Tour of HTT data

Brandon D. Gallas

Division of Imaging, Diagnostics, Software Reliability

Office of Science and Engineering Laboratories

Center for Devices and Radiological Health

U.S. Food and Drug Administration

Outline

- Different audiences in attendance, different goals
  - Rookies in R, Rstudio, Git, GitHub
  - Collaborators wanting clarity on the data so they can explore

  This is the priority today: 30-40 minutes

  - Teammates that are already analyzing the data
  - Collaborators and colleagues that want to see agreement for quantitative values

  10-20 minutes

- “Case” == region of interest (ROI)
- 64 Whole Slide Images (WSI)
- 10 ROIs per WSI

- <u>Modalities</u>
- caMicroscope == digital platform
- PathPresenter == digital platform
- eeDAP == microscope platform

sTIL = stromal tumor-infiltrating lymphocytes

Data-collection Description

- Classify the ROI
- If ROI label is appropriate

  … estimate percent (tumor-associated) stroma
- If percent stroma > 0

  … estimate density of sTILs

Getting Started

- Public Repository
  - https://github.com/DIDSR/HTT
  - See README and Release 1.0.0
- Private Repository
  - https://github.com/DIDSR/HTTdev
  - See README and Release 1.1.0
- Branch of Private Repository
  - https://github.com/DIDSR/HTTdev/tree/RenameTrainingCases
  - Current working branch

- Download options
  - Explicit download of R package from release page
  - Explicit download of repository as a zip file
  - Clone a copy (cannot push to repo)
  - Fork and clone a copy (can push to repo)

- How to compile R package from a copy or clone of the repository

What's your GitHub username

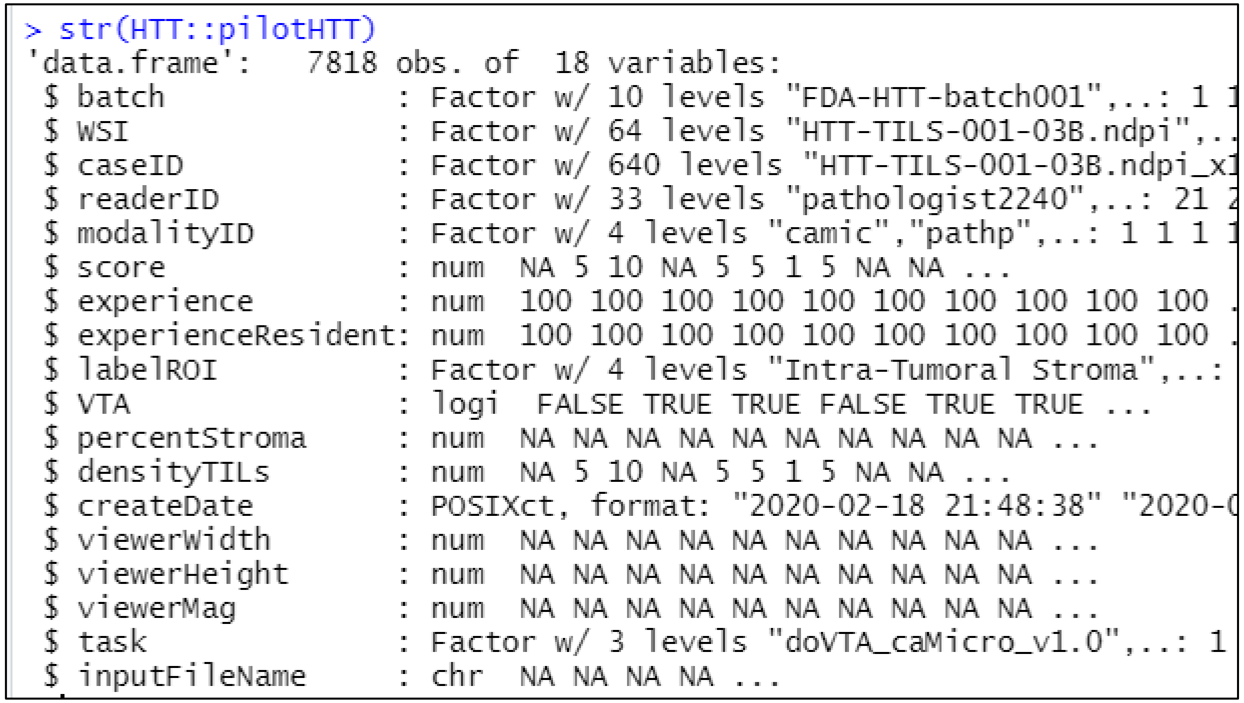
The repositories all have different names, but the R package is always HTT.

(As data and code move from one repository to the next, we don't want to have to rename variables etc.)

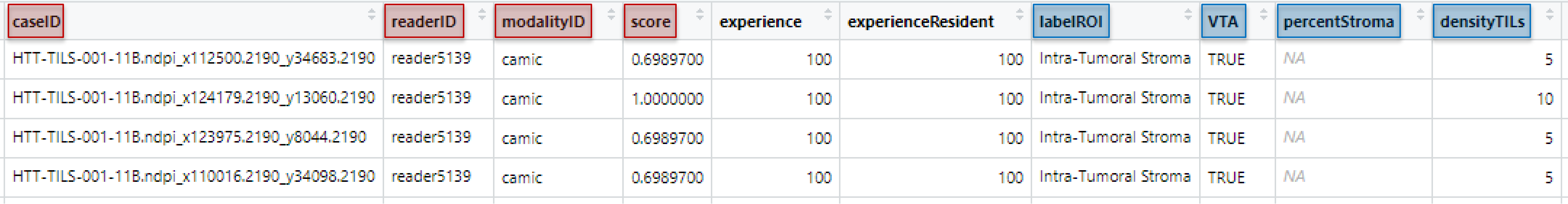
Stay up to date with the source of your download

R package objects

- The primary data object
  - HTT::pilotHTT
- Generic name: `mrmcDF`
  - Primary variables:
  - readerID
  - caseID
  - modalityID
  - score



```
> str(HTT::pilotHTT)
'data.frame':   7818 obs. of  18 variables:
 $ batch             : Factor w/ 10 levels "FDA-HTT-batch001",..: 1 1
 $ WSI               : Factor w/ 64 levels "HTT-TILS-001-03B.ndpi",..
 $ caseID            : Factor w/ 640 levels "HTT-TILS-001-03B.ndpi_x1
 $ readerID          : Factor w/ 33 levels "pathologist2240",..: 21 2
 $ modalityID        : Factor w/ 4 levels "camic","pathp",..: 1 1 1 1
 $ score             : num  NA 5 10 NA 5 5 1 5 NA NA ...
 $ experience        : num  100 100 100 100 100 100 100 100 100 100 .
 $ experienceResident: num  100 100 100 100 100 100 100 100 100 100 .
 $ labelROI          : Factor w/ 4 levels "Intra-Tumoral Stroma",..:
 $ VTA               : logi  FALSE TRUE TRUE FALSE TRUE TRUE ...
 $ percentStroma     : num  NA NA NA NA NA NA NA NA NA NA ...
 $ densityTILs       : num  NA 5 10 NA 5 5 1 5 NA NA ...
 $ createDate        : POSIXct, format: "2020-02-18 21:48:38" "2020-0
 $ viewerWidth       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ viewerHeight      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ viewerMag         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ task              : Factor w/ 3 levels "doVTA_caMicro_v1.0",..: 1
 $ inputFileName     : chr  NA NA NA NA ...
```

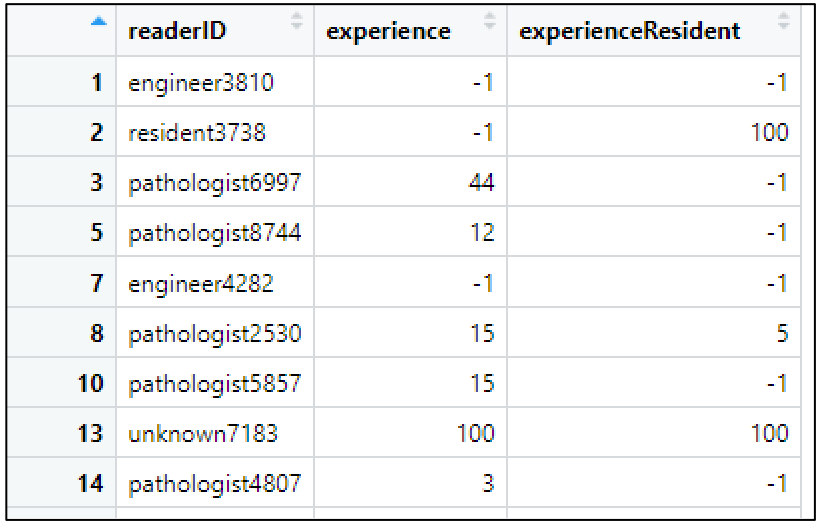


caseID	readerID	modalityID	score	experience	experienceResident	labelROI	VTA	percentStroma	densityTILs
HTT-TILS-001-11B.ndpi_x112500.2190_y34683.2190	reader5139	camic	0.6989700	100	100	Intra-Tumoral Stroma	TRUE	NA	5
HTT-TILS-001-11B.ndpi_x124179.2190_y13060.2190	reader5139	camic	1.0000000	100	100	Intra-Tumoral Stroma	TRUE	NA	10
HTT-TILS-001-11B.ndpi_x123975.2190_y8044.2190	reader5139	camic	0.6989700	100	100	Intra-Tumoral Stroma	TRUE	NA	5
HTT-TILS-001-11B.ndpi_x110016.2190_y34098.2190	reader5139	camic	0.6989700	100	100	Intra-Tumoral Stroma	TRUE	NA	5

Possible "Scores"

R package objects

- `View(HTT::cleanReaders)`
- `library(HTT)`
- `View(cleanReaders)`
  - Name reflects pathologist experience
  - Name includes a random 4-digit number



	readerID	experience	experienceResident
1	engineer3810	-1	-1
2	resident3738	-1	100
3	pathologist6997	44	-1
5	pathologist8744	12	-1
7	engineer4282	-1	-1
8	pathologist2530	15	5
10	pathologist5857	15	-1
13	unknown7183	100	100
14	pathologist4807	3	-1

- These primary variables are factors
  - `levels(pilotHTT$readerID)`
  - `levels(pilotHTT$caseID)`
  - `levels(pilotHTT$modalityID)`

- Factors are often used in R
  - Two states: index and value
  - `as.character(pilotHTT$readerID[1000])`
  - `as.numeric(pilotHTT$readerID[1000])`

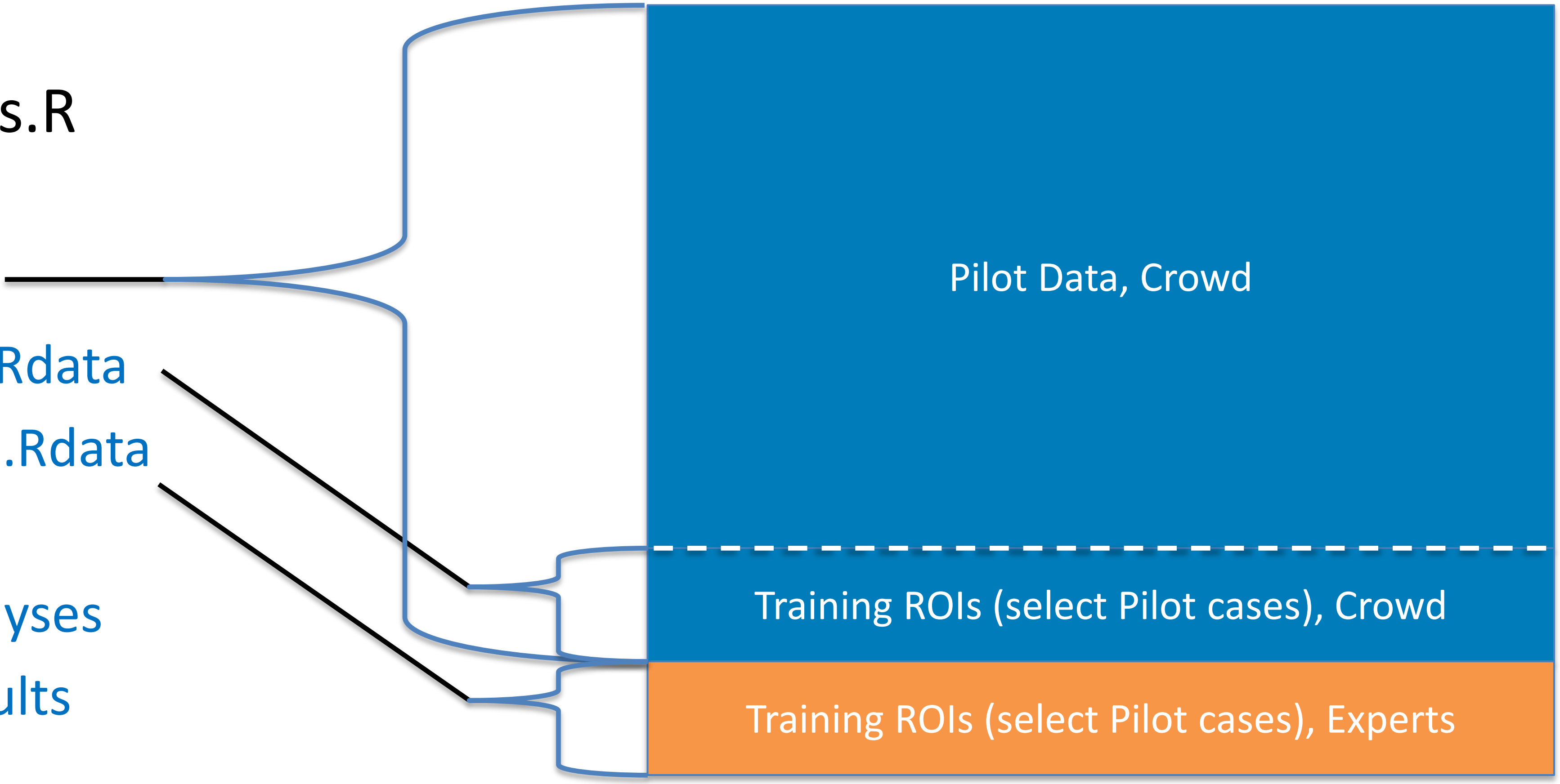
R package objects

- I often use factors to split the data
  - Create a list of reader-specific data frames
  - `result <- split(pilotHTT, pilotHTT$readerID)`
  - `names(result)`

- Be careful How you reference a list (str = structure function)
  - This returns a list element
  - `str(result[1])`
  - This returns the content of a list element
  - `str(result[[1]])`

R package objects

- analyzeTrainingSets.R
  - Curates the data
  - resultsPilot.Rdata
  - resultsTrainCrowd.Rdata
  - resultsTrainExperts.Rdata
  - Sends data for analyses
  - Shows analysis results
  - `names(resultsTrainCrowd)`


Pilot Data, Crowd
Training ROIs (select Pilot cases), Crowd
Training ROIs (select Pilot cases), Experts

R package objects

- initData.R
  - Discuss statsByCase
- binShow.R
- aucShow.R