Statistical vs. Machine Learning approaches
– a debate

Antoine Chambaz

MAP5 (UMR CNRS 8145), Université Paris Cité

September 21st, 2022

Statistics & Biopharmacy conference

“when Machine Learning meets Statistics

for drug development and evaluation”

Setting the scene

We are gathered to debate on the theme

Statistical vs. Machine Learning approaches

Let's set the scene

Common ground (1/6)

Common ground (1/6)

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

Common ground (1/6)

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$
- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$

Common ground (1/6)

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$
- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$

Common ground (1/6)

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$
- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

Common ground (1/6)

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$
- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$
  - a loss function ℓ and the risk R_P^ℓ it induces

Common ground (1/6)

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$
- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$
  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (2/6)

Example: regression

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$
  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (2/6)

Example: regression

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (X_i, Y_i)$, $Y_i \in \mathbb{R}$

  $O_1, \ldots, O_n$ i.i.d

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$
  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (2/6)

Example: regression

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (X_i, Y_i)$, $Y_i \in \mathbb{R}$

  $O_1, \ldots, O_n$ i.i.d

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$

- scientific question $\rightsquigarrow$

  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    $\Psi(P')(X) = E_{P'}(Y|X)$

  - a loss function ℓ and the risk R_P^ℓ it induces

  - a specific objective

Common ground (2/6)

Example: regression

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (X_i, Y_i)$, $Y_i \in \mathbb{R}$

  $O_1, \ldots, O_n$ i.i.d

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$

- scientific question $\rightsquigarrow$

  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    $\Psi(P')(X) = E_{P'}(Y|X)$

  - a loss function ℓ and the risk R_P^ℓ it induces

    $\ell(f)(x, y) = (y - f(x))^2$

  - a specific objective

Common ground (2/6)

Example: regression

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (X_i, Y_i)$, $Y_i \in \mathbb{R}$

  $O_1, \ldots, O_n$ i.i.d

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$

- scientific question $\rightsquigarrow$

  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    $\Psi(P')(X) = E_{P'}(Y|X)$

  - a loss function ℓ and the risk R_P^ℓ it induces

    $\ell(f)(x, y) = (y - f(x))^2$

  - a specific objective

    learn $\Psi(P)$

Common ground (3/6)

Example: classification

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$
  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (3/6)

Example: classification

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (X_i, Y_i)$, $Y_i \in \{0, 1\}$

  $O_1, \ldots, O_n$ i.i.d

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$
  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (3/6)

Example: classification

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (X_i, Y_i)$, $Y_i \in \{0, 1\}$

  $O_1, \ldots, O_n$ i.i.d

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    $\Psi(P')(X) = P'(Y = 1|X)$

  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (3/6)

Example: classification

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \dots, O_n)$, $O_i = (X_i, Y_i)$, $Y_i \in \{0, 1\}$

  $O_1, \dots, O_n$ i.i.d

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$

- scientific question $\rightsquigarrow$

  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    $\Psi(P')(X) = P'(Y = 1|X)$

  - a loss function ℓ and the risk R_P^ℓ it induces

    $\ell(f)(x, y) = -y \log f(x) - (1 - y) \log(1 - f(x))$

  - a specific objective

Common ground (3/6)

Example: classification

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (X_i, Y_i)$, $Y_i \in \{0, 1\}$

  $O_1, \ldots, O_n$ i.i.d
- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    $\Psi(P')(X) = P'(Y = 1|X)$
  - a loss function ℓ and the risk R_P^ℓ it induces

    $\ell(f)(x, y) = -y \log f(x) - (1 - y) \log(1 - f(x))$
  - a specific objective

    learn $\Psi(P)$

Common ground (4/6)

Example: average treatment effect

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$
  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (4/6)

Example: average treatment effect

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (W_i, A_i, Y_i)$, $A_i \in \{0, 1\}$ (binary treatment)

  $O_1, \ldots, O_n$ i.i.d

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$
  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (4/6)

Example: average treatment effect

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (W_i, A_i, Y_i)$, $A_i \in \{0, 1\}$ (binary treatment)

  $O_1, \ldots, O_n$ i.i.d
- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    $\Psi(P') = E_{P'}\left[E_{P'}(Y|A=1, W) - E_{P'}(Y|A=0, W)\right]$
  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (4/6)

Example: average treatment effect

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (W_i, A_i, Y_i)$, $A_i \in \{0, 1\}$ (binary treatment)

  $O_1, \ldots, O_n$ i.i.d

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$

- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    $\Psi(P') = E_{P'} \left[E_{P'}(Y|A = 1, W) - E_{P'}(Y|A = 0, W) \right]$

  - a loss function ℓ and the risk R_P^ℓ it induces

    ℓ chosen to target the relevant nuisance parameters...

  - a specific objective

Common ground (4/6)

Example: average treatment effect

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (W_i, A_i, Y_i)$, $A_i \in \{0, 1\}$ (binary treatment)

  $O_1, \ldots, O_n$ i.i.d
- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    $\Psi(P') = E_{P'}\left[E_{P'}(Y|A=1, W) - E_{P'}(Y|A=0, W)\right]$
  - a loss function ℓ and the risk R_P^ℓ it induces

    ℓ chosen to target the relevant nuisance parameters...
  - a specific objective

    estimate $\Psi(P)$, build confidence interval

Common ground (5/6)

Example: optimal recommendation rule

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$
  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (5/6)

Example: optimal recommendation rule

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (W_i, A_i, Y_i)$, $A_i \in \{0, 1\}$ (binary treatment)

  $O_1, \ldots, O_n$ drawn sequentially from known adaptive design $(g_1, \ldots, g_n)$

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$

- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$
  - a loss function ℓ and the risk R_P^ℓ it induces
  - a specific objective

Common ground (5/6)

Example: optimal recommendation rule

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (W_i, A_i, Y_i)$, $A_i \in \{0, 1\}$ (binary treatment)

  $O_1, \ldots, O_n$ drawn sequentially from known adaptive design $(g_1, \ldots, g_n)$

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$

- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    given a class $\mathcal{F}$ of candidate rules

    $$\Psi(P') = \arg\max_{f \in \mathcal{F}} \mathcal{V}_{P'}(f)$$

    where $\mathcal{V}_{P'}(f)$ is the **value of f** under P'

  - a loss function ℓ and the risk R_P^ℓ it induces

  - a specific objective

Common ground (5/6)

Example: optimal recommendation rule

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (W_i, A_i, Y_i)$, $A_i \in \{0, 1\}$ (binary treatment)

  $O_1, \ldots, O_n$ drawn sequentially from known adaptive design $(g_1, \ldots, g_n)$

- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$

- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    given a class $\mathcal{F}$ of candidate rules

    $$\Psi(P') = \arg\max_{f \in \mathcal{F}} \mathcal{V}_{P'}(f)$$

    where $\mathcal{V}_{P'}(f)$ is the **value of** f under P'
  - a loss function ℓ and the risk R_P^ℓ it induces

    ℓ chosen to target the relevant nuisance parameters. . .
  - a specific objective

Common ground (5/6)

Example: optimal recommendation rule

- experiment of interest, $P \rightsquigarrow$ observation $O \sim P$

  $O = (O_1, \ldots, O_n)$, $O_i = (W_i, A_i, Y_i)$, $A_i \in \{0, 1\}$ (binary treatment)

  $O_1, \ldots, O_n$ drawn sequentially from known adaptive design $(g_1, \ldots, g_n)$
- expert knowledge $\rightsquigarrow$ statistical model $\mathcal{M} \ni P$
- scientific question $\rightsquigarrow$
  - a statistical parameter $\Psi : \mathcal{M} \to \Theta$

    given a class $\mathcal{F}$ of candidate rules

    $$\Psi(P') = \arg\max_{f \in \mathcal{F}} \mathcal{V}_{P'}(f)$$

    where $\mathcal{V}_{P'}(f)$ is the **value of f** under P'
  - a loss function ℓ and the risk R_P^ℓ it induces

    ℓ chosen to target the relevant nuisance parameters...
  - a specific objective

    learn $\Psi(P)$

    possibly estimate and build a confidence interval for $\mathcal{V}_P(\Psi(P))$

Common ground (6/6)

In all these examples (and beyond)

Common ground (6/6)

In all these examples (and beyond)

- one needs to

Common ground (6/6)

In all these examples (and beyond)

- one needs to
  - introduce a class of functions/working model $\mathcal{F}$

Common ground (6/6)

In all these examples (and beyond)

- one needs to
  - introduce a class of functions/working model $\mathcal{F}$
  - explicitely or implicitely

Common ground (6/6)

In all these examples (and beyond)

- one needs to
  - introduce a class of functions/working model $\mathcal{F}$
  - explicitely or implicitely
  - solve

$$\arg\min_{f\in\mathcal{F}} \text{Criterion}(f; O) \tag{$\star$}$$

Common ground (6/6)

In all these examples (and beyond)

- one needs to
  - introduce a class of functions/working model $\mathcal{F}$
  - explicitely or implicitely
  - solve

$$\arg\min_{f \in \mathcal{F}} \text{Criterion}(f; O) \qquad (\star)$$

- three related key-questions:

Common ground (6/6)

In all these examples (and beyond)

- one needs to
  - introduce a class of functions/working model $\mathcal{F}$
  - explicitely or implicitely
  - solve

$$\arg\min_{f \in \mathcal{F}} \text{Criterion}(f; O) \qquad (\star)$$

- three related key-questions:
  - how big is $\mathcal{F}$?

Common ground (6/6)

In all these examples (and beyond)

- one needs to
  - introduce a class of functions/working model $\mathcal{F}$
  - explicitely or implicitely
  - solve

$$\arg\min_{f \in \mathcal{F}} \text{Criterion}(f; O) \qquad (\star)$$

- three related key-questions:
  - how big is $\mathcal{F}$?
  - how difficult is it to explore $\mathcal{F}$?

Common ground (6/6)

In all these examples (and beyond)

- one needs to
  - introduce a class of functions/working model $\mathcal{F}$
  - explicitely or implicitely
  - solve

$$\arg\min_{f \in \mathcal{F}} \text{Criterion}(f; O) \qquad (\star)$$

- three related key-questions:
  - how big is $\mathcal{F}$?
  - how difficult is it to explore $\mathcal{F}$?
  - how much information does one need to solve $(\star)$? [how big should O be?]

Statistical vs. Machine Learning approaches

Statistical approach

Machine Learning approach

Statistical vs. Machine Learning approaches

Statistical approach

- $\mathcal{F}$ small-to-middle-sized

Machine Learning approach

Statistical vs. Machine Learning approaches

Statistical approach

- $\mathcal{F}$ small-to-middle-sized
- $\mathcal{F}$ (rather) easily explorable

Machine Learning approach

Statistical vs. Machine Learning approaches

Statistical approach

- $\mathcal{F}$ small-to-middle-sized
- $\mathcal{F}$ (rather) easily explorable
- solving $(\star)$ requires a “reasonable” amount of information

Machine Learning approach

Statistical vs. Machine Learning approaches

Statistical approach

- $\mathcal{F}$ small-to-middle-sized
- $\mathcal{F}$ (rather) easily explorable
- solving $(\star)$ requires a "reasonable" amount of information

Machine Learning approach

- $\mathcal{F}$ big to huge

Statistical vs. Machine Learning approaches

Statistical approach

- $\mathcal{F}$ small-to-middle-sized
- $\mathcal{F}$ (rather) easily explorable
- solving $(\star)$ requires a "reasonable" amount of information

Machine Learning approach

- $\mathcal{F}$ big to huge
- $\mathcal{F}$ (not so) easily explorable

Statistical vs. Machine Learning approaches

Statistical approach

- $\mathcal{F}$ small-to-middle-sized
- $\mathcal{F}$ (rather) easily explorable
- solving $(\star)$ requires a "reasonable" amount of information

Machine Learning approach

- $\mathcal{F}$ big to huge
- $\mathcal{F}$ (not so) easily explorable
- solving $(\star)$ requires a "(very) large" amount of information

Measuring the size of $\mathcal{F}$

How to measure the size of $\mathcal{F}$?

Measuring the size of $\mathcal{F}$

How to measure the size of $\mathcal{F}$?

- $N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,P})$, the minimal number of $\|\cdot\|_{2,P}$-balls of radius ε required to cover $\mathcal{F}$

Measuring the size of $\mathcal{F}$

How to measure the size of $\mathcal{F}$?

- $N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,P})$, the minimal number of $\|\cdot\|_{2,P}$-balls of radius ε required to cover $\mathcal{F}$
- if $\text{card}(\mathcal{F})$ is not finite $\varepsilon \downarrow 0$ implies $N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,P}) \uparrow \infty$

Measuring the size of $\mathcal{F}$

How to measure the size of $\mathcal{F}$?

- $N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,P})$, the minimal number of $\|\cdot\|_{2,P}$-balls of radius ε required to cover $\mathcal{F}$
- if $\text{card}(\mathcal{F})$ is not finite $\varepsilon \downarrow 0$ implies $N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,P}) \uparrow \infty$
- if

$$I_P(\mathcal{F}) = \int_0^1 \sqrt{\log(1 + N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,P}))} d\varepsilon < \infty$$

  then $\mathcal{F}$ is **tamable**

Measuring the size of $\mathcal{F}$

How to measure the size of $\mathcal{F}$?

- $N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,P})$, the minimal number of $\|\cdot\|_{2,P}$-balls of radius ε required to cover $\mathcal{F}$
- if $\mathrm{card}(\mathcal{F})$ is not finite $\varepsilon \downarrow 0$ implies $N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,P}) \uparrow \infty$
- if
$$I_P(\mathcal{F}) = \int_0^1 \sqrt{\log(1 + N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,P}))} d\varepsilon < \infty$$
  then $\mathcal{F}$ is **tamable**
- otherwise, that depends on the objective...

A few questions

A few questions

1. How to define Machine Learning approaches in contrast with Statistical approaches?

A few questions

1. How to define Machine Learning approaches in contrast with Statistical approaches?
2. Is the answer to 1. technical or cultural?

   (vocabulary, topics of main interest; conceptual tools)

A few questions

1. How to define Machine Learning approaches in contrast with Statistical approaches?
2. Is the answer to 1. technical or cultural?

   (vocabulary, topics of main interest; conceptual tools)
3. If part of the answer to 1. is technical, is the size of $\mathcal{F}$ the major distinction?

   (Cambridge Dictionary: a distinction is "a difference between two similar things")

A few questions

1. How to define Machine Learning approaches in contrast with Statistical approaches?
2. Is the answer to 1. technical or cultural?

   (vocabulary, topics of main interest; conceptual tools)
3. If part of the answer to 1. is technical, is the size of $\mathcal{F}$ the major distinction?

   (Cambridge Dictionary: a distinction is "a difference between two similar things")
4. Is interpretability better guaranteed in Statistics than in Machine Learning?

   (or the other way around)

A few questions

1. How to define Machine Learning approaches in contrast with Statistical approaches?
2. Is the answer to 1. technical or cultural?

   (vocabulary, topics of main interest; conceptual tools)
3. If part of the answer to 1. is technical, is the size of $\mathcal{F}$ the major distinction?

   (Cambridge Dictionary: a distinction is "a difference between two similar things")
4. Is interpretability better guaranteed in Statistics than in Machine Learning?

   (or the other way around)
5. Is Machine Learning the future of Statistics?

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?

4. Is interpretability better guaranteed in Statistics than in Machine Learning? (or the other way around)

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?

4. Is interpretability better guaranteed in Statistics than in Machine Learning? (or the other way around)

5. Is Machine Learning the future of Statistics?

A few questions. . . and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?
   (a) one can do beautiful Statistics when $I_P(\mathcal{F})$ is **barely** finite

4. Is interpretability better guaranteed in Statistics than in Machine Learning? (or the other way around)

5. Is Machine Learning the future of Statistics?

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?
   (a) one can do beautiful Statistics when $I_P(\mathcal{F})$ is **barely** finite
   e.g. in examples 3 and 4

4. Is interpretability better guaranteed in Statistics than in Machine Learning? (or the other way around)

5. Is Machine Learning the future of Statistics?

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?
   (a) one can do beautiful Statistics when $I_P(\mathcal{F})$ is **barely** finite
   e.g. in examples 3 and 4
   (b) one even can do beautiful Statistics when $I_P(\mathcal{F})$ is infinite!

4. Is interpretability better guaranteed in Statistics than in Machine Learning?
   (or the other way around)

5. Is Machine Learning the future of Statistics?

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?
   (a) one can do beautiful Statistics when $I_P(\mathcal{F})$ is **barely** finite

   e.g. in examples 3 and 4

   (b) one even can do beautiful Statistics when $I_P(\mathcal{F})$ is infinite!

   e.g. in examples 3 and 4

4. Is interpretability better guaranteed in Statistics than in Machine Learning? (or the other way around)

5. Is Machine Learning the future of Statistics?

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?
   (a) one can do beautiful Statistics when $I_P(\mathcal{F})$ is **barely** finite
   e.g. in examples 3 and 4
   (b) one even can do beautiful Statistics when $I_P(\mathcal{F})$ is infinite!
   e.g. in examples 3 and 4
   (c) given $f^\star \in \mathcal{F}^{\mathrm{ML}}$, a feature built by Machine Learning, a statistician may use a small-to-middle-sized working model of the form $f^\star + \mathcal{F}$

4. Is interpretability better guaranteed in Statistics than in Machine Learning?
   (or the other way around)

5. Is Machine Learning the future of Statistics?

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?
   (a) one can do beautiful Statistics when $I_P(\mathcal{F})$ is **barely** finite
   e.g. in examples 3 and 4
   (b) one even can do beautiful Statistics when $I_P(\mathcal{F})$ is infinite!
   e.g. in examples 3 and 4
   (c) given $f^\star \in \mathcal{F}^{\mathrm{ML}}$, a feature built by Machine Learning, a statistician may use a small-to-middle-sized working model of the form $f^\star + \mathcal{F}$
   (d) super learning/aggregation/stacking is now routinely used in Statistics

4. Is interpretability better guaranteed in Statistics than in Machine Learning? (or the other way around)

5. Is Machine Learning the future of Statistics?

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?
   (a) one can do beautiful Statistics when $I_P(\mathcal{F})$ is **barely** finite
   e.g. in examples 3 and 4
   (b) one even can do beautiful Statistics when $I_P(\mathcal{F})$ is infinite!
   e.g. in examples 3 and 4
   (c) given $f^\star \in \mathcal{F}^{\mathrm{ML}}$, a feature built by Machine Learning, a statistician may use a small-to-middle-sized working model of the form $f^\star + \mathcal{F}$
   (d) super learning/aggregation/stacking is now routinely used in Statistics

4. Is interpretability better guaranteed in Statistics than in Machine Learning? (or the other way around)
   (e) not really/no...

5. Is Machine Learning the future of Statistics?

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?
   (a) one can do beautiful Statistics when $I_P(\mathcal{F})$ is **barely** finite
   e.g. in examples 3 and 4
   (b) one even can do beautiful Statistics when $I_P(\mathcal{F})$ is infinite!
   e.g. in examples 3 and 4
   (c) given $f^\star \in \mathcal{F}^{\mathrm{ML}}$, a feature built by Machine Learning, a statistician may use a small-to-middle-sized working model of the form $f^\star + \mathcal{F}$
   (d) super learning/aggregation/stacking is now routinely used in Statistics
4. Is interpretability better guaranteed in Statistics than in Machine Learning? (or the other way around)
   (e) not really/no... if interpretability is not addressed specifically
5. Is Machine Learning the future of Statistics?

A few questions… and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?
   (a) one can do beautiful Statistics when $I_P(\mathcal{F})$ is **barely** finite
   e.g. in examples 3 and 4
   (b) one even can do beautiful Statistics when $I_P(\mathcal{F})$ is infinite!
   e.g. in examples 3 and 4
   (c) given $f^\star \in \mathcal{F}^{\mathrm{ML}}$, a feature built by Machine Learning, a statistician may use a small-to-middle-sized working model of the form $f^\star + \mathcal{F}$
   (d) super learning/aggregation/stacking is now routinely used in Statistics

4. Is interpretability better guaranteed in Statistics than in Machine Learning? (or the other way around)
   (e) not really/no… if interpretability is not addressed specifically

5. Is Machine Learning the future of Statistics?
   (f) yes…

A few questions... and elements of answer

1. Is the frontier between Statistics and Machine Learning so clearly drawn?
   (a) one can do beautiful Statistics when $I_P(\mathcal{F})$ is **barely** finite
   e.g. in examples 3 and 4
   (b) one even can do beautiful Statistics when $I_P(\mathcal{F})$ is infinite!
   e.g. in examples 3 and 4
   (c) given $f^\star \in \mathcal{F}^{\mathrm{ML}}$, a feature built by Machine Learning, a statistician may use a small-to-middle-sized working model of the form $f^\star + \mathcal{F}$
   (d) super learning/aggregation/stacking is now routinely used in Statistics

4. Is interpretability better guaranteed in Statistics than in Machine Learning? (or the other way around)
   (e) not really/no... if interpretability is not addressed specifically

5. Is Machine Learning the future of Statistics?
   (f) yes... if Statistics is the future of Machine Learning!

What do **you** think?