Explanation v Prediction

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2018
Columbia University

Data challenge #1: a recap

Data challenge #1

a quick review...

86.1% of dead civilians who presumably participated in confrontations with federal armed forces were killed in events of "perfect lethality" where there were only dead and no wounded. [...] Mexico has the terrible situation of having lethality indices of 2.6. The lethality index of the Federal Police is 2.6 dead for every wounded, the Navy's reaches 17.3 dead for every wounded, and the Army's is 9.1 dead for every wounded.

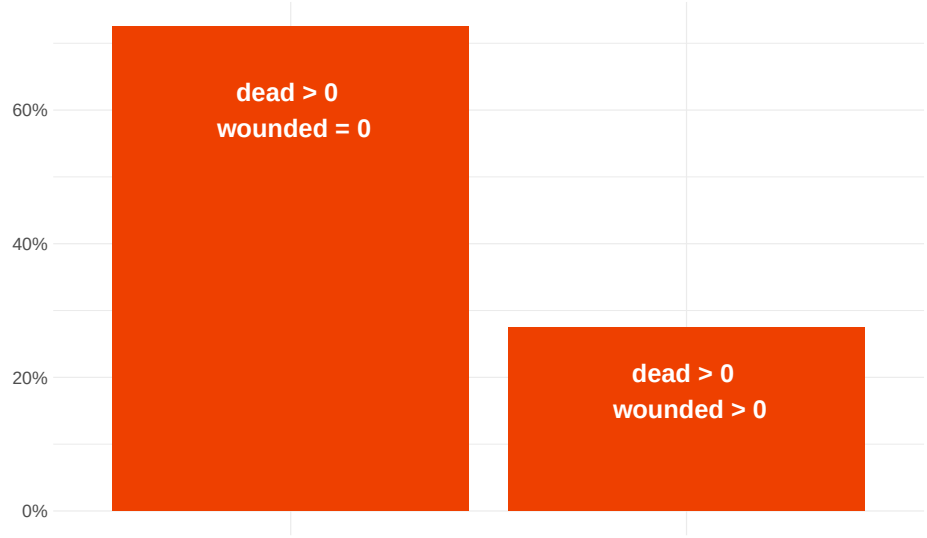
1. **Can you replicate the 86.1% number? the overall lethality ratio? the ratios for the Federal Police, Navy and Army?**

Data challenge #1

a quick review...

First, based on the database, we can compute that:

- **78% of all organized crime deaths happened in events of “perfect lethality”**



- ... we were aiming to find **86.1 %**

Data challenge #1

a quick review...

Second, we need to calculate a lethality indices:

- How do you compute a lethality index?
  - i) the ratio of **total** deaths over **total** wounded among organized crime

$$\frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} w_i}$$

  - ii) the average of the **individual** ratios computed at the event level

$$\frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{w_i}$$

  with d_i dead on event i, and w_i wounded on event i
- what is the **substantive difference?**

Data challenge #1

a quick review...

- And there is also a **numerical difference!**

	overall	**army**	**navy**	**federal police**
index (original)	2.6	9.1	17.3	2.6
index (total)	3.0	5.4	4.6	3.0
index (avg)	0.74	0.63	0.68	0.45

Data challenge #1

a quick review...

- There is a difference because the numbers were quoted from a study unrelated to this data!

Tabla 5. Índice de letalidad de presuntos delincuentes fallecidos sobre presuntos delincuentes heridos

Policía Federal	2.6
Ejército	9.1
Marina	17.3
Policía Federal y Ejército	4.8
Fuerzas de seguridad	7.3

Fuente: Base de datos de enfrentamientos (prensa, enero 2008-mayo 2011).

Figure: Silva et al. (2012)

Data challenge #1

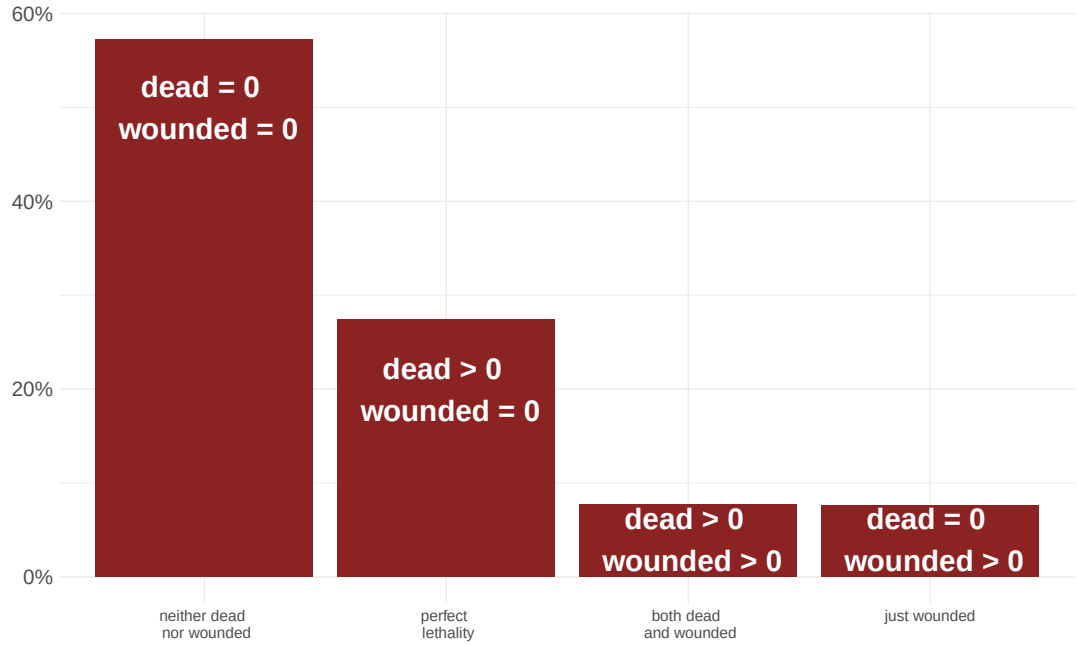
a quick review...

2. The additional questions:
   - Is this the right metric to look at? Why or why not?
   - What is the "lethality index" showing explicitly? What is it not showing? What is the definition assuming?
   - With the same available data, can you think of an alternative way to capture the same construct? Is it "better"?
   - What additional information would you need to better understand the data?
   - What additional information could help you better capture the construct behind the "lethality index"

Data challenge #1

a few steps forward...

- Is there more in the data that we may be missing?

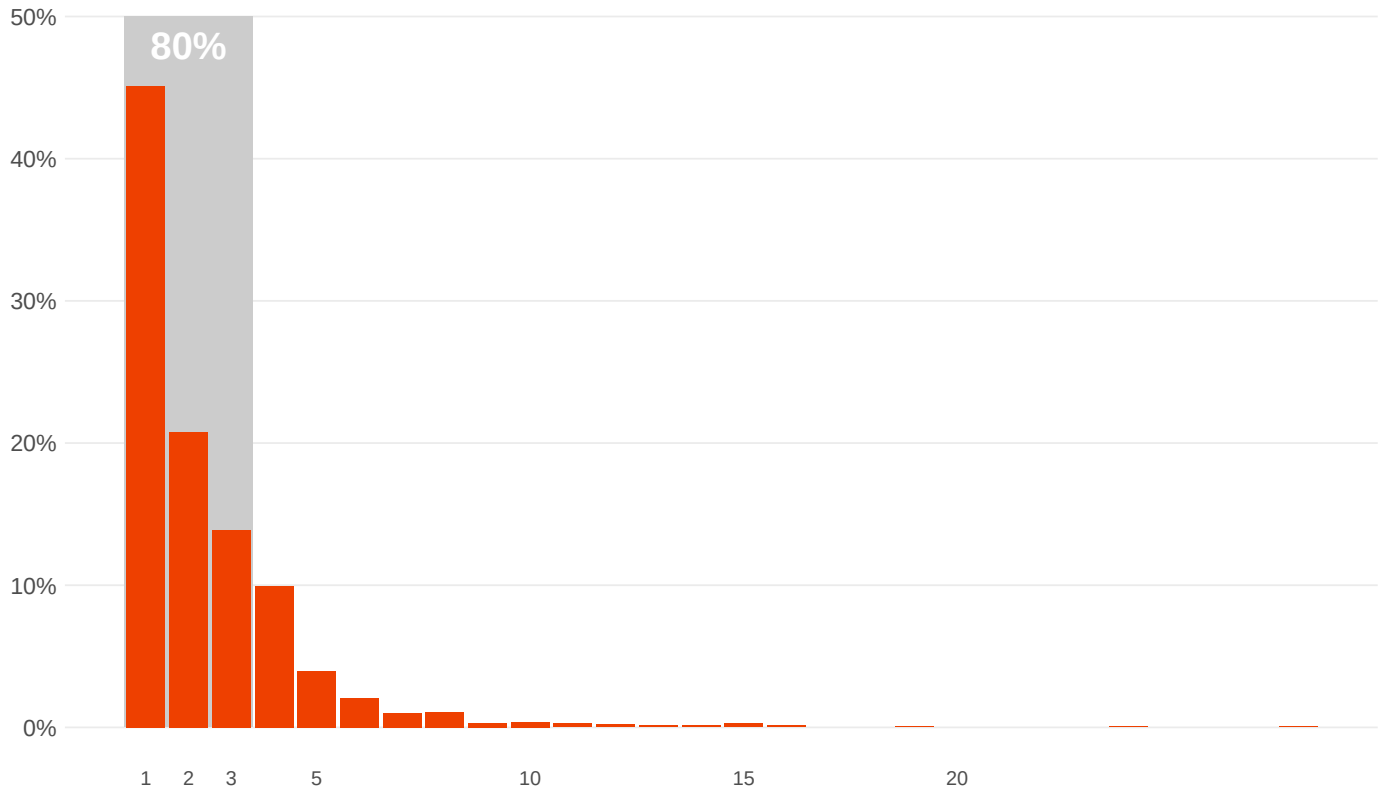


- the full database contains **5,396** events
  - in **57%** of events there were **neither deaths nor wounded**
  - in **27%** of events there is **perfect lethality**
  - in **8%** of events there were **both dead and wounded**
  - in **8%** of events there were **just wounded**

Data challenge #1

a few steps forward...

- 80% of events of “perfect lethality” had between 1 and 3 deaths



Data challenge #1

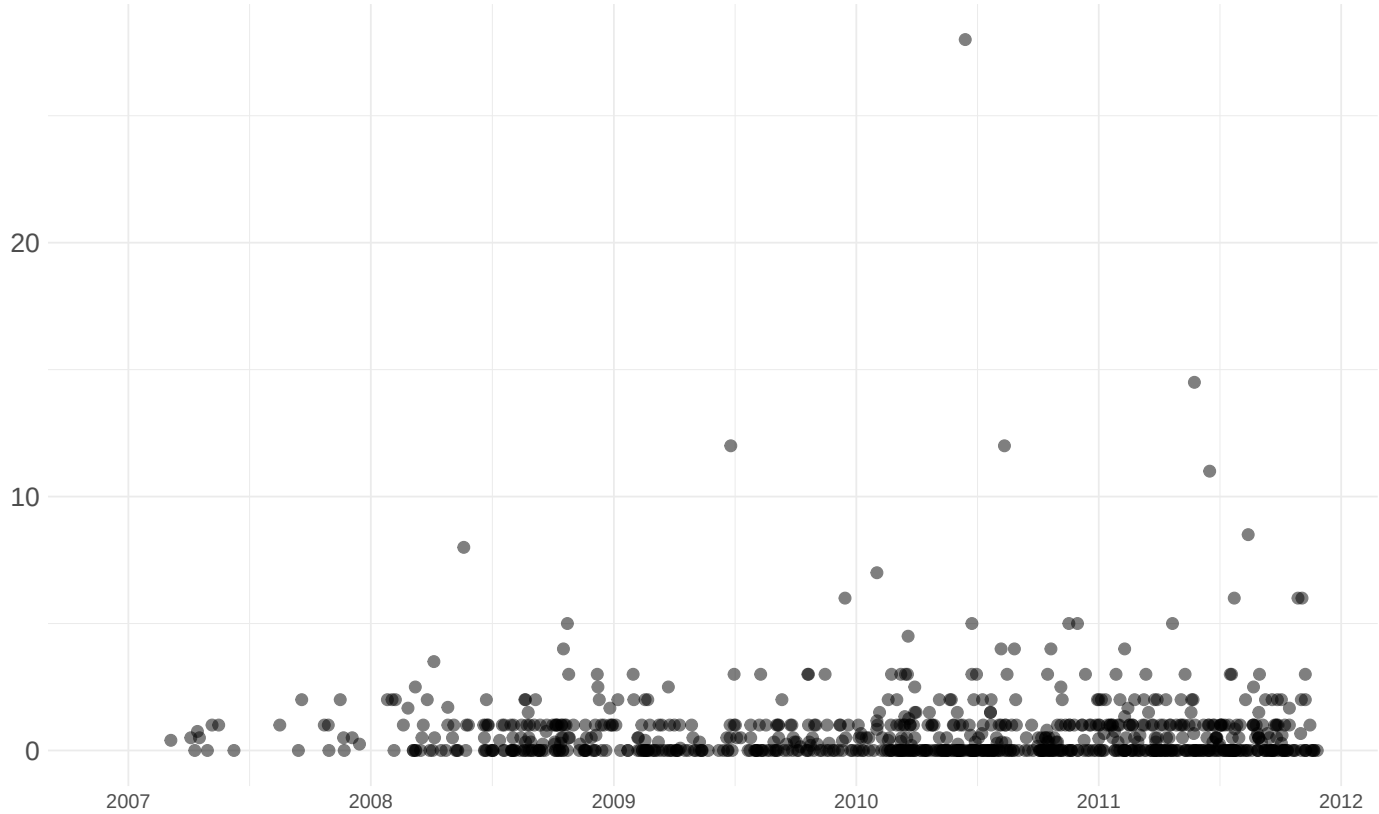
a few steps forward...

- All summary statistics have pros and cons
- Lethality index shows the proportionality between dead and wounded
- some edge cases
  - **all dead, no wounded** is excluded since the ratio is undefined (e.g. $\frac{8}{0}$ is undefined)
  - **no dead, no wounded** is also excluded since the ratio is again undefined (e.g. $\frac{0}{0}$ is undefined)
  - **no dead, all wounded** is misleading as it gives the same value whether it was one or a thousand wounded

Data challenge #1

a few steps forward...

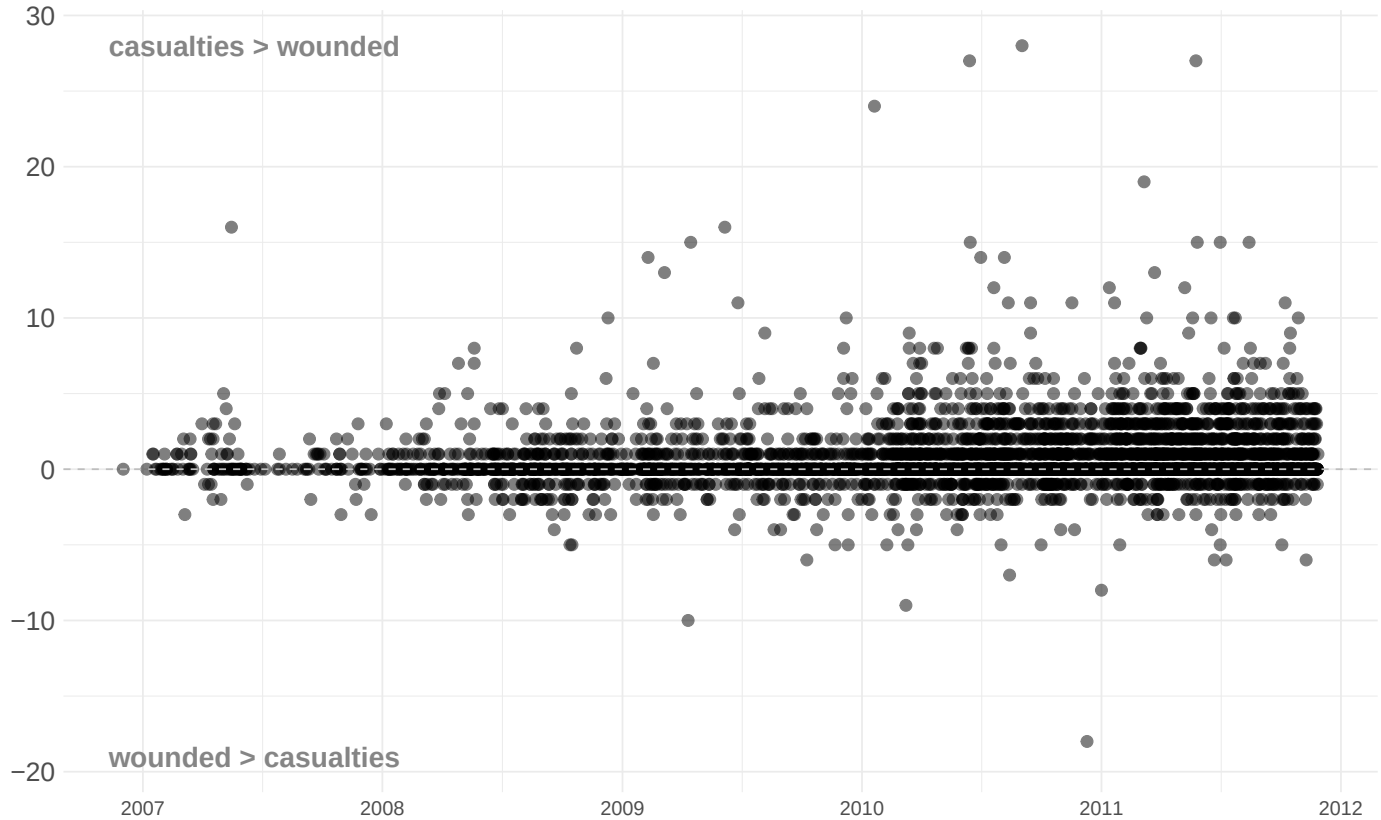
- the lethality index can only be computed for **825 cases (16% of events)**



Data challenge #1

a few steps forward...

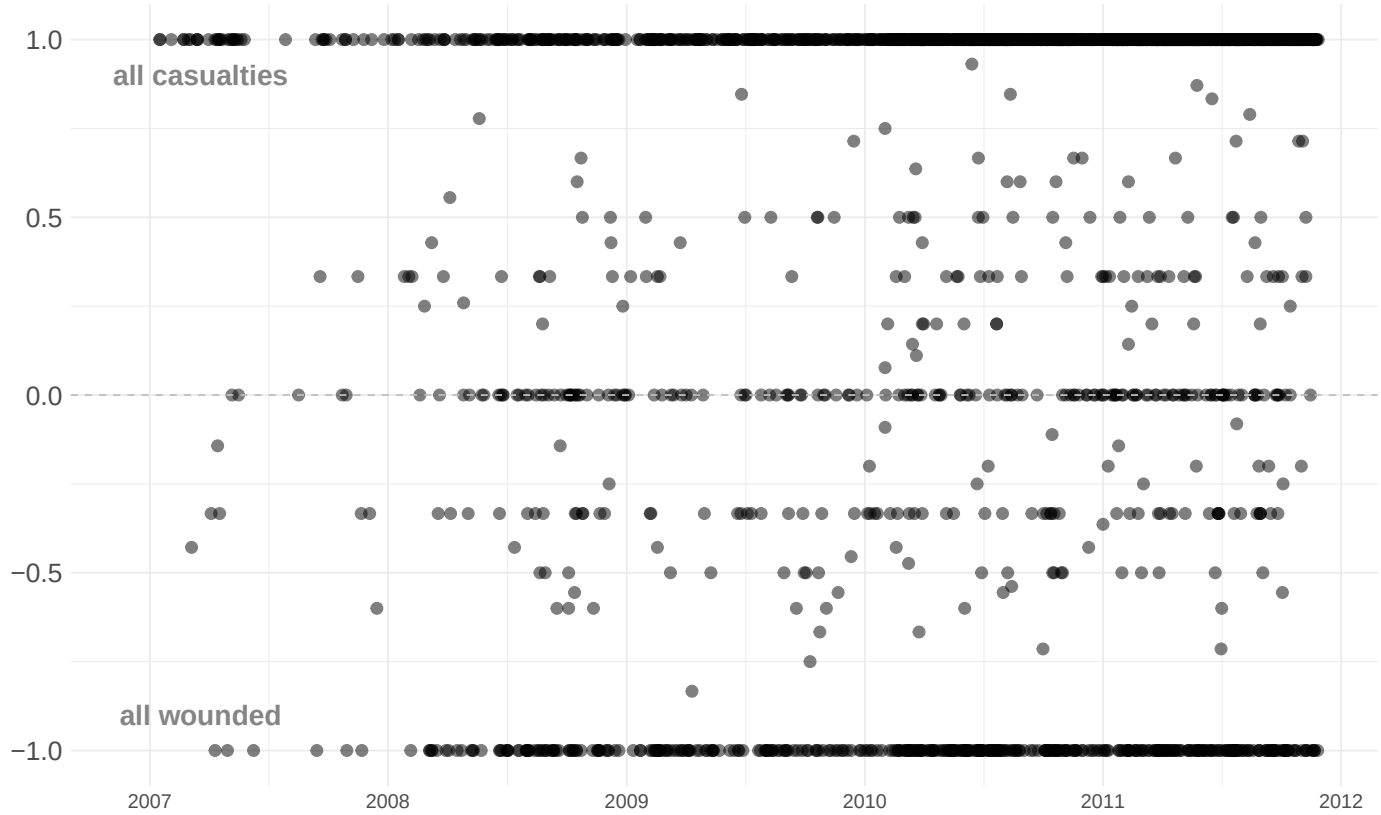
- if we're interested in the relation between dead and wounded, a simple difference may be illustrative: $d_i - w_i$



Data challenge #1

a few steps forward...

- with such variable range, perhaps we need to normalize:
  $\frac{d_i - w_i}{d_i + w_i}$



- what did we lose with the normalization? what did we gain?

Explanation vs Prediction

Explanation v Prediction

a framework...(Shmueli 2010)

- Let $\mathcal{X}$ cause $\mathcal{Y}$ through the function $\mathcal{F}$

$$\mathcal{Y} = \mathcal{F}(\mathcal{X})$$

- empirically...
  - $\mathbf{X}$ and Y operationalize $\mathcal{X}$ and $\mathcal{Y}$
  - f is the statistical model that operationalizes $\mathcal{F}$
- **explanatory modeling** seeks an f close to $\mathcal{F}$

$$E(Y) = f(\mathbf{X})$$

- **predictive modeling** seeks an $\hat{f}$ that best predicts Y_{new}

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

Explanation v Prediction

Expected Prediction Error (Hastie et al. 2009)

$$EPE = Var(Y) + Bias^2 + Var(\hat{f}(x)) \tag{1}$$

- where:

  $Var(Y) = E\{Y - f(x)\}^2$: random error
  $Bias^2 = \{E(\hat{f}(x)) - f(x)\}^2$: model mispecification
  $Var(\hat{f}(x)) = E\{\hat{f}(x) - E(\hat{f}(x))\}^2$: sample estimation

- **explanatory modeling**

$$min\{Bias^2\}$$

- **predictive modeling**

$$min\{Bias^2 + Var(\hat{f}(x))\}$$

Explanation v Prediction

a framework...(Shmueli 2010)

Explanatory Modeling	**Predictive Modeling**
f resembles $\mathcal{F}$	$\hat{f}$ links $\mathbf{X}, Y$
theory-selected $\mathbf{X}$	**association**-selected $\mathbf{X}$
may use **alternate** $\mathbf{X}$ and Y	requires **exact** $\mathbf{X}$ and Y
backward-looking	**forward**-looking
model fit validation	**predictive error** validation
$min(Bias^2)$ on (1)	$min\{Bias^2 + Var(\hat{f}(x))\}$ on (1)

Explanation v Prediction

... in sum

- any model will contain a combination of degrees of:
  - **explanatory power**
  - **predictive accuracy**
- two different dimensions or one with tradeoffs?
- a "good" model is **sophisticatedly simple** (Zellner 2001)

Prediction

"Unfortunately, we were a little off-target again this quarter."

Prediction

What do we mean by predictions?

- ***predict*:** *prae-* before + *dicere* to say
- ***forecast*:** *fore-* before + *casten* to prepare
- ***prognosticate*:** *pro-* before + *gnoscere* to know
- generically, the use of **observed information** to estimate **new information**

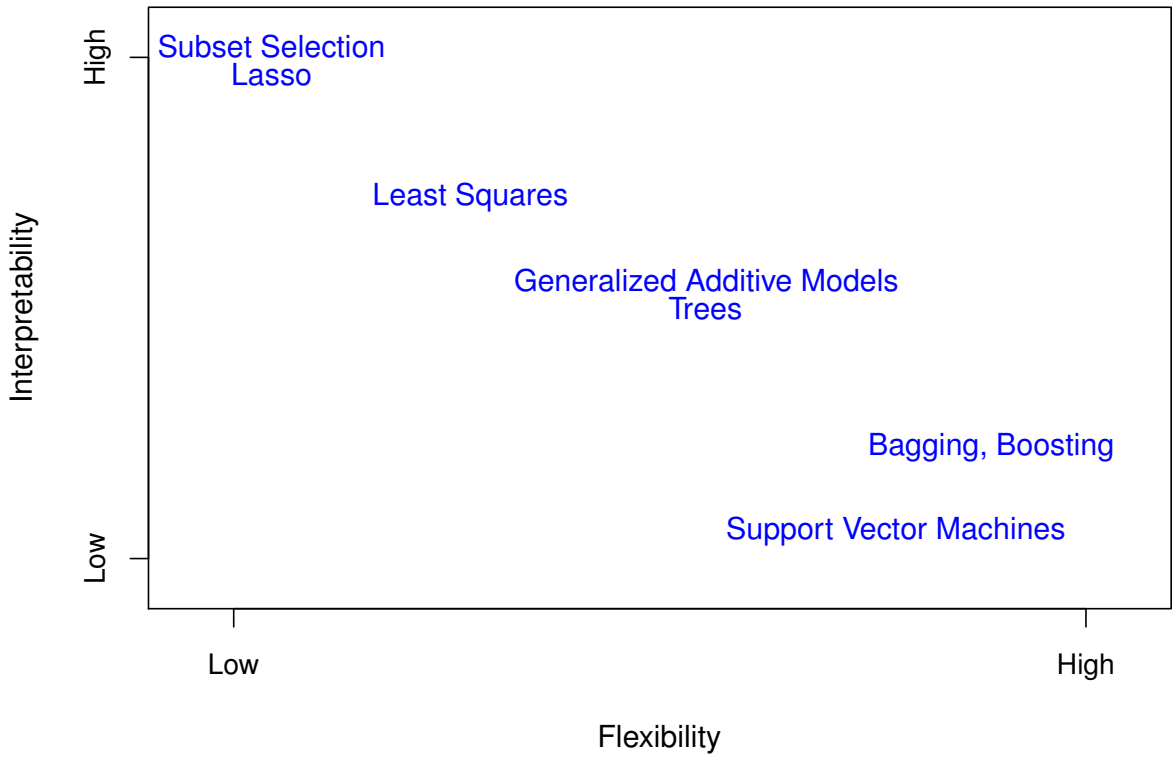
$$\hat{Y}_{new} = f(Y_{obs}, X_{obs})$$

Prediction

- **Predictability** depends on (Hyndman & Athanasopoulos 2013):
  - how well we know factors that influence the predictions
  - how much data (and of what quality!)
  - recursive influence of predictions (especially forecasts)
- Key question: **what to predict?**
  - every item?
  - at what level of aggregation?
  - at what frequency? daily? weekly? quarterly? yearly?
- **Remember:** explain $\neq$ predict

Prediction

Ways, Means and Tools to Predict...



- **Cross-Sectional models**
  - regression-based
  - ML-based

Prediction

Ways, Means and Tools to Predict...

- **Time-Series models**
  - Naïve

$$\hat{Y}_{t+1} = Y_t$$

  - Exponential Smoothing

$$\hat{Y}_{t+1|t} = \sum_{j=0}^{t-1} \alpha(1-\alpha)^j Y_{t-j} + (1-\alpha)^t \ell_0$$

  - ARIMA models

$$\hat{Y}_{t+1} = c + \phi_1 Y_t + ... + \phi_p Y_{t-p} + \theta_1 \epsilon_t + ...\theta_q \epsilon_{t-q} + \epsilon_{t+1}$$

Prediction

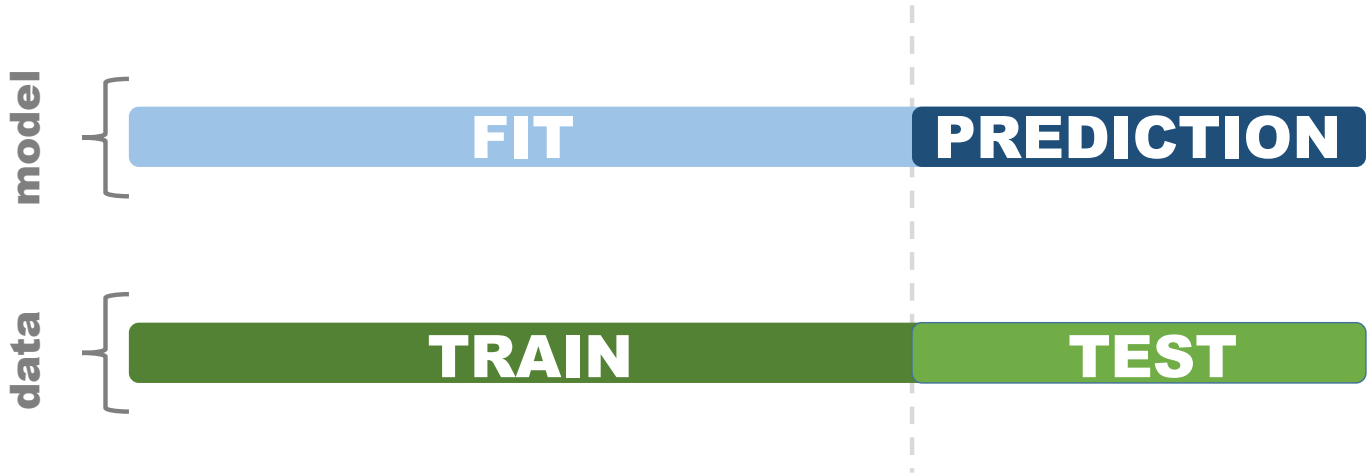
Some (empirically validated) rules of thumb

1. **keep it simple:**
   - start parsimonious and add complexity (*iff* called for)
   - increased complexity typically reduces accuracy
2. **rely on domain expertise to select inputs**
   - statistical significance a faulty guide for inclusion
   - domain expertise should drive variables to include
3. **include more (useful) information**
   - high correlation in predictors not an issue
4. **fit $\neq$ accuracy**
   - well-fitting models may impose unwarranted "structure" and "certainty" to the forecast
5. **update models constantly**
   - update parameters as new information arrives

Prediction

Cross-Validation and Overfitting

- goal of prediction is **generalization**
  - a model that **overfits** the data is not generalizable
- assessing **predictive accuracy**:



i) estimate model on a **training set**

ii) measure error on a **test set**

Prediction

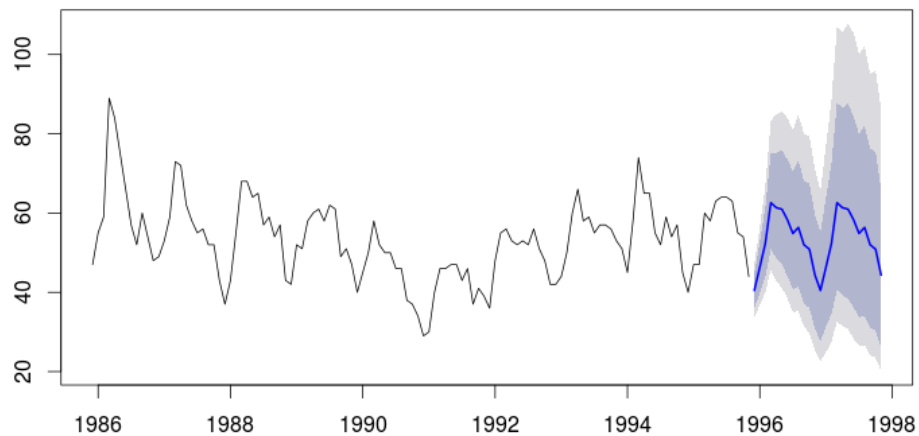
Cross-Validation and Overfitting

- performance on **test set** should be as good as on **training set**
- **caveat:** training and test sets should come **from the same population**
- **cross-validation** can take many flavors (e.g. **k-fold validation**, **leave p-out cross-validation**...)

Prediction

Other issues: prediction uncertainty...

- by definition, predictions are uncertain
  - we should be interested in **point estimates** of predictions and their **prediction intervals**
- it is possible to estimate the **range of values** where predictions may lie **with a given probability**



Prediction

Other issues: Forecast Ensembles

Prediction

Other issues: Forecast Ensembles

- we typically think of a single model to produce forecasts
  - what if we have various “informative” models?
- simple averaging of forecasts has proven in many cases superior to single forecasts
  - complex methods have been devised to optimize forecast weights, not always best
- particularly useful when models/methods are sufficiently different

Prediction

Other issues: Time-Series cross-validation

- usual cross-validation **inadequate** for time-series because of lagged values in these models
- an appropriate (rolling) **time-series cross-validation algorithm** (Hyndman):
  1. fit your time-series model and compute the error (ϵ^*_{t+h}) for the forecasted observation ($\hat{Y}_{t+h}$) h steps into the future per

$$\epsilon^*_{t+h} = Y_{t+h} - \hat{Y}_{t+h}$$

  2. repeat step 1 for $t = m + h, ..., n - 1$ where m is the minimal number of obs to estimate model
  3. compute appropriate error measure (i.e. MAPE, RMSE..) with estimated errors

Prediction

Other issues: Time-Series cross-validation

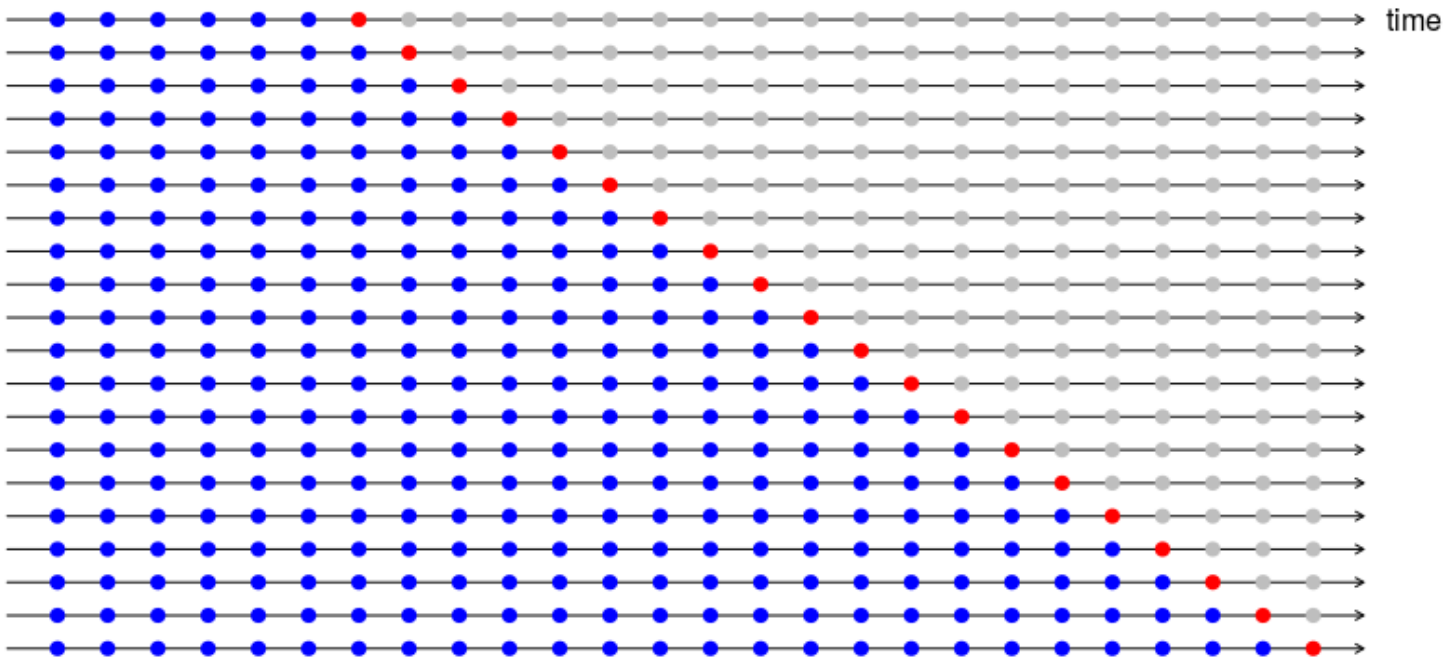


Figure: Rob Hyndman

Prediction

Other issues: Time-Series cross-validation

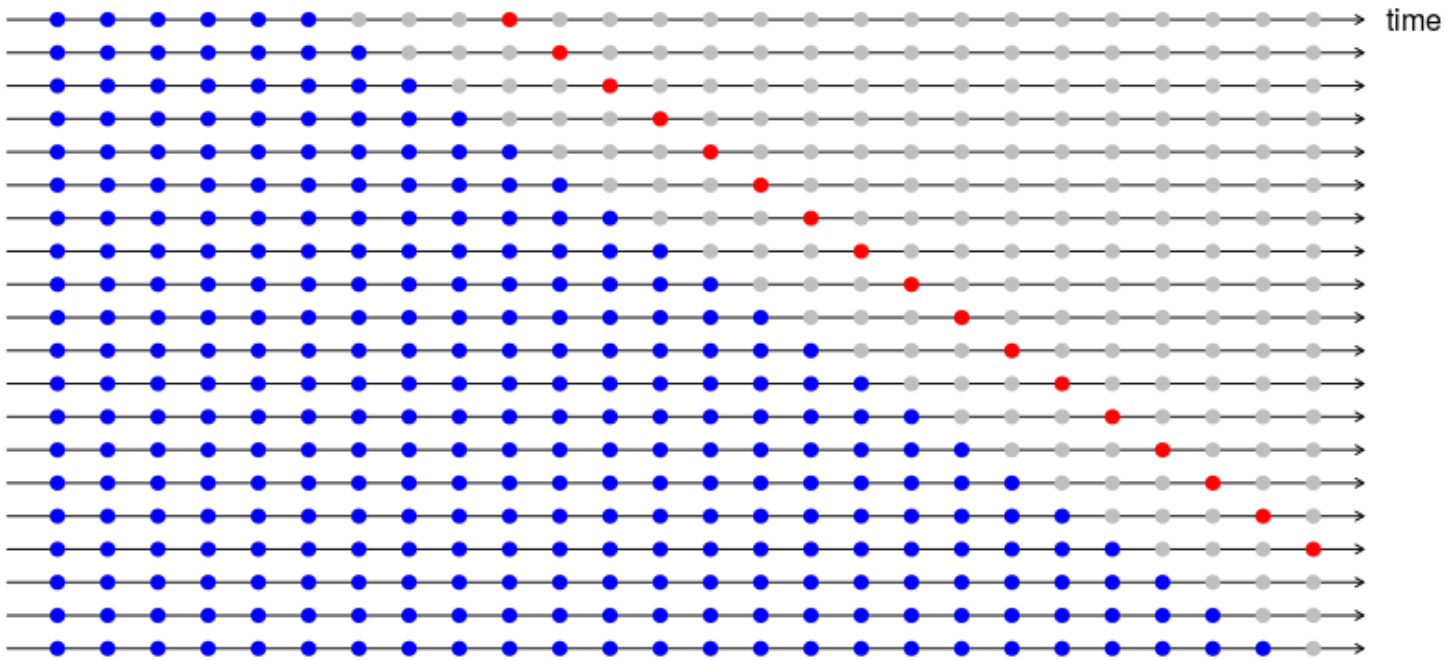


Figure: Rob Hyndman

Prediction

Other issues: Time-Series cross-validation

Error Measure	Definition
Root Mean Squared Error (RMSE)	$\sqrt{\frac{\sum_{i=1}^{n}(Y_{t+i}-\hat{Y}_{t+i})^2}{n}}$
Mean Absolute Percent Error (MAPE)	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\lvert Y_{t+i}-\hat{Y}_{t+i}\rvert}{Y_{t+i}} * 100\right)$

- when evaluating forecasts remember:
  - is the measure **valid** (makes sense to experts)?
  - is the measure **sensitive to outliers**?
  - is the measure **affected by scale**?
  - **do not use R^2 to assess models**

Team Planning

Explanation v Prediction

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2018
Columbia University