Selection by Prediction

Prediction-assisted screening and discovery with conformal p-values

Ying Jin

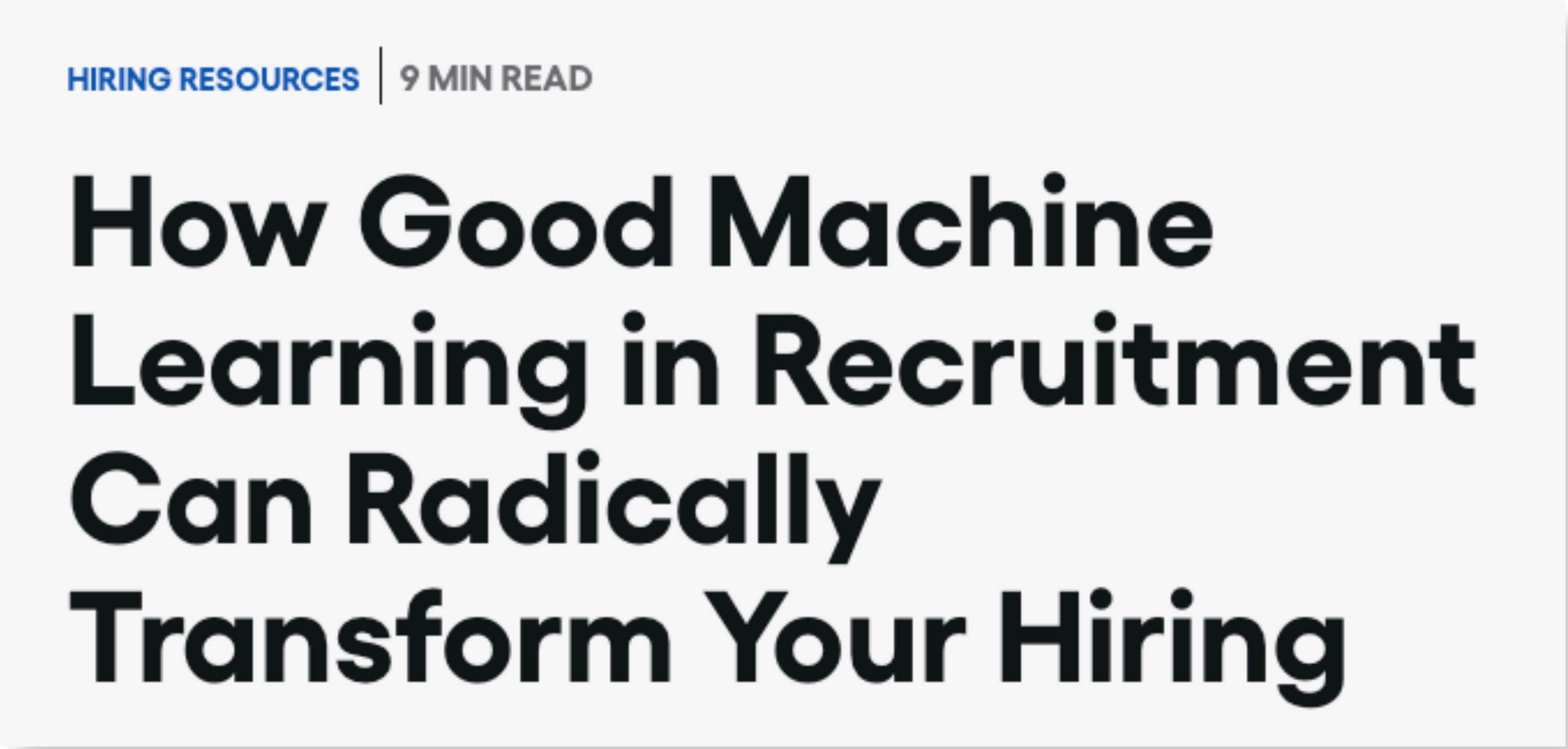
Department of Statistics

Stanford University

Joint work with Emmanuel Candès

One World Seminar on Mathematics of Machine Learning, April 19, 2023

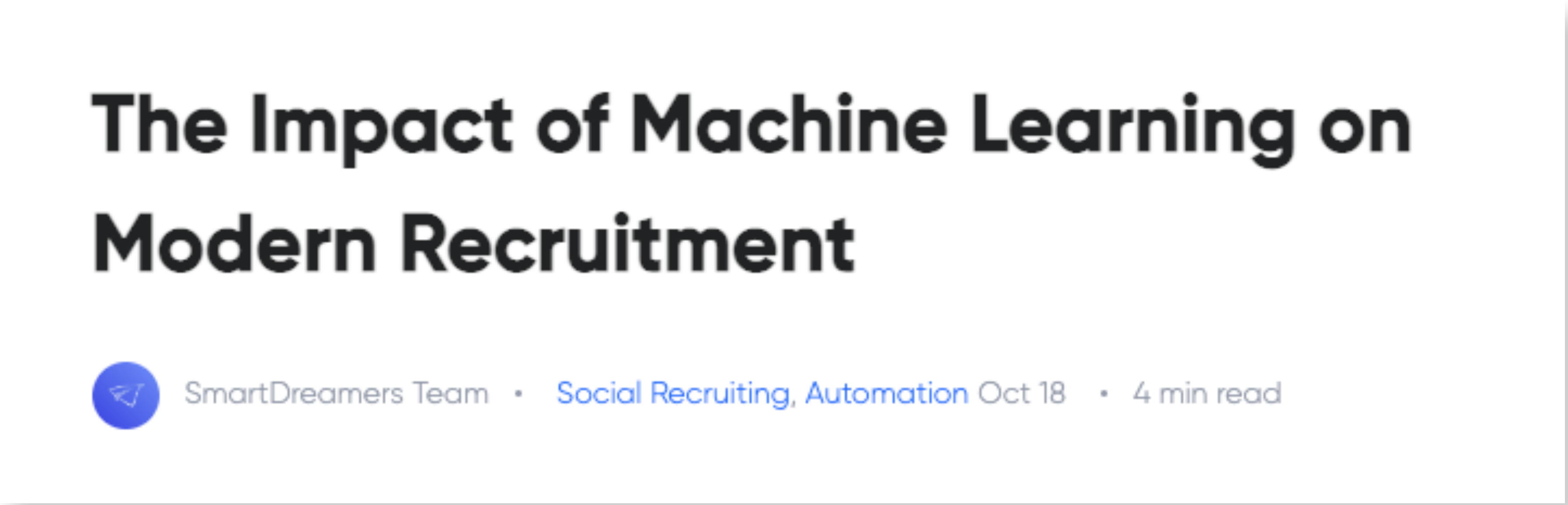
ML prediction assists decision



HIRING RESOURCES | 9 MIN READ

How Good Machine Learning in Recruitment Can Radically Transform Your Hiring

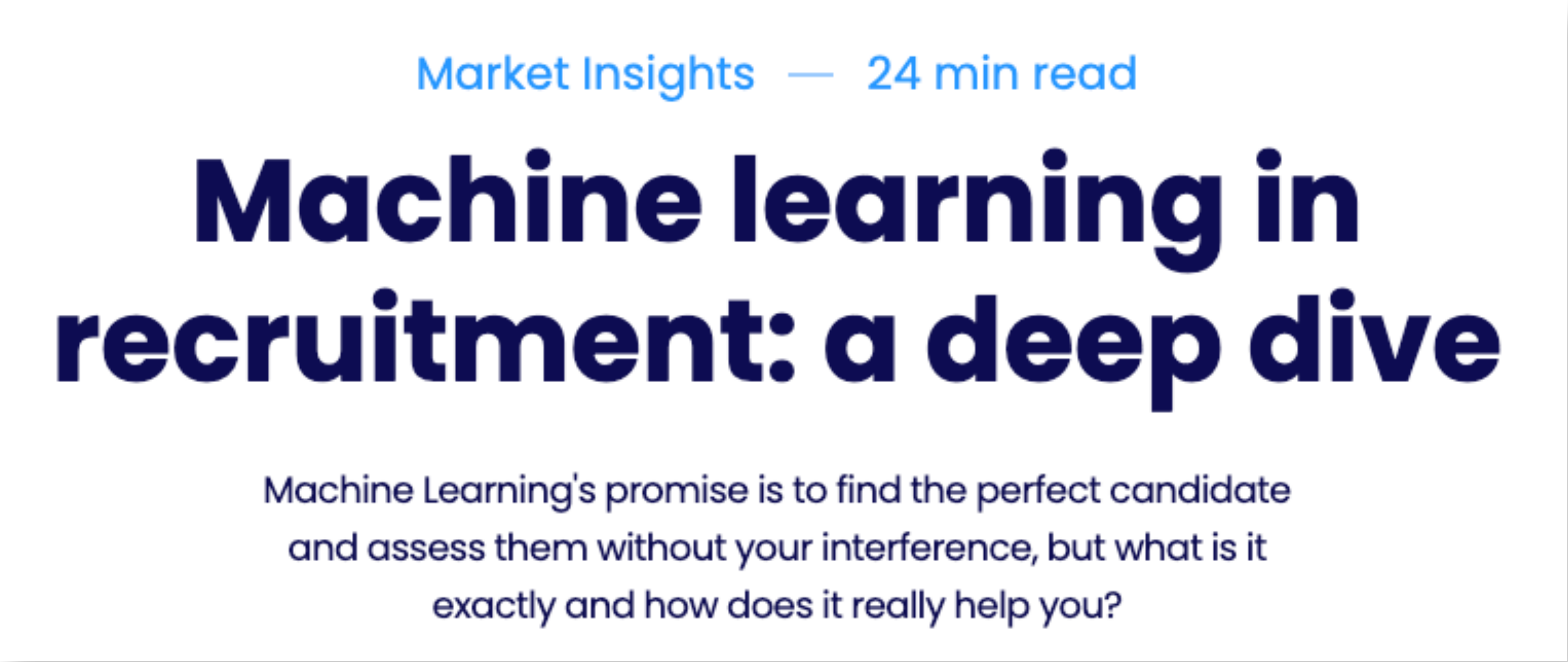
[VerVoe.com]



The Impact of Machine Learning on Modern Recruitment

SmartDreamers Team • Social Recruiting, Automation Oct 18 • 4 min read

[smartdreamers.com]



Market Insights — 24 min read

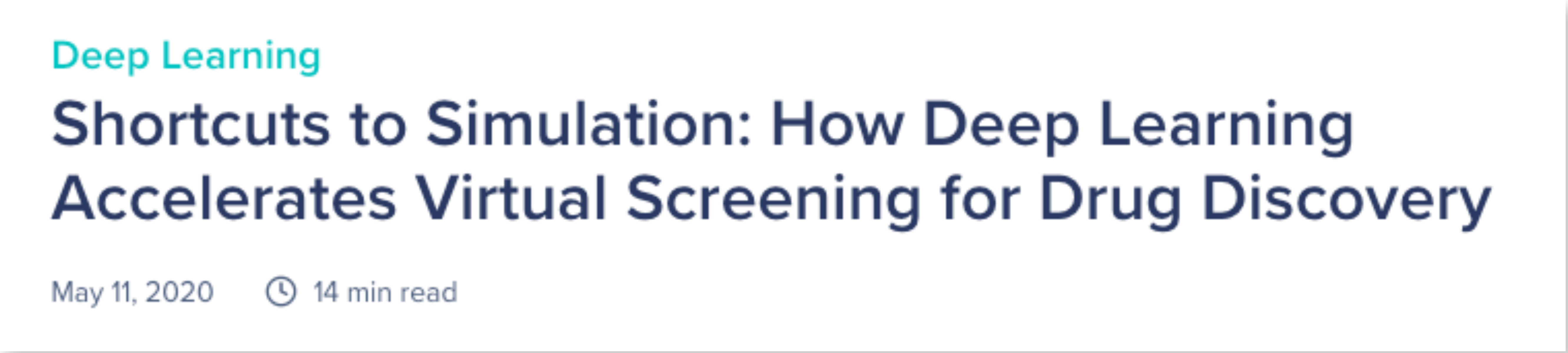
Machine learning in recruitment: a deep dive

Machine Learning's promise is to find the perfect candidate and assess them without your interference, but what is it exactly and how does it really help you?

[HeroHunt.ai]

Job hiring: Who to reach out to?
Who to proceed to interview?

ML prediction assists discovery

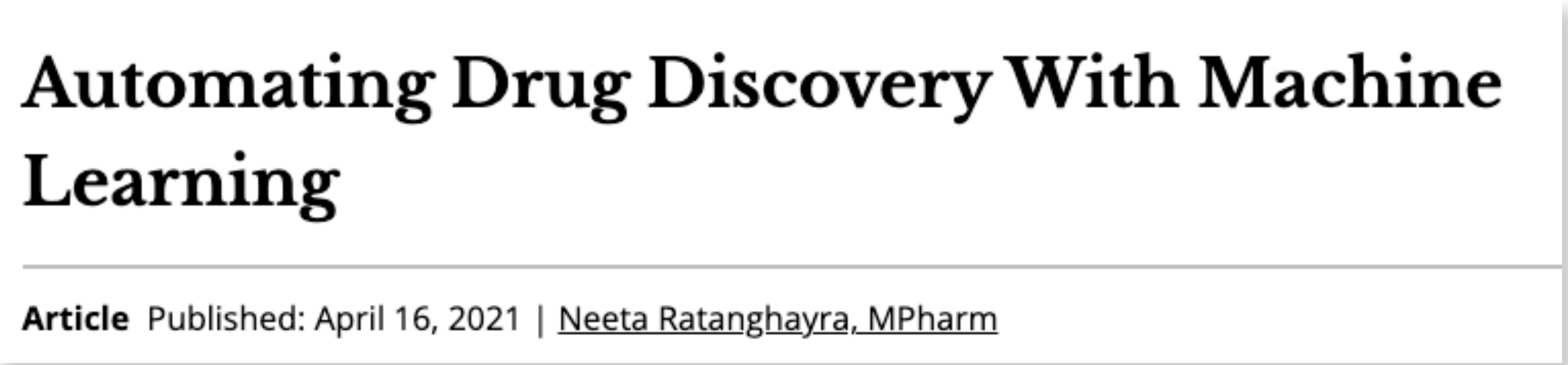


Deep Learning

Shortcuts to Simulation: How Deep Learning Accelerates Virtual Screening for Drug Discovery

May 11, 2020 14 min read

[DZone.com]



Automating Drug Discovery With Machine Learning

Article Published: April 16, 2021 | Neeta Ratanghayra, MPharm

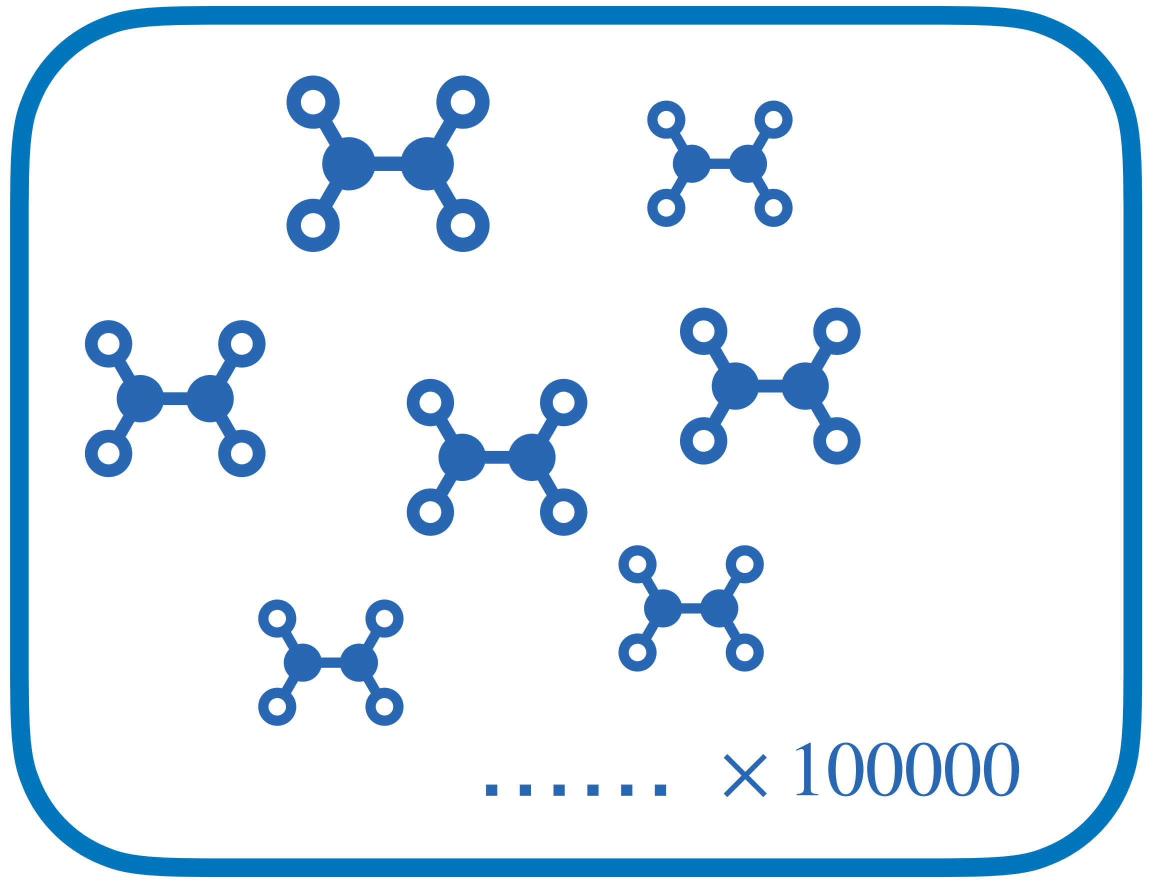
[technologynetworks.com]

Drug discovery: Which molecules/compounds to proceed to physical screening and clinical trials?

Decision and discovery processes

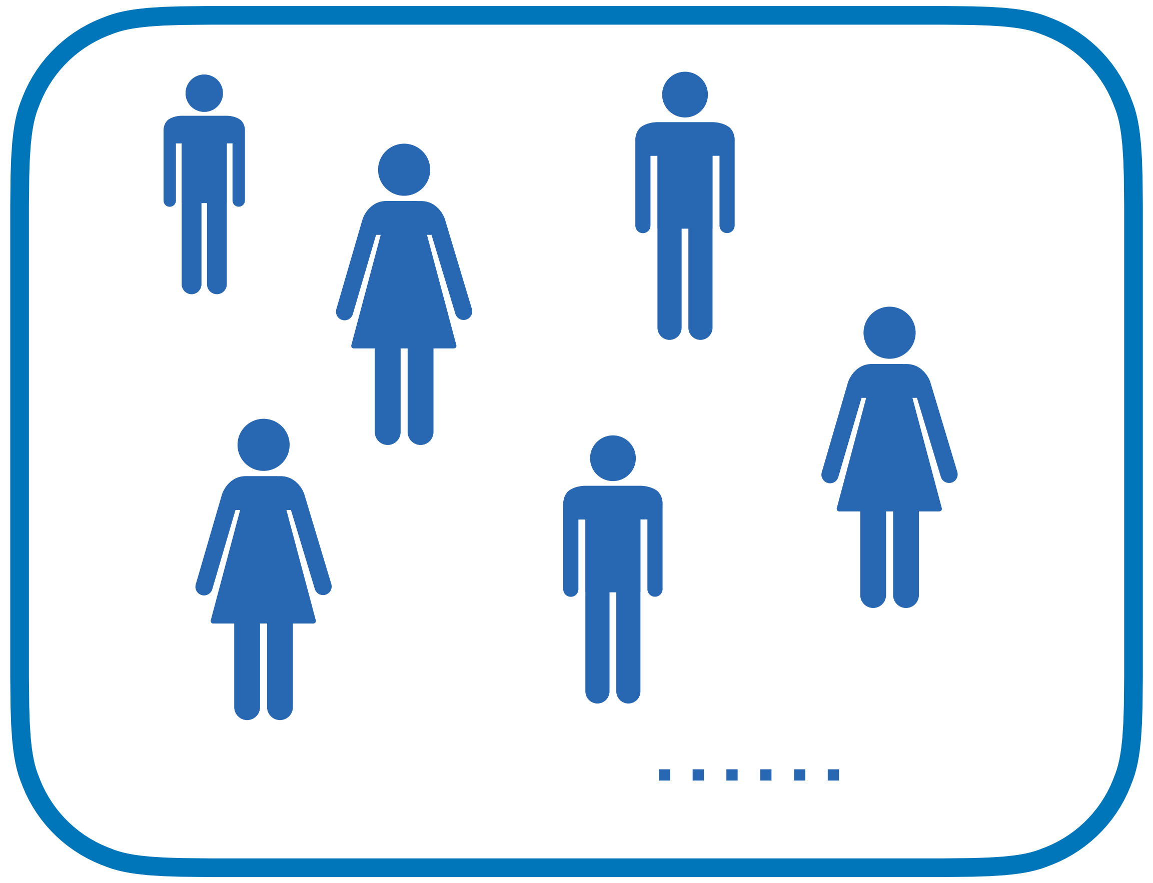
- Find a few interesting cases from a huge pool

Disease (COVID)



Candidate drugs

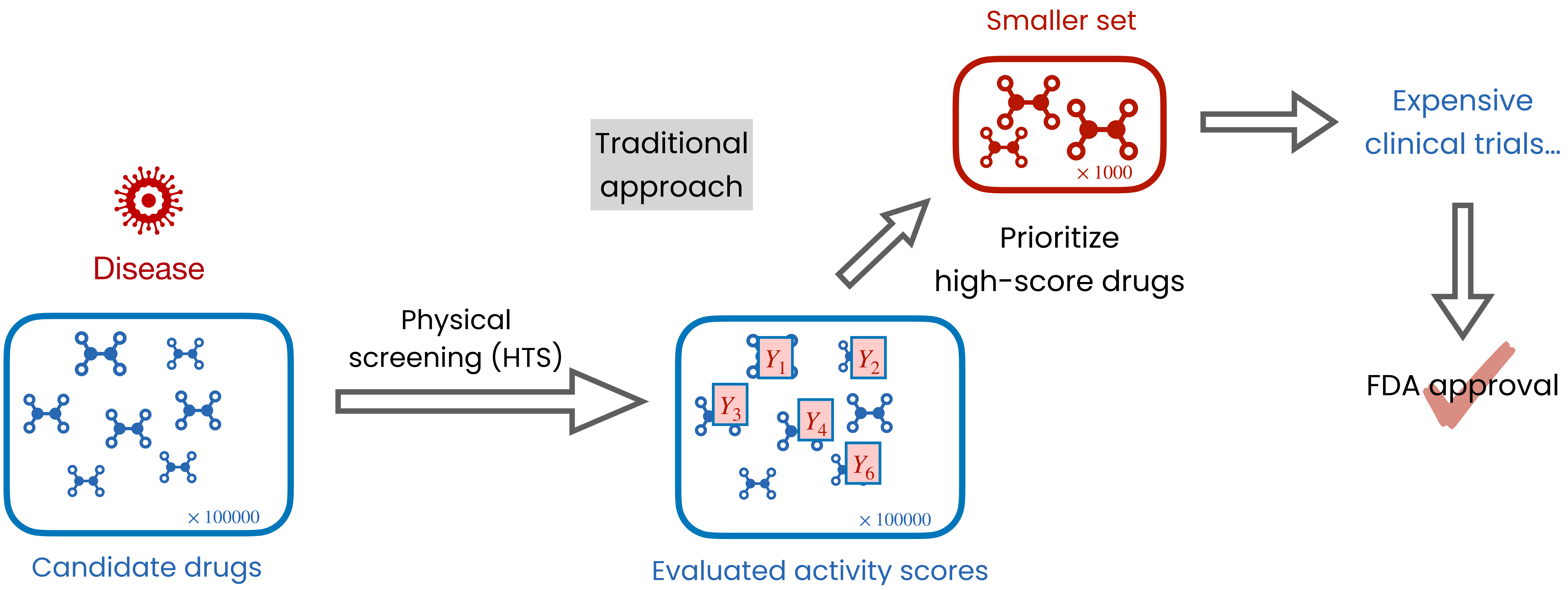
Position



Job applicants

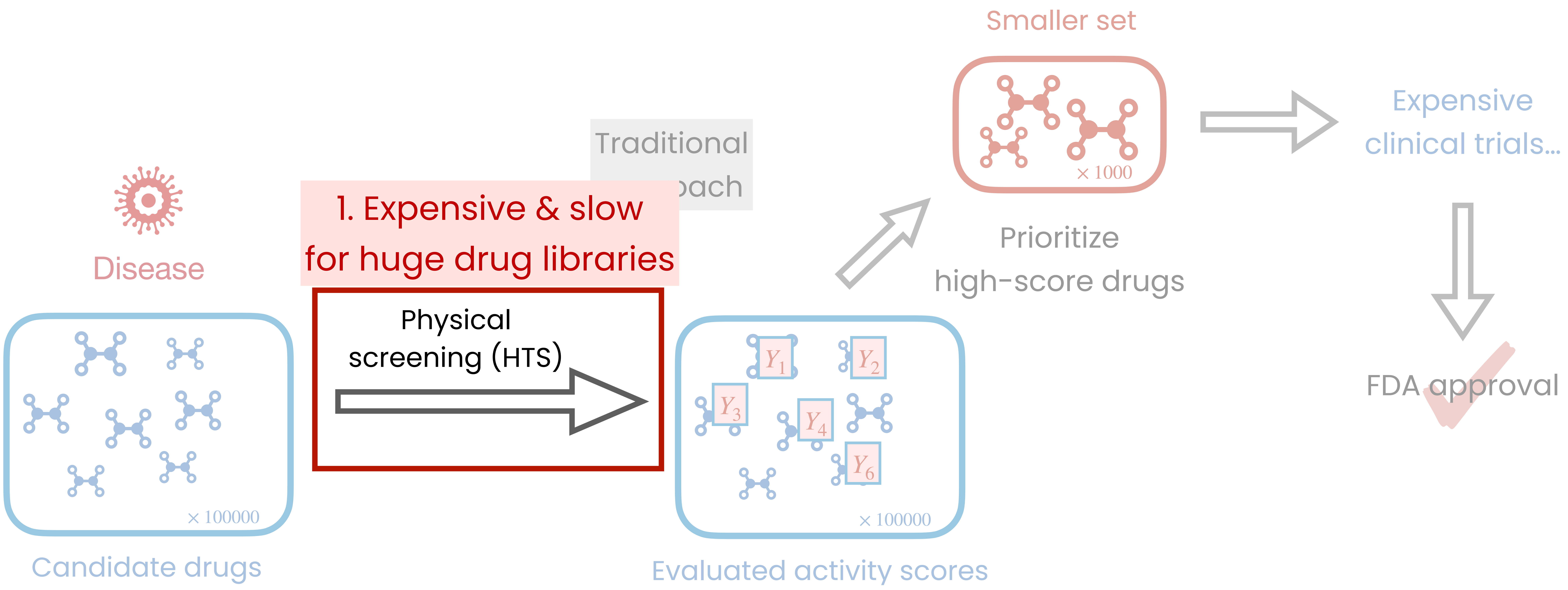
Decision and discovery processes

- Find a few interesting cases from a huge pool



Decision and discovery processes

- Find a few interesting cases from a huge pool



Smaller set

Expensive clinical trials...

Traditional oach

1. Expensive & slow for huge drug libraries

Disease

Physical screening (HTS)

Prioritize high-score drugs

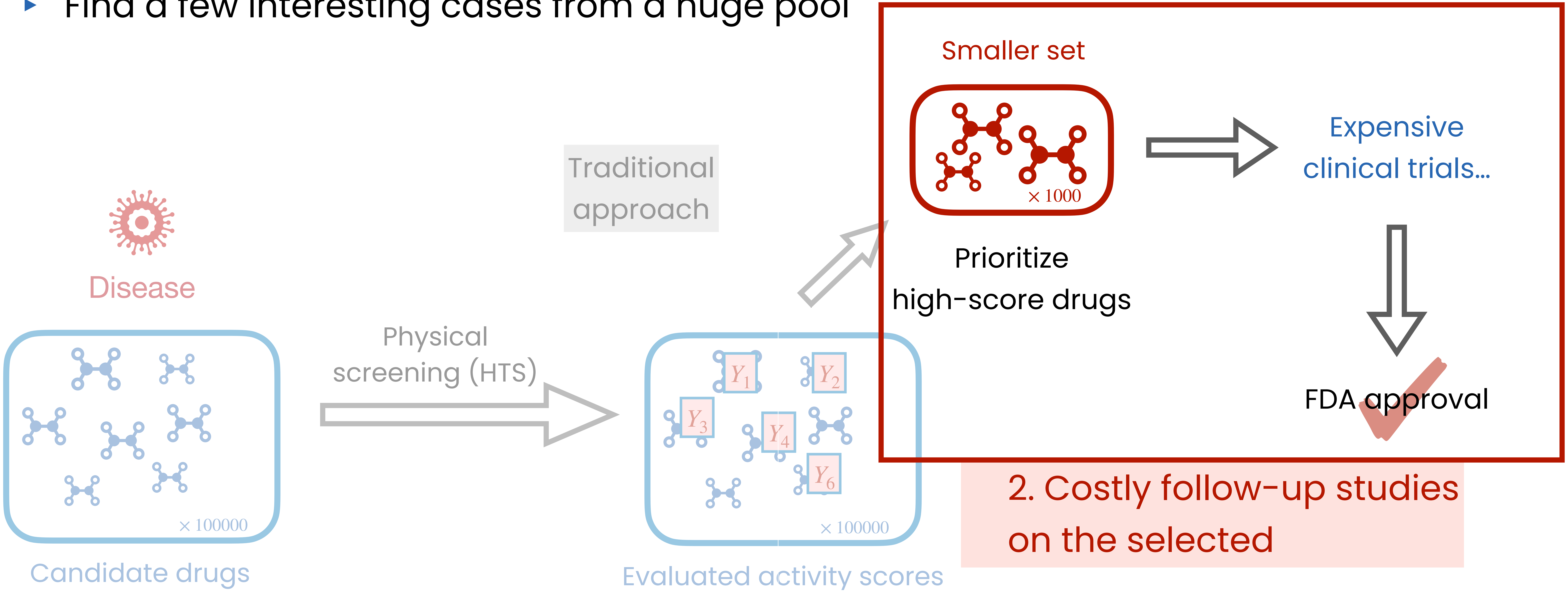
FDA approval

Candidate drugs

Evaluated activity scores

Decision and discovery processes

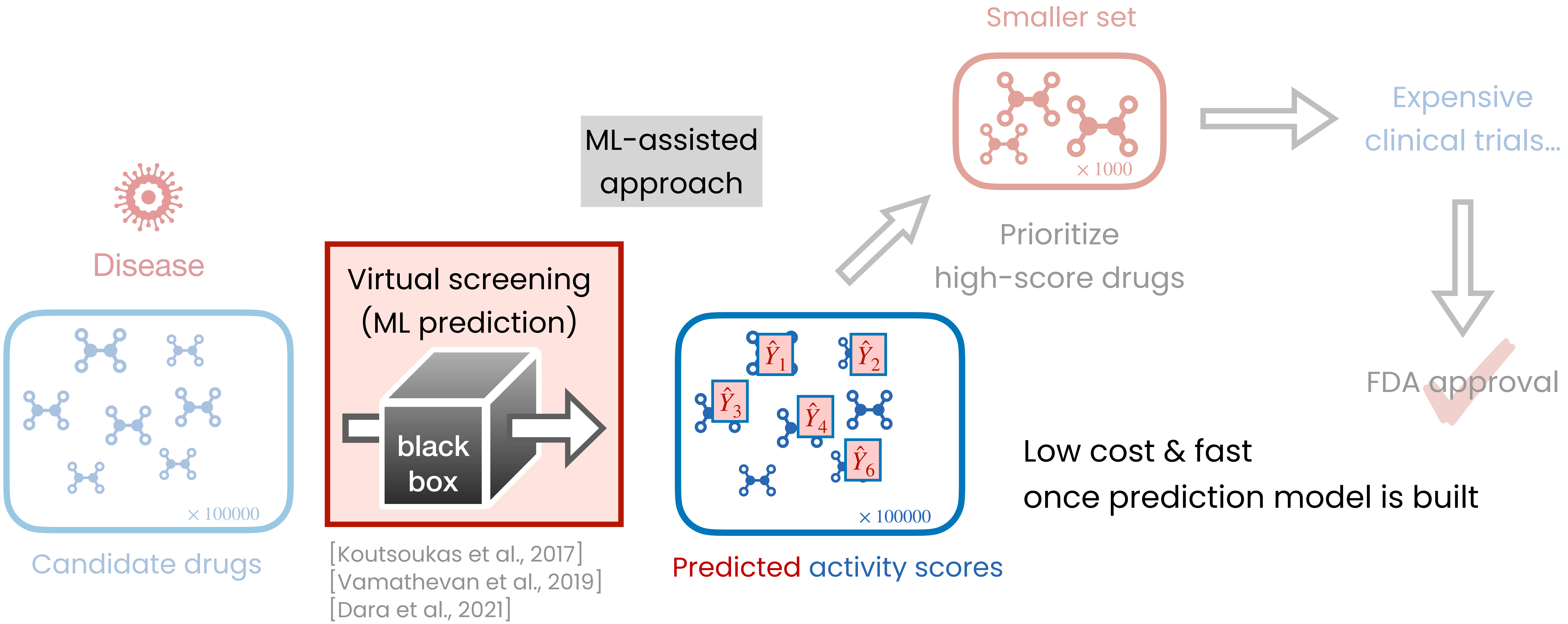
- Find a few interesting cases from a huge pool



2. Costly follow-up studies on the selected

The role of ML in decision and discovery processes

- Find a few interesting cases from a huge pool



Candidate drugs

ML-assisted approach

Virtual screening (ML prediction)

black box

[Koutsoukas et al., 2017]
[Vamathevan et al., 2019]
[Dara et al., 2021]

Predicted activity scores

Smaller set

Prioritize high-score drugs

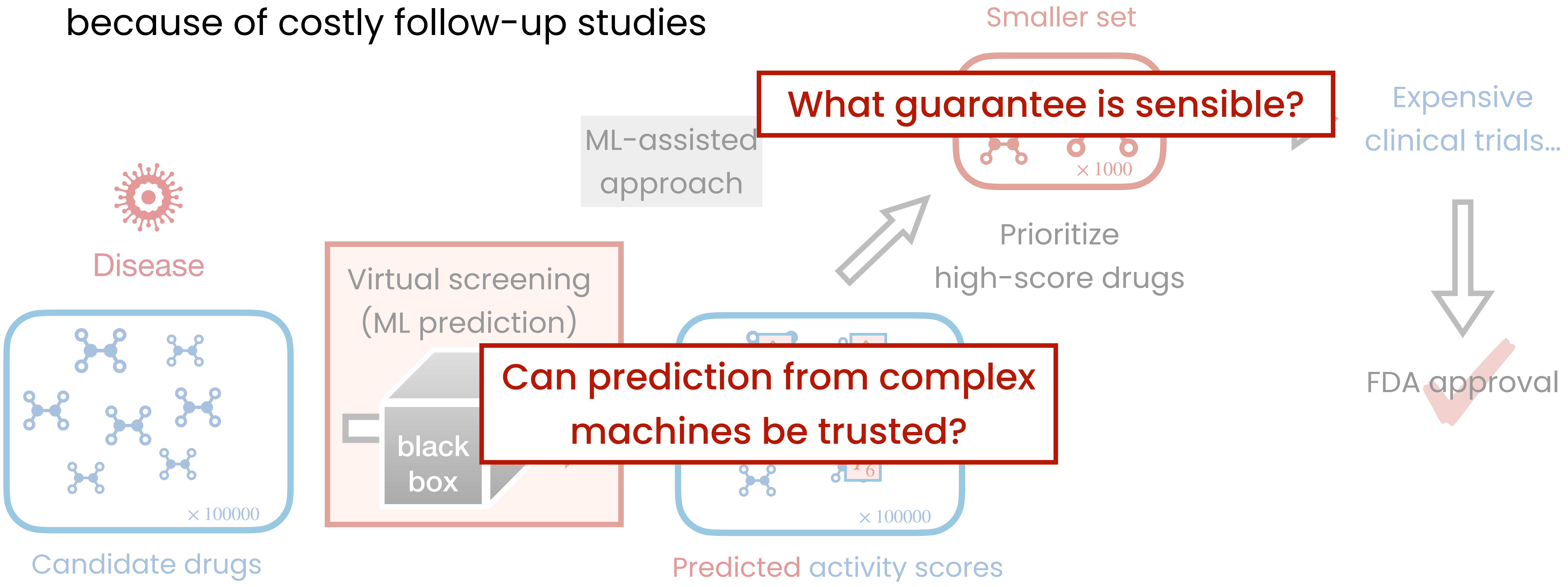
Expensive clinical trials...

FDA approval

Low cost & fast
once prediction model is built

The role of ML in decision and discovery processes

- Error on the selected is concerning because of costly follow-up studies

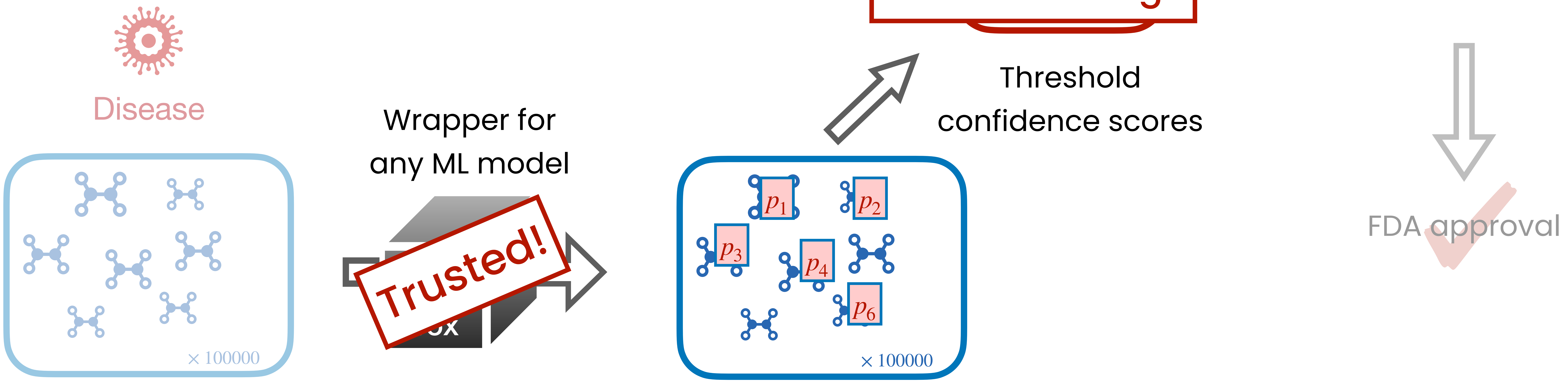


What guarantee is sensible?

Can prediction from complex machines be trusted?

This work

- Screening with error control on the selected candidates



Smaller set

90% of this set are active drugs

Expensive clinical trials...

Disease

Wrapper for any ML model

Threshold confidence scores

FDA approval

Candidate drugs

Calibrated confidence scores in $[0,1]$

Mathematical setup

- Any pre-trained ML model $\hat{\mu}: \mathscr{X} \to \mathscr{Y}$
- Training data $\{(X_i, Y_i)\}_{i=1}^n$ (already-screened drugs)
- Test samples $\{(X_{n+j}, Y_{n+j})\}_{j=1}^m$, only observe covariates $\{X_{n+j}\}_{j=1}^m$ (new drugs)

- For now: assume training and test samples are i.i.d. from an unknown distribution
  - Drugs drawn from a diverse drug library
  - Will be relaxed later on to allow for distribution shift

- Interested in large outcomes: $Y_{n+j} > c_{n+j}$ for some user-specified thresholds c_{n+j}

Guarantees we seek for

- Recall: Interested in large outcomes: $Y_{n+j} > c_{n+j}$ for some user-specified c_{n+j}

- Our goal is to find a subset $\mathscr{R} \subseteq \{1,\ldots,m\}$ as "promising candidates"
- While controlling the false discovery rate (FDR) below some $q \in (0,1)$

$$FDR = \mathbb{E}[FDP], \quad FDP = \frac{\boxed{\sum_{j=1}^{m} \mathbf{1}\{j \in \mathscr{R}, Y_{n+j} \leq c_{n+j}\}}}{\boxed{1 \vee |\mathscr{R}|}}$$

Number of selected but uninteresting units

[Benjamini and Hochberg, 1995]

$\approx$ Number of selected units

- FDR measures the proportion of follow-up resources wasted on uninteresting cases

Reliable prediction: conformal inference

- Conformal prediction for reliable predictive inference [Vovk et al., 2005]
  - Build any score function $V(x, y)$ based on the ML model, such as $V(x, y) = -|y - \hat{\mu}(x)|$
  - Compute $V_i = V(X_i, Y_i)$ for $i = 1,2,\ldots,n$
  - Construct prediction interval

$$\hat{C}(X_{n+j}; \alpha) = \left\{y : V(X_{n+j}, y) \geq \text{Quantile}(\alpha, \hat{P}_n(V_1, \ldots, V_n))\right\}$$

  - Assumption-free guarantee:

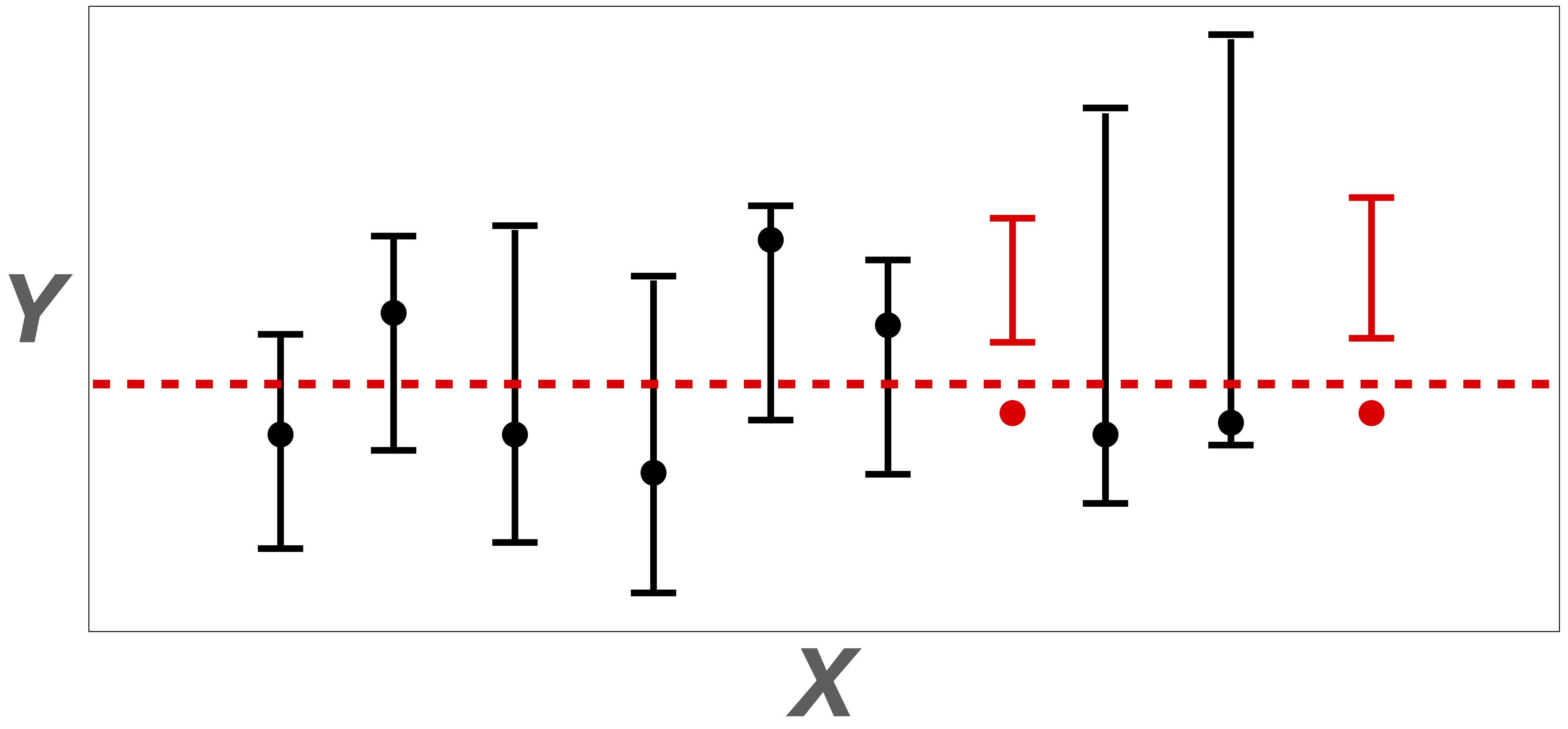
$$\mathbb{P}\left(Y_{n+j} \in \hat{C}(X_{n+j}; \alpha)\right) \geq 1 - \alpha, \quad \forall j = 1,\ldots,m$$

  - True for any score function $V(x, y)$ that builds on any (independently trained) ML model

- A literature on using conformal prediction intervals for drug discovery [Norinder et al., 2014, Svensson et al., 2017, Ahlberg et al., 2017, Svensson et al., 2018, Cortes-Ciriano and Bender, 2019, Wang et al., 2022]

Validity for one single point is not sufficient

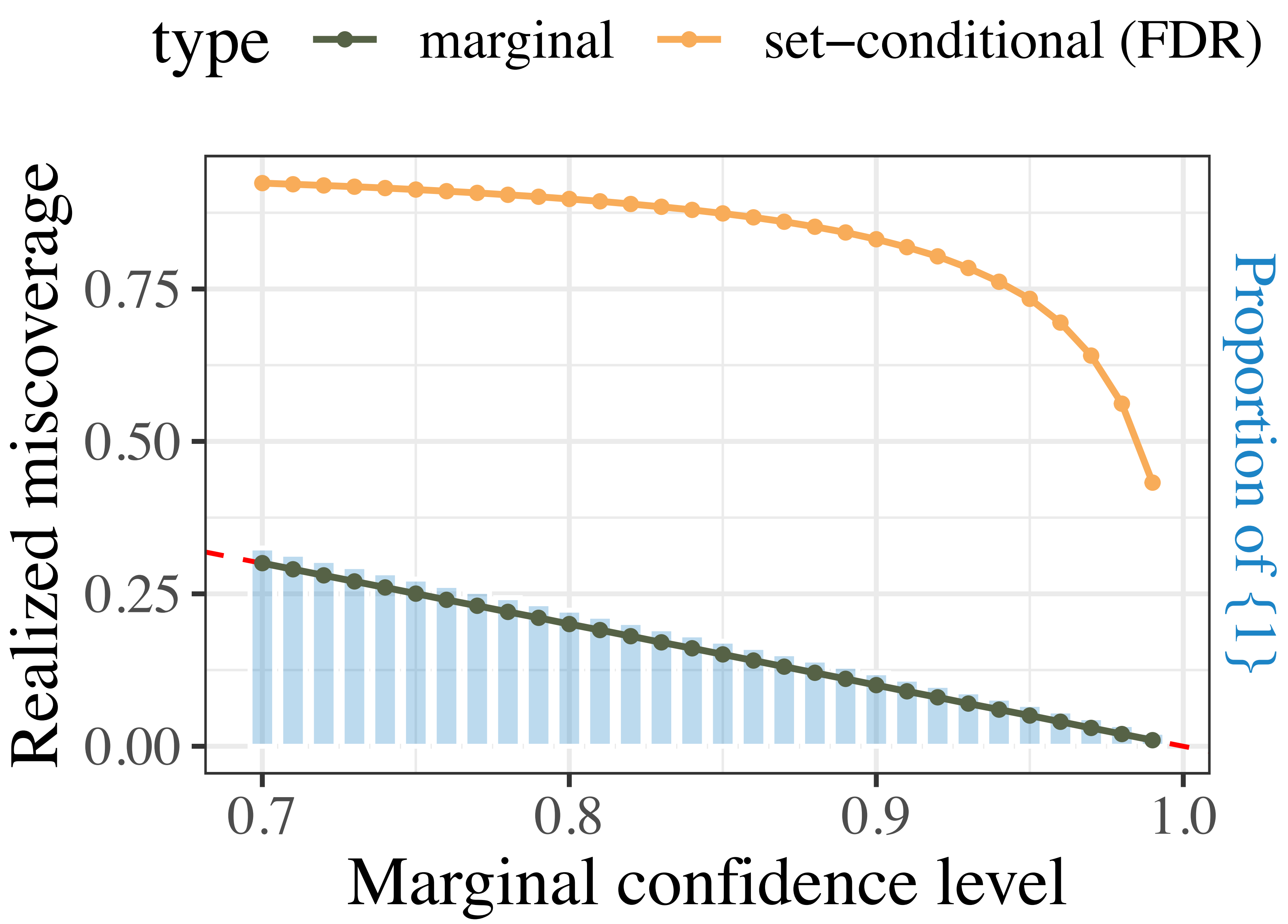
- $\mathbb{P}(Y_{n+j} \in \hat{C}(X_{n+j}; \alpha)) \geq 1 - \alpha$ over the randomness in training data and the j-th test data
- In binary classification, to find $Y_{n+j} = 1$ with $\leq q$ error, choose $\hat{C}(X_{n+j}, q) = \{1\}$?
  - Valid if those $\hat{C}(X_{n+j}, q) = \{1\}$ covers Y_{n+j} with probability $1 - q$
  - Coverage on average does not imply coverage on selected



- Constructing prediction intervals and then selecting promising ones is the approach in most works regarding conformal inference for drug discovery

Validity for one single point is not sufficient

- $\mathbb{P}(Y_{n+j} \in \hat{C}(X_{n+j}; \alpha)) \geq 1 - \alpha$ over the randomness in training data and the j-th test data
- In binary classification, to find $Y_{n+j} = 1$ with $\leq q$ error, choose $\hat{C}(X_{n+j}, q) = \{1\}$?
    - Valid if those $\hat{C}(X_{n+j}, q) = \{1\}$ covers Y_{n+j} with probability $1 - q$
    - Coverage on average does not imply coverage on selected



Build $(1 - \alpha)$ prediction sets taking the form $\{0\}, \{1\}, \{0,1\}$

Select those $\hat{C}(X_{n+j}; \alpha) = \{1\}$ to get the **orange** curve

Marginal miscoverage for the **dark** curve

Our approach: thresholding confidence measure

- Recall: Interested in large outcomes: $Y_{n+j} > c_{n+j}$ for some user-specified c_{n+j}

- Main idea: use a sequence of prediction intervals to decide a confidence measure, then leverage multiple testing ideas to threshold the confidence measure

  - Build any monotone score function $V(x, y)$, i.e., $y \le y'$ implies $V(x, y) \le V(x, y')$
    - One-sided residual $V(x, y) = y - \hat{\mu}(x)$
    - Fitted cumulative distribution function $V(x, y) = \hat{\mathbb{P}}(Y \le y \mid X = x)$
  - Compute $V_i = V(X_i, Y_i)$ for $i = 1,2,\ldots,n$
  - Compute test scores $\hat{V}_{n+j} = V(X_{n+j}, c_{n+j})$ for $j = 1,2,\ldots,m$
  - Compute confidence measures (p-value in statistics)

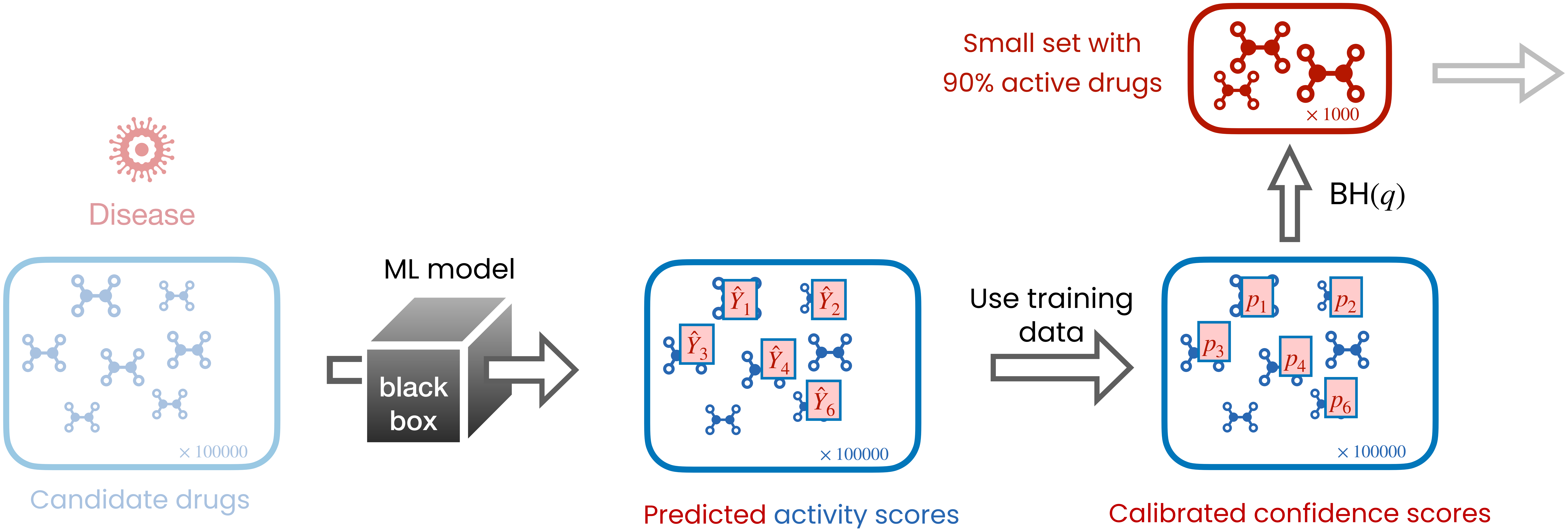
$\approx$ rank of $\hat{V}_{n+j}$ among training scores $\{V_i\}_{i=1}^{n}$

$$\boxed{p_j = \frac{\sum_{i=1}^{n} \mathbf{1}\{V_i < \hat{V}_{n+j}\} + U_j}{n+1}, \quad U_j \sim \text{Unif}[0,1]}$$

  - Get selection set $\mathcal{R}$ by Benjamini-Hochberg procedure applied to $\{p_j\}$ at level q

Our approach: thresholding confidence measure

- Back to the implied pipeline in drug discovery



Interpreting the confidence measure

- Recall: Interested in large outcomes: $Y_{n+j} > c_{n+j}$ for some user-specified c_{n+j}

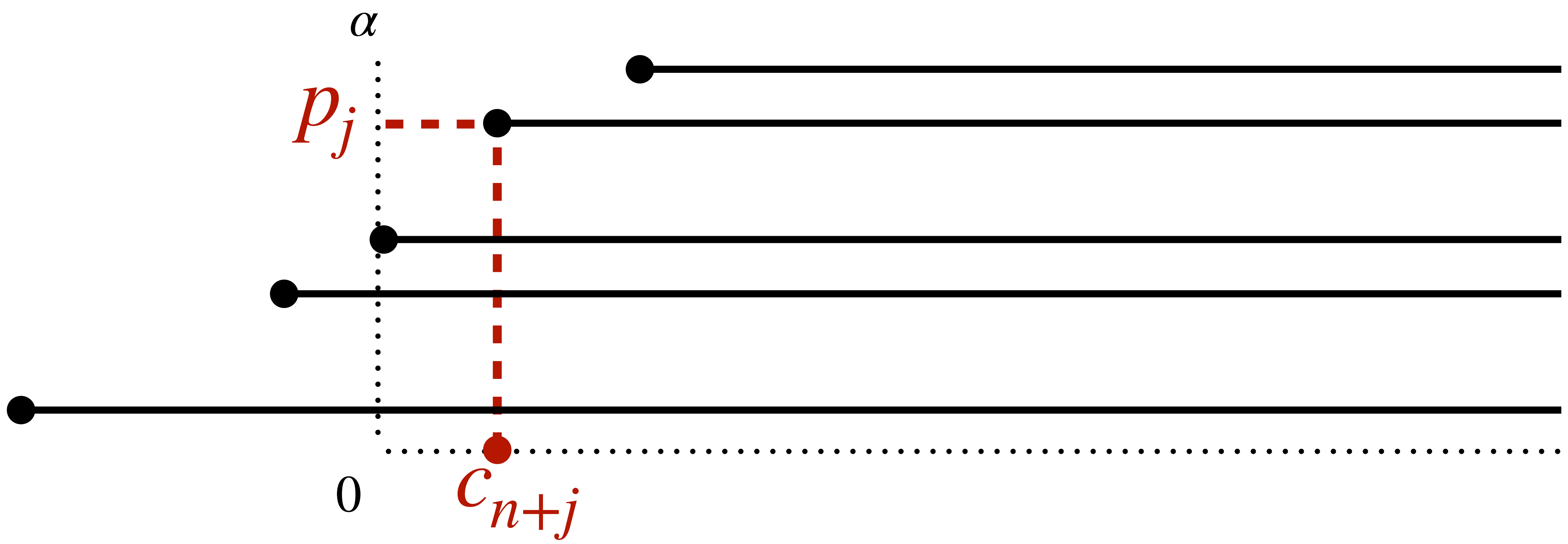
$$p_j = \frac{\sum_{i=1}^{n} \mathbf{1}\{V_i < \hat{V}_{n+j}\} + U_j}{n+1}, \quad U_j \sim \text{Unif}[0,1]$$

$$p_j \approx \inf\left\{\alpha \colon c_{n+j} \notin \hat{C}(X_{n+j};\alpha)\right\}$$

$$\hat{C}(X_{n+j};\alpha) = \{y \colon V(X_{n+j}, y) \geq \text{Quantile}(\alpha, \hat{P}_n(V_1, \ldots, V_n))\}$$

$\approx$ critical point α such that $\hat{C}(X_{n+j};\alpha)$ is all larger than c_{n+j}

A smaller p_j means c_{n+j} is smaller than the typical behavior of Y_{n+j}



By monotonicity,
$\hat{C}(X_{n+j};\alpha) = [\eta(X_{n+j};\alpha), \infty)$

A bit more statistics

- Recall: Interested in large outcomes: $Y_{n+j} > c_{n+j}$ for some user-specified c_{n+j}

- This can be viewed as testing the random null hypotheses

$$H_j\colon\ Y_{n+j} \le c_{n+j}$$

- Our confidence measure p_j is a valid p-value for testing H_j

$$\mathbb{P}(p_j \le t,\ H_j \text{ is true}) \le t, \quad \forall t \in [0,1]$$

Valid type-I control that accounts for the randomness in H_j

FDR control with the confidence measure

- Get selection set $\mathscr{R}$ by Benjamini-Hochberg procedure applied to $\{p_j\}$ at level q
  - Set $\mathscr{R} = \{j : p_j \leq q\hat{k}/m\}$, where $\hat{k} = \max\left\{k : \sum_{j=1}^{m} \mathbf{1}\{p_j \leq qk/m\} \geq k\right\}$

Theorem (J. and Candès, 2022)

If $V(x, y)$ is monotone, the training and test data are i.i.d., and for each j, data in $\{Z_i\}_{i=1}^{n} \cup \{\tilde{Z}_{n+\ell}\}_{\ell \neq j} \cup \{Z_{n+j}\}$ are mutually independent for $Z_i = (X_i, Y_i)$ and $\tilde{Z}_{n+j} = (X_{n+j}, c_{n+j})$,
Then for any $q \in (0,1)$, the output $\mathscr{R}$ at level q obeys $FDR \leq q$.

- True for random c_{n+j} (will my health risk tomorrow be higher than today?)

A bit more math

- This is a new statistical problem: random p-values for random hypotheses
- Also, p-values are mutually dependent, which is typically challenging for FDR control
- Why it works: the p_j are “positively dependent”, which ensures FDR control

- Proof step 1: Leave-one-out

$$FDR \leq \sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbf{1}\{j \in \mathscr{R}_{j\to *}\}}{1 \vee |\mathscr{R}_{j\to *}|}\right]$$

- Proof step 2: Uniform distribution + positive dependence

$$\mathbb{E}\left[\frac{\mathbf{1}\{j \in \mathscr{R}_{j\to *}\}}{1 \vee |\mathscr{R}_{j\to *}|}\right] \leq \frac{q}{m}$$

A bit more math

- Proof step 1: Leave-one-out
  - Define $p_j^* = \dfrac{\sum_{i=1}^n \mathbf{1}\{V_i < V_{n+j}\} + U_j}{n+1}$ with the "true test score" $V_{n+j} = V(X_{n+j}, Y_{n+j})$ (uncomputable, just for analysis)
  - Let $\mathscr{R}_{j\to *}$ be the rejection set of BH applied to $p_j^* \cup \{p_\ell\}_{\ell \neq j}$ at level q
  - Because of monotonicity, one can show that $\mathscr{R} = \mathscr{R}_{j\to *}$ on the event $\{Y_{n+j} \leq c_{n+j}$ and $j \in \mathscr{R}\}$
  - This implies

$$FDR = \sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbf{1}\{j \in \mathscr{R}, Y_{n+j} \leq c_{n+j}\}}{1 \vee |\mathscr{R}|}\right] \leq \sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbf{1}\{j \in \mathscr{R}_{j\to *}, Y_{n+j} \leq c_{n+j}\}}{1 \vee |\mathscr{R}_{j\to *}|}\right] \leq \sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbf{1}\{j \in \mathscr{R}_{j\to *}\}}{1 \vee |\mathscr{R}_{j\to *}|}\right]$$

A bit more math

- Proof step 1: Leave-one-out

$$FDR \leq \sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbf{1}\{j \in \mathscr{R}_{j\to *}\}}{1 \vee |\mathscr{R}_{j\to *}|}\right]$$

- Proof step 2: Uniform distribution + positive dependence
  - For i.i.d. data, the oracle p-value is uniformly distributed $p_j^* \sim \text{Unif}[0,1]$
  - Also, $\{p_\ell\}_{\ell \neq j}$ are PRDS on p_j^*
  - This implies for every j, [Benjamini and Yekutieli, 2001]

$$\mathbb{E}\left[\frac{\mathbf{1}\{j \in \mathscr{R}_{j\to *}\}}{1 \vee |\mathscr{R}_{j\to *}|}\right] \leq \frac{q}{m}$$

A random vector $X = (X_1, \ldots, X_m)$ is PRDS on x_i if for any increasing set D, the probability $\mathbb{P}(X \in D \mid X_i = x)$ is increasing in x

A set D is increasing if $a \in D$ and $b \geq a$ implies $b \in D$

A bit more math

- This is a new statistical problem: random p-values for random hypotheses
- Also, p-values are mutually dependent, which is typically challenging for FDR control
- Why it works: the p_j are "positively dependent", which ensures FDR control

- Proof step 1: Leave-one-out

$$FDR \leq \sum_{j=1}^{m} \mathbb{E}\left[\frac{\mathbf{1}\{j \in \mathscr{R}_{j\to *}\}}{1 \vee |\mathscr{R}_{j\to *}|}\right]$$

Takeaway:

- p_j controls the false selection error for each test sample j
- p_j's are PRDS so they work well together

- Proof step 2: Uniform distribution + positive dependence

$$\mathbb{E}\left[\frac{\mathbf{1}\{j \in \mathscr{R}_{j\to *}\}}{1 \vee |\mathscr{R}_{j\to *}|}\right] \leq \frac{q}{m}$$

Power boosting

- While FDR is controlled for any monotone score $V(x, y)$, some makes it powerful
- If the thresholds are constant $c_{n+j} \equiv c$, a particularly powerful choice is `clipped' score

$$V(x, y) = +\infty \cdot \mathbf{1}\{y > c\} + c \cdot \mathbf{1}\{y \leq c\} - \hat{\mu}(x)$$

- In binary case and $c = 0$, the ideal score is monotone in $\mathbb{P}(Y = 1 \mid X = x)$ (see paper)

Real application: drug property prediction for HIV

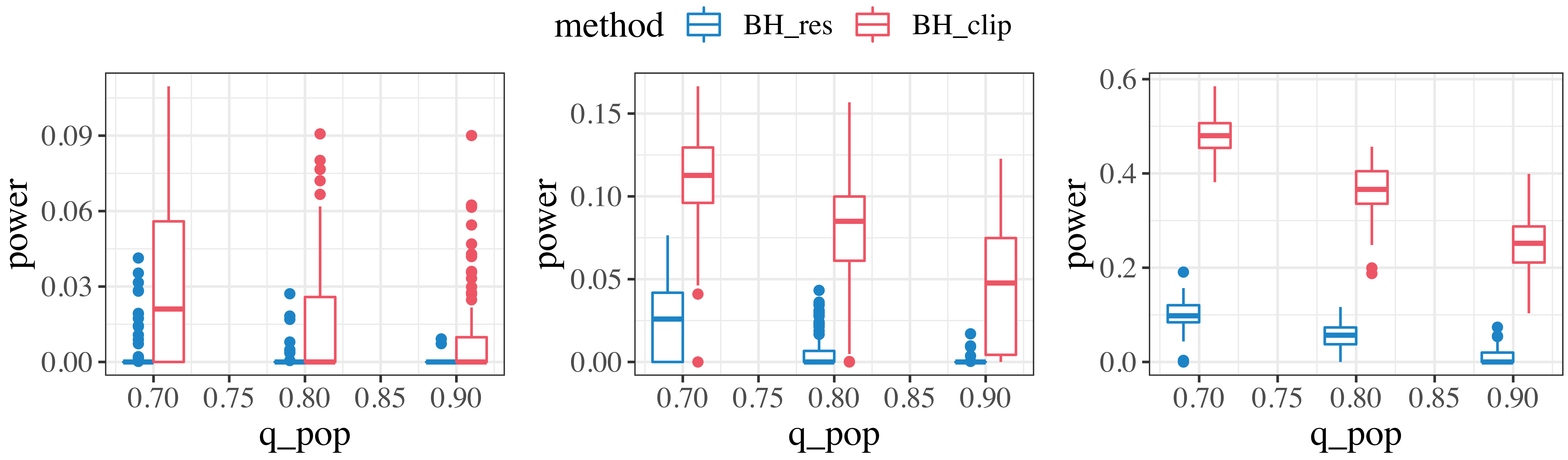
- Binary $Y \in \{0,1\}$: whether the drug interacts with the disease
- The drug library is $n_{tot} = 41127$ in total, use $6:2:2$ split
- Very sparse data: only 3% drugs are active
- Our hope: find a smaller subset to proceed so that $(1-q)$ of the subset are active drugs
- FDR level $q \in \{0.1,0.2,0.5\}$, use a small neural network (can be more complicated)

	Realized FDR			**Power**			$\lvert \mathscr{R} \rvert$		
FDR level	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5
Powerful score	**0.0957**	**0.196**	**0.495**	0.0788	0.174	0.410	26.5	64.2	240
Score $V(x,y) = y - \hat{\mu}(x)$	**0.0989**	**0.196**	**0.494**	0.0766	0.174	0.410	25.8	64.4	239

Real application: drug-target-interaction prediction

- Davis dataset, $Y \in \mathbb{R}$ continuous binding affinities, X feature for a drug-target pair
- The drug library is $n_{tot} = 30060$ in total, use $2:2:6$ split
- Set c_{n+j} as the q_{pop}-th quantile of the outcomes in the first training fold with the same binding target as test sample j, where $q_{pop} \in \{0.7, 0.8, 0.9\}$
- FDR level $q \in \{0.1, 0.2, 0.5\}$

$V(x, y) = y - \hat{\mu}(x)$ Powerful score



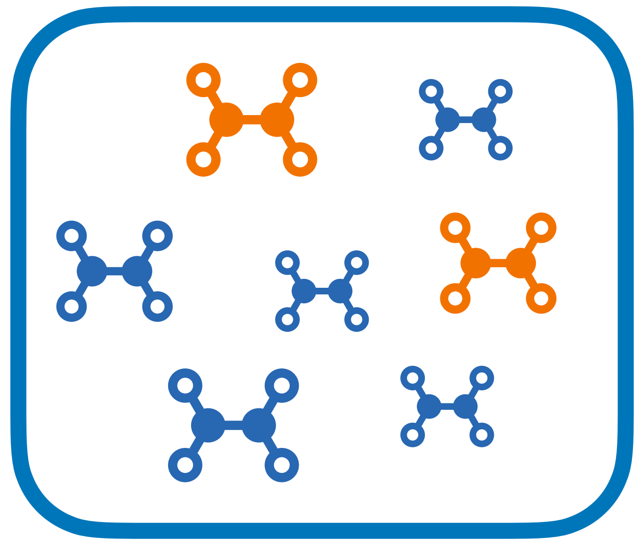
So far, and next

- A method that turns any prediction model into a reliable selection procedure
- Theoretically, FDR control due to monotonicity and positive dependence (PRDS)
- Works reasonably well in real drug discovery tasks
    - + job hiring tasks in paper
    - + more benchmarks and applications in ongoing work

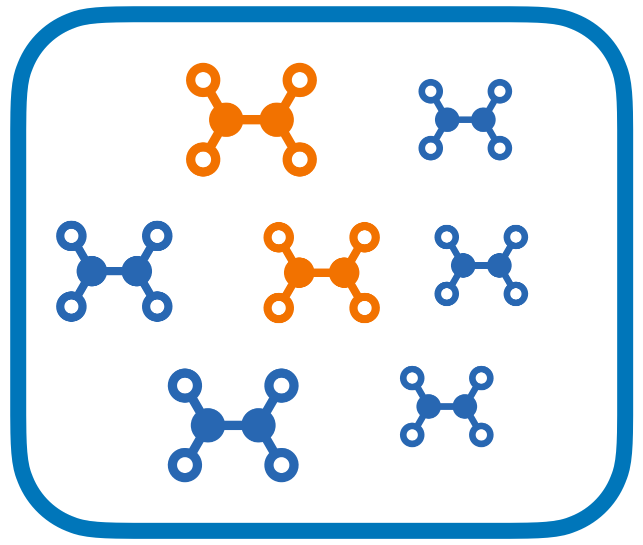
- **Next: dealing with distribution shifts**

Distribution shifts

- The only assumption for this method to work is i.i.d. data
- Are my evaluated drugs comparable to the unknown drugs?
  - **Yes** if the evaluated ones are drawn without preference from your library



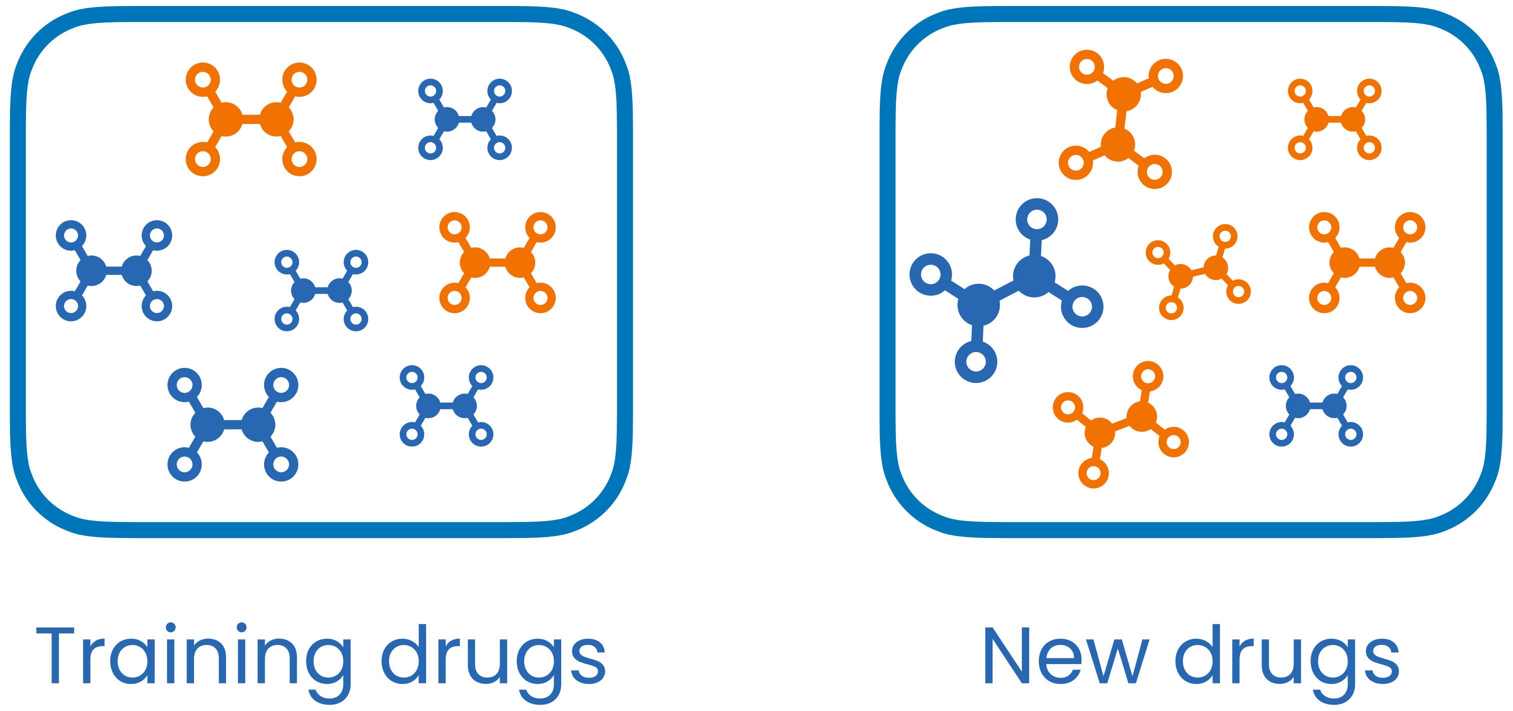
Training drugs



New drugs

Distribution shifts

- The only assumption for this method to work is i.i.d. data
- Are my evaluated drugs comparable to the unknown drugs?
  - **Yes** if the evaluated ones are drawn without preference from your library
  - **No** if you preferred drugs with some specific structures, etc



- Similar issues happen in job hiring, health monitoring, counterfactual inference…
  - Candidates documented last year may differ from current
  - Patients may differ in demographics across hospitals
  - People under treatment may be different than those under control

Extending the setting to covariate shifts

- Formally, we assume the test data are i.i.d. from some unknown $\mathbb{Q}$
- And the training data are i.i.d. from some unknown $\mathbb{P}$
- We only know that they are related by a covariate shift:

[Tibshirani et al., 2019]

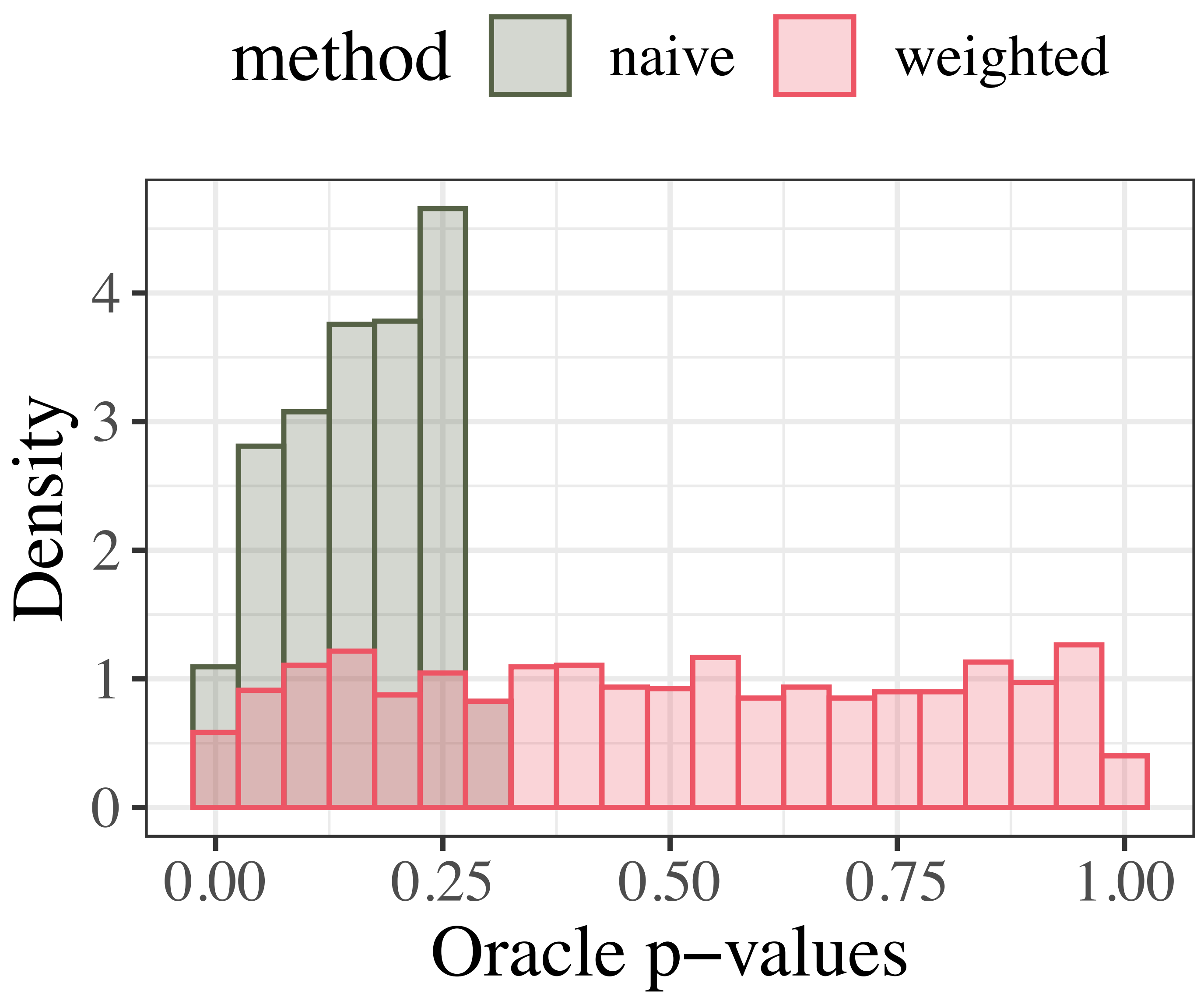
$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(x, y) = w(x)$$

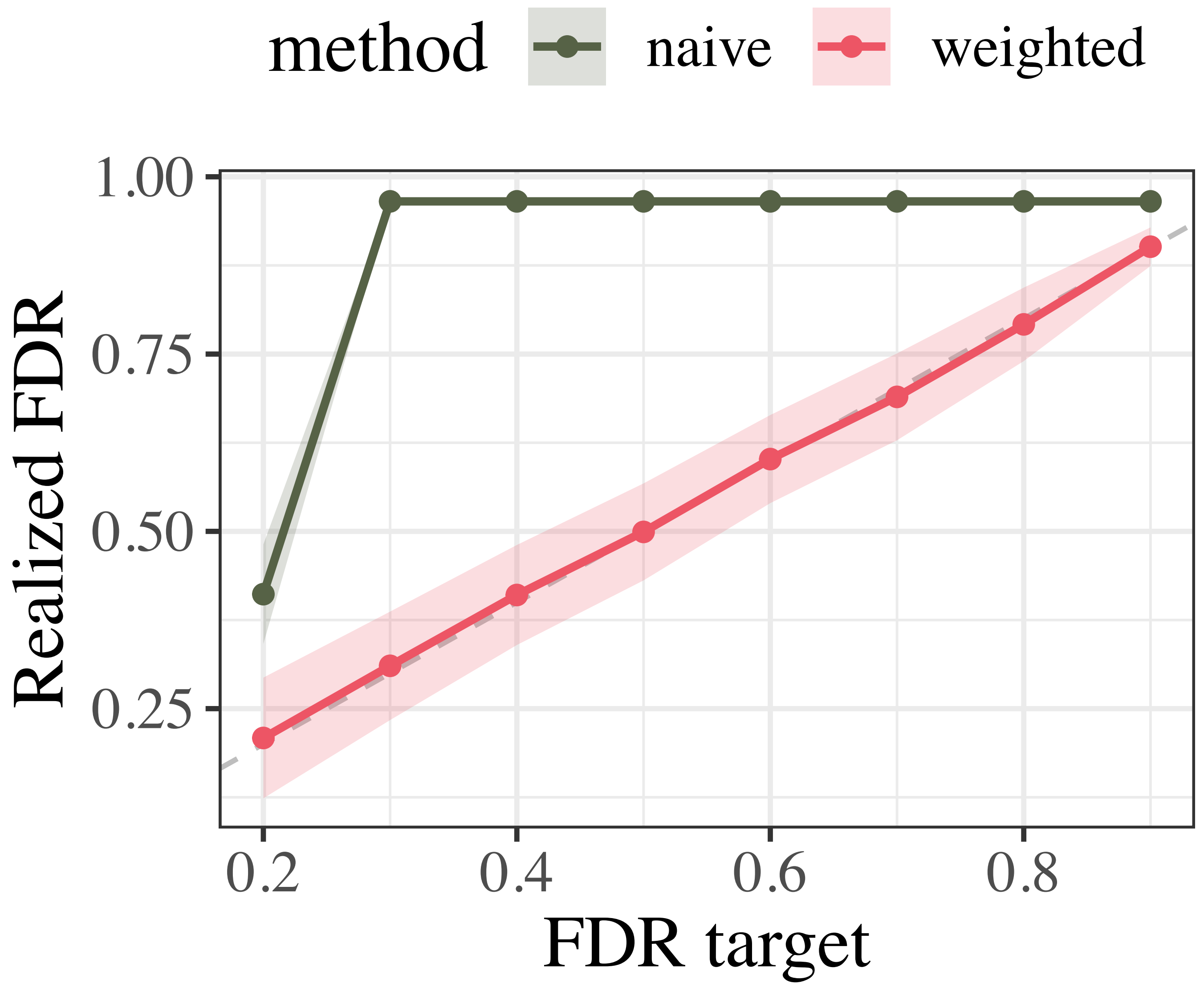
- The distribution shift is fully attributed to covariates

Confidence measure under covariate shift

- Under covariate shift, we need a new confidence measure

If use previous confidence measures (p-values) when there is covariate shift





p-values are no longer valid

FDR can be violated

Confidence measure under covariate shift

- Under covariate shift, we need a new confidence measure
  - Build any monotone score function $V(x, y)$, i.e., $y \leq y'$ implies $V(x, y) \leq V(x, y')$
    - One-sided residual $V(x, y) = y - \hat{\mu}(x)$
    - Fitted cumulative distribution function $V(x, y) = \hat{\mathbb{P}}(Y \leq y \mid X = x)$
  - Compute $V_i = V(X_i, Y_i)$ for $i = 1,2,\ldots, n$
  - Compute test scores $\hat{V}_{n+j} = V(X_{n+j}, c_{n+j})$ for $j = 1,2,\ldots, m$
  - Compute weighted confidence measures (p-value in statistics)

$$\boxed{p_j = \frac{\sum_{i=1}^n w(X_i)\mathbf{1}\{V_i < \hat{V}_{n+j}\} + w(X_{n+j})}{\sum_{i=1}^n w(X_i) + w(X_{n+j})}}$$

$\approx$ weighted rank of $\hat{V}_{n+j}$ among training scores $\{V_i\}_{i=1}^n$

Statistical properties

- The new confidence measure has similar statistical properties as before
- Still, we are testing the random null hypotheses

$$H_j:\ Y_{n+j} \le c_{n+j}$$

- Our p_j is a valid p-value for testing H_j under covariate shift

$$\mathbb{P}(p_j \le t,\ H_j \text{ is true}) \le t, \quad \forall t \in [0,1]$$

Probability over both training data and the test sample j

[Asserting $Y_{n+j} > c_{n+j}$ if $p_j \le \alpha$] controls type-I error **for a single test point**

Statistical properties

- The new confidence measure has similar statistical properties as before
- Still, we are testing the random null hypotheses

$$H_j\colon Y_{n+j} \le c_{n+j}$$

- Our p_j is a valid p-value for testing H_j under covariate shift

$$\mathbb{P}(p_j \le t,\ H_j \text{ is true}) \le t, \quad \forall t \in [0,1]$$

Probability over both training data and the test sample j

[Asserting Y_{n+j} point

- Does the previous recipe work?

(Recall for i.i.d.) **Takeaway:**

- p_j controls the false selection error for each test sample j ✓
- p_j's are PRDS so they work well together ?

Statistical properties

- Weighted conformal p-values are **not PRDS**

Theorem (J. and Candès, in preparation, 2023+)

Suppose we construct p_j assuming $Y_{n+j} = c_{n+j}$. Then there exists a weight function $w(\cdot)$, a monotone score function $V(\cdot,\cdot)$, such that for training and test samples obeying a covariate shift, the p-va

(Recall for i.i.d.) **Takeaway:**

- p_j controls the false selection error for each test sample j ✓
- p_j's are PRDS so they work well together ✗

- Does the previous recipe work?
  - Not sure theoretically, but works in our numerical experiments
  - In forthcoming paper: A new procedure **exactly** controlling FDR in **finite** samples

Other applications of this framework

- Detecting positive individual treatment effects
  - Individual treatment effects are random variables that describe the difference in outcomes with treatment $O(1)$ versus without treatment $O(0)$
  - We are interested in whether $O_{n+j}(1) > O_{n+j}(0)$ or not
  - Equivalent to taking $Y_{n+j} = O_{n+j}(1)$ and $c_{n+j} = O_{n+j}(0)$ for a control unit

- Detecting outliers and concept drifts
  - Given a set of normal transactions from $\mathbb{P}$ and a set of new transactions
  - We are interested in whether the new transactions are from $\mathbb{Q}$ (covariate shift from $\mathbb{P}$)

Summary

- In prediction-assisted screening problems, FDR can be a sensible measure
- A method that turns any prediction model into a reliable selection procedure
  - Useful if interested in “large” outcomes
  - Builds confidence scores (p-values) upon any prediction model
  - Controls FDR so that your follow-up investigations are well-deserved
- Extension to situations with covariate shifts
  - Some more complicated methodology & theory

arXiv: 2210.01408

any ML model
Trusted!
× 100000
× 1000

Candidate drugs

Small set with (1-q) true discovery