Casualty Actuarial Society

Machine Learning Part 2- Webinar

Thursday February 27, 2020

Ben Williams, Graham Wright

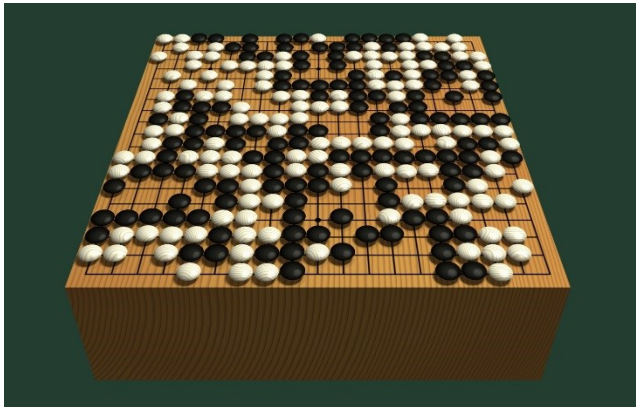
Willis Towers Watson

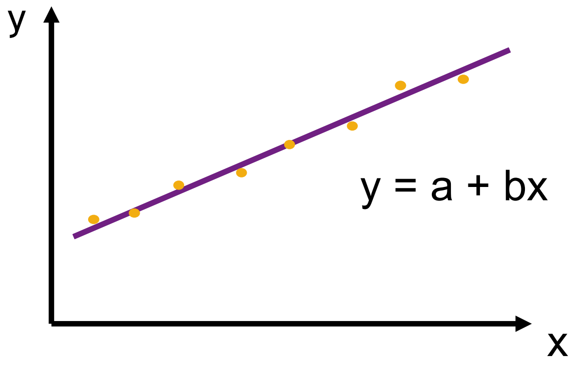
Agenda

Agenda
Context of machine learning in pricing
Session 1:
Decision trees **Random forests** **Gradient boosting machines**
Session 2:
"Earth" **Penalized regression** **Neural networks**
Conclusions
Q&A

Objective: to give you a working knowledge of some machine learning methods that may be used to improve GLM results and/or offer valuable insights in their own right in the field of P&C insurance pricing

Who’s interested in what?





Applications of machine learning in the insurance sector

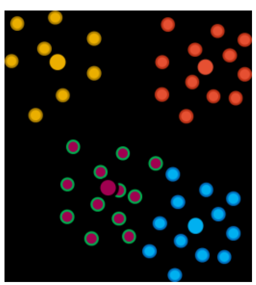
Marketing and Distribution → **Customer management** → **Underwriting and risk management** → **Pricing** → **Claims management** → **Asset management**

Product development (UBI)

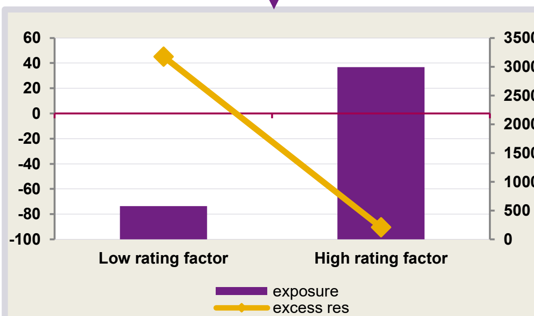
Retention segmentation and program design

Agent/Broker performance evaluation

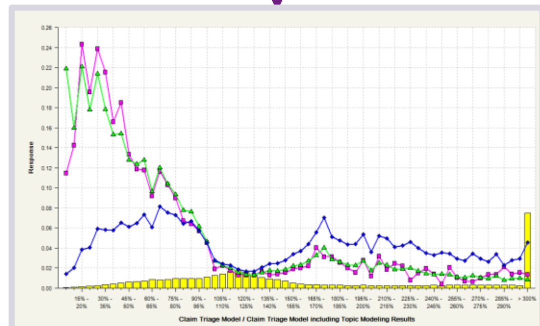
Fraud detection



P&C to Life cross-selling

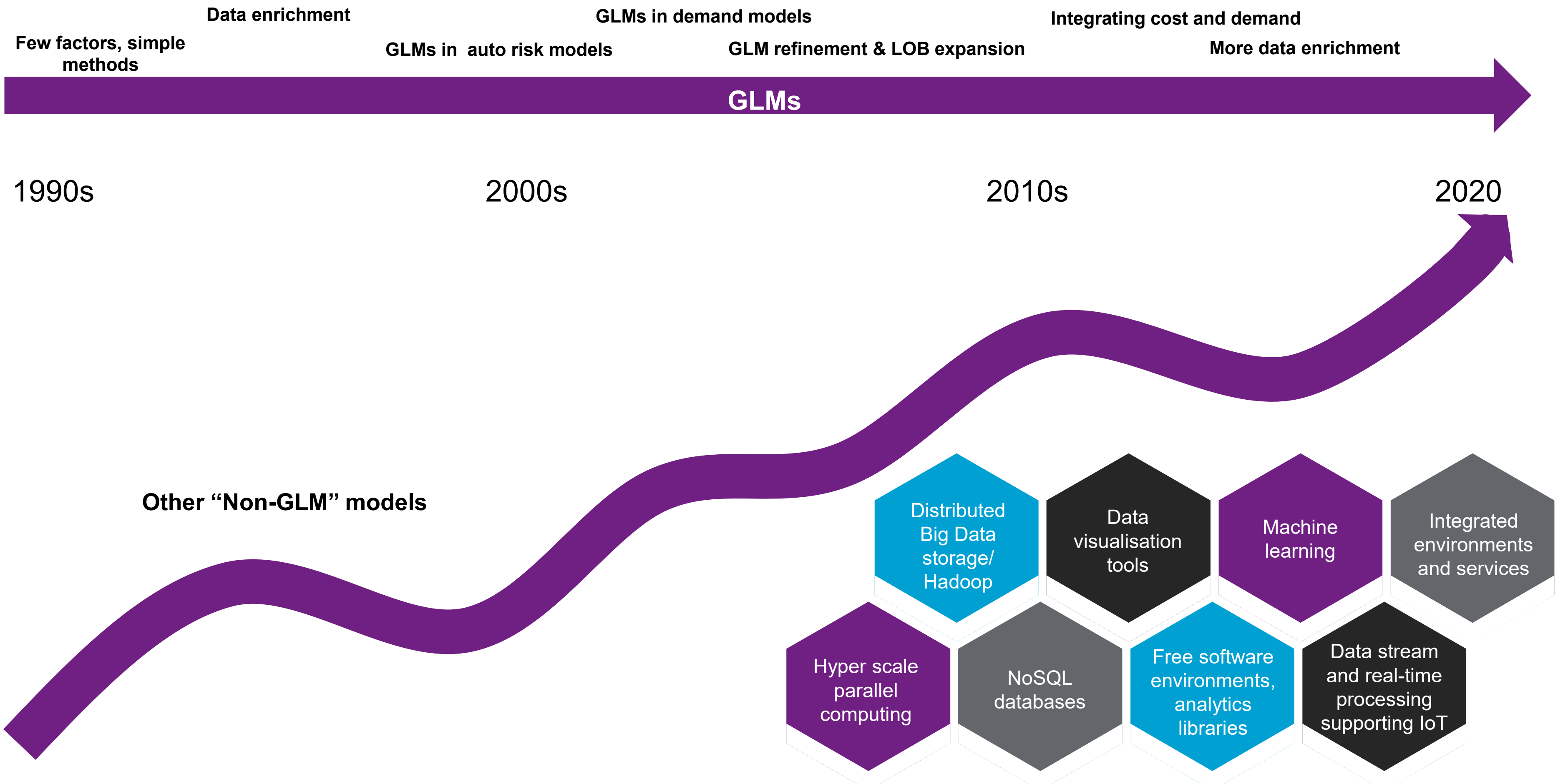


Web-scraping and feature design in commercial lines



Topic modeling and large loss modeling

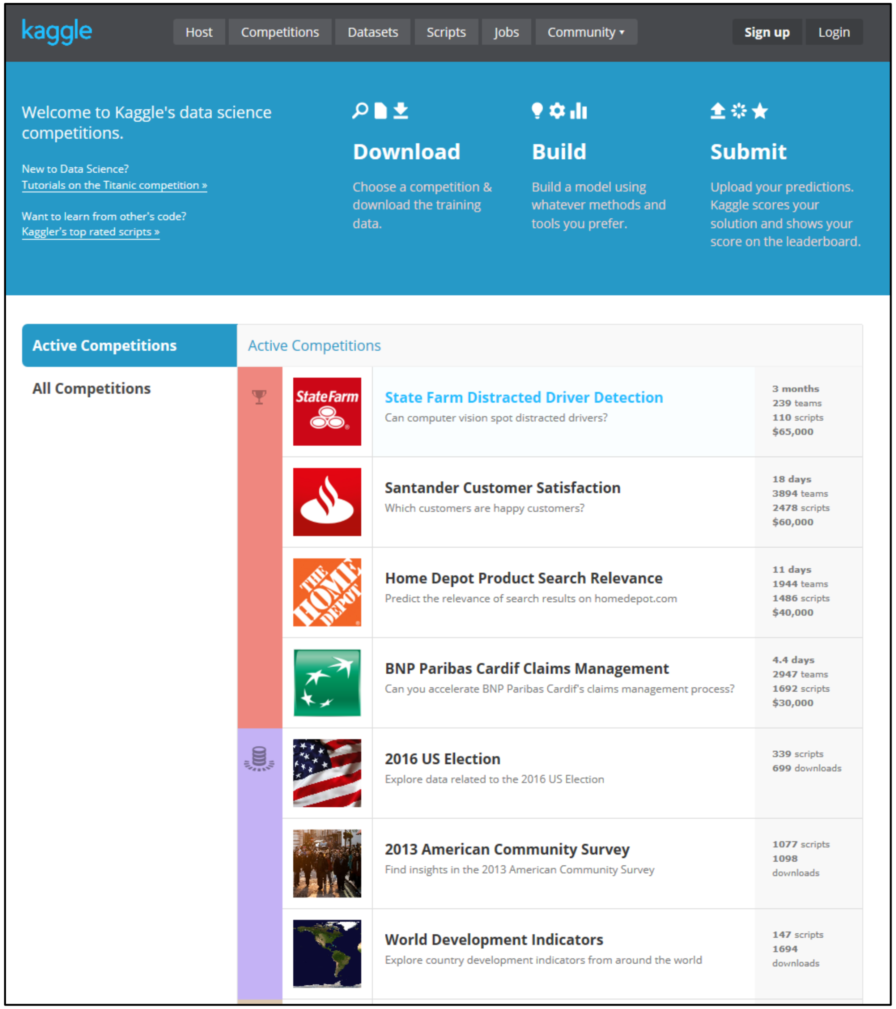
This is not new….

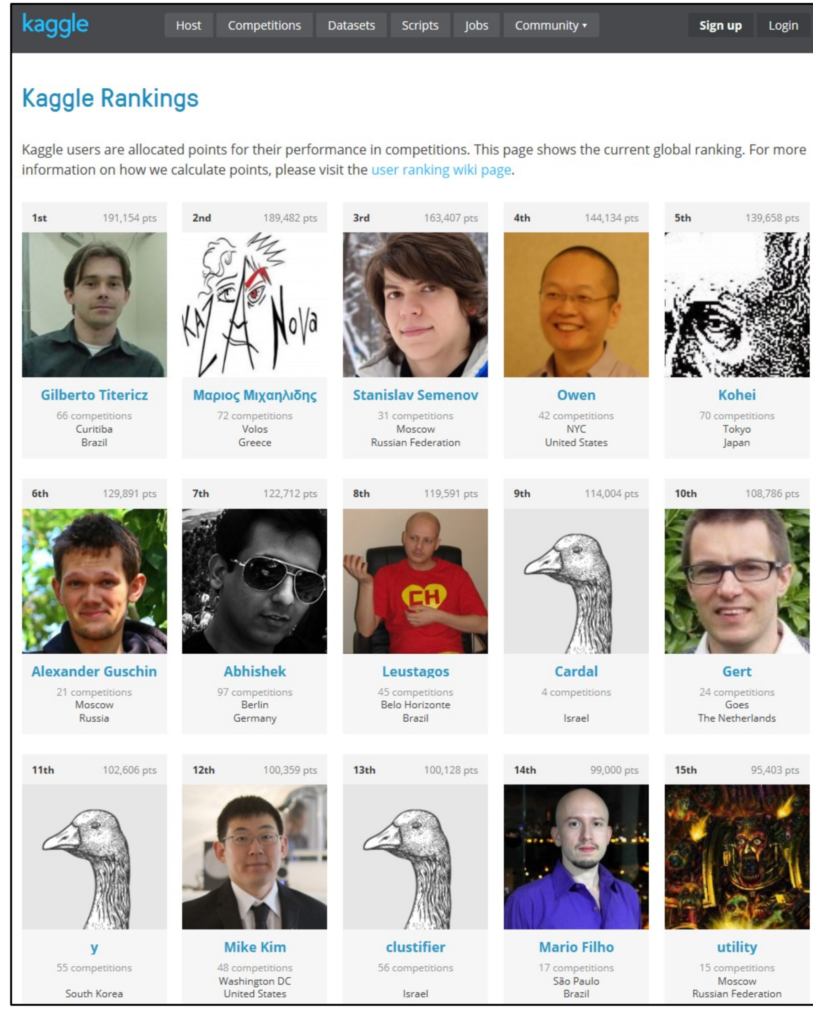


What are these machine learning methods?

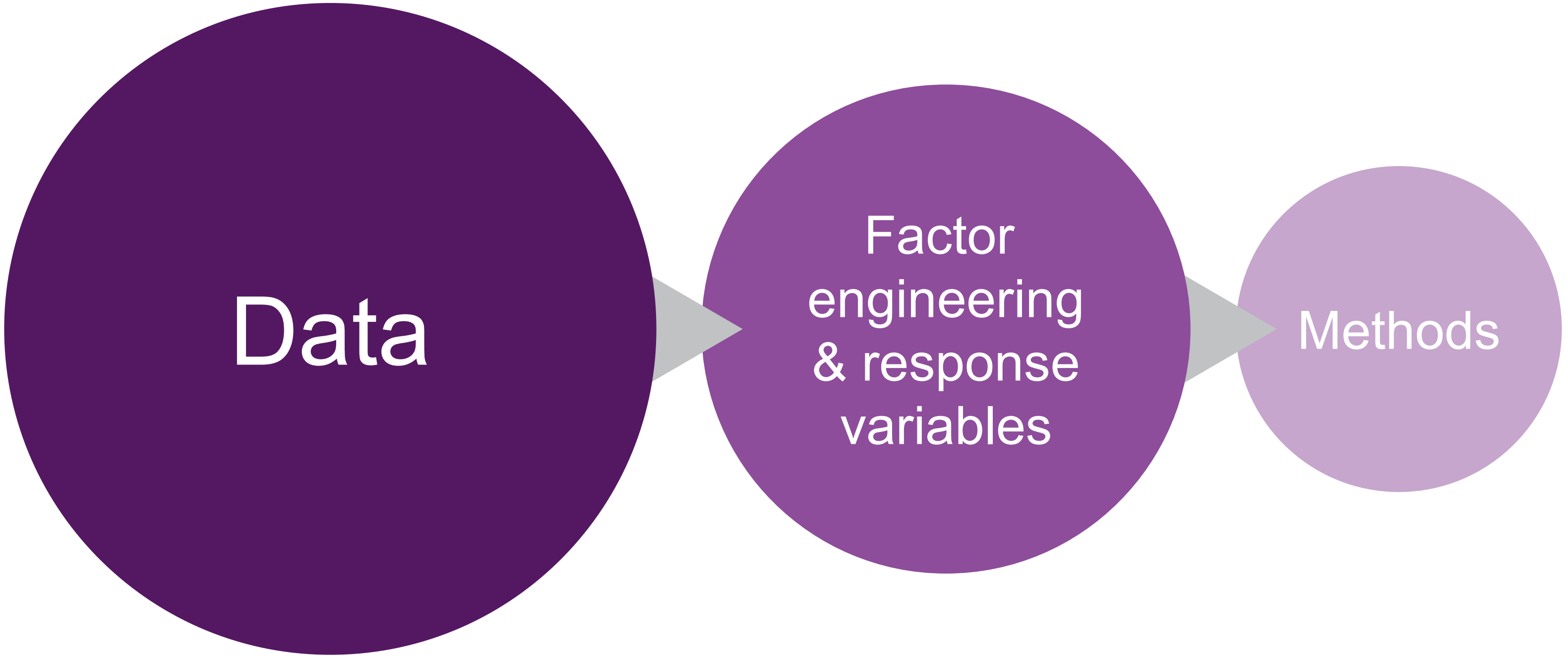
- Ensembles
- Classifications Trees
- "Earth"
- Regression Trees
- Gradient Boosting Machines
- K-nearest Neighbors
- Elastic Net
- Neural Networks
- Naïve Bayes
- Random Forests
- K-Means Clustering
- Principal Components Analysis
- Lasso
- Support Vector Machines
- Ridge Regression

Kaggle

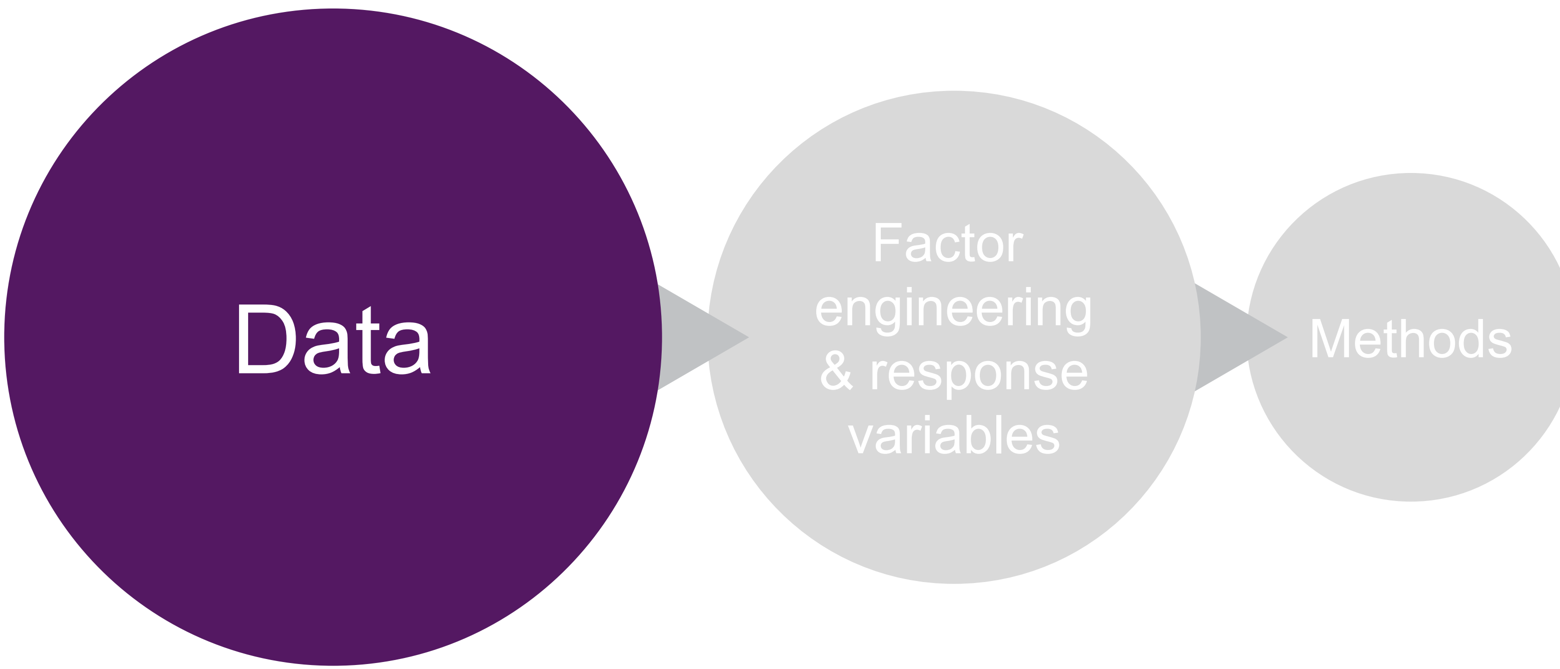

kaggle
Host
Competitions
Datasets
Scripts
Jobs
Community
Sign up
Login
Welcome to Kaggle's data science competitions.
New to Data Science?
Tutorials on the Titanic competition »
Want to learn from other's code?
Kaggler's top rated scripts »
Download
Choose a competition & download the training data.
Build
Build a model using whatever methods and tools you prefer.
Submit
Upload your predictions. Kaggle scores your solution and shows your score on the leaderboard.
Active Competitions
All Competitions
Active Competitions
State Farm Distracted Driver Detection
Can computer vision spot distracted drivers?
3 months
239 teams
110 scripts
$65,000
Santander Customer Satisfaction
Which customers are happy customers?
18 days
3894 teams
2478 scripts
$60,000
Home Depot Product Search Relevance
Predict the relevance of search results on homedepot.com
11 days
1944 teams
1486 scripts
$40,000
BNP Paribas Cardif Claims Management
Can you accelerate BNP Paribas Cardif's claims management process?
4.4 days
2947 teams
1692 scripts
$30,000
2016 US Election
Explore data related to the 2016 US Election
339 scripts
699 downloads
2013 American Community Survey
Find insights in the 2013 American Community Survey
1077 scripts
1098 downloads
World Development Indicators
Explore country development indicators from around the world
147 scripts
1694 downloads


kaggle
Host
Competitions
Datasets
Scripts
Jobs
Community
Sign up
Login
Kaggle Rankings
Kaggle users are allocated points for their performance in competitions. This page shows the current global ranking. For more information on how we calculate points, please visit the user ranking wiki page.
1st 191,154 pts
Gilberto Titericz
66 competitions
Curitiba
Brazil
2nd 189,482 pts
Μαριος Μιχαηλιδης
72 competitions
Volos
Greece
3rd 163,407 pts
Stanislav Semenov
31 competitions
Moscow
Russian Federation
4th 144,134 pts
Owen
42 competitions
NYC
United States
5th 139,658 pts
Kohei
70 competitions
Tokyo
Japan
6th 129,891 pts
Alexander Guschin
21 competitions
Moscow
Russia
7th 122,712 pts
Abhishek
97 competitions
Berlin
Germany
8th 119,591 pts
Leustagos
45 competitions
Belo Horizonte
Brazil
9th 114,004 pts
Cardal
4 competitions
Israel
10th 108,786 pts
Gert
24 competitions
Goes
The Netherlands
11th 102,606 pts
y
55 competitions
South Korea
12th 100,359 pts
Mike Kim
48 competitions
Washington DC
United States
13th 100,128 pts
clustifier
56 competitions
Israel
14th 99,000 pts
Mario Filho
17 competitions
São Paulo
Brazil
15th 95,403 pts
utility
15 competitions
Moscow
Russian Federation

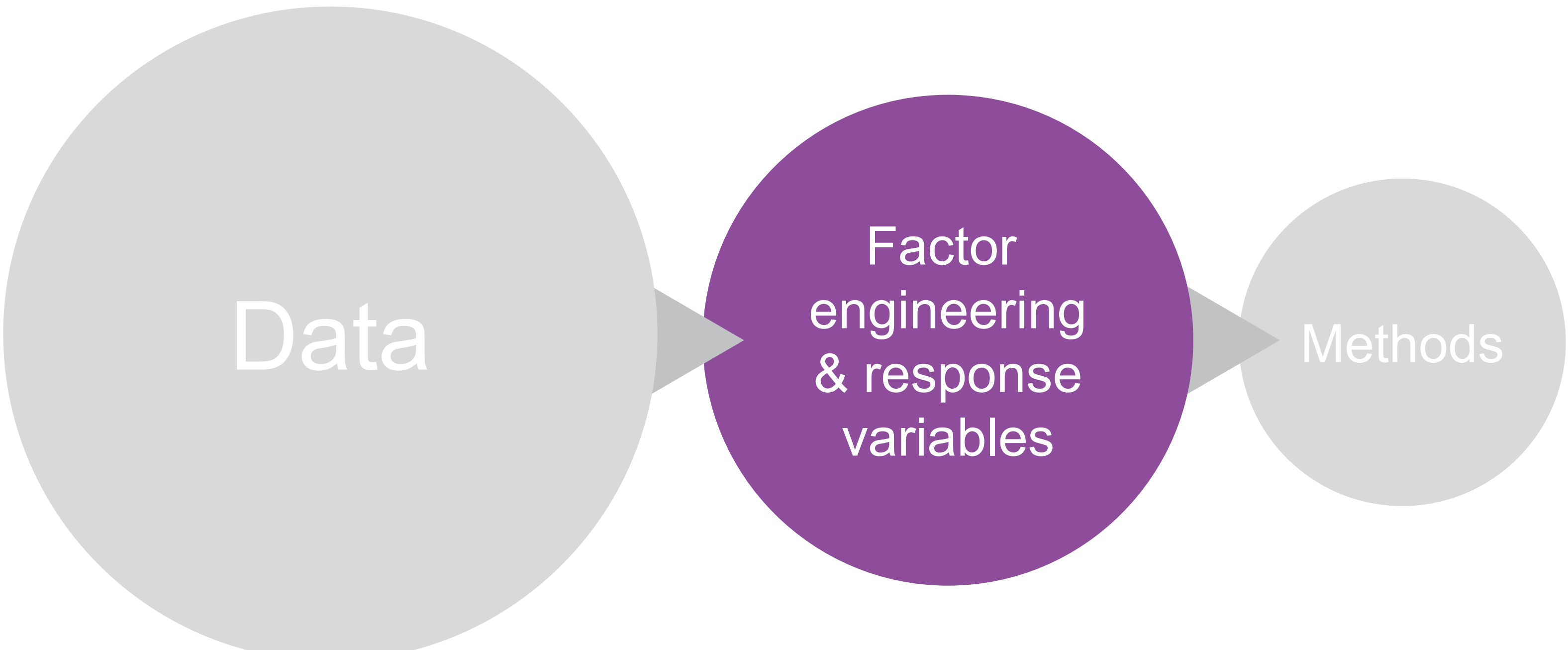
Is it really all about the method?



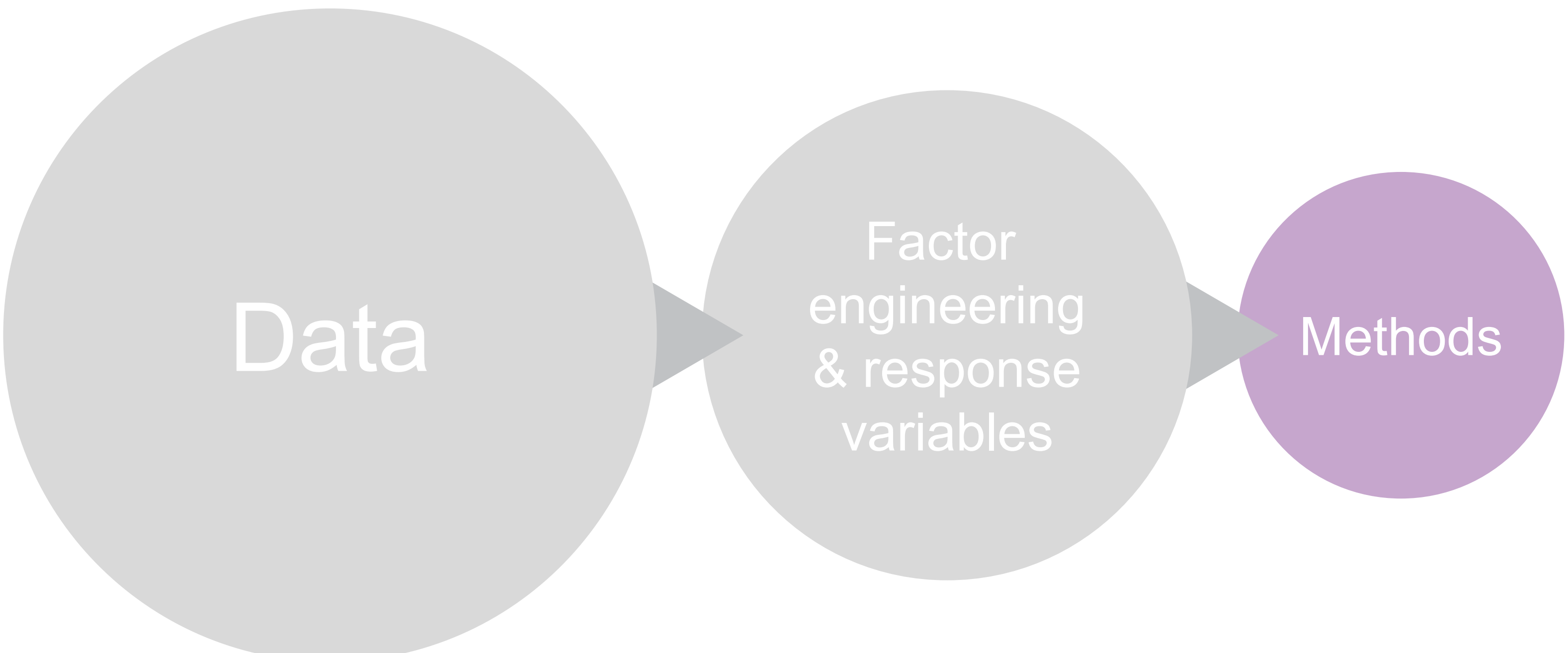
Is it really all about the method?



Is it really all about the method?



Is it really all about the method?

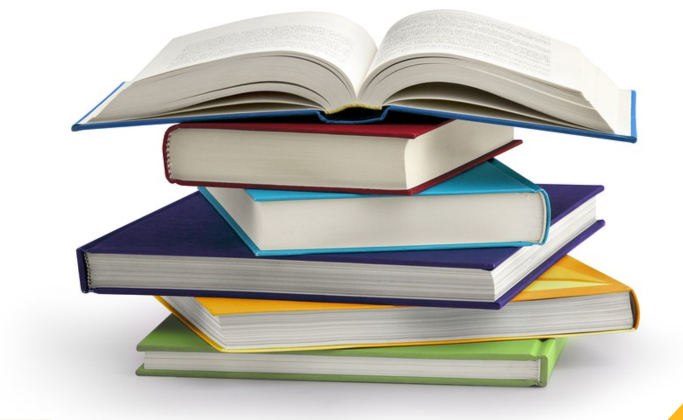


How do you know if a method works?

AIC

Gini

RMSE

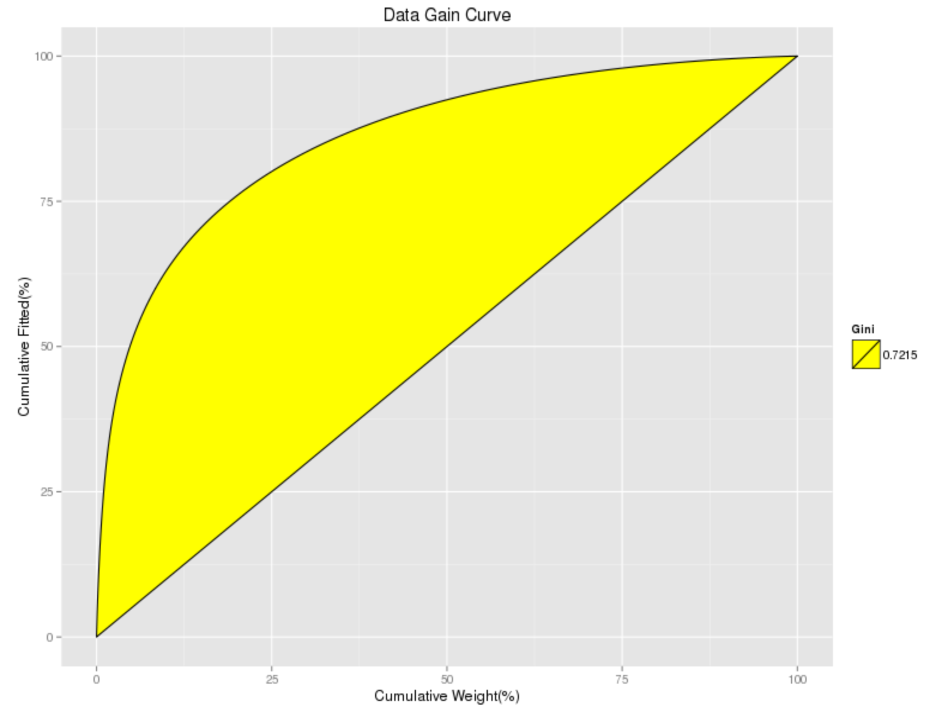


MAE

Log loss

How do you measure value?

Gini



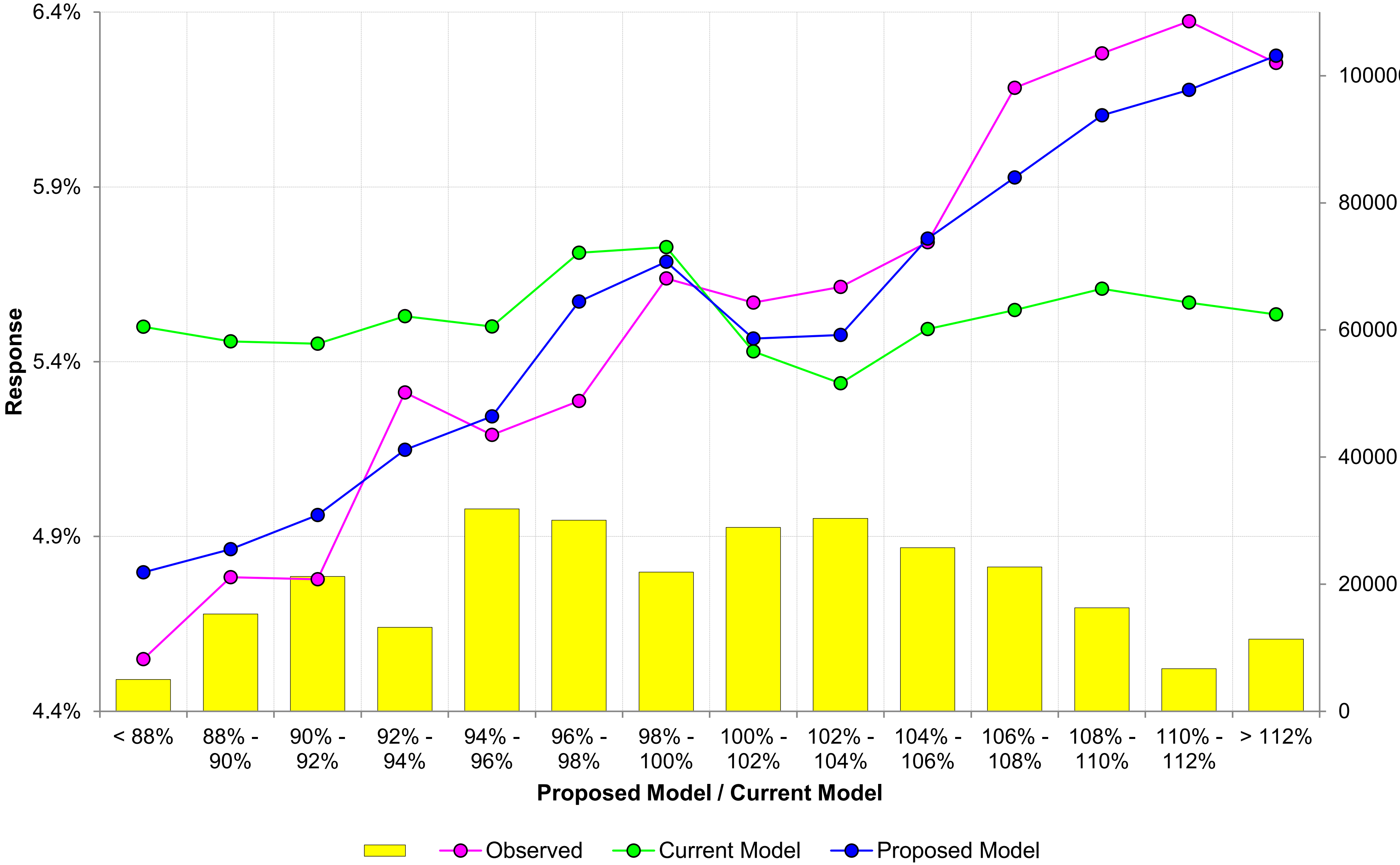
- Rank hold out observations by their **fitted values** (high to low)
- **Plot cumulative response** by cumulative exposure
- A **better model** will explain a **higher proportion of the response** with a **lower proportion of exposure**
- …and will give a **higher Gini coefficient** (yellow area)

But…

- Think of a model…
- Multiply it by 123
- Square it
- Add 74½ billion

- …and you get the same Gini coefficient!

Double lift chart

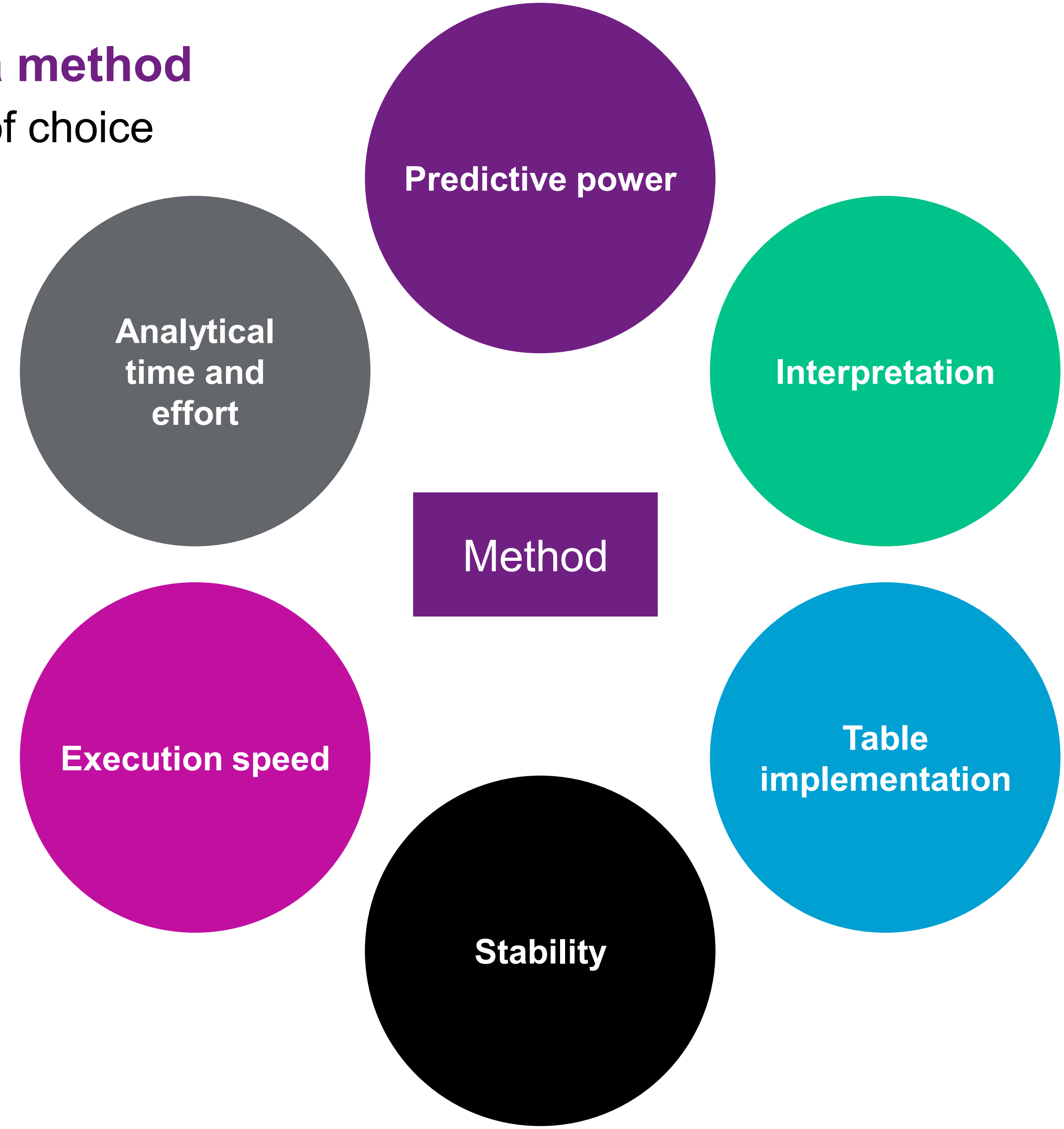


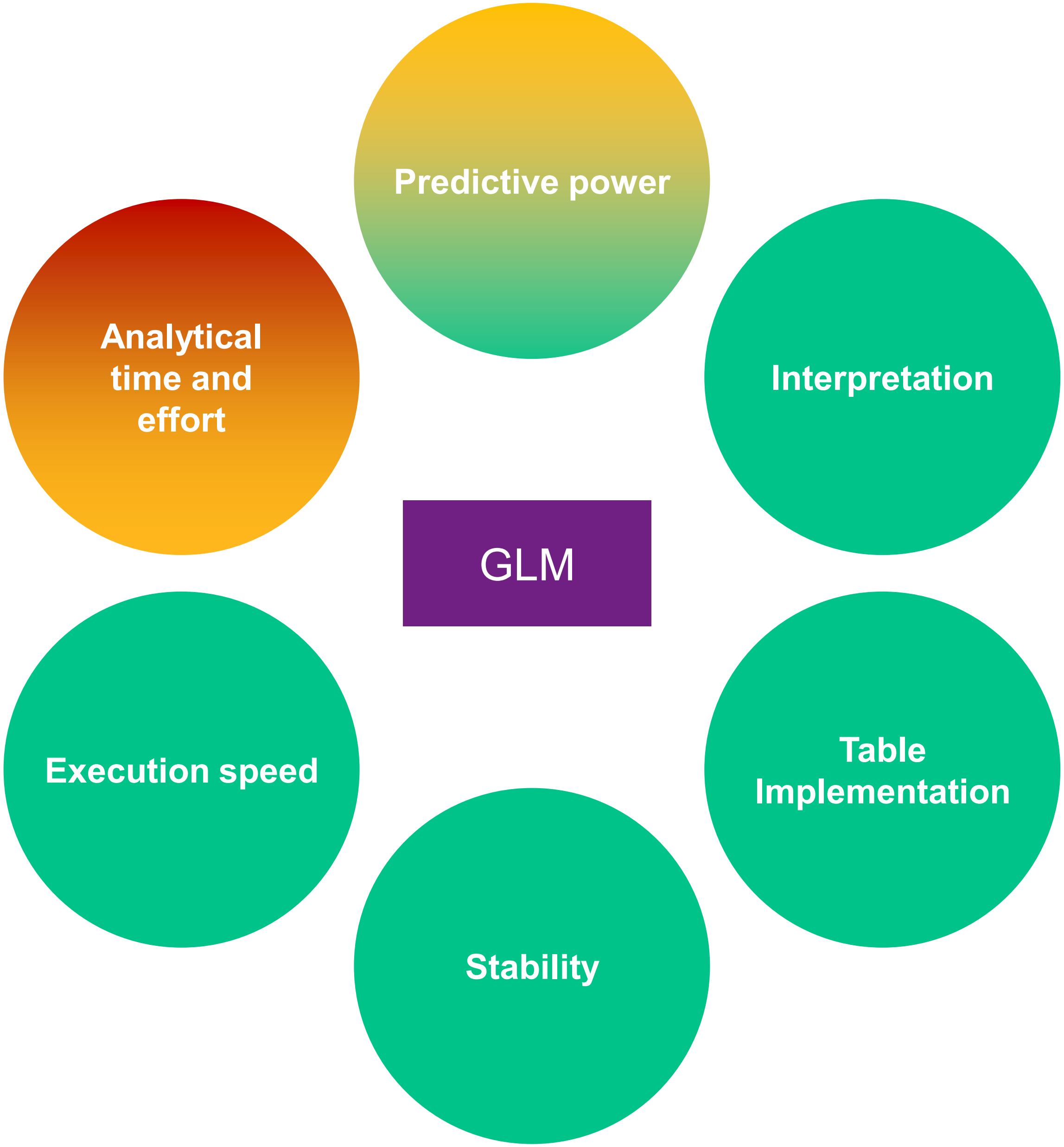
Is there more to it…?

Predictive power

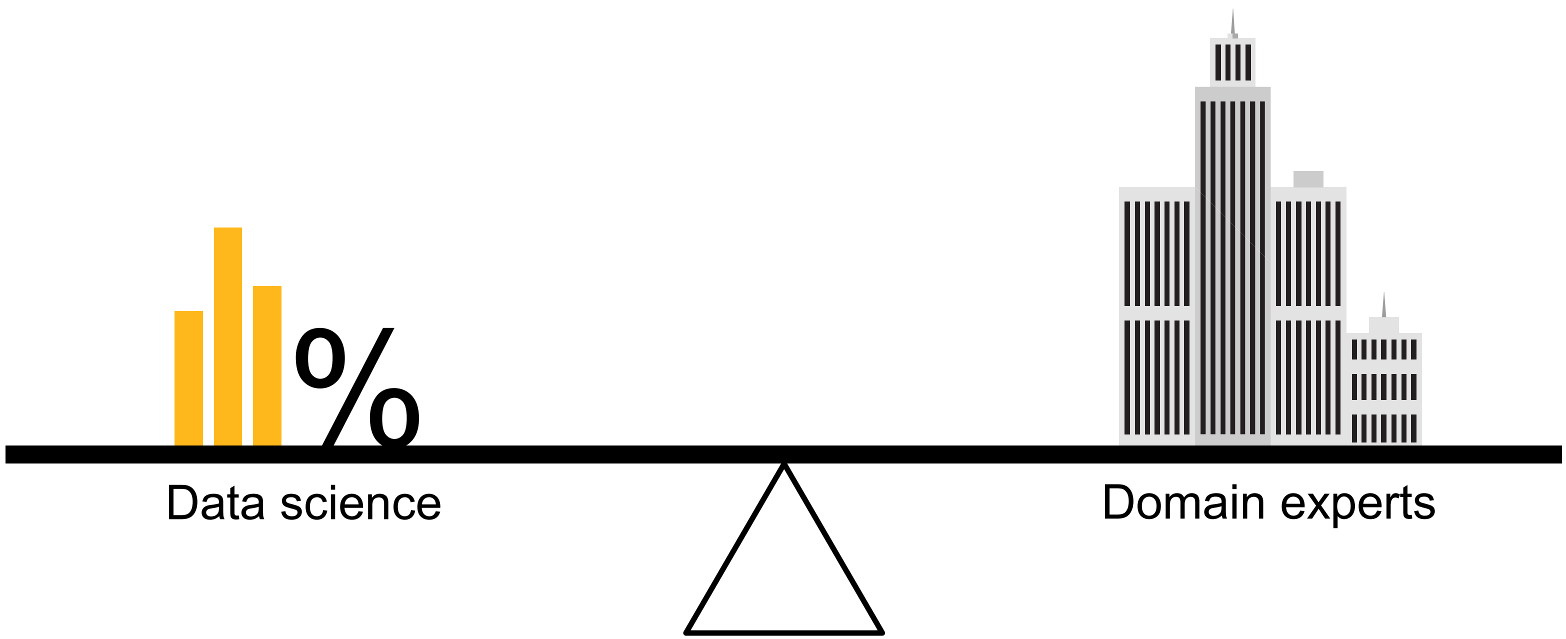
Choosing a method

Dimensions of choice

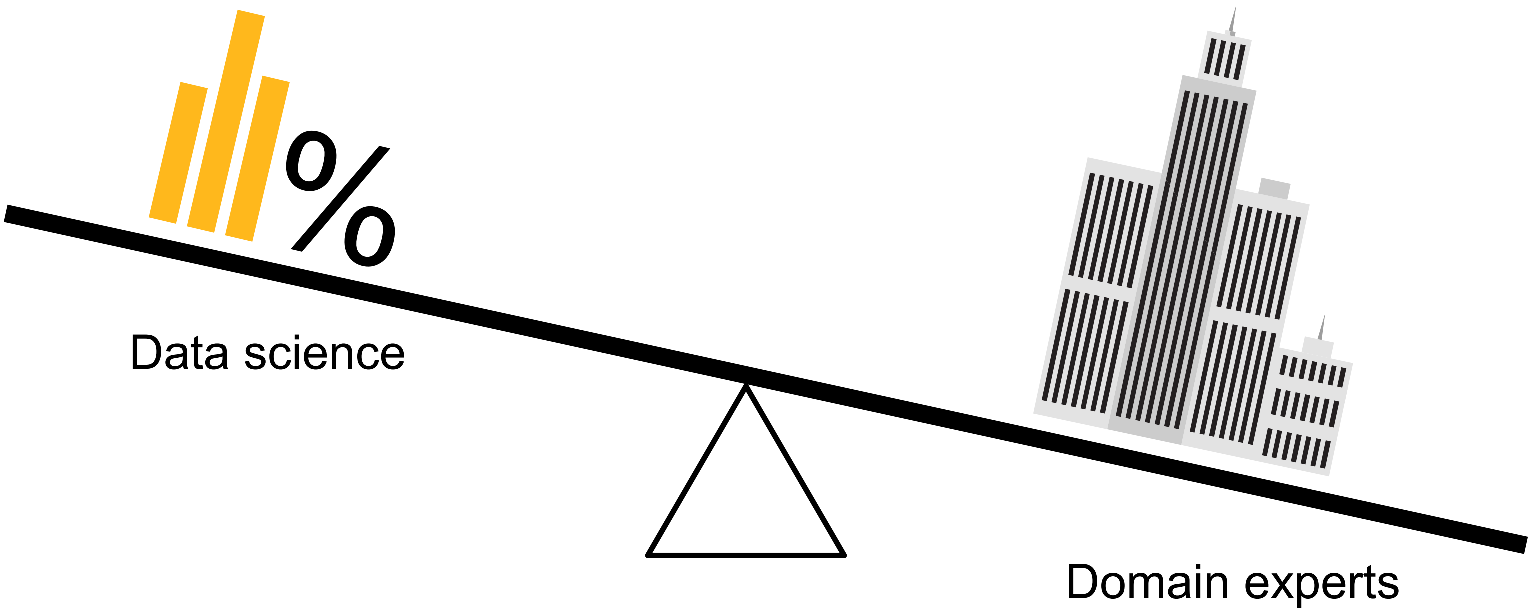



Predictive power
Analytical time and effort
Interpretation
GLM
Execution speed
Table Implementation
Stability

What do you use where?



It’s domain expertise that helps decide



What are these machine learning methods?

Ensembles

Classifications Trees

"Earth"

Regression Trees

Gradient Boosting Machines

K-nearest Neighbors

Elastic Net

Neural Networks

Naïve Bayes

Random Forests

K-Means Clustering

Principal Components Analysis

Lasso

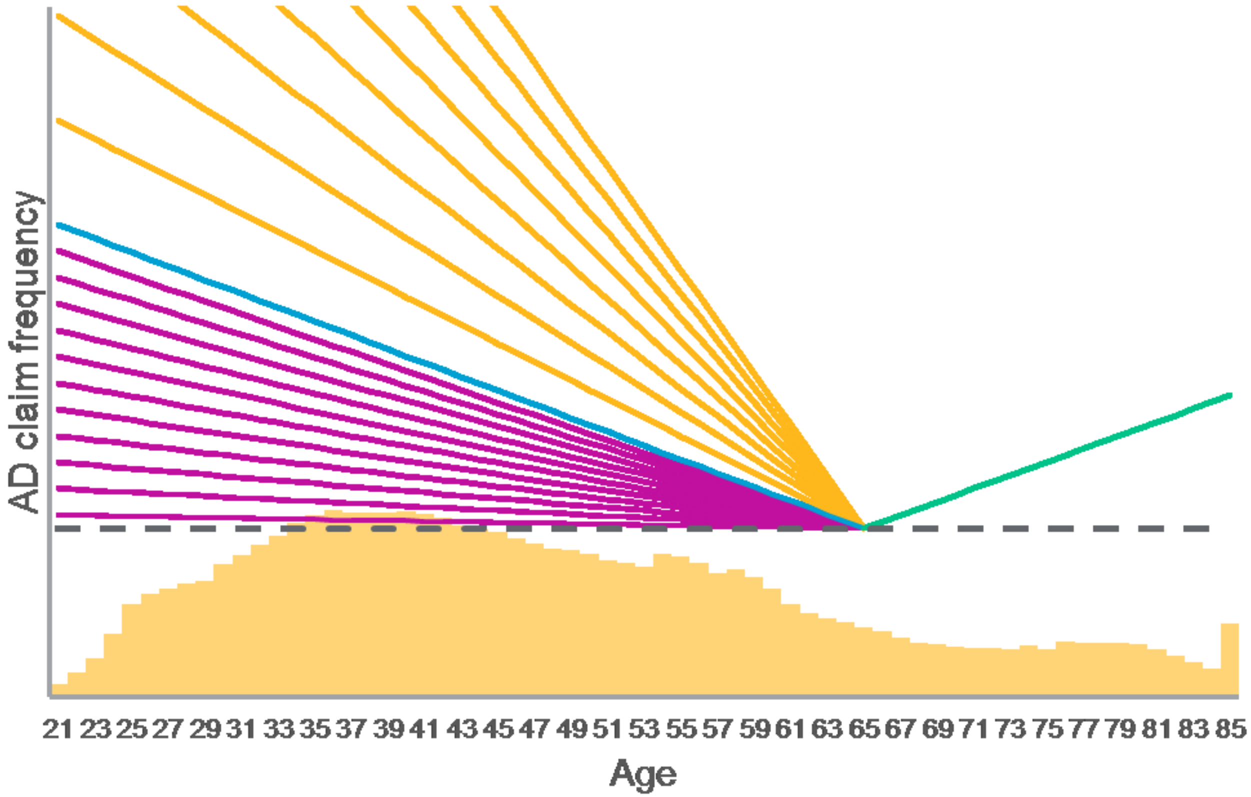
Support Vector Machines

Ridge Regression

Focus on “Earth”

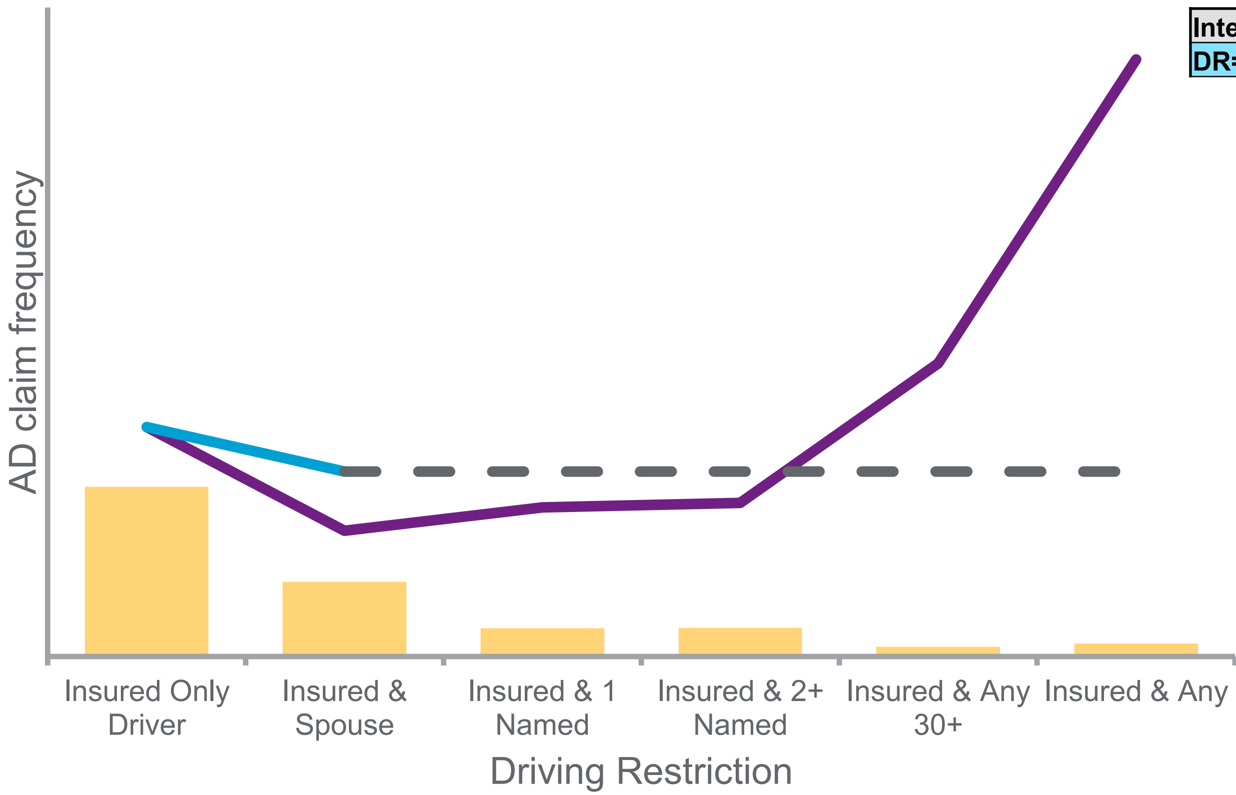
"Earth"

Multivariate adaptive regression splines ("Earth")



Multivariate adaptive regression splines ("Earth")

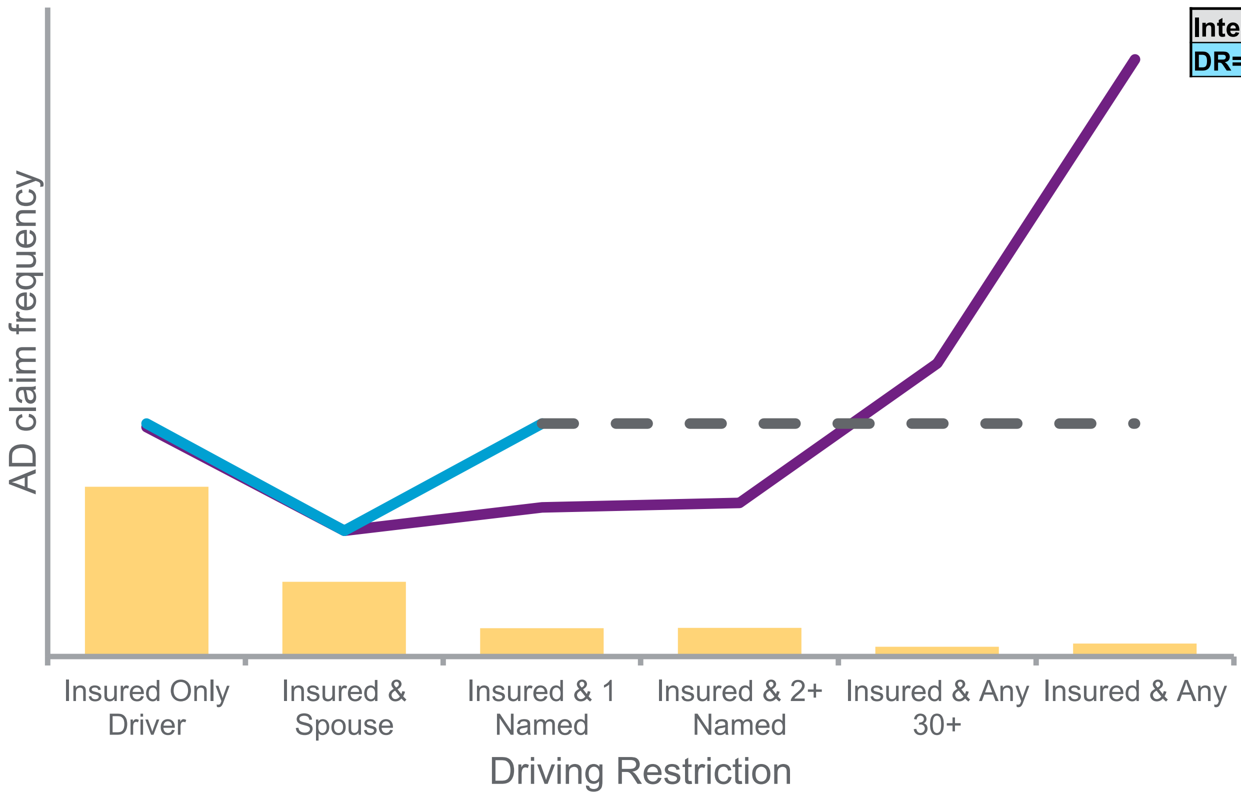
Categorical factors



Intercept	-2.958
DR=IOD	0.082

Multivariate adaptive regression splines ("Earth")

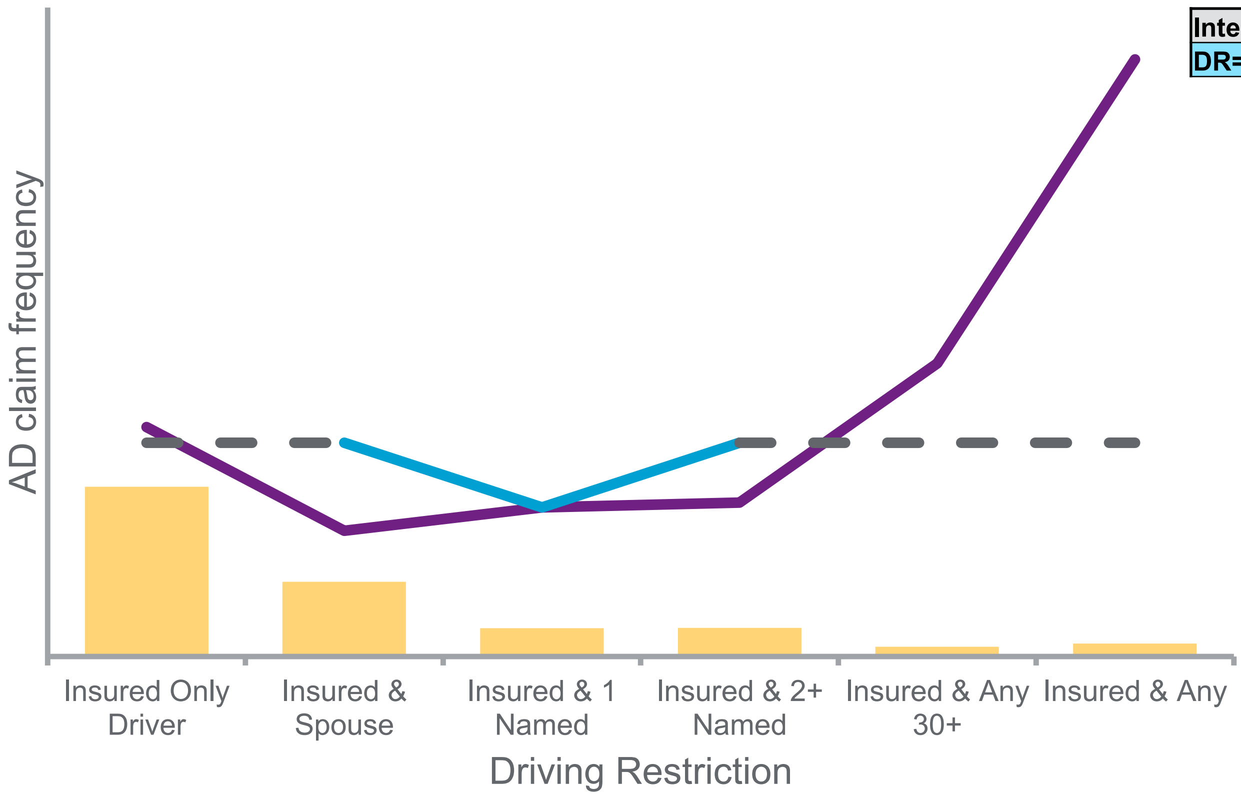
Categorical factors



Intercept	-2.869
DR=I&S	-0.198

Multivariate adaptive regression splines ("Earth")

Categorical factors

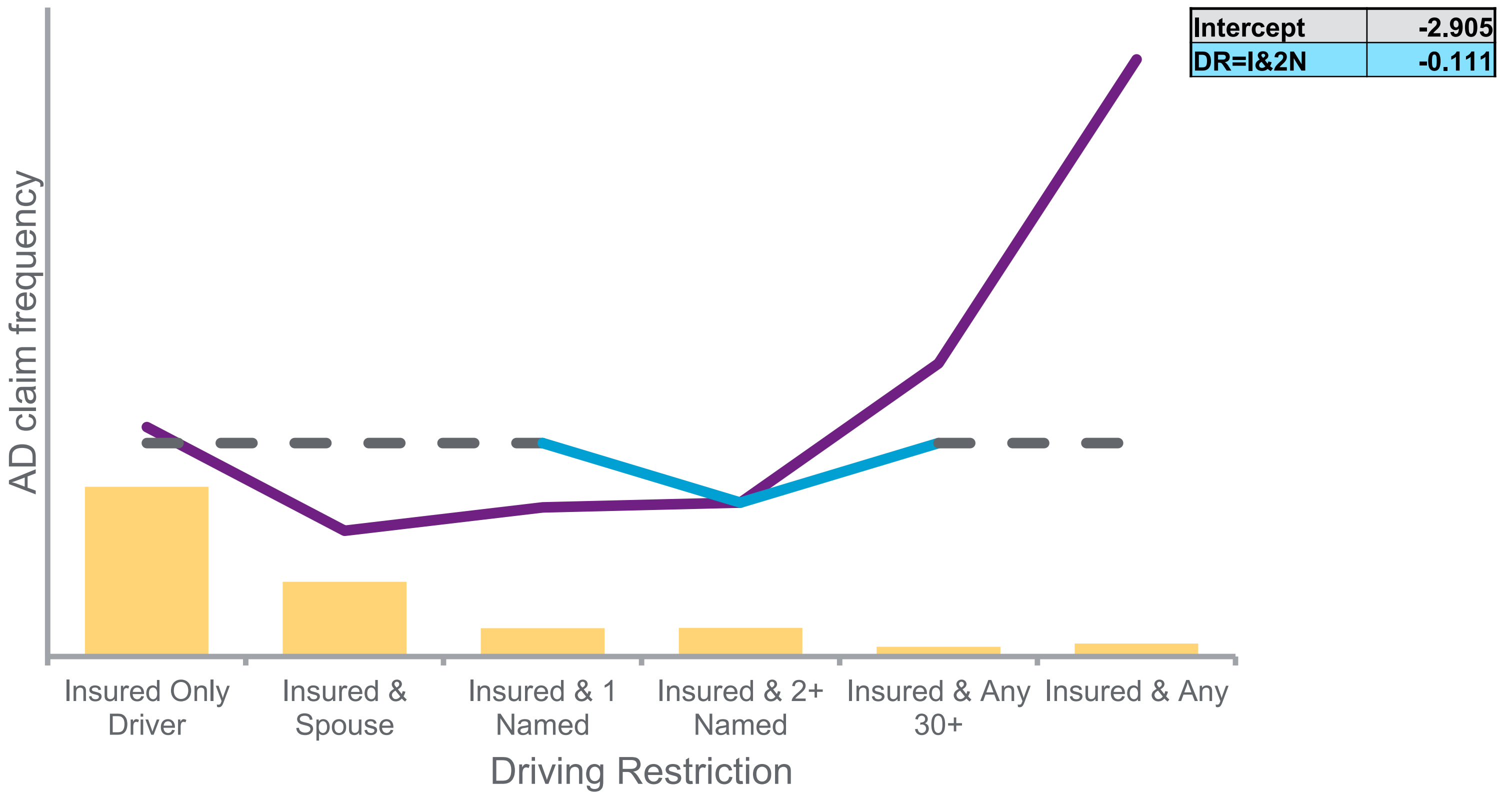


Intercept	-2.905
DR=I&1N	-0.120

Willis Towers Watson

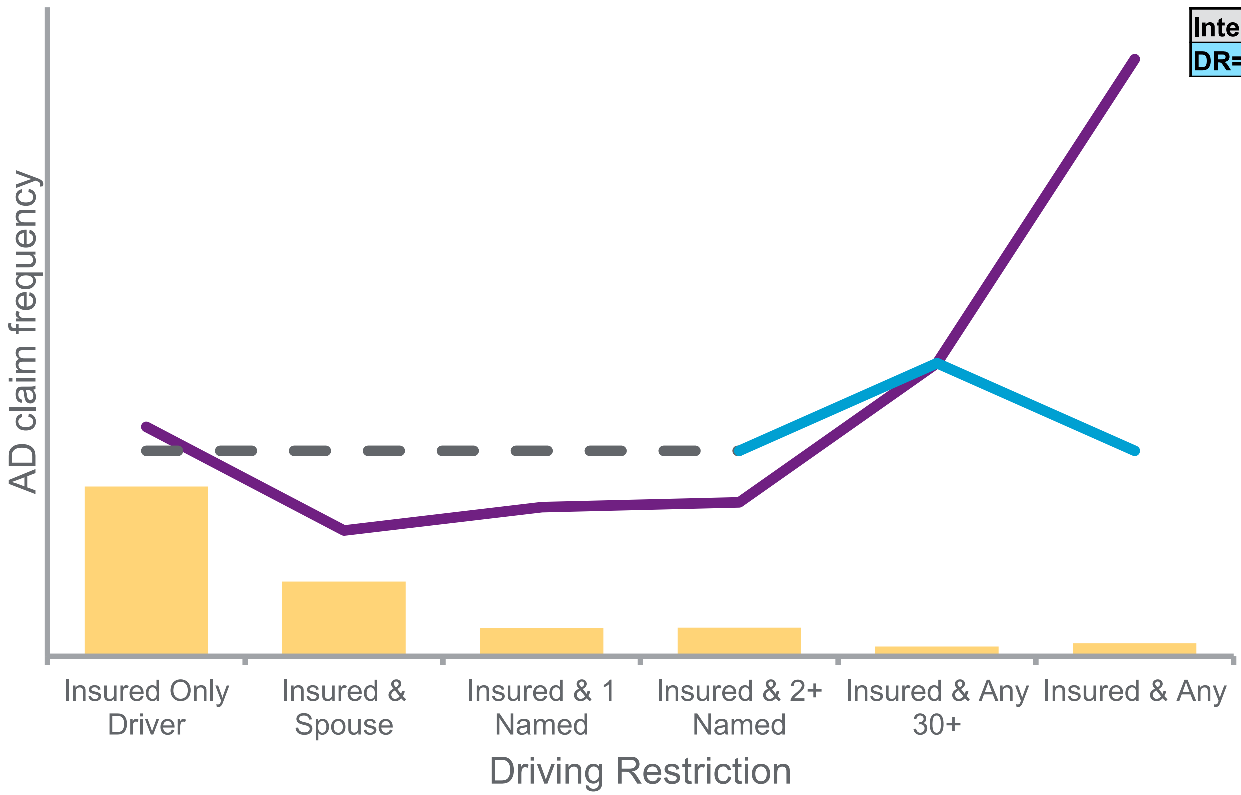
Multivariate adaptive regression splines ("Earth")

Categorical factors



Multivariate adaptive regression splines ("Earth")

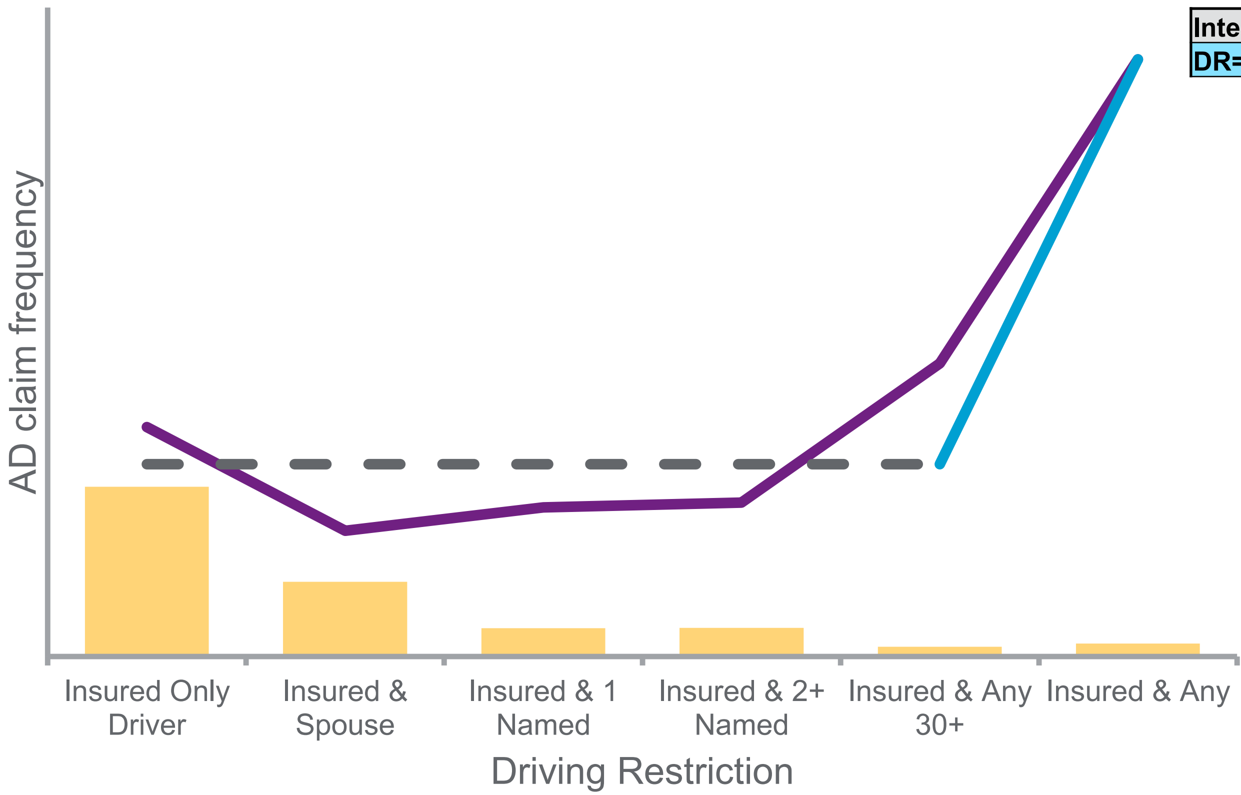
Categorical factors



Intercept	-2.920
DR=I&Any30+	0.162

Multivariate adaptive regression splines ("Earth")

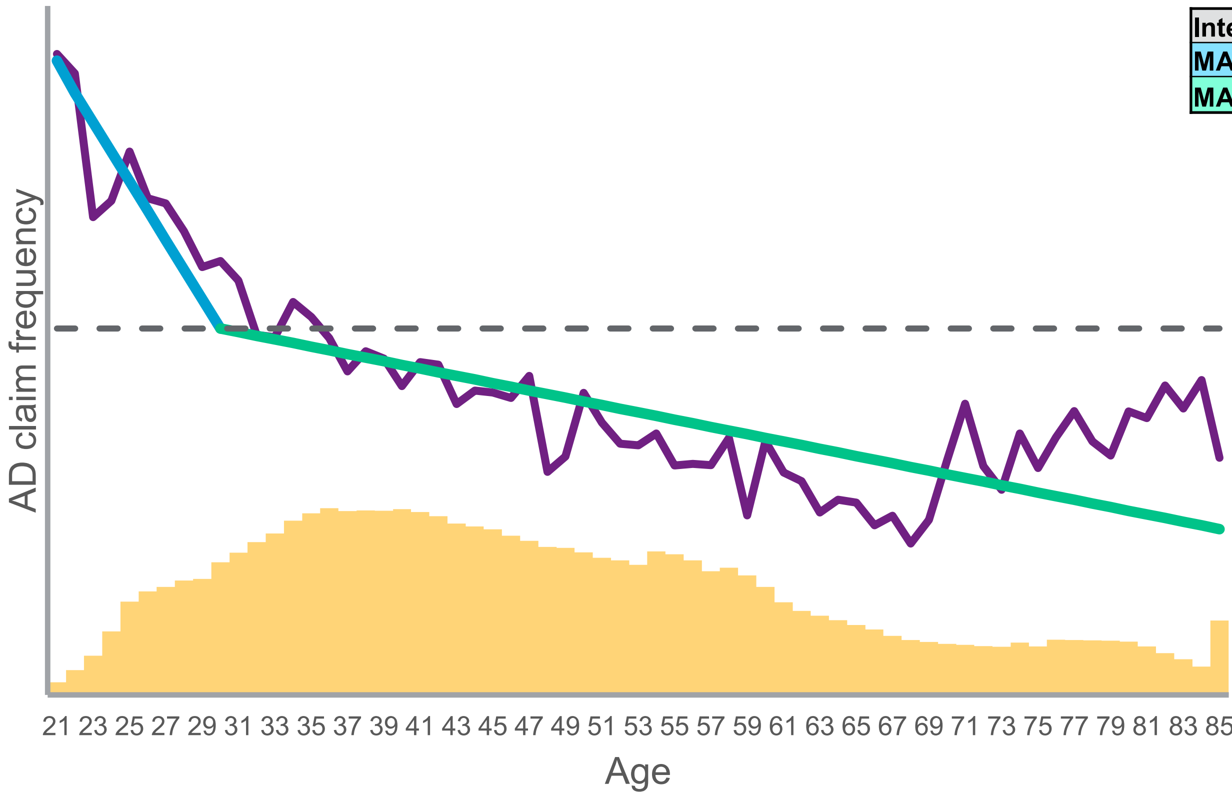
Categorical factors



Intercept	-2.945
DR=I&Any	0.749

Multivariate adaptive regression splines (“Earth”)

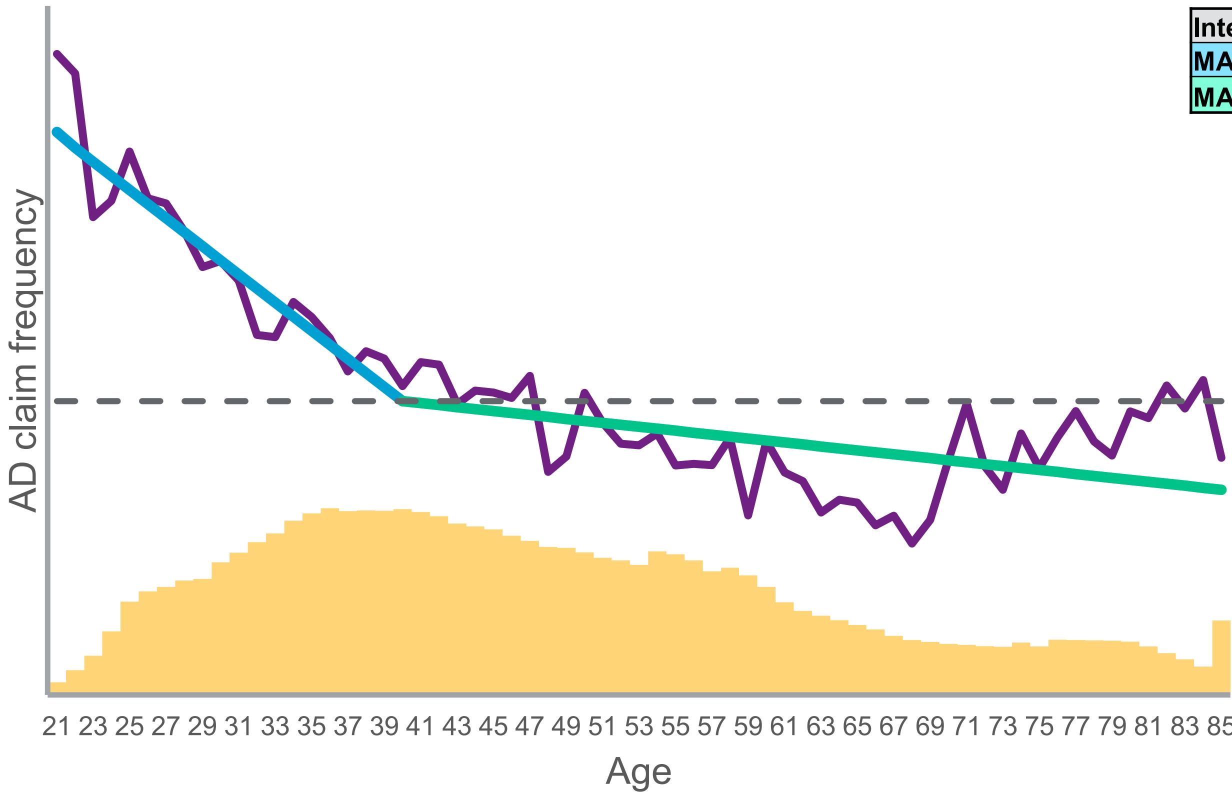
Numerical factors



Intercept	**-2.815**
MAX(30-Age,0)	**0.051**
MAX(Age-30,0)	**-0.006**

Multivariate adaptive regression splines (“Earth”)

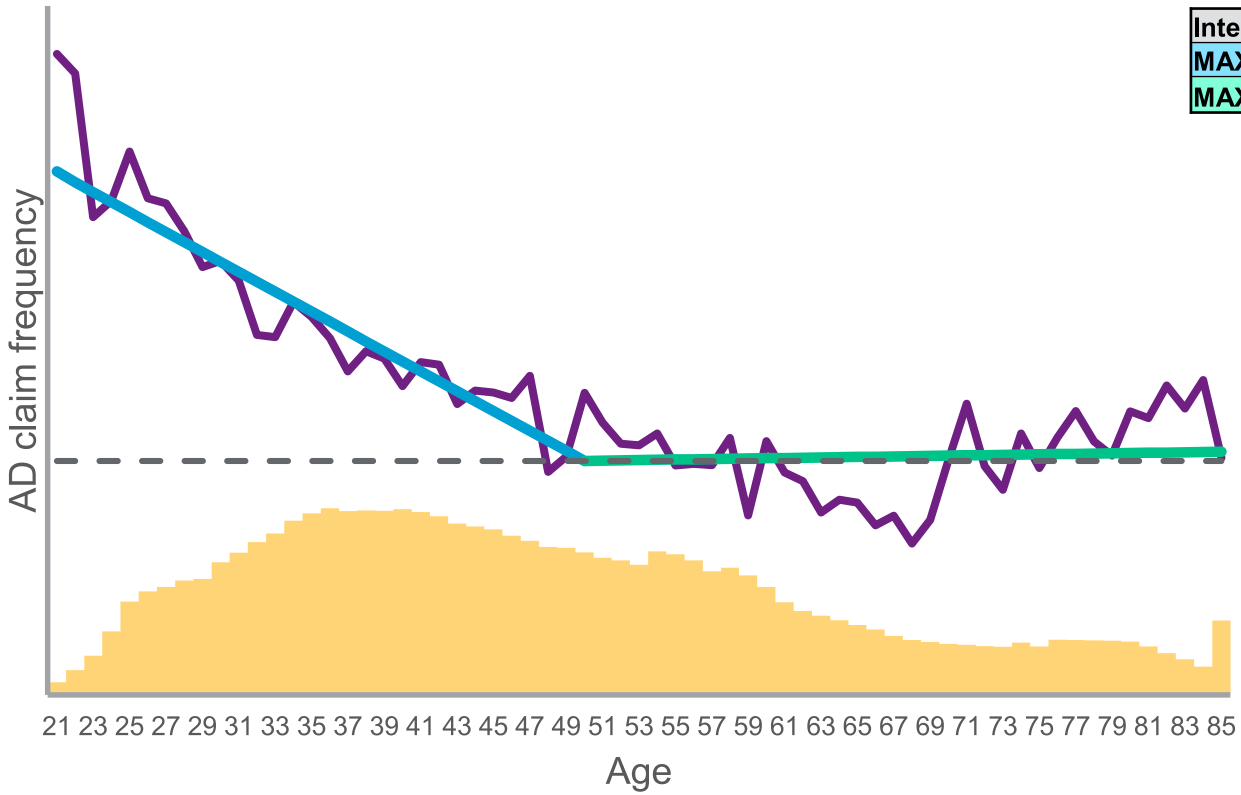
Numerical factors



Intercept	-2.931
MAX(40-Age,0)	0.025
MAX(Age-40,0)	-0.003

Multivariate adaptive regression splines (“Earth”)

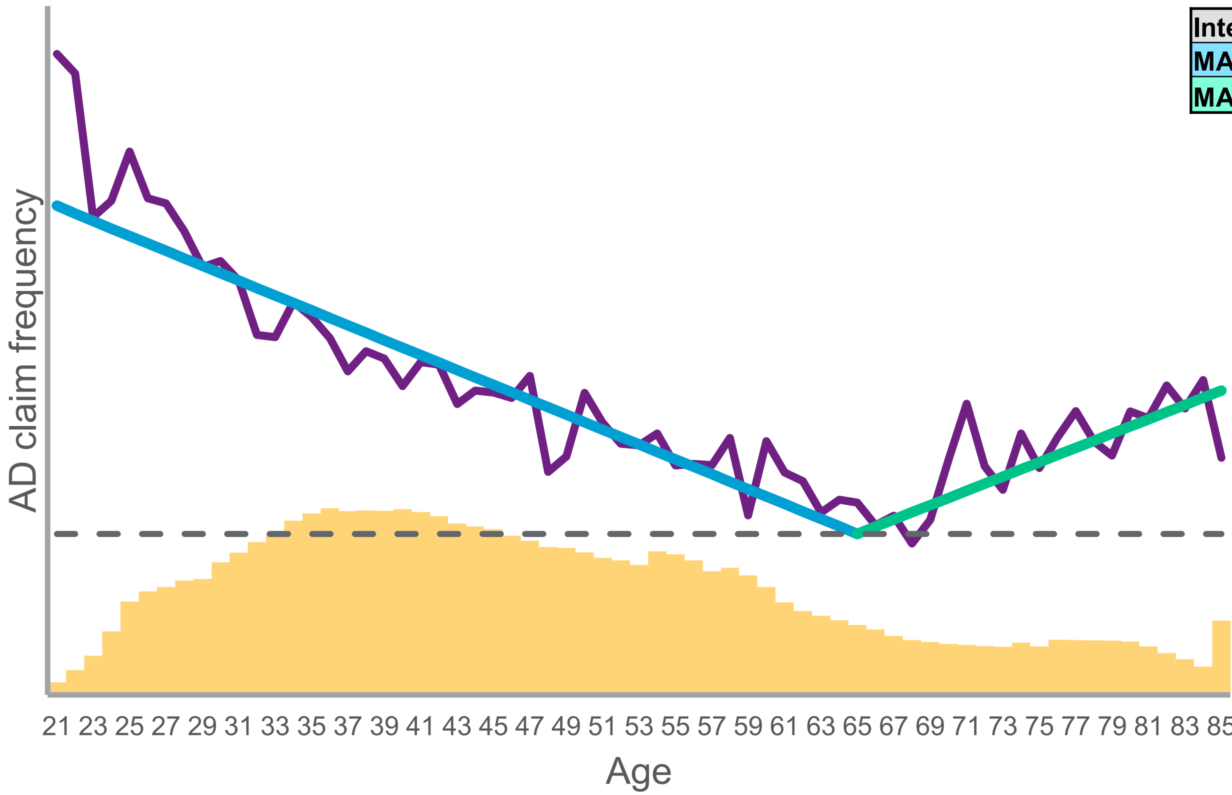
Numerical factors



Intercept	-3.026
MAX(50-Age,0)	0.017
MAX(Age-50,0)	0.000

Multivariate adaptive regression splines (“Earth”)

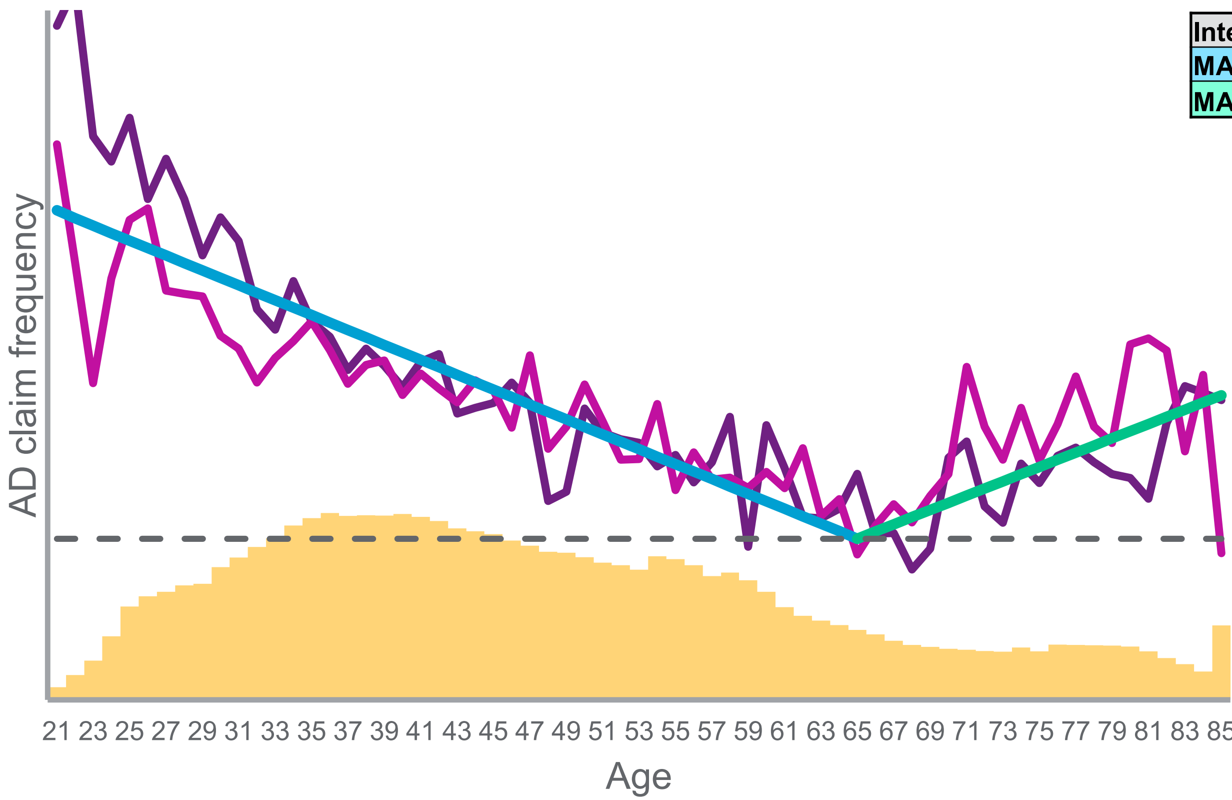
Numerical factors



Intercept	-3.143
MAX(65-Age,0)	0.013
MAX(Age-65,0)	0.011

Multivariate adaptive regression splines (“Earth”)

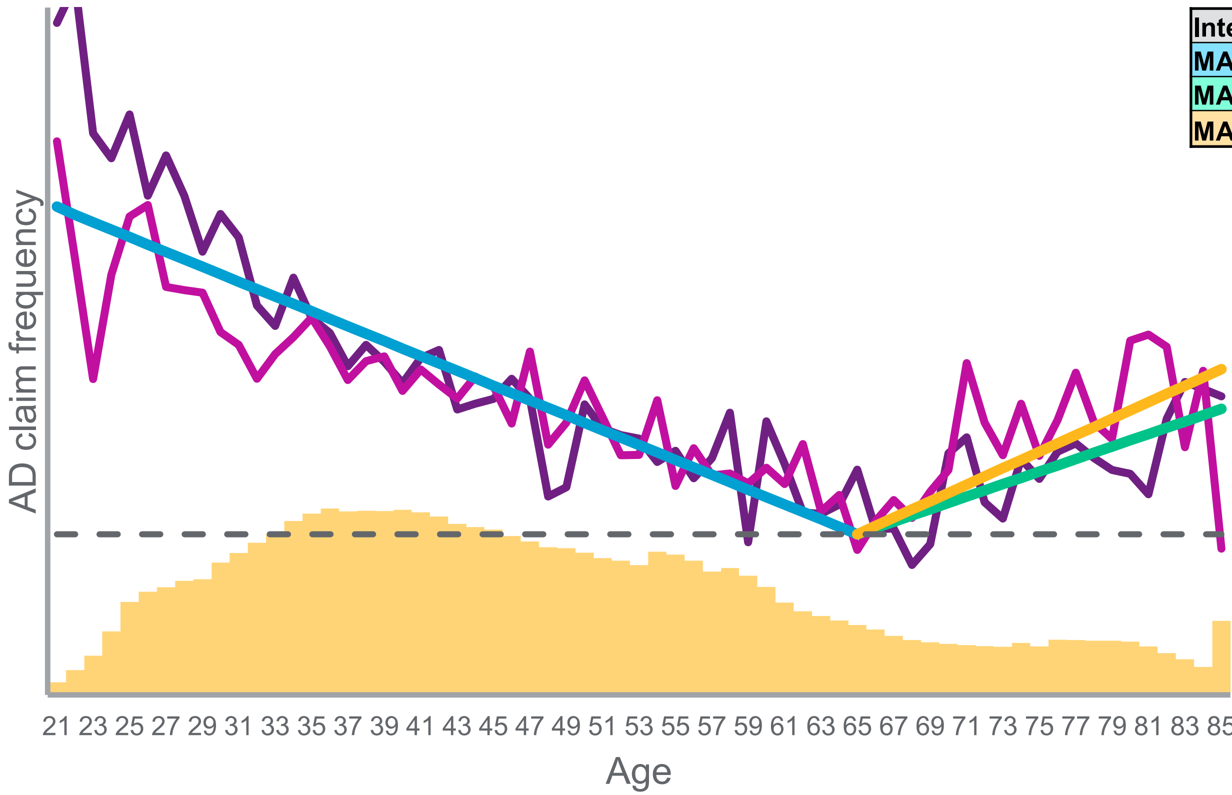
Interactions



Intercept	**-3.143**
MAX(65-Age,0)	**0.013**
MAX(Age-65,0)	**0.011**

Multivariate adaptive regression splines (“Earth”)

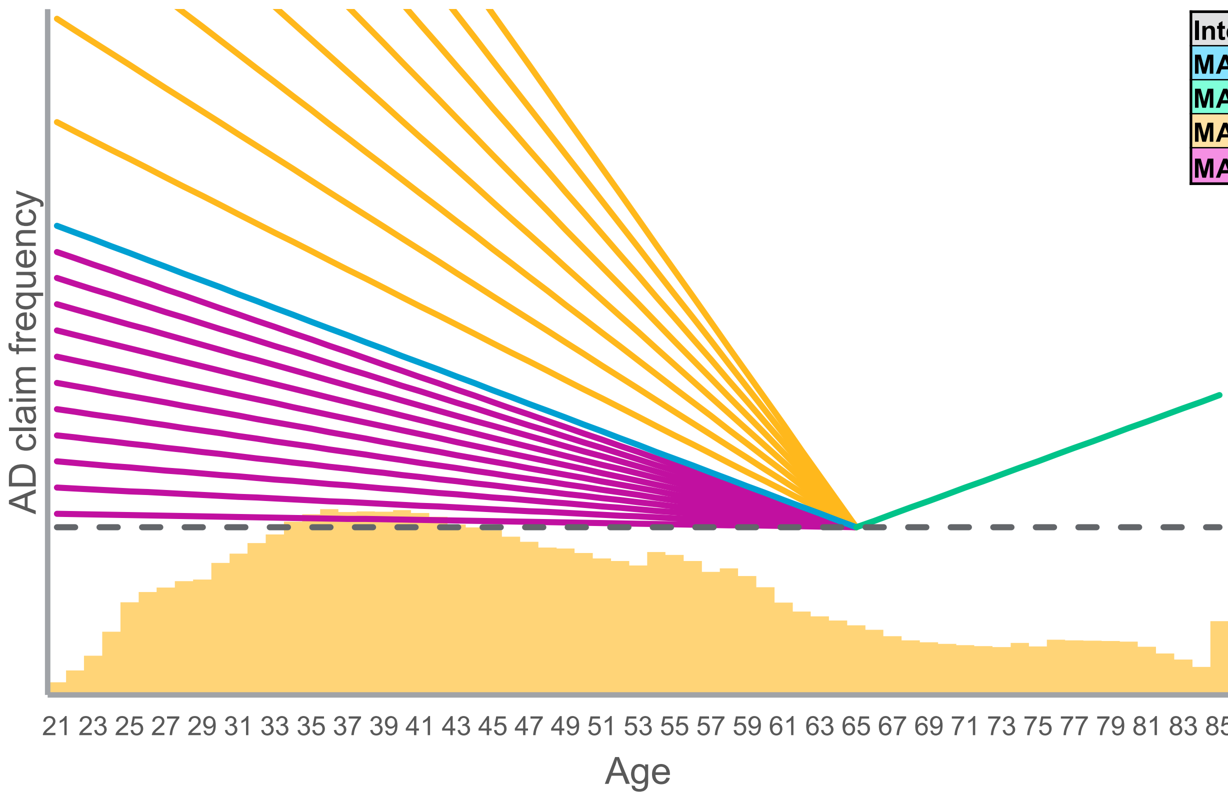
Interactions



Intercept	**-3.143**
MAX(65-Age,0)	**0.013**
MAX(Age-65,0)	**0.010**
MAX(Age-65,0)*(Gender=F)	**0.003**

Multivariate adaptive regression splines (“Earth”)

Interactions



Intercept	**-3.131**
MAX(65-Age,0)	**0.011**
MAX(Age-65,0)	**0.011**
MAX(65-Age,0)*MAX(VG-12,0)	**0.004**
MAX(65-Age,0)*MAX(12-VG,0)	**-0.001**

Multivariate adaptive regression splines (“Earth”)

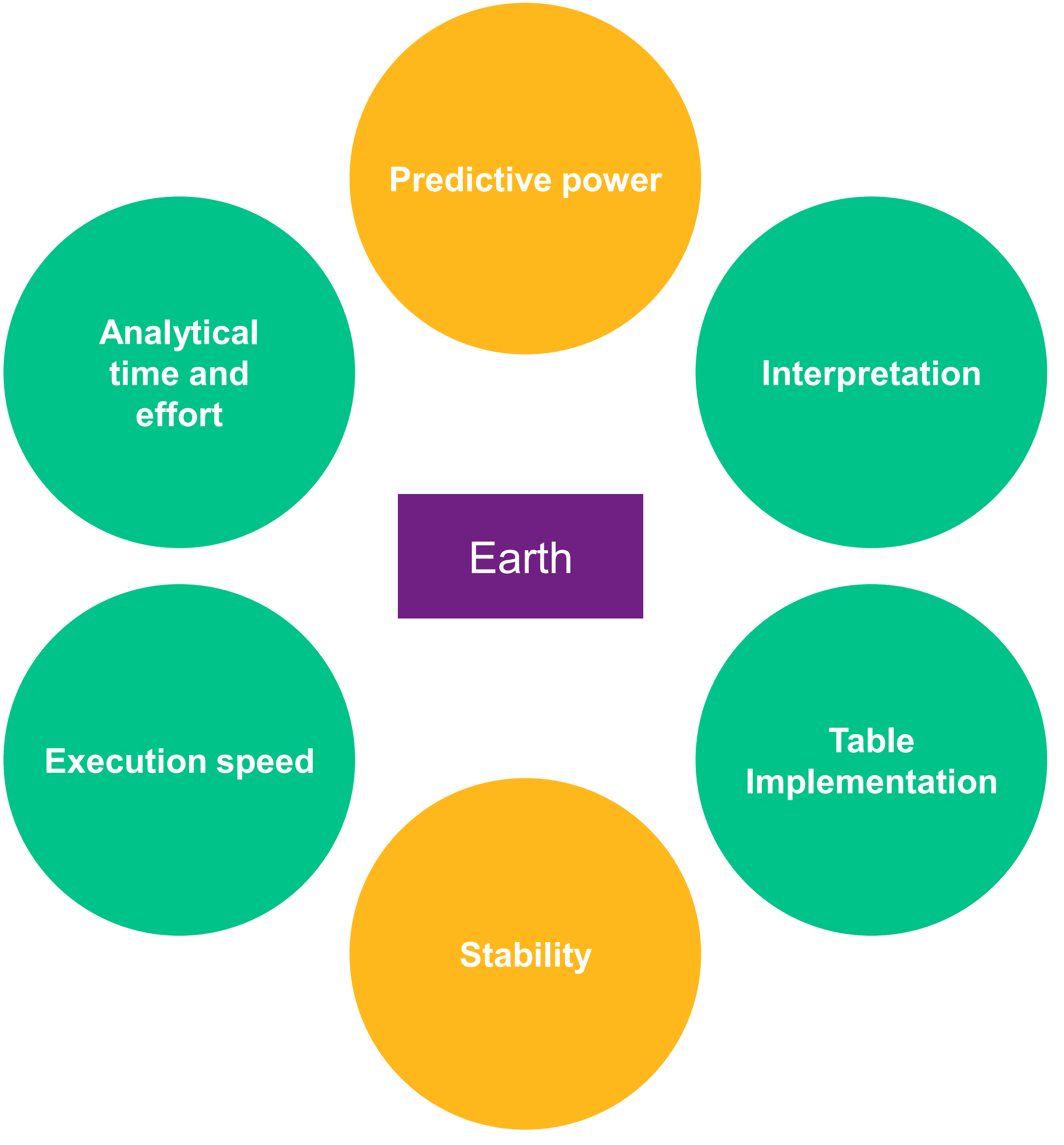
Advantages

- Minimum manual setup required
- Fast run time
- Highly interpretable results

Disadvantages

- Model will contain discontinuities around knot points
- Hand-crafting likely to improve results

Intercept	0.412
UsuallyPayANNUAL	0.543
h(Log_Premium – 6.314)	0.432
h(Age-35)	-0.329
UsuallyPayANNUAL * h(Log_Premium-6.5673)	0.00654
Homeowner	-0.0291
etc	….


Predictive power
Analytical time and effort
Interpretation
Earth
Execution speed
Table Implementation
Stability

Some machine learning methods

- Ensembles
- Classifications Trees
- "Earth"
- Regression Trees
- Gradient Boosting Machines
- K-nearest Neighbors
- Elastic Net
- Neural Networks
- Naïve Bayes
- Random Forests
- K-Means Clustering
- Principal Components Analysis
- Lasso
- Support Vector Machines
- Ridge Regression

Focus on Penalized Regression

Elastic Net

Lasso

Ridge Regression

Penalized Regression

Overview

GLMs

- Predictions are given by $f(\underline{x}) = g^{-1}(\mathbf{X}.\underline{\beta})$
- $\underline{\beta}$ is estimated by minimizing a loss function $L(\underline{\beta}|\mathbf{X},\underline{y})$ ($\mathbf{X}$ is data & model, $\underline{y}$ the response)

Penalized regression

- The same, except the objective function becomes $L(\underline{\beta}|\mathbf{X},\underline{y}) + \lambda.$ "Penalty on $\underline{\beta}$"

Elastic Net

Minimize: $L(\beta|X,y) + \boxed{\lambda_1 \sum_i |\beta_i|} + \boxed{\lambda_2 \sum_i \beta_i^2}$

Lasso **-** just the **blue** part

- Penalty reduces insignificant parameter values to zero – useful for variable selection

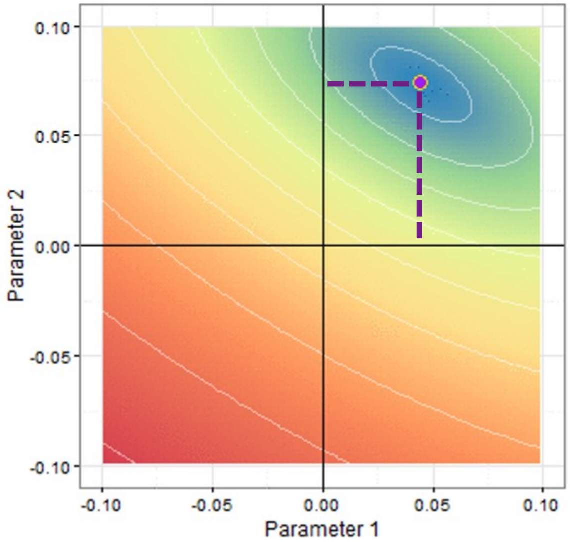
Ridge **-** just the **purple** part regression models

- Penalty heavily penalize extreme parameters, but do not reduce parameters to zero

Penalized Regression

GLM

f(x) = $g^{-1}(\mathbf{X}.\underline{\beta})$ where $\underline{\beta}$ estimated by minimizing $L(\beta|X,y)$



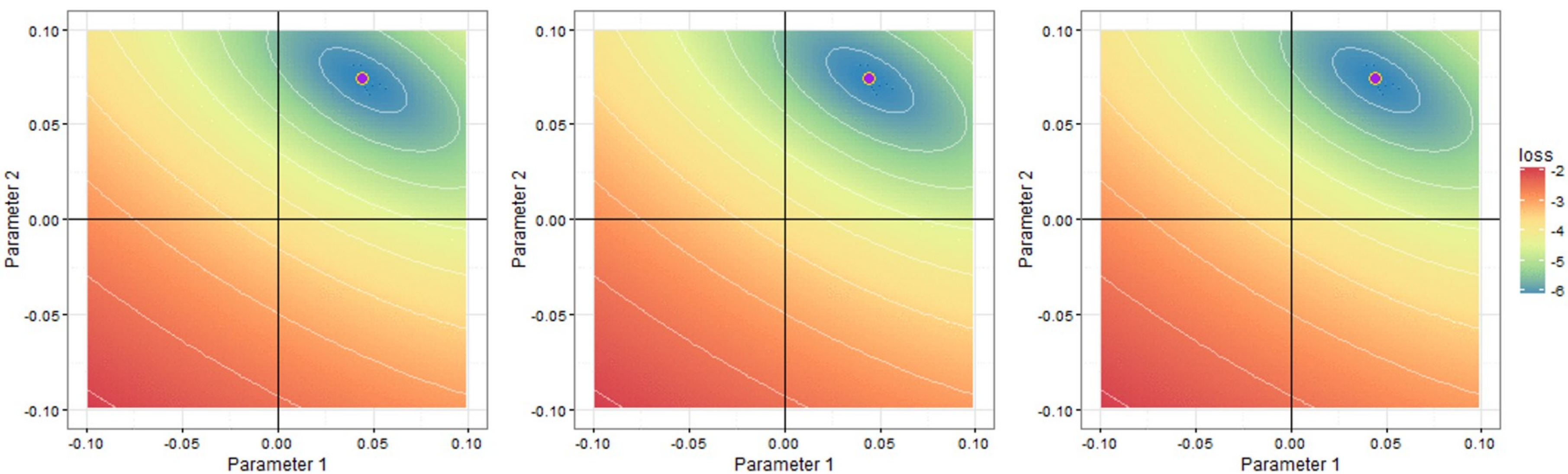
Penalized Regression

GLM Lasso Ridge

f(x̲) = g^{-1}(**X**.β̲) where β̲ estimated by minimizing $L(\beta|X,y) + \boxed{\boxed{\lambda_1 \sum_i |\beta_i|} + \boxed{\lambda_2 \sum_i \beta_i^2}}$

Elastic Net

Ridge $\sum_i \beta_i^2$ **Elastic Net** **Lasso** $\sum_i |\beta_i|$



Heavily penalize large parameters, but does not reduce parameters to zero

Mix of the two

Penalty reduces insignificant parameter values to zero - useful for variable selection

Penalized Regression

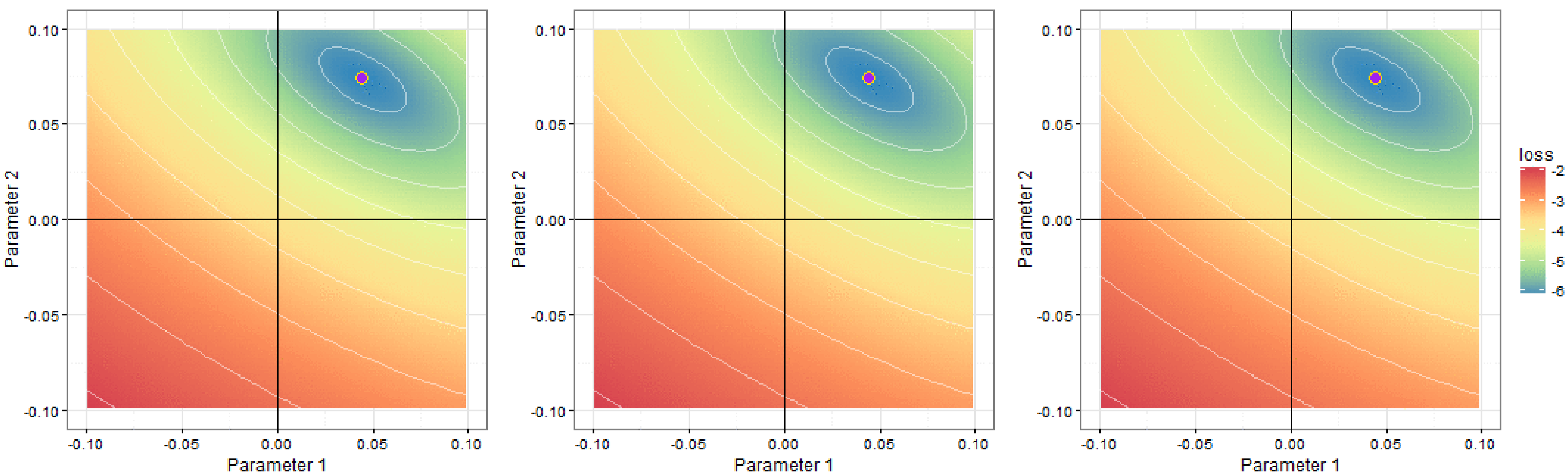
f(x) = $g^{-1}(\mathbf{X}.\beta)$ where β estimated by minimizing

GLM **Lasso** **Ridge**

$$L(\beta|X,y) + \boxed{\boxed{\lambda_1 \sum_i |\beta_i|} + \boxed{\lambda_2 \sum_i \beta_i^2}}$$

Elastic Net

Ridge $\sum_i \beta_i^2$ | **Elastic Net** | **Lasso $\sum_i |\beta_i|$**



Heavily penalize large parameters, but does not reduce parameters to zero

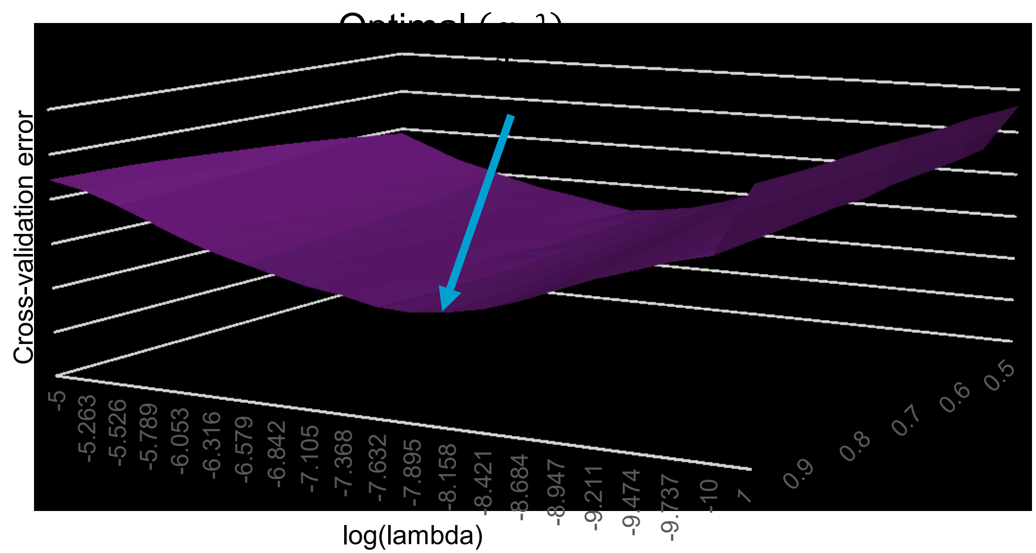
Mix of the two

Penalty reduces insignificant parameter values to zero - useful for variable selection

Penalized Regression

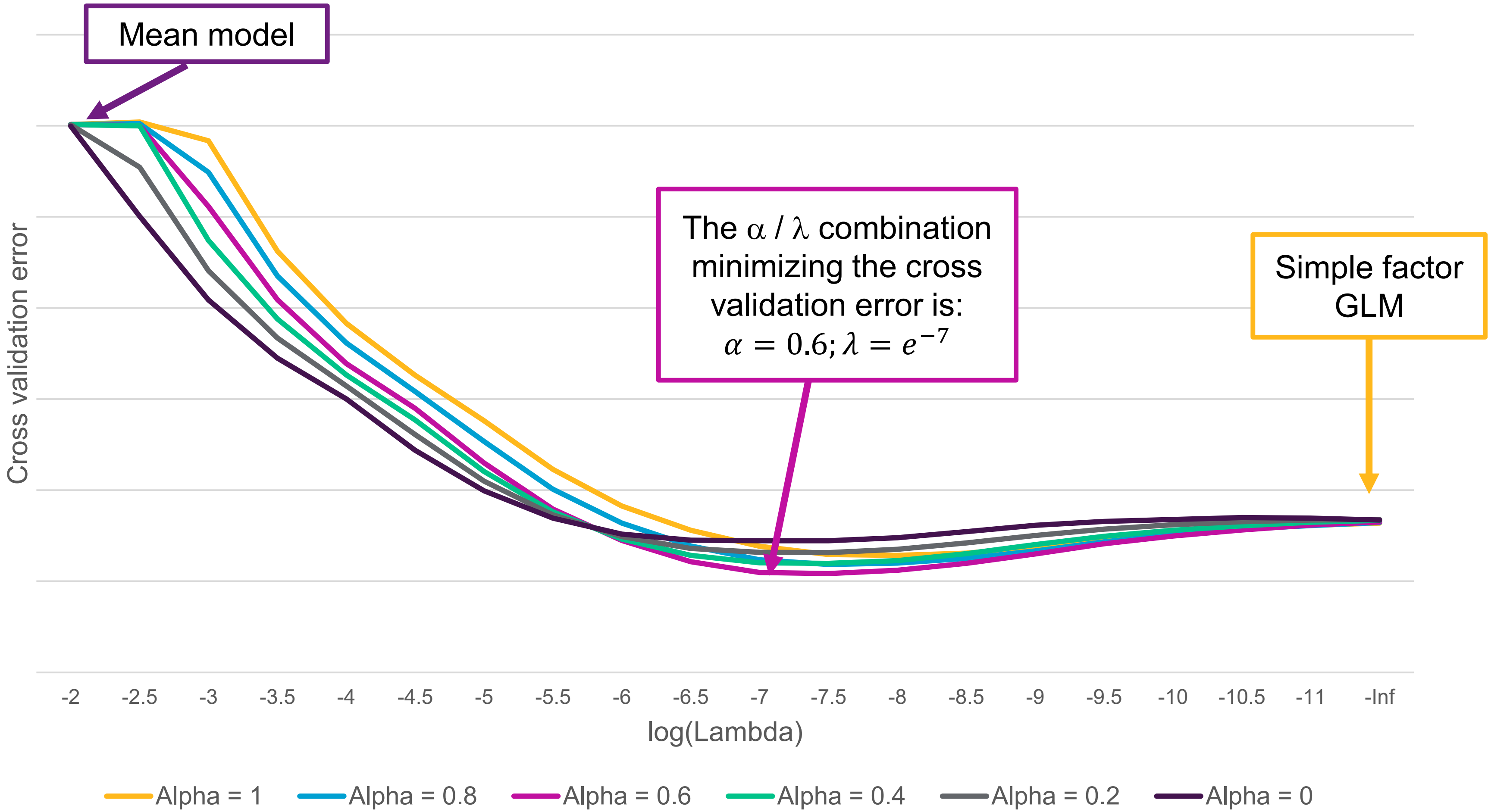
Parameter selection

- Minimize: $L(\beta|X, y) + \lambda_1 \sum_i |\beta_i| + \lambda_2 \sum_i \beta_i^2$
- Penalty parameters can be re-written: $\lambda_1 = \lambda\alpha, \quad \lambda_2 = \lambda\left(\frac{1-\alpha}{2}\right)$
- α controls the mixture between Lasso ($\alpha = 1$) and Ridge ($\alpha = 0$)
- λ controls the overall size of the penalty
- λ, α selected using cross-validation
- Factors automatically selected from initial set!



Penalized Regression

Parameter selection - example

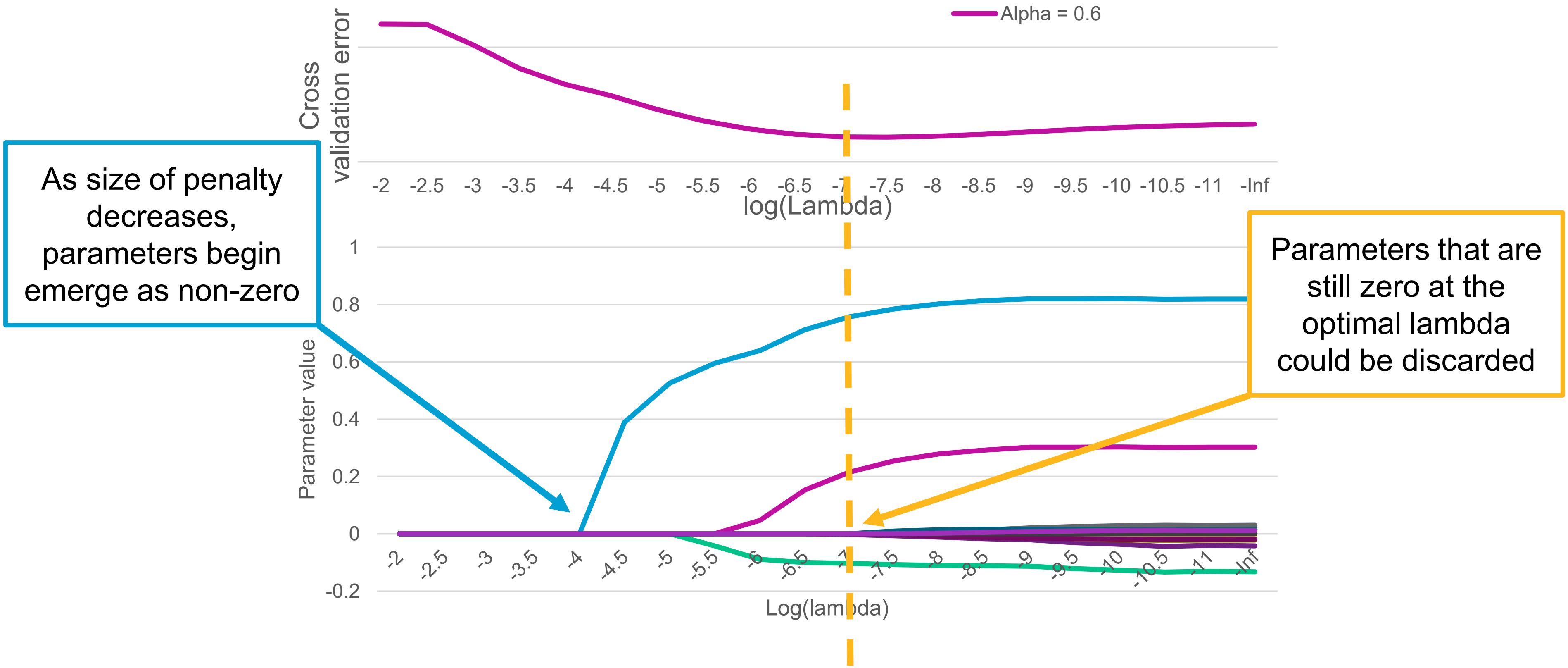


Models range from Lasso (α =1) to Ridge (α =0)

Penalized Regression

Parameter selection - example

- The fitting process can be investigated to help with feature selection



Penalized Regression

Parameter selection

There are costs to allowing too many factors in our models

- Computational cost of processing more data / fitting more parameters
- Time cost of analysts needing to consider more potential effects
- Reduced comprehensibility of interplay of many different correlated effects in our models
- Financial cost of licensing and maintaining many different data sources, and hosting/updating tables to use them in rating
- Performance cost as increased number of tests makes it more likely that we will find false-positives and overfit to noise in our data

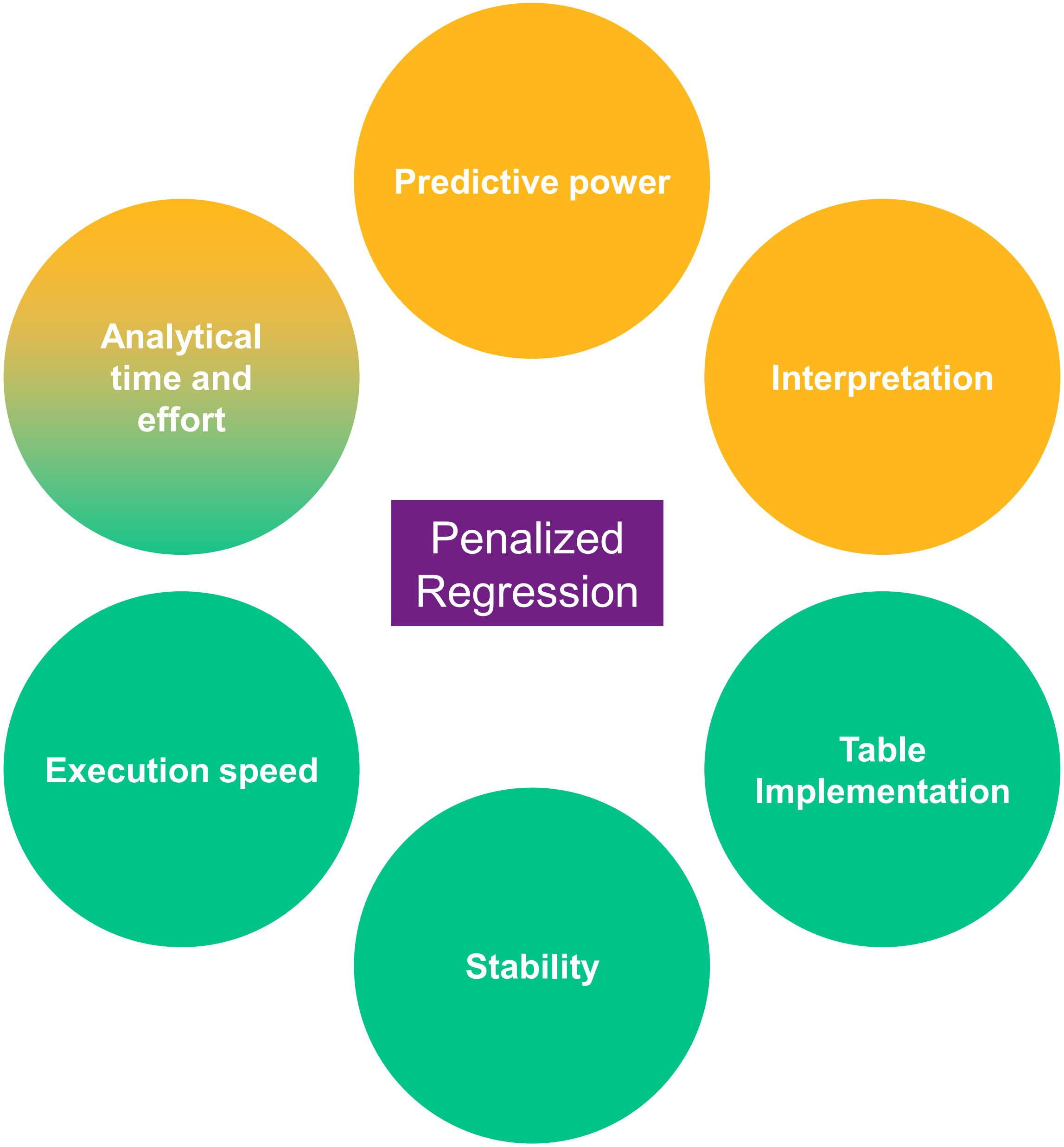
Deploying Penalized Regression

Same as GLMs!

	Age	Exposure	Loss Cost
1	<=20	1,720	179
2	21-30	34,893	122
3	31-50	118,182	102
4	51+	127,054	70
5	Age Total	281,849	91

	Vehicle Group	Exposure	Loss Cost
1	1-10	164,107	77
2	11-14	84,859	101
3	15-18	28,952	116
4	19-20	3,931	272
5	VG Total	281,849	91

	Gender	Exposure	Loss Cost
1	Male	197,339	92
2	Female	84,510	87
3	Gender Total	281,849	91


Predictive power
Analytical time and effort
Interpretation
Penalized Regression
Execution speed
Table Implementation
Stability

Some machine learning methods

- Ensembles
- Classifications Trees
- "Earth"
- Regression Trees
- Gradient Boosting Machines
- K-nearest Neighbors
- Elastic Net
- Neural Networks
- Naïve Bayes
- Random Forests
- K-Means Clustering
- Principal Components Analysis
- Lasso
- Support Vector Machines
- Ridge Regression

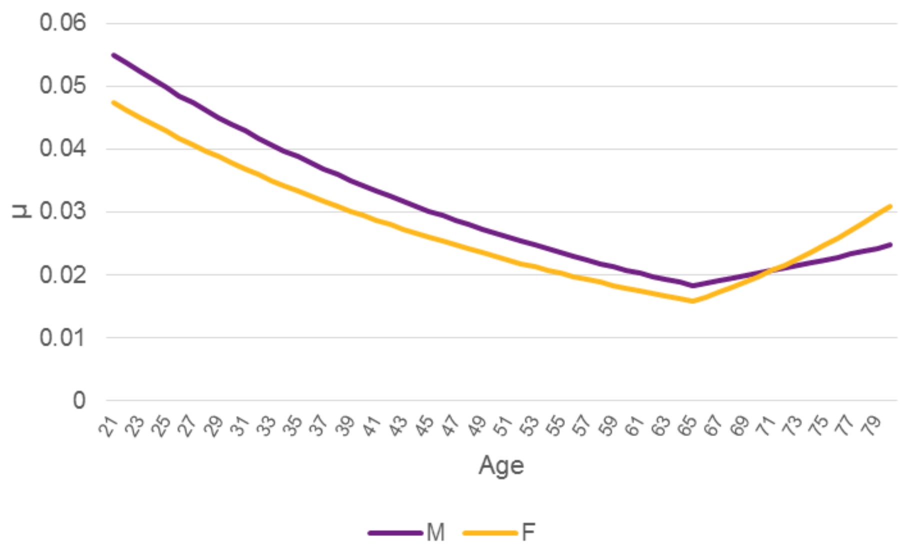
Focus on Neural Networks

Neural Networks

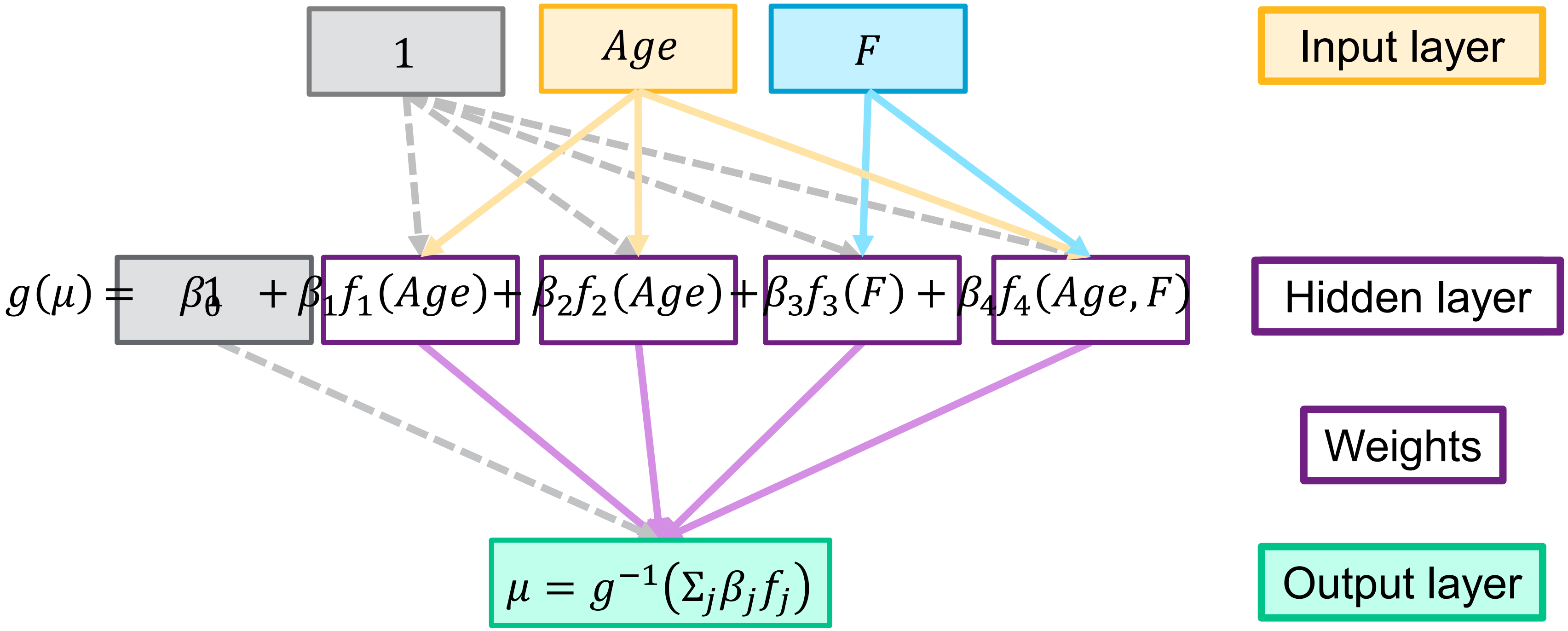
Start with a simple GLM…

- Log link function, g
- Age (piecewise-linear variates)
- F (indicator of Gender = Female)
- Age x Gender interaction

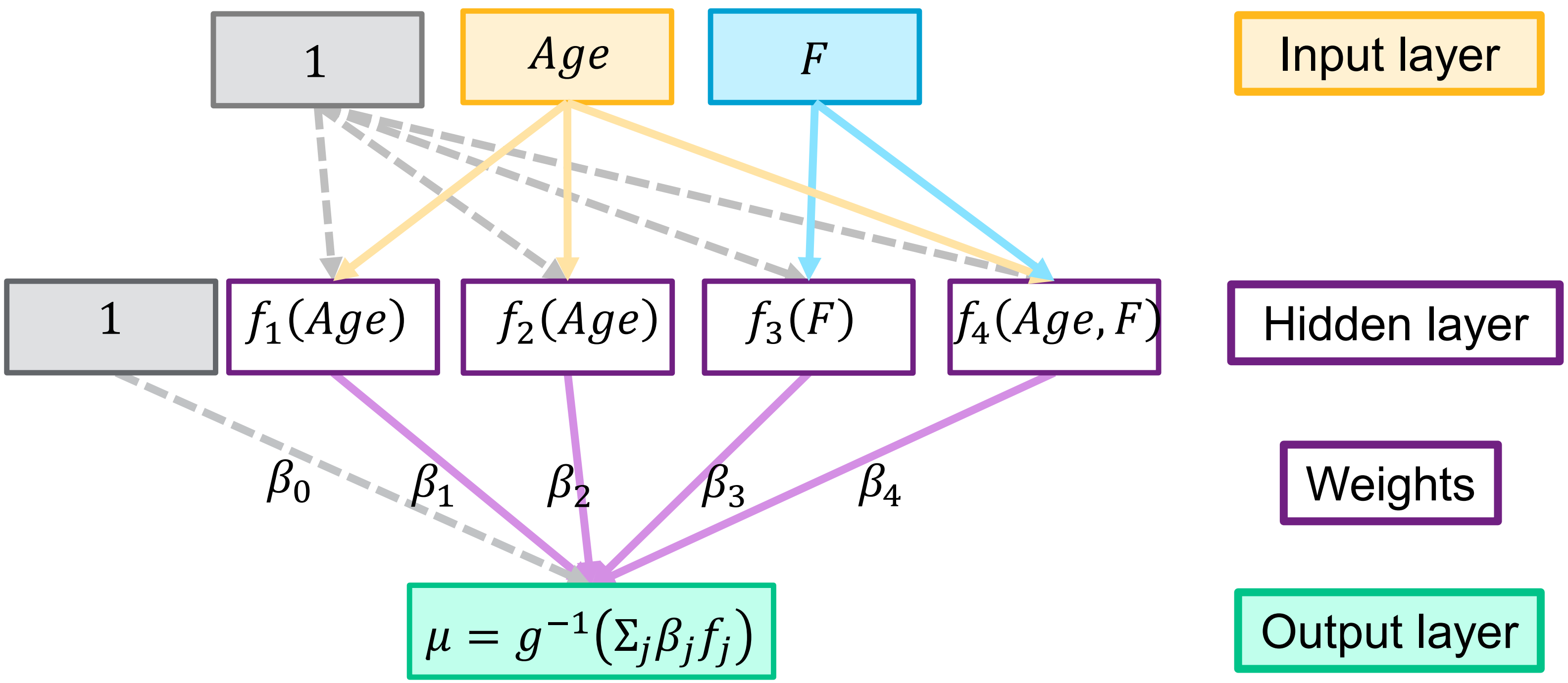
$$g(\mu) = \quad \beta_0 \quad + \beta_1 f_1(Age) + \beta_2 f_2(Age) + \beta_3 f_3(F) + \beta_4 f_4(Age, F)$$



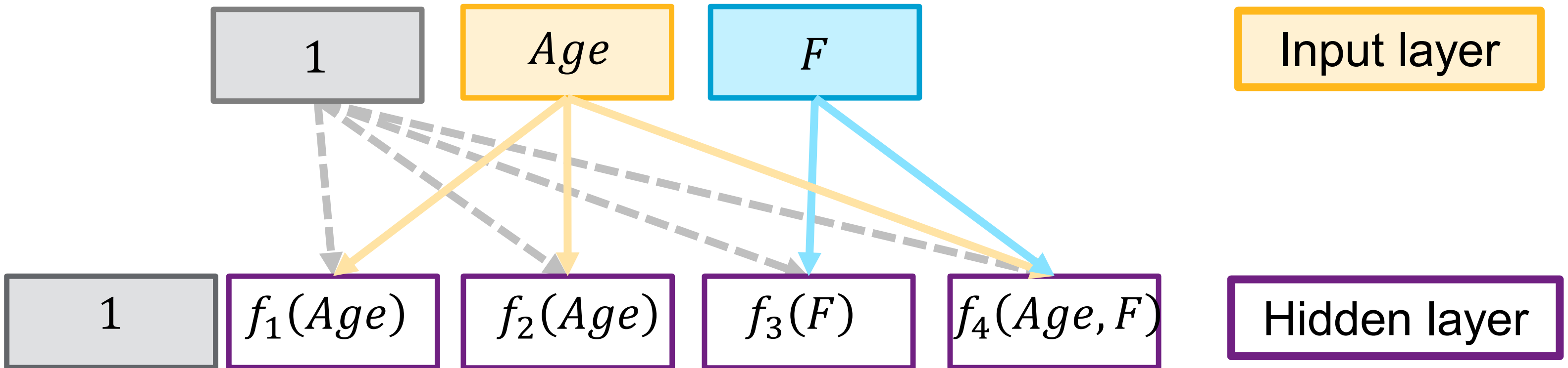
We can represent GLMs as a network…



We can represent GLMs as a network…



We can represent GLMs as a network…



Hidden layer represents our manually engineered features:

- $f_0 = 1$
- $f_1 = \max(65 - Age, 0)$
- $f_2 = \max(Age - 65, 0)$
- $f_3 = F$
- $f_4 = \max(Age - 65 - 100(1 - F), 0)$

General form:

- $f_i = \text{ReLU}(w_{i,0} + w_{i,1}Age + w_{i,2}F)$

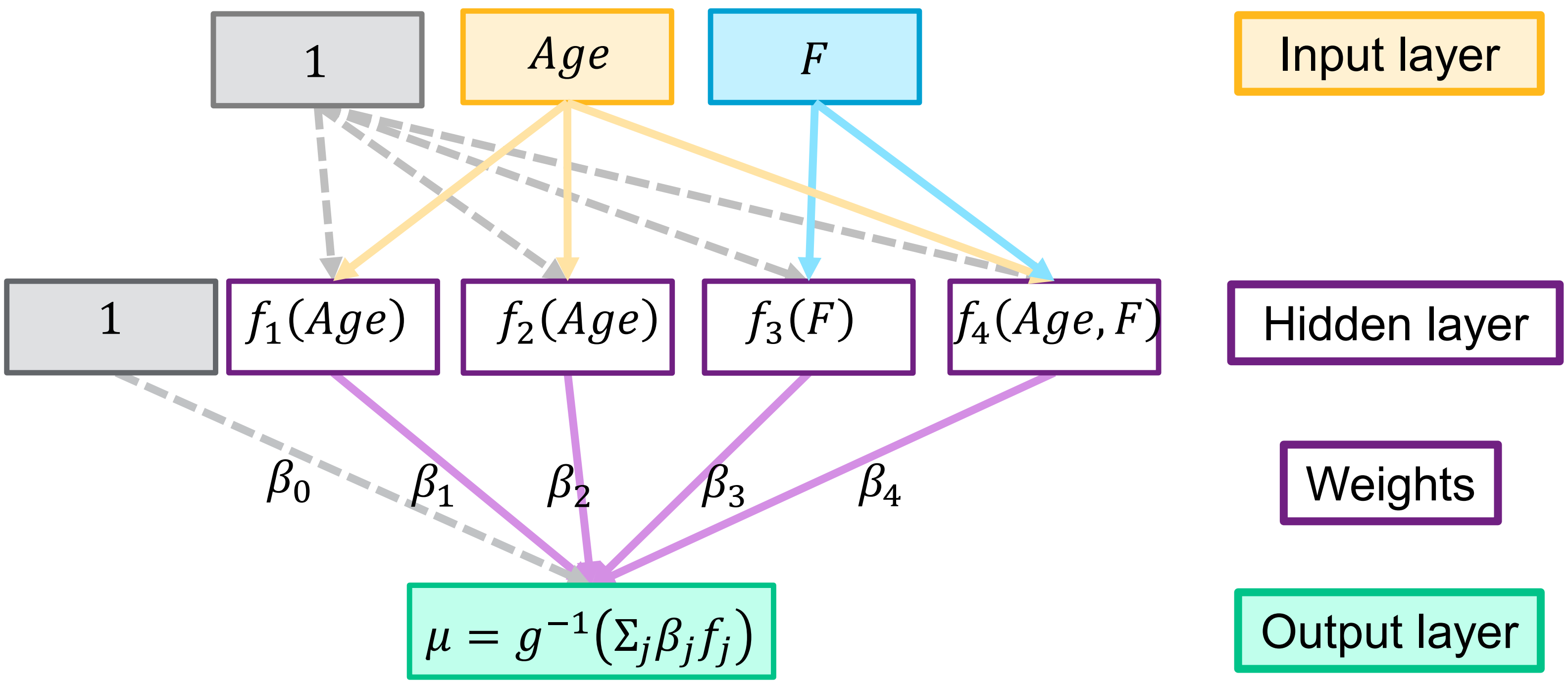
Activation function breaks linearity:

$$\text{ReLU}(x) = \max(x, 0)$$

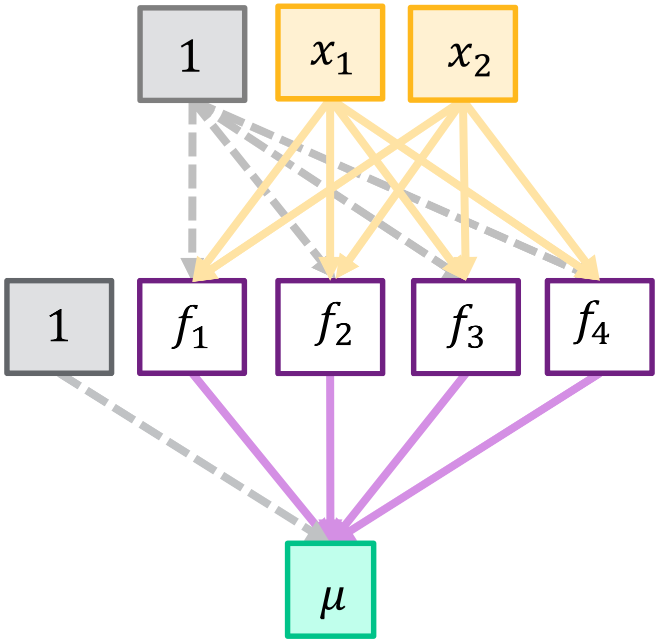
*(****R****ectified* ***L****inear* ***U****nit)*

Universal approximation theorem:
We can approximate (almost*) any function arbitrarily well with a single hidden layer
*(*continuous, on compact subsets)*

We can represent GLMs as a network…



Generalizing to neural networks



Input layer (L0): x_i

Weights: w_{ij}

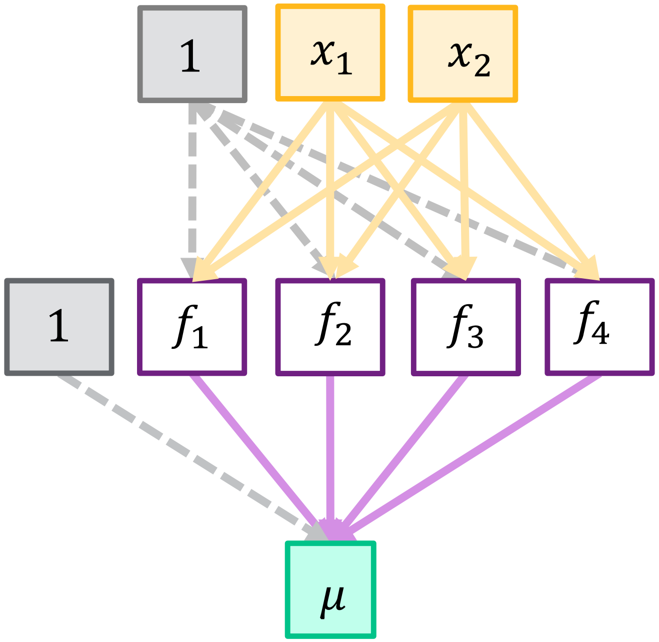
Hidden layer (L1): $f_j = h_1\left(\Sigma_i w_{ij} x_i\right)$

Weights: β_j

Output layer (L2): $\mu = h_2\left(\Sigma_j \beta_j f_j\right)$

Generalizing to neural networks

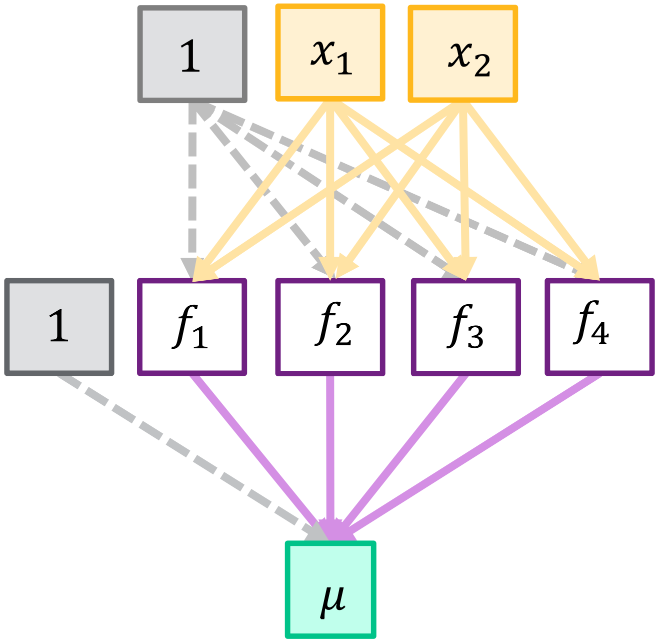
Model structure decisions



- Input features
- Number of hidden layers
- Size of each hidden layer
- Activation functions
  - Typically specified by layer
  - ReLU is most commonly used
- Connectivity of layers and weight sharing
  - Typically **fully connected** with **unique weights**
  - Many variants exist, eg: **Convolutional Neural Networks** for image classification connect nearby blocks of pixels and apply the same shared weights across each block

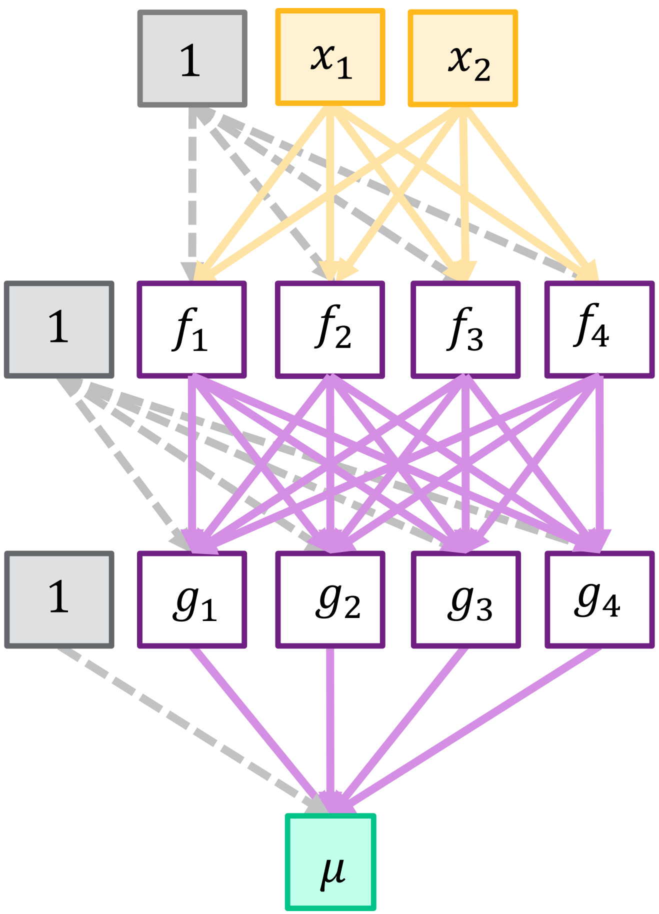
Generalizing to neural networks

Key model fitting decisions



- **Optimization algorithm**
  - Typically variants of **Back-Propagation**
- **Loss function** – to be minimized
- **Batch size** – number of rows to consider in each iteration
- **Epochs** – number of passes through full data
- **Initial weights**
- **Regularization parameters**, eg:
  - L1 / L2 penalties
  - Learning rate and decay
  - Dropout

Generalizing to neural networks



Input layer (L0): x_i

Weights: w_{ij}

Hidden layer (L1): $f_j = h_1\left(\Sigma_i w_{ij} x_i\right)$

Weights: v_{jk}

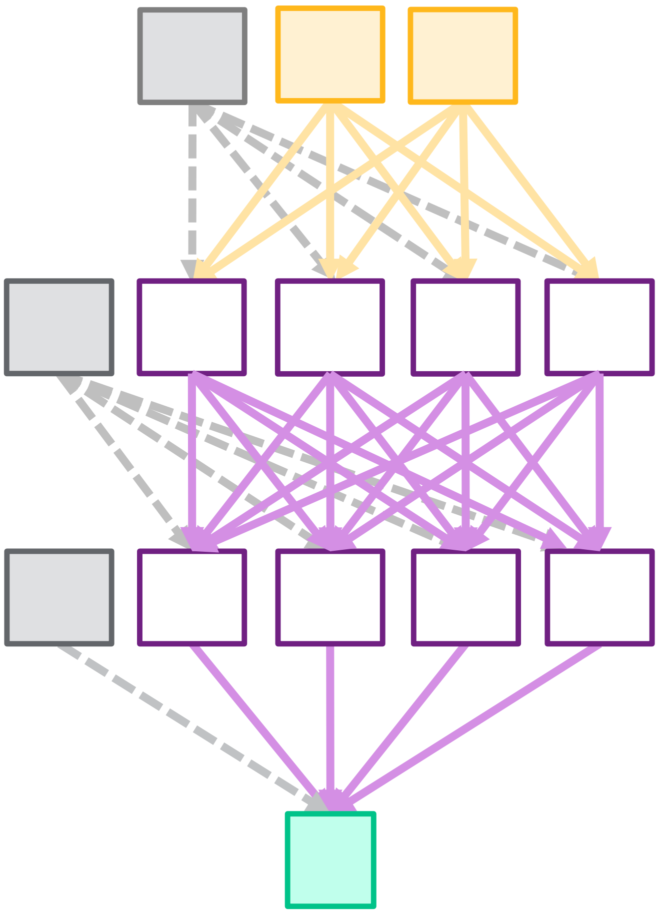
Hidden layer (L2): $g_k = h_2\left(\Sigma_j v_{jk} f_j\right)$

Weights: β_k

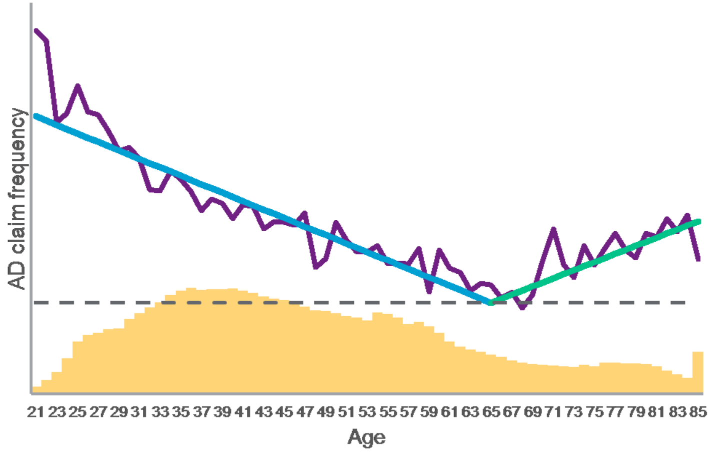
Output layer (L3): $\mu = h_3(\Sigma_k \beta_k g_k)$

"Deep learning" refers to multiple hidden layers

Generalizing to neural networks

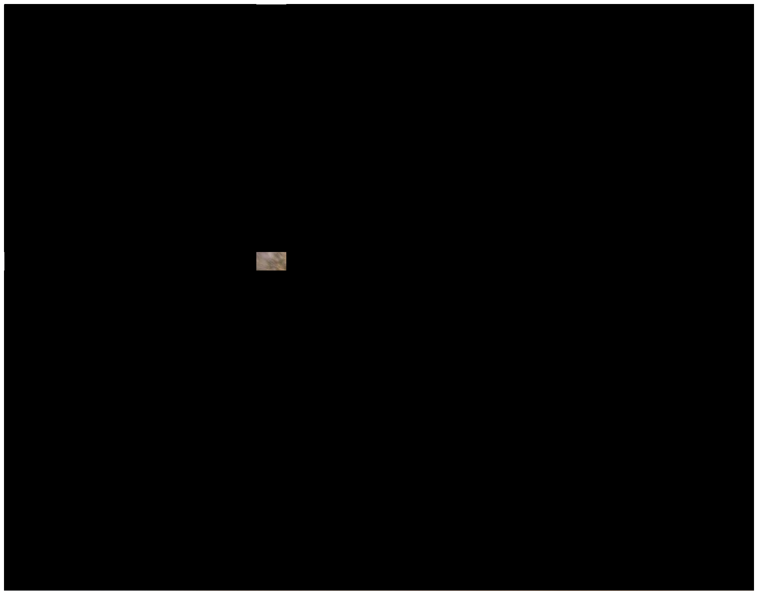


Where is the value?

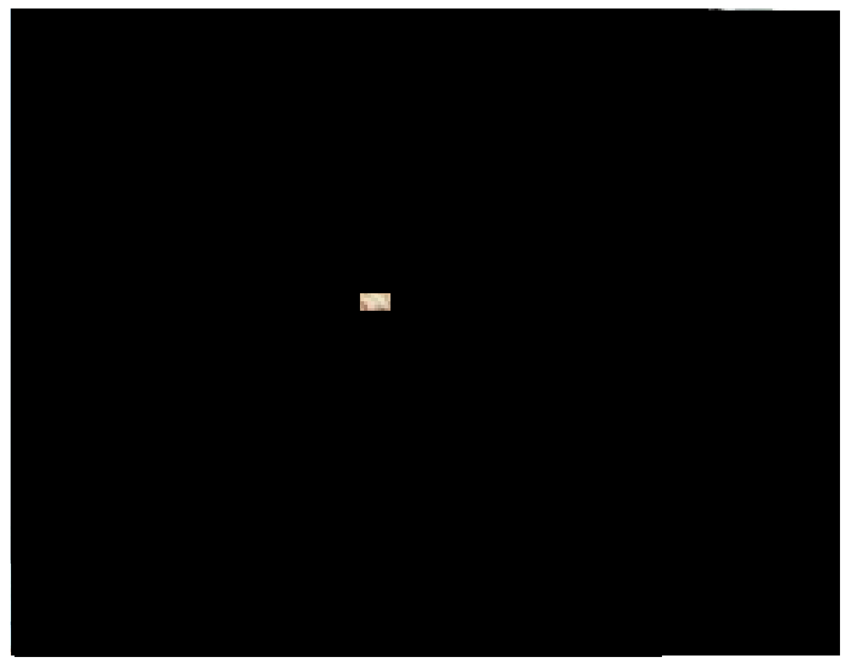


Which policyholder is more likely to make a claim?

Where is the value?



Which picture is more likely to be of a cat?

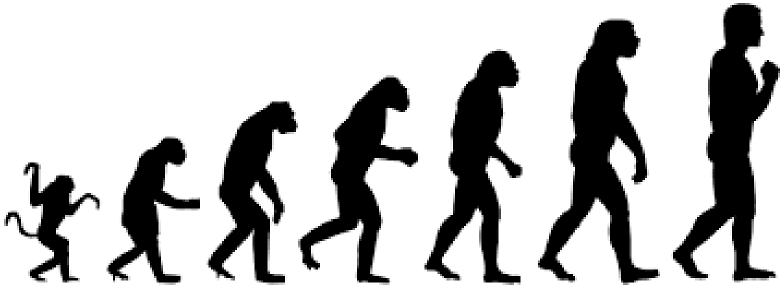


Where is the value?

Which picture is more likely to be of a cat?

Neural networks

Evolution or revolution?



Neural networks

Case study – market models

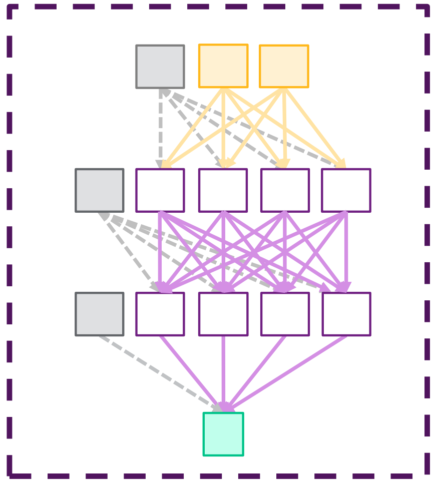
Context

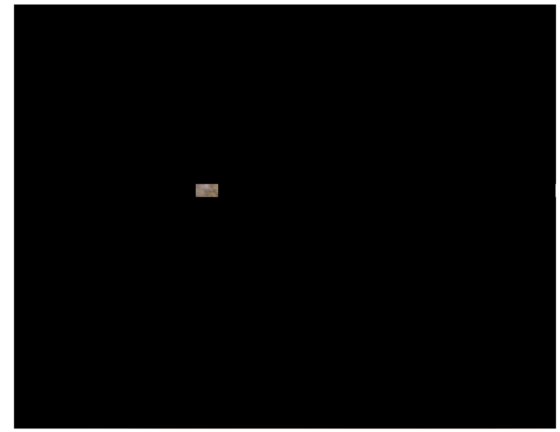
- UK aggregator sites provide some historic quote data
- We wanted a model of "Average top 5 premium" for auto quotes to understand the market's pricing structure
- One month of data (~1m quotes)
- Limited subset of factors (no data enrichment beyond simple rating area & vehicle group)

Approach

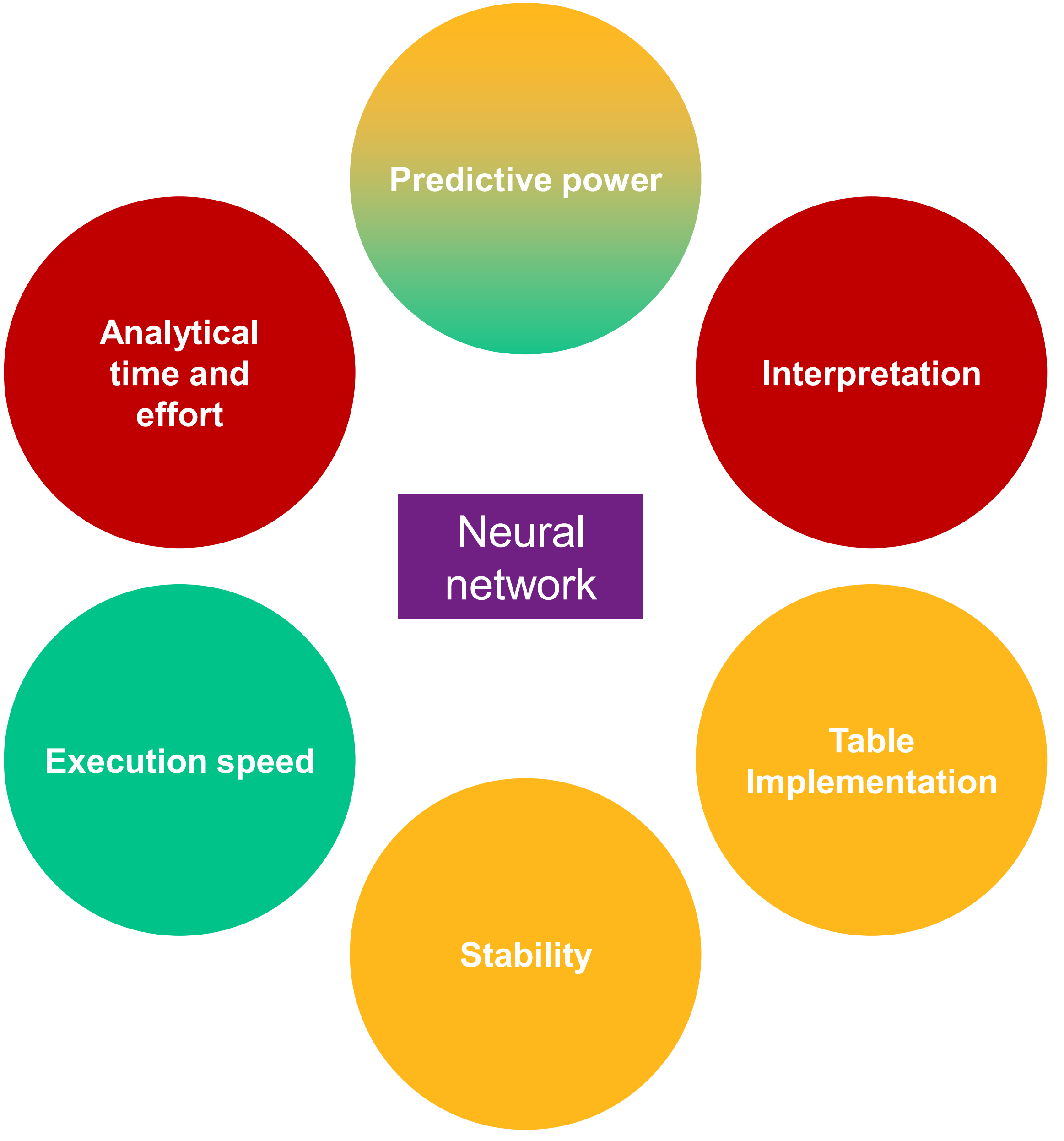
- 60/40 split for training and holdout data
- Modelled as Log-Normal (ie $\ln(\text{Premium}) \sim N(\mu, \sigma^2)$) as Normal distributions well supported across packages
- Compare Neural Network performance to GLM (using existing model parameterizations) and GBM with RMSE of log-Premium on holdout data

Practical applications of neural networks in pricing

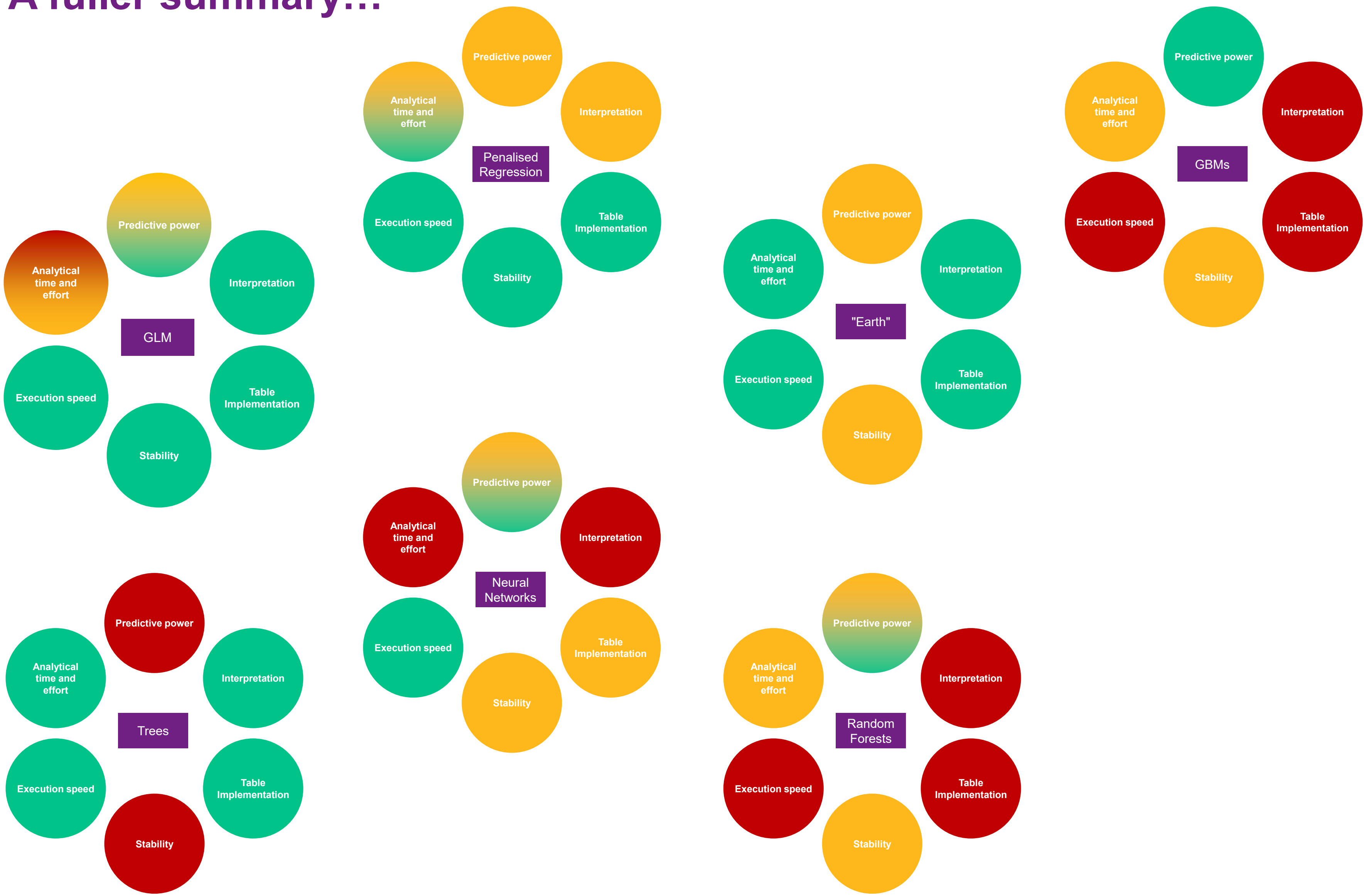




Marketing and Distribution → Customer management → Underwriting and risk management → Pricing → Claims management → Asset management

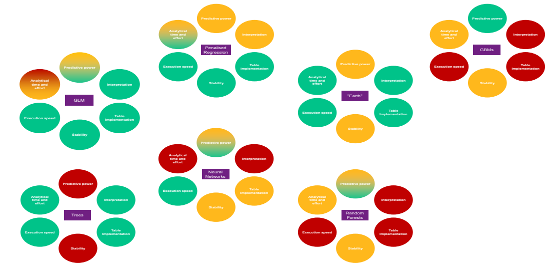

Predictive power
Analytical time and effort
Interpretation
Neural network
Execution speed
Table Implementation
Stability

A fuller summary…



Machine learning in pricing

Conclusions (Part 2)



- Machine learning brings a proliferation of new methods
- Improving models is more than just finding the best method. Consider:
  - What data are available and how can data be transformed to give insight
  - What is the optimal model structure and target variable?
  - How can information be transferred between models?

- Earth is a fast, interpretable method that can improve overall lift by informing when/where to segment models
- Neural networks are complex and require numerous input decisions; analyzing unstructured data (e.g., imagery) is an intuitive application for this method … but where else may it be helpful?
- Penalized regression can aid in factor selection decisions and may in fact be a good method in its own right – particularly when the modeler has less of a "feel" for the data

- Machine learning in pricing is not all about improving predictive power. Consider:
  - Fast investigation of new data
  - Quick assessment and response of emerging experience

So what? How the North American market is doing with machine learning

Some critical success factors

Component	Rating	Directional trend
Data availability		Static
Appetite to try new approaches		
Modelling tools and platforms		
Internal skills sets		
Measuring value		
Application		

So what? How the North American market is doing with machine learning

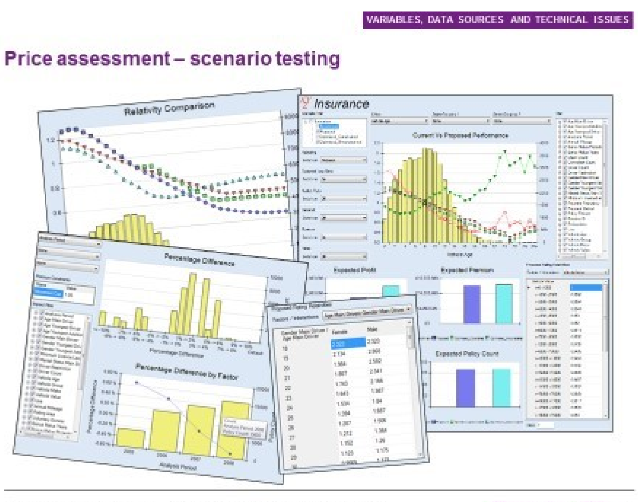
Some critical success factors

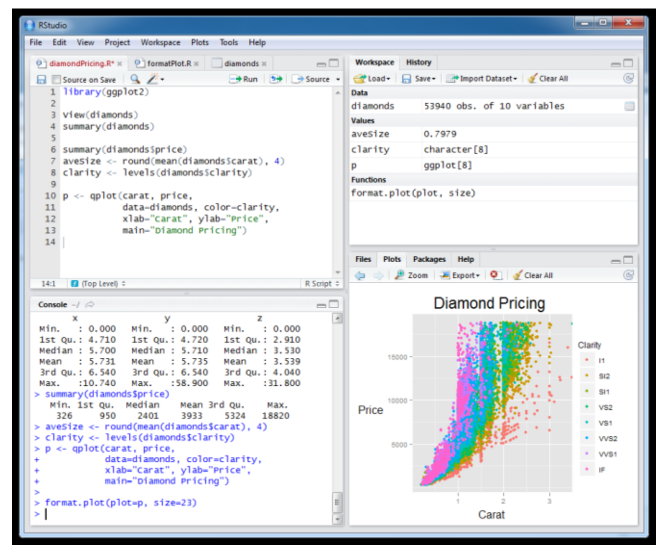
Component	Rating	Directional trend
Data availability		Static
Appetite to try new methods		Slowly upward
Modeling tools and platforms		
Internal skills sets		
Measuring value		
Application		

So what? How the North American market is doing with machine learning

Some critical success factors

Component	Rating	Directional trend
Data availability		Static
Appetite to try new approaches		
Modeling tools and platforms		Slowly upward
Internal skills sets		
Measuring value		
Application		





So what? How the North American market is doing with machine learning

Some critical success factors

Component	Rating	Directional trend
Data availability		Static
Appetite to try new approaches		
Modeling tools and platforms		
Internal skill sets	?	Static
Measuring value		
Application		

So what? How the North American market is doing with machine learning

Some critical success factors

Component	Rating	Directional trend
Data availability		Static
Appetite to try new approaches		
Modelling tools and platforms		
Internal skills sets		
Measuring value		Static
Application		

So what? How the North American market is doing with machine learning

Some critical success factors

Component	Rating	Directional trend
Data availability		Static
Appetite to try new approaches		
Modelling tools and platforms		
Internal skills sets		
Measuring value		Slowly upward
Application	?	Slowly upward

Machine learning beyond pricing

- Carriers are experimenting with ML, it is becoming established within insurance analytics
- It opens up a broader set of problems to analytics, and offers a broader tool set for familiar problems
- New (wider) data beats new methods – think UBI!
- Factor definition, problem specification and method selection are critical for success
- There's opportunity to reveal actionable, first-order insights in applications to which analytics have not been deployed previously
- With this broad new opportunity, spotting strong initial use cases is important

Questions

?