Machine Learning on Belgian Health Expenditure Data

Data-Driven Screening for Type 2 Diabetes

Marc Claesen

STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics
Department of Electrical Engineering (ESAT)
KU Leuven

iMinds Medical IT Department

Mission statement

Demonstrate potential applications of health expenditure data

- with clinical relevance to improve healthcare
- enabled by advanced machine learning techniques

We focused on (type 2) diabetes mellitus

- large-scale survival analysis
- development of a screening tool

Machine learning contributions

- semi-supervised learning
- automated hyperparameter optimization
- open-source software ecosystem

Outline

Outline


DIABETES

https://www.idf.org/diabetesatlas

Case-finding for type 2 diabetes

The problem:

- long, asymptomatic period (in contrast to T1D)
- many patients remain undiagnosed for years
- many patients present signs of complications at diagnosis

Early detection:

- complications can be delayed or avoided
- universal screening is infeasible

Recommendations by WHO, ADA, IDF, Diabetes Liga, . . . :

- focus on case-finding (= identify persons at high risk)
- forward high risk patients to diagnostic test

Early detection $\rightarrow$ early treatment $\rightarrow$ long-term health benefits.

Case-finding guidelines for type 2 diabetes

Persons of 18–45 years of age and one of these conditions:

- prior history of gestational diabetes
- prior history of stress-induced hyperglycemia

or two of the following conditions:

- prior history of giving birth to a baby of over 4.5 kg
- diabetes in first-line relatives
- high BMI $\geq$ 25 kg/m^2
- large waist circumference
- treated for high blood pressure or with corticoids

Persons of 45–64 years of age with $\geq$ 1 of above conditions.
Persons above 64 years old.


HEALTH
INSURANCE


E-HEALTH
INSURANCE?

AAN TE VULLEN DOOR DE GERECHTIGDE

HIERNA INVULLEN OF KLEEFBRIEFJE V.I. AANBRENGEN

Naam en voornaam van de gerechtigde/patiënt : .....

.....

Verzekeringsinstelling : .....

Inschrijvingsnummer : ☐☐☐☐☐☐☐☐☐☐☐☐☐

Adres van de gerechtigde : .....

.....

GETUIGSCHRIFT VOOR VERSTREKTE HULP

IN TE VULLEN DOOR DE VERSTREKKER

Naam en voornaam van de patiënt : [redacted]

Gerechtigde - Echtgenote - Kind - Ascendent (1)

Raadpleging – Bezoek (2)

Datum : 04 / 11 / 11 Nomenclatuur nr. : 101076

Reiskosten : .....

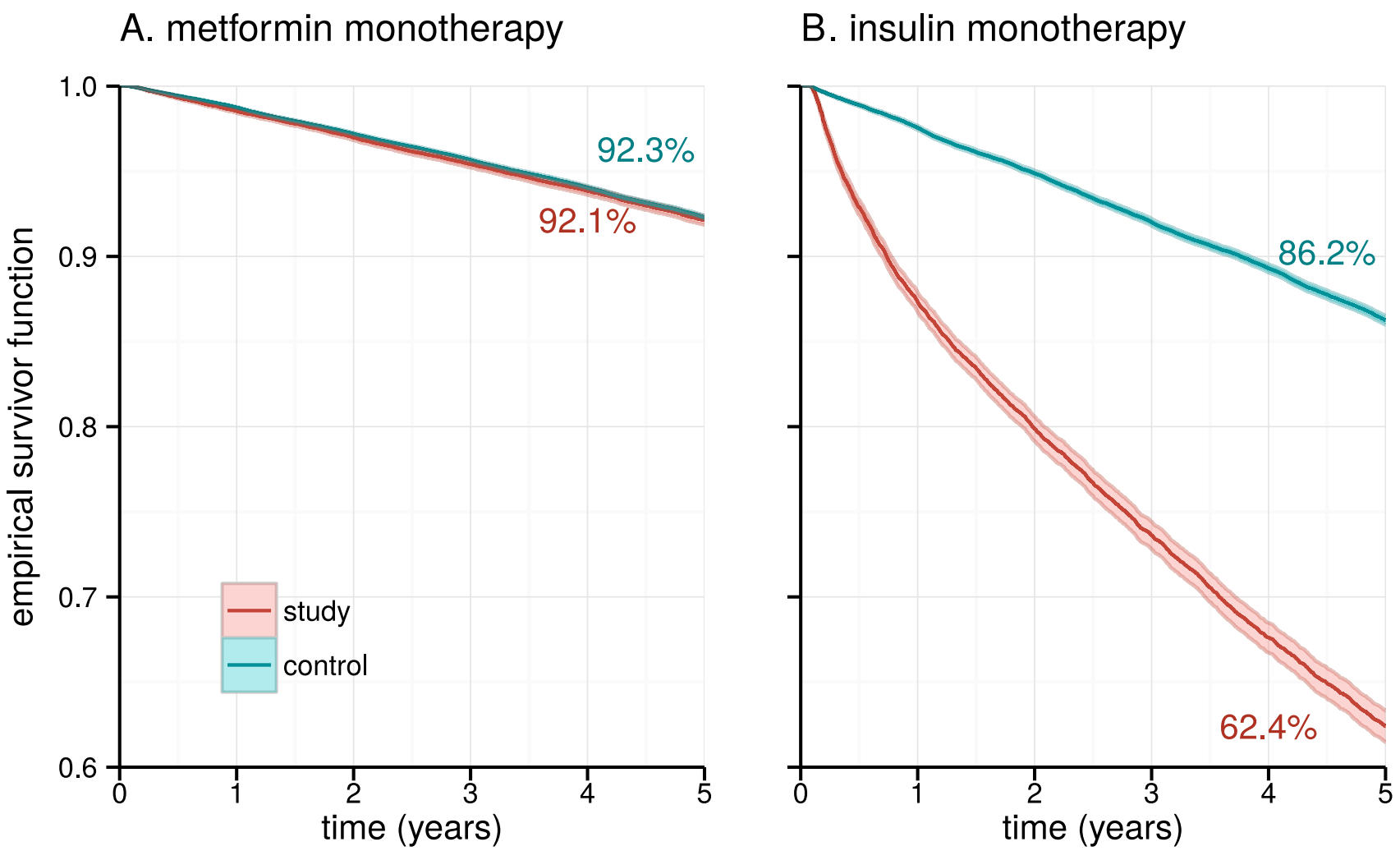
Andere verstrekkingen (2)

Datum van de verstrekking	Nummer v. d. nomenclatuur	Voorbehouden aan V.I.	Datum van de verstrekking	Nummer v. d. nomenclatuur	Voorbehouden aan V.I.
					(2)

Voorgeschreven door : .....

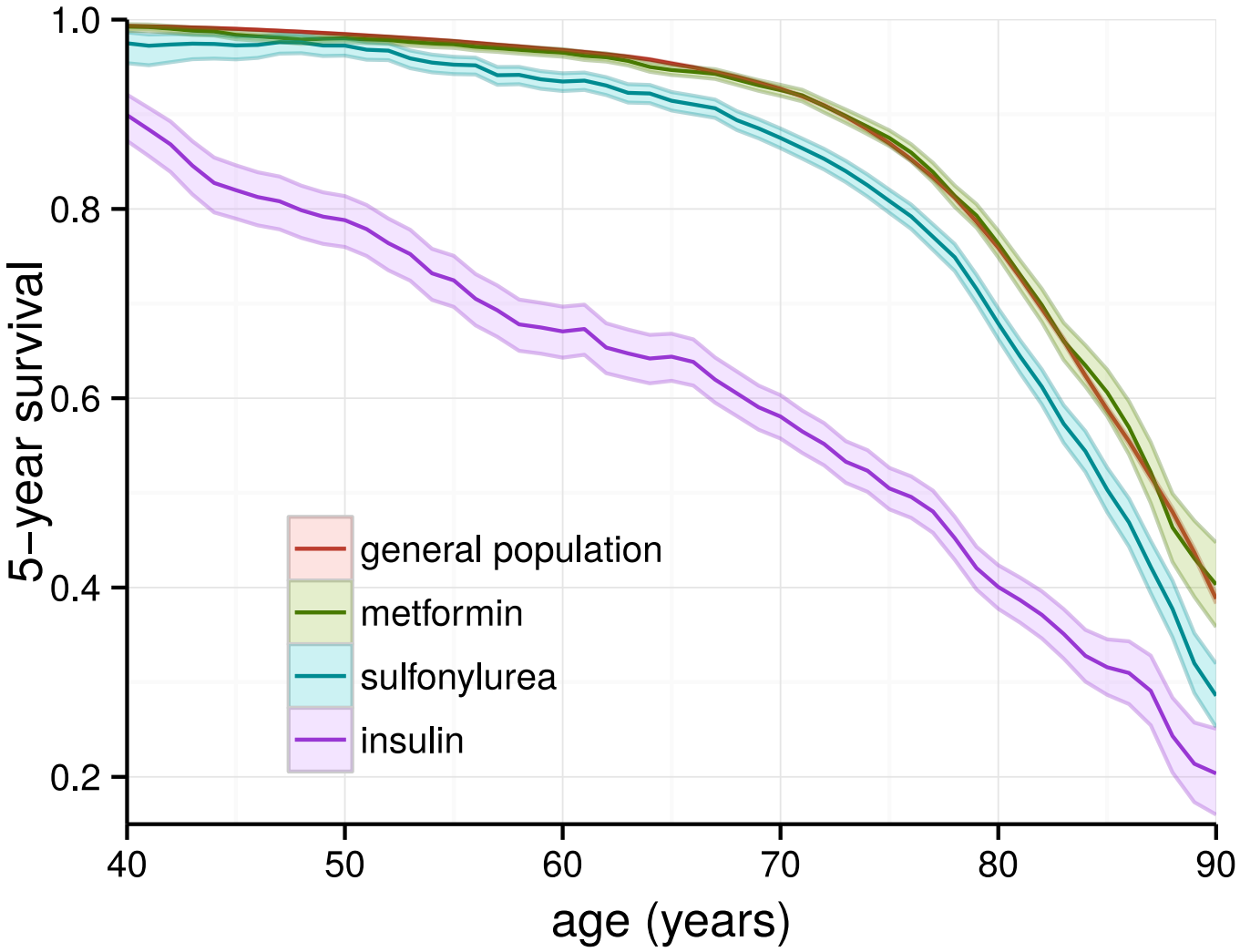
(Naam en voornaam)

Survival analysis based on health expenditure data

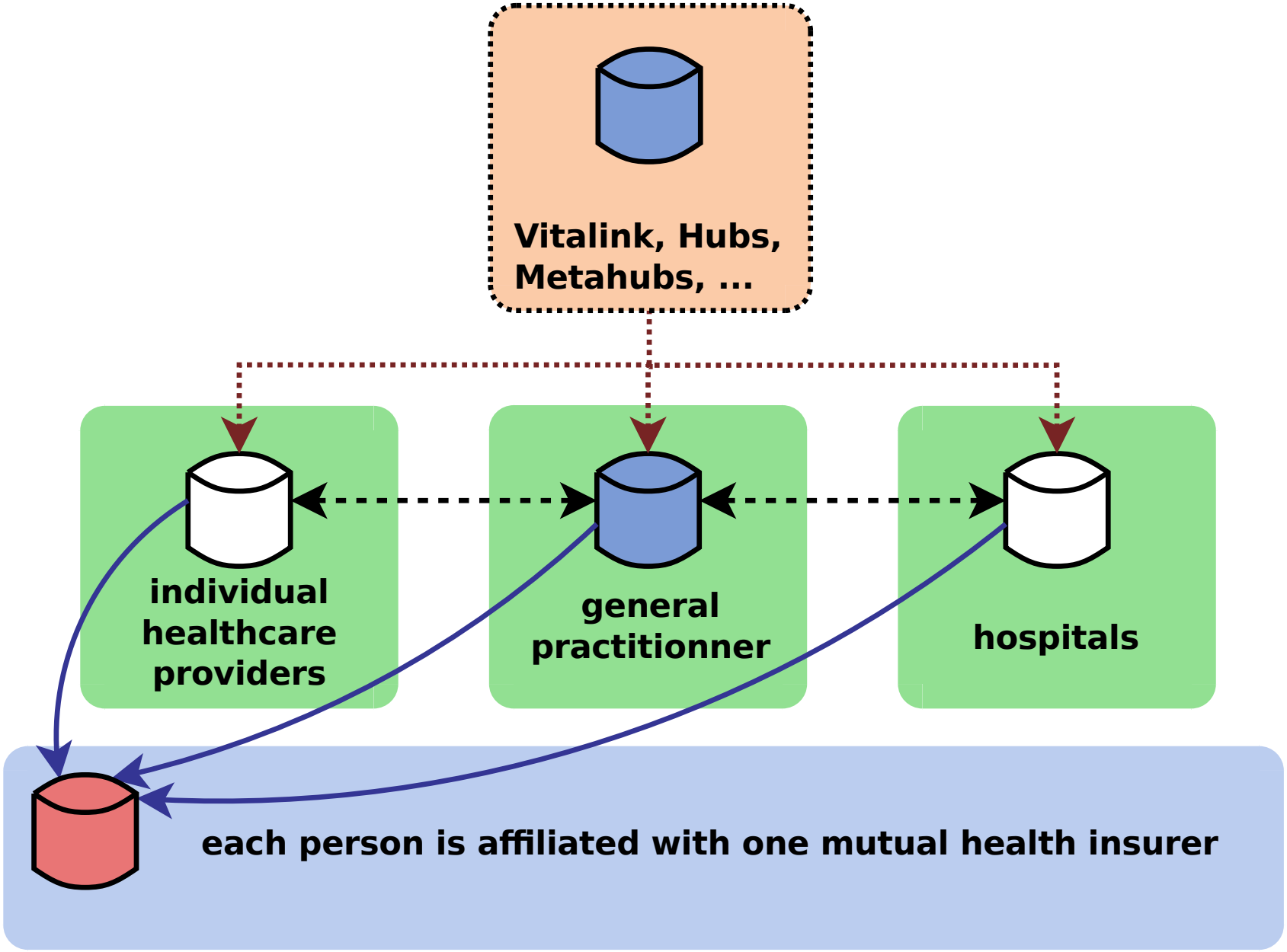


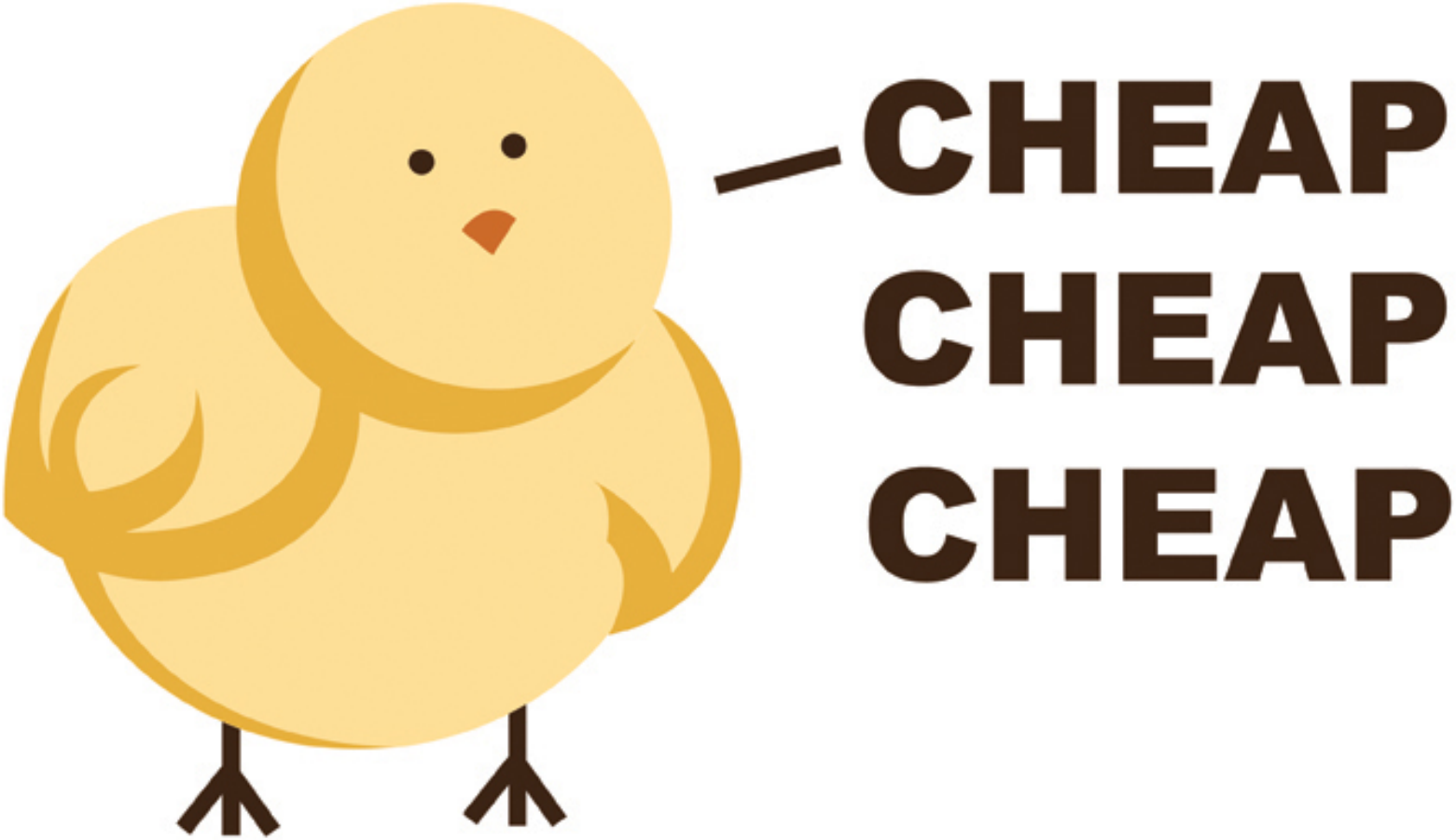
Marc Claesen†, Pieter Gillard†, Frank De Smet, Michael Callens, Bart De Moor & Chantal Mathieu (2015).
Mortality in individuals treated with glucose lowering agents: a large, controlled cohort study.
Provisionally accepted with minor revisions at *Journal of Clinical Endocrinology & Metabolism (JCEM)*.

5-year survival by age via health expenditure data



Marc Claesen†, Pieter Gillard†, Frank De Smet, Michael Callens, Bart De Moor & Chantal Mathieu (2015).
Mortality in individuals treated with glucose lowering agents: a large, controlled cohort study.
Provisionally accepted with minor revisions at *Journal of Clinical Endocrinology & Metabolism (JCEM)*.


Vitalink, Hubs, Metahubs, ...
individual healthcare providers
general practitionner
hospitals
each person is affiliated with one mutual health insurer


CHEAP
CHEAP
CHEAP

Screening for T2D using health expenditure data

Advantages:

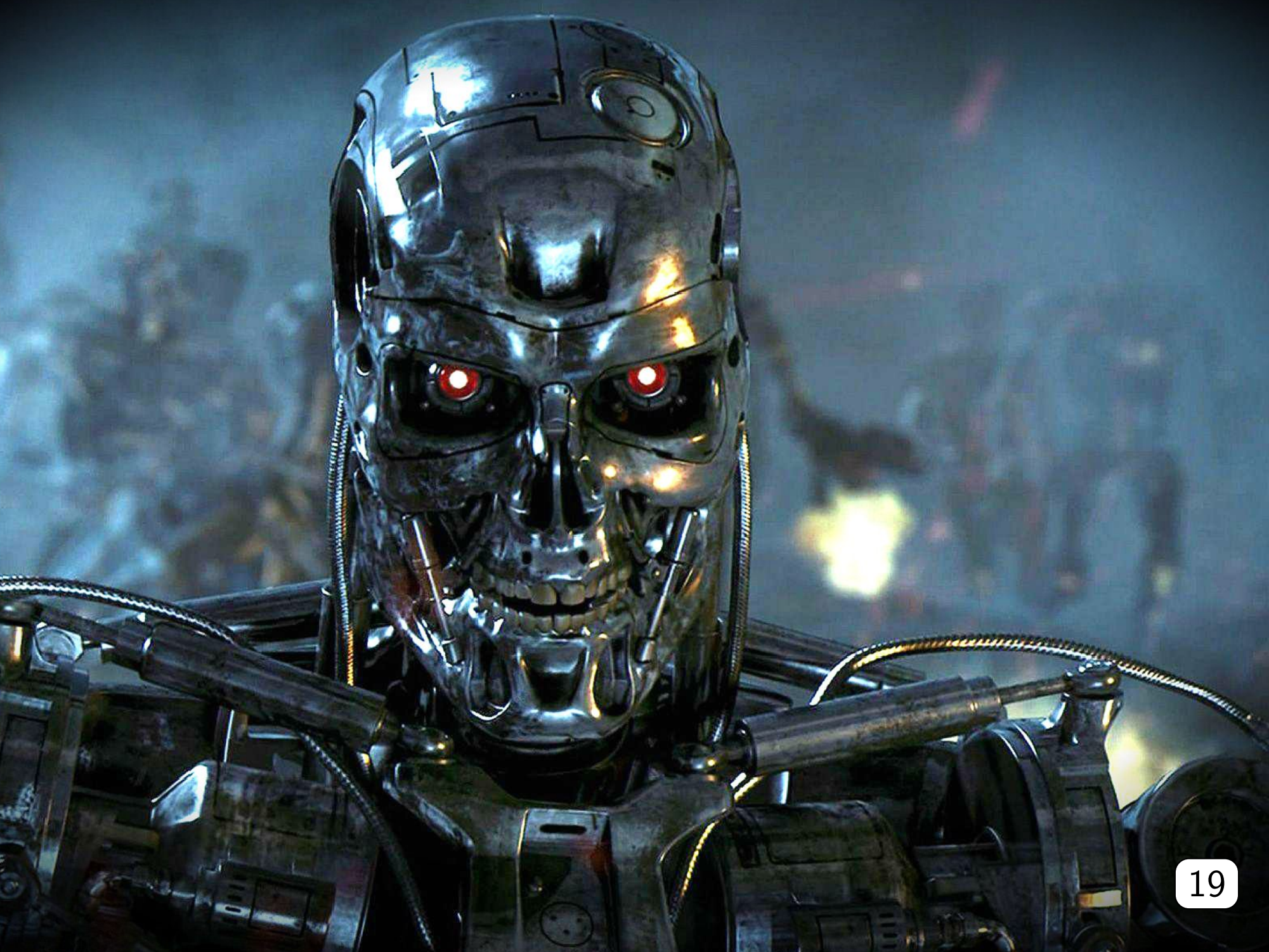
- long-term longitudinal overview of patients' medical history
- diabetics identifiable via routine use of glucose-lowering agents
- screening would be essentially free, as data is already available

Disadvantages:

- lack of info related to several important known risk factors
- some false positives are induced by labeling via GLAs

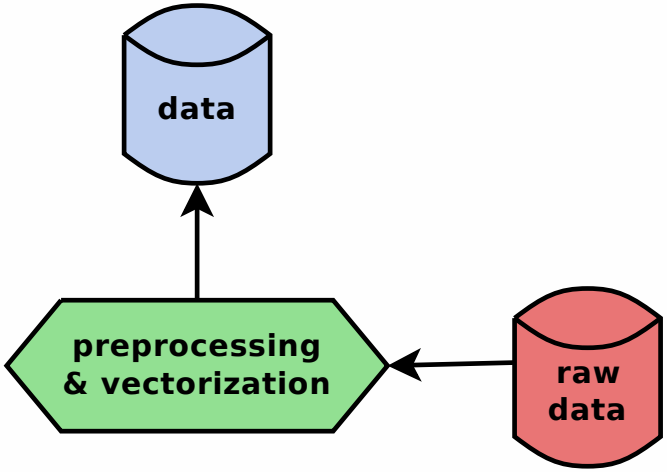
Key challenge: impossible to identify non-diabetics
$\rightarrow$ requires special learning methods (no known negatives)

Outline

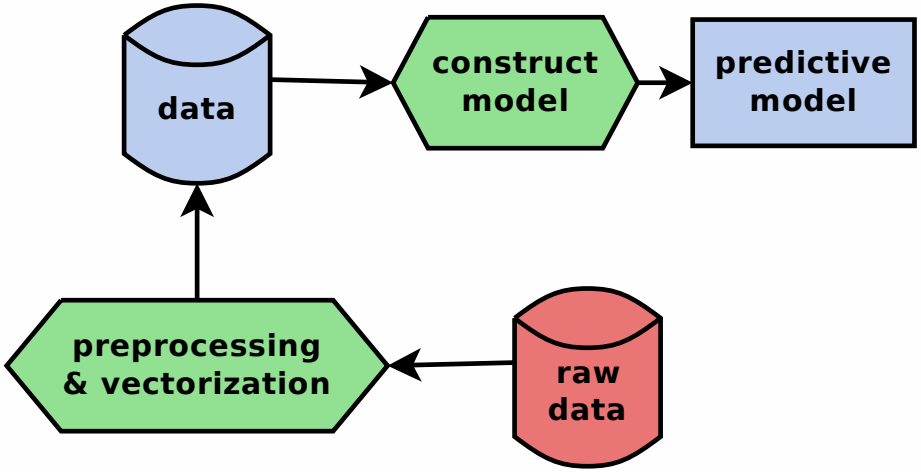


Machine learning pipeline

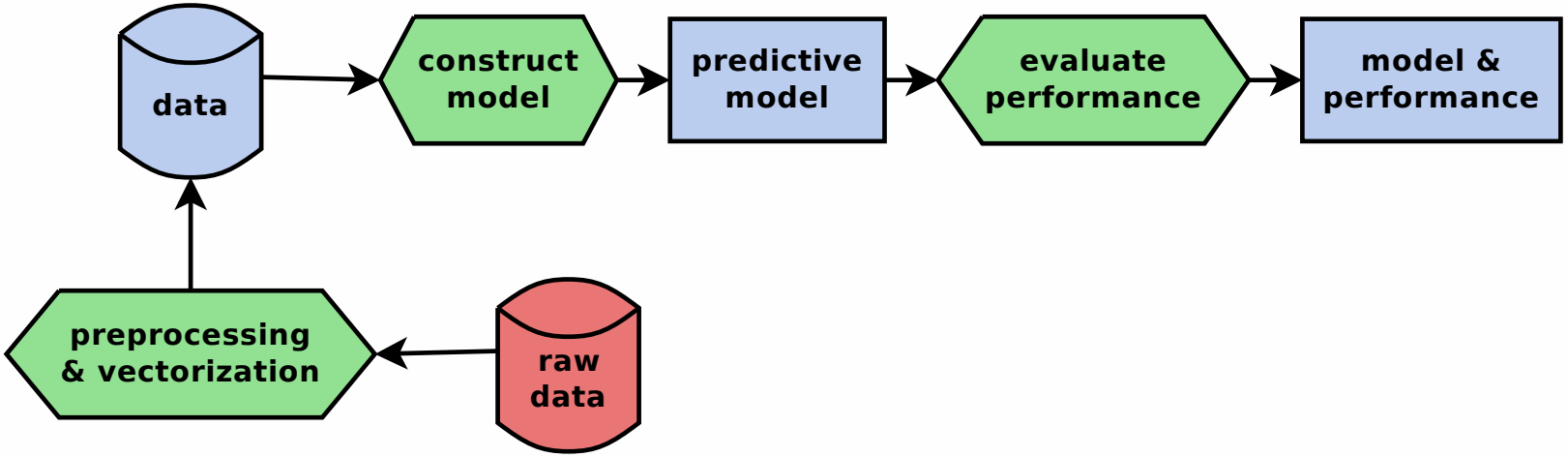
Machine learning pipeline



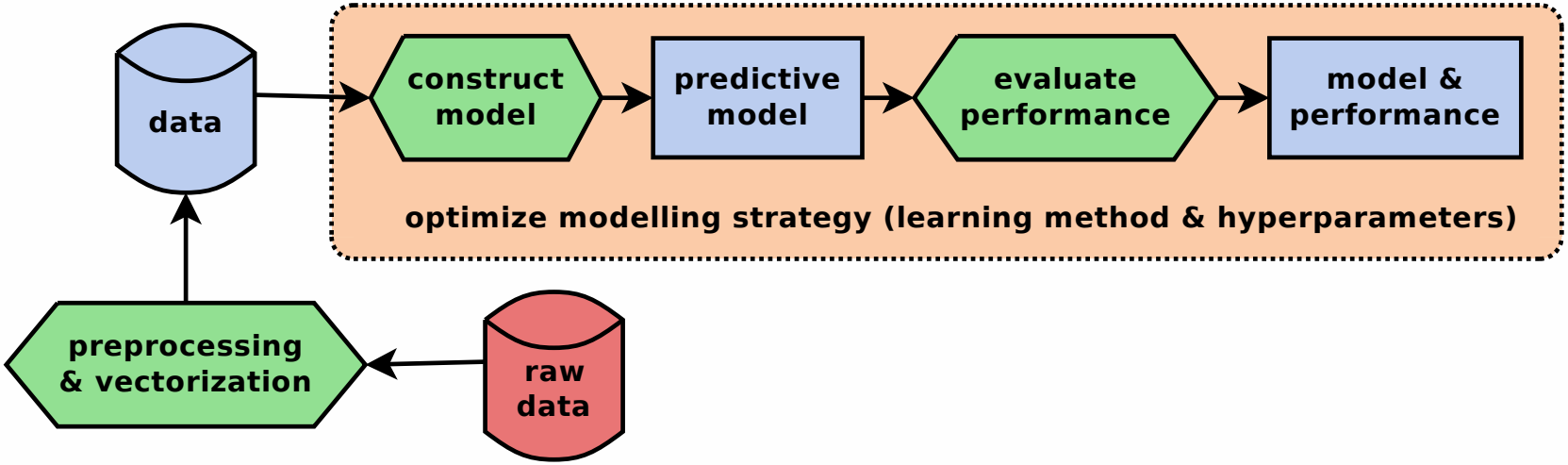
Machine learning pipeline



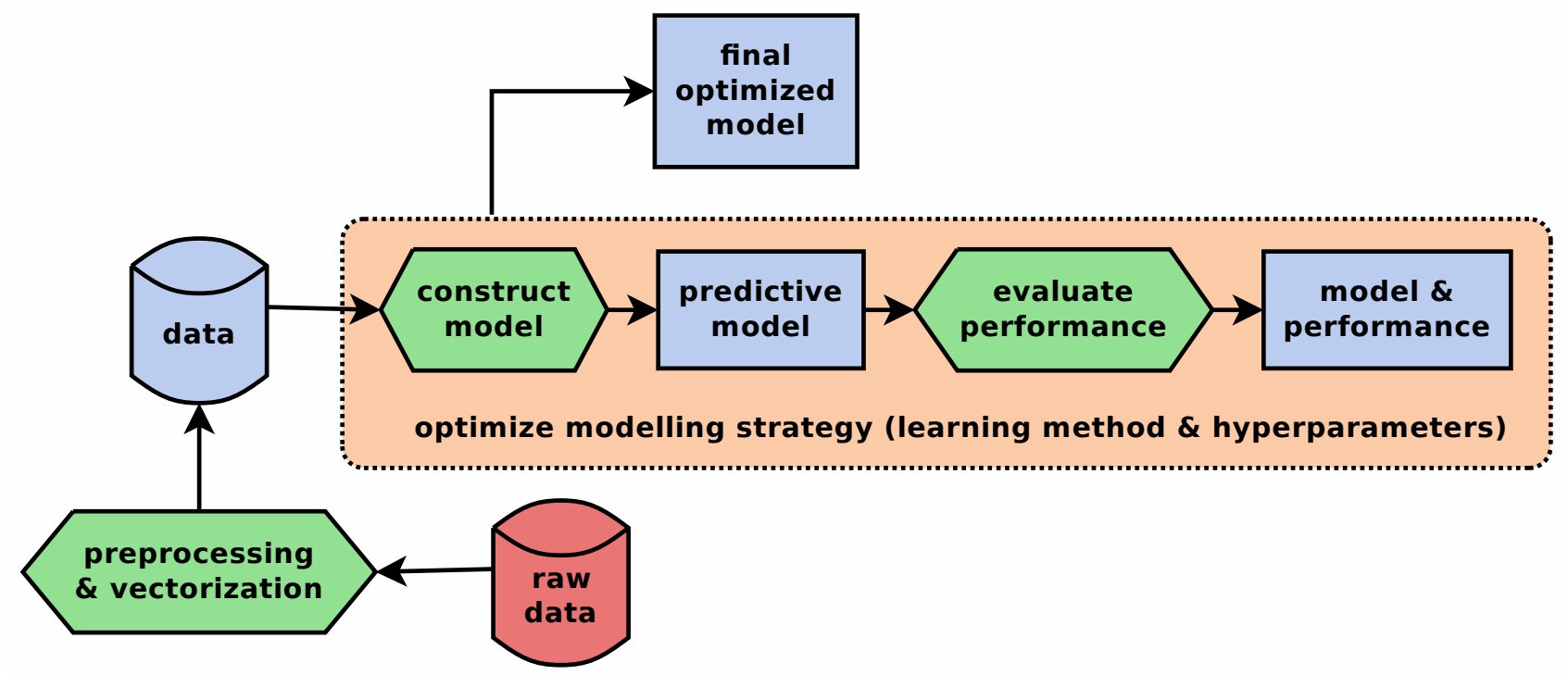
Machine learning pipeline



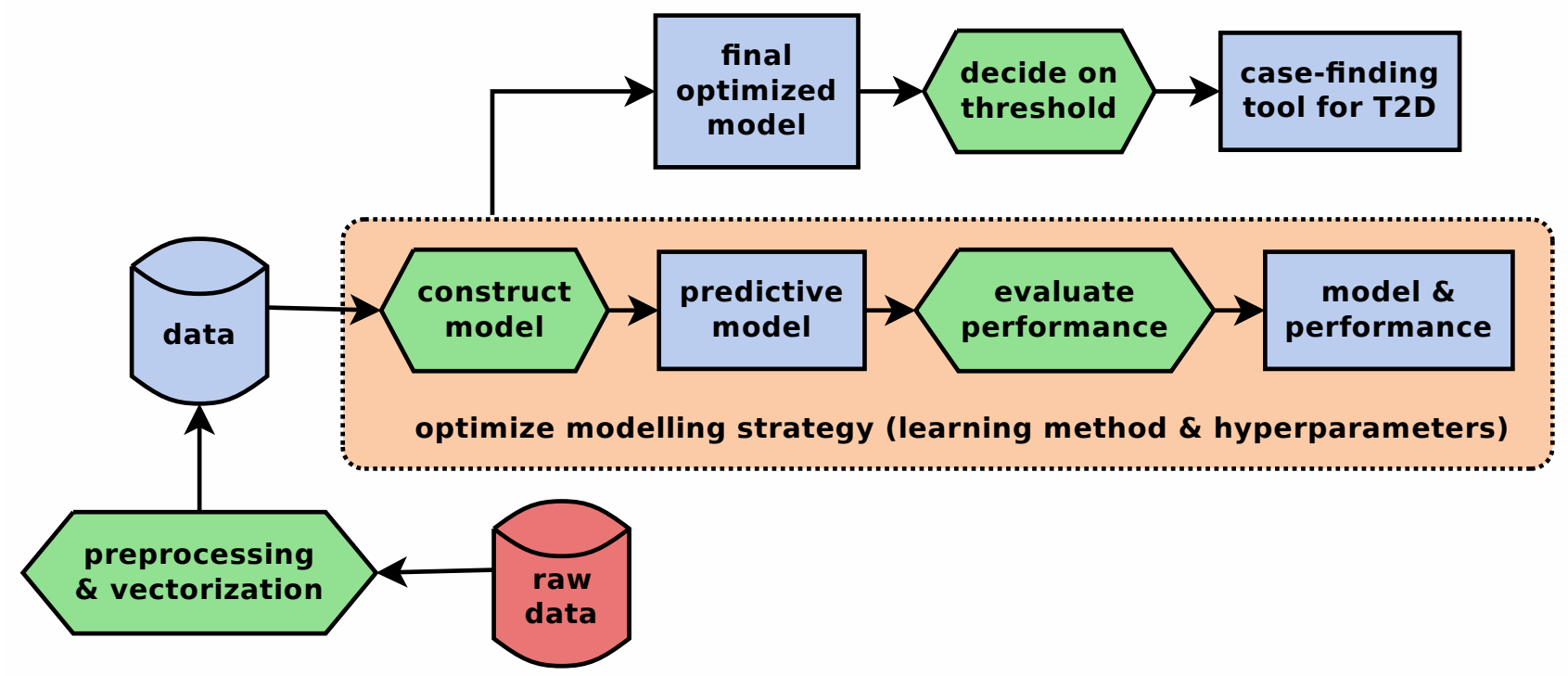
Machine learning pipeline



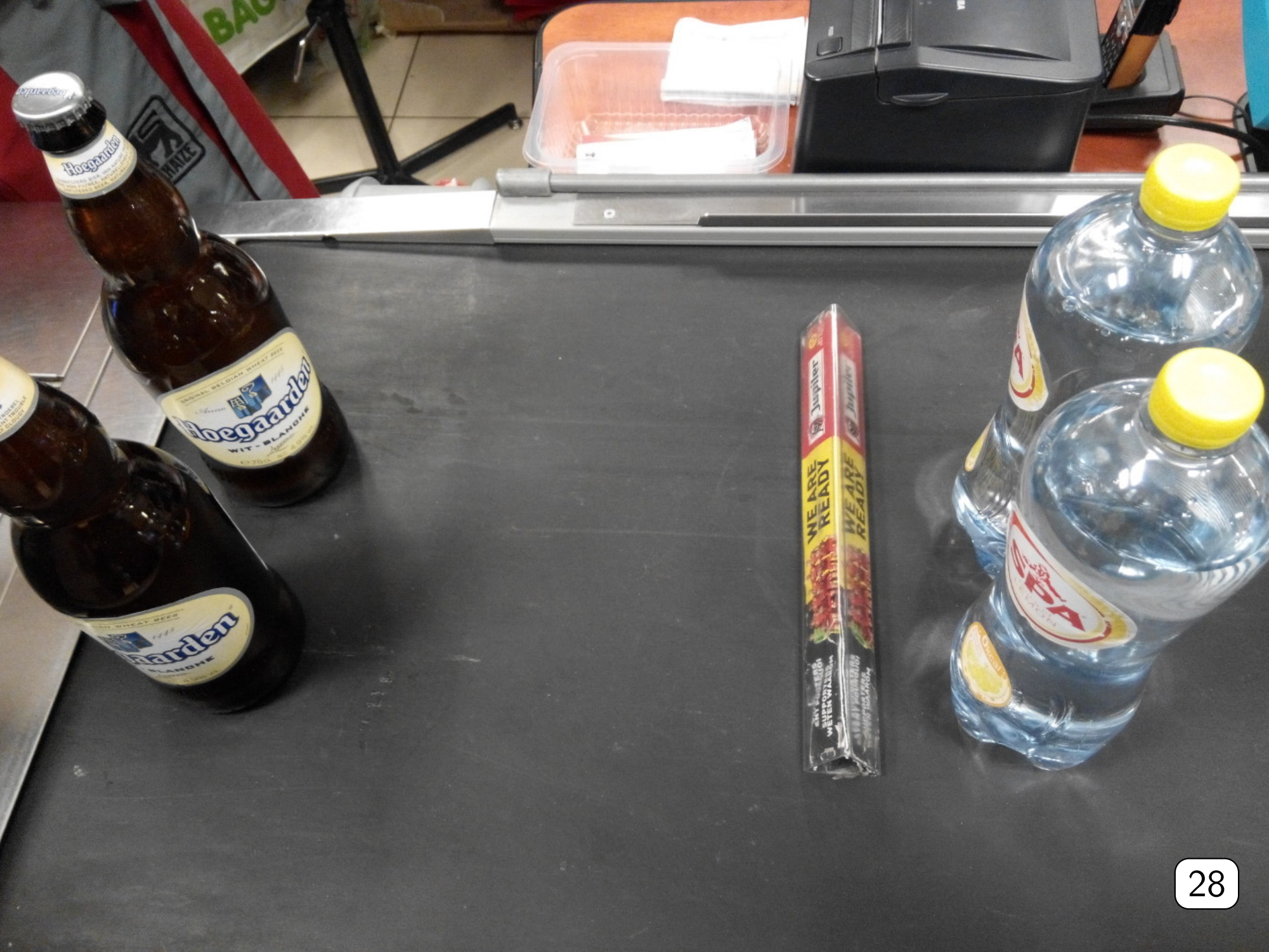
Machine learning pipeline



Machine learning pipeline

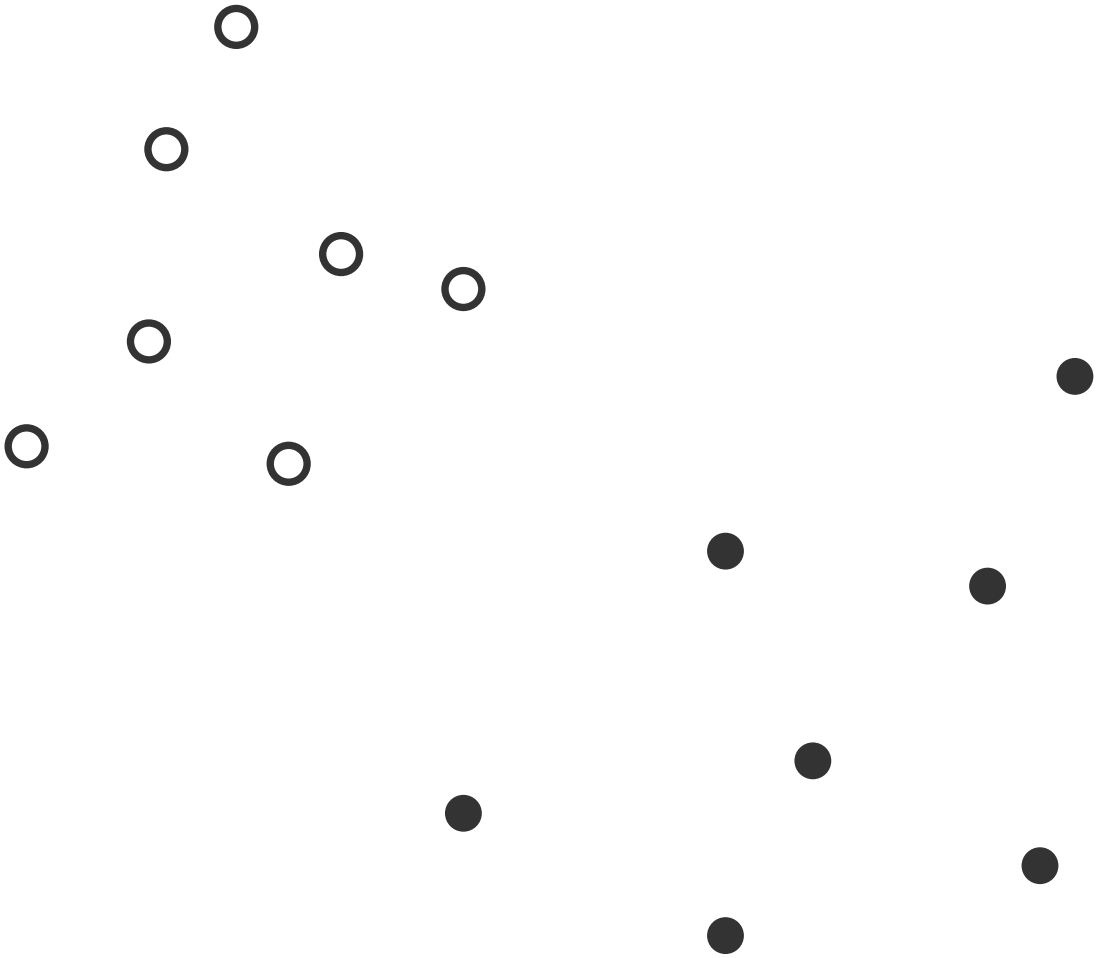


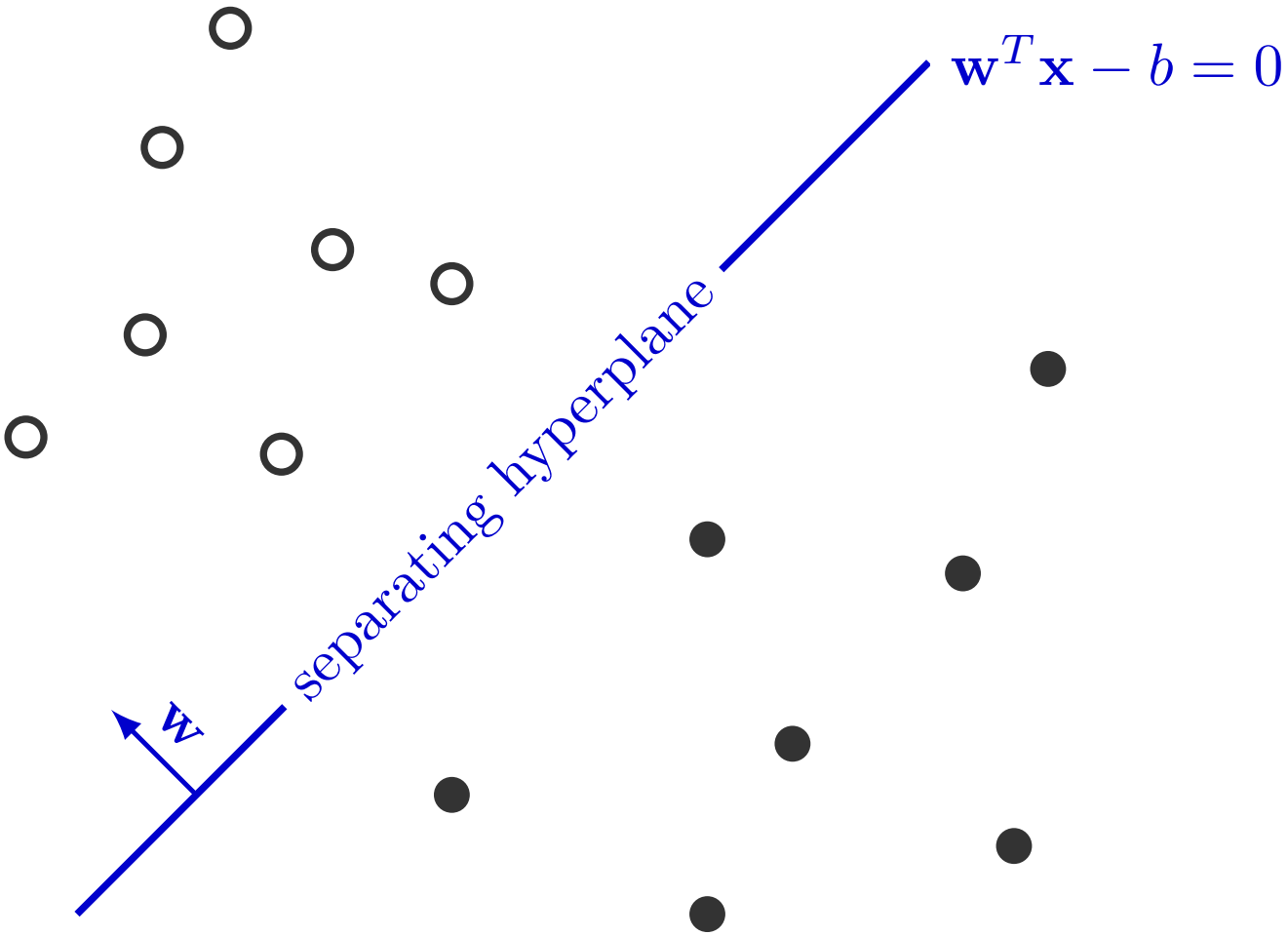

COM QUEIJO PARMIGGIANO REGGIANO
e molho vinaigrette
Carne de Vitela
CARPACCIO
CARNE DE VITELA MARINADA
Entrada para 2 pessoas
Fatiado fino
Pronto para servir
DC
DELICORTE
Ekro
Hoegaarden
WIT - BLANCHE

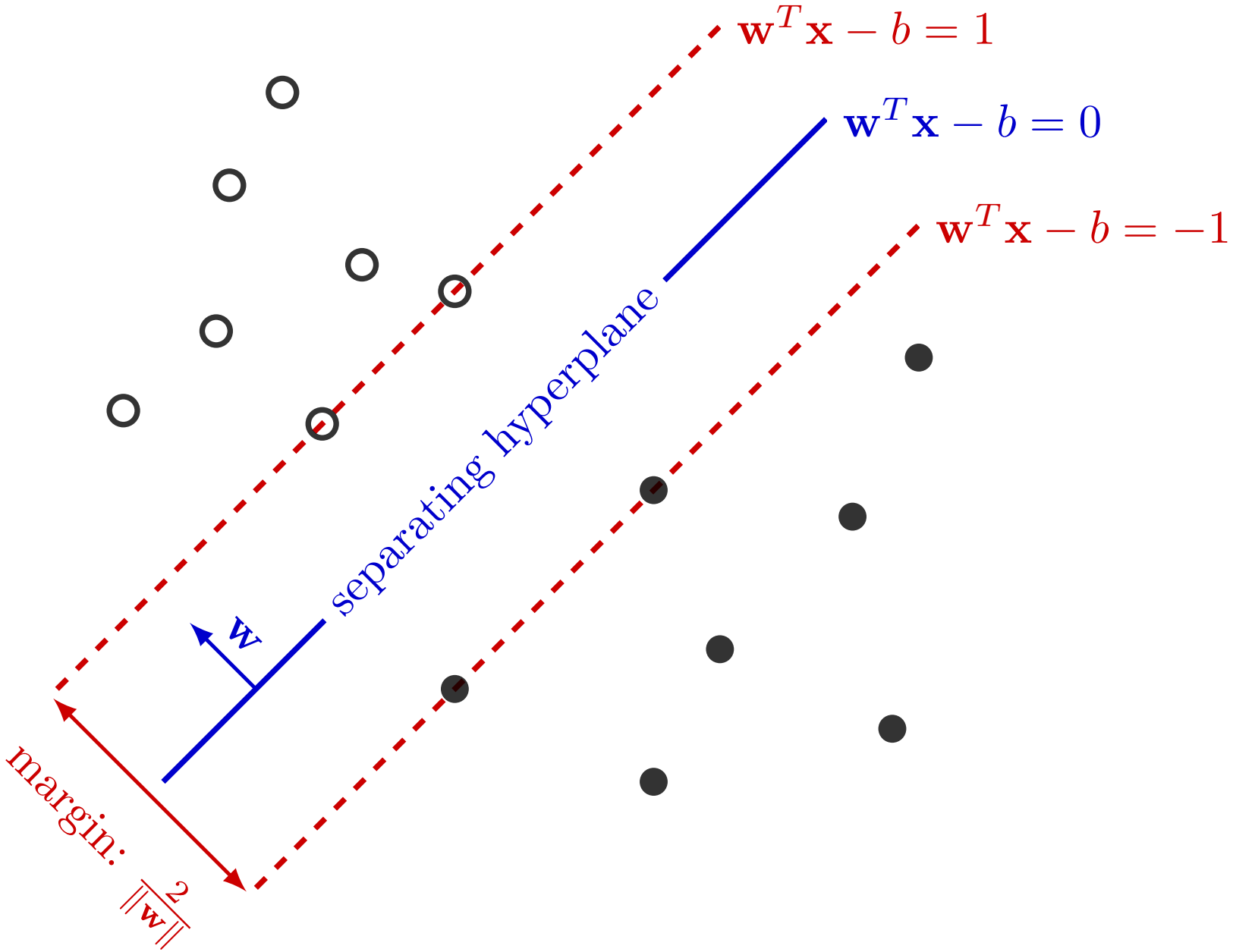

Hoegaarden
WIT - BLANCHE
Jupiler
WE ARE READY
SPA

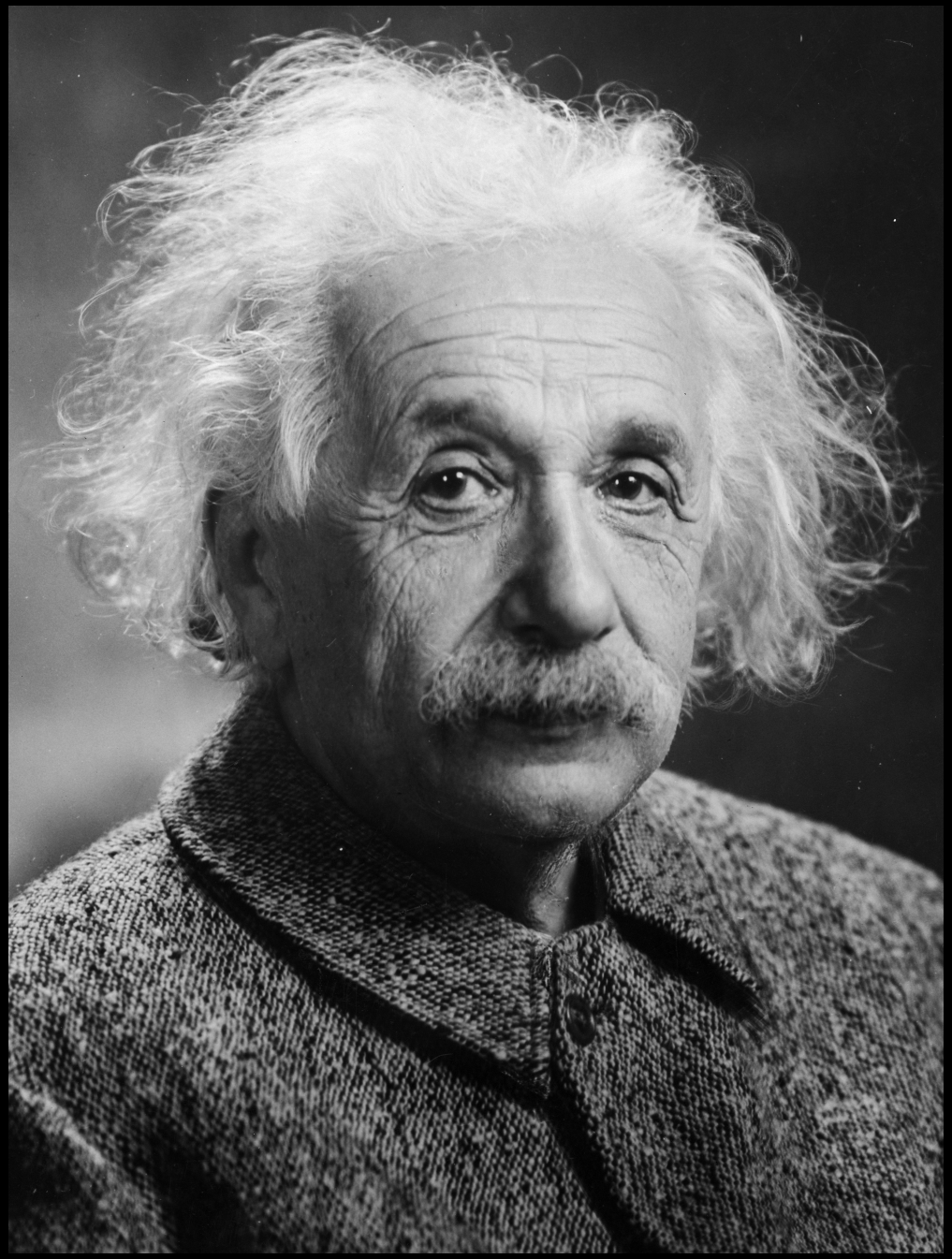

Hoegaarden
WIT · BLANCHE
Jupiler
WE ARE READY
SPA

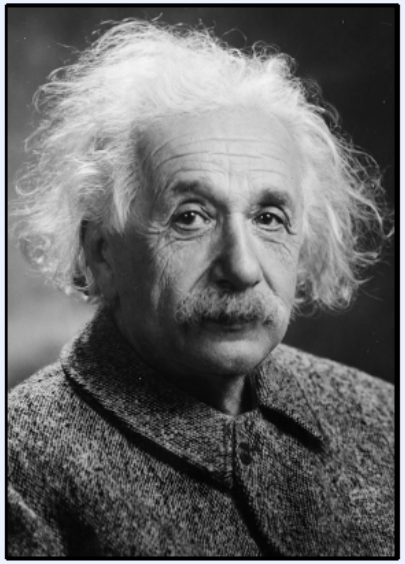

Hoegaarden
WIT - BLANCHE
WE ARE READY
Jupiler
SPA




$\mathbf{w}^T\mathbf{x} - b = 0$
separating hyperplane
$\mathbf{w}$

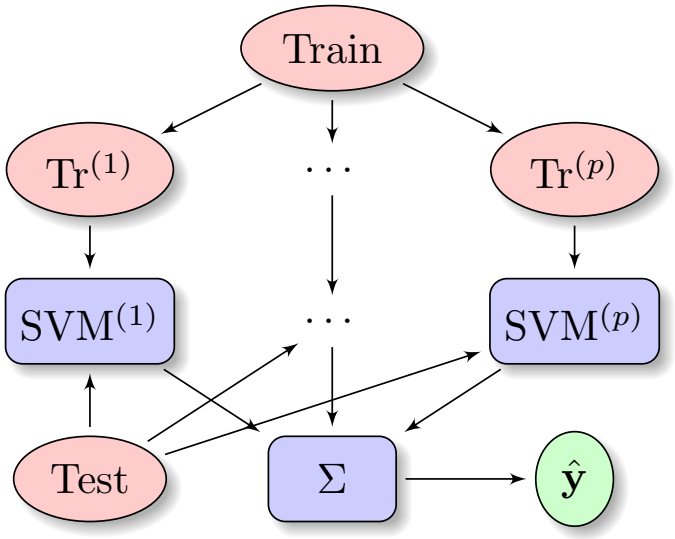

$\mathbf{w}^T\mathbf{x} - b = 1$
$\mathbf{w}^T\mathbf{x} - b = 0$
$\mathbf{w}^T\mathbf{x} - b = -1$
separating hyperplane
$\mathbf{w}$
margin: $\frac{2}{\|\mathbf{w}\|}$





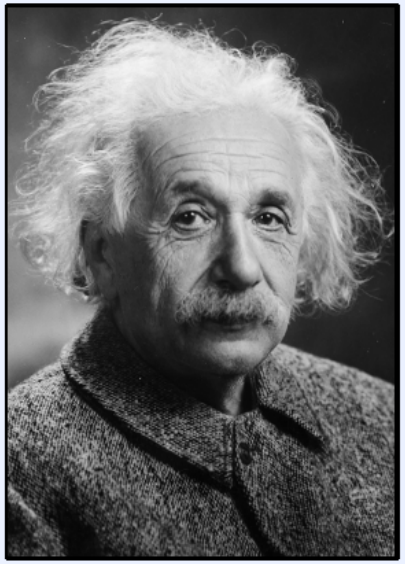
Ensemble learning with SVM base models

- aggregate many SVM models trained on small resamples
- facilitates nonlinear learning on large-scale data sets
- resulting models are robust to label noise

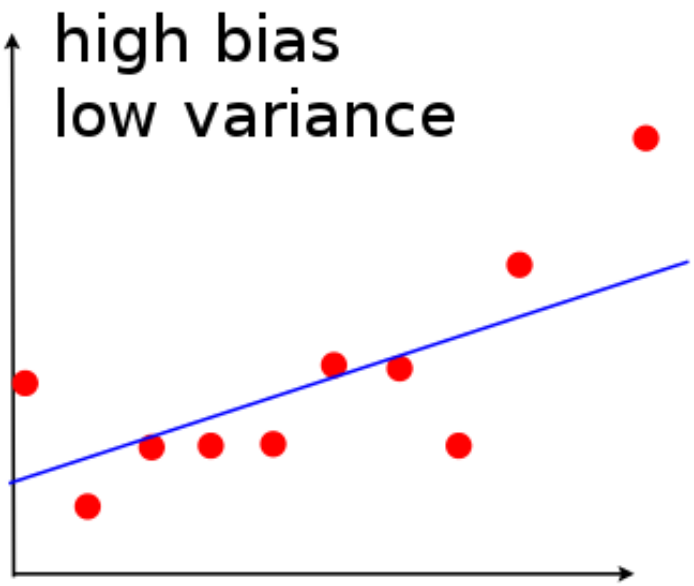


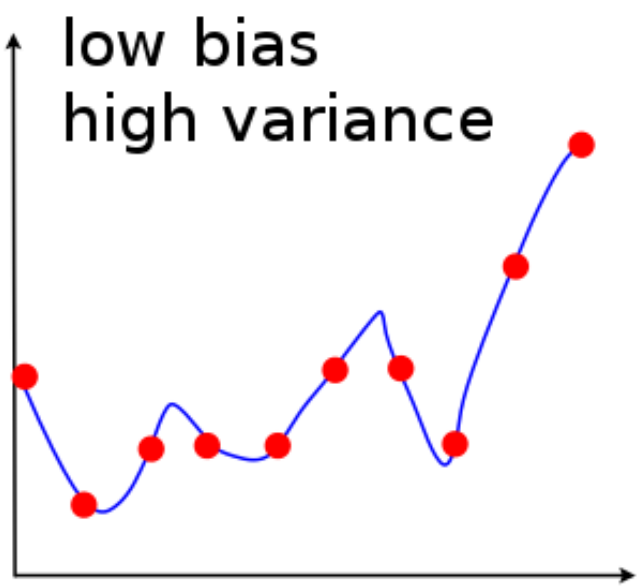
Marc Claesen, Frank De Smet, Johan Suykens & Bart De Moor (2014). *EnsembleSVM: A library for ensemble learning using support vector machines*. Journal of Machine Learning Research, 15(1), 141-145. Publication available at `http://www.jmlr.org/papers/volume15/claesen14a/claesen14a.pdf`.

Marc Claesen, Frank De Smet, Johan Suykens & Bart De Moor (2015). *A robust ensemble approach to learn from positive and unlabeled data using SVM base models*. Neurocomputing, 160, 73-84. Publication available at `http://dx.doi.org/10.1016/j.neucom.2014.10.081`.



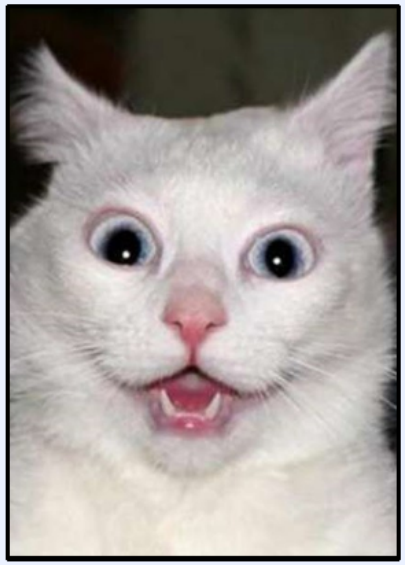


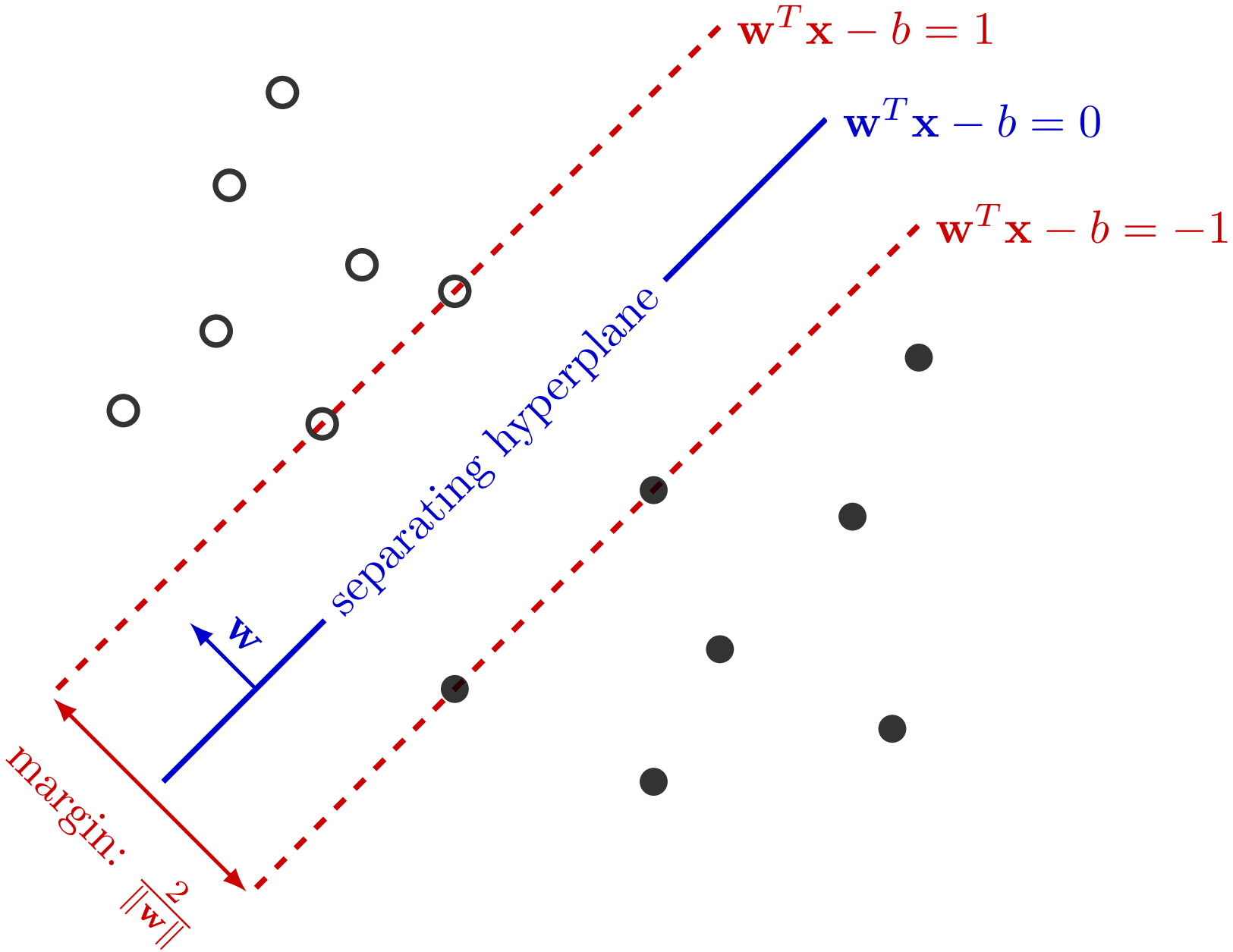


high bias
low variance

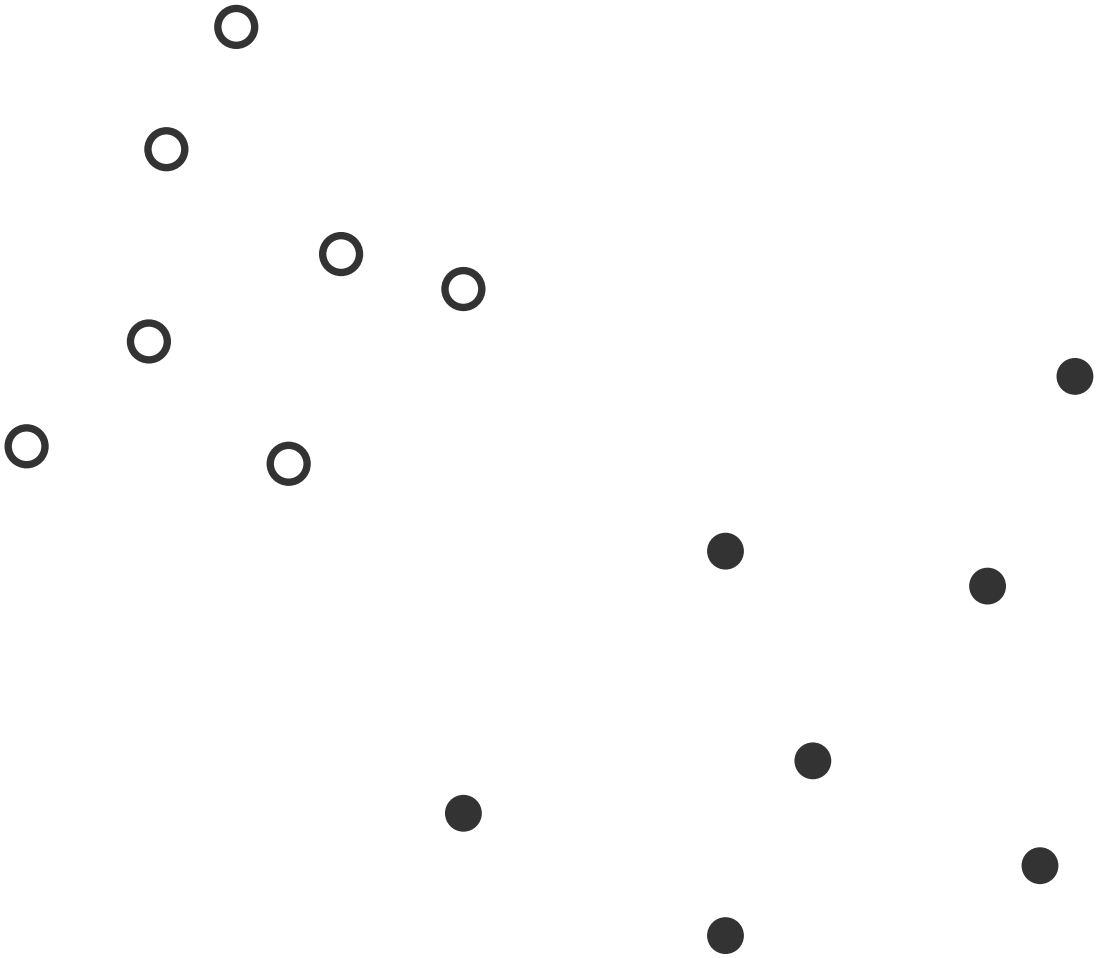

low bias
high variance

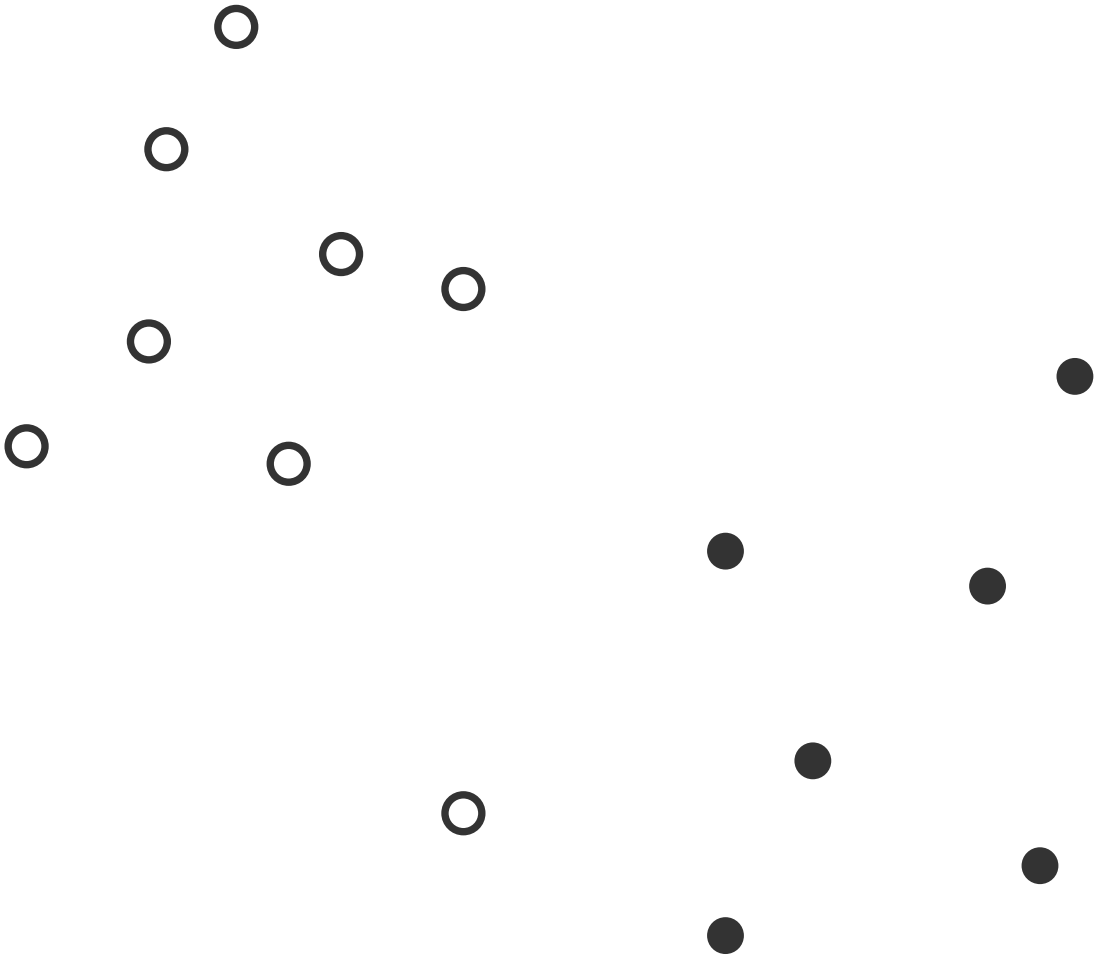


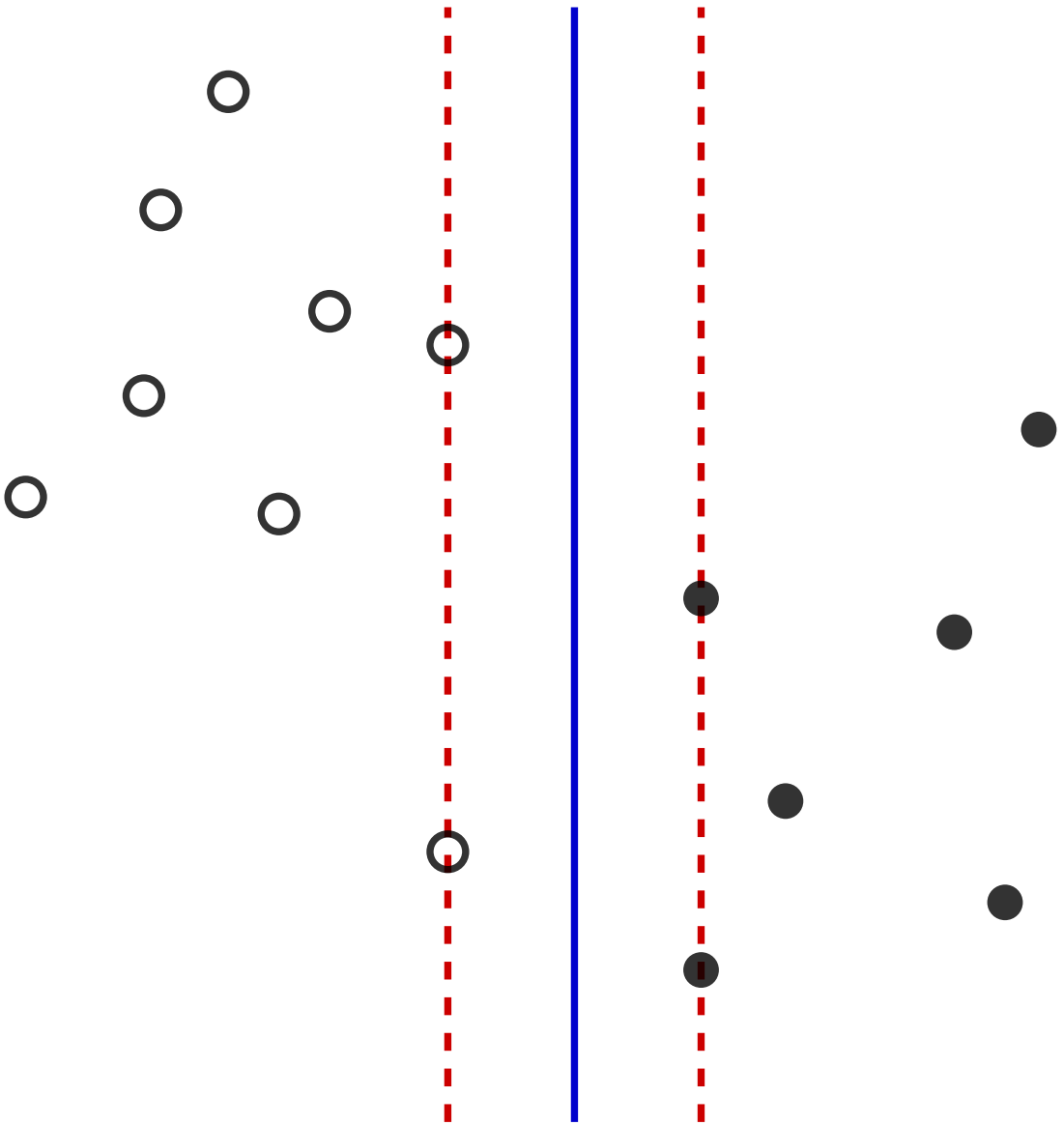


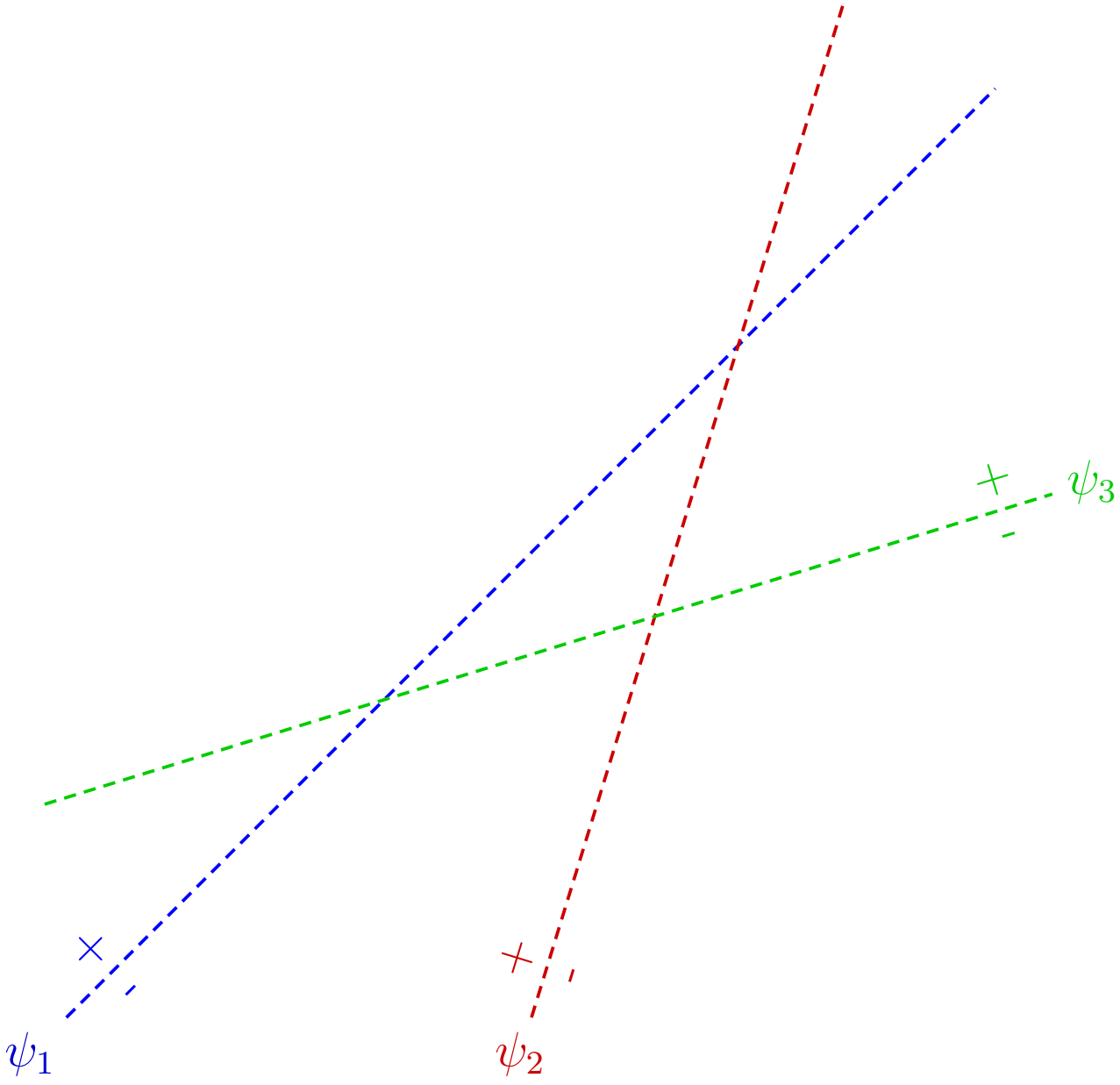



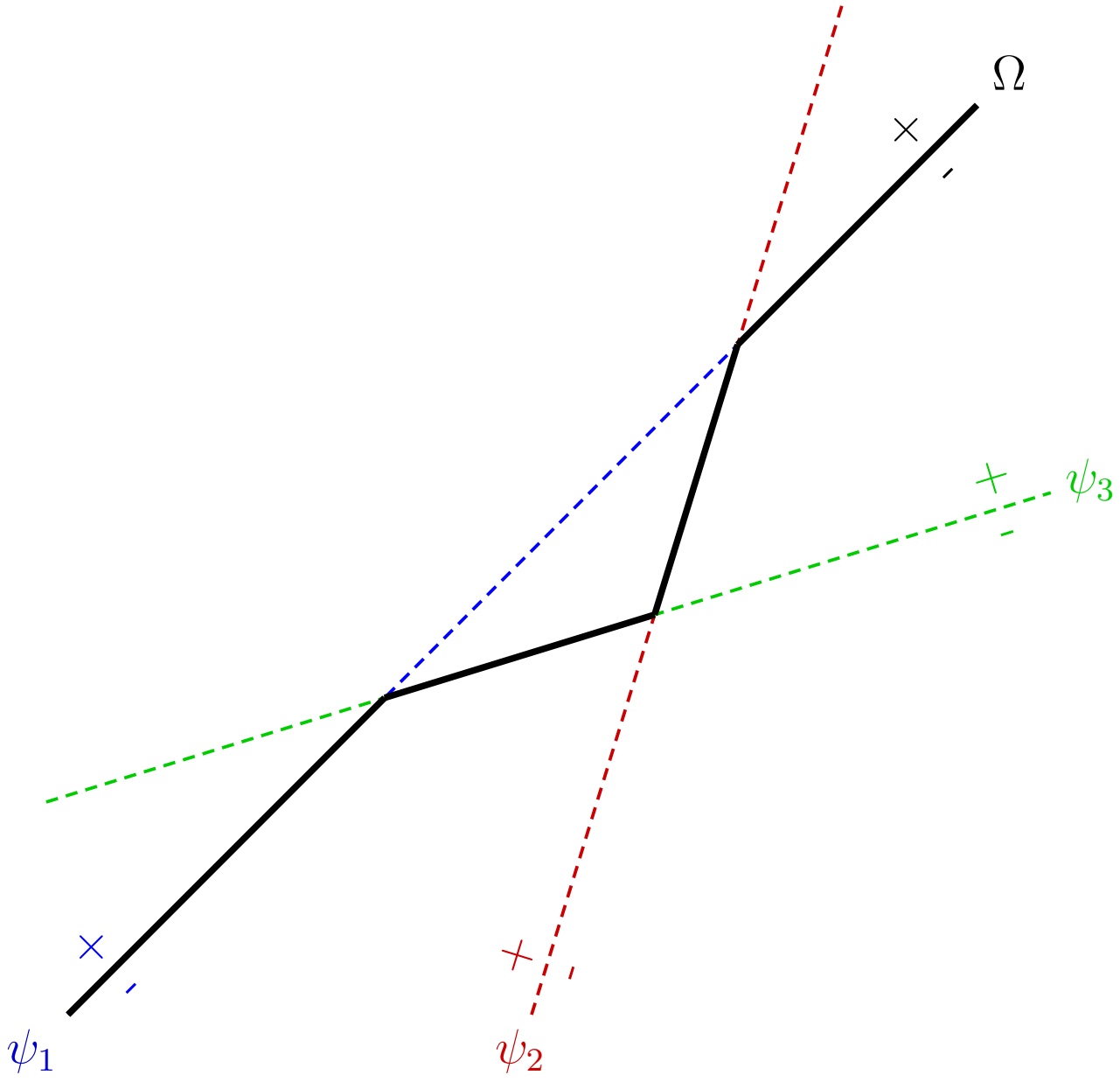

$\mathbf{w}^T\mathbf{x} - b = 1$
$\mathbf{w}^T\mathbf{x} - b = 0$
$\mathbf{w}^T\mathbf{x} - b = -1$
separating hyperplane
$\mathbf{w}$
margin: $\frac{2}{\|\mathbf{w}\|}$

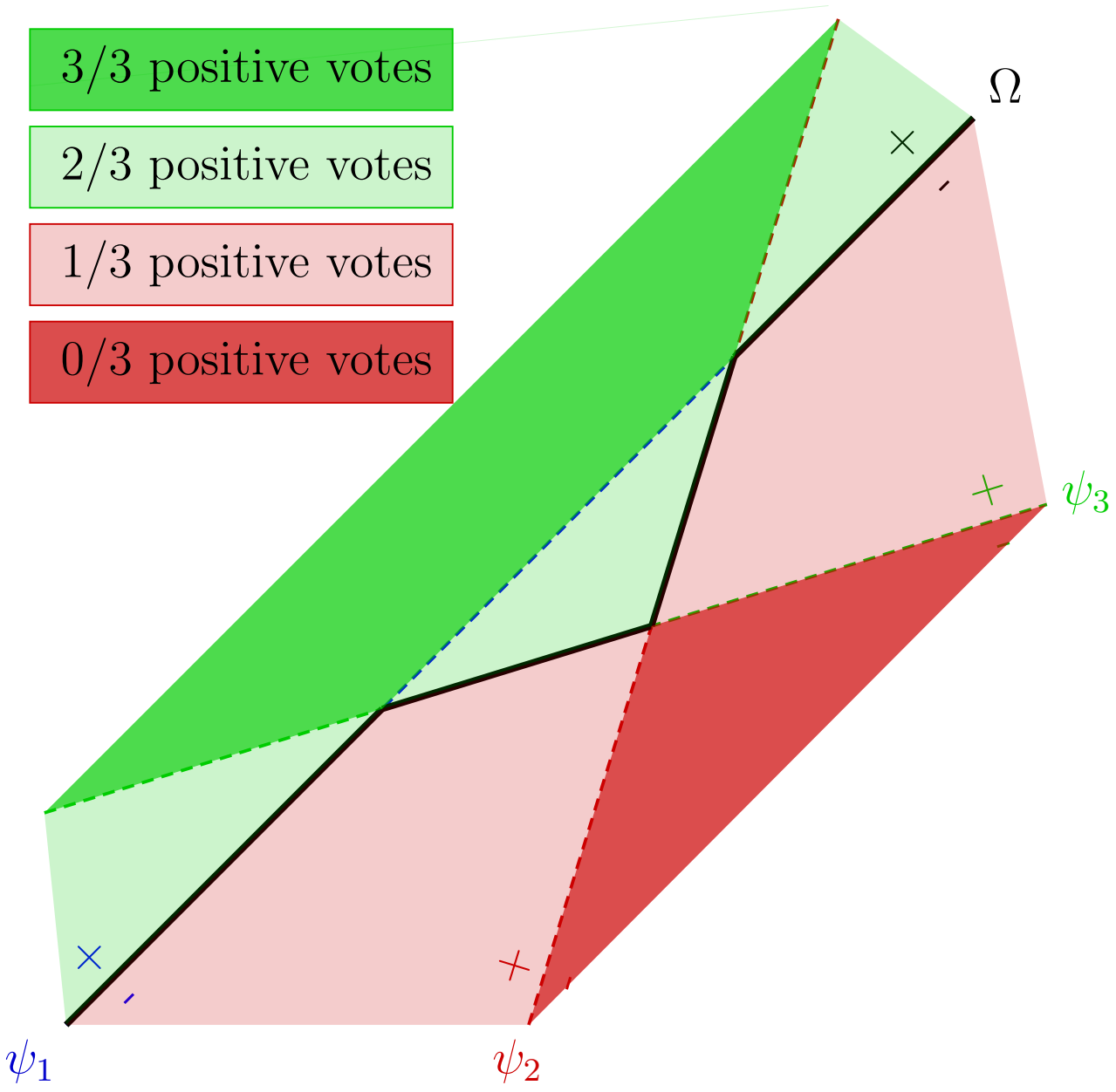




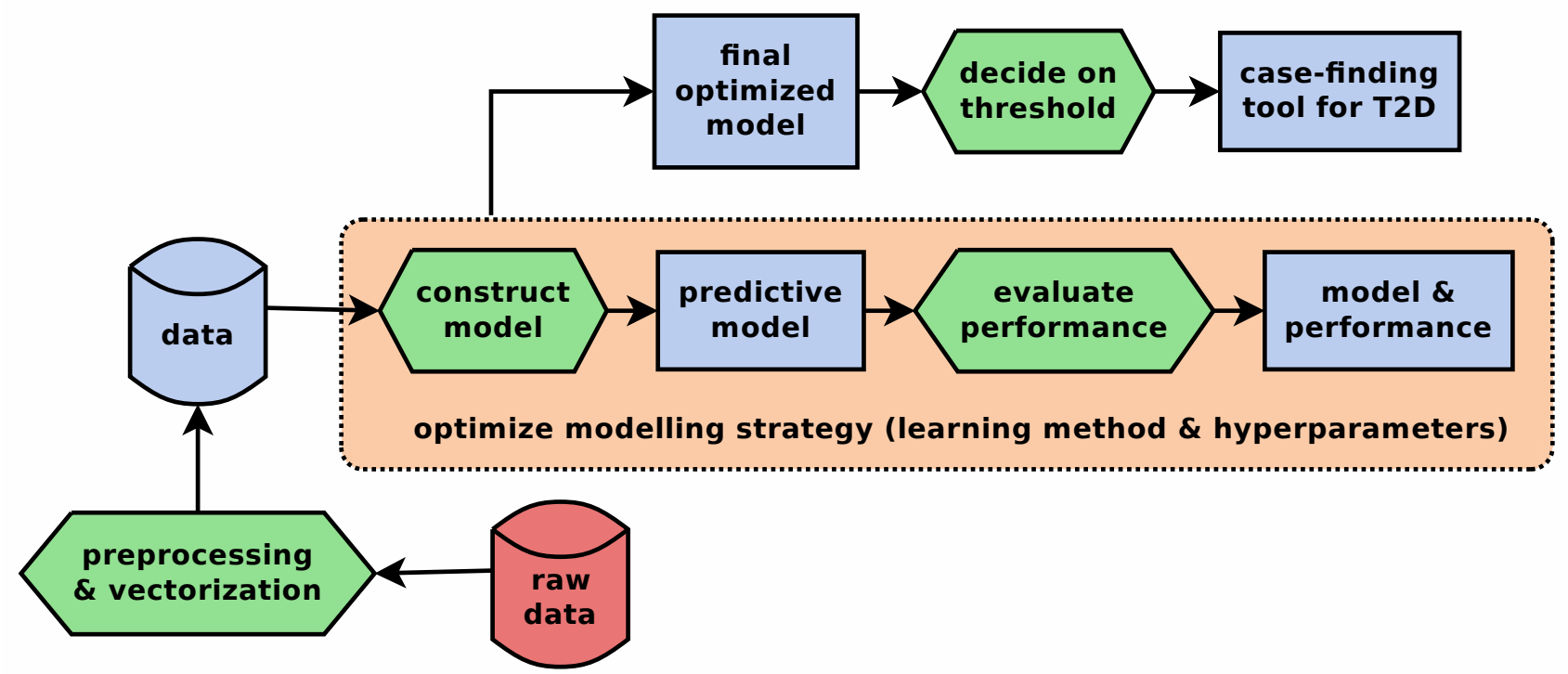



ψ_1
ψ_2
ψ_3


Ω
ψ_3
ψ_1
ψ_2


3/3 positive votes
2/3 positive votes
1/3 positive votes
0/3 positive votes
Ω
ψ_3
ψ_1
ψ_2

Machine learning pipeline



	example's true label is positive	example's true label is negative
model predicts positive	*true positive*	*false positive*
model predicts negative	*false negative*	*true negative*

	example's true label is positive	example's true label is negative
model predicts positive	*true positive*	*false positive*
model predicts negative	*false negative*	*true negative*

- 0.1
+ 0.8
- 0.2
+ 0.3
+ 0.9
- 0.7

	example's true label is positive	example's true label is negative
model predicts positive	*true positive*	*false positive*
model predicts negative	*false negative*	*true negative*

-	0.1
+	0.8
-	0.2
+	0.3
+	0.9
-	0.7

$$\mathcal{R} = \{ \; 0.9 \quad 0.8 \quad 0.7 \quad 0.3 \quad 0.2 \quad 0.1 \; \}$$

	example's true label is positive	example's true label is negative
model predicts positive	*true positive*	*false positive*
model predicts negative	*false negative*	*true negative*

-	0.1
+	0.8
-	0.2
+	0.3
+	0.9
-	0.7

$$\mathcal{R} = \{ \underbrace{0.9 \quad 0.8 \quad 0.7} \quad 0.3 \quad 0.2 \quad 0.1 \}$$

2 true positives	1 false positives
1 false negatives	2 true negatives

	example's true label is positive	example's true label is negative
model predicts positive	*true positive*	*false positive*
model predicts negative	*false negative*	*true negative*

-	0.1
+	0.8
-	0.2
+	0.3
+	0.9
-	0.7

$$\mathcal{R} = \{ \underbrace{0.9 \quad 0.8 \quad 0.7 \quad 0.3 \quad 0.2} \quad 0.1 \}$$

3 true positives	2 false positives
0 false negatives	1 true negatives

Performance evaluation without known negatives

Task: compute classifier performance without known negatives.

Commonly used (bad) approximation: treat unlabeled as negatives.

Performance evaluation without known negatives

Task: compute classifier performance without known negatives.

Commonly used (bad) approximation: treat unlabeled as negatives.

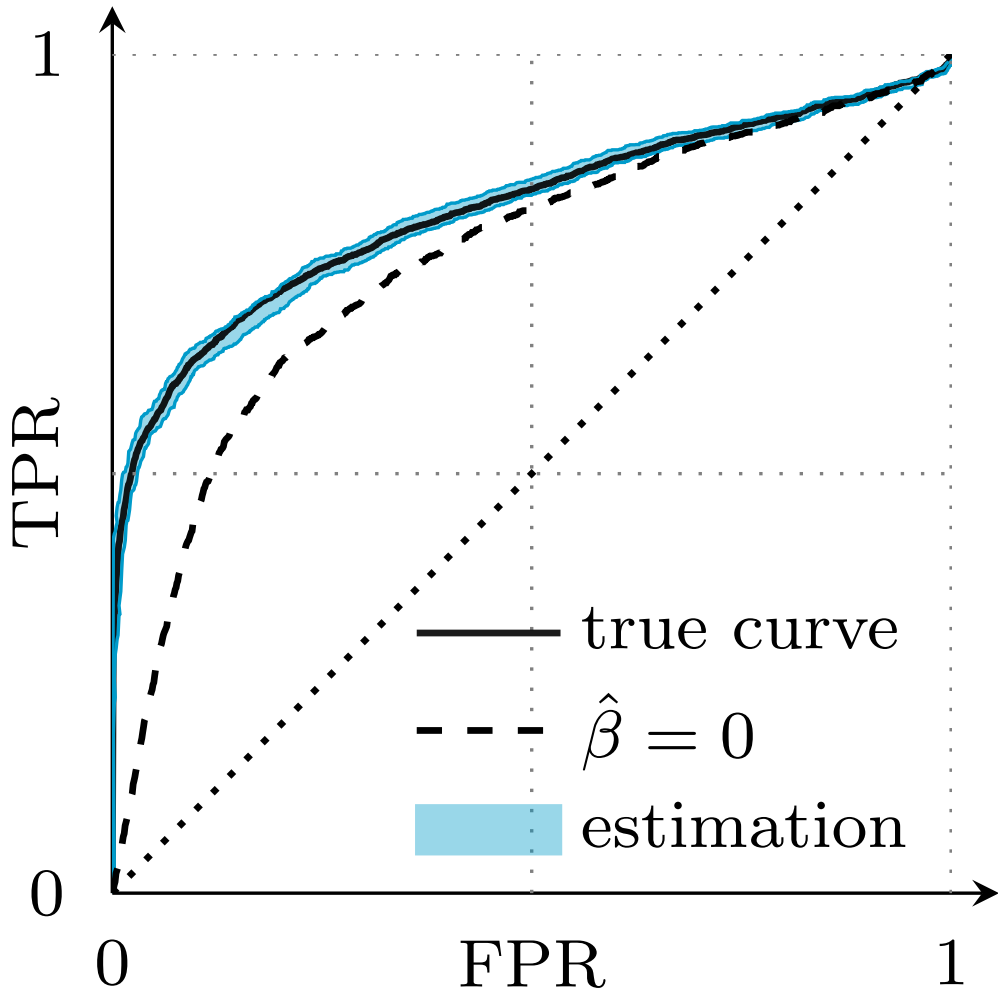
We developed a novel method:

- computes bounds on contingency tables (+ related metrics)
- assuming known positives are sampled at random
- given the fraction of latent positives in unlabeled set

Performance evaluation without negatives was uncharted territory.

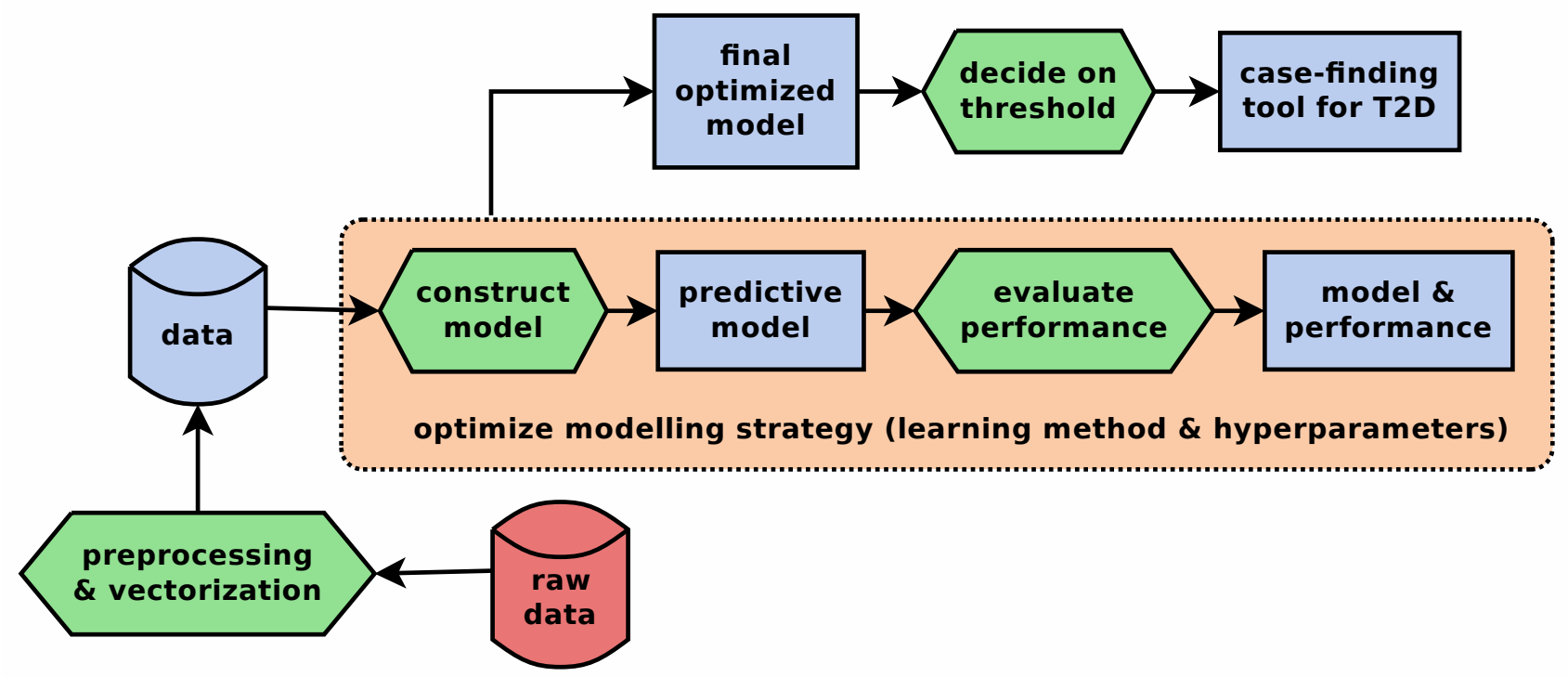
Marc Claesen, Jesse Davis, Frank De Smet & Bart De Moor (2015).
Assessing binary classifiers using only positive and unlabeled data.
Will be submitted to ACM SIGKDD in 2016.
Preprint available at `http://arxiv.org/abs/1504.06837`.

Example of performance bounds on ROC curve



Our approach accurately bounds the true performance.

Machine learning pipeline



Hyperparameter optimization aka "tuning"

Many learning algorithms are hyperparameterized.

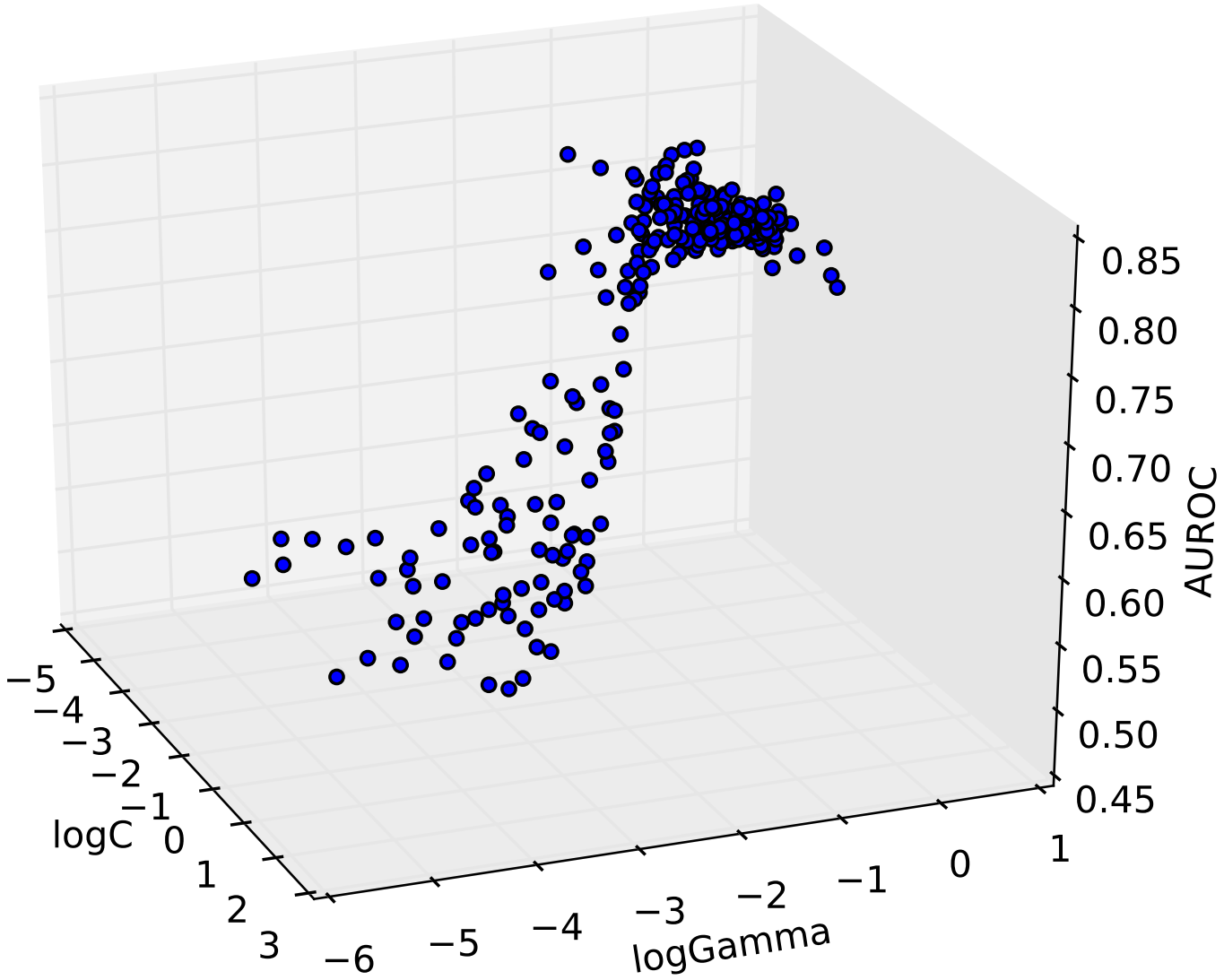
- regularization, learning rate, kernel bandwidth, ...

Suitable values must be found, but this is difficult.

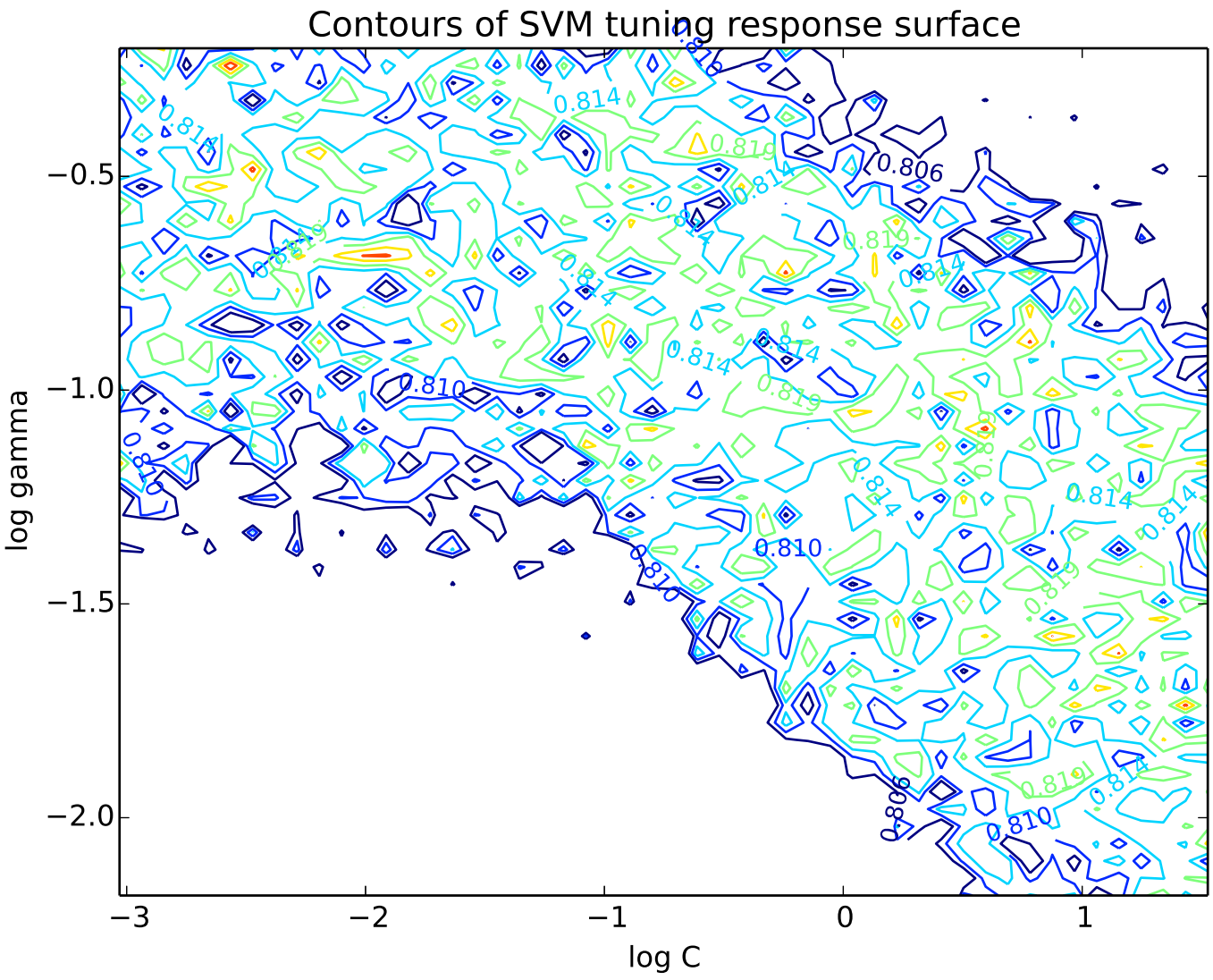
- non-convex, expensive, black-box objective function
- commonly done manually or via grid or random search

Marc Claesen & Bart De Moor (2015). *Hyperparameter Search in Machine Learning*.
In Proceedings of the 11th Metaheuristics International Conference (MIC), Agadir, Morocco.
Paper available at `http://arxiv.org/abs/1502.02127`.

Hyperparameter response surface of SVM with RBF kernel



Contours for tuning an SVM with RBF kernel



http://docs.optunity.net

Python library for automated hyperparameter optimization

- offers a wide variety of metaheuristic approaches
- design focus on easy deployment & intuitive API
- direct support for R, Julia, MATLAB & Octave
- $> 1,000$ downloads/month via Python package index

Marc Claesen, Jaak Simm, Dusan Popovic, Yves Moreau & Bart De Moor (2014). *Easy hyperparameter search using Optunity*. Under review at Journal of Machine Learning Research (4th revision ...). Preprint available at http://arxiv.org/abs/1412.1114.

Outline

Goal

Identify persons likely to start T2D therapy

- based exclusively on health expenditure data
- = patients with similar medical histories to known diabetics

Labeling:

- positives: patients that start routine use of GLAs
- negatives: not directly available
- discard medical history once diabetes therapy is started

Note: not all diabetics use GLAs (even when diagnosed!).

- our labeling approach identifies more progressed patients

Marc Claesen, Frank De Smet, Pieter Gillard, Chantal Mathieu & Bart De Moor (2015). *Building Classifiers to Predict the Start of Glucose-Lowering Pharmacotherapy Using Belgian Health Expenditure Data*.
Under review at Journal of Machine Learning Research (revision).
Preprint available at `http://arxiv.org/abs/1504.07389`.

The data set

We used records from the period 2008 up to 2012:

- of patients without prior use of GLAs before 2012
- patients of 40 years or older in 2012
- frequency counts per provision and drug

Labeling based on 2012 up to and including 2014:

- 31,066 known positives, 79,243 unlabeled (random)
- known positives have minimum 30 days of GLA use

Drugs & provisions

Drug purchases:

- encoded per package, with info about substances and doses
- each record can be mapped onto ATC system w/ DDDs

Provisions:

- each provision has a unique code
- each medical consultation $\rightarrow$ list of nomenclature codes

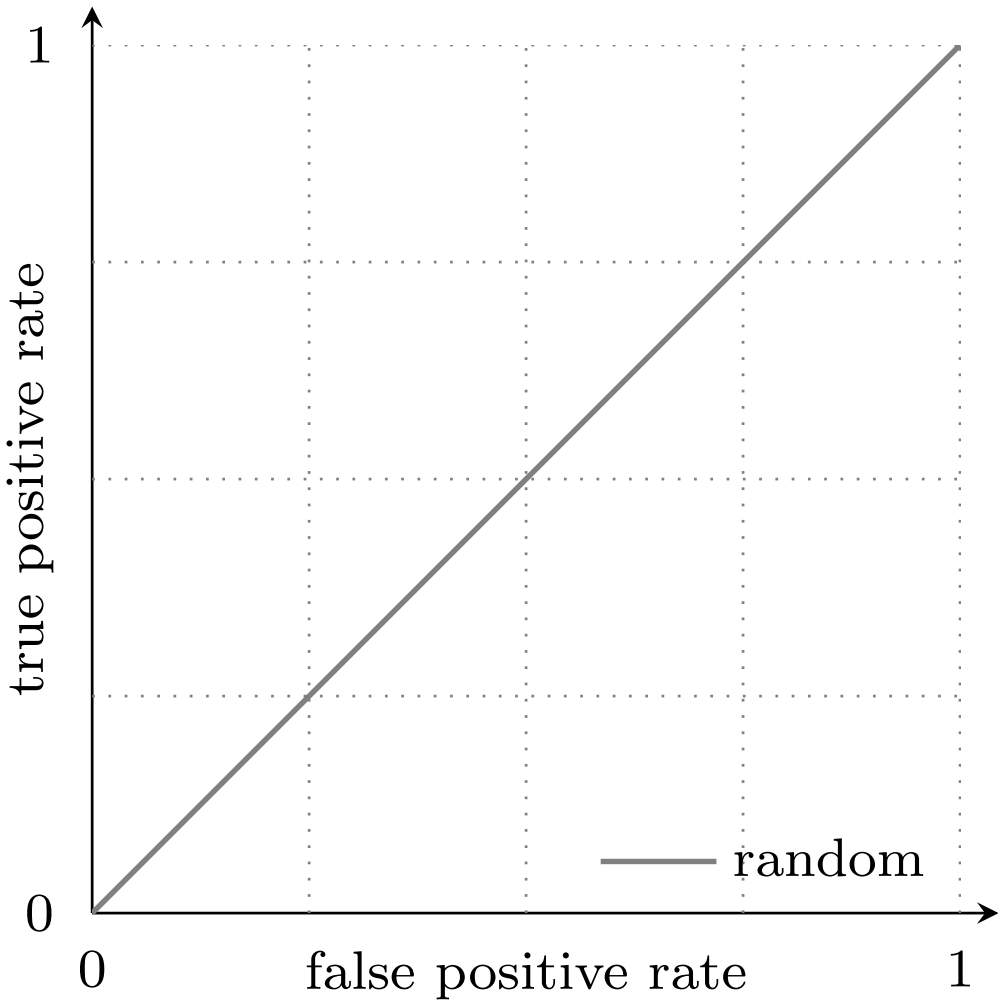
Example: ATC codes related to metformin

The Anatomical Therapeutic Chemical classification system:

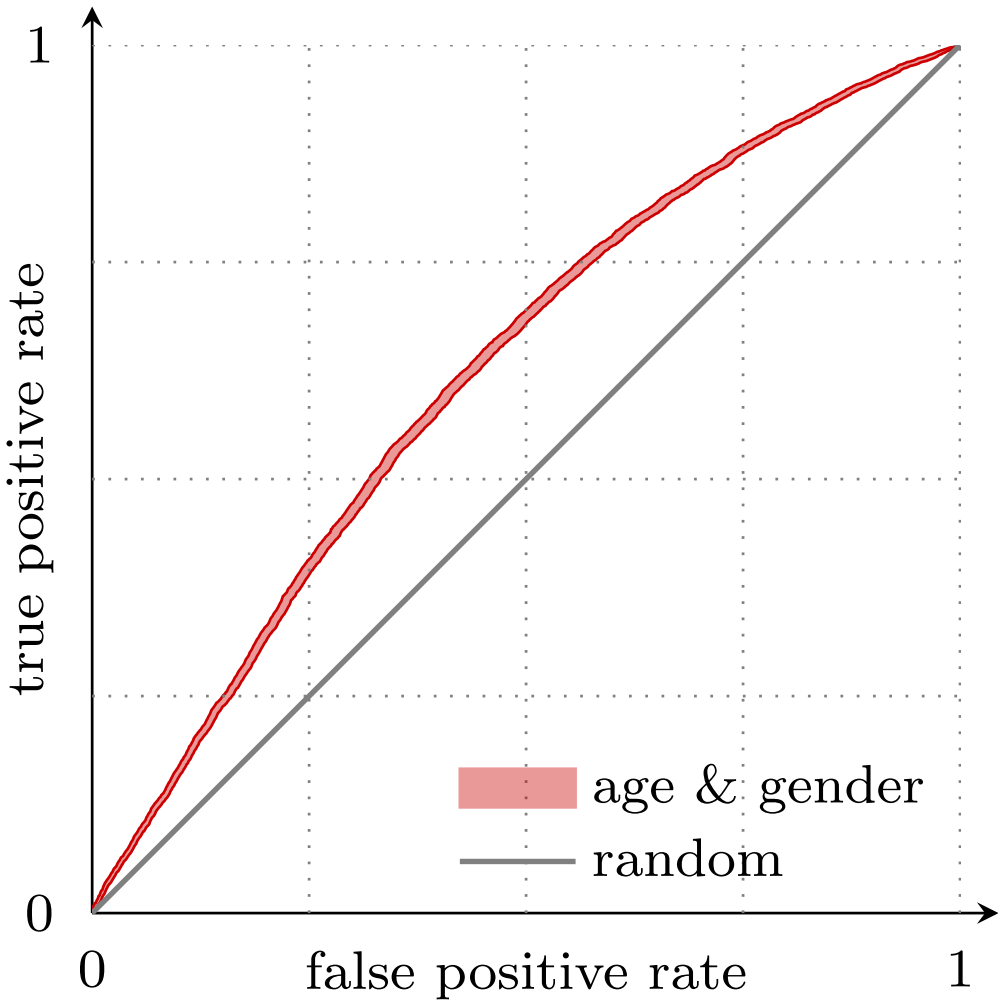
- tree structure which classifies medication into 5 levels
- 14 main groups (1st level)
- $\approx$ 1400 active substances (5th level)

level	ATC code	description
1	A	alimentary tract and metabolism
2	A10	drugs used in diabetes
3	A10B	blood glucose lowering drugs, excluding insulins
4	A10BA	biguanides
5	A10BA02	metformin

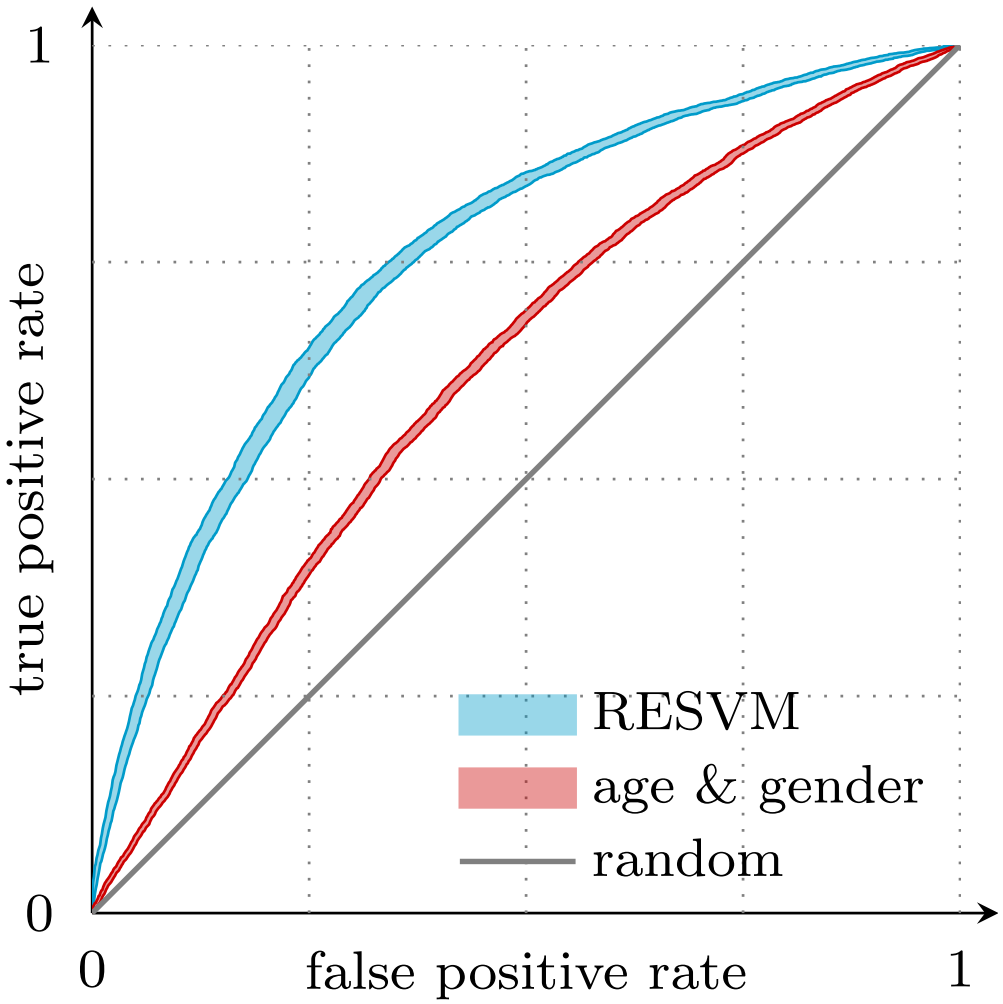
Receiver Operating Characteristic curves



Receiver Operating Characteristic curves



Receiver Operating Characteristic curves



Our approach beats Belgian guidelines under all configurations.

Performance in AUROC (%)

Our approaches based only on health expenditure data:

age & gender	61% – 62%
+ medication	74% – 76%
+ provisions	75% – 77%

International state-of-the-art based on surveys/primary care:

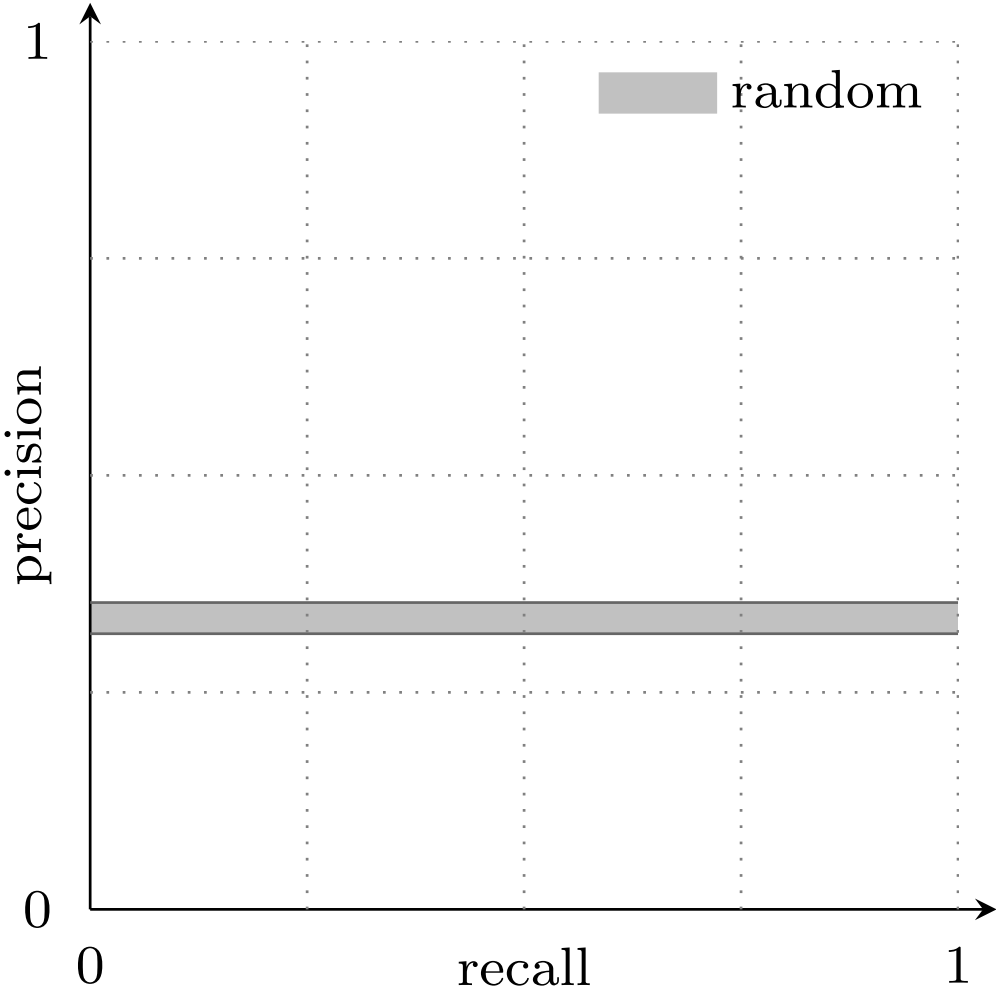
Cambridge Risk Score	67% – 80%
Danish diabetes risk score	76% – 80%
Diabetes Risk Calculator	75% – 85%

International state-of-the-art with clinical data:

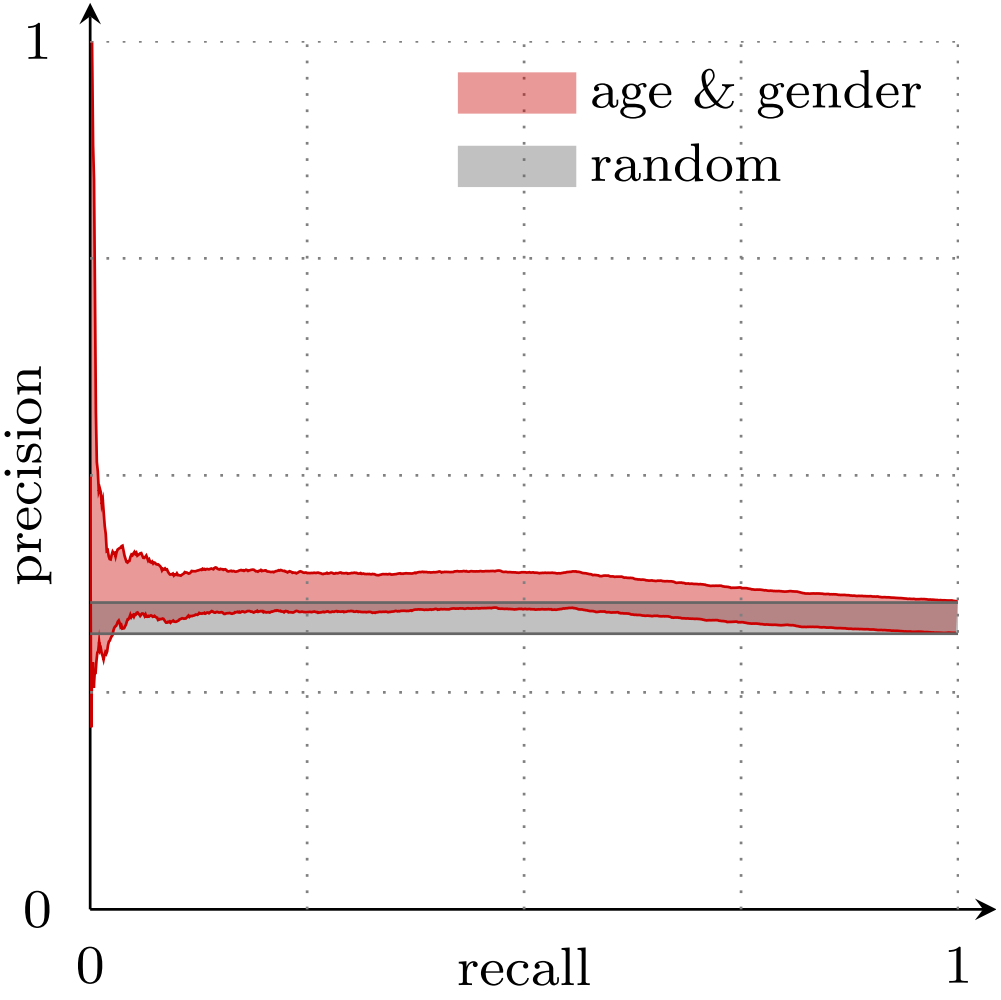
FINDRISC	85% – 87%
German diabetes risk score	75% – 83%

Key risk factors: BMI, waist circumference, family history, diet, ...

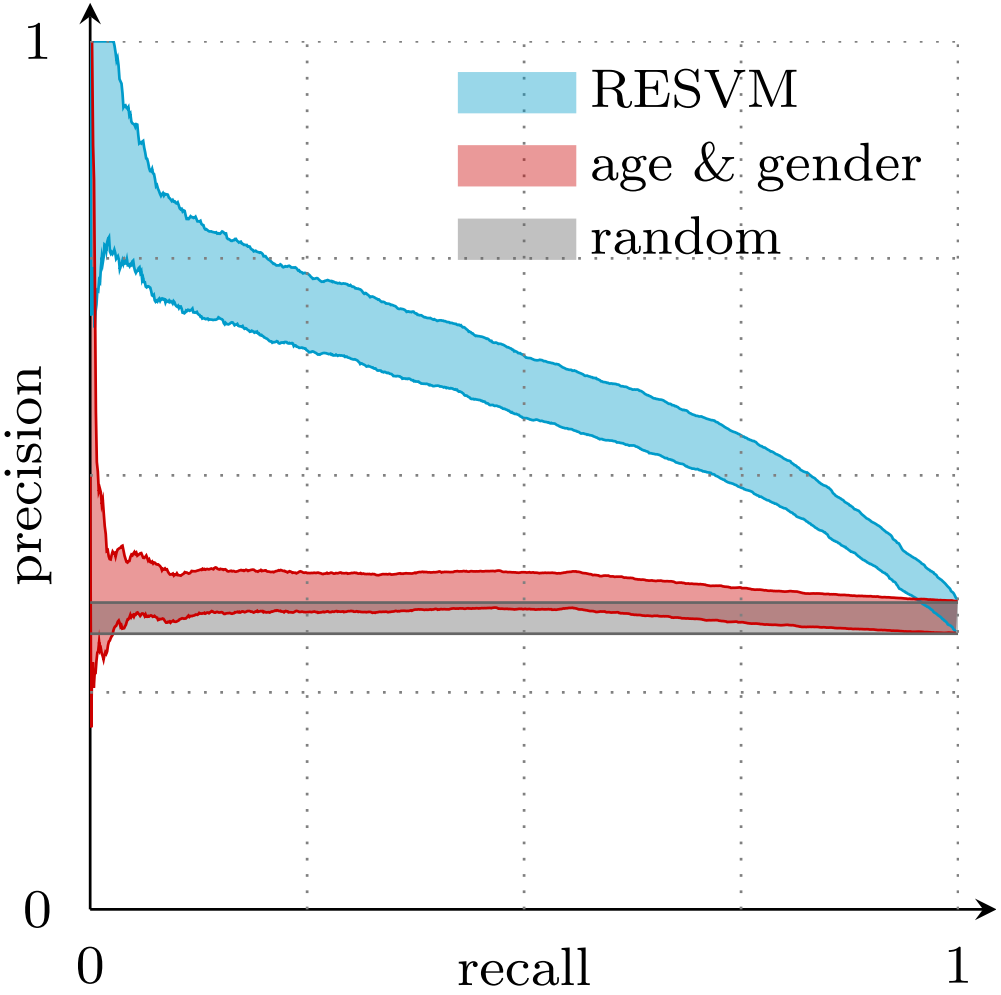
Precision-recall curves



Precision-recall curves



Precision-recall curves



Our approach has suitable characteristics for case-finding.

Biggest risk factors in level 3 ATC codes

feature	normalized weight	description
C09B		ACE inhibitors (combo)
C09D		angiotensin II antagonists
C10A		statins
N05A		antipsychotics
C03E		diuretics & potassium-sparing
C03C		high-ceiling diuretics
N06A		antidepressants
M04A		antigout preparations
A02B		peptic ulcer & GORD
C09A		ACE inhibitors (plain)

Outline

Contributions

Machine learning:

- learning method for positive and unlabeled data
- evaluating classifiers without known negatives

Open-source software:

- ensemble learning with SVM base models
- automated hyperparameter optimization

Medical:

- survival analysis of patients taking GLAs in Belgium
- approach to identify patients that require GLA therapy

Main result & future research

We developed a good case-finding approach for T2D.

- competitive with state-of-the-art screening approaches which typically use surveys or GP data (expensive)
- without direct info about important known risk factors BMI, lifestyle, diet, genetic predisposition, ...
- which can predict start of medication years in advance

Main result & future research

We developed a good case-finding approach for T2D.

- competitive with state-of-the-art screening approaches which typically use surveys or GP data (expensive)
- without direct info about important known risk factors BMI, lifestyle, diet, genetic predisposition, ...
- which can predict start of medication years in advance

Operational cost of our approach = hosting a simple website

- because the data is already digitally available

Main result & future research

We developed a good case-finding approach for T2D.

- competitive with state-of-the-art screening approaches which typically use surveys or GP data (expensive)
- without direct info about important known risk factors BMI, lifestyle, diet, genetic predisposition, ...
- which can predict start of medication years in advance

Operational cost of our approach = hosting a simple website

- because the data is already digitally available

Future research

- enrich the data with known risk factors
- clinical validation via other data sources

Potential healthcare savings

Type 2 diabetes is expensive

- 5.000 EUR/year excess cost per patient vs. non-diabetic
- mainly pharma + complications

Many new patients every year:

- 10.000 new known drug-treated patients/year in CM
- we can reliably identify 5 to 10% of these patients years earlier

Potential healthcare savings

Type 2 diabetes is expensive

- 5.000 EUR/year excess cost per patient vs. non-diabetic
- mainly pharma + complications

Many new patients every year:

- 10.000 new known drug-treated patients/year in CM
- we can reliably identify 5 to 10% of these patients years earlier

early detection of 500 patients/year

Potential healthcare savings

Type 2 diabetes is expensive

- 5.000 EUR/year excess cost per patient vs. non-diabetic
- mainly pharma + complications

Many new patients every year:

- 10.000 new known drug-treated patients/year in CM
- we can reliably identify 5 to 10% of these patients years earlier

early detection of	500	patients/year
save	$\times$ 10.000	EUR/patient (in lifetime)

Potential healthcare savings

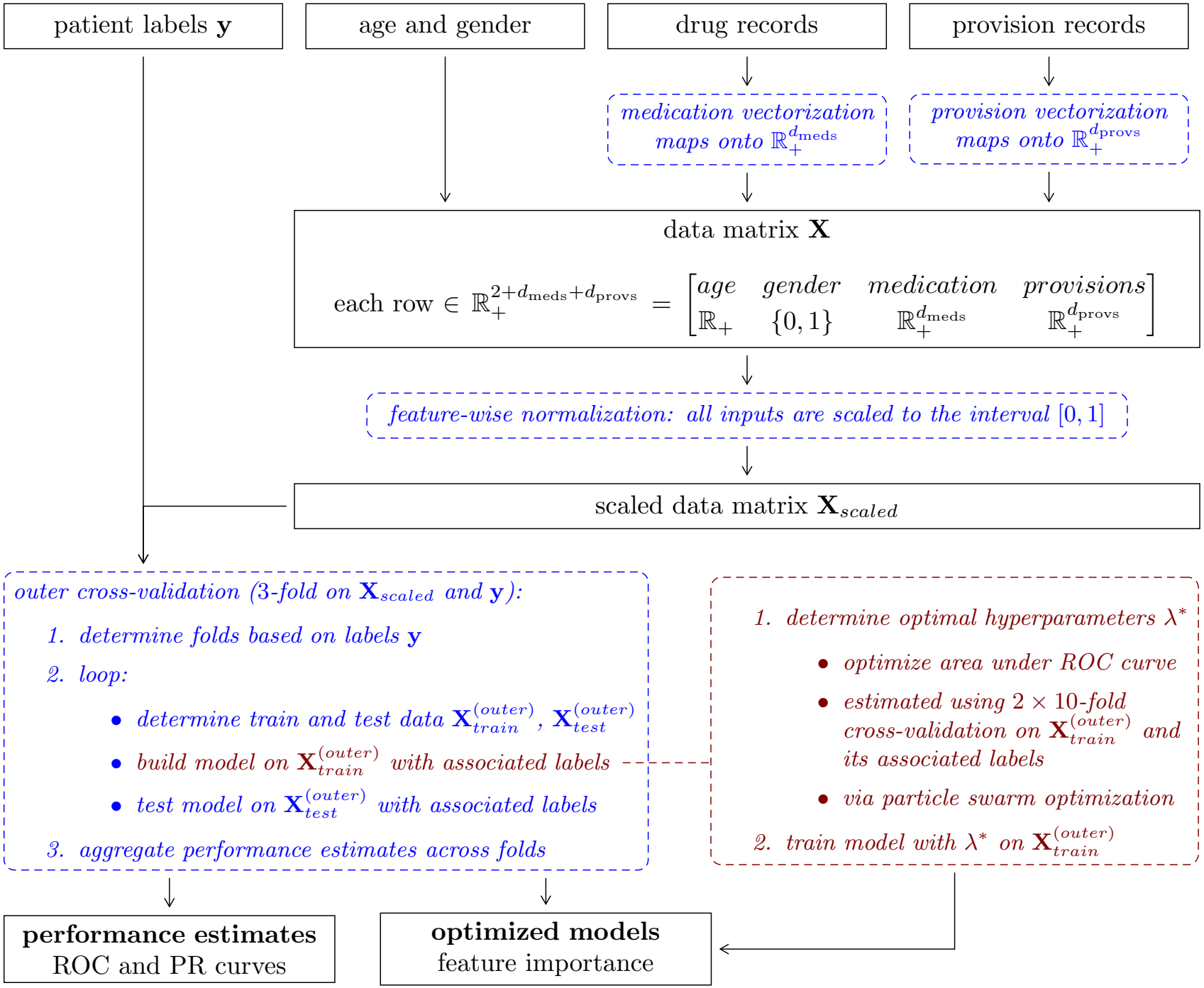
Type 2 diabetes is expensive

- 5.000 EUR/year excess cost per patient vs. non-diabetic
- mainly pharma + complications

Many new patients every year:

- 10.000 new known drug-treated patients/year in CM
- we can reliably identify 5 to 10% of these patients years earlier

early detection of	500	patients/year
save	$\times$ 10.000	EUR/patient (in lifetime)
healthcare cost reduction	5.000.000	EUR/year


patient labels $\mathbf{y}$
age and gender
drug records
provision records
medication vectorization maps onto $\mathbb{R}_+^{d_{\text{meds}}}$
provision vectorization maps onto $\mathbb{R}_+^{d_{\text{provs}}}$
data matrix $\mathbf{X}$
each row $\in \mathbb{R}_+^{2+d_{\text{meds}}+d_{\text{provs}}} = \begin{bmatrix} \textit{age} & \textit{gender} & \textit{medication} & \textit{provisions} \\ \mathbb{R}_+ & \{0,1\} & \mathbb{R}_+^{d_{\text{meds}}} & \mathbb{R}_+^{d_{\text{provs}}} \end{bmatrix}$
feature-wise normalization: all inputs are scaled to the interval $[0,1]$
scaled data matrix $\mathbf{X}_{scaled}$
outer cross-validation (3-fold on $\mathbf{X}_{scaled}$ *and* $\mathbf{y}$*):*
1. *determine folds based on labels* $\mathbf{y}$
2. *loop:*
   - *determine train and test data* $\mathbf{X}_{train}^{(outer)}$, $\mathbf{X}_{test}^{(outer)}$
   - *build model on* $\mathbf{X}_{train}^{(outer)}$ *with associated labels*
   - *test model on* $\mathbf{X}_{test}^{(outer)}$ *with associated labels*
3. *aggregate performance estimates across folds*
1. *determine optimal hyperparameters* λ^*
   - *optimize area under ROC curve*
   - *estimated using* 2×10*-fold cross-validation on* $\mathbf{X}_{train}^{(outer)}$ *and its associated labels*
   - *via particle swarm optimization*
2. *train model with* λ^* *on* $\mathbf{X}_{train}^{(outer)}$
performance estimates
ROC and PR curves
optimized models
feature importance

Key papers

Published:

- EnsembleSVM (*Journal of Machine Learning Research*)
- PU learning method (*Neurocomputing*)

Under review:

- Optunity (*Journal of Machine Learning Research*)
- Survival analysis (*J. of Clinical Endocrinology & Metabolism*)
- Diabetes screening (*Journal of Machine Learning Research*)

To be submitted/in preparation:

- Evaluating models without negatives (*ACM SIGKDD 2016*)
- Diabetes screening, medical interpretation (*tier 1 medical*)