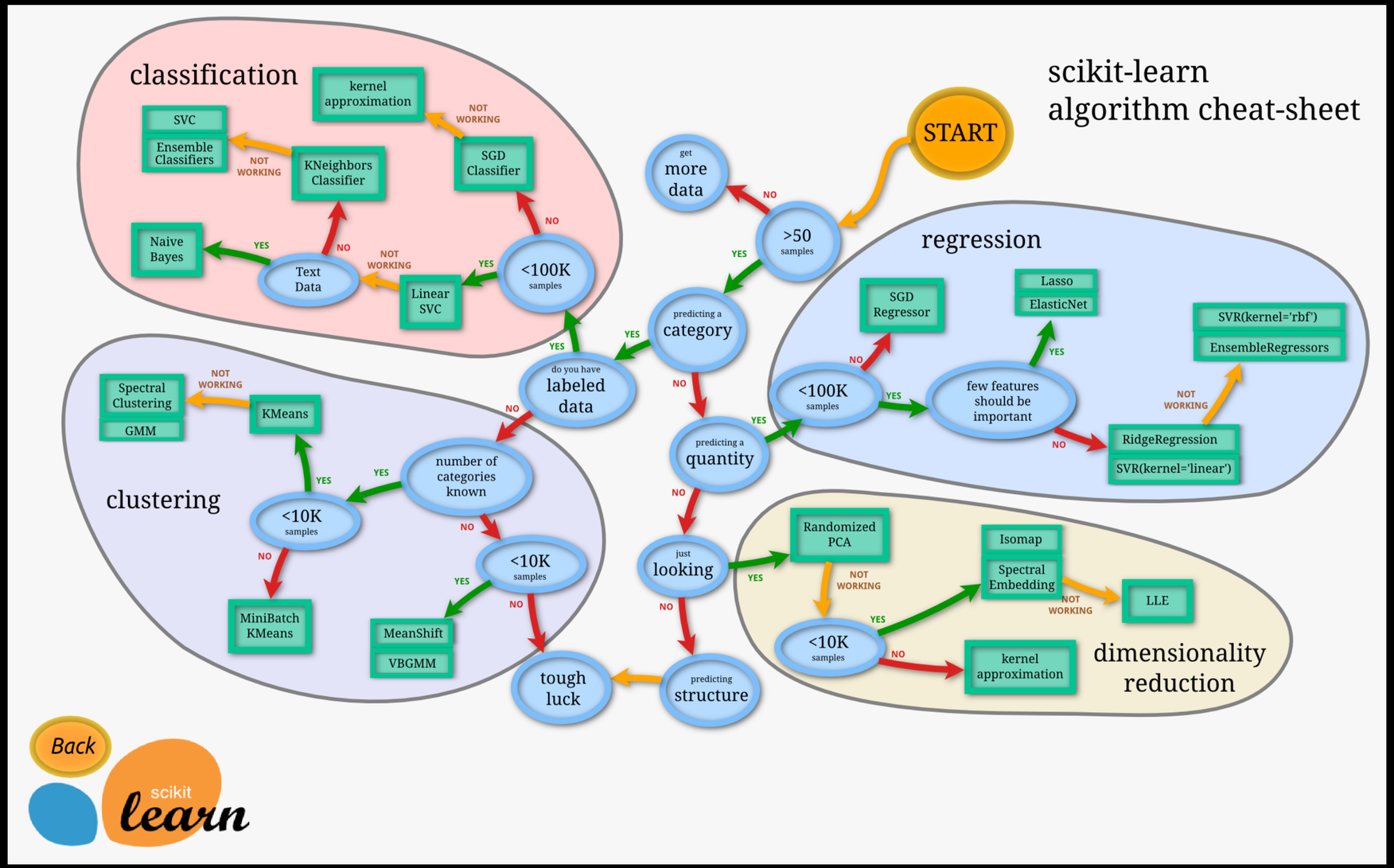
Supervised Learning
Regression Models

X-Ed
Claire Boyer PhD; Demian Wassermann, PhD

Goals for Today

- Familiarize ourselves with regression models as synthesis of our surrounding reality
- Understand the bias / variance tradeoff
- Learn how to quantify their performance
- Practical work

Where are we in the Data Science Landscape?



Motivating Example

Le taux de chômage peut-il s'expliquer par la qualité de l'éducation ?

Une étude récente décèle une relation statistique forte entre qualité d'un système éducatif et vitalité économique. Décryptage et explications sur la « corrélation statistique ».

LE MONDE I 29.08.2016 à 14h07 • Mis à jour le 30.08.2016 à 17h57 I Par Romain Damian

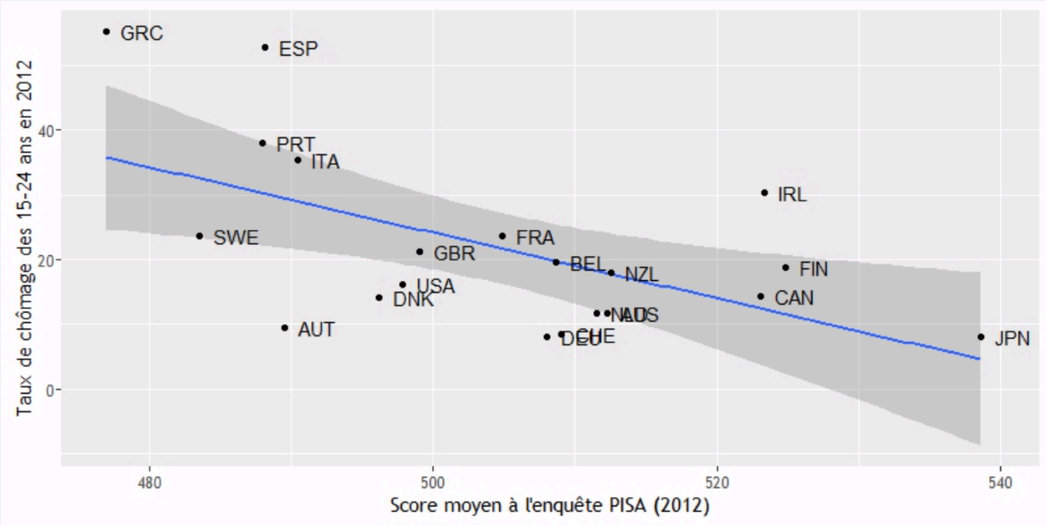
La qualité d'un système éducatif peut-elle expliquer le taux de chômage d'un pays ? Une étude récente l'affirme.
FRANCOIS GUILLOT / AFP

Plus d'écoles, moins de chômage ? C'est ce que semble affirmer une étude publiée récemment par la banque Natixis. A l'approche de la rentrée, l'économiste Patrick Artus, dans son « Flash économie » (https://research.natixis.com/GlobalResearchWeb/main/globalresearch/viewdocument?id=CYgPvF23oqm0m6y6rDUfDg==) du 3 août 2016 (un papier non académique destiné aux clients de la banque), affirme que « *la qualité du système éducatif et du système de formation professionnelle joue […] un rôle majeur pour expliquer la performance économique et sociale* » d'un pays.

▶ Une corrélation significative dans trois cas sur quatre

D'après les régressions linéaires simples des grandeurs considérées sur le score PISA, seul le poids de l'industrie dans l'emploi total ne donne pas une corrélation *significative*.

Chômage des jeunes | **Production industrielle** | **Poids de l'industrie** | **Dépenses de R&D**



SOURCE : OCDE

Regression Implies a Model of Reality

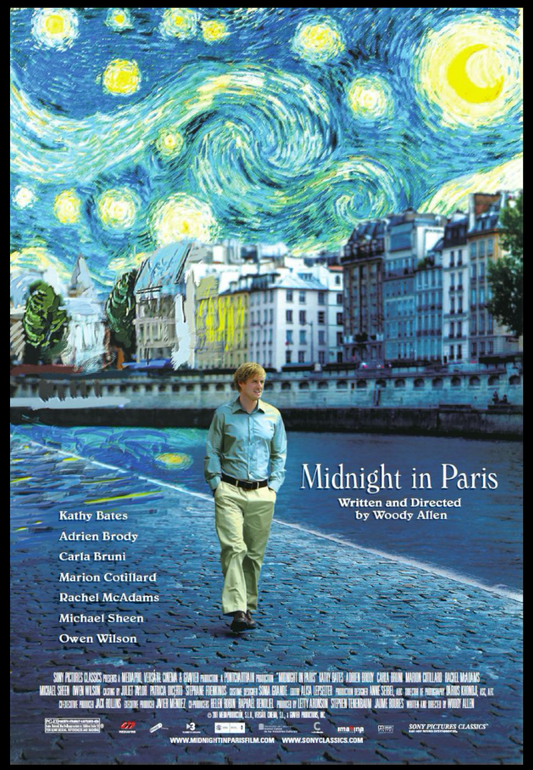
“A movie is better when there are more explosions”

Regression Implies a Model of Reality

“A movie is better when there are more explosions”

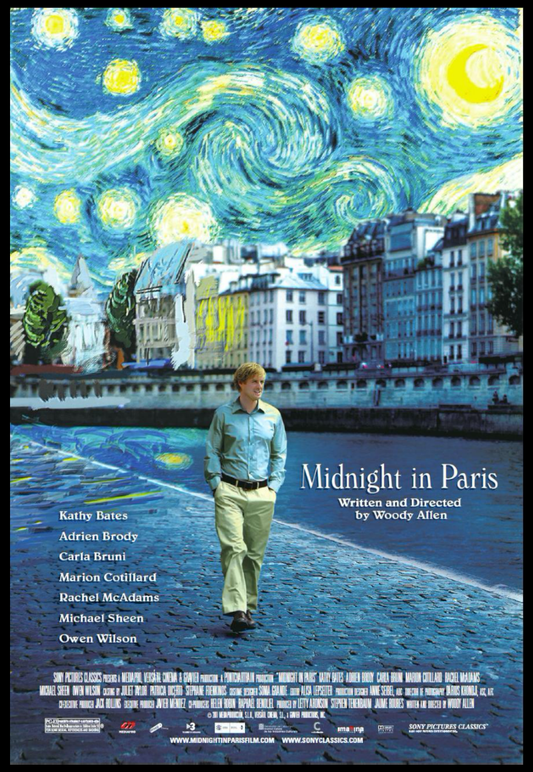
Regression Implies a Model of Reality

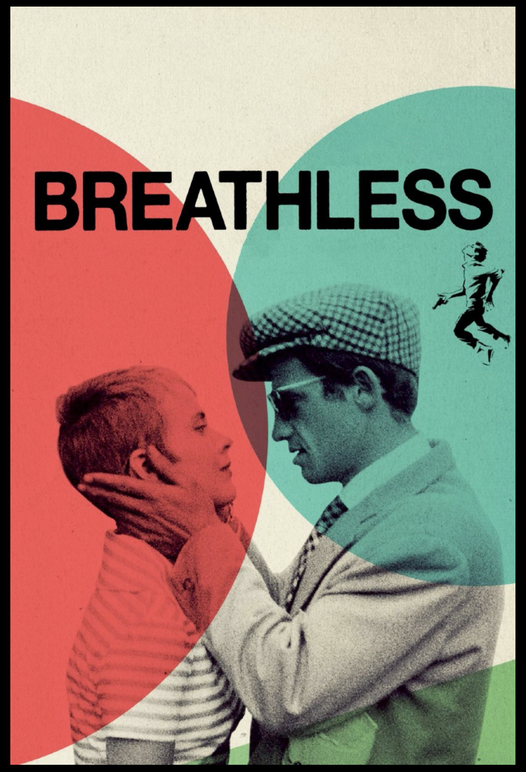
“A movie is better when there are more explosions”



Regression Implies a Model of Reality

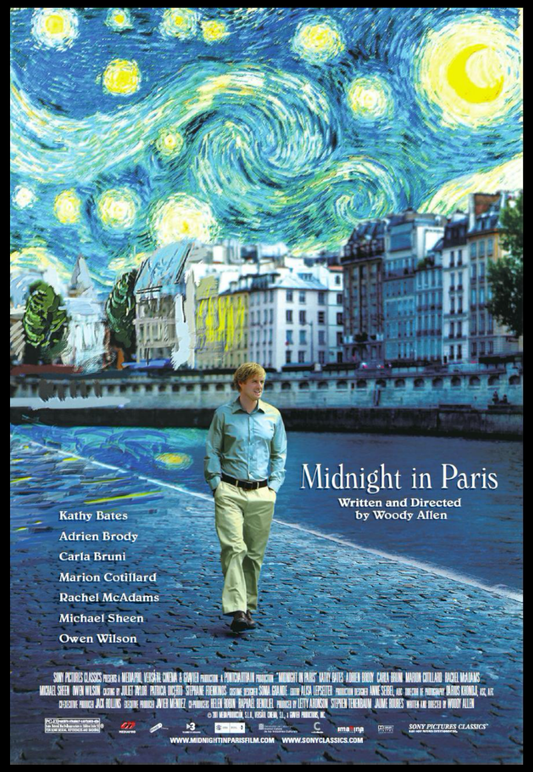
“A movie is better when there are more explosions”

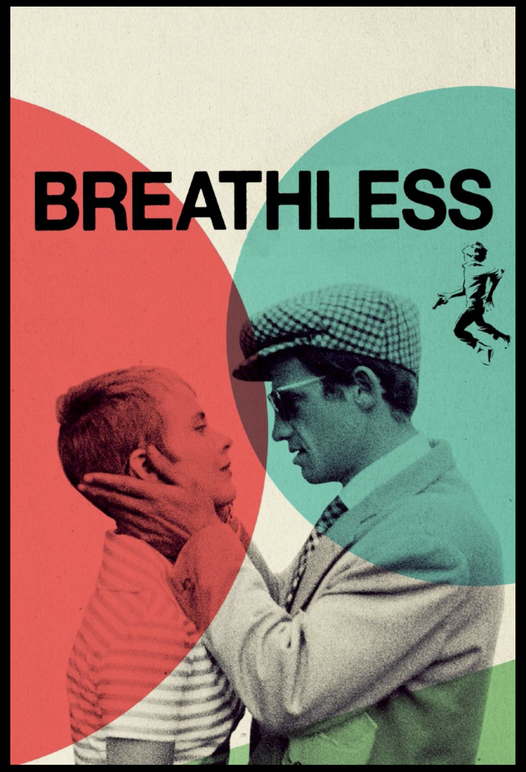


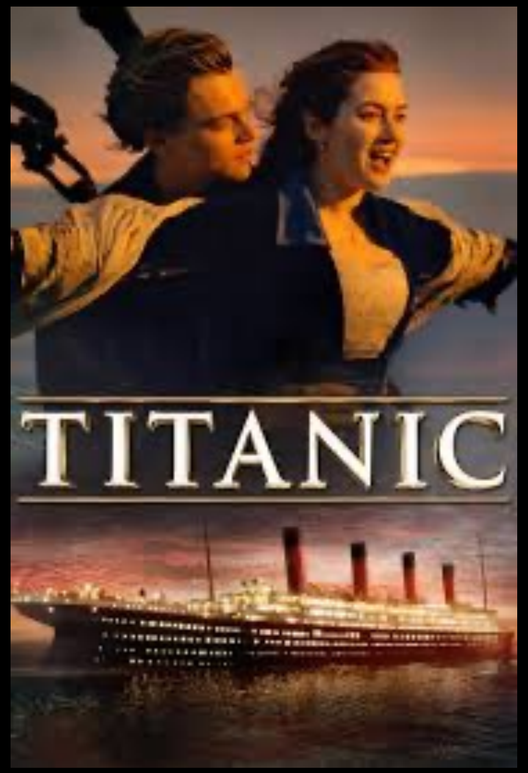


Regression Implies a Model of Reality

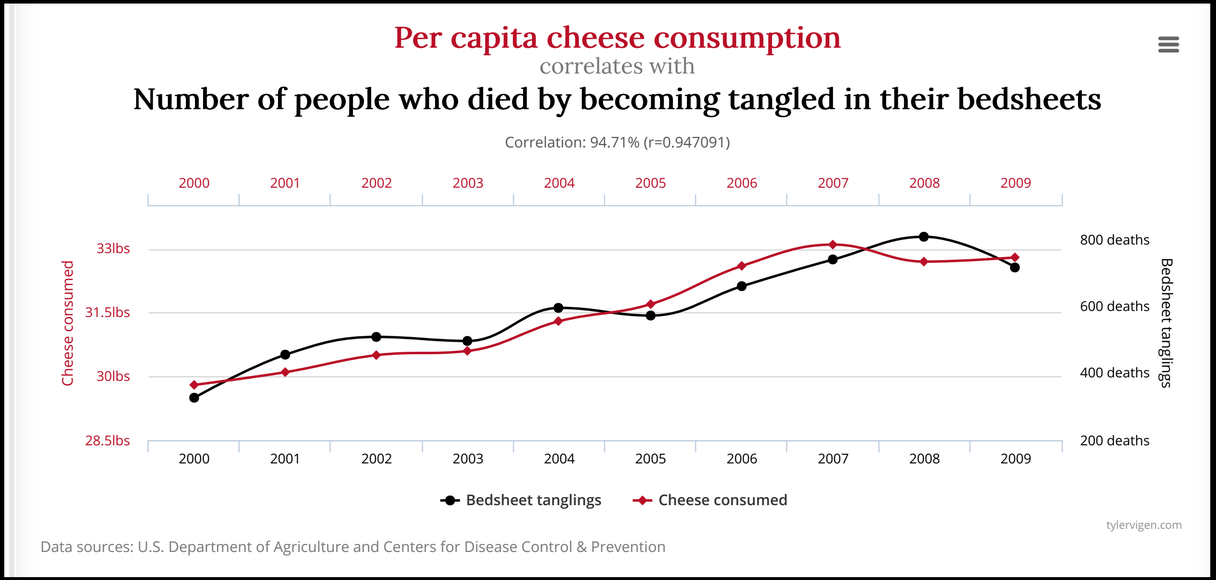
“A movie is better when there are more explosions”

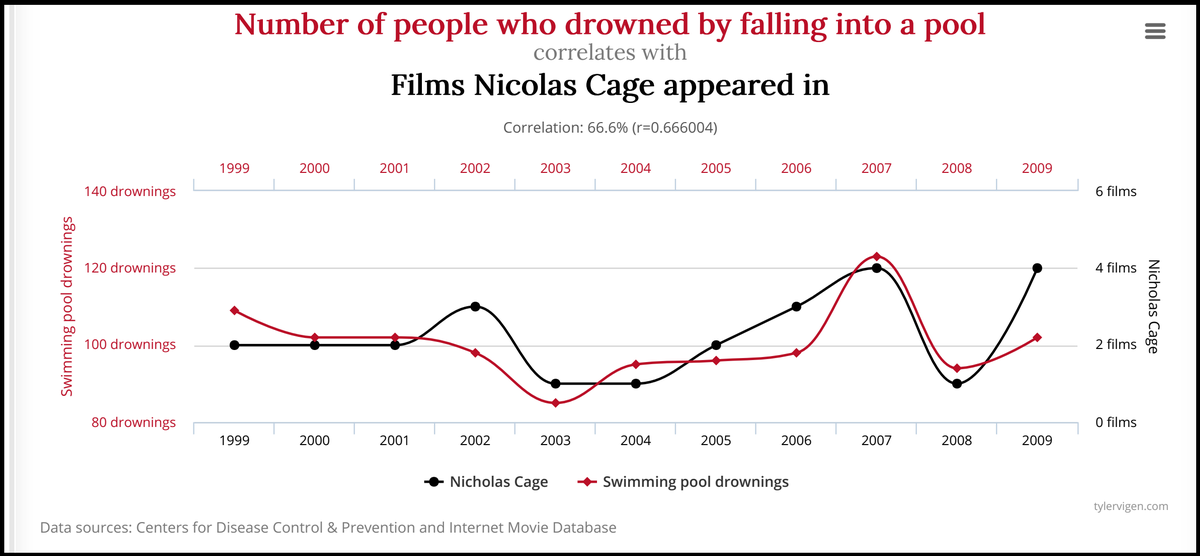






Models can be (and often are) Spurious





Models can be “Indirect”

The NEW ENGLAND JOURNAL of MEDICINE

OCCASIONAL NOTES

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

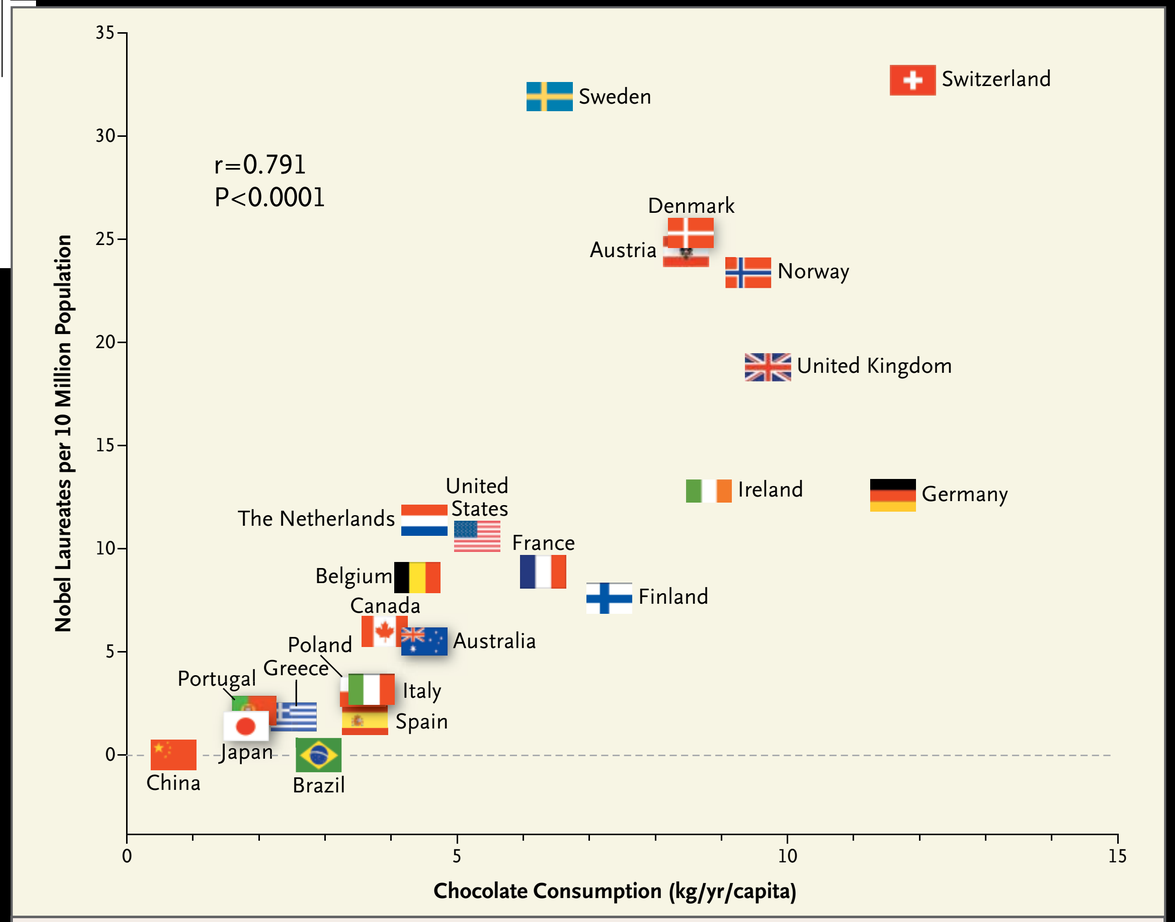
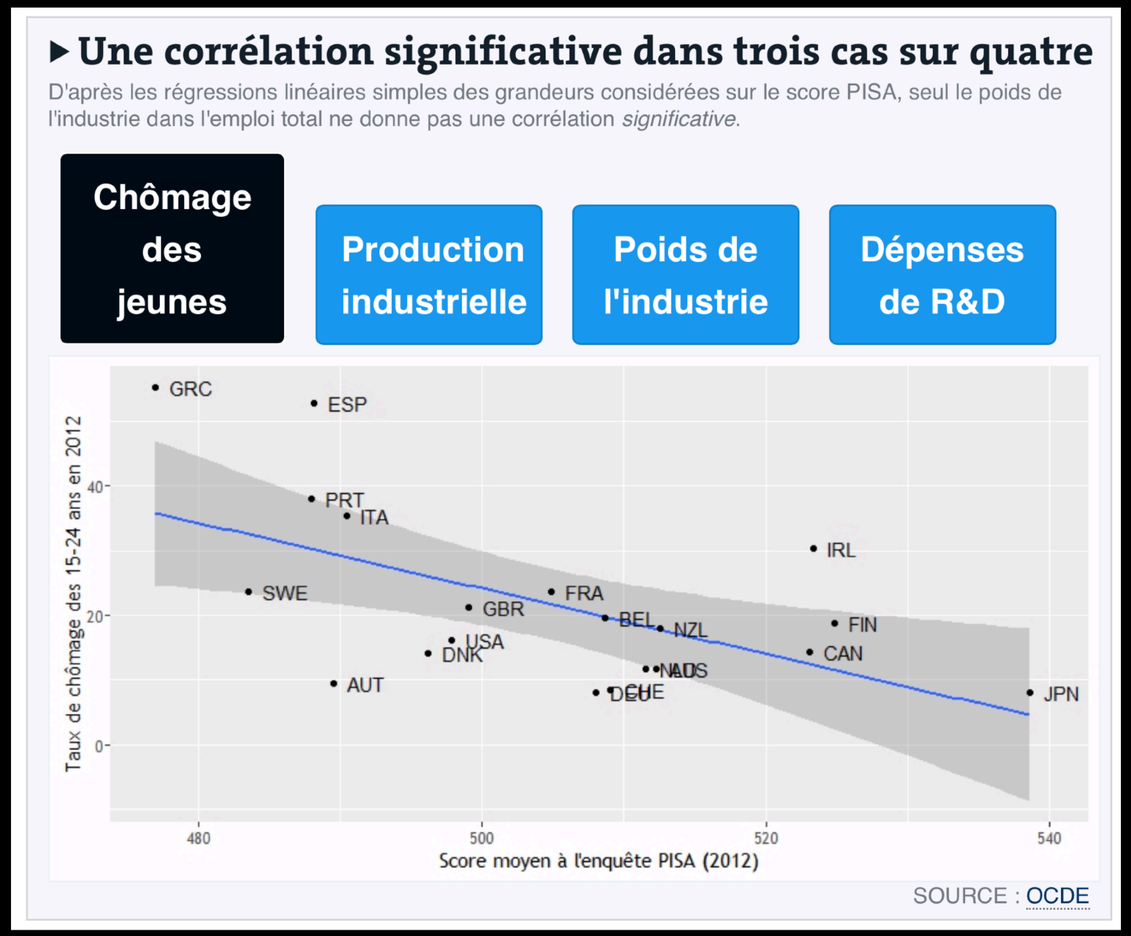


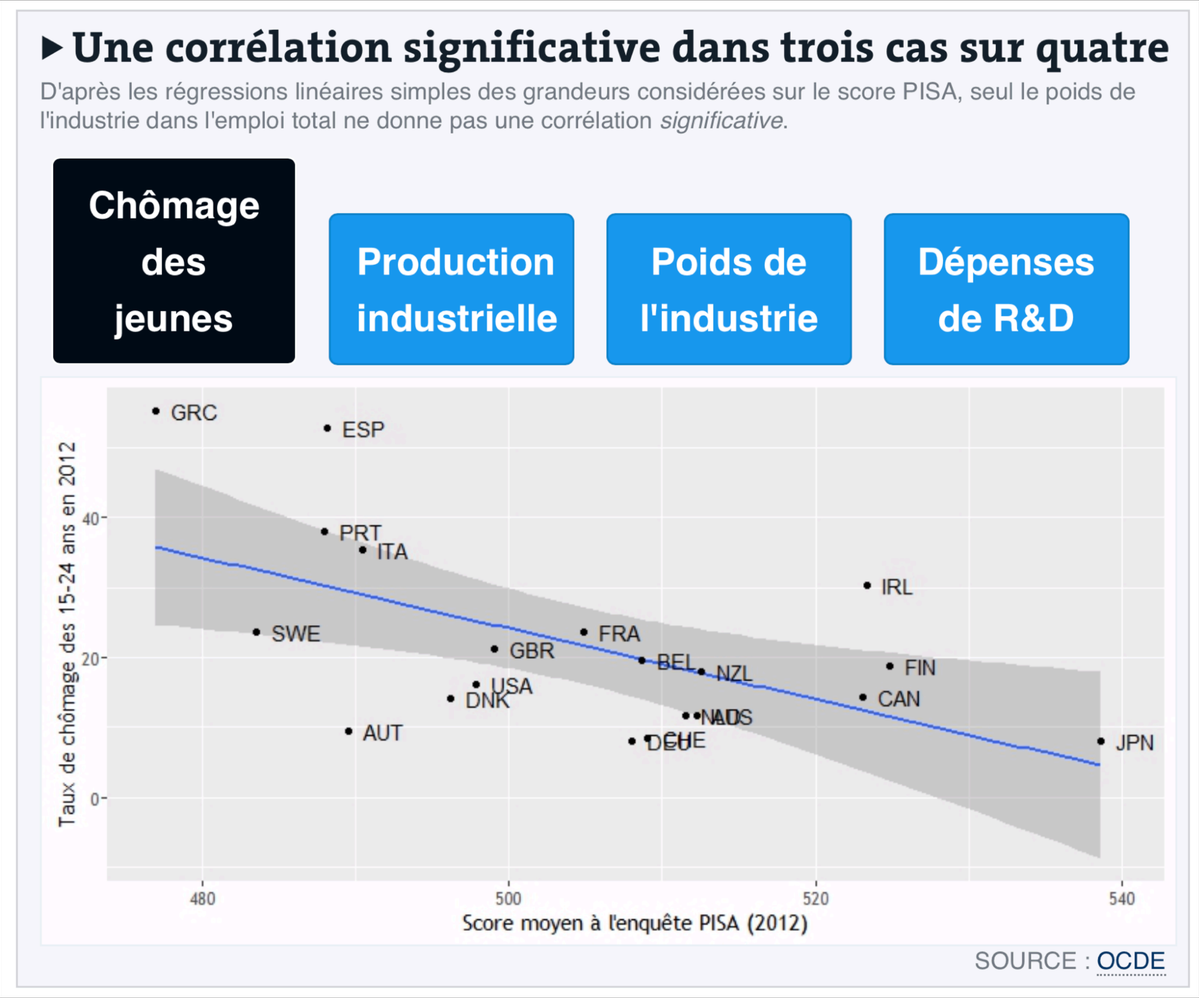
Figure 1. Correlation between Countries’ Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

How do we make sure our model is appropriate and correct?

- Visual inspection
- Model Selection
  - Feature selection
  - Model comparison
- Model Validation
  - Statistical tests
  - Residual Analysis



Visual Inspection



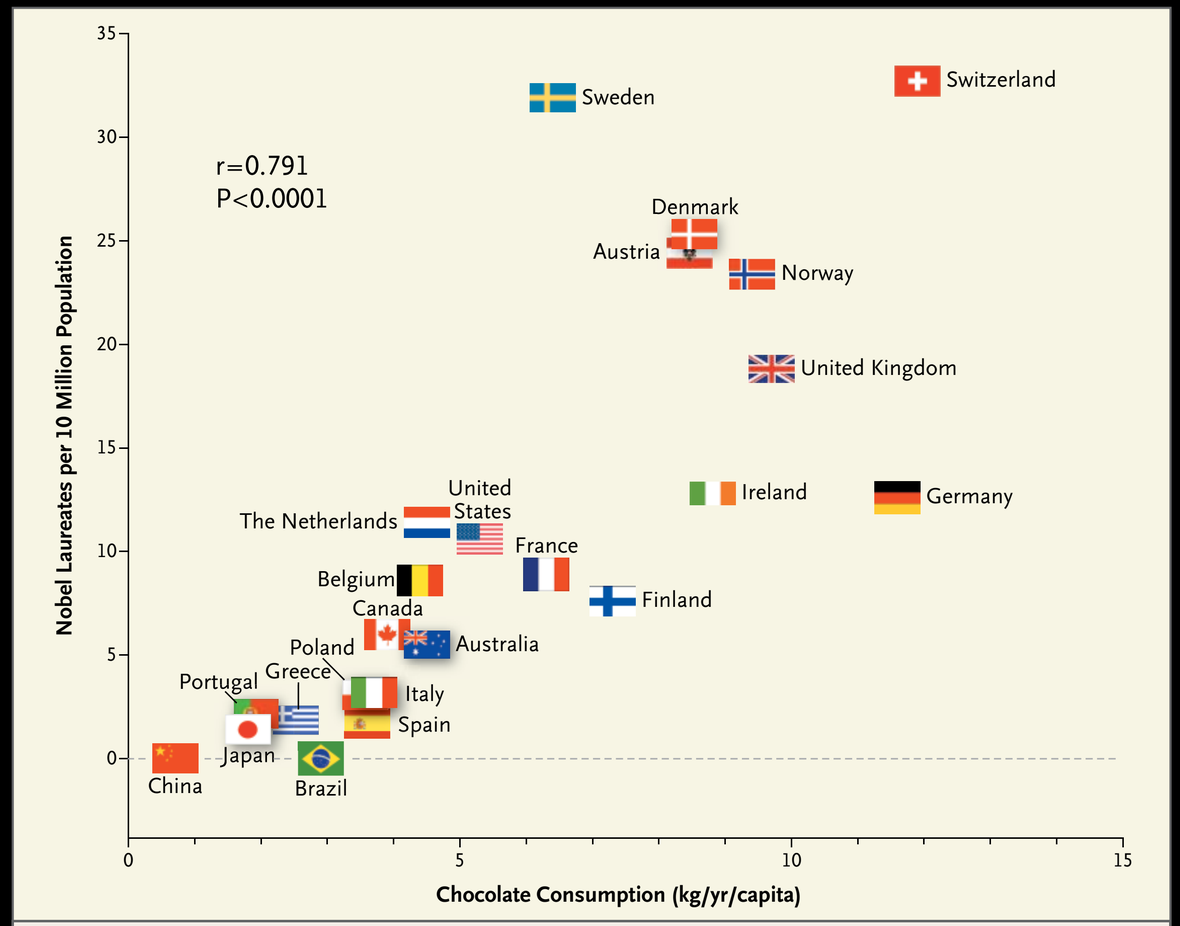
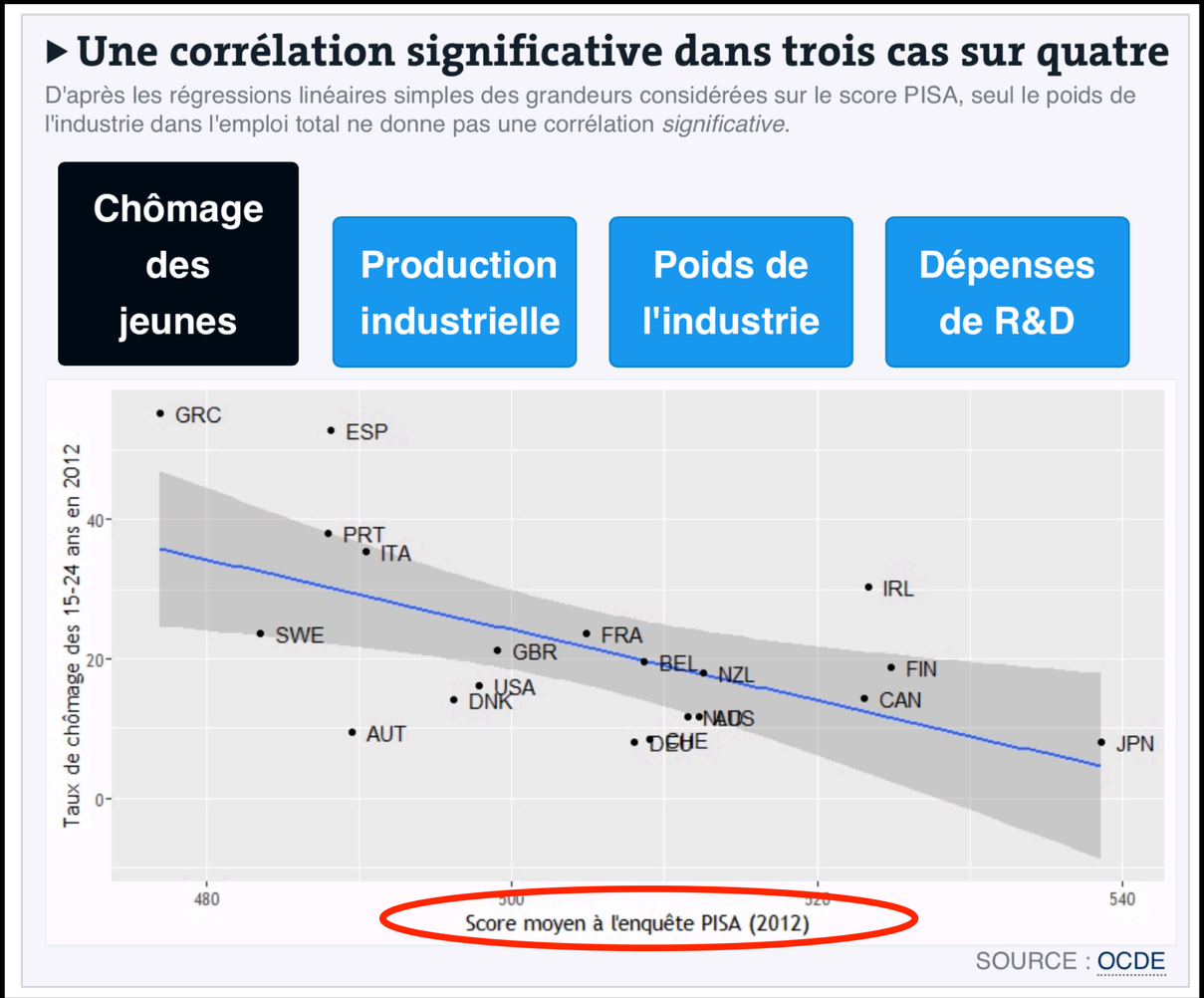


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Model Selection



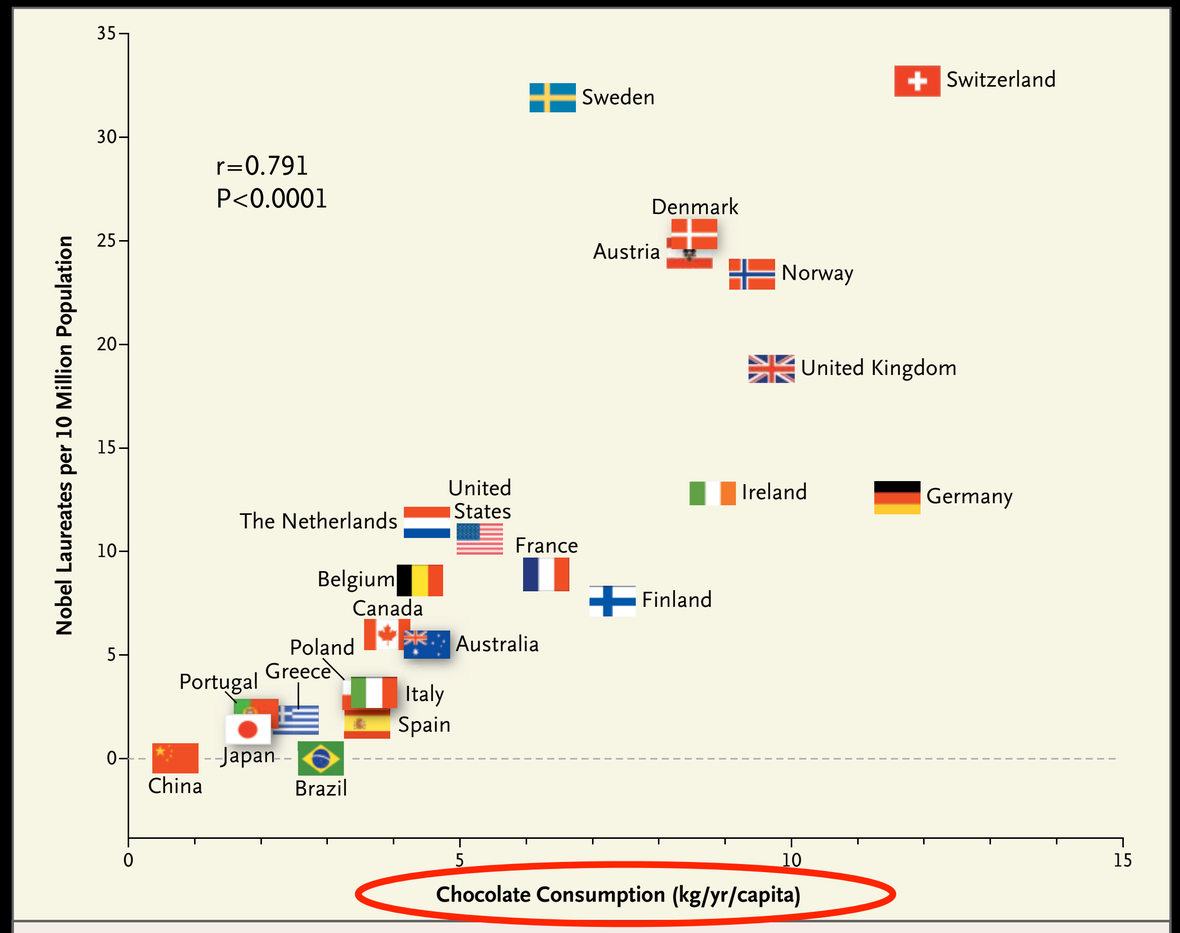
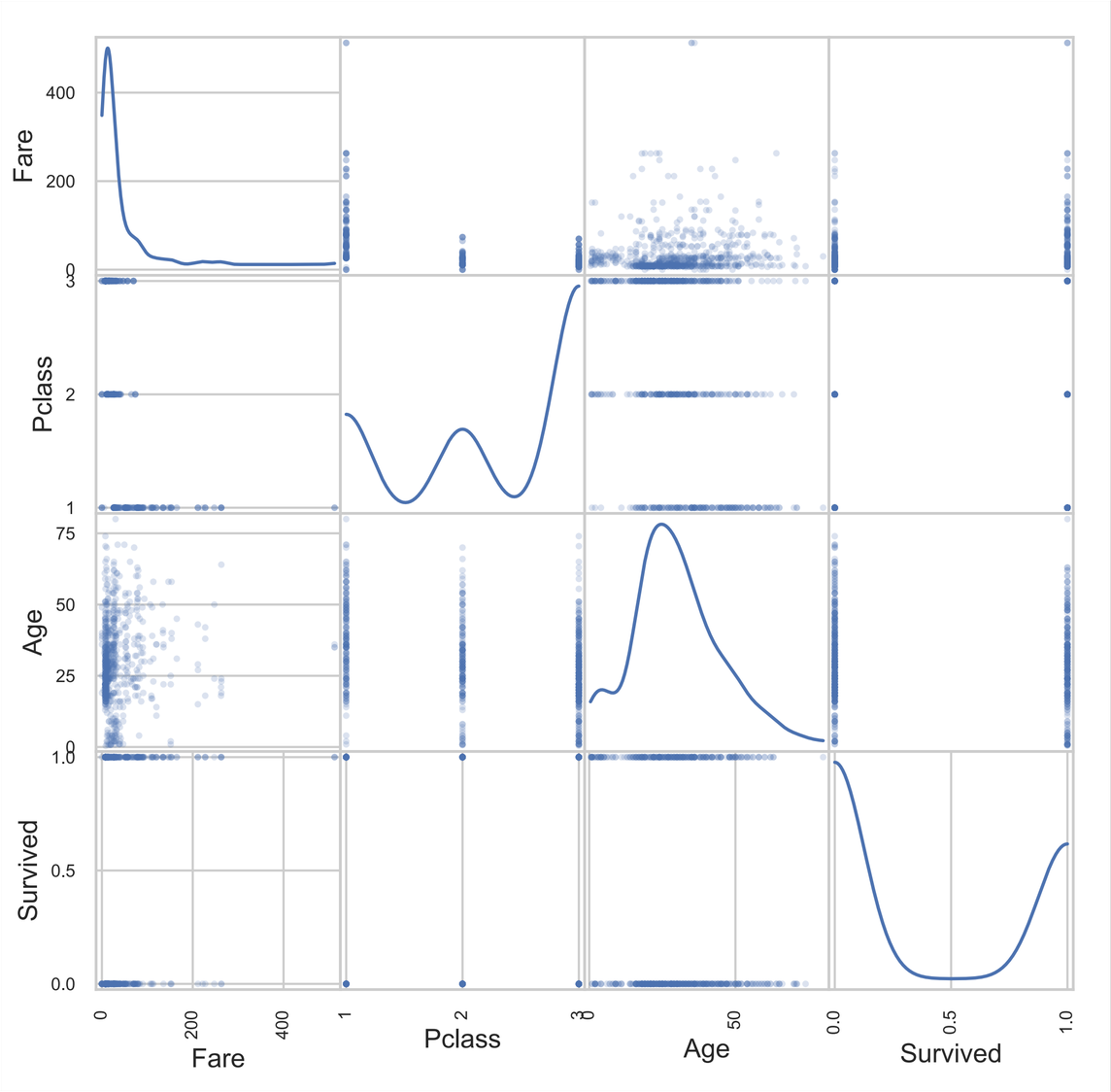


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

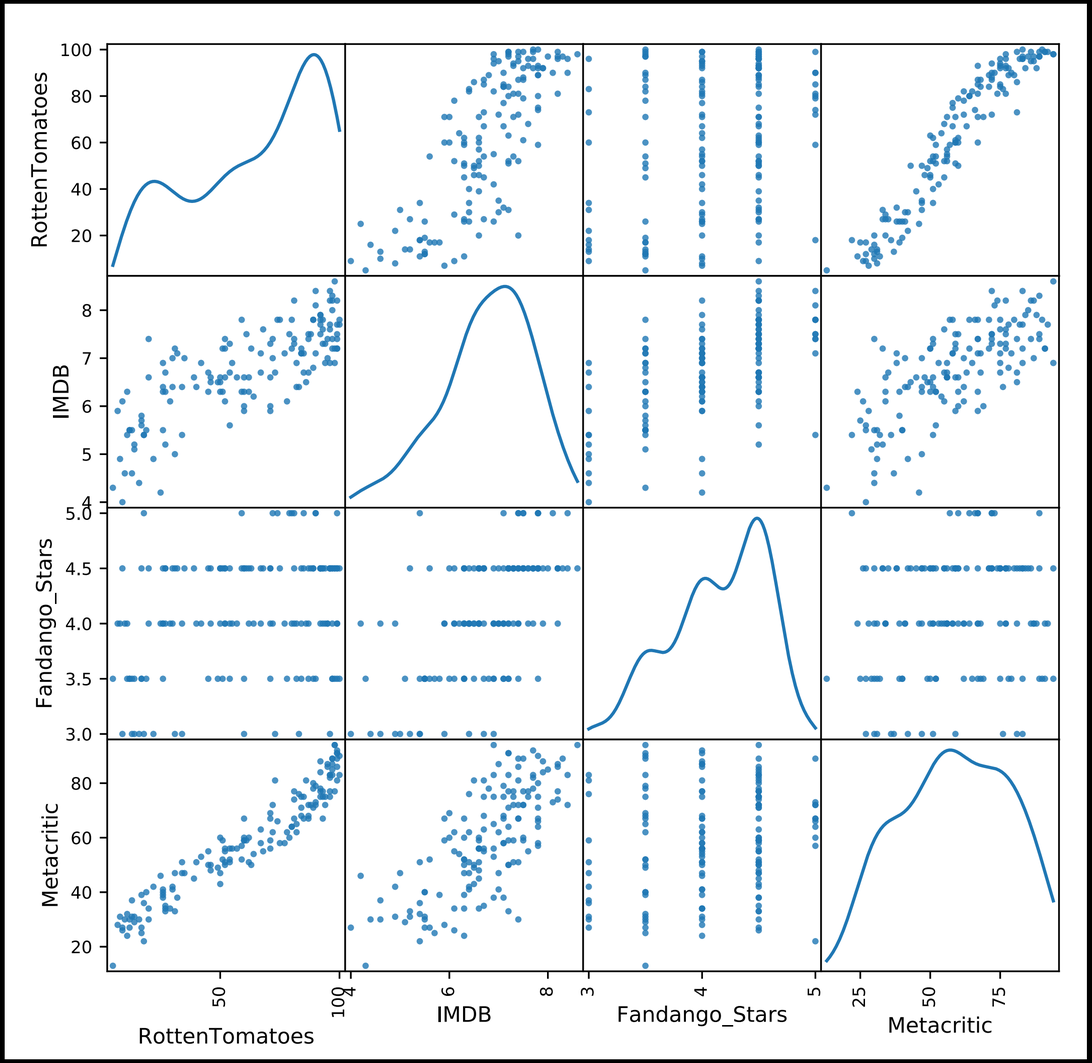
Model Selection - Feature Selection

Who survived in the Titanic?



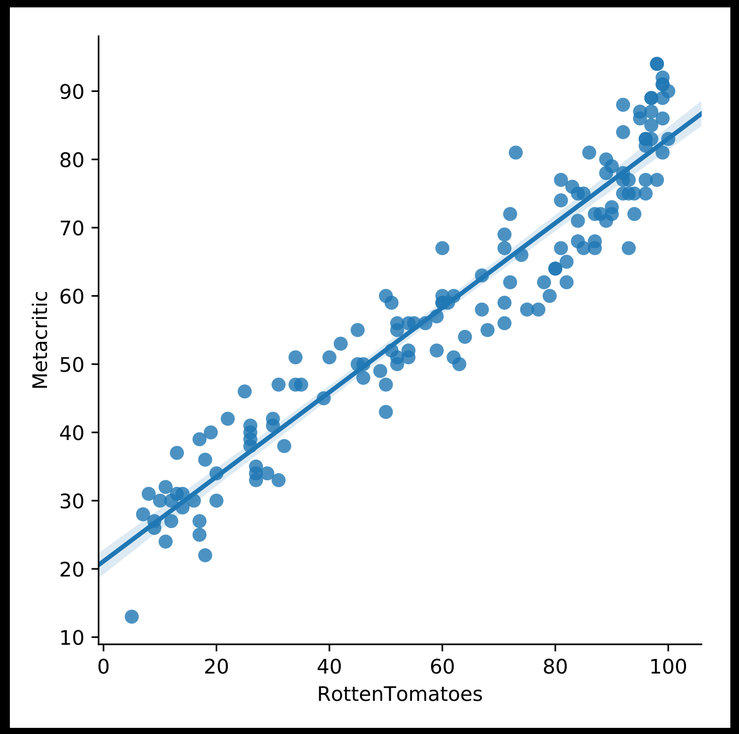
Model Selection - Regression Model

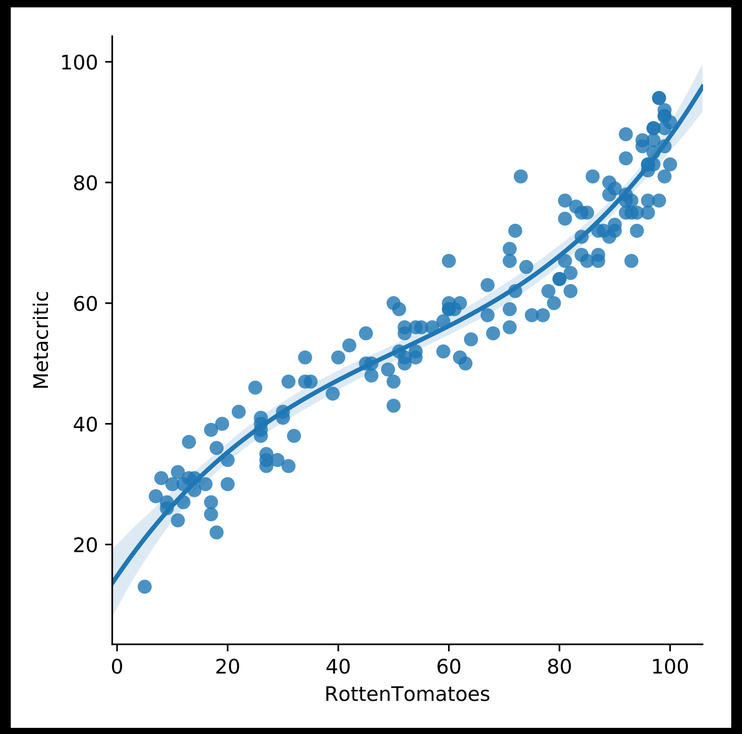
Are movie recommendation systems consistent?

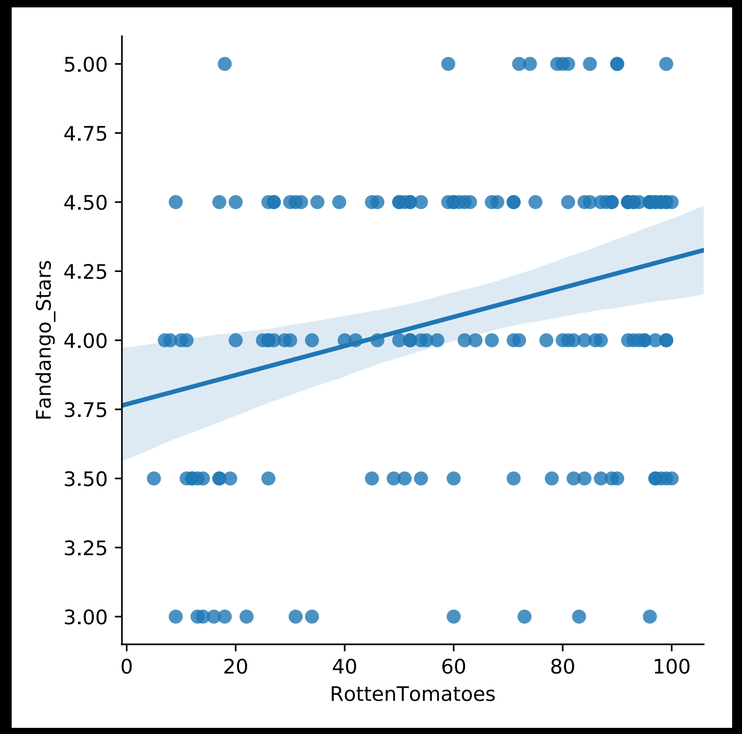


Model Selection - Regression Model

Are movie recommendation systems consistent?

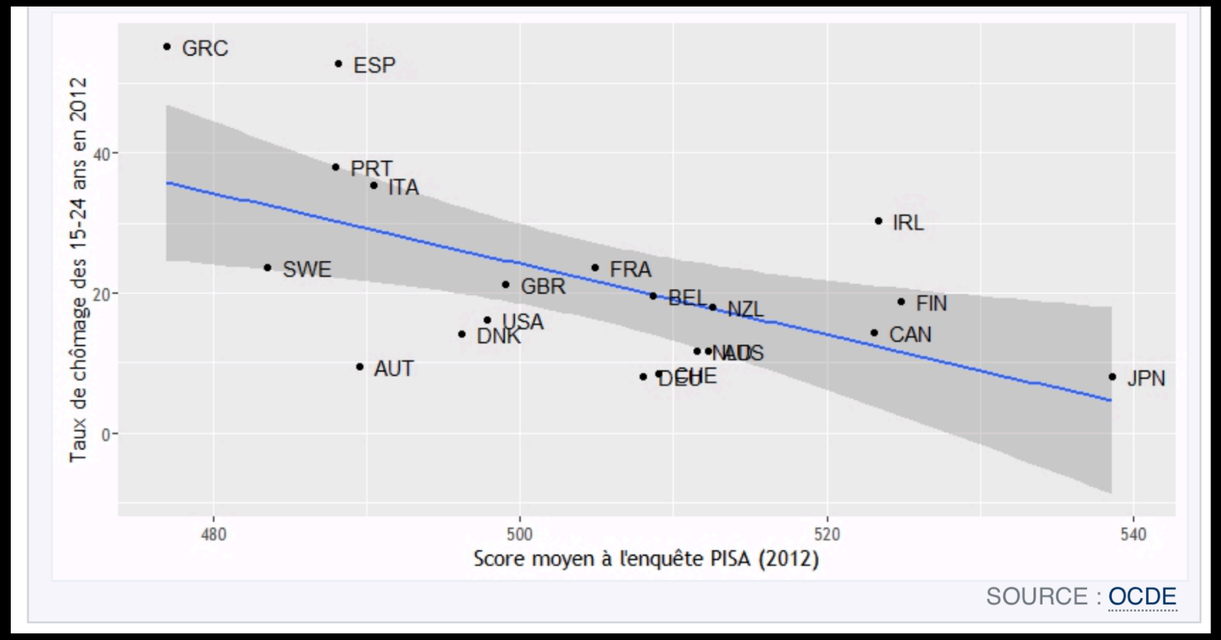




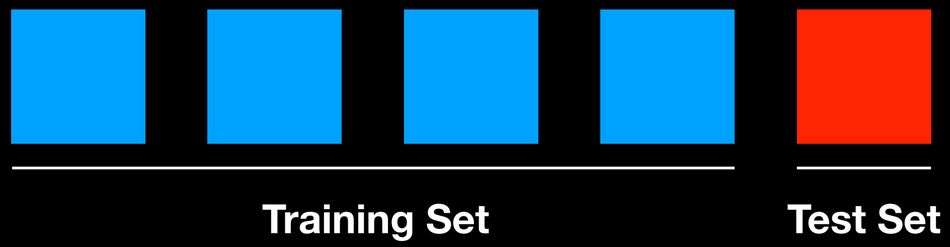


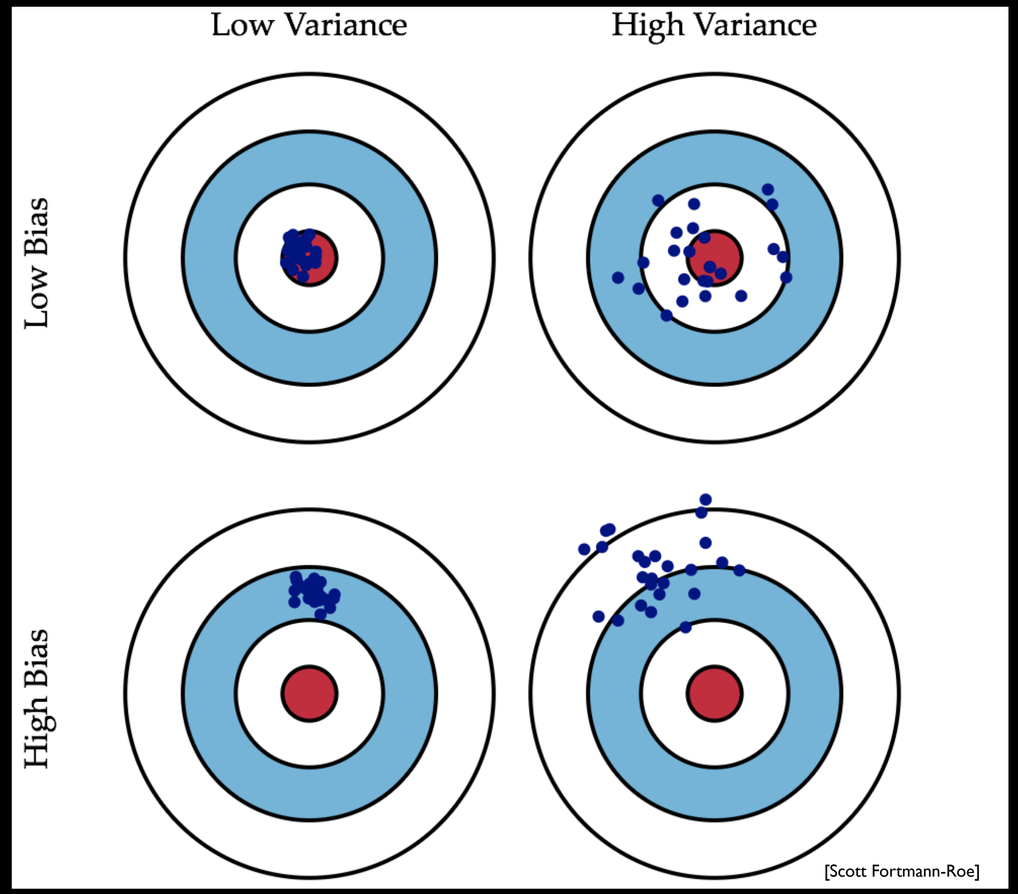
Model Selection - How to Evaluate Models

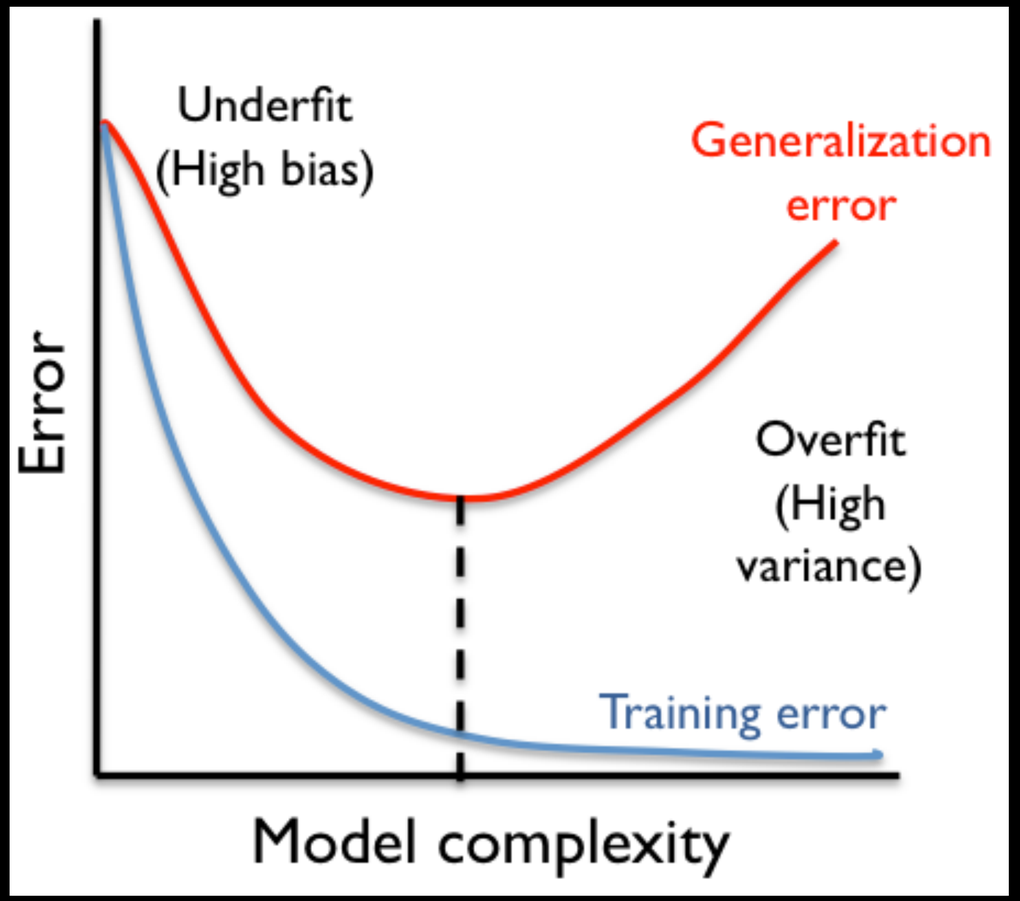
Analysing Predictive Power



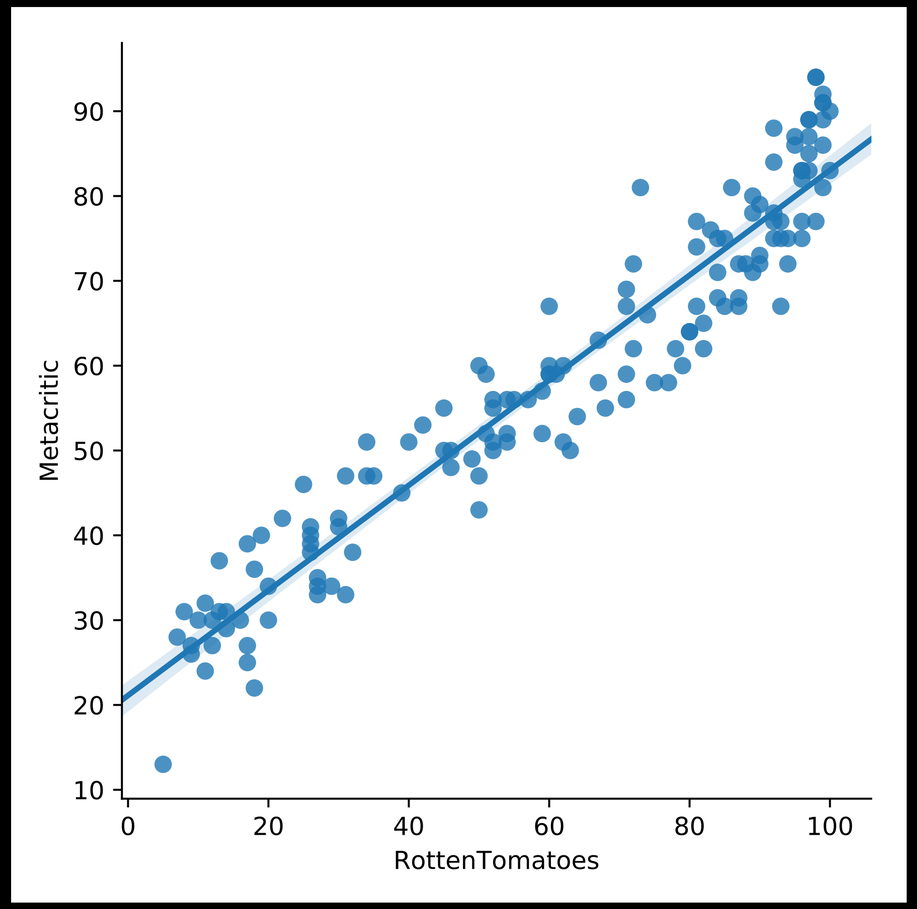
Model Selection - What should we look for?







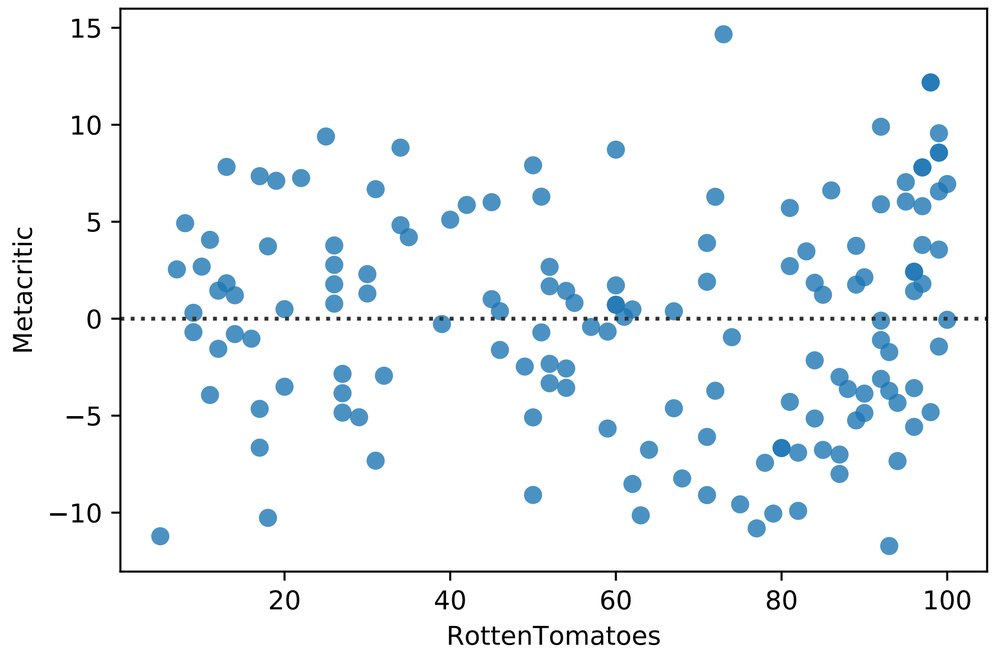
Model Validation - Statistical Tests



Dep. Variable:	Metacritic	**R-squared:**	0.917
Model:	OLS	**Adj. R-squared:**	0.916
Method:	Least Squares	**F-statistic:**	1581.
Date:	Wed, 17 Oct 2018	**Prob (F-statistic):**	1.54e-79
Time:	07:20:37	**Log-Likelihood:**	-459.19
No. Observations:	146	**AIC:**	922.4
Df Residuals:	144	**BIC:**	928.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	21.1210	1.057	19.980	0.000	19.032	23.210
RottenTomatoes	0.6194	0.016	39.766	0.000	0.589	0.650

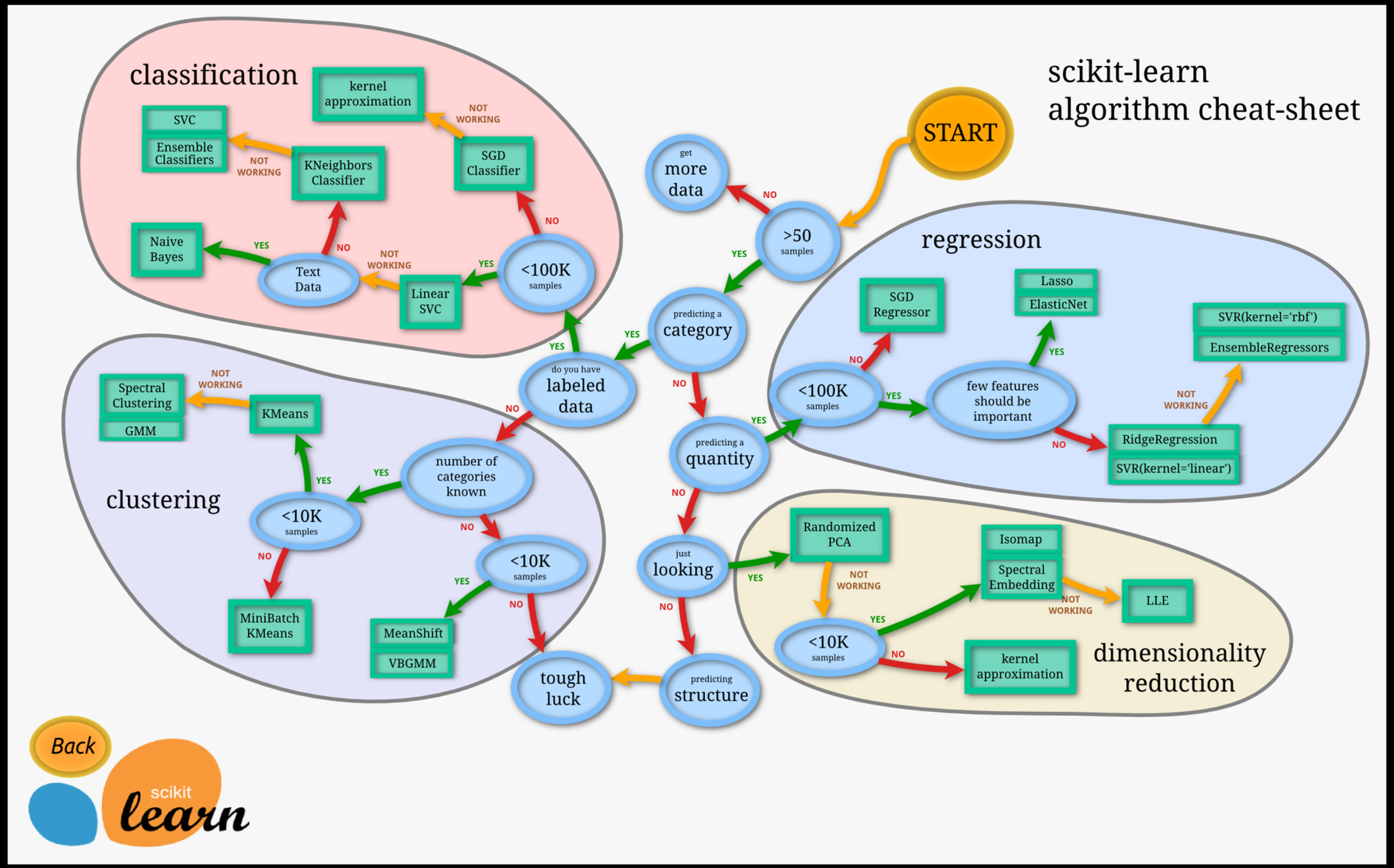
Model Validation - Residual Analysis



Dep. Variable:	Metacritic	**R-squared:**	0.917
Model:	OLS	**Adj. R-squared:**	0.916
Method:	Least Squares	**F-statistic:**	1581.
Date:	Wed, 17 Oct 2018	**Prob (F-statistic):**	1.54e-79
Time:	07:20:37	**Log-Likelihood:**	-459.19
No. Observations:	146	**AIC:**	922.4
Df Residuals:	144	**BIC:**	928.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	21.1210	1.057	19.980	0.000	19.032	23.210
RottenTomatoes	0.6194	0.016	39.766	0.000	0.589	0.650

Where are we in the Data Science Landscape?



Thanks!

So, now we need to separate signal from noise

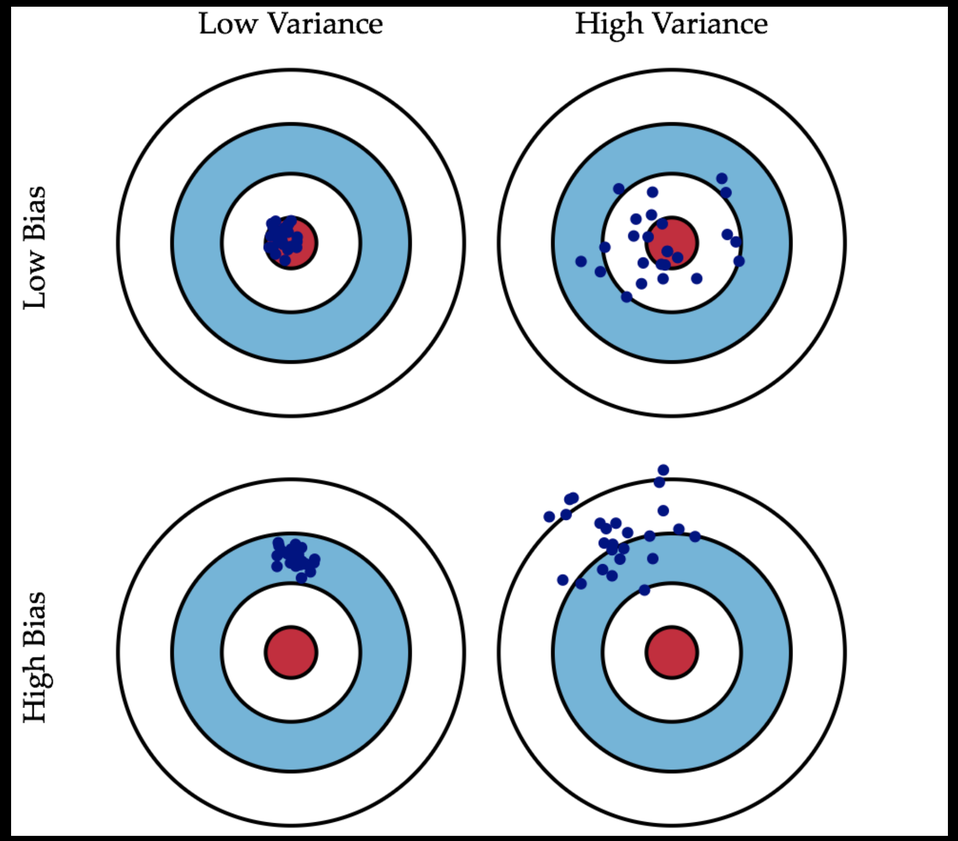
Regression model

$$Y = f(X, \epsilon) = f(X) + \epsilon, \quad \epsilon \sim \mathcal{G}(0, \sigma^2)$$

Bayesian model

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Y	Measured signal
X	Hidden parameter
$f(\ldots)$	Function model
ϵ	Measurement noise



[Scott Fortmann-Roe]

What do we do about Variance?

Regression model

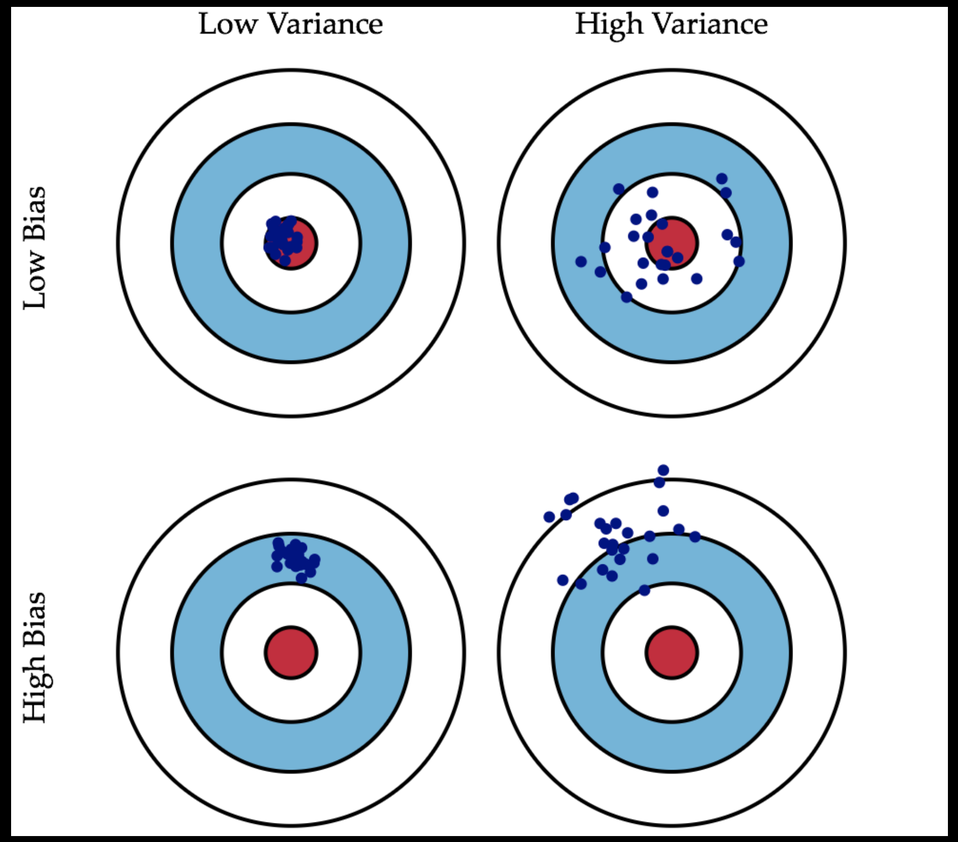
$$Y = f(X, \epsilon) = f(X) + \epsilon, \quad \epsilon \sim \mathcal{G}(0, \sigma^2)$$

$\hat{X}$ is estimated from the measured signal, then

$$\hat{Y} = f(\hat{X})$$

Error Analysis

$$\begin{aligned}
MSE &= E[(Y - \hat{Y})^2] \\
&= E[(f(X) + \epsilon - f(\hat{X}))^2] \\
&= E[(f(X) - f(\hat{X}))^2] + 2E[(f(X) - f(\hat{X}))]E[\epsilon] + E[\epsilon^2] \\
&= E[(f(X) - f(\hat{X}))^2] + \sigma^2 \\
&= (f(X) - E[f(\hat{X})])^2 + E[(f(\hat{X}) - E[f(\hat{X})])^2] + \sigma^2 \\
&= \text{Bias}^2 + \text{Variance} + \text{Error}
\end{aligned}$$



Y Measured signal

X Hidden parameter

$f(\ldots)$ Function model

ϵ Measurement noise

What do we do about Variance?

Regression model

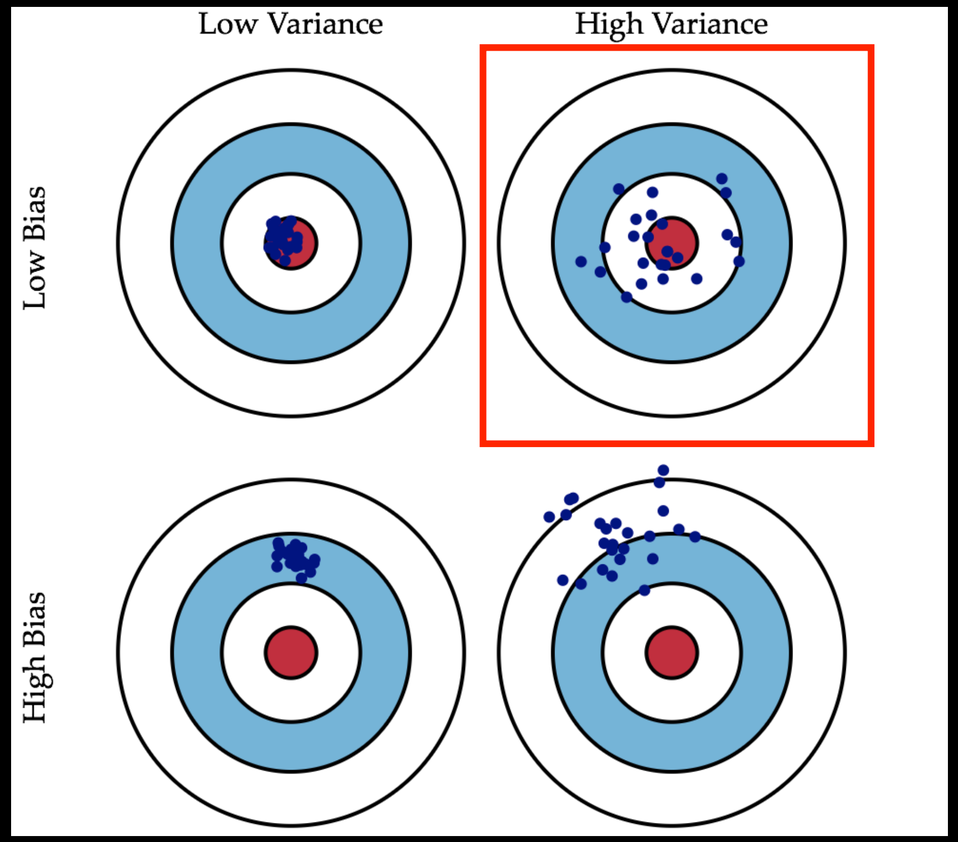
$$Y = f(X, \epsilon) = f(X) + \epsilon, \quad \epsilon \sim \mathcal{G}(0, \sigma^2)$$

$\hat{X}$ is estimated from the measured signal, then

$$\hat{Y} = f(\hat{X})$$

Error Analysis

$$\begin{aligned}
MSE &= E[(Y - \hat{Y})^2] \\
&= E[(f(X) + \epsilon - f(\hat{X}))^2] \\
&= E[(f(X) - f(\hat{X}))^2] + 2E[(f(X) - f(\hat{X}))]E[\epsilon] + E[\epsilon^2] \\
&= E[(f(X) - f(\hat{X}))^2] + \sigma^2 \\
&= (f(X) - E[f(\hat{X})])^2 + E[(f(\hat{X}) - E[f(\hat{X})])^2] + \sigma^2 \\
&= \text{Bias}^2 + \boxed{\text{Variance}} + \text{Error}
\end{aligned}$$



Y	Measured signal
X	Hidden parameter
$f(\ldots)$	Function model
ϵ	Measurement noise

Bayesian Equivalence

Regression model

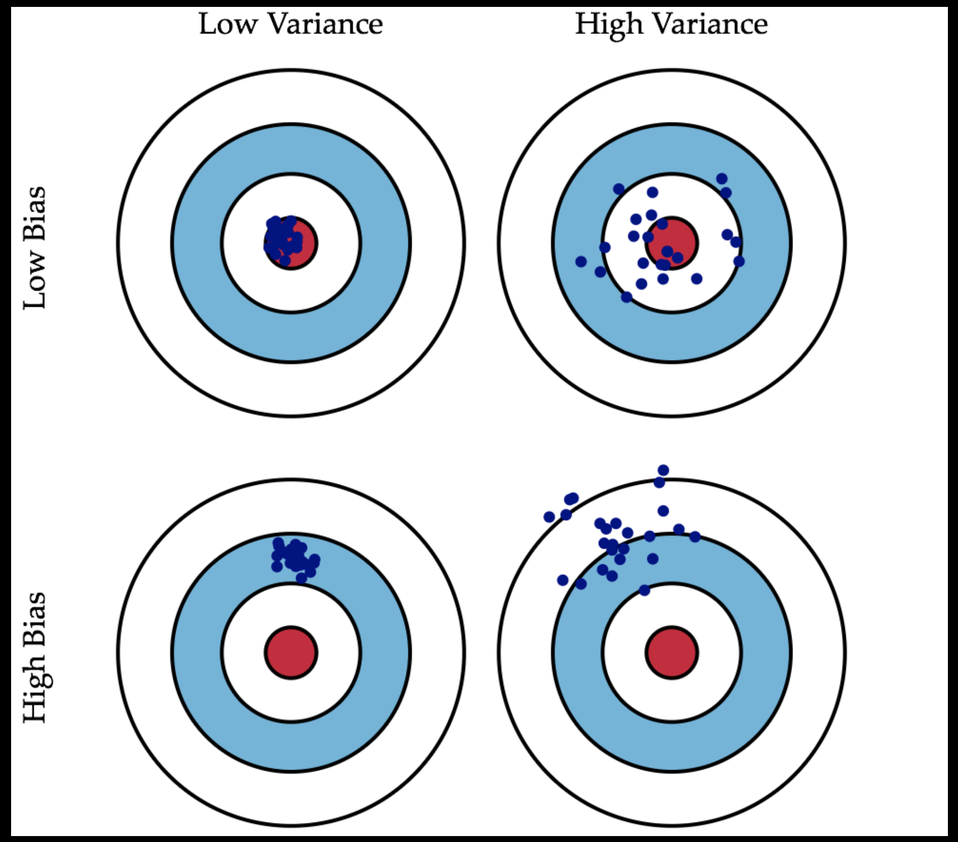
$$Y = f(X, \epsilon) = f(X) + \epsilon, \quad \epsilon \sim \mathcal{G}(0, \sigma^2)$$

$\hat{X}$ is estimated from the measured signal, then

$$\hat{Y} = f(\hat{X})$$

Bayesian Equivalence — No Priors

$$\begin{aligned}
-\log P(X|Y) &\propto -\log P(Y|X) - \log P(X) + \log P(Y) \\
&\propto \|Y - f(X, \epsilon)\|^2 - \log P(X) + \log P(Y) \\
&\propto \|Y - f(X, \epsilon)\|^2 \\
P(Y|X) &\propto \exp(-\|Y - f(X, \epsilon)\|^2) \\
Y|X &\sim \mathcal{N}(f(X), \sigma^2)
\end{aligned}$$



Y	Measured signal
X	Hidden parameter
$f(\ldots)$	Function model
ϵ	Measurement noise

Bayesian Equivalence

Regression model

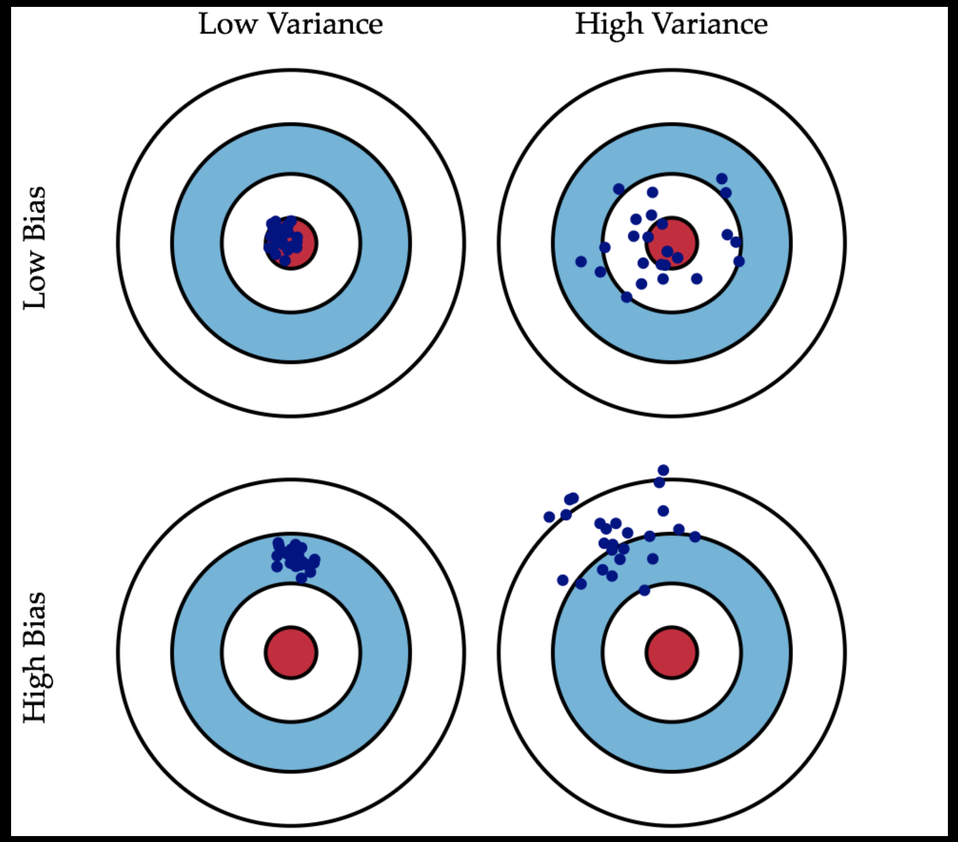
$$Y = f(X, \epsilon) = f(X) + \epsilon, \quad \epsilon \sim \mathcal{G}(0, \sigma^2)$$

$\hat{X}$ is estimated from the measured signal, then

$$\hat{Y} = f(\hat{X})$$

Bayesian Equivalence

$$\begin{aligned} -\log P(X|Y) &\propto -\log P(Y|X) - \log P(X) + \log P(Y) \\ &\propto \|Y - f(X, \epsilon)\|^2 - \log P(X) + \log P(Y) \\ &\propto \|Y - f(X, \epsilon)\|^2 - \log P(X) \\ P(X|Y) &\propto \exp(-\|Y - f(X, \epsilon)\|^2) P(X) \end{aligned}$$



Y	Measured signal
X	Hidden parameter
$f(\ldots)$	Function model
ϵ	Measurement noise

Ridge Regression

Regression model

$$Y = f(X, \epsilon) = f(X) + \epsilon, \quad \epsilon \sim \mathcal{G}(0, \sigma^2)$$

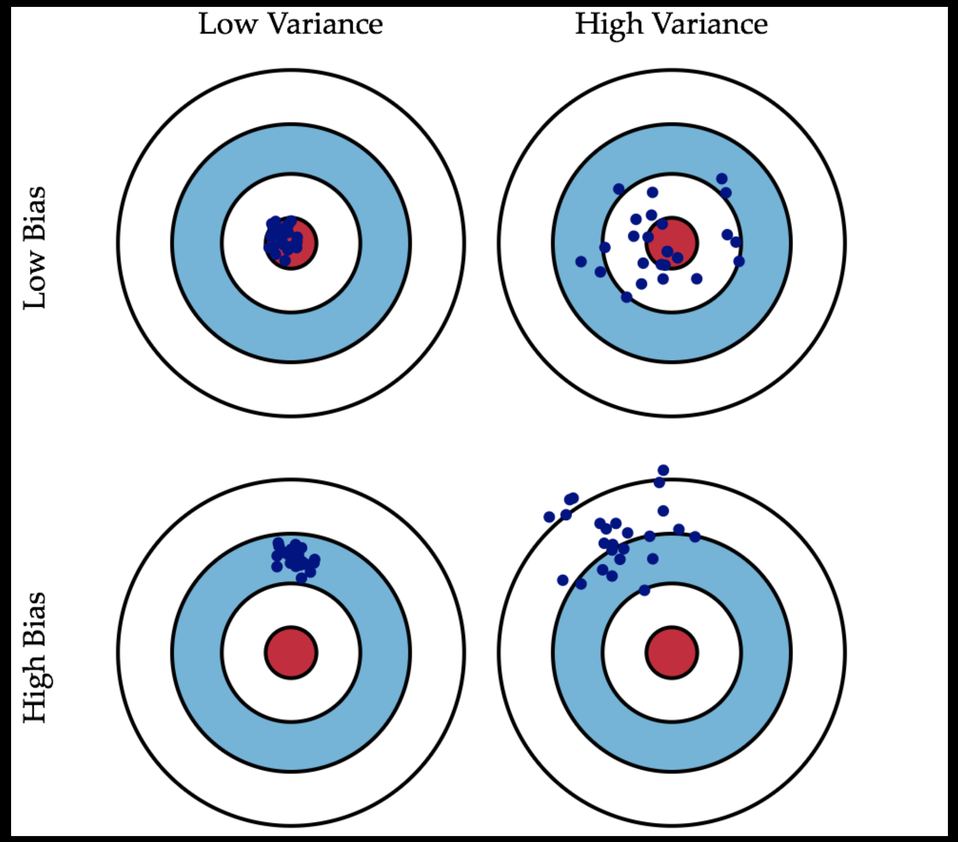
$\hat{X}$ is estimated from the measured signal, then

$$\hat{Y} = f(\hat{X})$$

Ridge Regression

$$P(X|Y) \propto \exp(-\|Y - f(X, \epsilon)\|^2) \exp(-\lambda \|X\|^2)$$

$$\arg\min_X \|Y - f(X, \epsilon)\|^2 + \lambda \|X\|^2$$



Y Measured signal

X Hidden parameter

$f(\ldots)$ Function model

ϵ Measurement noise

Ridge Regression

Ridge Regression

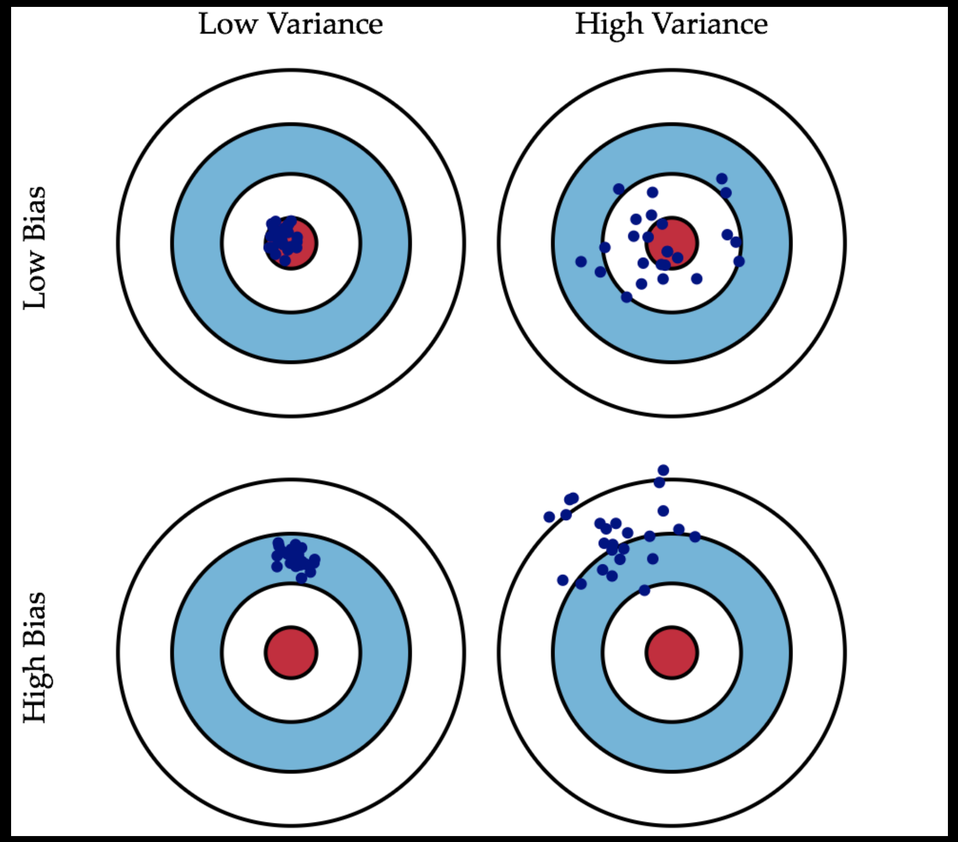
$$P(X|Y) \propto \exp(-\|Y - f(X,\epsilon)\|^2)\exp(-\lambda\|X\|^2)$$

$$\arg\min_X \|Y - f(X,\epsilon)\|^2 + \lambda\|X\|^2$$

Ridge Regression — Linear Problem

$$\arg\min_X \|Y - AX\|^2 + \lambda\|X\|^2$$

$$\hat{X} = (A^T A + \lambda\,\mathrm{Id})^{-1} A^T Y$$



Y Measured signal

X Hidden parameter

$f(\ldots)$ Function model

ϵ Measurement noise

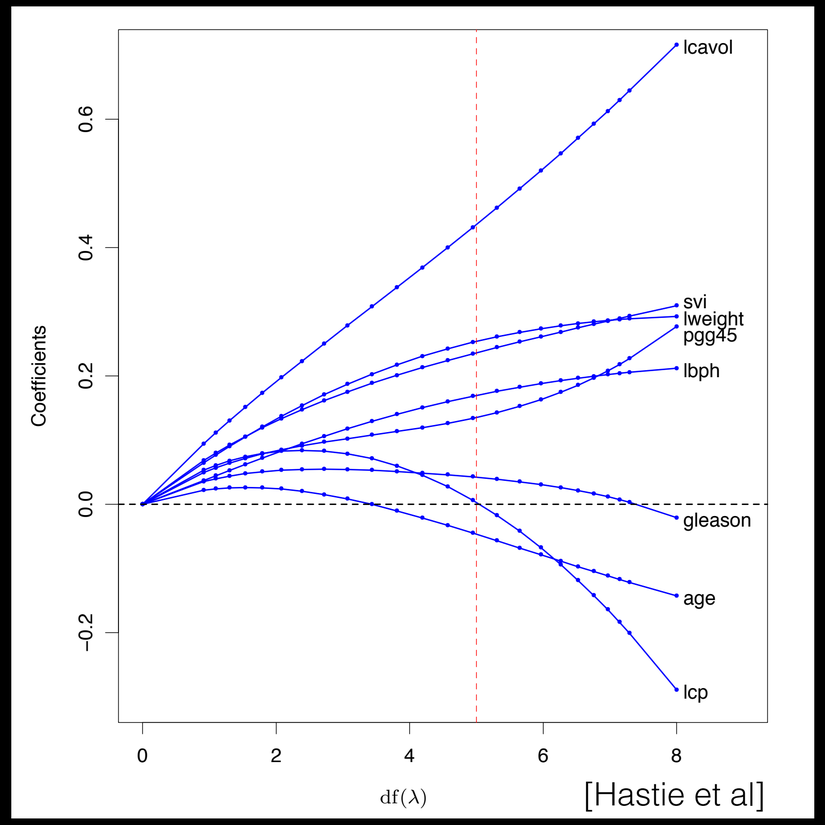
Ridge Regression

Ridge Regression — Linear Problem

$$\arg\min_{X} \|Y - AX\|^2 + \lambda \|X\|^2$$

$$\hat{X} = (A^T A + \lambda \,\mathrm{Id})^{-1} A^T Y$$

$$\mathrm{df}(\lambda) = \mathrm{tr}(A(A^T A + \lambda \,\mathrm{Id})^{-1} A^T)$$



Y Measured signal

X Hidden parameter

$f(\ldots)$ Function model

ϵ Measurement noise

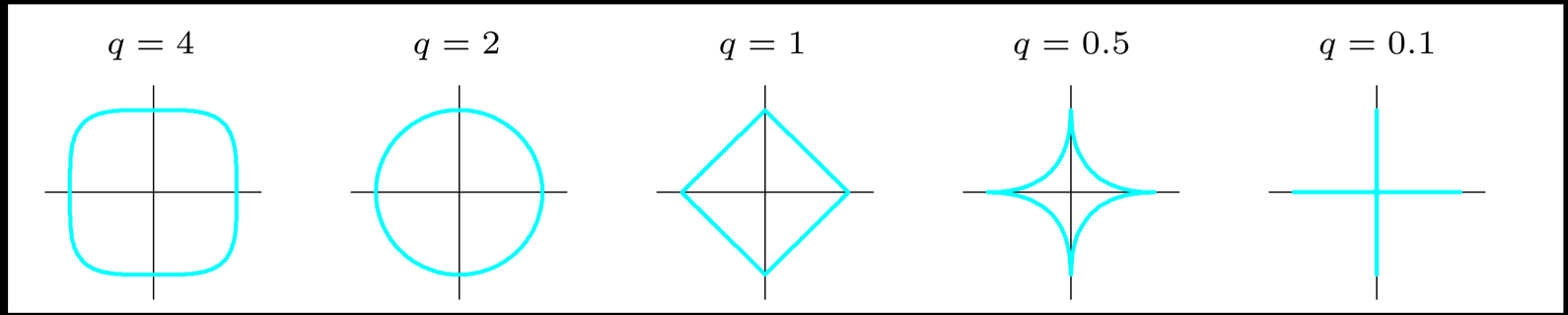
Enforcing Sparsity — The Lasso

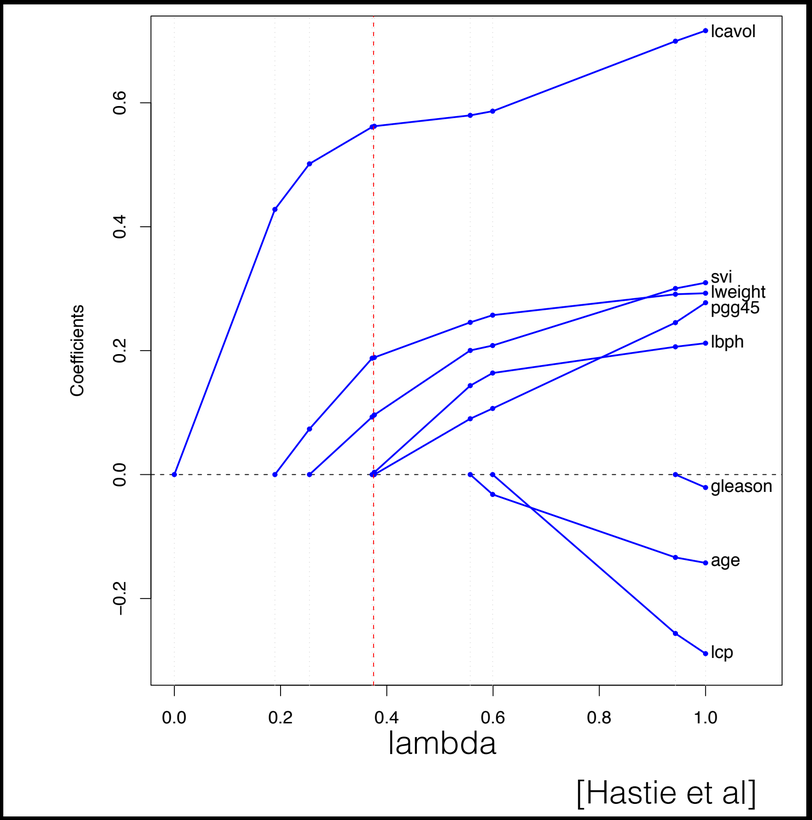
Lasso Regression — Linear Problem

$$\arg\min_{X} \|Y - AX\|^2 + \lambda |X|$$

$$\cancel{\hat{X} = (A^T A + \lambda \mathrm{Id})^{-1} A^T Y}$$

$$|X|^q$$





Y Measured signal

X Hidden parameter

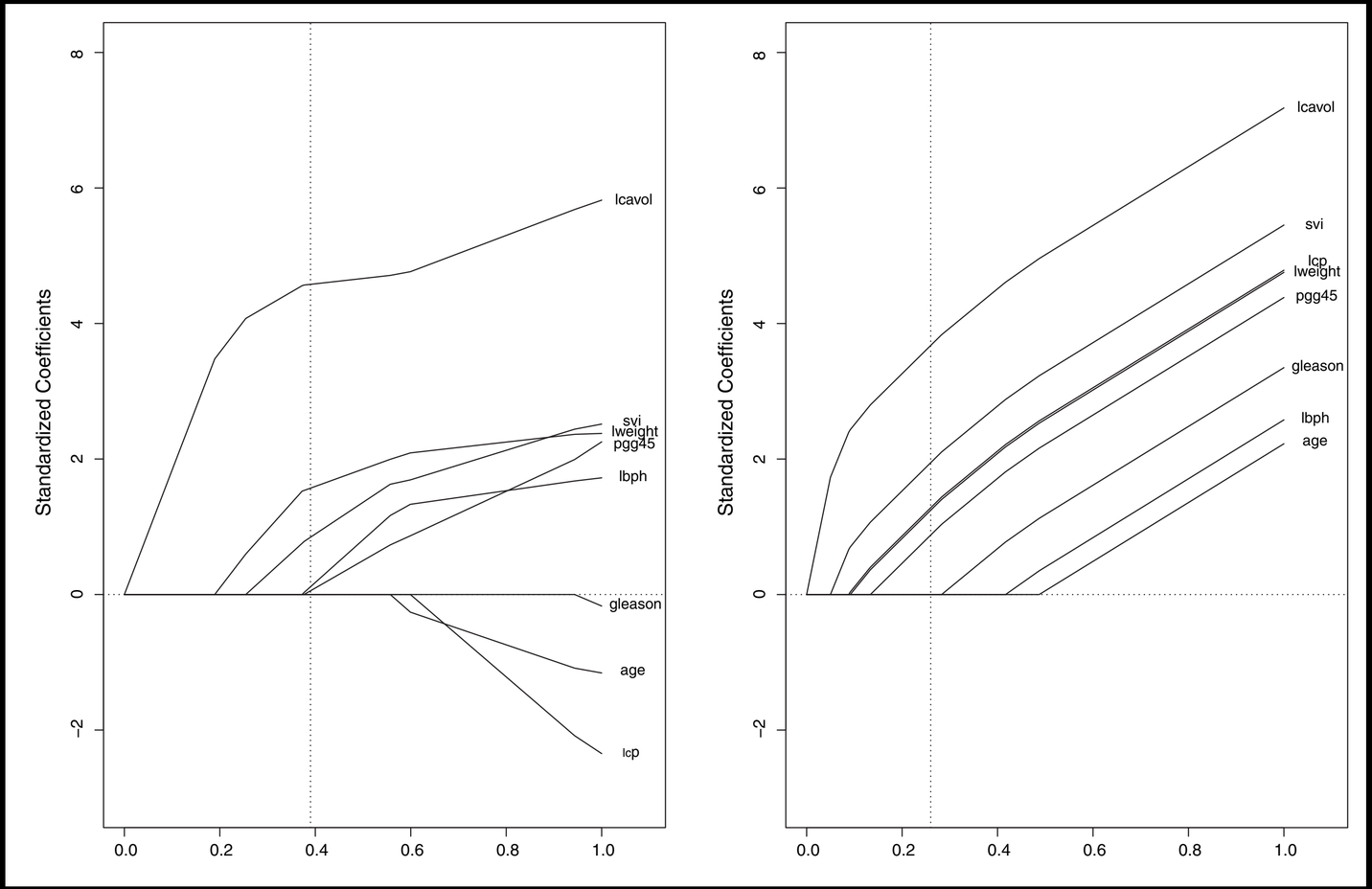
$f(\ldots)$ Function model

ϵ Measurement noise

Enforcing Sparsity and Smoothness — Elastic Net

Elastic Net Regression — Linear Problem

$$\arg\min_{X} \|Y - AX\|^2 + \lambda(\beta|X| + (1-\beta)\|X\|^2)$$



[Zhou et al.]