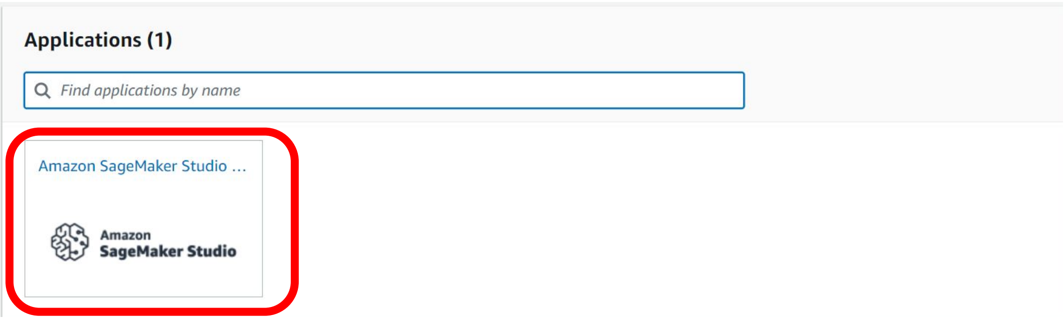
Session 2: Longitudinal Trajectory Analysis

Open a browser and visit: https://healthdatascience.awsapps.com/start/ Enter the email you used to register for the workshop as your username.

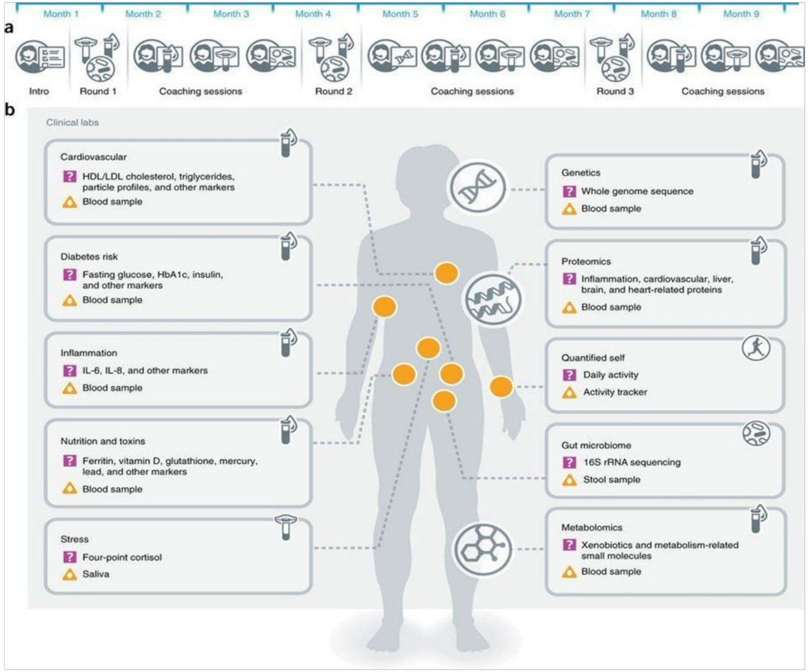
Check your email for a verification code and then follow the instructions to create your account password. After that you should be signed in.

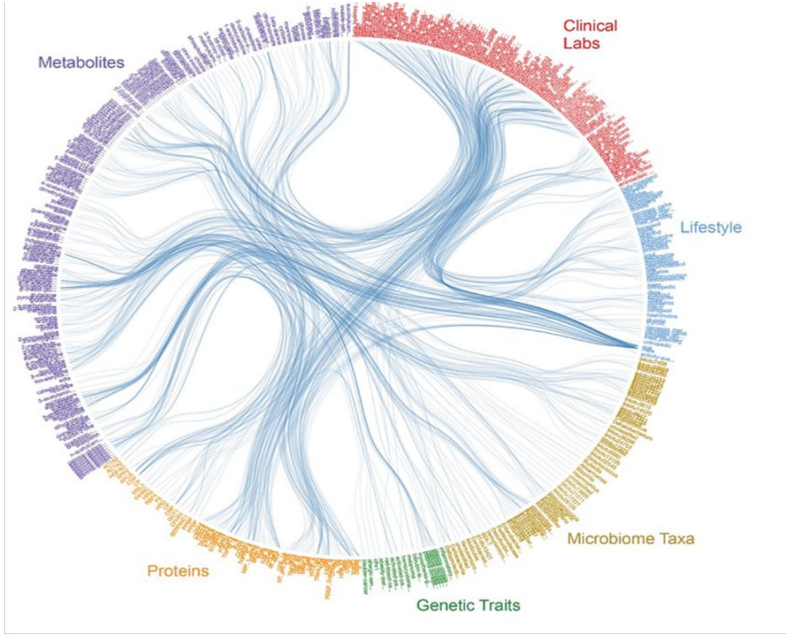
<u>On your dashboard under **Applications**, click on Amazon SageMaker Studio link to open SageMaker.</u>



- ***Look for the email also in your junk folder***

Review: The Arivale Dataset



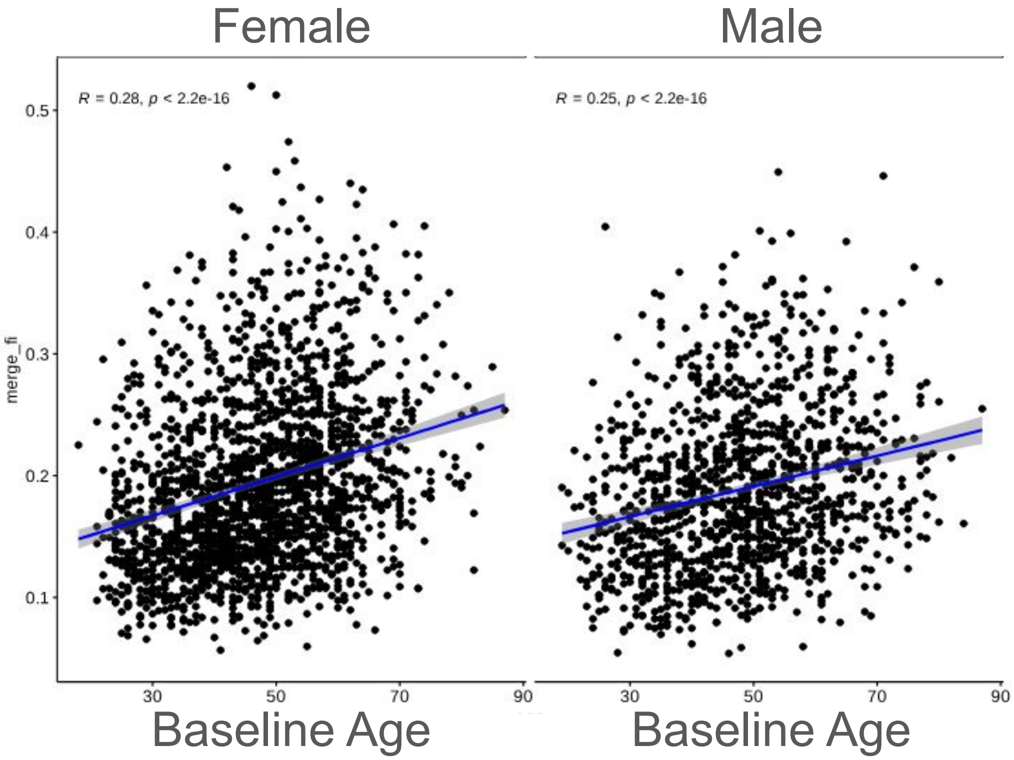


+Wellness coaching

Price et al, Nat Biotechnol, 2018

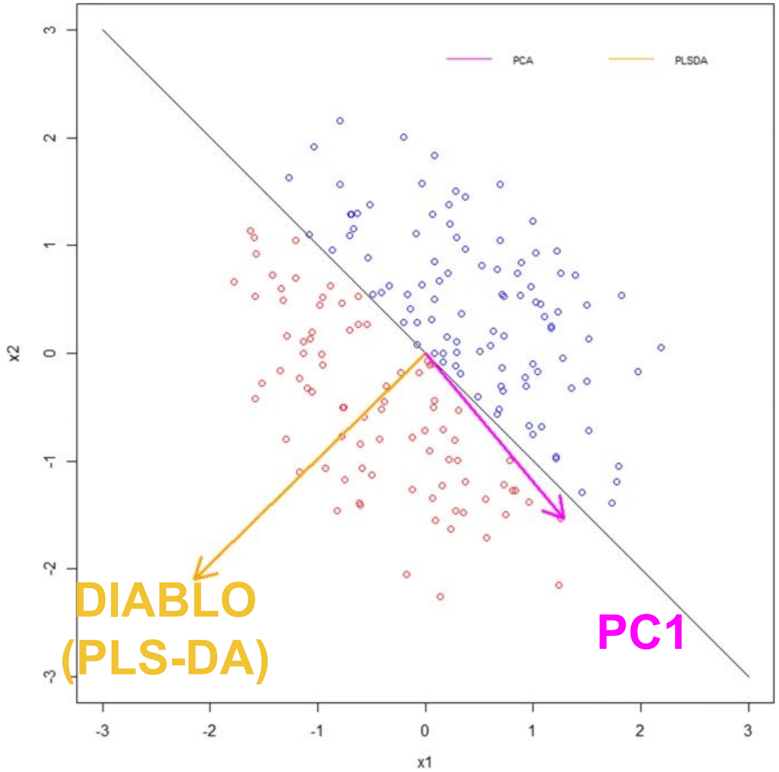
Review: Frailty Index is a Fraction of Health Defects

- Self-Report FI (35 items)
    - Disease (15 items)
    - Activity (9 items)
    - Satisfaction (6 items)
    - Medication (3 items)
    - Digestion (2 items)
- Lab FI (34 items)
    - Blood test items (29 items)
    - Blood pressure items (5 items)
- Combined FI (69 items)
    - The combination of the above two



Session 2.1: DIABLO Analysis

- Session 1
  - Trends in data (PCA)
  - Multi-omic correlations
  - Identified clusters
  - Cluster eigenvalues
- Session 2.1
  - Trends in the outcome
  - Multi-omic model
  - Cross-validation



DIABLO Overview

1. Data preparation
   a. Outcome: self-reported frailty index
   b. Baseline data only
   c. Proteomics, metabolomics, and lab tests
2. sPLS-DA
   a. Proteomics
   b. Metabolomics
   c. Lab tests
3. Block sPLS-DA (all three ‘omics)
   a. Sparsity parameter optimization
   b. Model fitting

Frailty Index (FI): fraction of health deficits

- Self-Report FI (35 items) – Baseline questionnaire (once)
  - Disease (15 items)
  - Activity (9 items)
  - Satisfaction (6 items)
  - Medication (3 items)
  - Digestion (2 items)
- Lab FI (34 items) – Longitudinal, every 6 months
  - Blood test items (29 items)
  - Blood pressure items (5 items)
- Combined FI (69 items) ⇦ Session 1 (baseline labs)
  - The combination of the above two
- Comparison (Spearman's rank correlation of quintiles)
  - Self x Lab 0.367
  - Self x Combined 0.730
  - Lab x Combined 0.843

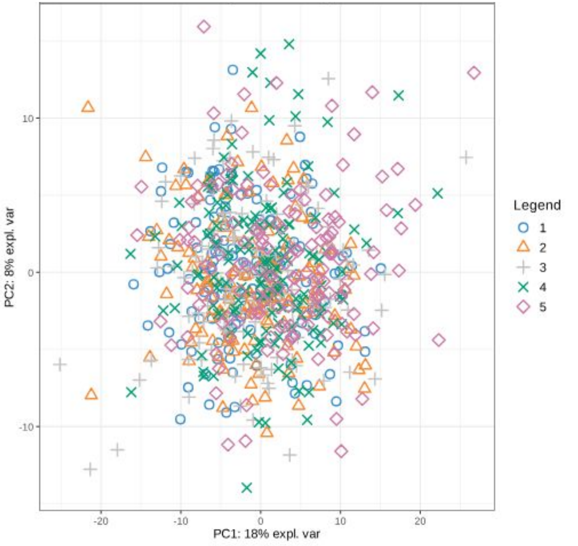
Frailty Index (FI): fraction of health deficits

- Self-Report FI (35 items) – Baseline questionnaire (once) ⇐ Session 2.1
  - Disease (15 items)
  - Activity (9 items)
  - Satisfaction (6 items)
  - Medication (3 items)
  - Digestion (2 items)
- Lab FI (34 items) – Longitudinal, every 6 months ⇐ Session 2.2
  - Blood test items (29 items)
  - Blood pressure items (5 items)
- Combined FI (69 items) ⇐ Session 1 (baseline labs)
  - The combination of the above two
- Comparison (Spearman's rank correlation of quintiles)
  - Self x Lab 0.367
  - Self x Combined 0.730
  - Lab x Combined 0.843

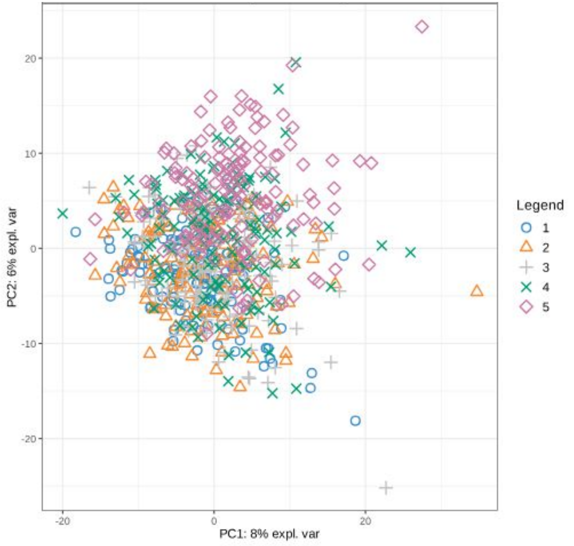
To the Notebook!

PCA of Baseline ‘Omics Data

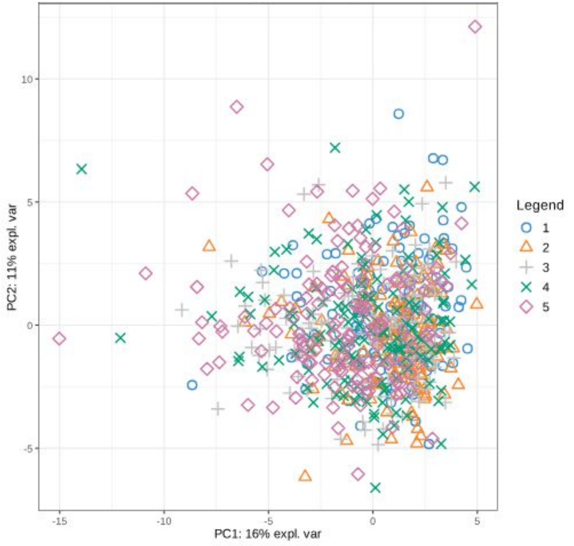
Proteomics (274)



Metabolomics (703)



Clinical Tests (47)



Color and shape: Frailty Index Quintile (e.g. Q5 are the most frail 20%)

sPLS-DA of Baseline ‘Omics Data

```
[27]: plsda.met <- mixOmics::plsda(mets_mat, Outcome, ncomp = 5)

      perf.plsda.met  <- mixOmics::perf(plsda.met, validation = 'Mfold', folds = 3,
                                         progressBar = TRUE,
                                         nrepeat = 10)   ### This is a low number of repeats that␣
       ↪should be increased for a better analysis. Its slow.

      plot(perf.plsda.met, sd = TRUE, legend.position = 'horizontal')
```

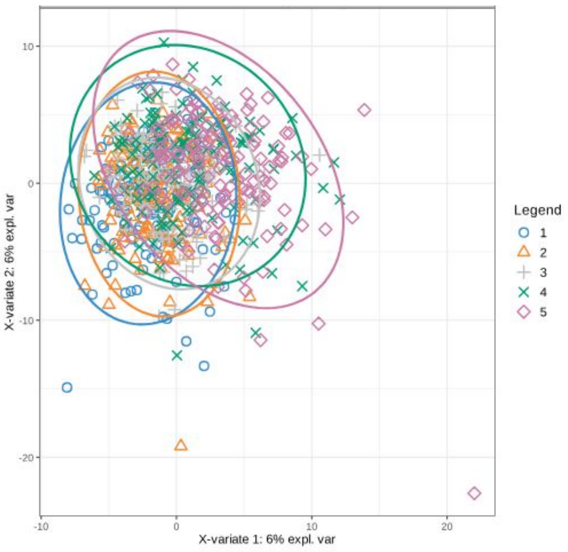
To the Notebook!

sPLS-DA of Baseline ‘Omics Data

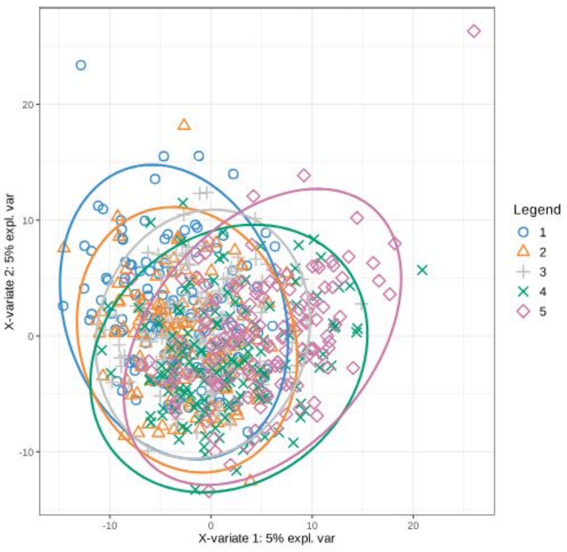
Area Under ROC, predicting each quintile

SelfFI	Block Metabolites	Proteins	Clinical
Q1	0.7583	0.7519	0.7122
Q2	0.6429	0.6525	0.6236
Q3	0.5567	0.5419	0.5498
Q4	0.6800	0.6505	0.5912
Q5	0.8010	0.8223	0.7745

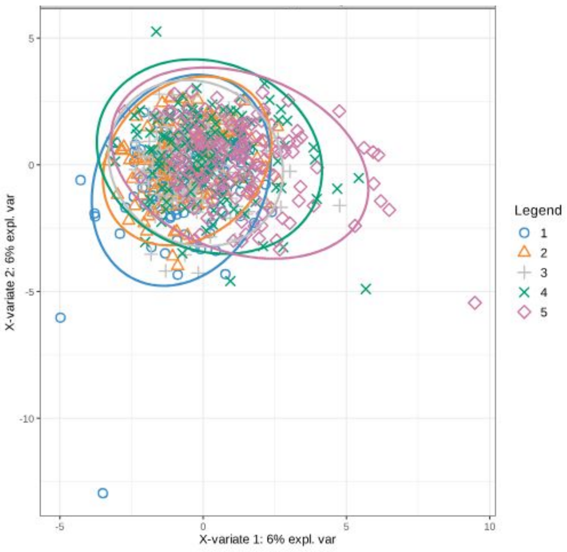
Proteomics (274)



Metabolomics (703)



Clinical Tests (47)



Block sPLS-DA

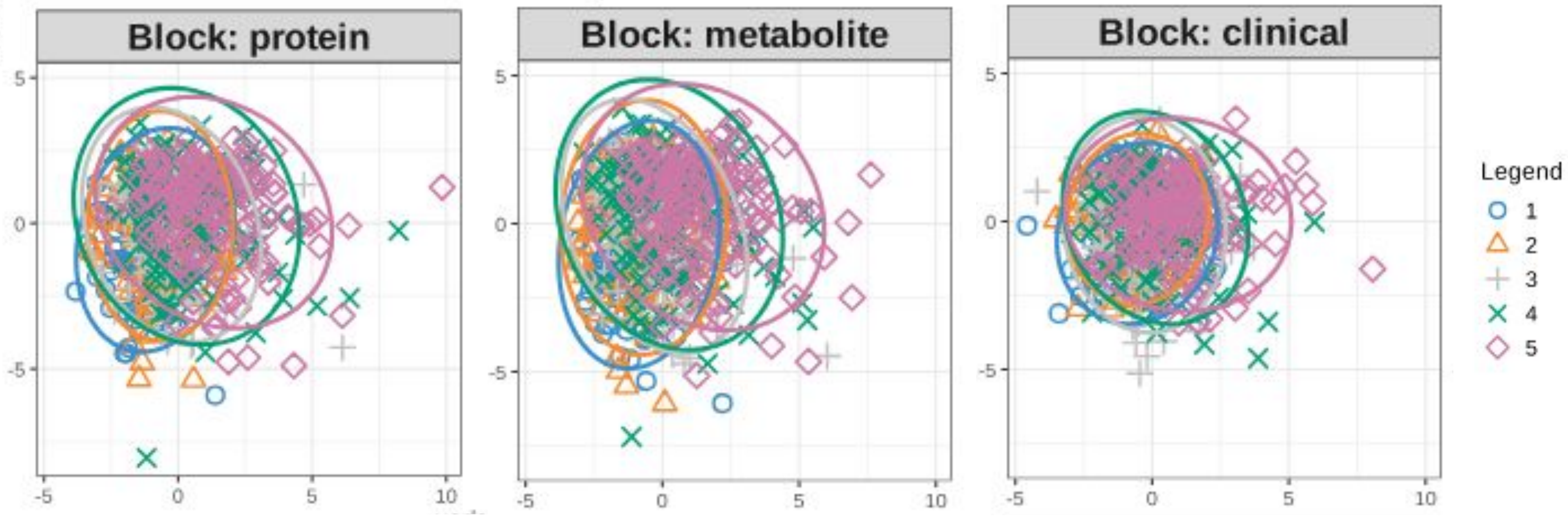
A matrix: 3×3 of type dbl

	metabolite	protein	clinical
metabolite	0.0	0.1	0.1
protein	0.1	0.0	0.1
clinical	0.1	0.1	0.0

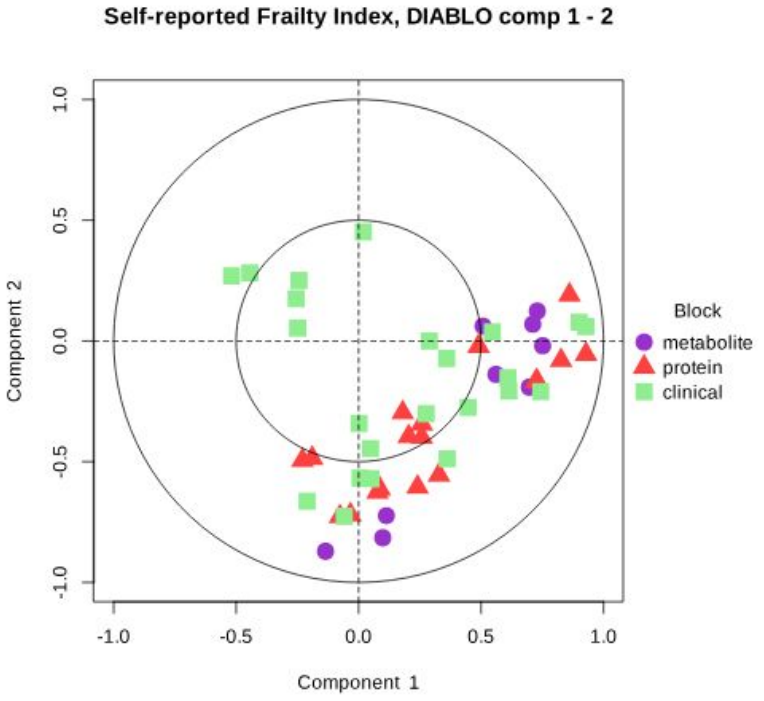
[51]:
```
# This takes a 20 min to run!
diablo.selfFI <- block.plsda(X, Outcome, ncomp = 5, design = design)

perf.diablo.selfFI = mixOmics::perf(diablo.selfFI, validation = 'Mfold',
                                    progressBar = TRUE,
```

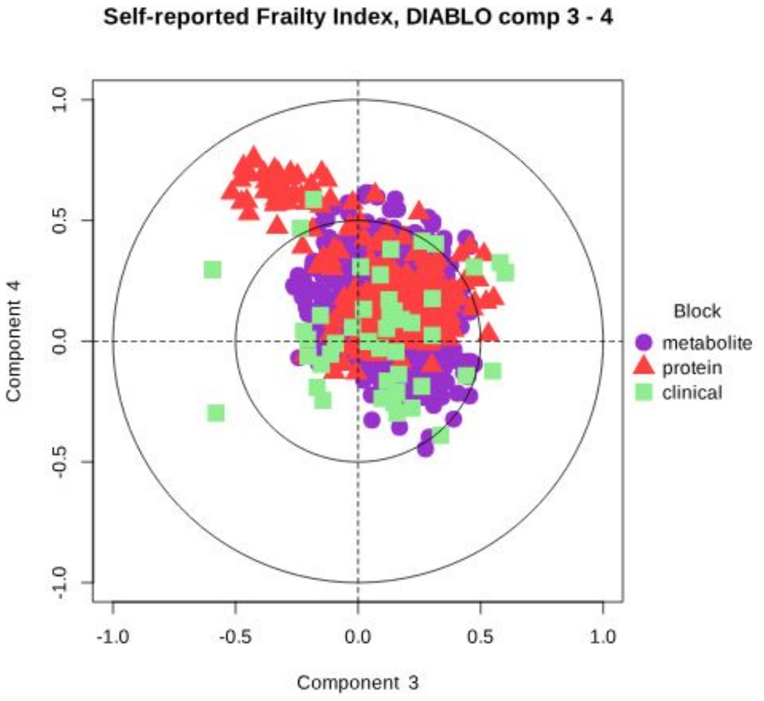
To the Notebook!

Block sPLS-DA



Visualizing DIABLO model components





DIABLO Model: Component 1

Proteins		Lab Tests		Metabolites	
FABP4	**0.930**	INSULIN	**0.689**	hydroxyasparagine	**0.522**
LEP	**0.347**	HOMA-IR	**0.533**	N-stearoyl-sphinganine	**0.438**
		LPIR	**0.385**	cortolone glucuronide	**0.366**
				N-stearoyl-sphingosine	**0.311**
				5-methylthioadenosine	**0.275**
				1-carboxyethylphenylalanine	**0.253**

DIABLO Model: Component 2

Proteins	
AGRP	-0.763
NT-proBNP	**0.413**
NPPB	**0.354**
GDF15	**0.242**

Lab Tests	
BUN/CREAT	**0.561**
POTASSIUM	-0.318
ALBUMIN	-0.453
GFR, MDRD	-0.246
PROTEIN	-0.232
CREATININE	-0.228
VIT D, 25-OH	**0.227**
EPA	**0.124**

Metabolites	
DHEA-S	-0.557
androstenediol(3b,17b)S2	-0.452
pregnenediol-S	-0.292
pregnenetriol-S2	-0.284
pregnenediol-S2	-0.259
androstenediol(3a,17a)S	-0.253

Session 2.2: Longitudinal Data Analysis

Session 2.2: Longitudinal Data Analysis

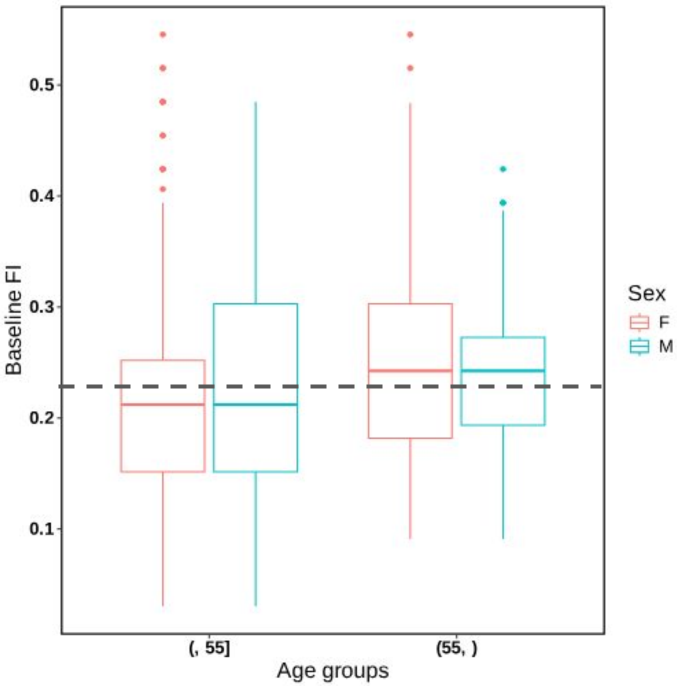
- Measurements repeated over time are often *correlated*, not independent
- Appropriate analytical methods depend on…
  - Number of repetitions (as few as 2, as many as thousands, millions)
  - Data type (categorical, counts, money, continuous) and missingness
  - Synchronous (e.g. stock prices, EEG) vs asynchronous (e.g. blood tests) data collection
  - Handling of collection failures (e.g. recollection of one failed tube out of 5 in a blood draw)
  - Systems biology (‘omics) data can be expected to have a variety of data collection issues!
- Potential goals
  - Predict future values
  - Identify unknown correlations among measured analytes (systems analysis)
  - Determine differences between groups (e.g. treated vs untreated patients; same treatment of different patient subgroups)
  - Identify causal relationships (causes precede effects in time).

Generalized Linear Mixed-Effects Models (GLMMs)

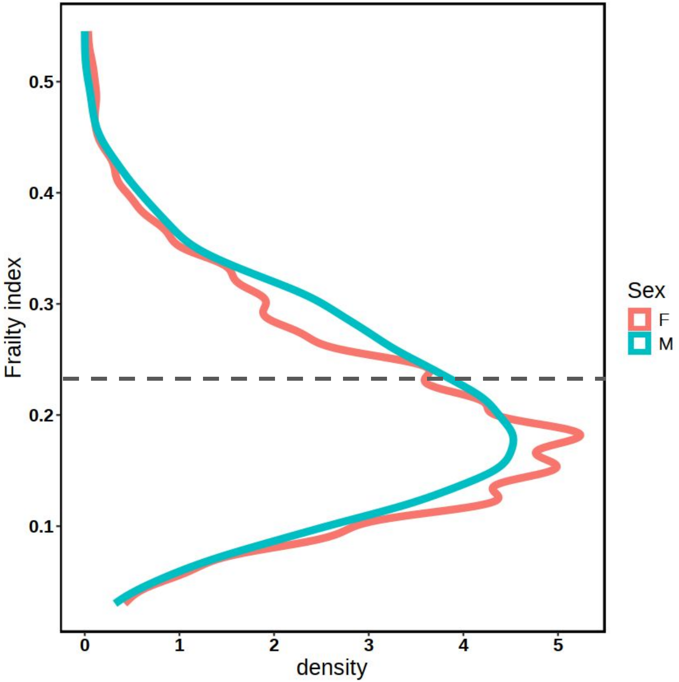
- **Generalized Linear Model (GLM):** $g(Y) \sim A + B + \ldots + \epsilon$
  - Response **Y** may be **binary**, **categorical**, or **continuous**
  - **Link function g(*x*)** maps **Y** to a **continuous** value
    - Must be invertible: **Y** $\sim g^{-1}(g(Y))$
    - g(x) = x (Identity), = *1/x* (inverse), = *log(x)*, = x / (1+x) (logit), = Pr{ N(0,1) < x } (Probit)
  - **Fixed effects** (A + B + …)
    - Explicit parameters
  - **Error ϵ distribution** may be any distribution in the Exponential family
    - Gaussian, Poisson, Binomial, Negative Binomial, Bernoulli, Exponential, Gamma
- **GLMM**: $g(Y) \sim A + B + \ldots + (C + D + \ldots \mid \text{Grouping}) + \epsilon$
  - **Random effects** (C + D + … | Grouping)
    - Identical *within* but independent *across* groups
    - Implicit parameters (no point estimates; average over a distribution)

Additional information: https://stats.oarc.ucla.edu/wp-content/uploads/2022/03/generalized_linear_new.html#(1)

Frailty Index increases with age

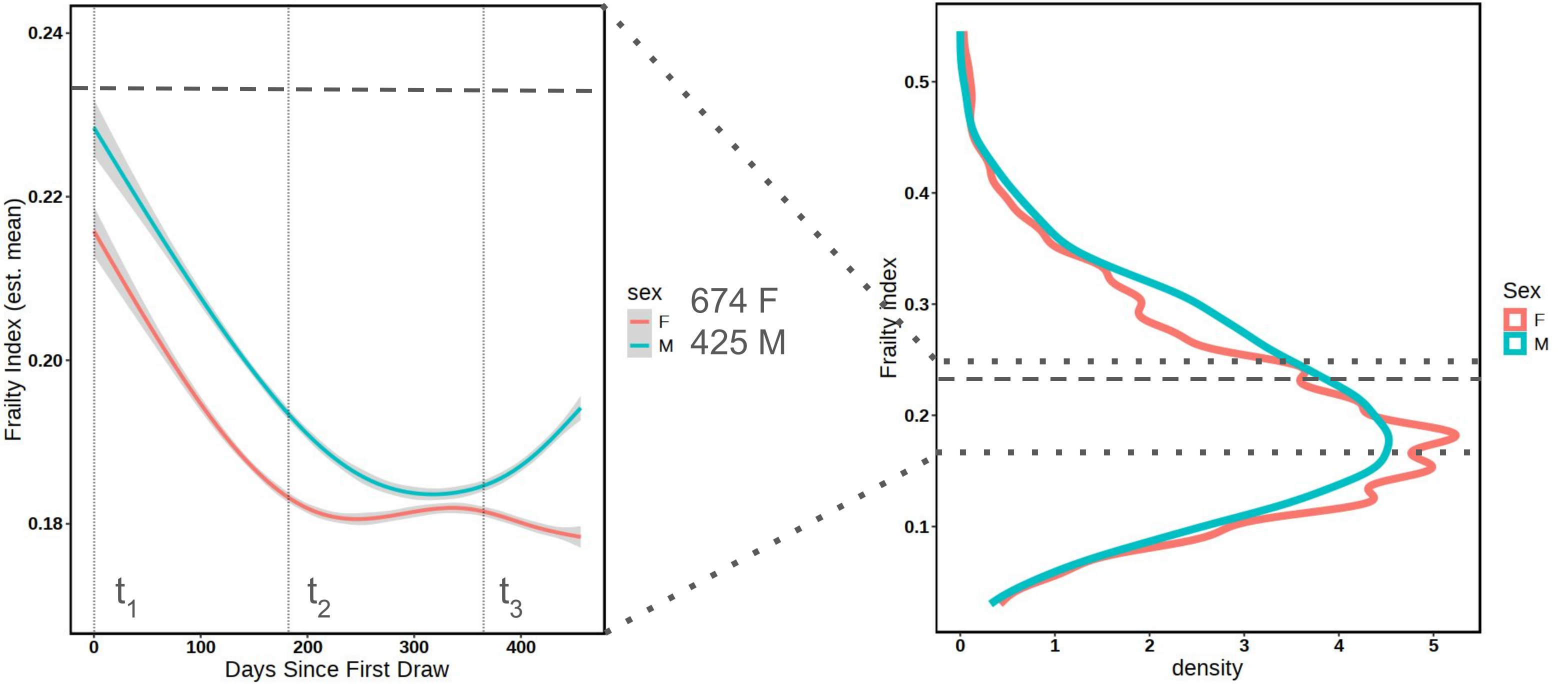


674 F
425 M



$$FI \sim \underbrace{t_1 + t_2 + t_3}_{\text{— spline —}} + Age + PC_1 + PC_2 + PC_3 + PC_4 + \underbrace{(1 + t_1 \mid client)}_{\text{— AIC —}}$$

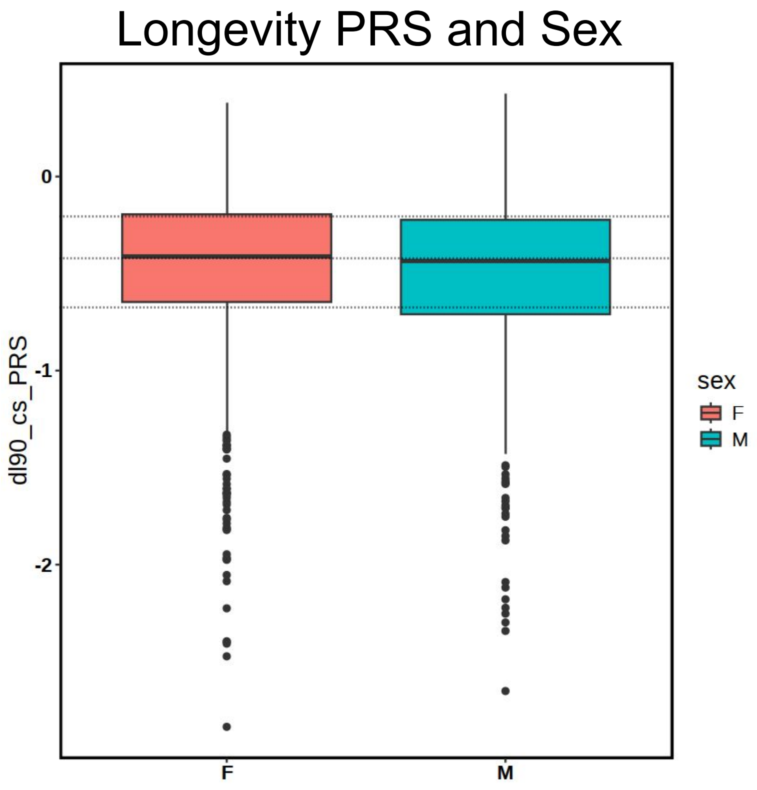
Frailty Index can be lowered by lifestyle changes

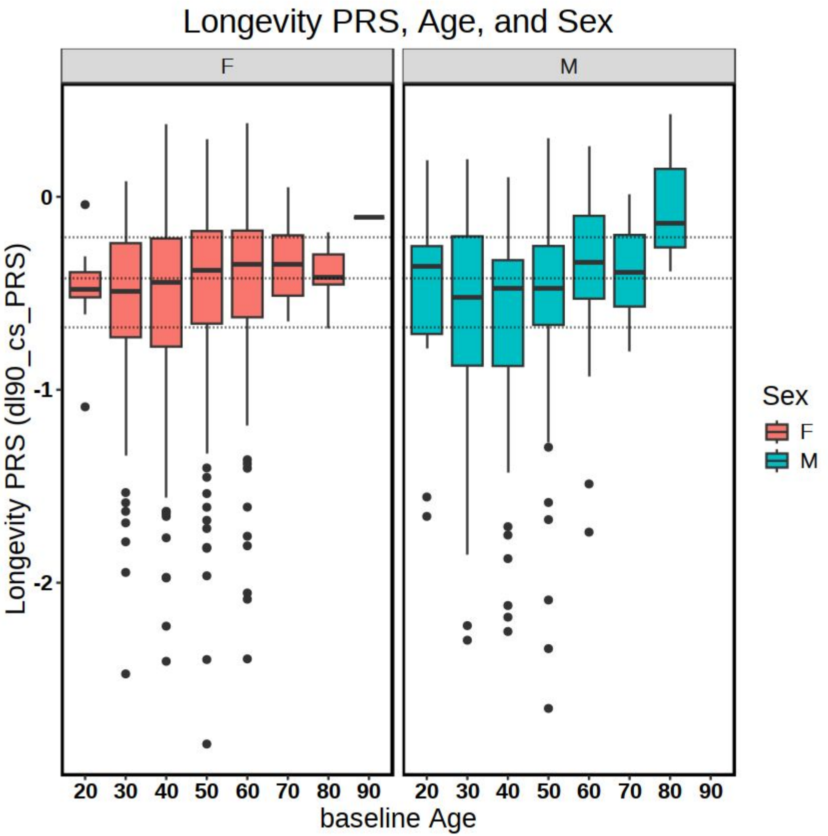


$$FI \sim t_1 + t_2 + t_3 + Age + PC_1 + PC_2 + PC_3 + PC_4 + (1 + t_1 \mid client)$$

— spline — — AIC —

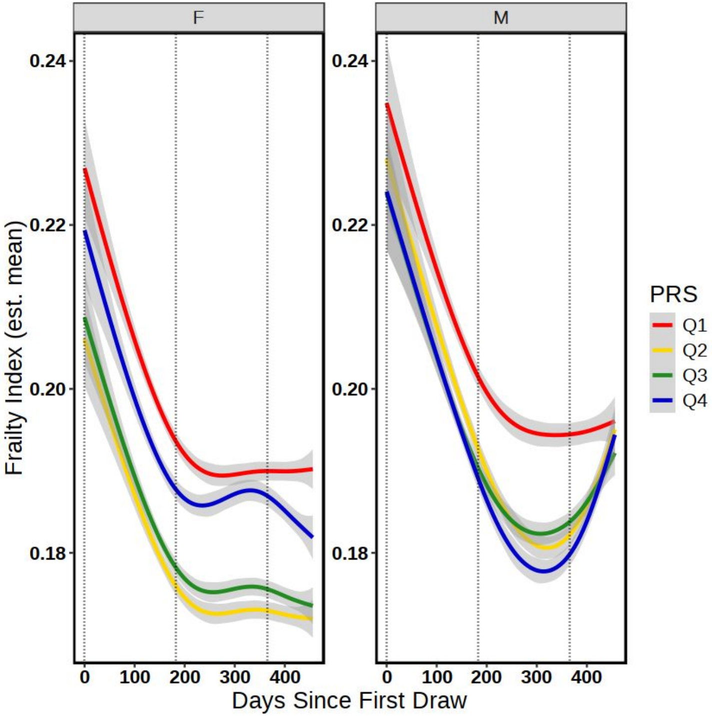
Polygenic Risk Score for Longevity (dl90_cs_PRS)



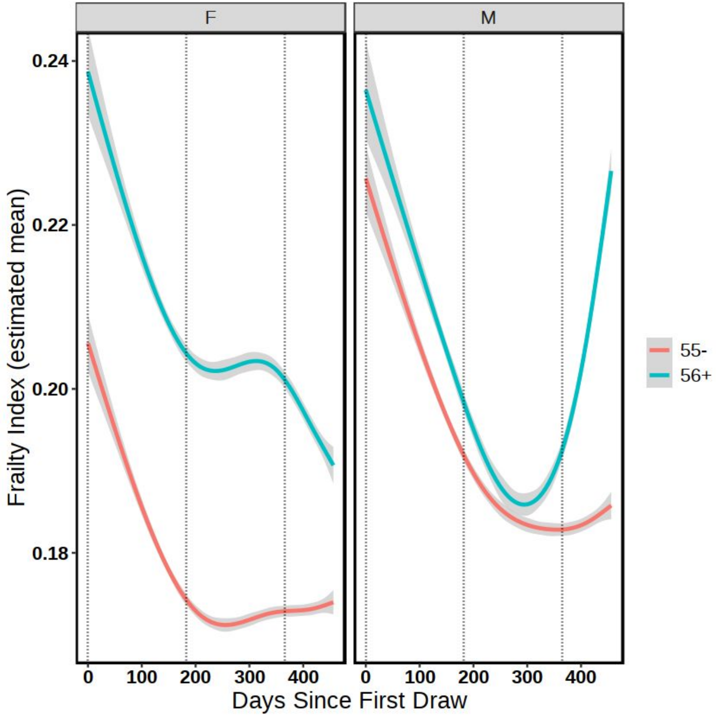


Frailty, Age, and Longevity

Frailty and Longevity PRS



Frailty by Sex and Age



F	M
472	343
202	109

BREAK