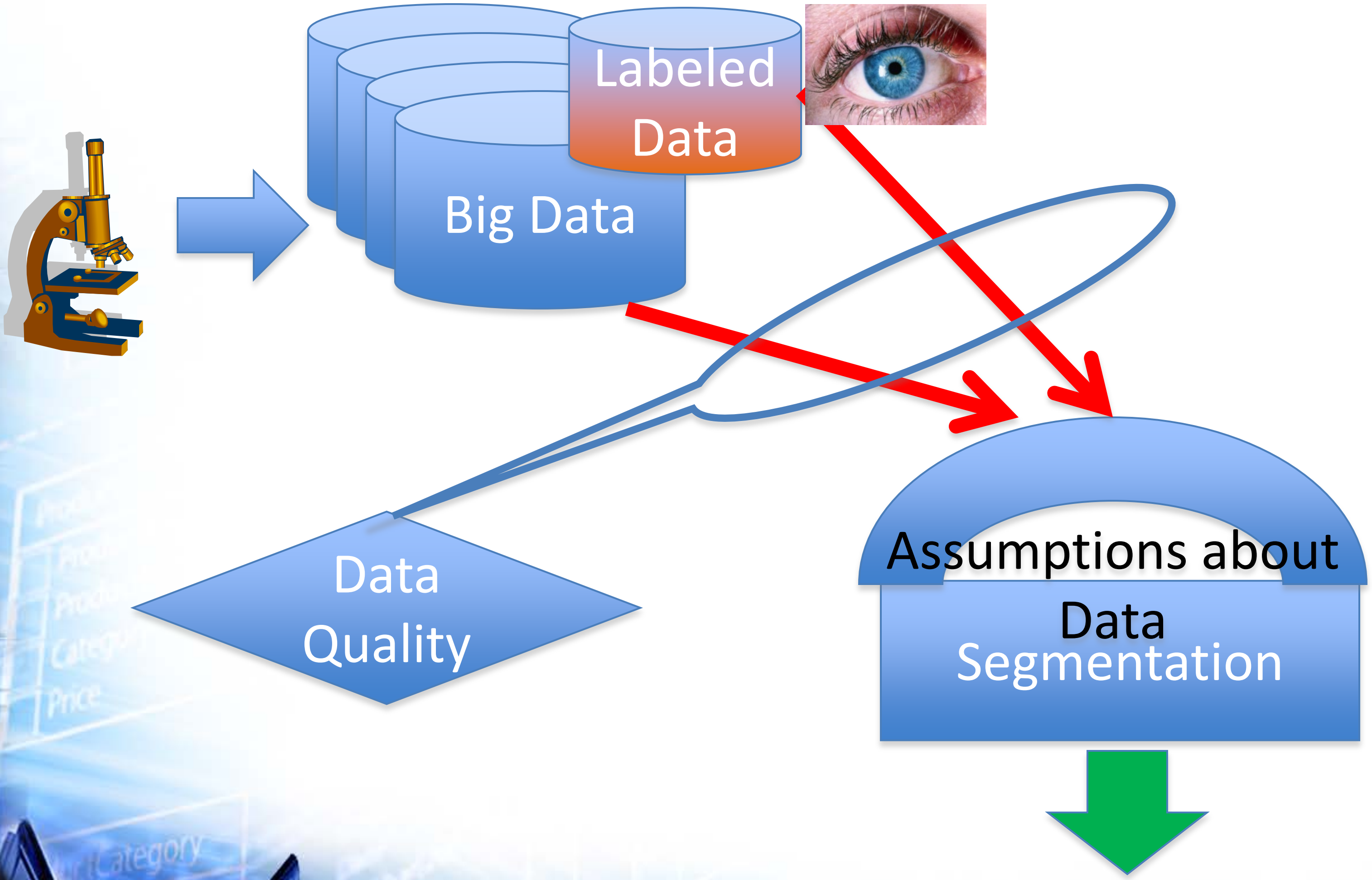
- E mation of uracy and f cell stics

BIG DATA

LIMITED HUMAN INPUT

CONFIDENCE IN CS METROLOGY

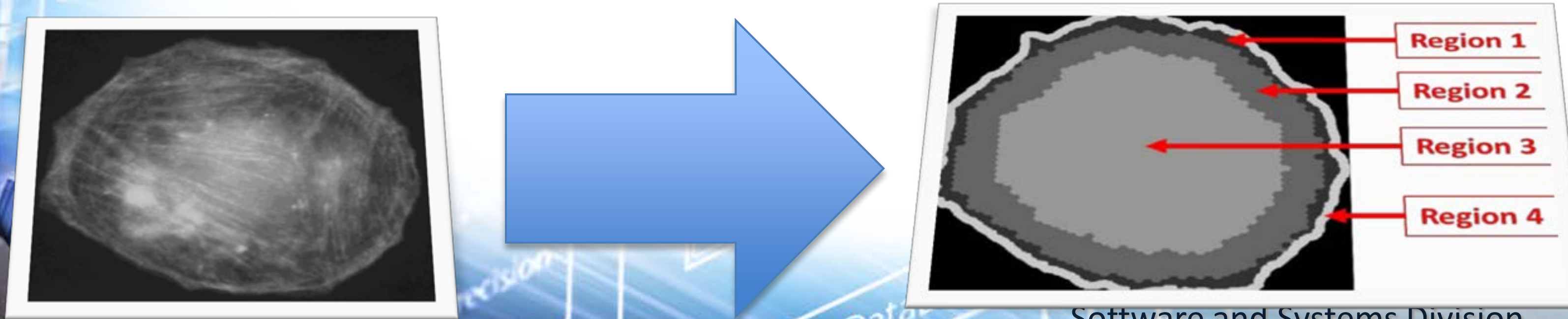
Central Problem:
Image Segmentation Accuracy



ACCURACY/ERROR

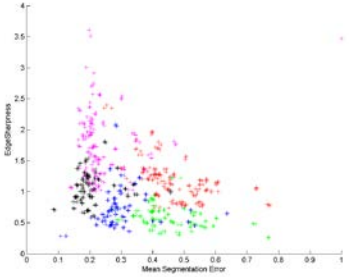
Image Segmentation Baseline

- What is the baseline for segmentation results in cell microscopy?
  - Disagreement between biologists, statisticians, and computer vision researchers on the baseline criteria
  - Manual versus mathematically grounded approach

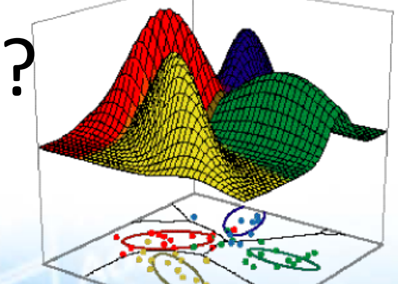


Data Quality

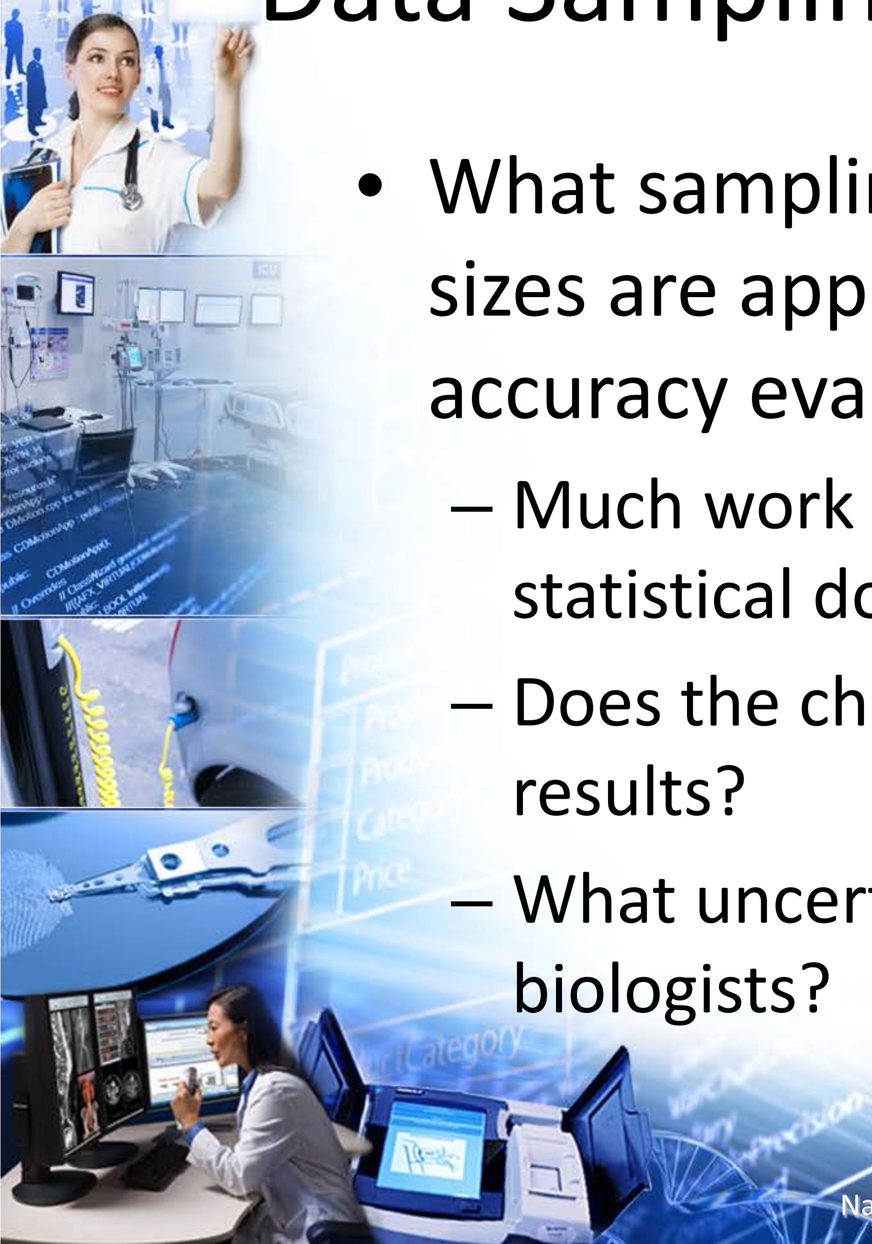
- How does data quality relate to segmentation accuracy?
- What data quality methods are appropriate for driving optimal microscopy settings?



- How does one detect and measure a mismatch between segmentation method assumptions and the input data?



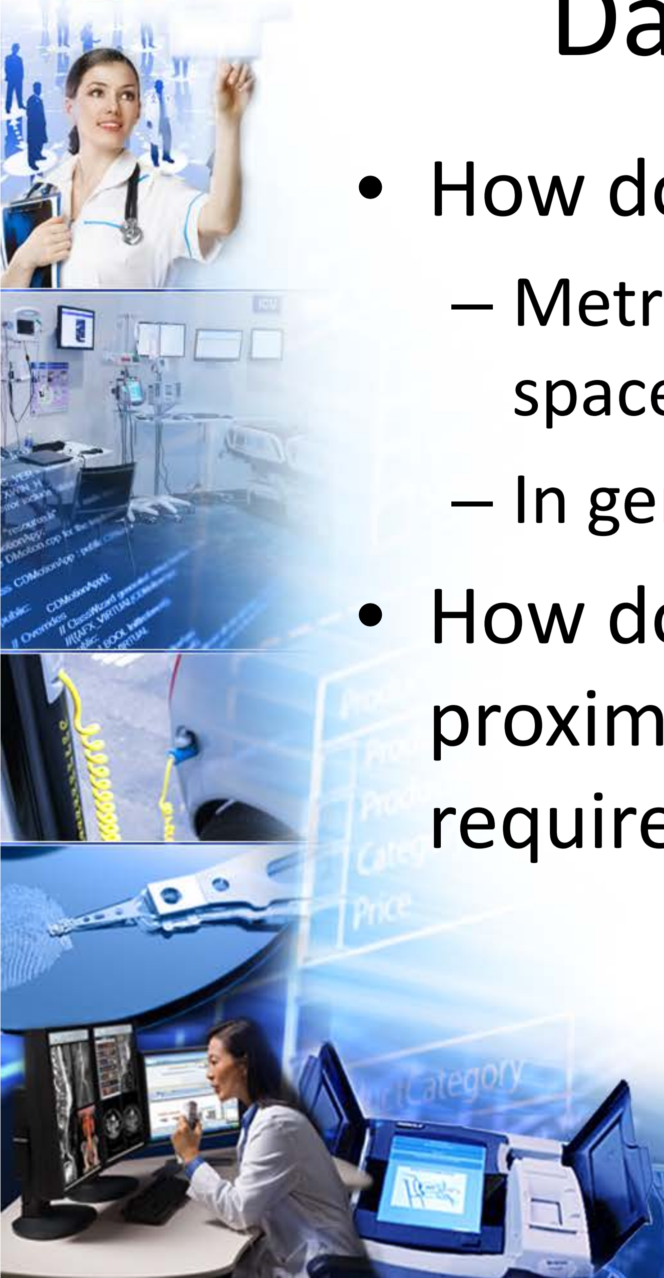
Data Sampling (Image or Cell)

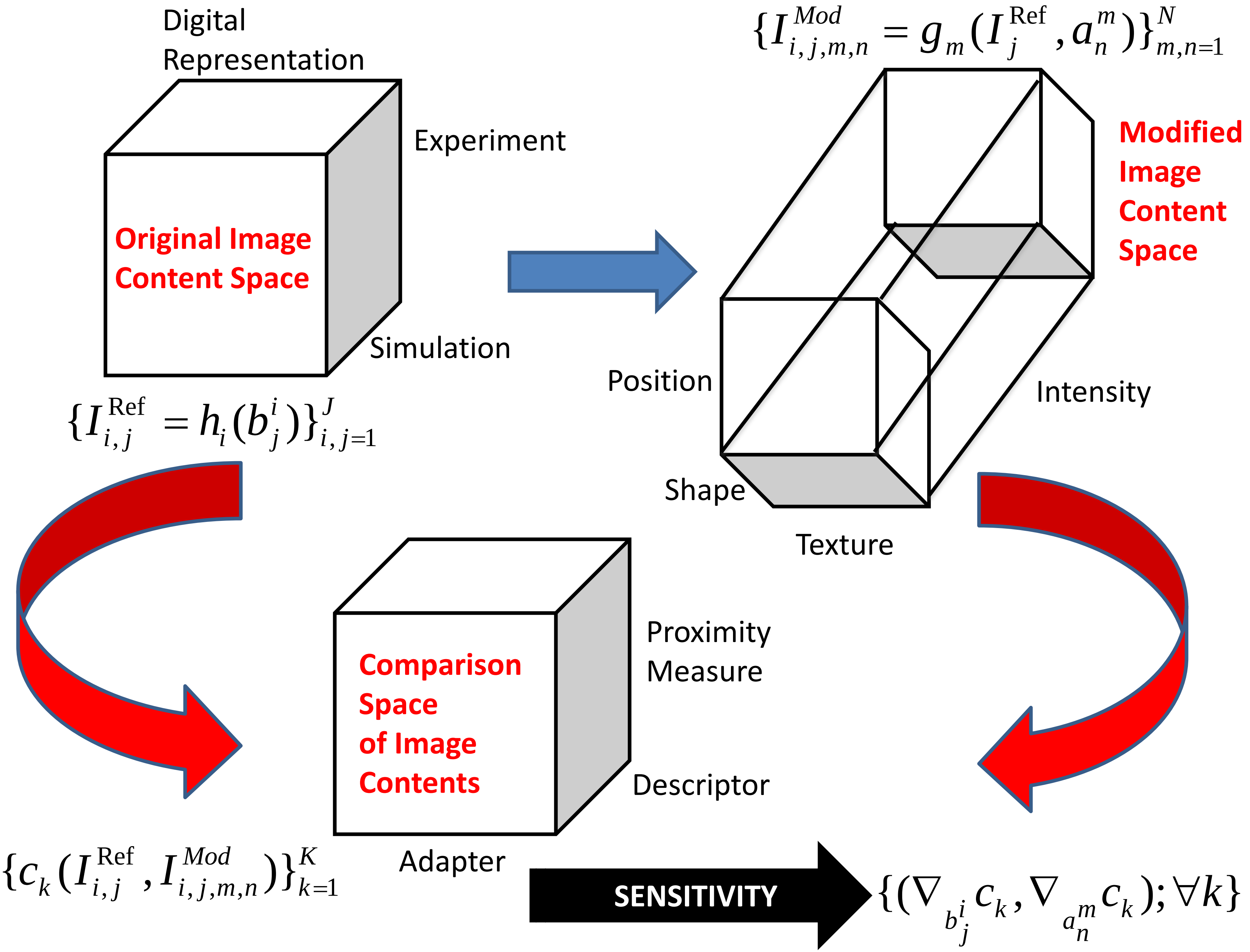


- What sampling techniques and sample sizes are appropriate for segmentation accuracy evaluations of Big Data?
  - Much work in the signal processing and statistical domains
  - Does the choice of sampling method bias results?
  - What uncertainty is acceptable to biologists?

Data Comparison

- How does one compare two images?
  - – Metrics: Euclidean versus Riemannian spaces
  - – In general, any two digital data sets?
- How does one choose the most suitable proximity metrics given application requirements?





Digital Representation

Experiment

Original Image Content Space

Simulation

$\{I^{\mathrm{Ref}}_{i,j} = h_i(b^i_j)\}^J_{i,j=1}$

$\{I^{Mod}_{i,j,m,n} = g_m(I^{\mathrm{Ref}}_j, a^m_n)\}^N_{m,n=1}$

Modified Image Content Space

Position

Intensity

Shape

Texture

Comparison Space of Image Contents

Proximity Measure

Descriptor

Adapter

$\{c_k(I^{\mathrm{Ref}}_{i,j}, I^{Mod}_{i,j,m,n})\}^K_{k=1}$

SENSITIVITY

$\{(\nabla_{b^i_j} c_k, \nabla_{a^m_n} c_k); \forall k\}$

Sensitivity Signatures of Similarity Metrics

$$\vec{\mathrm{s}}_{\mathrm{k}} = \left(\frac{\delta c_k}{\delta I_{i,j}^{\mathrm{Ref}}}, \frac{\delta c_k}{\delta I_{i,j,m,n}^{\mathrm{Mod}}} \right)_{(I_{i,0}^{\mathrm{Ref}}; I_{i,0,m,0}^{Mod})}$$

$$\frac{\delta c_k}{\delta I_{i,j}^{\mathrm{Ref}}} = \delta\Phi[c_k(I_{i,j}^{\mathrm{Ref}}, \delta I_{i,j,m,n}^{\mathrm{Mod}}); I_{i,j}^{\mathrm{Ref}}]$$

$$\frac{\delta c_k}{\delta I_{i,j,m,n}^{\mathrm{Mod}}} = \delta\Phi[c_k(I_{i,j}^{\mathrm{Ref}}, \delta I_{i,j,m,n}^{\mathrm{Mod}}); I_{i,j,m,n}^{\mathrm{Mod}}]$$

The differentials are the 1st Frechet derivative of a similarity metric C_k along the direction of reference image content changes (application dependent space of reference image content) and along the direction of any possible modification of reference image content (application task specific) at the referenced point $(I_{i,0}^{\mathrm{Ref}}; I_{i,0,m,0}^{Mod})$

Similarity Metric Recommendation

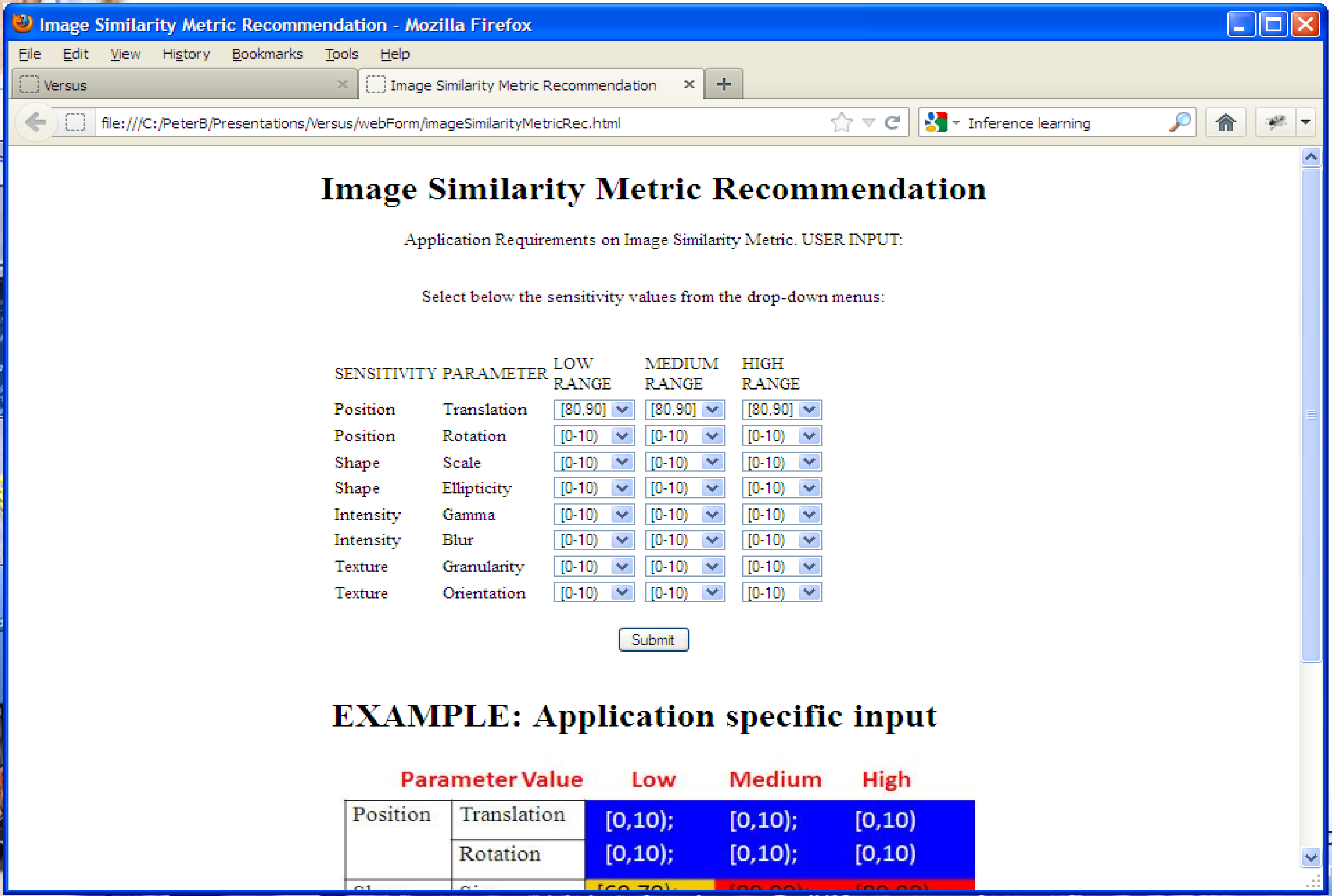


Image Similarity Metric Recommendation - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Versus
Image Similarity Metric Recommendation
file:///C:/PeterB/Presentations/Versus/webForm/imageSimilarityMetricRec.html
Inference learning

Image Similarity Metric Recommendation

Application Requirements on Image Similarity Metric. USER INPUT:

Select below the sensitivity values from the drop-down menus:

SENSITIVITY	PARAMETER	LOW RANGE	MEDIUM RANGE	HIGH RANGE
Position	Translation	[80,90]	[80,90]	[80,90]
Position	Rotation	[0-10)	[0-10)	[0-10)
Shape	Scale	[0-10)	[0-10)	[0-10)
Shape	Ellipticity	[0-10)	[0-10)	[0-10)
Intensity	Gamma	[0-10)	[0-10)	[0-10)
Intensity	Blur	[0-10)	[0-10)	[0-10)
Texture	Granularity	[0-10)	[0-10)	[0-10)
Texture	Orientation	[0-10)	[0-10)	[0-10)

Submit

EXAMPLE: Application specific input

Parameter Value		Low	Medium	High
Position	Translation	[0,10);	[0,10);	[0,10)
	Rotation	[0,10);	[0,10);	[0,10)

Access, Access, Access

- What is the most efficient Big Data access and retrieval protocol?
  - RESTful web services, Web applications, …

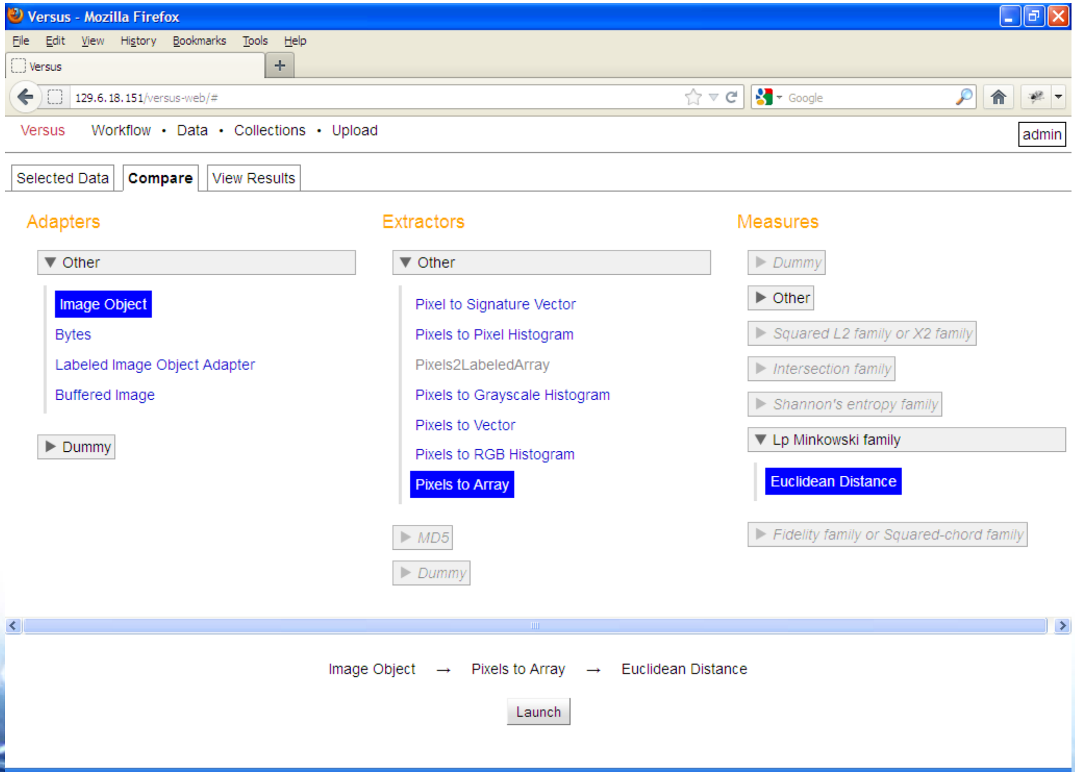
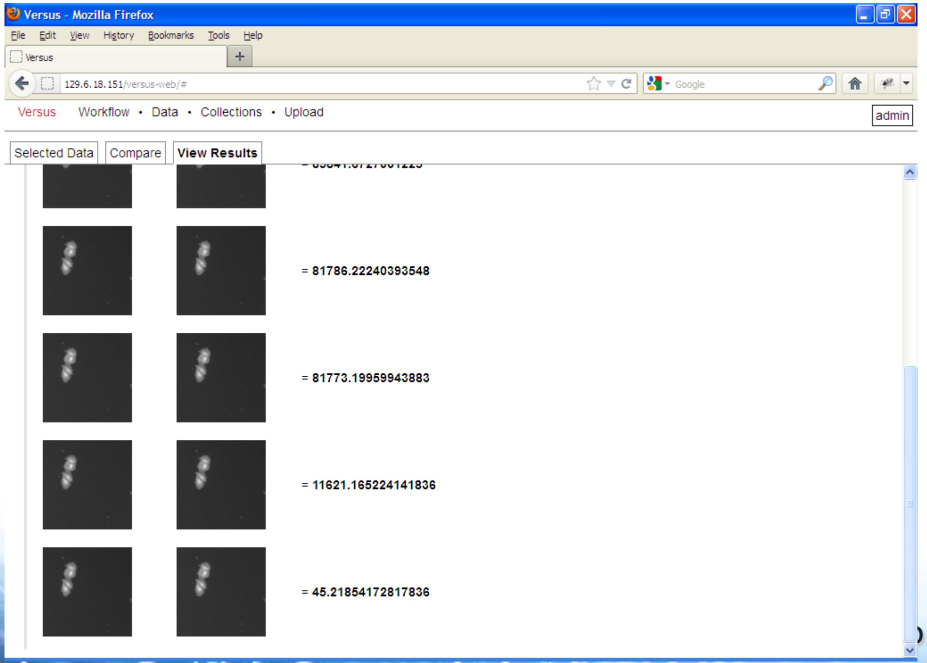


Image Object → Pixels to Array → Euclidean Distance

Visualization

- How to visualize comparisons over Big Data?
  - Human Computer Interfaces for Big Data?



Versus - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Versus
129.6.18.151/versus-web/#
Google
Versus
Workflow • Data • Collections • Upload
admin
Selected Data
Compare
View Results

= 83841.6727001225

= 81786.22240393548

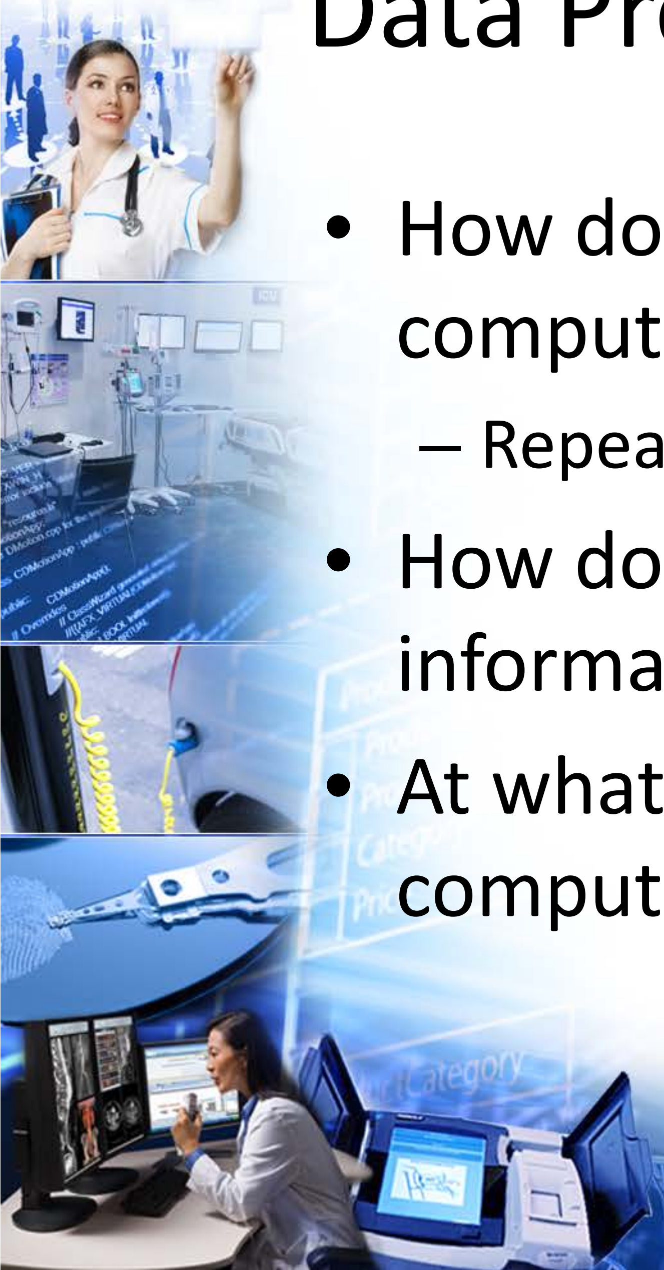
= 81773.19959943883

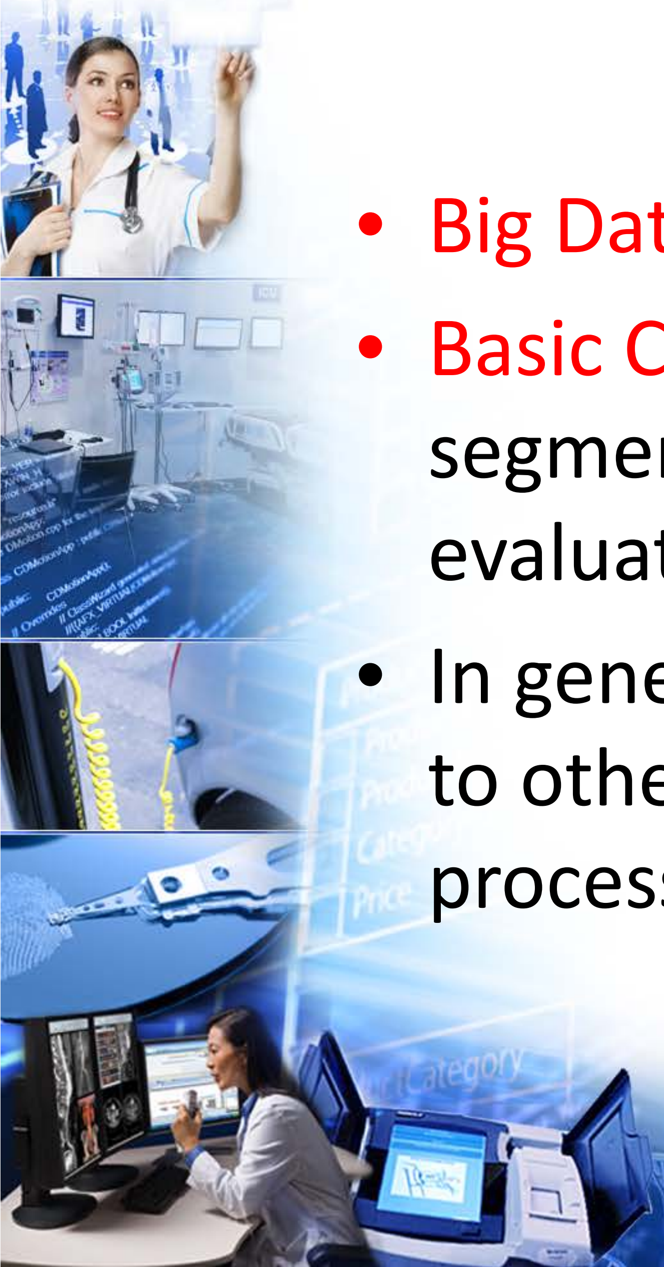
= 11621.165224141836

= 45.21854172817836

Data Provenance Gathering

- How do we automate gathering of computational provenance information?
  - Repeatability
- How do we represent provenance information?
- At what granularity should we gather computational provenance?





Summary

- Big Data: Cell microscopy images
- Basic Challenge: automated segmentation and its accuracy evaluations over Big Data
- In general, the basic challenge applies to other domains for other automated processing operations

Summary

- Accuracy evaluations of automated processing over Big Data include at least
  - Data quality, sampling, and comparison measurements
  - Standards for accessing Big Data remotely
  - Standards for computational provenance information representation

Acknowledgement

CS

- Peter Bajcsy
- Adele Peskin
- Mary Brady
- Joe Chalfoun
- Antonio Cardone
- Frederic de Vaulx
- Afzal Godil
- Ben Long
- Ya-Shian Li-Baboud
- Julien Amelot
- Antoine Vandecreme
- Paul Khouri Saba
- Bertrand Stivalet
- Walid Keyrouz
- Doug Foxvog
- Alden Dima

STATS

- James Filliben
- Dan Samarov

BIO

- Anne Plant
- Kiran Bhadriraju
- Michael Halter
- John Elliott
- TN Bhat

External Collaborators

- Marcin Kociolek(1)
- Carol Parent (2)
- Christina Stuelten (2)
- Michael Weiger (2)
- Wolfgang Loster (3)
- Daniel Hoeppner (5)
- Rachel Errington (4)
- Imtiaz Khan (4)
- Justin Martineau (6)
- David Chapman (6)
- Mike Grasso (7)

(1) Technical University of Lodz, Poland (Marcin Kociolek – image processing

(2) NIH NCI (Carol Parent, Christina Stuelten, Michael Weiger, – cancer cells),

(3) University of Maryland at College Park (Wolfgang Loster – cancer cells),

(4) Broad Institute (Rachel Errington, Imtiaz Khan - bone cells)

(5) The Lieber Institute for Brain Development (Dan Hoeppner - live cell imaging)

(6) University of Maryland, Baltimore County (UMBC) (image processing)

(7) University of Maryland, Baltimore (image processing)

Questions

- peter.bajcsy@nist.gov