A Framework for Understanding Unintended Consequences of Machine Learning

Robert Różański

Harini Suresh & John V. Guttag 2019
https://arxiv.org/pdf/1901.10002.pdf

unintended and potentially harmful consequences:

- underperforming consumer products/services
- unequal access to resources
- legal failure/injustice

unintended and potentially harmful consequences:

- underperforming consumer products/services
- unequal access to resources
- legal failure/injustice

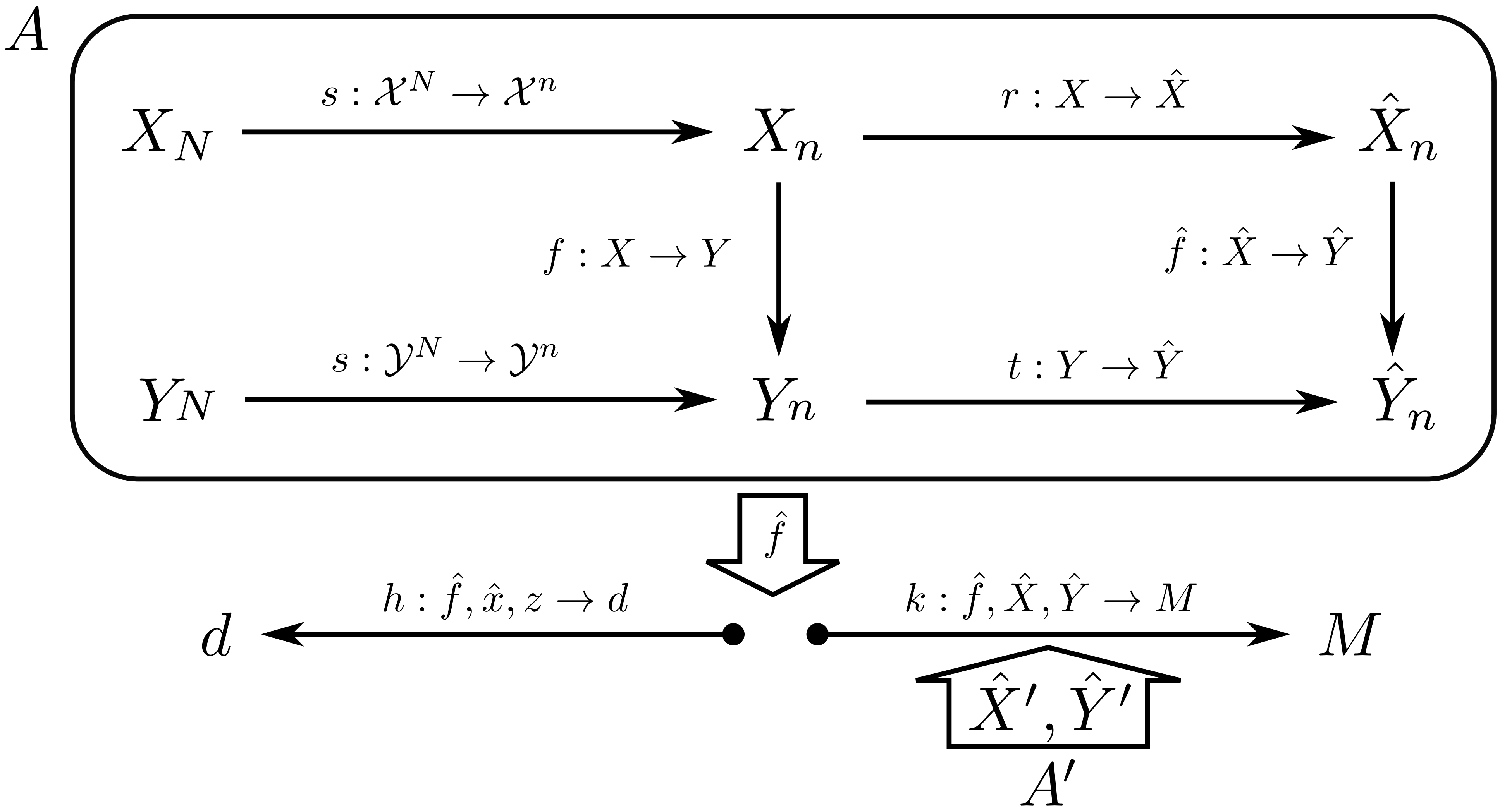
no universal solution:
e.g. in the context of protected attributes, sometimes their use is prohibited (sex & job application), while in others they must be used to ensure good treatment (sex & medical diagnosis)

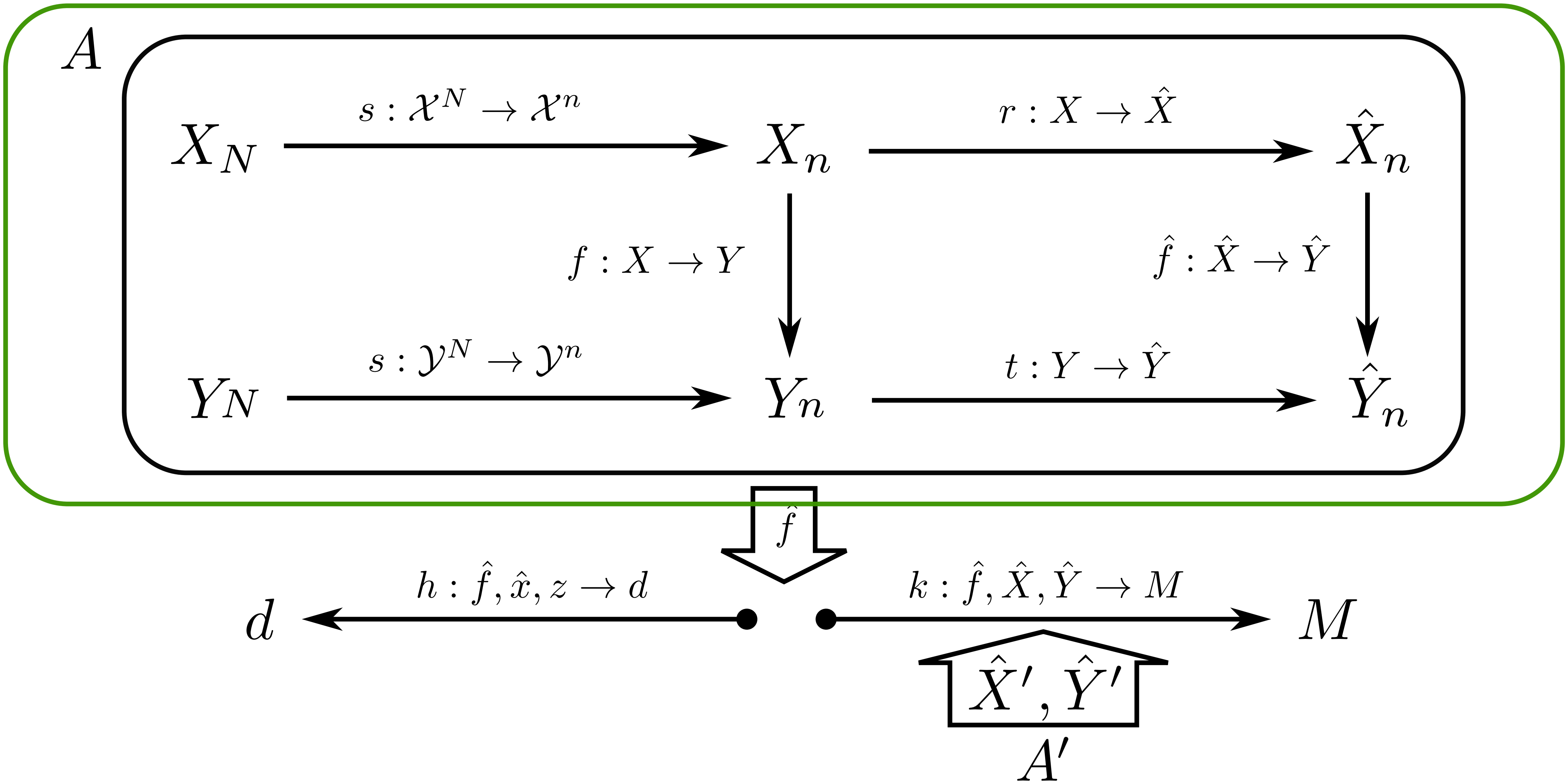
unintended and potentially harmful consequences:

- underperforming consumer products/services
- unequal access to resources
- legal failure/injustice

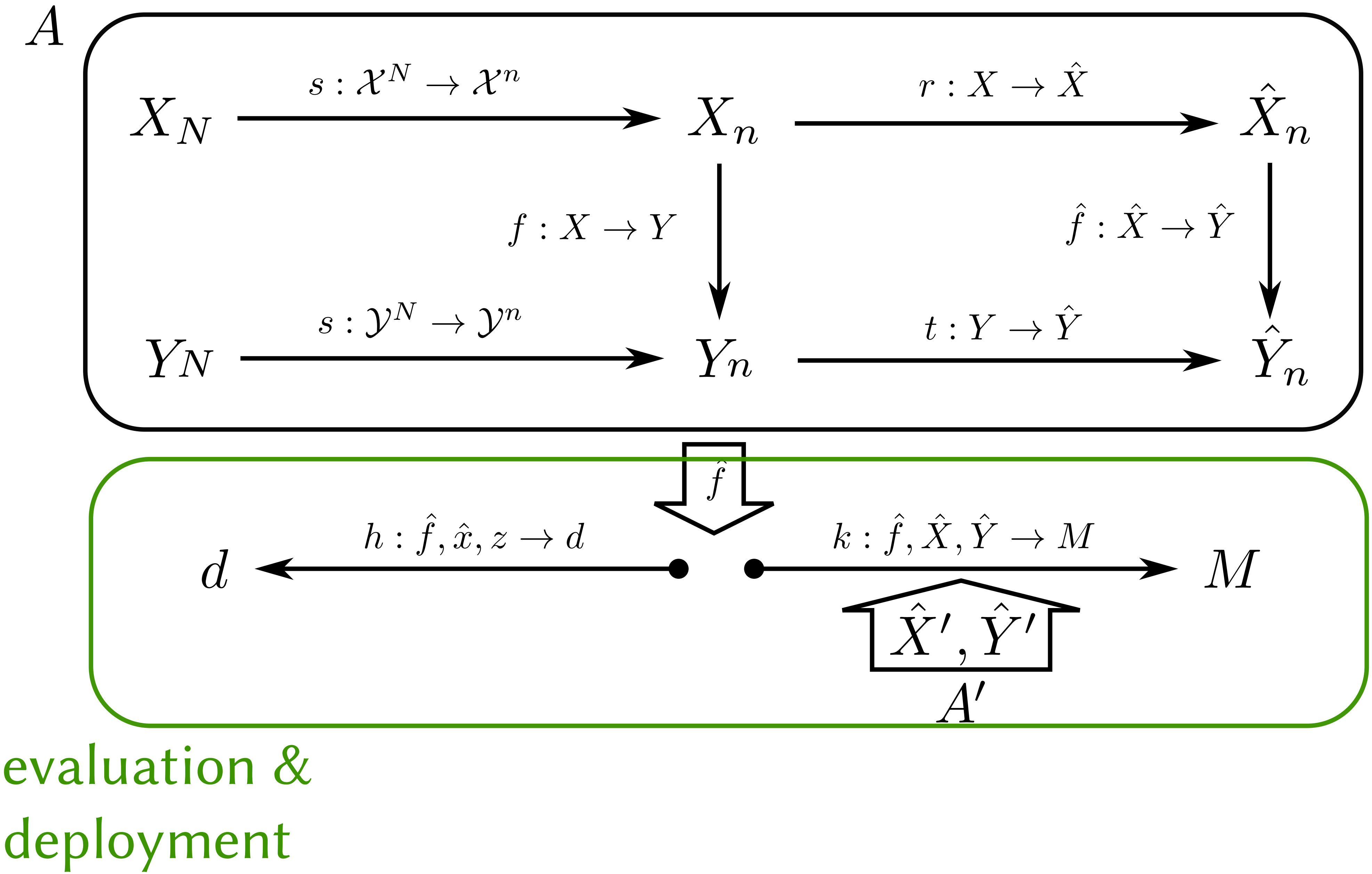
no universal solution:
e.g. in the context of protected attributes, sometimes their use is prohibited (sex & job application), while in others they must be used to ensure good treatment (sex & medical diagnosis)

there is a lack of common framework that would allow to compare problems and solutions

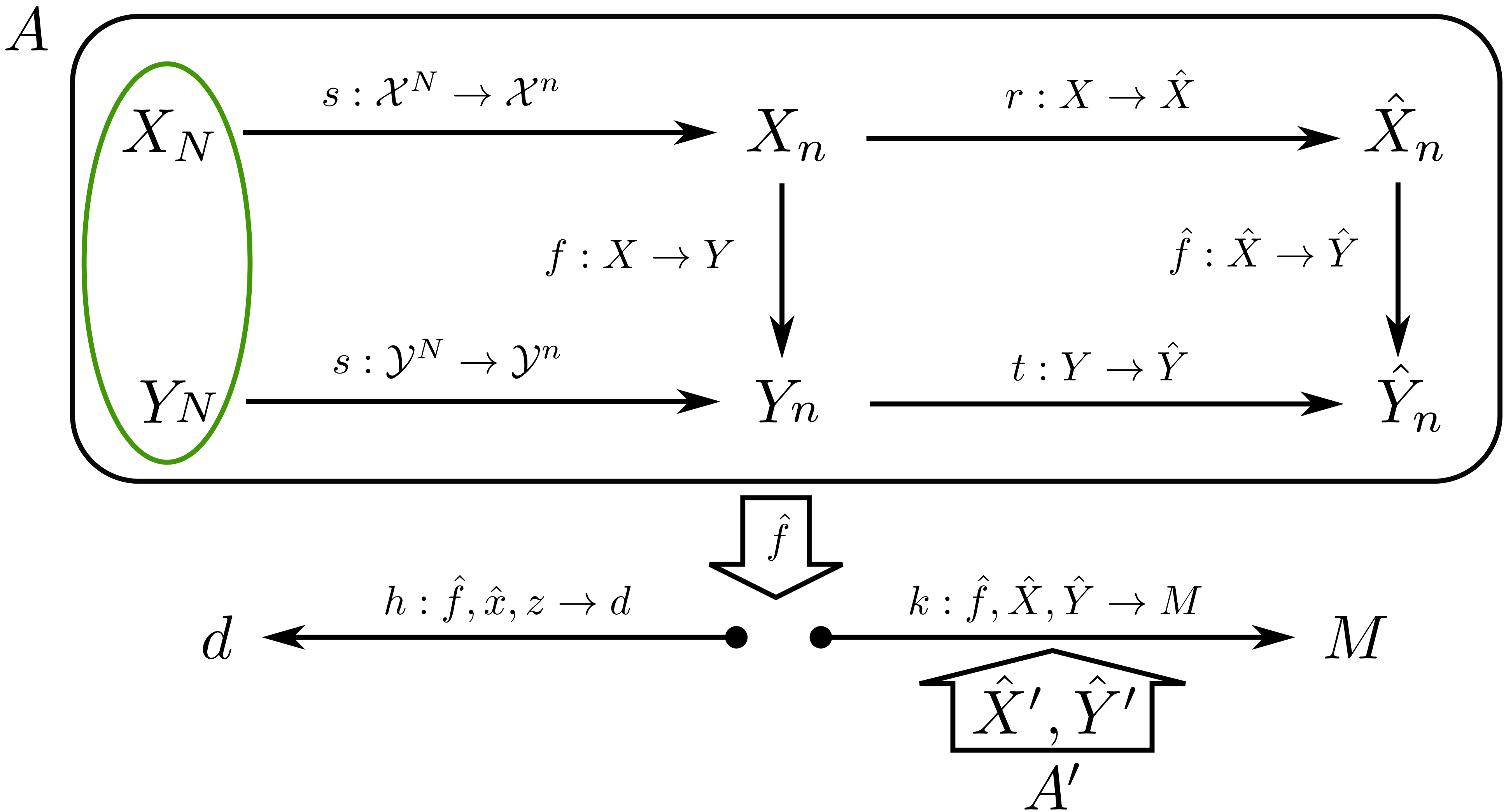

A
$s : \mathcal{X}^N \to \mathcal{X}^n$
$r : X \to \hat{X}$
X_N
X_n
$\hat{X}_n$
$f : X \to Y$
$\hat{f} : \hat{X} \to \hat{Y}$
$s : \mathcal{Y}^N \to \mathcal{Y}^n$
$t : Y \to \hat{Y}$
Y_N
Y_n
$\hat{Y}_n$
$\hat{f}$
$h : \hat{f}, \hat{x}, z \to d$
$k : \hat{f}, \hat{X}, \hat{Y} \to M$
d
M
$\hat{X}', \hat{Y}'$
A'



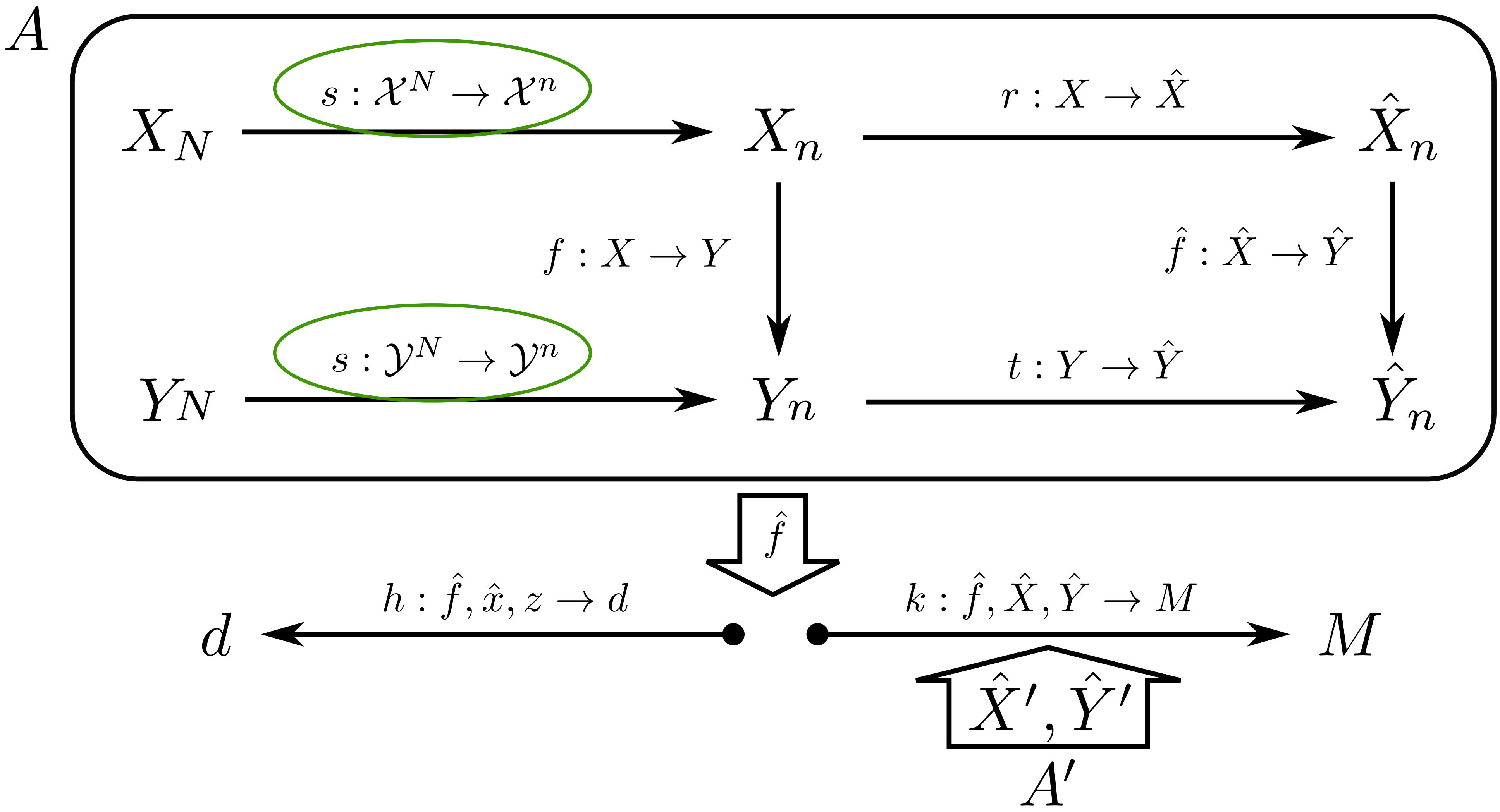
model building
process



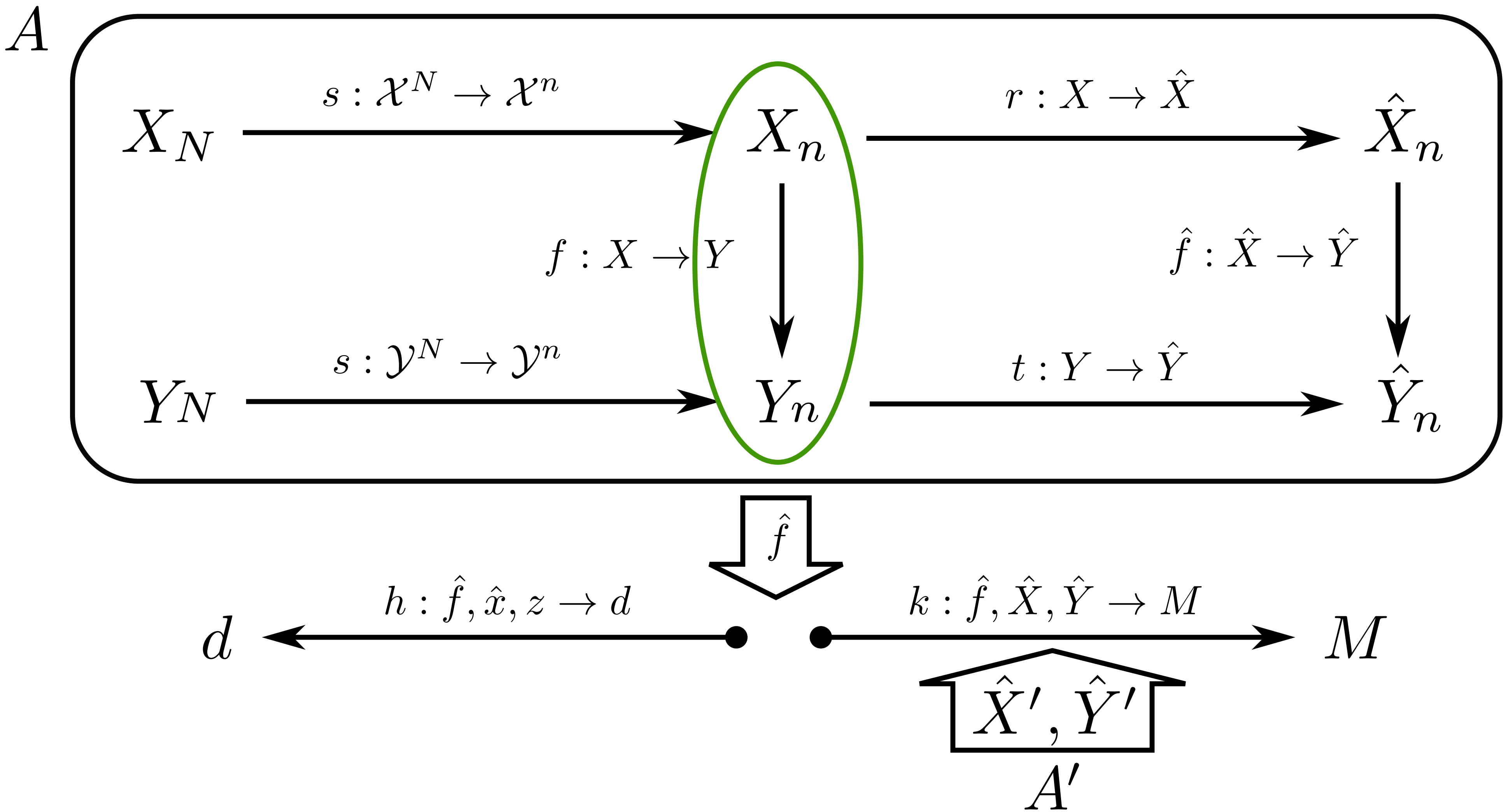
A
X_N
$s : \mathcal{X}^N \to \mathcal{X}^n$
X_n
$r : X \to \hat{X}$
$\hat{X}_n$
$f : X \to Y$
$\hat{f} : \hat{X} \to \hat{Y}$
Y_N
$s : \mathcal{Y}^N \to \mathcal{Y}^n$
Y_n
$t : Y \to \hat{Y}$
$\hat{Y}_n$
$\hat{f}$
d
$h : \hat{f}, \hat{x}, z \to d$
$k : \hat{f}, \hat{X}, \hat{Y} \to M$
M
$\hat{X}', \hat{Y}'$
A'
evaluation &
deployment



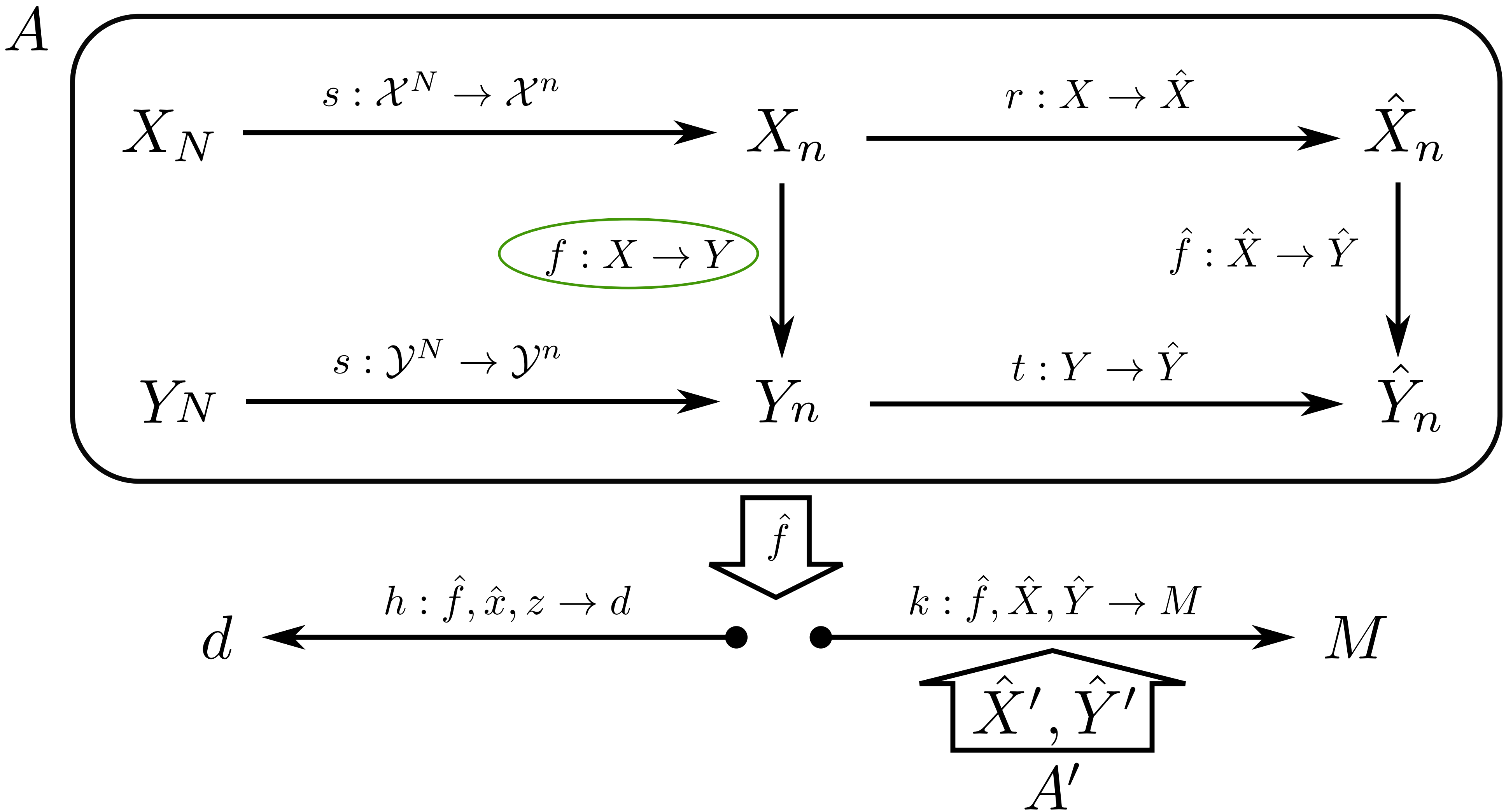
Ideal, underlying
features on the
whole population



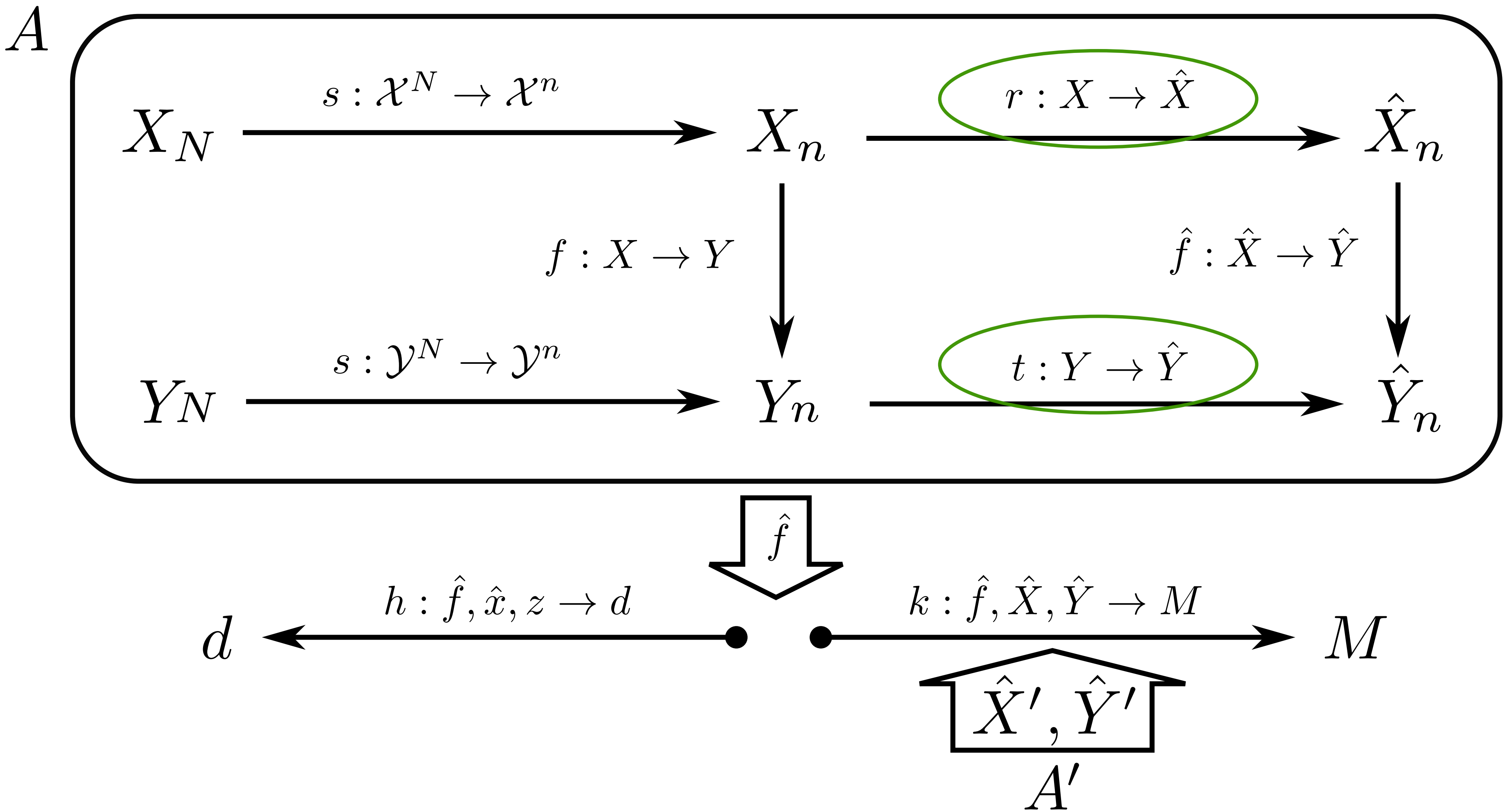
sampling functions



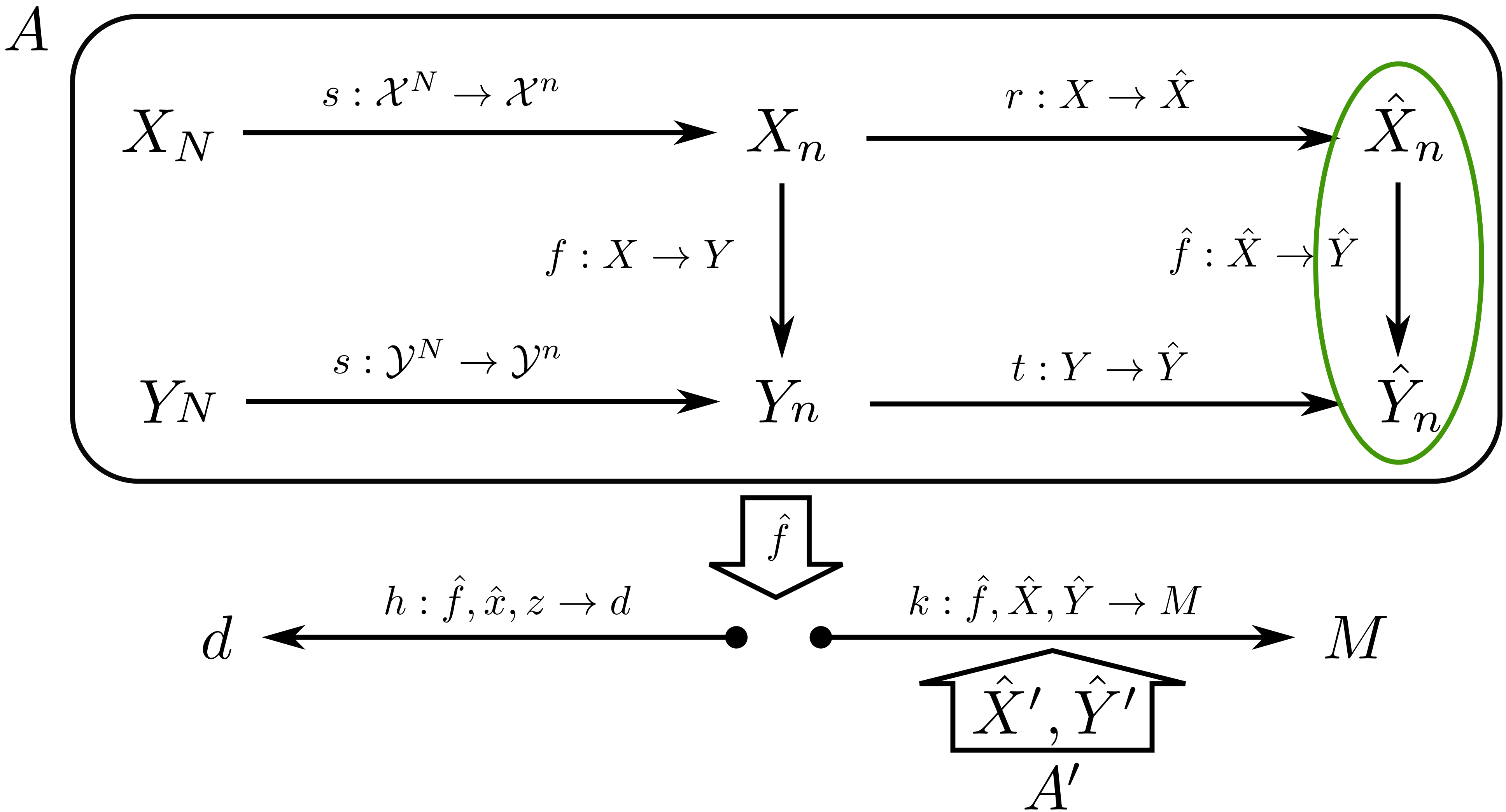
Ideal, underlying features on the sample



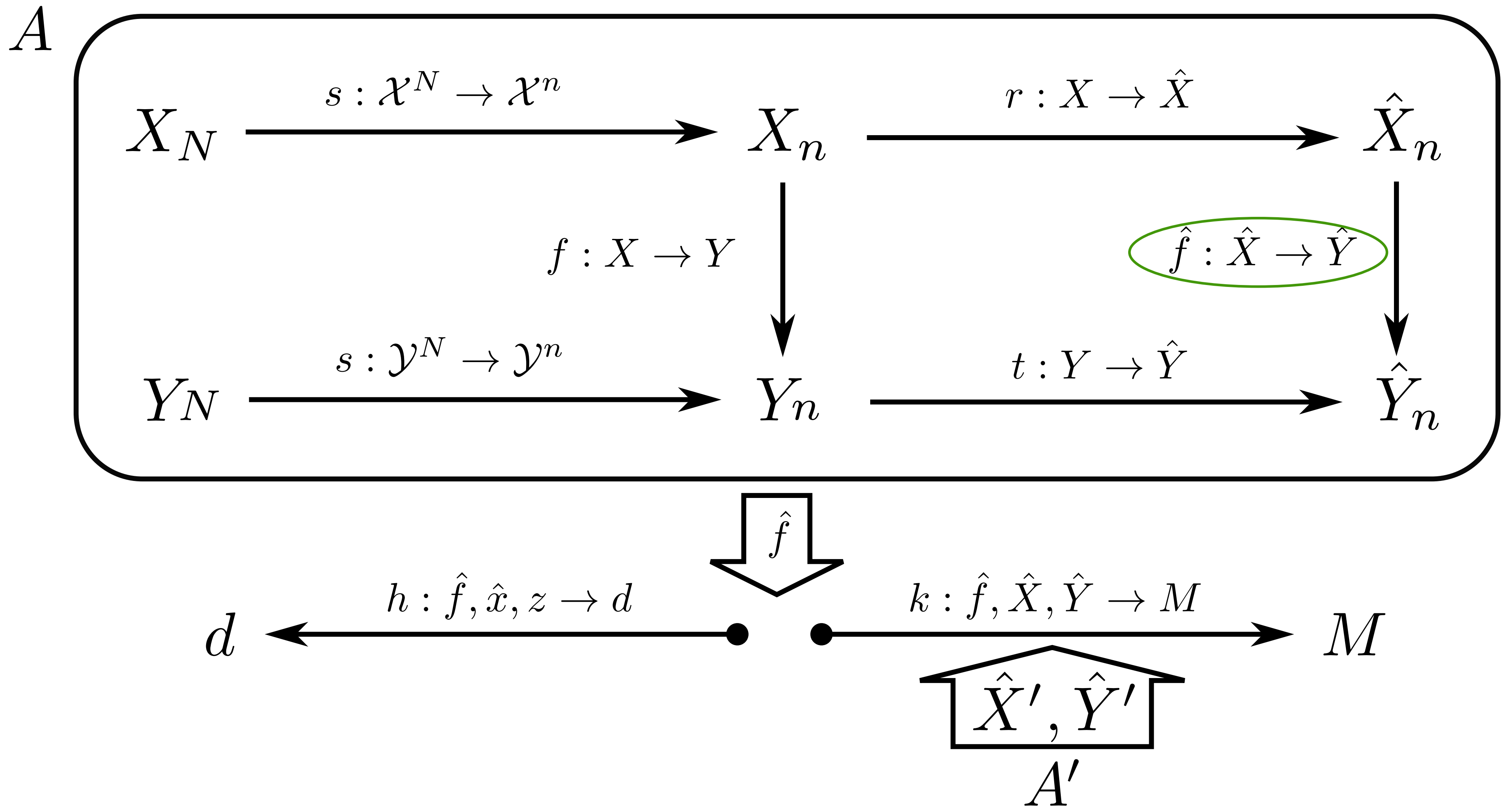
model based on the ideal, underlying features



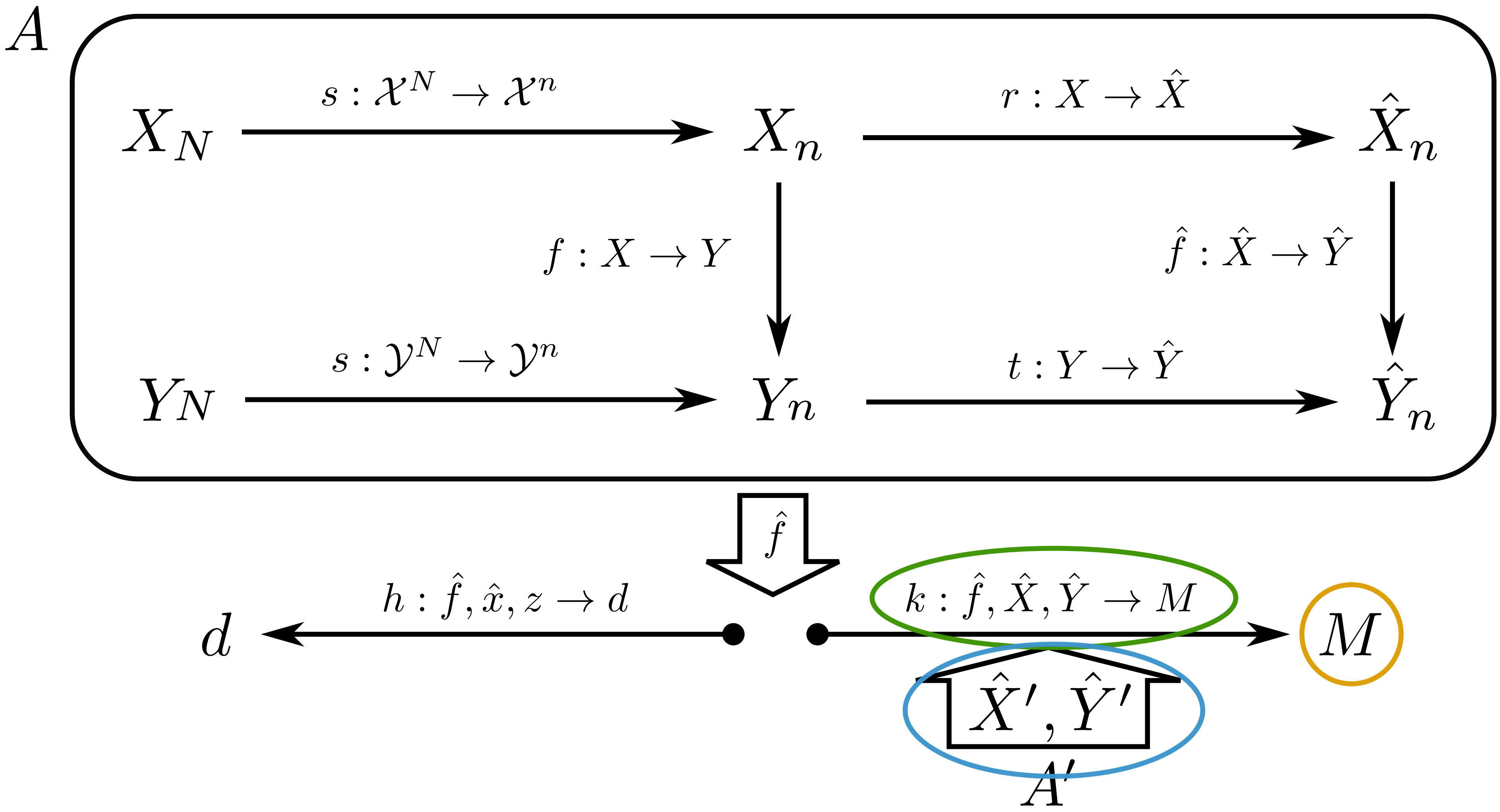
projections from the
ideal features to the
measured ones



Actual, measured features on the sample



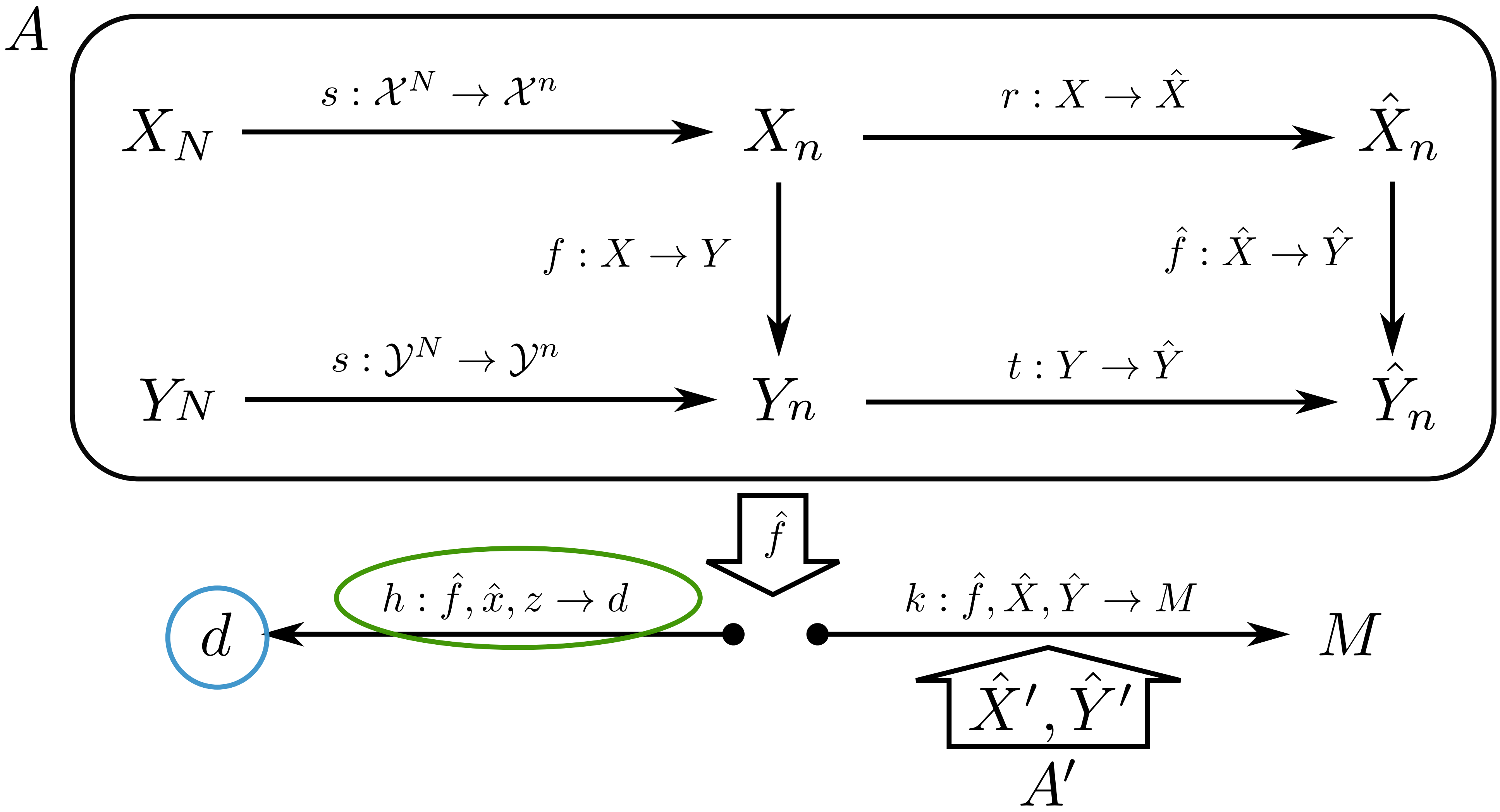
actual model, based
on the measured
features



scoring function

test data

measure of success



real world decision

process

decision

A

$$
\begin{tikzcd}
X_N \arrow[r, "s : \mathcal{X}^N \to \mathcal{X}^n"] & X_n \arrow[r, "r : X \to \hat{X}"] \arrow[d, "f : X \to Y"'] & \hat{X}_n \arrow[d, "\hat{f} : \hat{X} \to \hat{Y}"'] \\
Y_N \arrow[r, "s : \mathcal{Y}^N \to \mathcal{Y}^n"] & Y_n \arrow[r, "t : Y \to \hat{Y}"] & \hat{Y}_n
\end{tikzcd}
$$

historical bias: if the world as it is or was leads a model to produce outcomes that are not wanted.

Example: representation bias in image search results:

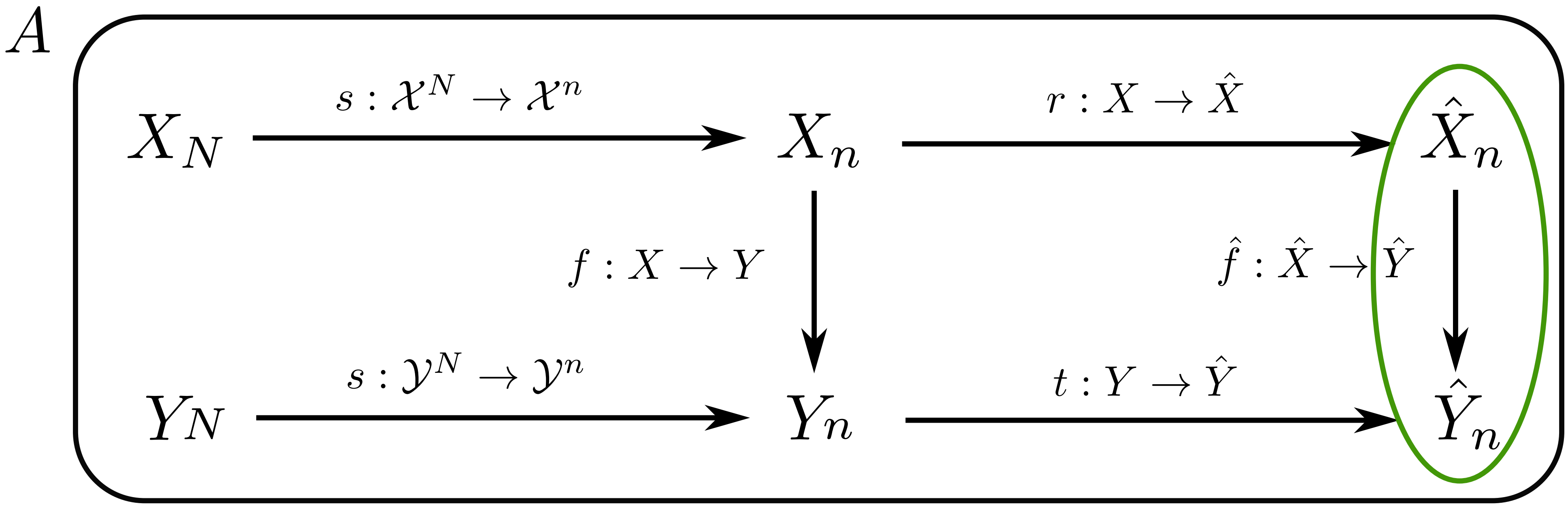
https://www.theguardian.com/technology/2016/dec/05/google-alters-search-autocomplete-remove-are-jews-evil-suggestion

A

$$
\begin{tikzcd}
X_N \arrow[r, "s : \mathcal{X}^N \to \mathcal{X}^n"] & X_n \arrow[r, "r : X \to \hat{X}"] \arrow[d, "f : X \to Y"'] & \hat{X}_n \arrow[d, "\hat{f} : \hat{X} \to \hat{Y}"'] \\
Y_N \arrow[r, "s : \mathcal{Y}^N \to \mathcal{Y}^n"] & Y_n \arrow[r, "t : Y \to \hat{Y}"] & \hat{Y}_n
\end{tikzcd}
$$

representation bias: certain parts of the input
space are underrepresented. This could be caused not only by deficient sampling methods, but, for example, also because the population of interest has changed, etc.

Example: geographic diversity in the ImageNet (45% from USA)



measurement bias: available/measureable features and labels are noisy proxies for the features and labels of interest. Could be a result of differences in measurement process or data quality accross groups, etc.

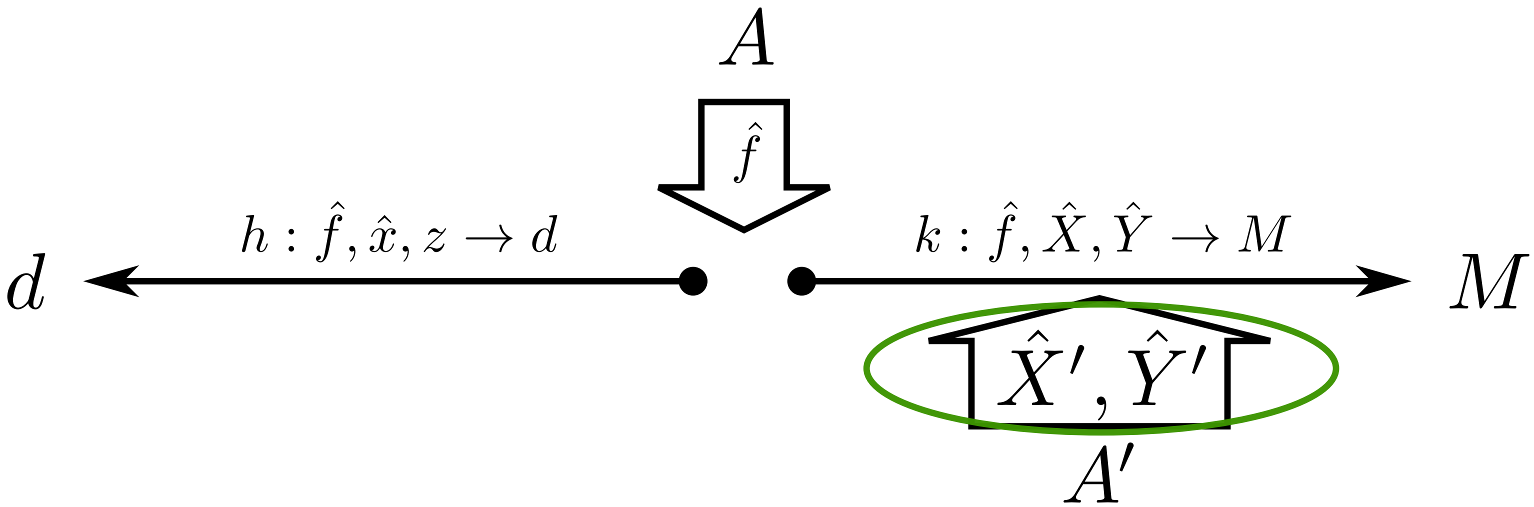
Example: in predictive policing and recidivism prediction proxy variable "arrest" is used to measure "crime"

A

$$\begin{tikzcd}
X_N \arrow[r, "s : \mathcal{X}^N \to \mathcal{X}^n"] & X_n \arrow[r, "r : X \to \hat{X}"] \arrow[d, "f : X \to Y"'] & \hat{X}_n \arrow[d, "\hat{f} : \hat{X} \to \hat{Y}"'] \\
Y_N \arrow[r, "s : \mathcal{Y}^N \to \mathcal{Y}^n"] & Y_n \arrow[r, "t : Y \to \hat{Y}"] & \hat{Y}_n
\end{tikzcd}$$

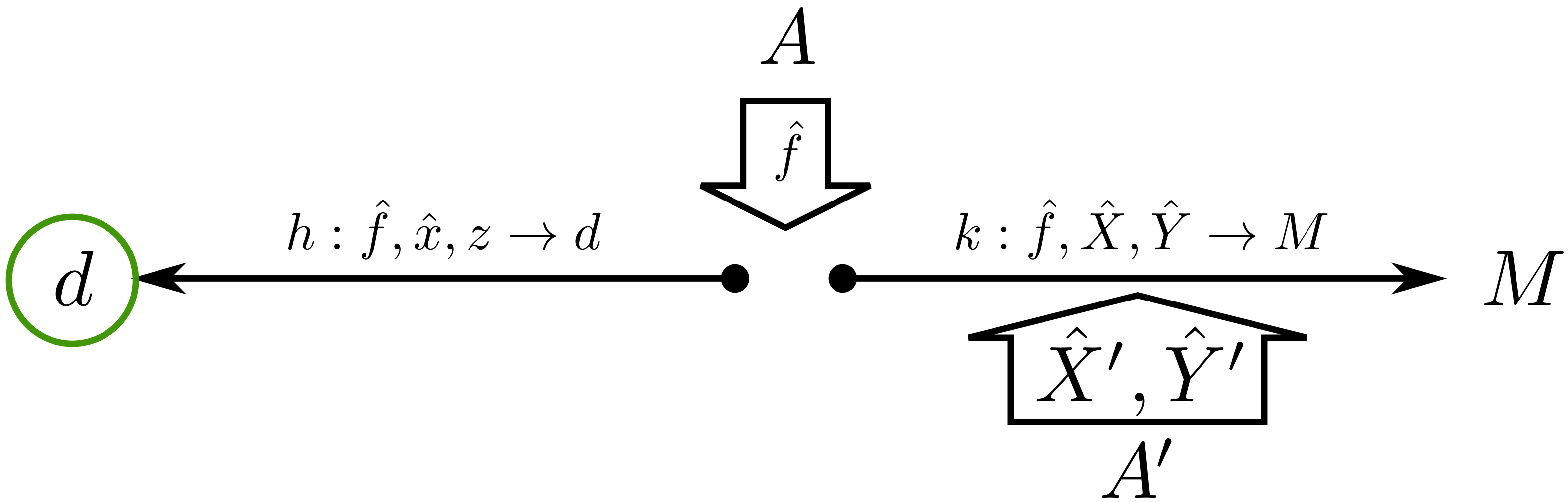
aggregation bias: when a one-size-fit-all model is used
for groups with different conditional distributions. This can lead to underperforming models even if the sample is balanced.

Example: in diabetes diagnosis/monitoring, clinical meaning of HbA1c levels differs accross ethnicities and genders



evaluation bias: when the evaluation and/or benchmark data for an algorithm doesn't represent the target population.

Example: underperformance of facial analysis models on dark-skinned females (7.4% and 4.4% of the images in benchmark datasets such as Adience and IJB-A are of dark-skinned female faces.)



deployment bias: there is a mismatch between the problem a model is intended to solve and the way in which it is actually used.

Example: models predicting person's likelihood of commiting a future crime used in "off label" ways, e.g. to help determine the length of a sentence.

A

$$\begin{tikzcd}
X_N \arrow[r, "s : \mathcal{X}^N \to \mathcal{X}^n"] & X_n \arrow[r, "r : X \to \hat{X}"] \arrow[d, "f : X \to Y"'] & \hat{X}_n \arrow[d, "\hat{f} : \hat{X} \to \hat{Y}"'] \\
Y_N \arrow[r, "s : \mathcal{Y}^N \to \mathcal{Y}^n"] & Y_n \arrow[r, "t : Y \to \hat{Y}"] & \hat{Y}_n
\end{tikzcd}$$

case study: aggregation bias

A

$$
\begin{tikzcd}
X_N \arrow[r, "s : \mathcal{X}^N \to \mathcal{X}^n"] & X_n \arrow[r, "r : X \to \hat{X}"] \arrow[d, "f : X \to Y"'] & \hat{X}_n \arrow[d, "\hat{f} : \hat{X} \to \hat{Y}"'] \\
Y_N \arrow[r, "s : \mathcal{Y}^N \to \mathcal{Y}^n"] & Y_n \arrow[r, "t : Y \to \hat{Y}"] & \hat{Y}_n
\end{tikzcd}
$$

case study: aggregation bias

adjusting the model:

- using multitask learning
- using more complex model, which can capture the differences in conditional distribution between the groups (if enough data available)

A

$$\begin{tikzcd}
X_N \arrow[r, "s : \mathcal{X}^N \to \mathcal{X}^n"] & X_n \arrow[r, "r : X \to \hat{X}"] \arrow[d, "f : X \to Y"'] & \hat{X}_n \arrow[d, "\hat{f} : \hat{X} \to \hat{Y}"'] \\
Y_N \arrow[r, "s : \mathcal{Y}^N \to \mathcal{Y}^n"] & Y_n \arrow[r, "t : Y \to \hat{Y}"] & \hat{Y}_n
\end{tikzcd}$$

case study: aggregation bias

adjusting the model:
- using multitask learning
- using more complex model, which can capture the differences in conditional distribution between the groups (if enough data available)

adjusting the data:
- changing the projection functions *r* & *t* (representation learning)