Best Practices:

Reproducibility and code sharing in Quantitative Science / Machine Learning

2022-06-06

Jamie K. Teer, Ph.D.
Yi Luo, Ph.D.

MOFFITT CANCER CENTER

Review - FAIR

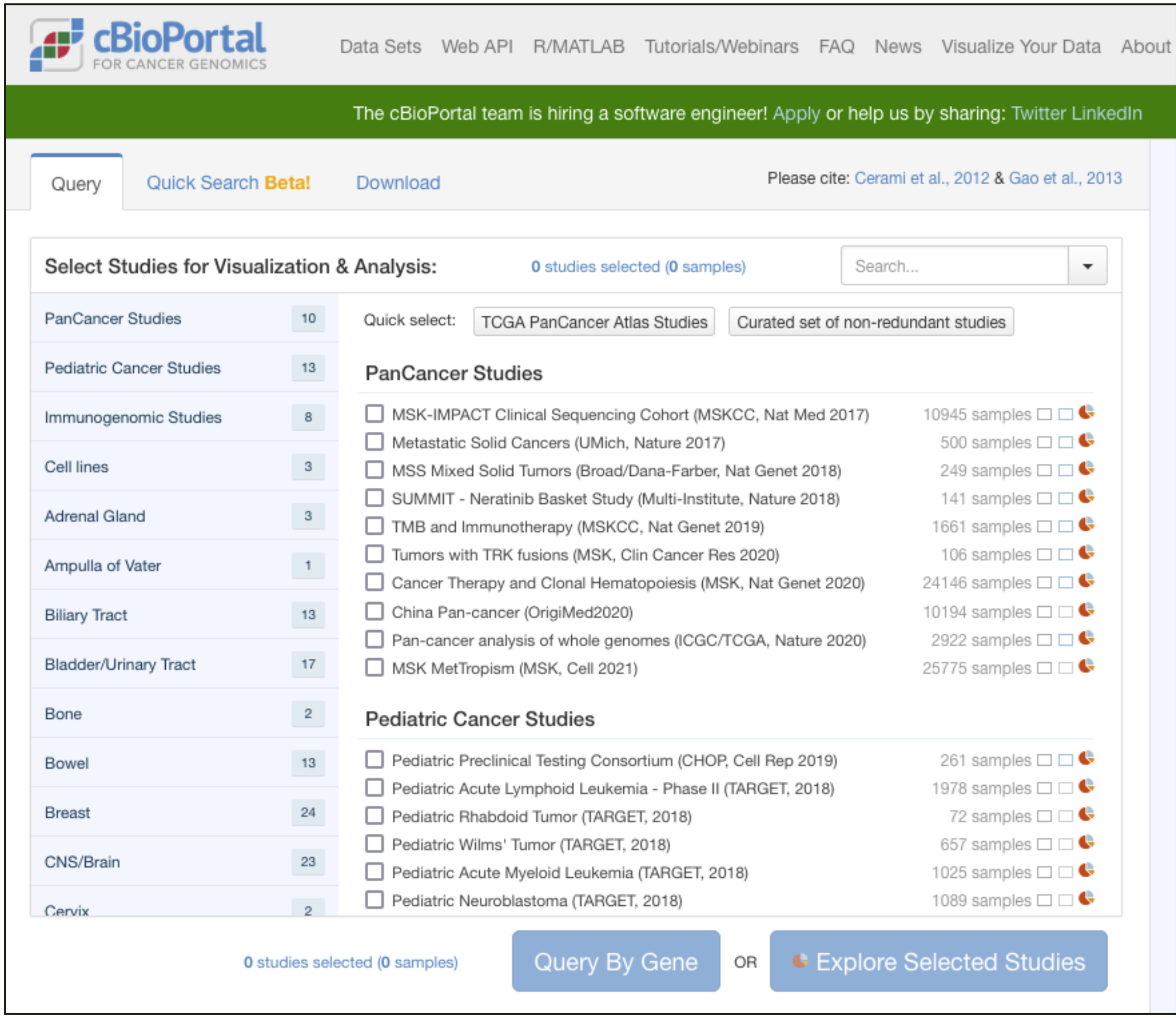
- '**Findability**' implies data can be **found** online, typically through indexing in search engines.
- '**Accessibility**' means data can be **retrieved** directly or via an approval process.
- '**Interoperability**' imposes data to follow **standards**.
- '**Reusability**' requires the context of the data generation (metadata) is **documented** so it can be compared to or integrated with other data sets.

Charles Vesteghem, Briefings in Bioinformatics, 2020,

MOFFITT CANCER CENTER

FAIR Sequence Data Example - cBioPortal

- Many sequencing projects share high-level results via a web-based tool.
- Includes The Cancer Genome Atlas (TCGA) and others.
- Can easily **find** different diseases.
- Can **access** clinical and molecular results.
- **Interoperability** via downloads and APIs
- **Reusable** data and software!

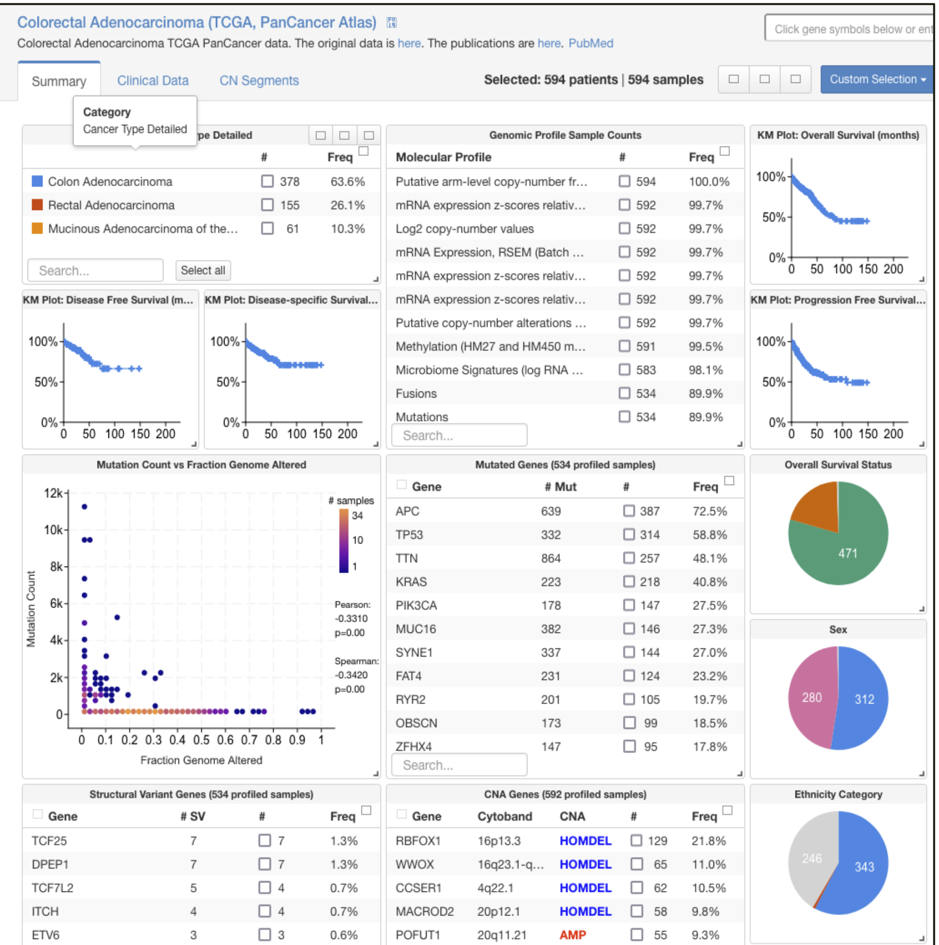


www.cbioportal.org

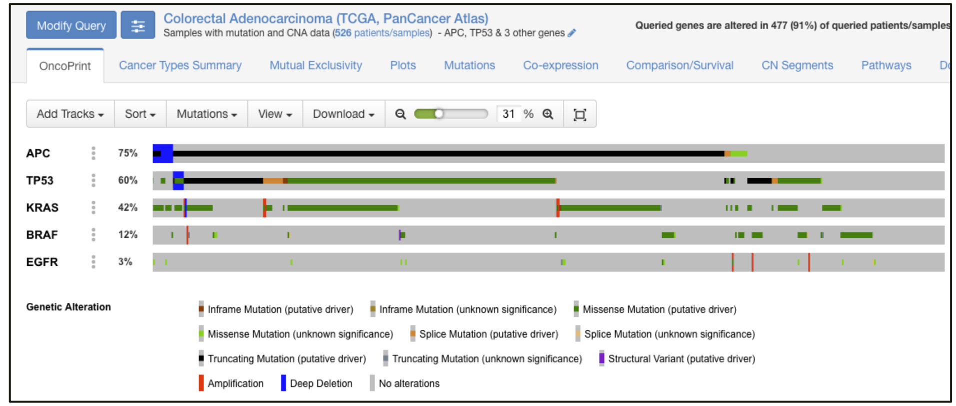
MOFFITT CANCER CENTER

cBioPortal, Colorectal Cancer Example

Clinical Data



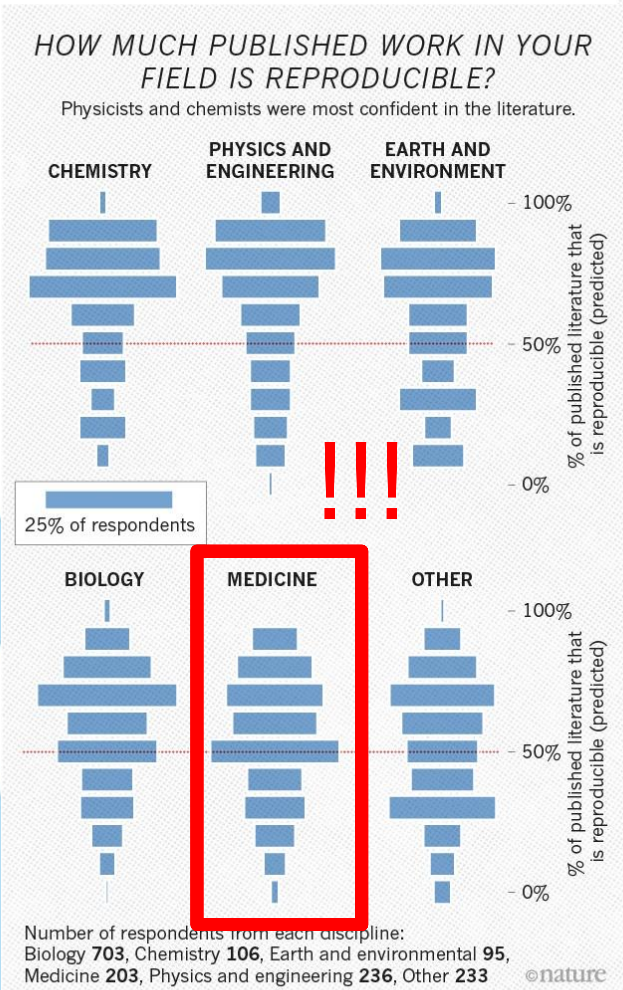
Sequence Results (Mutations)

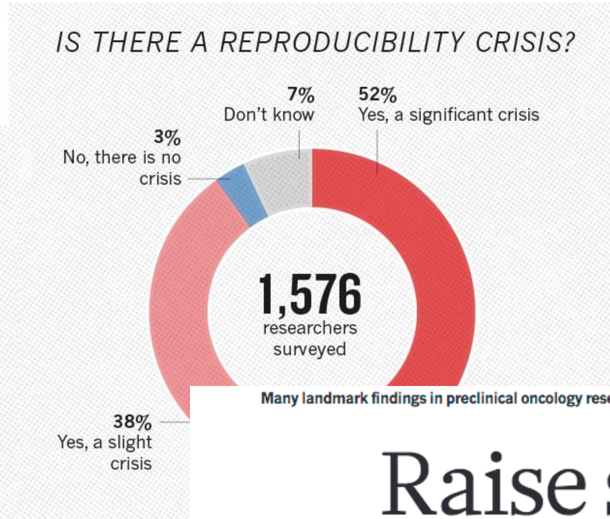


MOFFITT CANCER CENTER

The importance of reproducibility

1,500 scientists lift the lid on reproducibility





Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even

© 2012 Macmillan Publishers Limited. All rights reserved

"Nevertheless, scientific findings were confirmed in only 6 (11%) cases. Even knowing the limitations of preclinical research, this was a shocking result."

"Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings.

…

When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome."

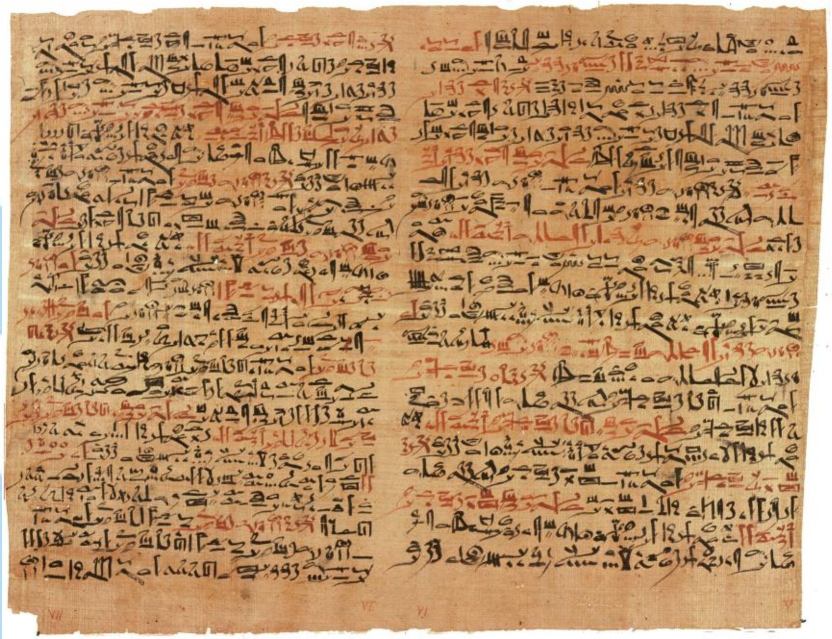
- NIH on Rigor and Reproducibility

https://www.nih.gov/research-training/rigor-reproducibility

MOFFITT CANCER CENTER

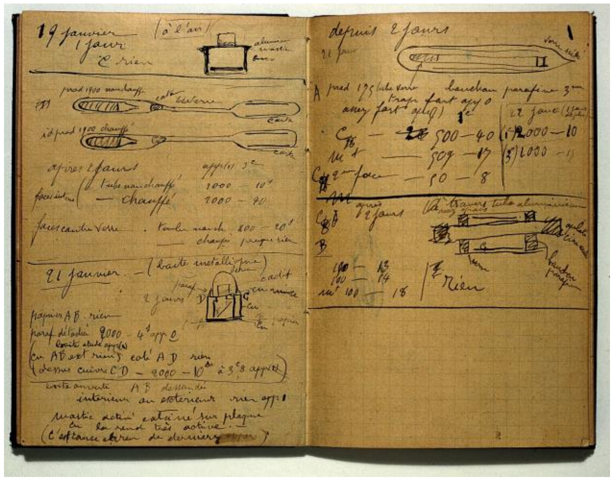
Fundamental Philosophy

- Reproducibility a critical part of any scientific investigation.
- Recording observations allows others to know what has been done and start from what has already been learned.
- There is a long history of recording observations:



Edwin Smith Papyrus, ca 1600 BCE
Egyptian surgery manual

Da Vinci
~1485-1510



Marie Curie
1899-1902

Requirements

Overall: record the practice of the scientific method. Allow an expert unfamiliar with the specific project to follow along.

- Describe project goals, and how experiment relates to goals
- Describe experiment, how it was performed, observations (classical understanding of a lab notebook)
- For data science, include (or link) code and settings used to output.
  - Electronic lab notebooks to describe experiment
  - Version control to maintain code
  - Workflow/pipeline tools to automate
- Describe conclusions. Helpful when going back to the notebook to understand why the experiment was done, and what was learned. Avoid re-interpreting data over and over.
- Future experiments: how to test remaining questions or apply what was learned in an experiment.
- Able to be cross-referenced, in order to find successful approaches from other projects to avoid re-inventing the wheel.

Poll

Have you taken a formal course on keeping a lab notebook?

Classic lab notebook

Components:

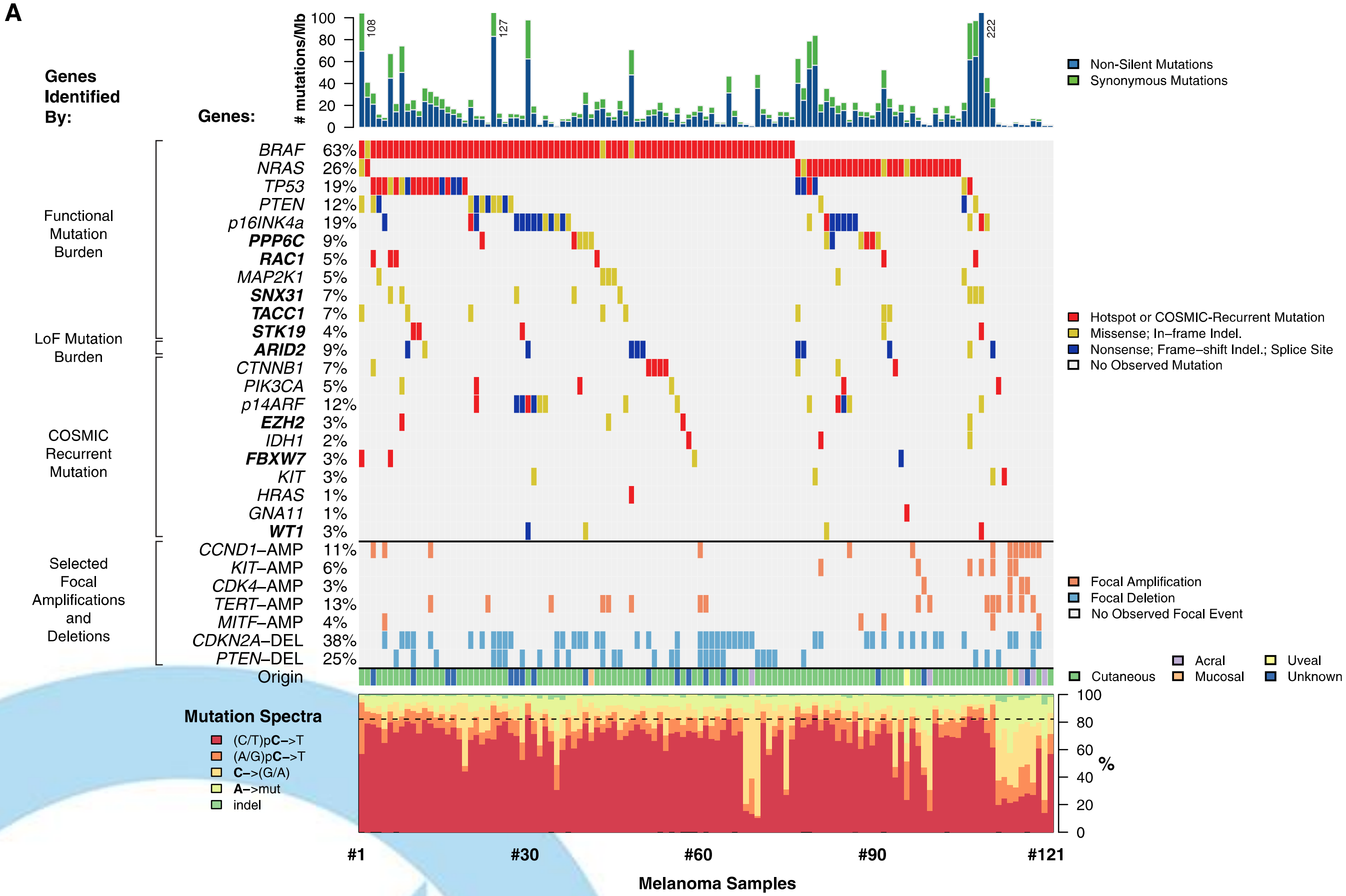
1. Date/Name
2. Hypothesis: goals of the experiment and expected outcome
3. Methods: What did you actually do?
4. Results: Actual results (blots, summary metrics, etc). Properly labeled!
5. Conclusions: Very helpful to have your conclusions about an experiment, including success/failure, reasons or improvements, overall conclusions
6. Locations of outputs: reagents, files, etc. Link products to the notebook.

7 scary words

"I have a question about your figure."

MOFFITT CANCER CENTER

Molecular data analysis – TCGA example

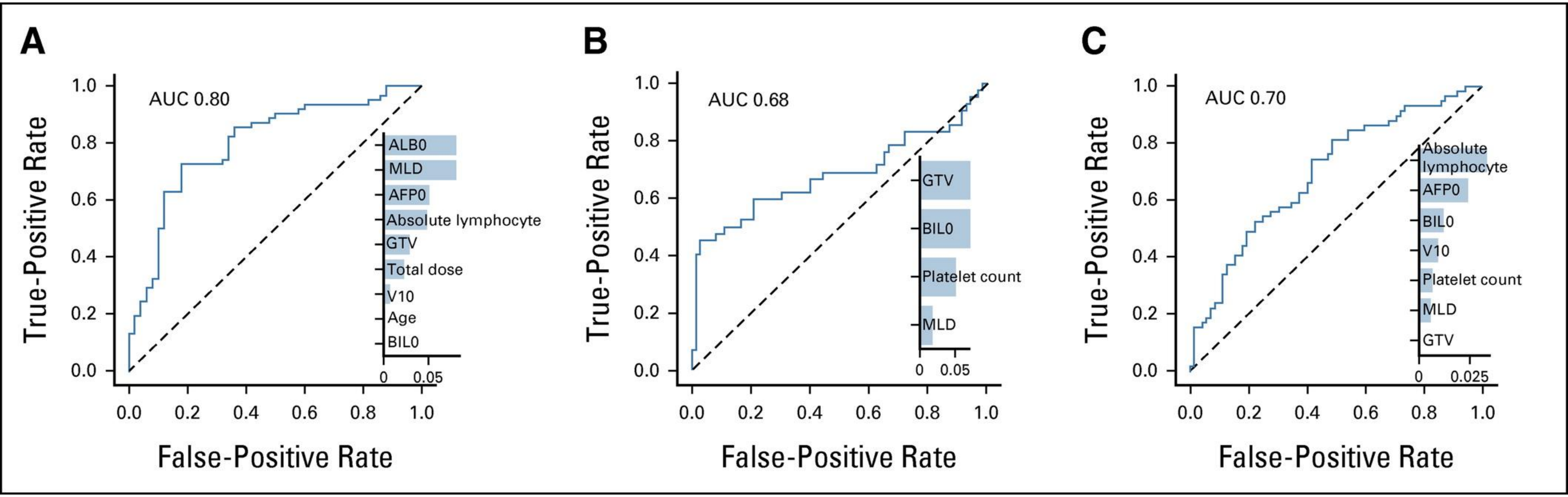


- WES on 121 Tumor/Normal pairs
- Mutations, CNV, counts, spectra
- ~20,000 genes
- Melanoma type

"Did you remember to exclude the patient with the odd toenail tumor?"

Hodis E et al Cell, 2012

Machine learning analysis – AUC example



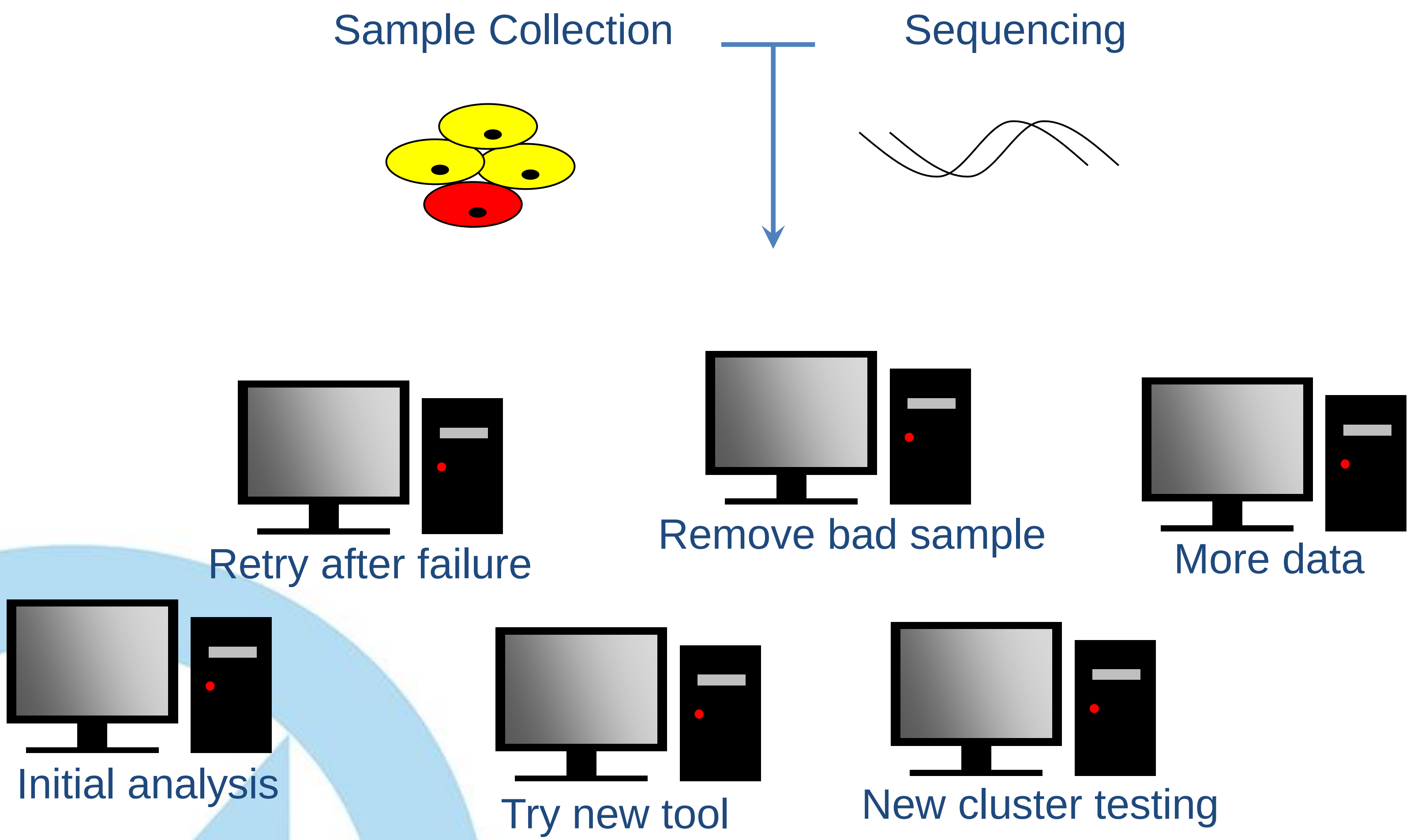
- External validation of models predicting tumor control and toxicity.
- <u>A:</u> Nonlocal failure Yr 1 <u>B:</u> 2+ increase in Child-Pugh score <u>C:</u> lymphopenia

<u>"Are you sure the results in Fig A are the Nonlocal failure?"</u>

Chamseddine etal, JCO Clinical Cancer Informatics, 2022

MOFFITT CANCER CENTER

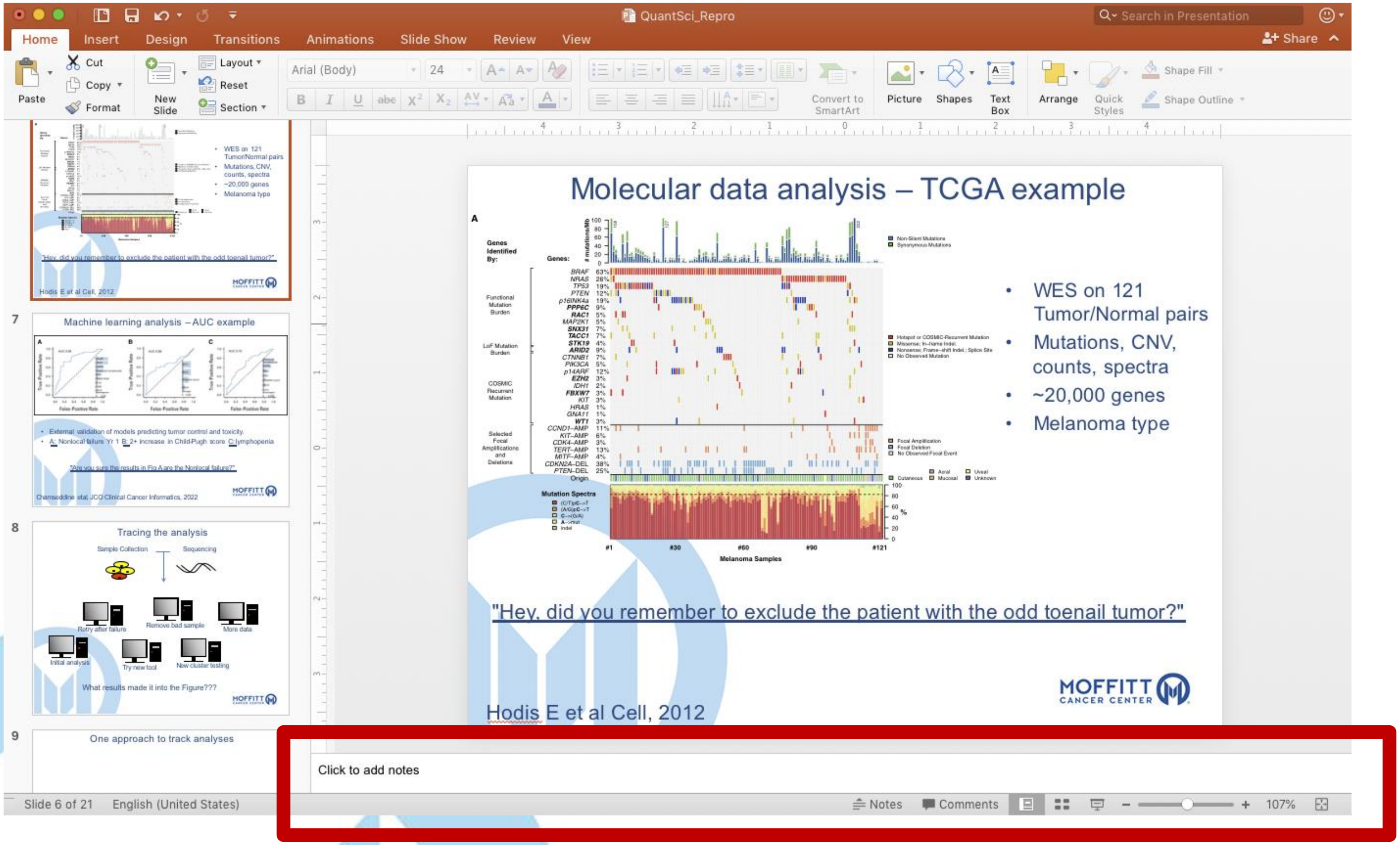
Tracing the analysis



What results made it into the Figure???

MOFFITT CANCER CENTER

One approach to track analyses



Full path to output figure goes here!

Record the location of the data in the "comments" field!

A link back to the working directory with <u>output</u> AND <u>documentation</u>!

Dry lab notebook examples – Text file

- README file – common in the software world
- With linux commandline editors, can automate things like date/user tags
- Can copy/paste the exact commands run.
- Comments allow for prose descriptions of results, conclusions, etc
- Can be stored together with the data
- Can make a list of all README files, write a simple linux script to search for keywords across all files!

Cons:

- Plaintext not as feature-rich as markdown options
- Documentation is dispersed across projects. Strategies needed for searching and backup
- Usability depends on the organization of the file as written

MOFFITT CANCER CENTER

Text file code examples - command line

.bashrc alias (type “newREADME” to create a README file and add it to ~/README.list)

```
alias newREADME='touch README; echo `pwd`/README >> ~/README.list'
```

Searching all README files

```
for i in `cat ~/README.list`; do grep -Hi $KEYWORD $i; done
```

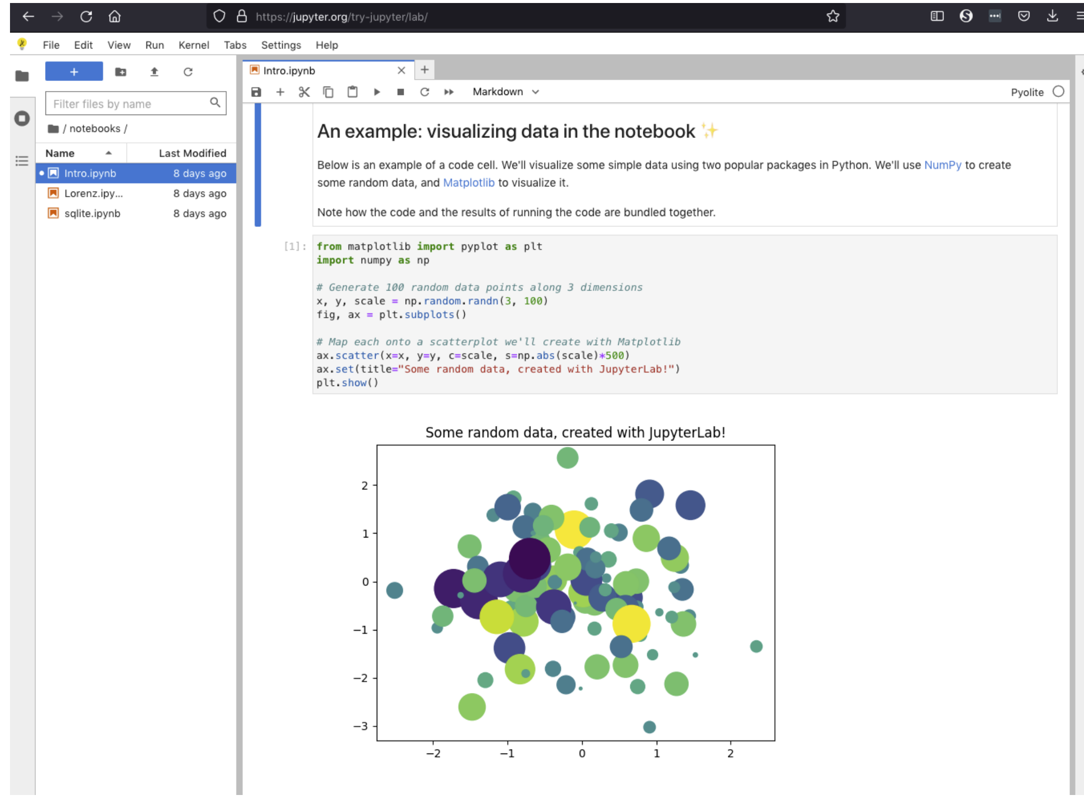
Backup all text in README files marked by full file path:

```
for i in `cat README.list`
  do echo -ne "****************FILE:\t"; echo "$i"; cat $i
done | gzip -c > README.backup_YYYY_MM_DD.txt.gz
```

- Can create text editor shortcuts to add your name, date, time
- Can search within README files using text editor commands, grep.
- Can write a script to test that all README files exist on the filesystem
  - Let's you know if files or folders have been deleted!

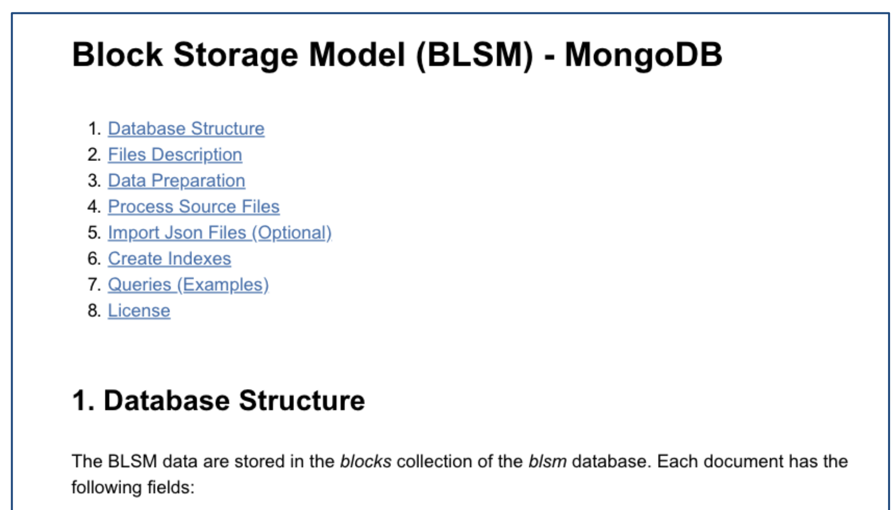
Dry lab notebook examples – Jupyter notebook

- Can have plugins to extend features: https://arxiv.org/abs/2111.00831
- Run code interactively
- Local and web-based instances
- GUI-based
- Can use Markdown to format text (like a Wiki)
- Can output to PDF



Cons:

- Not all languages supported
- May not be useful to run large projects



Block Storage Model (BLSM) - MongoDB

1. Database Structure
2. Files Description
3. Data Preparation
4. Process Source Files
5. Import Json Files (Optional)
6. Create Indexes
7. Queries (Examples)
8. License

1. Database Structure

The BLSM data are stored in the *blocks* collection of the *blsm* database. Each document has the following fields:

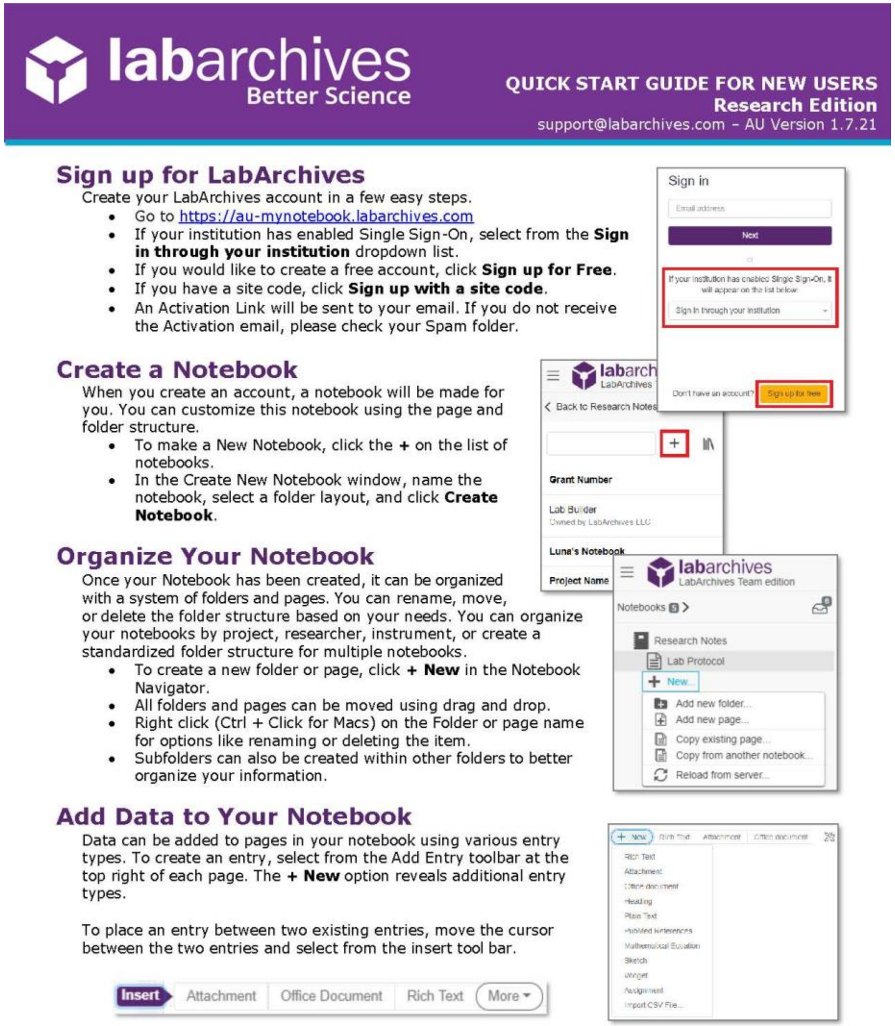
jupyter.org

Dry lab notebook examples – LabArchives

- Moffitt solution to electronic lab notebooks
- https://www.labarchives.com/labarchives-knowledge-base/
- https://intranet.moffitt.org/display/RELN/Lab+Archives

Cons:

- Designed for wet lab
- May not be available at all institutions


labarchives
Better Science

QUICK START GUIDE FOR NEW USERS
Research Edition
support@labarchives.com – AU Version 1.7.21

Sign up for LabArchives

Create your LabArchives account in a few easy steps.

- Go to https://au-mynotebook.labarchives.com
- If your institution has enabled Single Sign-On, select from the **Sign in through your institution** dropdown list.
- If you would like to create a free account, click **Sign up for Free**.
- If you have a site code, click **Sign up with a site code**.
- An Activation Link will be sent to your email. If you do not receive the Activation email, please check your Spam folder.

Create a Notebook

When you create an account, a notebook will be made for you. You can customize this notebook using the page and folder structure.

- To make a New Notebook, click the **+** on the list of notebooks.
- In the Create New Notebook window, name the notebook, select a folder layout, and click **Create Notebook**.

Organize Your Notebook

Once your Notebook has been created, it can be organized with a system of folders and pages. You can rename, move, or delete the folder structure based on your needs. You can organize your notebooks by project, researcher, instrument, or create a standardized folder structure for multiple notebooks.

- To create a new folder or page, click **+ New** in the Notebook Navigator.
- All folders and pages can be moved using drag and drop.
- Right click (Ctrl + Click for Macs) on the Folder or page name for options like renaming or deleting the item.
- Subfolders can also be created within other folders to better organize your information.

Add Data to Your Notebook

Data can be added to pages in your notebook using various entry types. To create an entry, select from the Add Entry toolbar at the top right of each page. The **+ New** option reveals additional entry types.

To place an entry between two existing entries, move the cursor between the two entries and select from the insert tool bar.

labarchives.com

MOFFITT CANCER CENTER

Data Science lab notebook: no "official" method

- You as a scientist will need to decide on an approach to use.
- Does a specific approach satisfy outlined requirements?
- Is an approach easy enough to actually use? (If too time consuming, will not endure.)
- Can others follow your experiments?

Version Control

A structured approach to preserve a history of changes to a file.

- Example (not great): word documents with rev1, rev2, final, final_a, really_final. Can know changes by comparing documents.
- Version control is designed around keeping a history of changes from MANY authors.
- Approaches to go back to previous versions, undo specific versions, etc.
- Usually used for software code, especially for larger teams making many changes at once.
- Client-Server (main repository is on a server) vs distributed (all users have a copy of main repository)

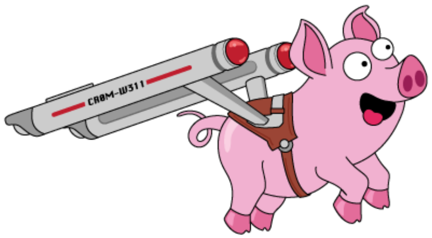
- Examples include:

- Git (distributed tool, developed to maintain Linux kernel)
- Subversion (older client-server tool)
- Mercurial (distributed tool, commands similar to Subversion)

MOFFITT CANCER CENTER

Reproducible workflows

- Pipeline development tools enable joining multiple processing steps together.
- Like a script, but separates engineering code from analysis code.
- Ideally allows for more consistent structures and practices
- Often open source, often support containers



WDL/Cromwell:
Workflow Descriptor Language describes a human readable language

Cromwell (Java) interprets and runs in a variety of compute environments

https://cromwell.readthedocs.io/en/stable/

Snakemake:
Python based language

https://snakemake.readthedocs.io/en/stable/index.html

Nextflow:
Custom human readable language

Java interpreter

https://snakemake.readthedocs.io/en/stable/index.html

MOFFITT CANCER CENTER

Containers

- An approach to bundle needed software together to run a program in a variety of compute environments
- Like a Virtual Machine, but lighter
- Separates software and dependencies from what is installed on the system
- May therefore allow for specific program versions to "last longer" (ie, not limited to installed dependency versions)

docker

- Leader in the field
- Rich environment of existing containers
- Often not permitted on HPC due to need for elevated permissions

Singularity

- Developed for HPC
- Better security model for HPC
- Can covert from Docker

Best Practices in Machine Learning:

Dr. Luo's approach

Bayesian Networks Approach for Personalized Adaptive Radiotherapy in Lung Cancer Patients

Personalized Adaptive Radiotherapy in Lung Cancer Patients

- Lung cancer (both small cell and non-small cell) is the second most common cancer, and it is by far the leading cause of cancer death among both men and women.
- Radiotherapy is the main treatment for locally advanced lung cancer. Outcomes of radiation treatment include patients' survival, tumor local control (LC) and radiation-induced toxicities (RITs), such as radiation pneumonitis, esophagitis, cardiac toxicity.
- The objective of personalized adaptive radiotherapy (pART) is a trade-off of obtaining LC while limiting RITs. In this study, radiation pneumonitis grade two and above (RP2) is considered as a representative of RITs.

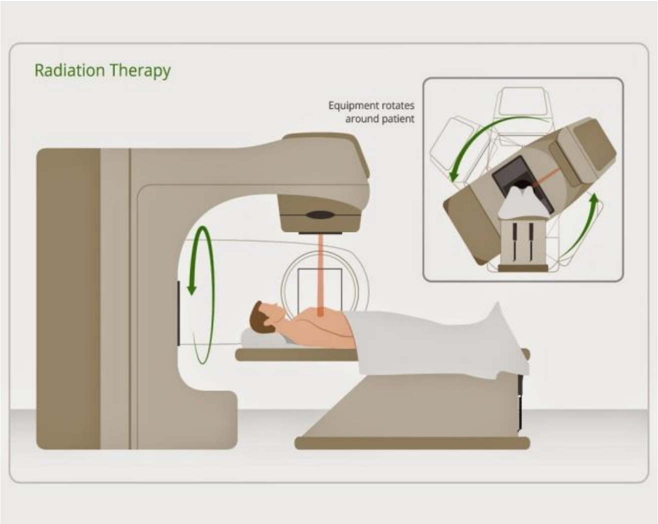
High Dimensional Retrospective Dataset

Retrospective dataset included 118 lung cancer patients, where 68 patients were selected for discovery dataset, and 50 patients were reserved for validation dataset. Number of biophysical features in the whole dataset for LC and RP prediction model development are listed as follows.

Categories	# of features in the whole dataset	# of features for LC prediction	# of features for RP2 Prediction
Common dosimetric information	15	6	9
Clinical factors	14	14	14
MicroRNAs (miRNAs)	62	62	62
Single nucleotide polymorphisms (SNPs)	60	60	60
Pre-treatment positron emission tomography (PET) radiomics	43	43	0*
Relative difference (RD) of PET radiomics during treatment	43	43	0*
Pre-treatment cytokines	30	30	30
Slopes (SLP) of cytokines change during treatment	30	30	30
Total	297	288	205

Data-Driven Approach to Achieve pART

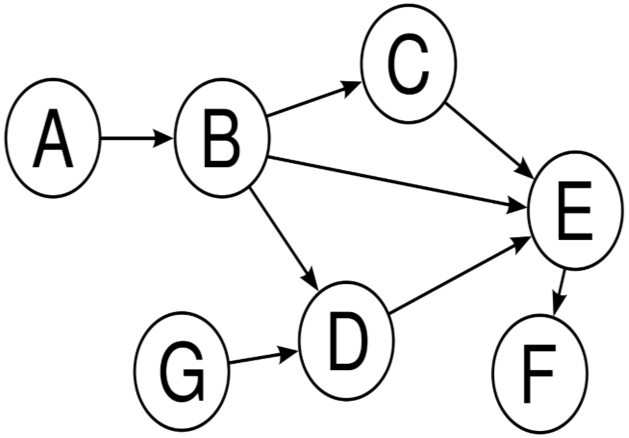
- Radiation outcomes such as LC and RP2 may depend on radiation dose, the patient's physical, clinical, biological, imaging, and genomic characteristics over the course of the radiotherapy. In an era of big data, pART can be achieved to improve patients' therapeutic satisfaction.
- The central challenge is how to integrate diverse, multimodal information in a quantitative manner
  i. to explore the biophysical relationship among radiation treatment, patients' characteristics, and their radiation outcomes?
  ii. to identify the optimal robust treatment plans before and during the radiotherapy?



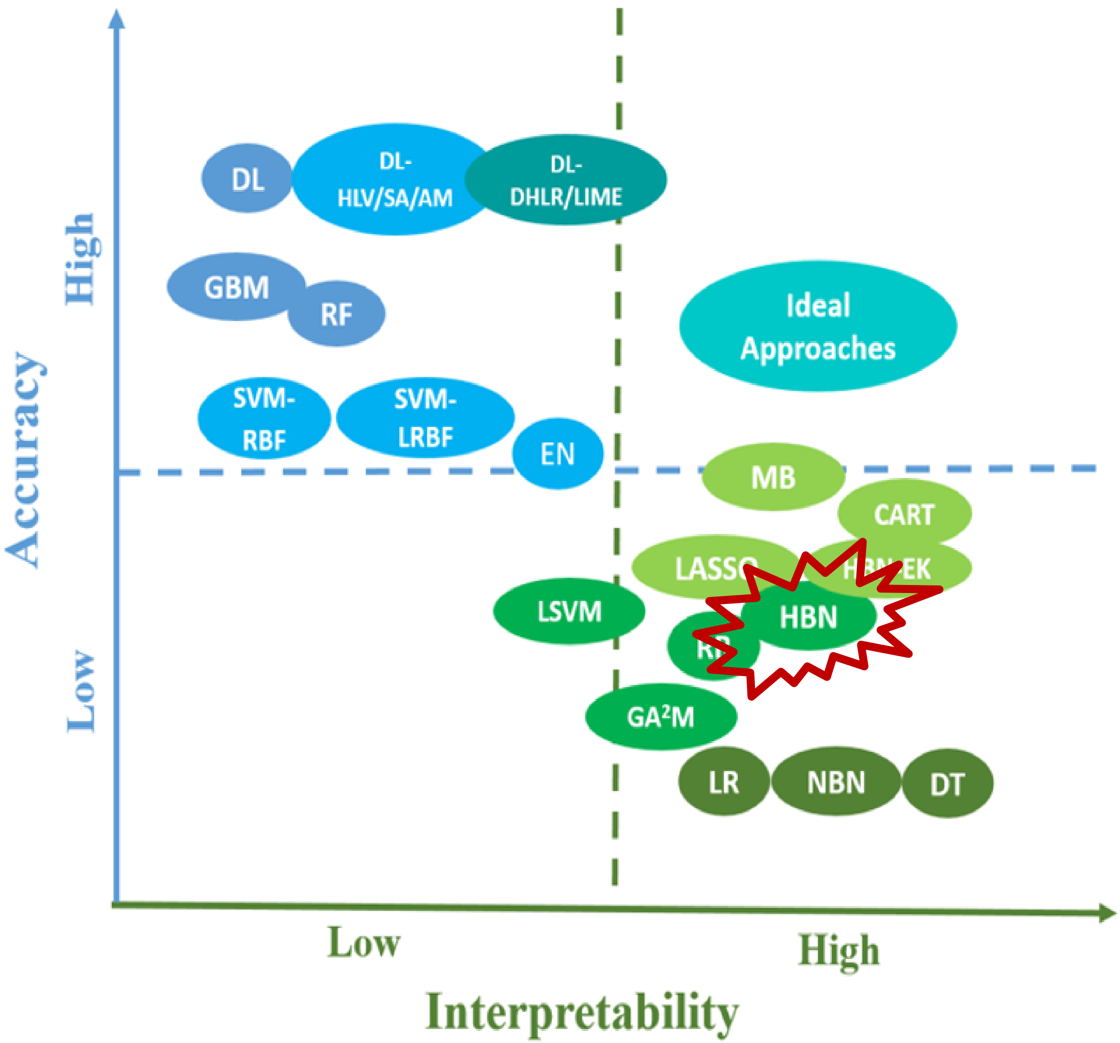
Bayesian Networks

A Bayesian network (BN) is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). BNs are promising approaches to support decision making for radiation therapy because they can be used to

- probe non-linear relationships among clinical, physical, imaging and genomic data;
- accommodate their hierarchical interactions;
- incorporate experts' insight and intuitions;
- simultaneously predict multiple objectives;
- handle missing data;
- represent a probabilistic dependency to help people understand how the variables are jointly related to each other to reach a final prediction.



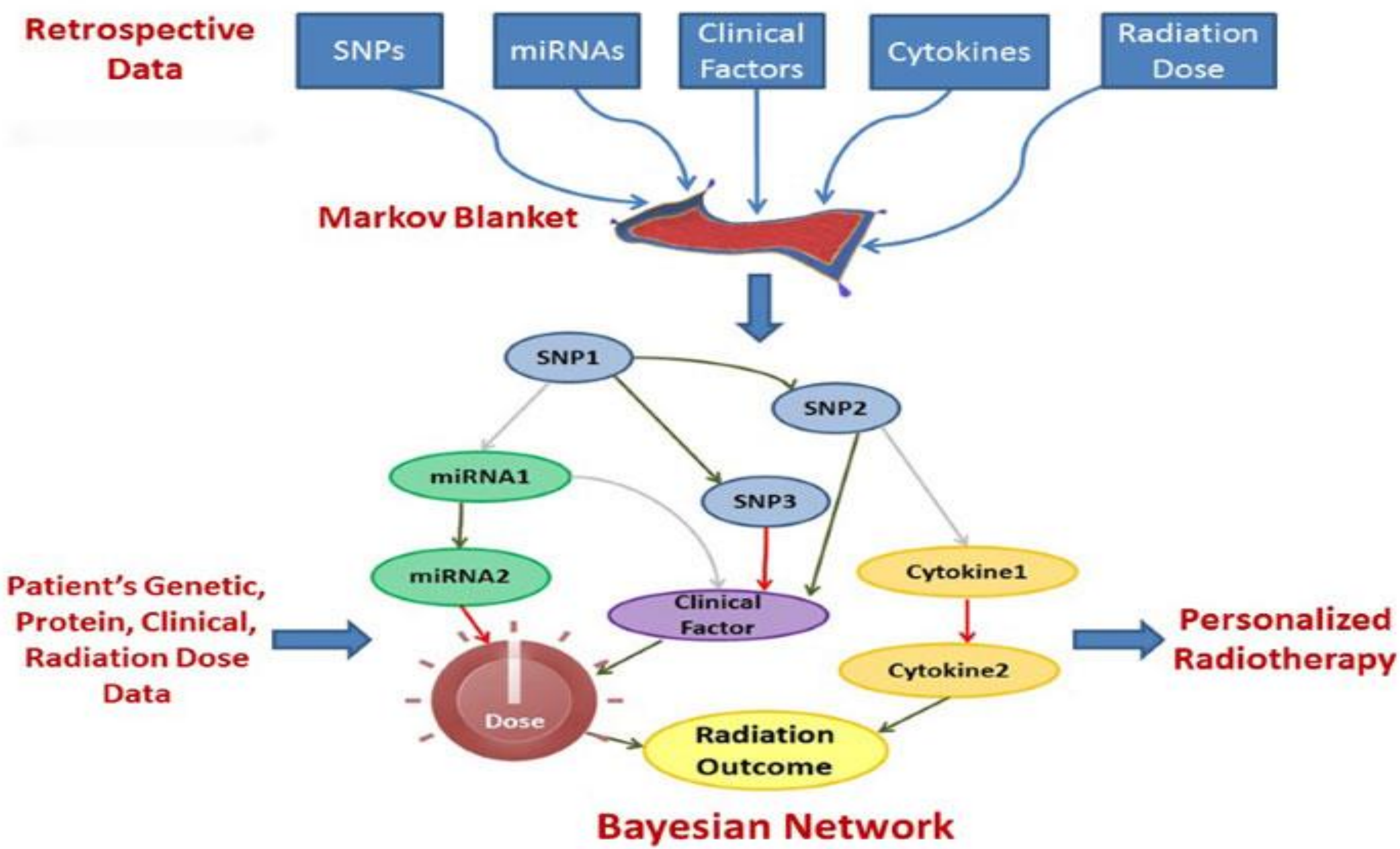
Accuracy Vs. Interpretability



Luo, et al., 2019

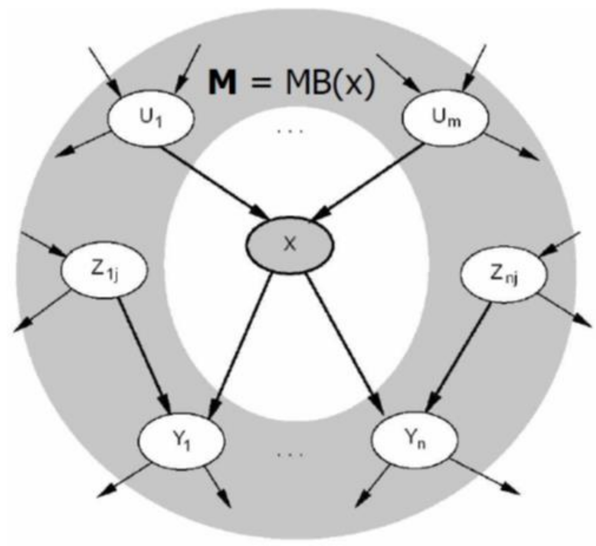
Abbreviation	
CART	Classification And Regression Trees
DL	Deep Learning
DL-AM	DL with Attention Mechanisms
DL-DHLR	DL with Disentangled Hidden Layer Representations
DL-HLV	DL with combination of Handcrafted features and Latent Variables
DL-LIME	DL with Local Interpretable Model-agnostic Explanations
DL-SA	DL with Sensitivity Analysis
DT	Decision Trees
EN	Elastic Net
GA²M	Generalized Additive Models pairwise interactions
GBM	Gradient Boosting Machines
HBN	Hierarchical Bayesian Networks
HBN-EK	HBN with Expert Knowledge
LR	Logistic Regression
LSVM	Linear Support Vector Machines
MB	MediBoost
NBN	Naïve Bayesian Networks
RF	Random Forests
RR	Ridge Regression
SVM	Support Vector Machines
SVM-LRBF	SVM with Localized Radial Basis Function kernel
SVM-RBF	SVM with Radial Basis Function kernel

Basic Idea of the Novel BN Approach

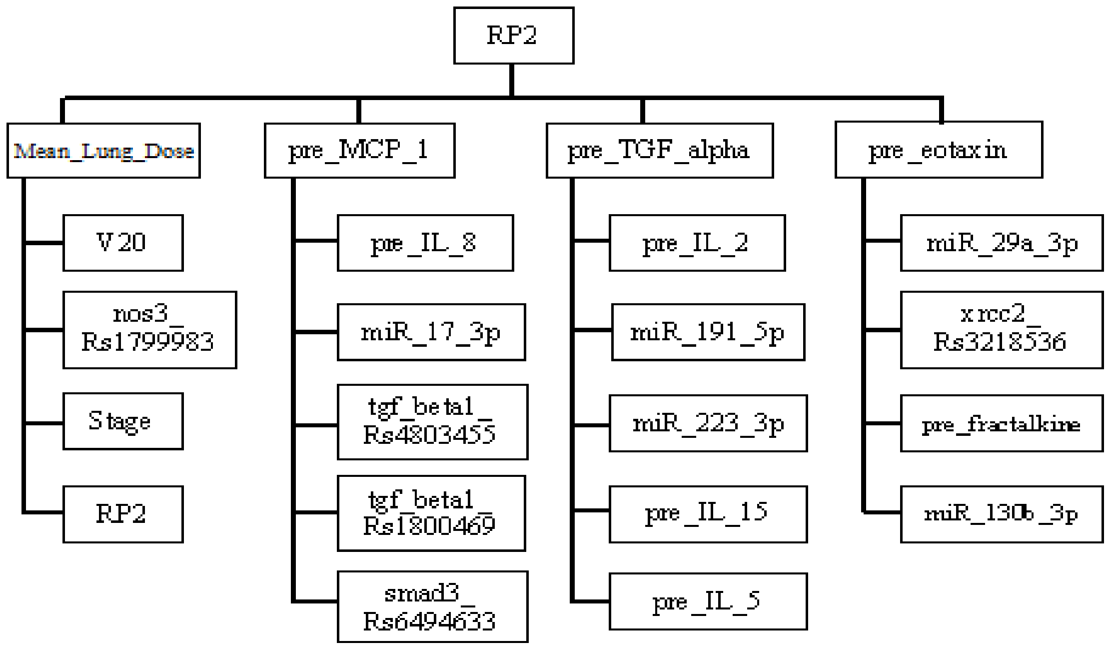


Step 1: Feature Selection

- In a BN, Markov blanket (MB) of a node is a set consisting of its inner family (parents, children, and spouses / parents of common children). Node *x* is conditionally independent of ALL other nodes given its MB.

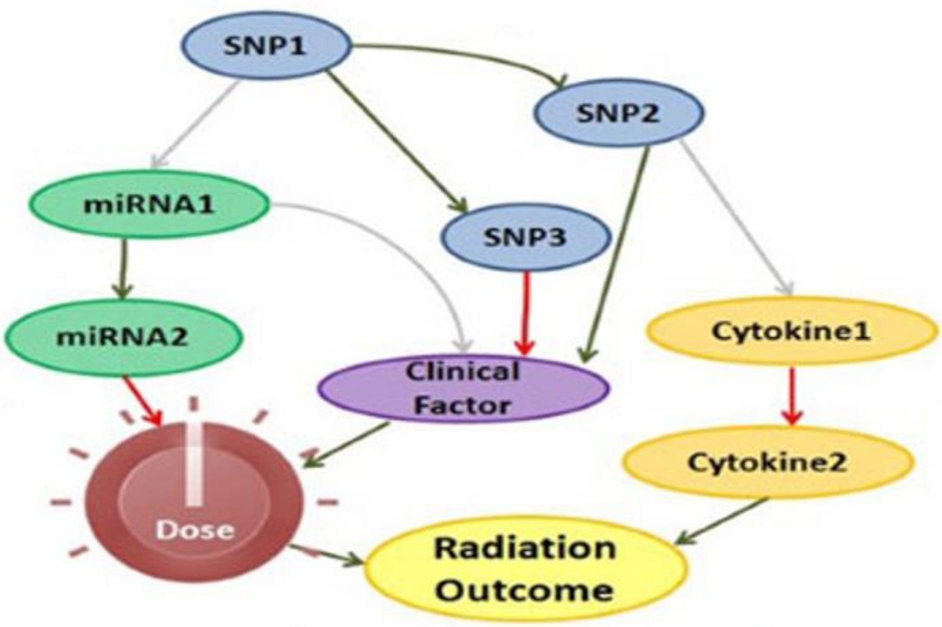


- The extended MB in the BN approach includes not only the inner family, but also each family member's next of kin.



Luo, et al., 2017

Step 2: Bayesian Networks Structure Learning



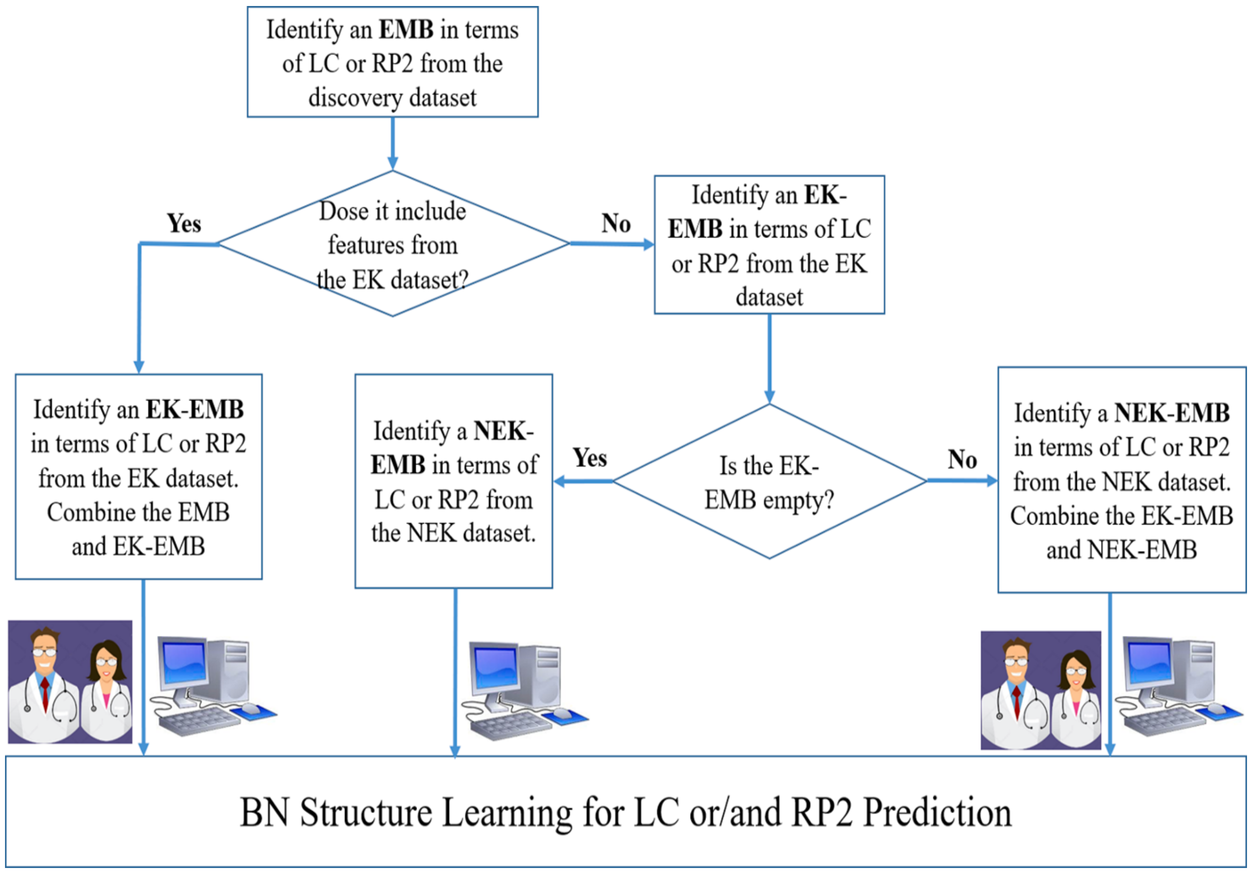
- Define biophysical and logical rules among the features;
- Compute arc strengths from bootstrap replicates;
- Identify an optimal BN structure via Tabu search;
- Eliminate leaf nodes to improve the BN's performance;
- Use cross-validation to guide the BN structure learning.

Expert Knowledge

- The physicians' trust comes from their accumulative knowledge gained from years of experience, reading articles, training, colleagues, which is named expert knowledge (EK).
- The EK bypasses complex systems and provides parsimonious solutions that focus on key aspects of a given situation.
- The EK adds new information to the models learned from data only.
- Thus, the EK has the potential for improving the choice of variables and the understanding, the trustworthiness, and the accuracy of the outcome prediction model.

Situation Awareness Bayesian Networks (SA-BNs)

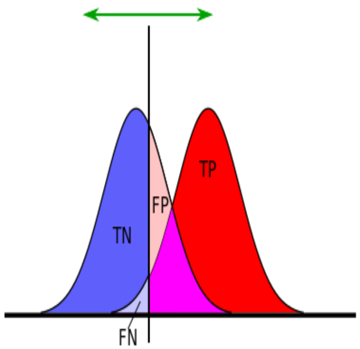
	LC	RP2
1	stage	total lung volume
2	GTV	smoking
3	tumor gEUD	lung gEUD
4	age	Chemo
5	Chemo	dose per fraction
6	GTVD95	V5
7	PTVD95	V20
8	PTV	
9	BED	
10	dose per fraction	

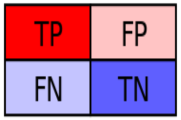


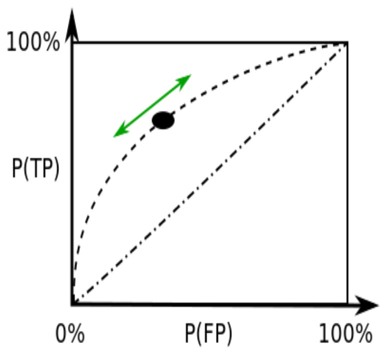
Luo, et al., 2021

Prediction Performance Measure with Multi-Focus

Prediction Performance Measure with Single Focus

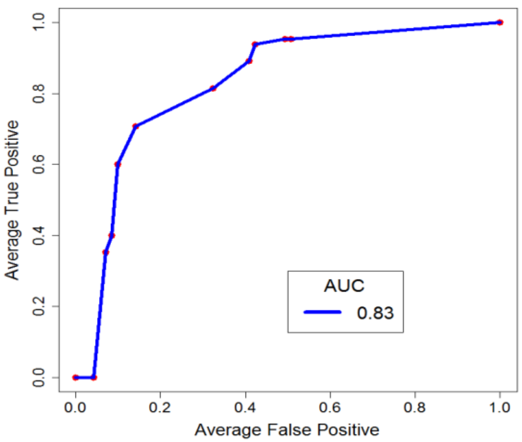




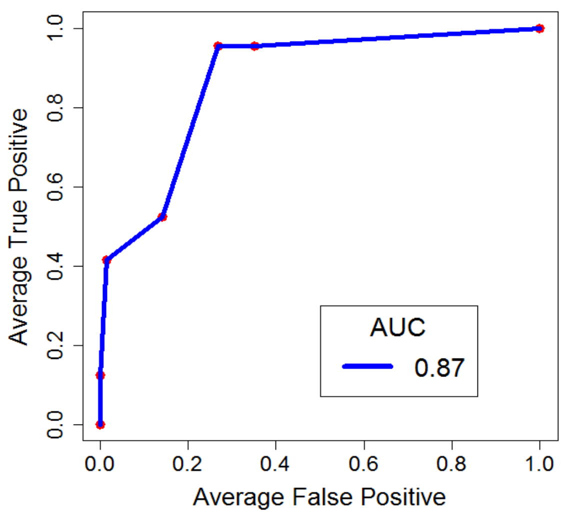


A free-response receiver operating characteristic (FROC) curve is a tool to evaluate the performance of classifying two or more characteristics within a subject simultaneously.

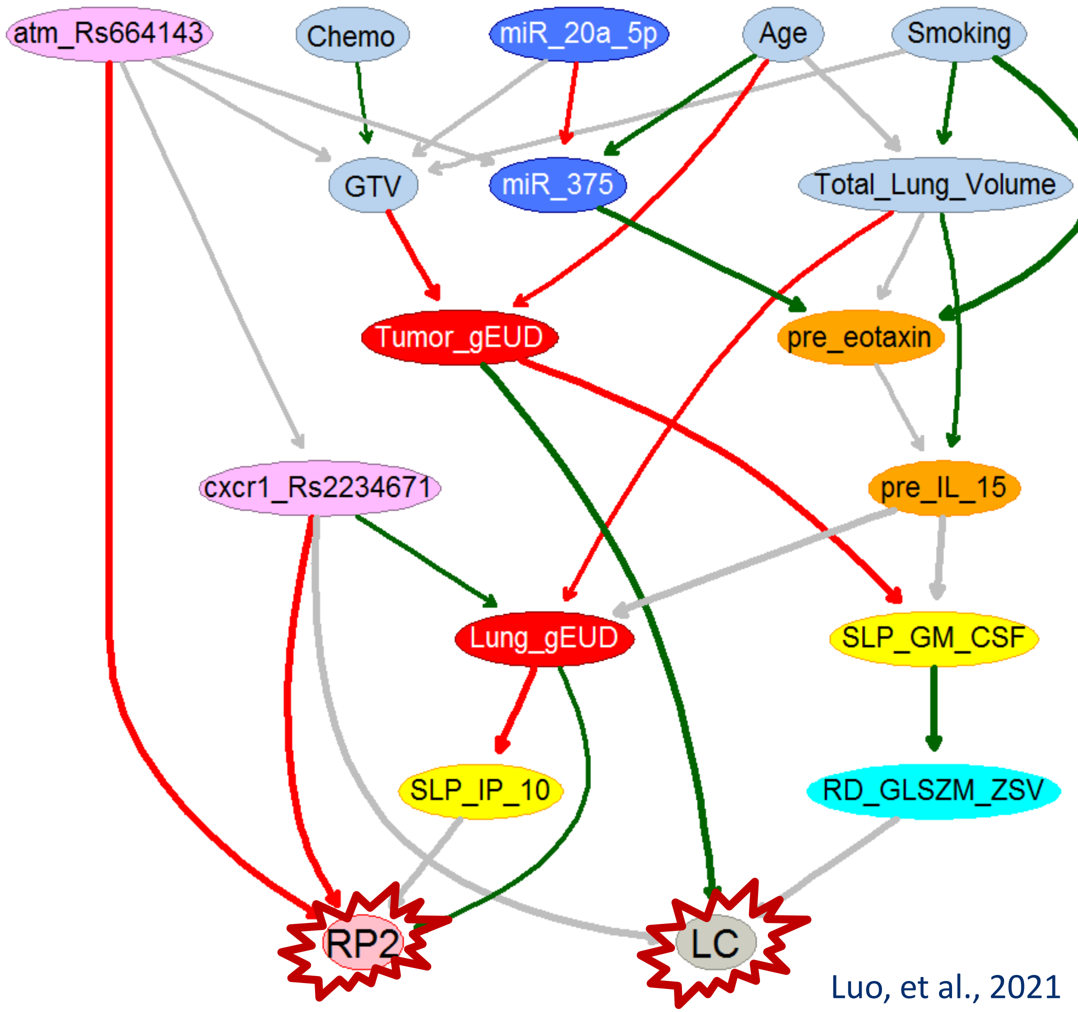
LC Prediction	RP2 Prediction	Score for FROC
correct	correct	1
correct	wrong	0.5
wrong	correct	0.5
wrong	wrong	0



During-Treatment SA-BNs for the Joint Prediction of LC and RP2

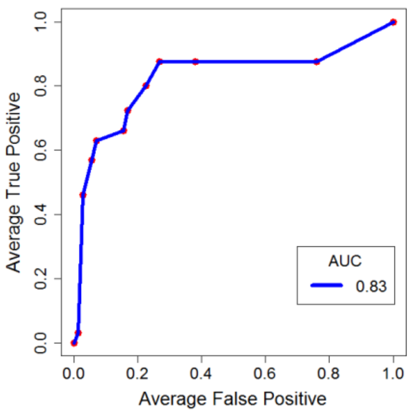


95% CI: 0.79-0.93
(2000 stratified bootstrap replicates)



Luo, et al., 2021

A Dynamic Bayesian Network to Identify Optimal Treatment Plans

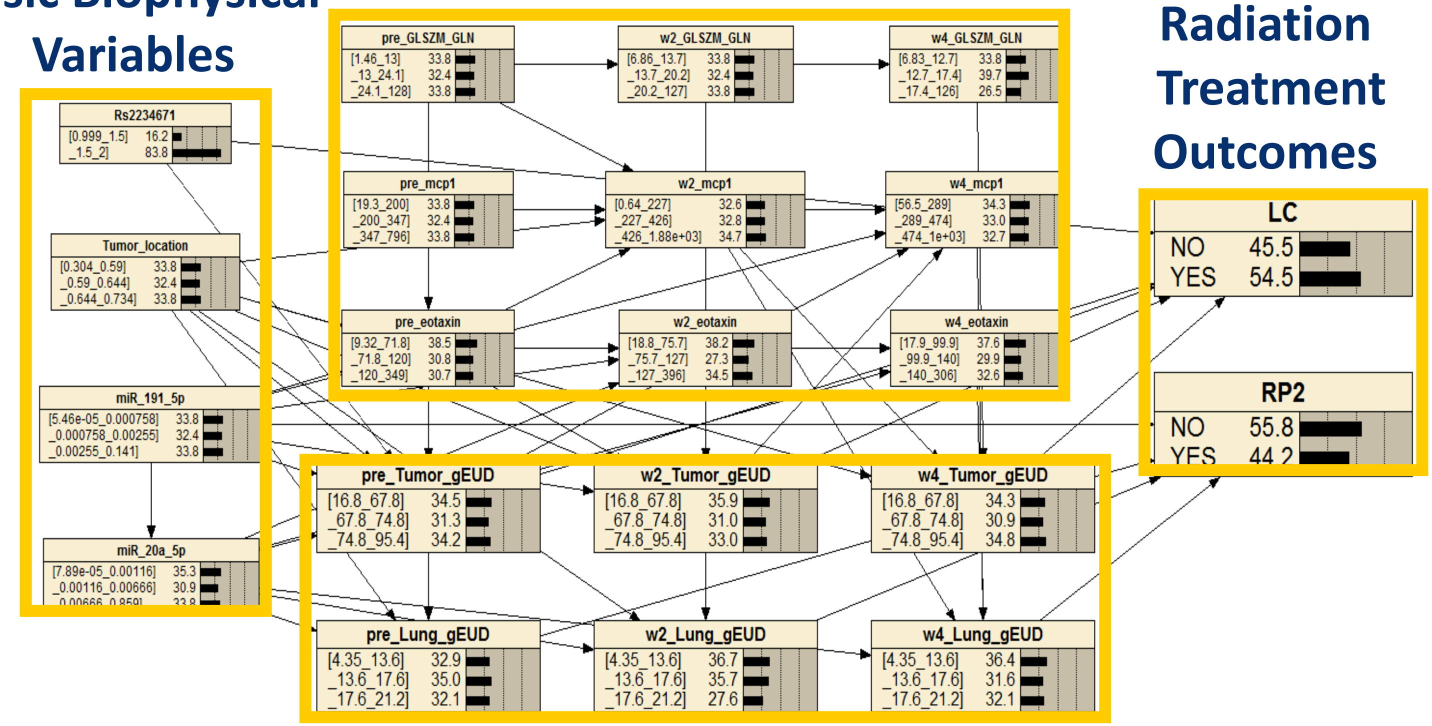


95% CI: 0.74-0.89 (2000 stratified bootstrap replicates)

Basic Biophysical Variables

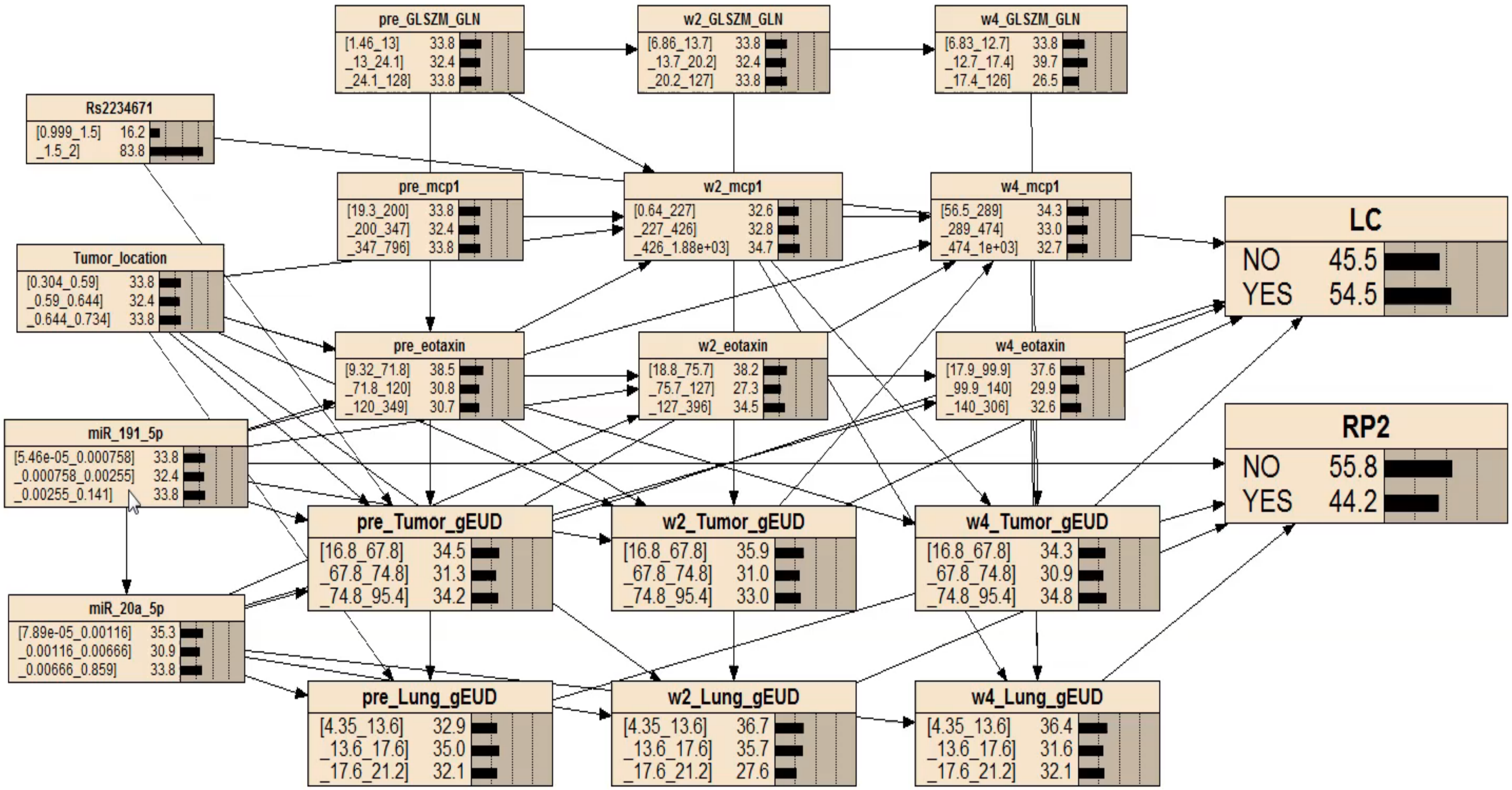
Sequential Biophysical Variables

Radiation Treatment Outcomes



Treatment Planning

Demo:
Personalized Radiation Treatment Planning



R Codes to Build a Bayesian Network after Feature Selection

\\isilfs04vs1.moffitt.org\UsersL$\44 4617\Computers Bridge\Teaching\T32\Teaching Slides\teaching codes - R Editor

```
library(cvTools)
library("MASS")
library(pROC)
library(base)
library(caret)
library(bnlearn)
library(gRbase)
library(gRain)
library("igraph")
library(Rgraphviz)
library(nlme)

########################
### Step 1: read in data
rawdata<-read.table(file.choose(),header=T)
ncol(rawdata)
nrow(rawdata)

##############################
### Step 2: data pre-processing
ABC<-rawdata[,1:2]
colnames(ABC)[1]<-paste(colnames(rawdata)[1])
colnames(ABC)[2]<-paste(colnames(rawdata)[2])

## One binary variable
AA<-as.numeric(rawdata[,3])
d1 = discretize(data.frame(AA), method = 'interval', breaks=2, ordered=TRUE, debug=TRUE)
ABC<-cbind(ABC, d1)
colnames(ABC)[3]<-paste(colnames(rawdata)[3])

## Five integer variables
d4=NULL
for (i in 1:3){
  A2<-as.numeric(as.matrix(rawdata[,3+i]))
  d4 = discretize(data.frame(A2), method = 'interval', breaks=3, ordered=TRUE, debug=TRUE)
  ABC<-cbind(ABC, d4)
  colnames(ABC)[3+i]<-paste(colnames(rawdata)[3+i])
}

## Twenty-two continuous varibales
d5 = discretize(data.frame(rawdata[,7:28]), method = 'hartemink', breaks=3, ordered=TRUE, ibreaks=3, debug=TRUE)
ABC<-cbind(ABC, d5)
A<-ABC
```

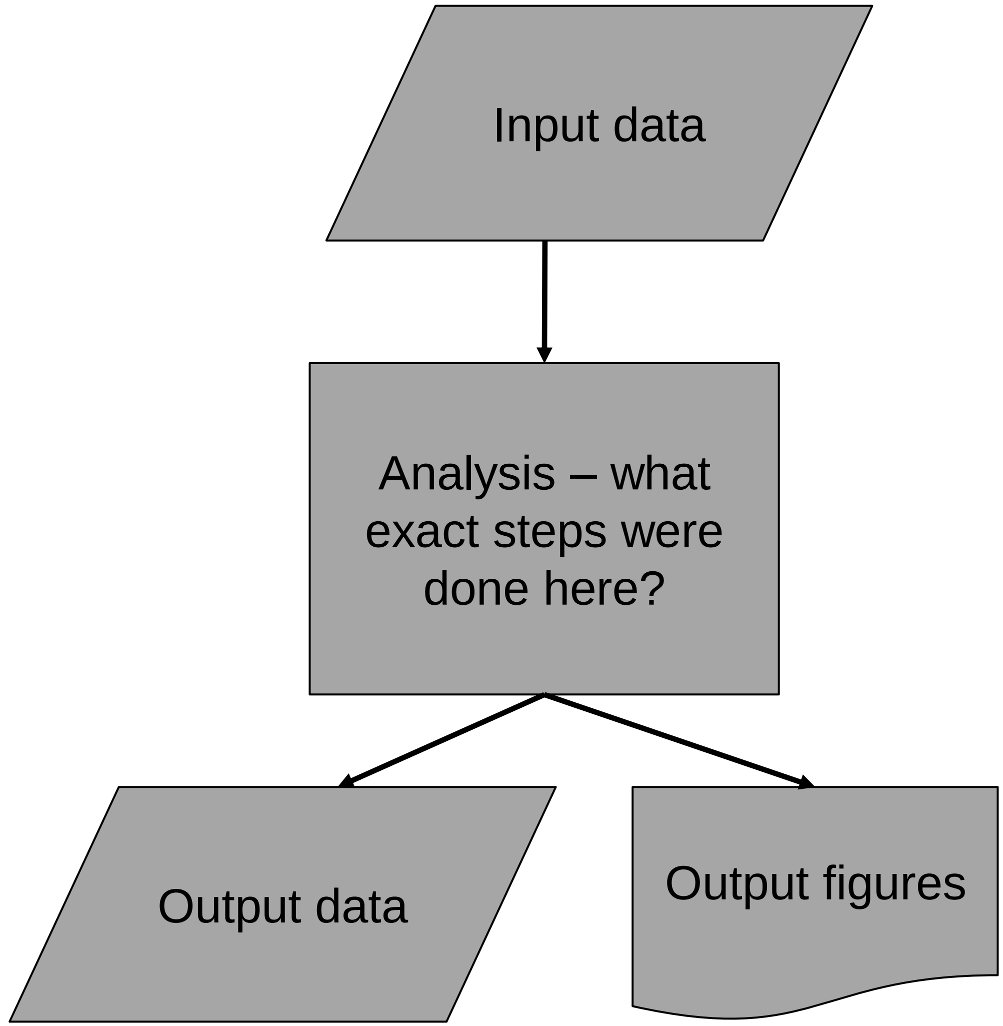
R Codes to Build a Bayesian Network (Cont.)

```
########################################
### Step 3: Develop Bayeain network(BN)

## Black list
BL<- data.frame(from=c( "LC", "LC", "LC", "LC", "LC","LC", "LC", "LC", "LC", "LC", "LC","LC",
"LC", "LC", "LC", "LC","LC","LC", "LC", "LC", "LC", "LC"......),

to=c("ercc5_Rs1047768", "miR_17_5p", "miR_191_5p","Tumor_location","pre_eotaxin","w2_eotaxin",
"w4_eotaxin","pre_MCP_1","w2_MCP_1","w4_MCP_1","pre_MTV","w2_MTV","w4_MTV","pre_GLSZM_GLN",
"w2_GLSZM_GLN","w4_GLSZM_GLN","Tumor_gEUD_F1", "Lung_gEUD_F1","Tumor_gEUD_F2", "Lung_gEUD_F2",
 "Tumor_gEUD_F3", "Lung_gEUD_F3"......))

## White list
WL = data.frame(from=c("w4_GLSZM_GLN","w4_MCP_1","Tumor_gEUD_F3", "Tumor_gEUD_F2","Tumor_gEUD_F1","Lung_gEUD_F3",
 "Lung_gEUD_F2", "Lung_gEUD_F1"), to=c("LC","RP2","LC","LC","LC","RP2","RP2","RP2"))

## Generate multiple BN structures based on bootstrap
bootBN = boot.strength(data =A[,3:ncol(A)], R = 300, algorithm = "tabu",
         algorithm.args = list(score = "bde", iss = 20, blacklist = BL, whitelist = WL, optimized=TRUE))

## Average network structure
BN_hc<- averaged.network(bootBN, threshold = 0.5)

## Plot the average Bayesian network and format its arcs
gr=strength.plot(BN_hc, bootBN, shape = "ellipse")
```

```
################################
### Step 4: Save as Netica file

BBN_hc<-as.graphNEL(BN_hc)
cad.cpt <- extractCPT(A, BBN_hc, smooth = 0.01)
BN_fit_grain <- grain(compileCPT(cad.cpt))
BN_fit<-as.bn.fit(BN_fit_grain)
write.dsc(BN_fit, file = "H:/2019/Lung Cancer/NETICA_May/DBN.dsc")
```

Testing an approach with a question

Can you go back to an analysis and recreate all the steps from the first data input file to the final output results and figures?



Homework!

- Consider how you are currently documenting your work.
- Could you go back and answer specific questions about settings, options, tools used, etc?
- If the output files were lost, could you recreate them exactly the same way?

- How might you improve your approach to reproducibility?

- **Many of us have learned the hard way! My goal for this lecture is that you won't have to!**

MOFFITT CANCER CENTER