Targeted Minimum Loss-based Estimation (TMLE) for Causal Inference (in Biostatistics)

Helene Charlotte Wiese Rytgaard (hely@sund.ku.dk)

Thomas Alexander Gerds (tag@biostat.ku.dk)

Anders Munch (a.munch@sund.ku.dk)

Ann-Sophie Buchardt (asbu@sund.ku.dk)

Overview of topics of course

Background theory

- Understanding key concepts of nonparametric efficiency theory.
- Estimation and inference based on the efficient influence function.

The TMLE procedure

- Targeted loss-based learning incorporating the efficient influence function.
- Data-adaptive estimation via machine learning.

Causal inference part

- Model-free (nonparametric) definition of statistical target parameter.
- Causal interpretation under certain assumptions.

Practical part

- Explore properties of estimation based on the efficient influence function.
- Assess model misspecification and estimator performance via simulations in `R`.

Overview of topics of course

We need a certain understand of relevant mathematical/statistical topics.

- Really foundational to understanding TMLE.

Overview of topics of course

We need a certain understand of relevant mathematical/statistical topics.

- Really foundational to understanding TMLE.
- On all levels this theory gives good insight, and it is needed to 'open the black box' of TMLE.

Overview of topics of course

We need a certain understand of relevant mathematical/statistical topics.

- Really foundational to understanding TMLE.
- On all levels this theory gives good insight, and it is needed to 'open the black box' of TMLE.

On the other hand, this is not a "math course".

- Theoretical background is quite deep, but too much for this course.
- (I will be pretty handwavy at times).

Overview of topics of course

We need a certain understand of relevant mathematical/statistical topics.

- Really foundational to understanding TMLE.
- On all levels this theory gives good insight, and it is needed to 'open the black box' of TMLE.

On the other hand, this is not a "math course".

- Theoretical background is quite deep, but too much for this course.
- (I will be pretty handwavy at times).

I will say many basic things.

Overview of topics of course

We need a certain understand of relevant mathematical/statistical topics.

- Really foundational to understanding TMLE.
- On all levels this theory gives good insight, and it is needed to 'open the black box' of TMLE.

On the other hand, this is not a "math course".

- Theoretical background is quite deep, but too much for this course.
- (I will be pretty handwavy at times).

I will say many basic things.

- For the larger part, we focus on the simple example of estimating an average treatment effect — with the general principles being similar for other parameters.

Overview of topics of course

We need a certain understand of relevant mathematical/statistical topics.

- Really foundational to understanding TMLE.
- On all levels this theory gives good insight, and it is needed to 'open the black box' of TMLE.

On the other hand, this is not a "math course".

- Theoretical background is quite deep, but too much for this course.
- (I will be pretty handwavy at times).

I will say many basic things.

- For the larger part, we focus on the simple example of estimating an average treatment effect — with the general principles being similar for other parameters.
- For many (biostatistical) applications, it gets more interesting when dealing with time-varying settings.

Overview of topics of course

Please give feedback ☺

Overview of topics of course

"Targeted learning"

- defining a (low-dimensional) (causal) target parameter to answer a specific scientific question.
- focus the statistical estimation procedure for estimation of that parameter specifically ... incorporating tools from nonparametric efficiency theory.

"Targeted minimum loss-based estimation (TMLE)"

- a particular tool for estimation.
- machine learning based substitution estimation.

We are interested in both.

(And it is hard to discuss one without the other).

Overview of topics of course

Across the days, we will move back and forth between theory and application.[1]

<u>Day 1:</u>

- targeted learning roadmap
- defining a (causal) parameter
- estimation, double robust estimation

<u>Day 2:</u>

- introduction to TMLE
- Targeting
- Super learning

<u>Day 3:</u>

- revisiting and broadening the theoretical basis
- bias/variance trade-off
- causal parameters in time-varying settings

<u>Day 4:</u>

- time-dependent confounding
- estimation in time-varying settings
- longitudinal TMLE

[1] Certain aspects and concepts will be repeated . . . multiple times.

Structure of the course

- Each day from 9:00 – 15:00.
- There is an extra hour in the morning for catching up; additionally, I am present in the room from 8:30–9:00, if you have any questions.
- Each day consists of lectures and practical exercises (mostly in `R`).
- There is not a sharp time-plan. Lessons take the time they require.
- You may not have time to finish all exercises during class, but all solutions are provided, and you can use the extra hour in the morning to catch up if you wish to do so.

Intended learning outcomes

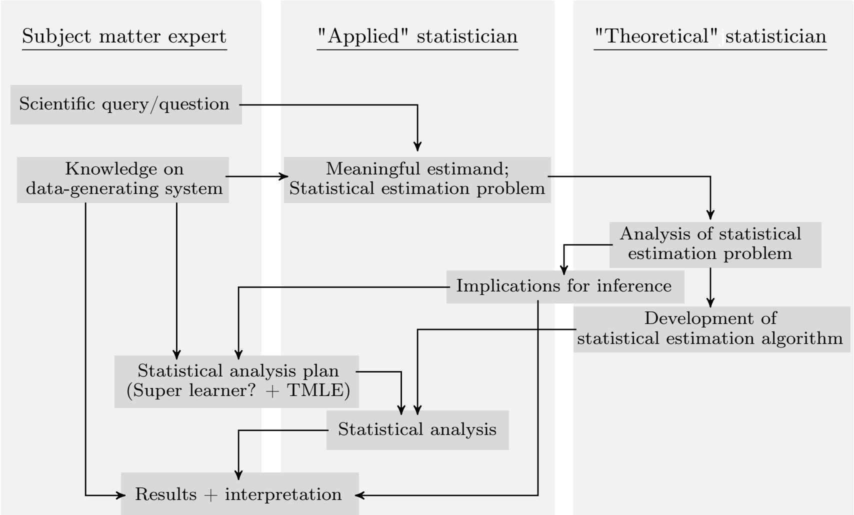
From the course description:

- *"The main focus of the course is to understand the overall concept, the theory, and the application of TMLE."*

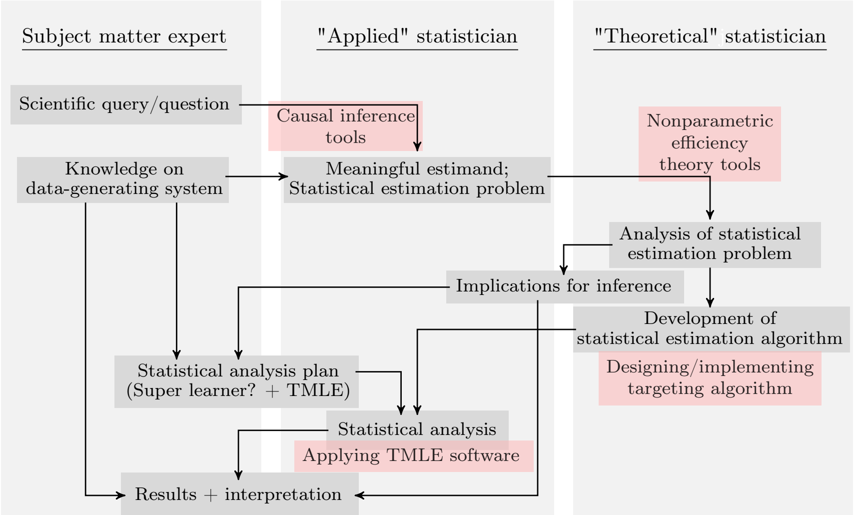
From the learning objectives:

- *"Explain the fundamental principles of statistical inference using TMLE and its application as a general framework for estimation of causal effects."*

Intended learning outcomes



Intended learning outcomes



Intended learning outcomes

My guess is that each of you have unique interests and distinct goals you aim to achieve from participating in this course ☺

- The theoretical basis of TMLE?
- Applying TMLE?
- The potential of TMLE?
- ...

Intended learning outcomes

My guess is that each of you have unique interests and distinct goals you aim to achieve from participating in this course ☺

- The theoretical basis of TMLE?
- Applying TMLE?
- The potential of TMLE?
- ...

What would you like to get out of the course?

- take 3 minutes to write down 1–5 sentences;
  (in a place where you can find them again).
- discuss briefly with the person sitting next to you.

If needed, the course description can be found here:
`https://phdcourses.ku.dk/detailkursus.aspx?id=110661&sitepath=SUND`.

Day 1, Lecture 1

Introduction: The roadmap of targeted learning

Overall statistical paradigm that TMLE is based on

Overview of today

Before lunch (9 – 12):

- Introduction to the roadmap of targeted learning.
- Brief introduction to causal inference.
- Estimation and double robust estimation.

* alignment with respect to "basic" (causal inference) concepts.
* introduction to critical notation.
* observed and counterfactual data simulation in `R`.
* simple application of software.

Overview of today

After lunch (13 – 15):

- Key theoretical concepts in analyzing an estimation problem.
- Construction of asymptotically linear estimation based on the efficient influence curve.
- The average treatment effect (ATE) as a concrete example.

* overall conditions for validity of (nonparametric) inference based on the efficient influence curve.
* Our focus today is practical: why this matters for understanding TMLE.

The roadmap of targeted learning

Theoretical angle The roadmap of targeted learning

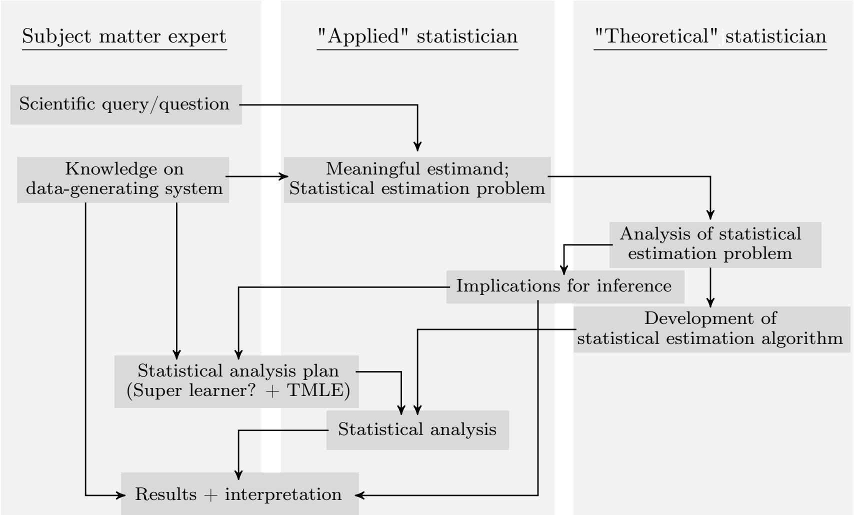
- data as a random variable having a probability distribution, scientific knowledge represented by a large statistical model, a statistical target parameter representing an answer to the question of interest.

Applied angle The roadmap of targeted learning / causal inference

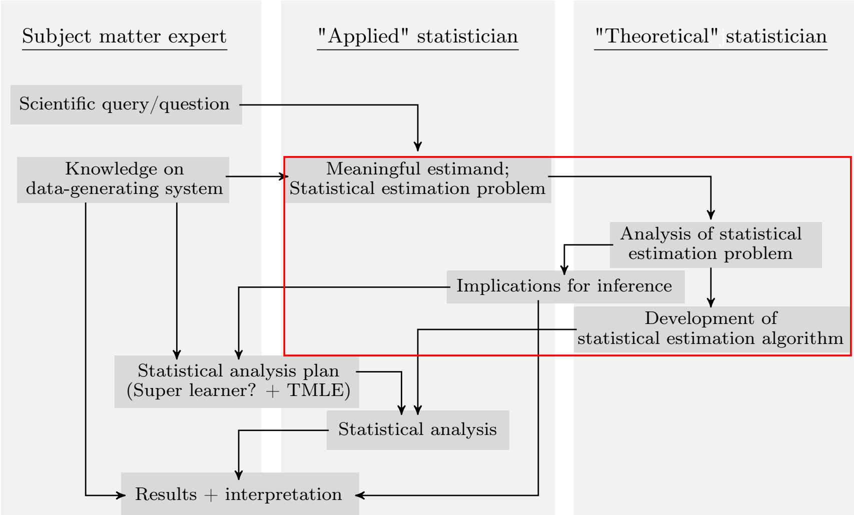
- translation from real-world data applications to a mathematical and statistical formulation of the relevant estimation problem.
- statistical analysis tailored towards answering that question.

Opposed to choosing a parametric model for the data-generating process and using that model to answer all questions.

The roadmap of targeted learning



The roadmap of targeted learning



The roadmap (theoretical)

1. Data is a random variable O with a probability distribution P_0
2. P_0 belongs to a statistical model $\mathcal{M}$
3. Our target is a parameter $\Psi : \mathcal{M} \to \mathbb{R}$
4. Construct estimator $\hat{P}_n$ for (relevant part of) P_0 and estimate the target parameter by $\hat{\psi}_n = \Psi(\hat{P}_n)$
5. Quantify uncertainty for the estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$

The roadmap (theoretical)

1. Data is a random variable O with a probability distribution P_0

$O_1, \ldots, O_n \overset{iid}{\sim} P_0$

O_i is the observation for individual i of the dataset

For example, O consists of

- Covariates: $X \in \mathcal{X} \subseteq \mathbb{R}^d$
- Exposure/treatment: $A \in \{0, 1\}$
- Outcome: $Y \in \{0, 1\}$ or $Y \in \mathbb{R}$

The roadmap (theoretical)

2. P_0 belongs to a statistical model $\mathcal{M}$

What do we know about the probability distribution of the data?

The statistical model $\mathcal{M}$ is the set of all probability distributions that we believe are possible for our observed data.

Limited statistical knowledge? $\Rightarrow \mathcal{M}$ should be large to reflect that.

The roadmap (theoretical)

Consider a parametric[2] model for the distribution of $Y \in \{0,1\}$ given $X \in \mathbb{R}^d$ and $A \in \{0,1\}$:

[2] i.e., distribution can be characterized by a finite number of parameters.

The roadmap (theoretical)

Consider a parametric[2] model for the distribution of $Y \in \{0, 1\}$ given $X \in \mathbb{R}^d$ and $A \in \{0, 1\}$:

$$\operatorname{logit} \mathbb{E}[Y \mid A, X] = \alpha_0 + \alpha_A A + \alpha_X^\top X \tag{M1}$$

- assumption of convenience?

[2] i.e., distribution can be characterized by a finite number of parameters.

The roadmap (theoretical)

Consider a parametric[2] model for the distribution of $Y \in \{0,1\}$ given $X \in \mathbb{R}^d$ and $A \in \{0,1\}$:

$$\operatorname{logit} \mathbb{E}[Y \mid A, X] = \alpha_0 + \alpha_A A + \alpha_X^\top X \tag{M1}$$

- assumption of convenience?

Another parametric model could be

$$\operatorname{logit} \mathbb{E}[Y \mid A, X] = \gamma_0 + \gamma_A A + \gamma_X^\top X + \gamma_{A,X}^\top AX \tag{M2}$$

- (M1) and (M2) cannot be true at the same time (except if $\gamma_{A,X} = 0$).

[2] i.e., distribution can be characterized by a finite number of parameters.

The roadmap (theoretical)

EXAMPLE:

- $O = (X, A, Y) \in [-2, 2] \times \{0, 1\} \times \{0, 1\}$
- True model is

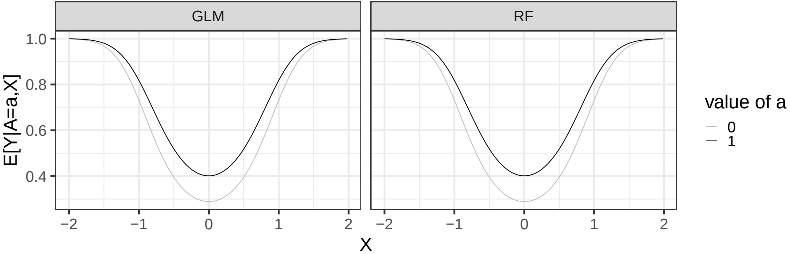
$$\operatorname{logit} \mathbb{E}[Y \mid A, X] = \beta_0 + \beta_A A + \beta_X X^2$$

The roadmap (theoretical)

EXAMPLE:

- $O = (X, A, Y) \in [-2,2] \times \{0,1\} \times \{0,1\}$
- True model is

$$\operatorname{logit} \mathbb{E}[Y \mid A, X] = \beta_0 + \beta_A A + \beta_X X^2$$



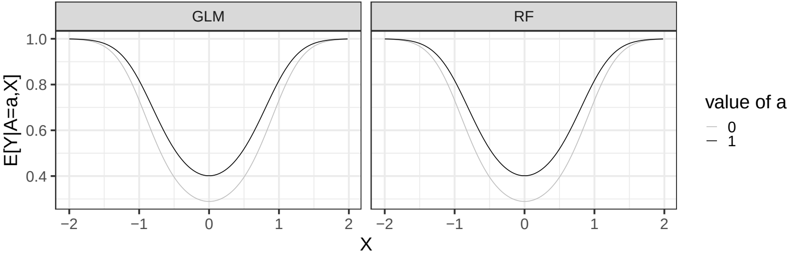
[Truth shown with solid lines]

The roadmap (theoretical)

EXAMPLE:

- $O = (X, A, Y) \in [-2, 2] \times \{0, 1\} \times \{0, 1\}$
- True model is

$$\operatorname{logit} \mathbb{E}[\, Y \mid A, X \,] = \beta_0 + \beta_A A + \beta_X X^2$$



[Truth shown with solid lines]

GLM: $\operatorname{logit} \mathbb{E}[\, Y \mid A, X \,] = \alpha_0 + \alpha_A A + \alpha_X X$

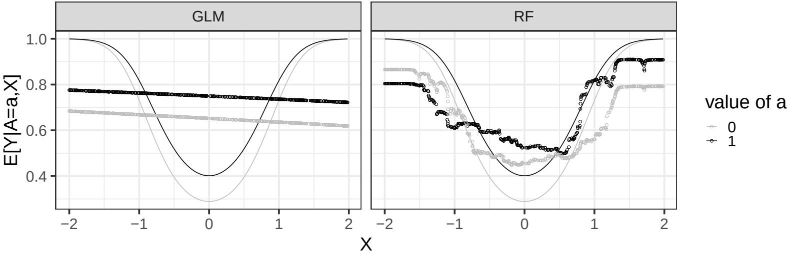
RF: Random forest (untuned)

The roadmap (theoretical)

EXAMPLE:

- $O = (X, A, Y) \in [-2, 2] \times \{0, 1\} \times \{0, 1\}$
- True model is

$$\operatorname{logit} \mathbb{E}[\, Y \mid A, X\,] = \beta_0 + \beta_A A + \beta_X X^2$$



[Truth shown with solid lines]

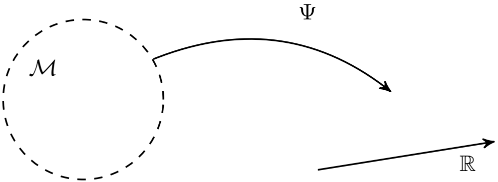
GLM: $\operatorname{logit} \mathbb{E}[\, Y \mid A, X\,] = \alpha_0 + \alpha_A A + \alpha_X X$

RF: Random forest (untuned)

The roadmap (theoretical)

3. Our target is a parameter (a functional) $\Psi : \mathcal{M} \to \mathbb{R}$

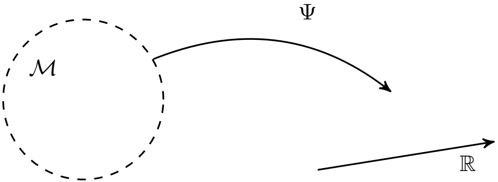
What are we trying to learn from the data?



The roadmap (theoretical)

3. Our target is a parameter (a functional) $\Psi : \mathcal{M} \to \mathbb{R}$

What are we trying to learn from the data?



EXAMPLE: Average treatment effect (ATE)

- $O = (X, A, Y) \in \mathbb{R}^d \times \{0,1\} \times \{0,1\}$
- The ATE is defined for $P \in \mathcal{M}$ as

$$\Psi(P) = \mathbb{E}_P[\mathbb{E}_P[Y \mid A = 1, X] - \mathbb{E}_P[Y \mid A = 0, X]]$$

The roadmap (theoretical)

EXAMPLE: Average treatment effect (ATE)

- $O = (X, A, Y) \in \mathbb{R}^d \times \{0,1\} \times \{0,1\}$
- The ATE is defined for $P \in \mathcal{M}$ as

$$\Psi(P) = \mathbb{E}_P[\mathbb{E}_P[Y \mid A = 1, X] - \mathbb{E}_P[Y \mid A = 0, X]]$$

The ATE can also be written, for $P \in \mathcal{M}$:

$$\Psi(P) = \tilde{\Psi}(\mu_X, f) = \int_{\mathbb{R}} \big(f(1,x) - f(0,x)\big) d\mu_X(x),$$

where $f(a,x) := \mathbb{E}_P[Y \mid A = a, X = x]$ and μ_X is the marginal distribution of X

f, μ_X are called *nuisance parameters*

The roadmap (theoretical)

This suggests a straightforward two-step estimation strategy:

1. estimate the nuisance parameters
2. plug estimates into the expression for the target parameter

A straightforward estimate of the ATE would be

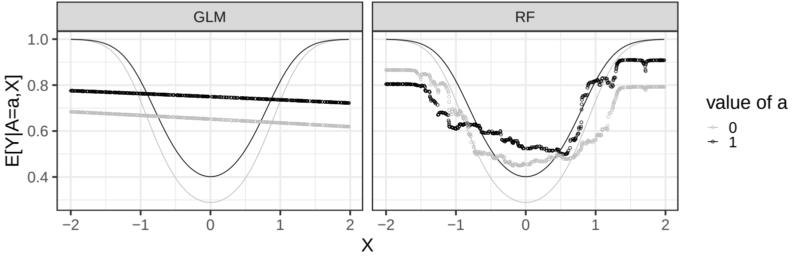
$$\hat{\psi}_n^{\mathrm{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}_n(1, X_i) - \hat{f}_n(0, X_i)\right\}$$

where $\hat{f}_n$ denotes some estimator for $f(a, x) = \mathbb{E}_P[Y \mid A = a, X = x]$

→ logistic regression, random forest, neural network, lasso, ...

The roadmap (theoretical)

In the previous example we had two different estimators for
$f(a,x) = \mathbb{E}_P[Y \mid A = a, X = x]$



$$\hat{\psi}_n^{\text{ATE,GLM}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}_n^{\text{GLM}}(1,X_i) - \hat{f}_n^{\text{GLM}}(0,X_i)\right\} = 0.0975$$

$$\hat{\psi}_n^{\text{ATE,RF}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}_n^{\text{RF}}(1,X_i) - \hat{f}_n^{\text{RF}}(0,X_i)\right\} = 0.0551$$

The roadmap (theoretical)

Contrast this to fitting a logistic regression model

$$\operatorname{logit} \mathbb{E}[\, Y \mid A, X\,] = \beta_0 + \beta_A A + \beta_X^\top X \tag{1}$$

to estimate the conditional odds ratio $\exp(\beta_A)$

- valid interpretation when model is correct
- statistical inference when model is correct
- *conditional* interpretation (crude and adjusted models target different parameters)

The roadmap (theoretical)

Contrast this to fitting a logistic regression model

$$\operatorname{logit} \mathbb{E}[\, Y \mid A, X\,] = \beta_0 + \beta_A A + \beta_X^\top X \tag{1}$$

to estimate the conditional odds ratio $\exp(\beta_A)$

- valid interpretation when model is correct
- statistical inference when model is correct
- *conditional* interpretation (crude and adjusted models target different parameters)

... and: (1) must be a priori specified (the same data cannot be used for testing and for fitting the final model).

The roadmap (theoretical)

4. Construct estimator $\hat{P}_n$ for (relevant part of) P_0 and estimate the target parameter by $\hat{\psi}_n = \Psi(\hat{P}_n)$

A priori specified algorithm that maps the data to an estimate in the parameter space for the target parameter

- a pre-specified logistic regression model
- a random forest
- cross-validated selection between a pre-specified library of different models ("super learning")

The roadmap (theoretical)

4. Construct estimator $\hat{P}_n$ for (relevant part of) P_0 and estimate the target parameter by $\hat{\psi}_n = \Psi(\hat{P}_n)$

A priori specified algorithm that maps the data to an estimate in the parameter space for the target parameter

- a pre-specified logistic regression model
- a random forest
- cross-validated selection between a pre-specified library of different models ("super learning")

+ "targeting" to yield the an estimator with improved properties

The roadmap (theoretical)

4. Construct estimator $\hat{P}_n$ for (relevant part of) P_0 and estimate the target parameter by $\hat{\psi}_n = \Psi(\hat{P}_n)$

A priori specified algorithm that maps the data to an estimate in the parameter space for the target parameter

"Initial estimation":

- a pre-specified logistic regression model
- a random forest
- cross-validated selection between a pre-specified library of different models ("super learning")

+ "targeting" to yield the an estimator with improved properties

The roadmap (theoretical)

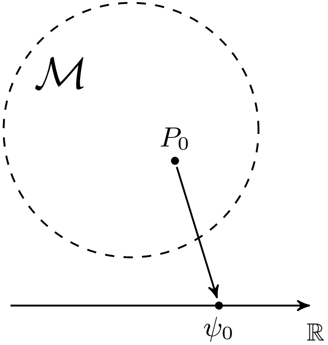
Estimation paradigm

1. P_0 is assumed to belong to a nonparametric model $\mathcal{M}$
2. Construction of $\sqrt{n}$-consistent and asymptotically linear estimation of $\psi_0 = \Psi(P_0)$ based the efficient influence function.

The roadmap (theoretical)

Estimation paradigm

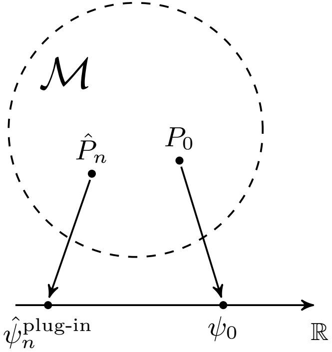
1. P_0 is assumed to belong to a nonparametric model $\mathcal{M}$
2. Construction of $\sqrt{n}$-consistent and asymptotically linear estimation of $\psi_0 = \Psi(P_0)$ based the efficient influence function.



The roadmap (theoretical)

Estimation paradigm

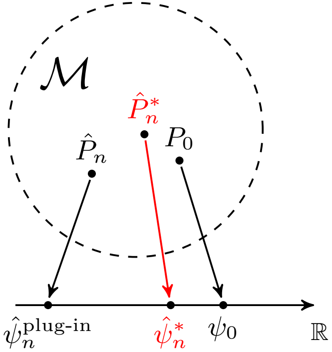
1. P_0 is assumed to belong to a nonparametric model $\mathcal{M}$
2. Construction of $\sqrt{n}$-consistent and asymptotically linear estimation of $\psi_0 = \Psi(P_0)$ based the efficient influence function.



The roadmap (theoretical)

Estimation paradigm

1. P_0 is assumed to belong to a nonparametric model $\mathcal{M}$
2. Construction of $\sqrt{n}$-consistent and asymptotically linear estimation of $\psi_0 = \Psi(P_0)$ based the efficient influence function.



Tools from semiparametric efficiency theory and empirical process theory tell us how conditions required for 2.

The roadmap (theoretical)

5. Quantify uncertainty for the estimator $\hat{\psi}_n = \Psi(\hat{P}_n)$

If we repeat the experiment of drawing n observations we would every time end up with a different realization of our estimator.

Across the repetitions, the estimator has a sampling distribution that we wish to quantify.

Under some conditions, we may use the asymptotic distribution

$$\hat{\psi}_n \overset{as}{\sim} N(\psi_0, \sigma^2/n)$$

to provide statistical inference.

The roadmap of targeted learning

Theoretical angle The roadmap of targeted learning

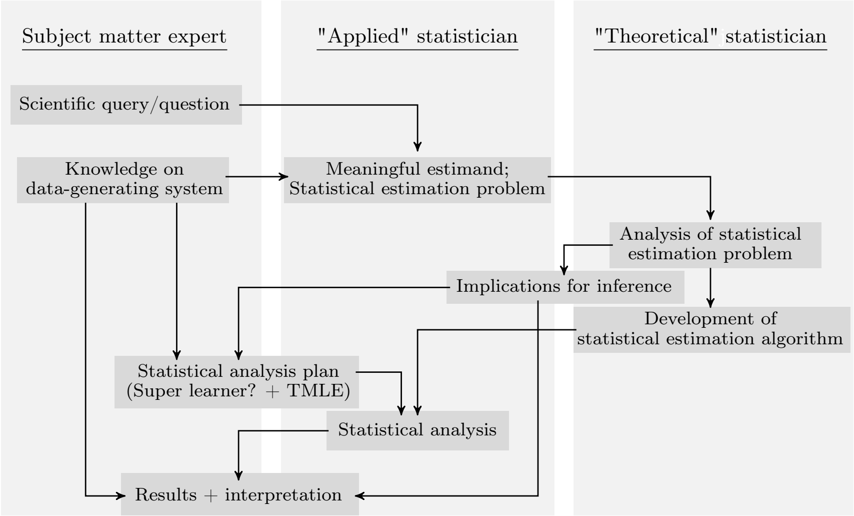
- data as a random variable having a probability distribution, scientific knowledge represented by a large statistical model, a statistical target parameter representing an answer to the question of interest.

Applied angle The roadmap of targeted learning / causal inference

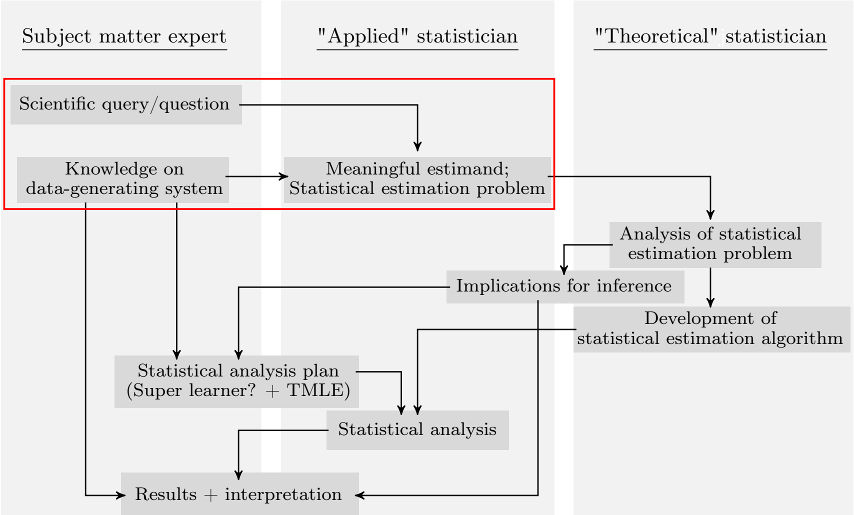
- translation from real-world data applications to a mathematical and statistical formulation of the relevant estimation problem.
- statistical analysis tailored towards answering that question.

Opposed to choosing a parametric model for the data-generating process and using that model to answer all questions.

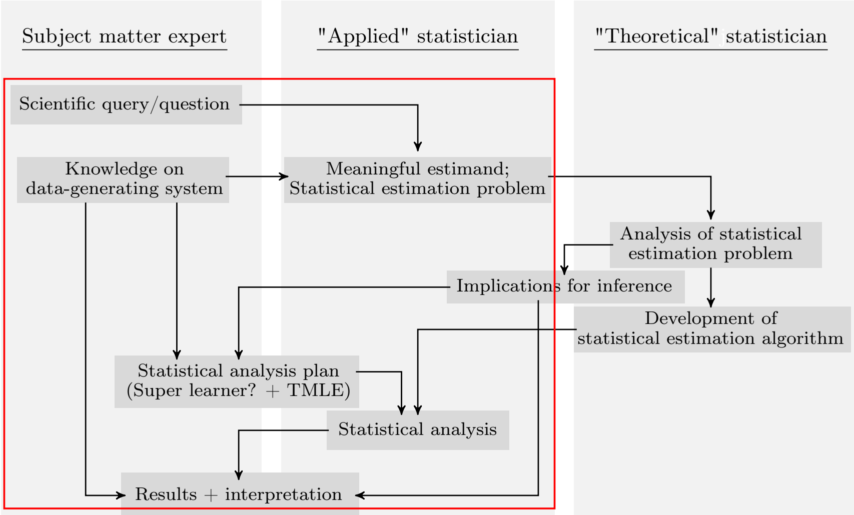
The roadmap (applied)



The roadmap (applied)



The roadmap (applied)



The roadmap (applied)

1. Observed data
2. Causal model
3. Causal question and target causal estimand
4. Identifiability
5. Stating the statistical estimation problem
6. Estimate
7. Interpret results

The roadmap (applied)

1. Observed data
2. Causal model
3. Causal question and target causal estimand
4. Identifiability
5. Stating the statistical estimation problem
6. Estimate
7. Interpret results

... putting things into the right boxes.

... make the statistical analysis about the targeted scientific question (and not the other way around).

... focus on statistical parameters that have a meaningful interpretation.

The roadmap (applied)

A formal causal framework can help us[3]

- designing a statistical analysis that come as close as possible to answering scientific/causal questions.
- understand how far away from a causal conclusion we may be.

[3] The output of the analysis is not causal just because we use causal inference methods.

The roadmap (applied)

A formal causal framework can help us[3]

- designing a statistical analysis that come as close as possible to answering scientific/causal questions.
- understand how far away from a causal conclusion we may be.

Clearly defining what an EFFECT is and WHAT effect we are interested in

[3] The output of the analysis is not causal just because we use causal inference methods.

The roadmap (applied)

A formal causal framework can help us[3]

- designing a statistical analysis that come as close as possible to answering scientific/causal questions.
- understand how far away from a causal conclusion we may be.

Clearly defining what an EFFECT is and WHAT effect we are interested in

- this gets even more relevant when we deal with time-varying settings.

[3] The output of the analysis is not causal just because we use causal inference methods.

The roadmap (applied)

1. Observed data — $O = (X, A, Y)$
2. Causal model — what we know/believe/assume about directions of effects
3. Causal question and target causal estimand — formulating the scientific question as a contrast between counterfactual outcomes (e.g., in terms of ideal hypothetical experiment)
4. Identifiability — is data sufficient to estimate the causal effect?

The roadmap (applied)

1. Observed data — $O = (X, A, Y)$
2. Causal model — what we know/believe/assume about directions of effects
3. Causal question and target causal estimand — formulating the scientific question as a contrast between counterfactual outcomes (e.g., in terms of ideal hypothetical experiment)
4. Identifiability — is data sufficient to estimate the causal effect?

This is the topic of the next lecture.

Summarizing comments

Statistical theory for parametric models

- meant for settings where the model is known a priori
  - the model is rarely known a priori
  - theory does not reflect how data are in fact analyzed (e.g., due to use of model selection strategies)
- the model is chosen for its simplicity and convenience
  - simple summary measures of associations

Targeted learning paradigm

- translating scientific question into predefined model-free target parameter
- machine learning based estimators can be constructed and still combined with valid/honest inference (allowing full prespecification of the statistical analysis)