WHERE DO WE STAND IN APPLICATIONS OF BIG DATA IN RAD. ONC.

Lei Xing, Jacob Haimson Professor

Departments of Radiation Oncology & Electrical Engineering
Stanford University School of Medicine

Disclosures

- The Department of Radiation Oncology at Stanford University Hospital has a research agreement with Varian Medical Systems.
- Dr. Lei Xing has received speakers honoraria from Varian Medical Systems.
- Research grants supports from NIH, DOD, Varian, Google Inc.

Big data

- Web sites: 40 billion indexed web pages
- Youtube: 100 hrs of videos are uploaded every minute
- WalMart: handels more than 1M transactions per hr.

How to make sense of big data?

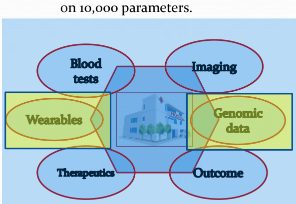
- Big is a reletive and semi-quantitative term
- How big is big? Big data does not necessarily provide a solution to the “missing data” or incomplete data problem.
- In general, it refers to “the study of pt cohorts an order of magnitude or larger than that of the largest prospective cooperative group clincial trials“ (J. Bibault et al, Big data and machine learning in radiation oncology: State of the art and future prospects, *Cancer Letters*, in press, 2016).
- Size of the data should not be the sole measure. It is problem specific.

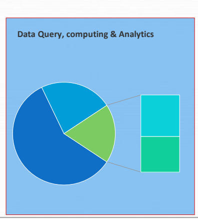
Applications of big data in radiation oncology

- Clinical studies.
- Radiomics & radiogenomics (images are data!).
- Knowledge-based treatment planning.
- Image analysis and CAD.

Big data in clinical studies

- Current practice is evidence-based medicine, relying on randomized trials.
  - Our cognitive capacity can integrate up to ~5 parameters for decision-making.
  - The number of parameters to be tested are dramatically increasing in modern medicine (by 2020, a decision for a single patient could depend on 10,000 parameters.





Big data in clinical studies

- Computing & storage power, especially parallel computing in cloud environment, have increased dramatically
  - MapReduce (Google, 2004) & Hadoop for distributed computing - Pratx G & Xing L, Monte Carlo simulation of photon migration in a cloud computing environment with MapReduce, *J. Biomed. Opt.* 16, 125003, 2011.
- Cost of genomic sequencing is decreasing
- These opens new opportunities for discovering factors influencing the disease's outcome
- Learning technique provides tools for us to meet the informatics challenges in precision medicine - large phenotyped cohorts & data science

Creation of a predictive model

- Consider both dosimetric and nondosimetric predictors
- Assesses multiple models and develop a method for automated predictor selection
- Manually curate predictors before automated analysis
- Consider how predictor multicollinealilty is affecting the model
- Cross-validation to improve prediction performance and generalization to provide model generalizability
- Compare results with established models

Kang et al, principles of modeling in rad onc, Int J Rad Onc Biol Phys, 2015

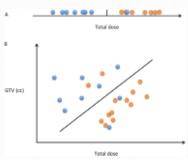
Machine leaning and deep learning

- Traditional statistical methods – logistic regression, Cox regression
- Machine learning methods –
  - decision trees (DT), a simple algorithm creates mutually exclusive classes by answering questions in a predefined orders;
  - Naïve Bayes (ND) classifiers, outputs probibilistic distributions among variables; k-nearest neighbors (k-NN);
  - Support vector machine (SVM), where a trained model will classify new data into categories;
  - Artificial neural net work (ANN), where models inspired by biological neural networks are used to approximate functions;
  - Deep learning, where multiple layers of neurons are used and is able to perform supervised and unsupervised learning.

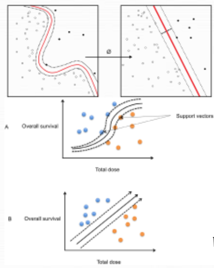
https://class.coursera.org/ml/class/index

J. Bibault et al, Big data and machine learning in radiation oncology: State of the art and future prospects, *Cancer Letters,* in press, 2016.
P. Lambin et al, Nature Reviews Clinical Oncology 10, 27-40, 2013

Linear regression and support vector machine



J. Bibault et al



Wikipedia

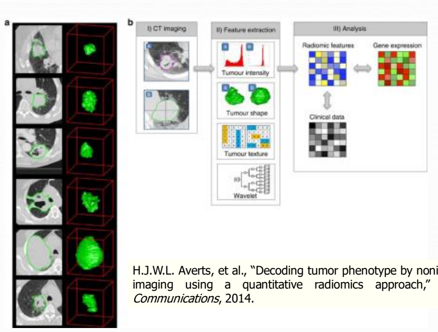
https://en.wikipedia.org/wiki/Support_vector_machine

P. Flach, Machine Learning, Cambridge Univ Press, 2012
J. Bibault et al, Big data and machine learning in radiation oncology: State of the art and future prospects, *Cancer Letters,* in press, 2016.

Supervised and unsupervised leaning

- Supervised learning will analyze a training dataset (where each example is a pair including an input feature and the desired output value) in order to create a function that best matches these training examples. The machine will generalize this function to pairs with unknown output value to predict them.
- Unsupervised learning – the data provided are unlabeled and the algorithm will try to find natural patterns or groups within data. In medicine, this will consist of characterizing each pt with vectors and values given to clinical features. Higher level features can be detected that would not have been seen as potential predictive or prognostic factors by a human intervention.

Radiomics for personalized medicine



H.J.W.L. Averts, et al., "Decoding tumor phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, 2014.

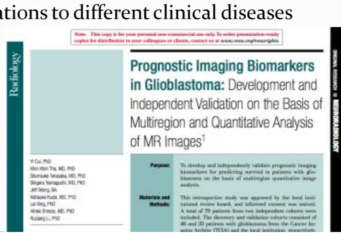
Radiomics for personalized medicine

- Biomarker: a measurable indicator of some biological state or condition
- Biomarker is a *key* element of personalized medicine.
  - Prognostic biomarkers: likelihood of disease progression – aggressive vs. indolent
  - Predictive biomarkers: sensitivity to therapy (drugs, radiation)
  - Early response biomarkers: spare patients ineffective treatment; speed up clinical trails.

Courtesy of Y. Cui

Role of big data and machine learning in radiomics

- Image feature definition and extraction
- Training and machine learning to establish predictive model
- Applications to different clinical diseases


Note: This copy is for your personal non-commercial use only. To order presentation-ready copies for distribution to your colleagues or clients, contact us at www.rsna.org/rsnarights.

Radiology

ORIGINAL RESEARCH ■ NEURORADIOLOGY

Prognostic Imaging Biomarkers in Glioblastoma: Development and Independent Validation on the Basis of Multiregion and Quantitative Analysis of MR Images[1]

Yi Cui, PhD
Khin Khin Tha, MD, PhD
Shunsuke Terasaka, MD, PhD
Shigeru Yamaguchi, MD, PhD
Jeff Wang, BA
Kohsuke Kudo, MD, PhD
Lei Xing, PhD
Hiroki Shirato, MD, PhD
Ruijiang Li, PhD

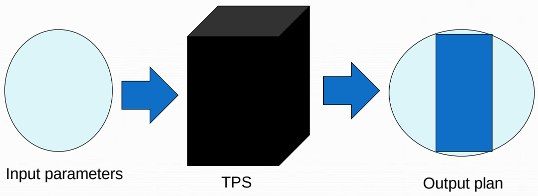
Purpose: To develop and independently validate prognostic imaging biomarkers for predicting survival in patients with glioblastoma on the basis of multiregion quantitative image analysis.

Materials and Methods: This retrospective study was approved by the local institutional review board, and informed consent was waived. A total of 79 patients from two independent cohorts were included. The discovery and validation cohorts consisted of 46 and 33 patients with glioblastoma from the Cancer Imaging Archive (TCIA) and the local institution, respectively.

Applications of big data in radiation oncology

- Clinical studies.
- Radiomics (images are data!).
- knowledge-based treatment planning.
- Machine learning/Deep learning for image registration and segmentation.

Inverse Treatment Planning



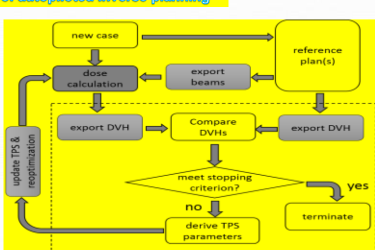
Problem: manual process with multiple trial and errors

Methods toward automated planning

- RapidPlan/Principle components
- Learning and deep-learning algorithm
- Multiobjective (RayStation)
- Automatic planning (Pinnacle)

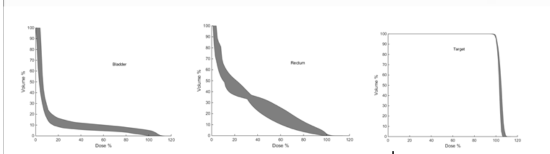
Autopiloted treatment planning guided by big-data

Workflow of autopiloted inverse planning

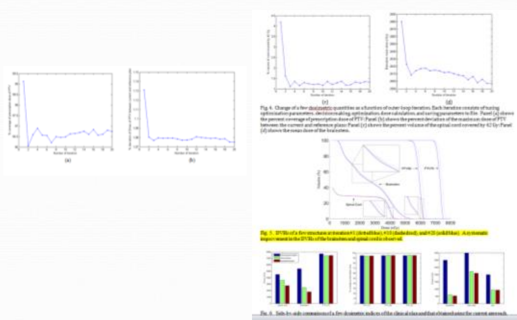


opt	TPS objective function with constraints	
s.t.	$\left\| d_{\sigma j} - d_{\sigma j}^{ref} \right\| < \varepsilon_{\sigma}$	$(\sigma \in 1,2,\ldots,N;\ j \in 1,2,\ldots,J)$

Prior knowledge for autopilot Eclipse (prostate IMRT)



Autopilot process of HN RapidAR case



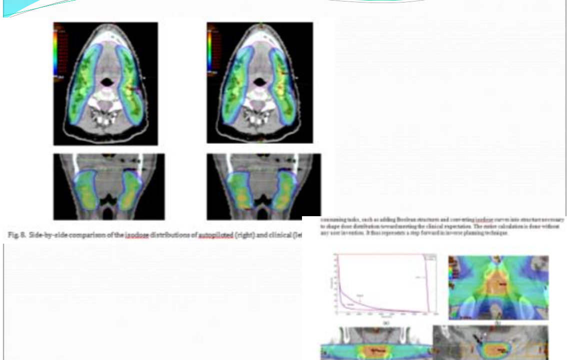


Fig 8. Side-by-side comparison of the isodose distributions of autopiloted (right) and clinical (le

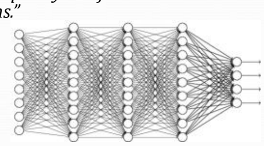
Applications of big data in radiation oncology

- Clinical studies.
- Radiomics (images are data!).
- Autopiloted and/or knowledge-based treatment planning.
- Machine learning/Deep learning for image registration and segmentation.

Deep learning

"machine learning technique to model high-level abstractions in data using multiple layers of linear or non-linear transformations."

Wikipedia



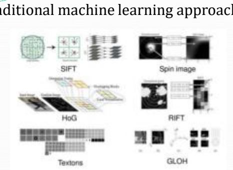
Major aspects of deep learning:

- Cascades multiple layers of processing units
- Units can comprise a broad family of linear/nonlinear functions for feature extraction and transformation
- Layers form a hierarchy from low- to high-level features
- Based on distributed representations assuming the observed data are generated by the interactions of factors

S Arik, B Ibragimov & L Xing, Med Phys, submitted, 2016

Traditional machine learning vs. deep learning

- Traditional machine learning approaches use pre-defined features:

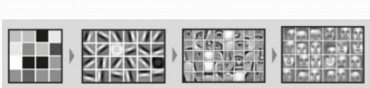


Schreibmann and Xing L, Med Phys. 2008

Xie Y, Chao M, Xing L, Tissue feature-based and segmented deformable image registration for improved modeling of shear movement of lungs, Int J Radiat Oncol Biol Phys.74, 1256-65, 2009

Ref. *Deep learning methods for vision*

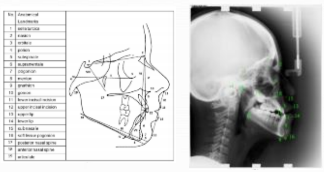
  - Non-flexible representation
  - Time-consuming hand-tuning
  - Problem specific
- Deep learning uses trainable feature extractors and optimizes the way to extract features.



Ref. *Convolutional deep belief networks, Honglak et al.*

Landmark detection in cephalometric analysis

- Analysis of X-ray images of the craniofacial area, i.e. cephalograms.



Ref. *Evaluation and Comparison of Anatomical Landmark Detection Methods for Cephalometric X-Ray Images: A Grand Challenge, C. Huang et al.*

- Marking of anatomical landmarks in cephalometric analysis is necessary as it provides the interpretation of patients' bony structures for surgery. Numerous clinical applications include:
  - Diagnosis and treatment of obstructive sleep apnea
  - Assessment of mandible/lower jaw
  - Assessment of soft facial tissue

S Arik, B Ibragimov & L Xing, Med Phys, submitted, 2016

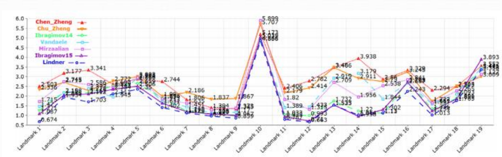
Landmark detection in cephalometric analysis

Technical (major) reasons for the choice of cephalometric analysis problem:

1)Availability of a sufficiently large image set (with ground-truth labels) in order to train a complex model

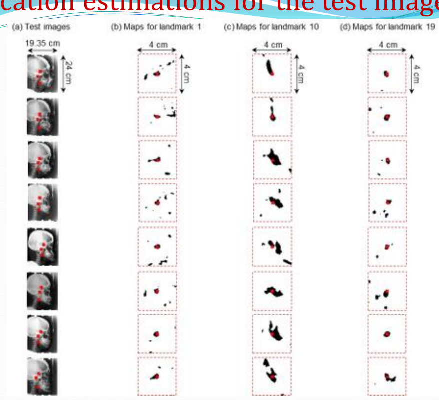
- 150 images for training and 250 images for testing

2)Availability of competitive benchmarks

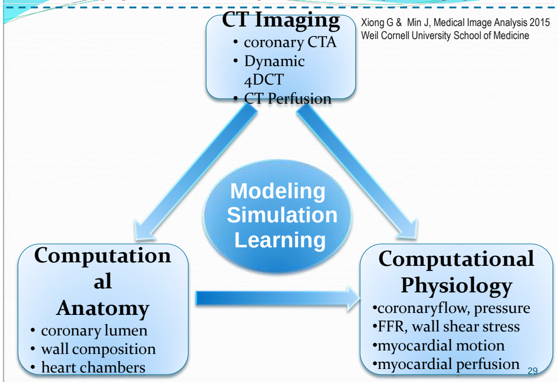


Ref. *Evaluation and Comparison of Anatomical Landmark Detection Methods for Cephalometric X-Ray Images: A Grand Challenge, C. Huang et al.*

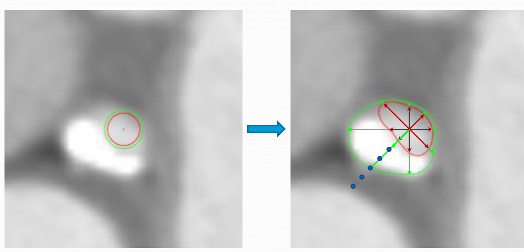
Location estimations for the test images



Towards Imaging Informatics for diagnosis of CAD using CT Imaging



Lumen and Wall Segmentation using Machine Learning



lumen

wall

Xiong G & Min J, Medical Image Analysis 2015

30

Summary

- What it takes for us to benefit from big data?
  - large database
  - Data science

    > predict outcome and guide treatments…..

- Machine learning tools to process data and extract meaningful information.
  - Clinical studies
  - Treatment planning
  - Image analysis

Acknowledgement

Bulat Ibragimov, Sercan Arik, Moteza Korani, Hongcheng Liu, Peng Dong, Anqi Liu, Ruijiang Li, Yi Cui, Guangwei Xiong, Ces Jenkins, Karl Bush, Chris Locke, Bin Han, Benjamin Fahimian, Albert Koong, Quynh Le, Stephen Boyd

National Institute of Health, Varian Medical Systems, Google Inc