- Je suis face à un problème de régression avec $n = 10\,000$ individus et $p = 100$ variables hétérogènes (quanti et quali), et j'aimerais savoir quelles sont les variables explicatives pertinentes :

  - [ ] je réalise une régression linéaire avec selection de variables backward
  - [ ] je discrétise les variables quantitatives en variables catégorielles
  - [ ] j'utilise la méthode de régression des forêts aléatoires
  - [ ] je transforme les variable catégorielles en variables binaire
  - [ ] je réalise une régression PLS
  - [ ] je réalise une régression LASSO
  - [ ] je réalise une régression ridge
  - [ ] je réalise une régression avec selection de variables forward.
  - [ ] je commence par trier les variables en enlevant toutes celles non significativement corrélées avec la réponse à prédire.

- Je dispose d'une base de données d'imagerie médicale, pour laquelles je sais dans quelles images il y a une tumeur. Je veux construire un modèle de scoring qui m'indique la probabilité de présence d'une tumeur.
  Je peux utiliser

- [ ] une régression linéaire
- [ ] une régression logistique
- [ ] une forêt aléatoire
- [ ] un SVM
- [ ] un KPPV
- [ ] un réseau de neurones

- Je dispose d'une base de données genomique, avec 10 000 gènes, et pour chaque individu la connaissance de si oui ou non ils ont une certaine maladie d'intérêt. Pour détecter les gènes d'intérêt, je veux réaliser :
  - [ ] une régression linéaire
  - [ ] une régression logistique
  - [ ] une régression logistique avec pénalité LASSO
  - [ ] un SVM
  - [ ] un Random Forest
  - [ ] une Artificial Neural Network

- Exercice : sur les données breast tumors, pour laquelle vous avez un début de correction :
  - → tester une autre technique d'imputation
  - → évaluer par validation croisée
  - → tester une régression logistique