L1 regularization & Intro to learning theory
Lecture 8

David Sontag
New York University

Feature Selection

Setting: Lots of possible features, many of which are irrelevant

Example:

When studying depression in teens, a researcher distributes a questionnaire of 250 different questions, many of them related or irrelevant.

Goal: Find a *small set* of questions that can be used to quickly determine whether or not a teen is depressed.

Feature Selection

Setting: Lots of possible features, many of which are irrelevant

Example:

When studying depression in teens, a researcher distributes a questionnaire of 250 different questions, many of them related or irrelevant.

Goal: Find a *small set* of questions that can be used to quickly determine whether or not a teen is depressed.

Mathematically:

$$\min_w \ell(w \cdot x, y) + \lambda(\text{non-zero elements in } w)$$

Feature Selection

Setting: Lots of possible features, many of which are irrelevant

Example:

When studying depression in teens, a researcher distributes a questionnaire of 250 different questions, many of them related or irrelevant.

Goal: Find a *small set* of questions that can be used to quickly determine whether or not a teen is depressed.

Mathematically:

$$\min_w \ell(w \cdot x, y) + \lambda(\underline{\text{non-zero elements in } w})$$

L1 regularization

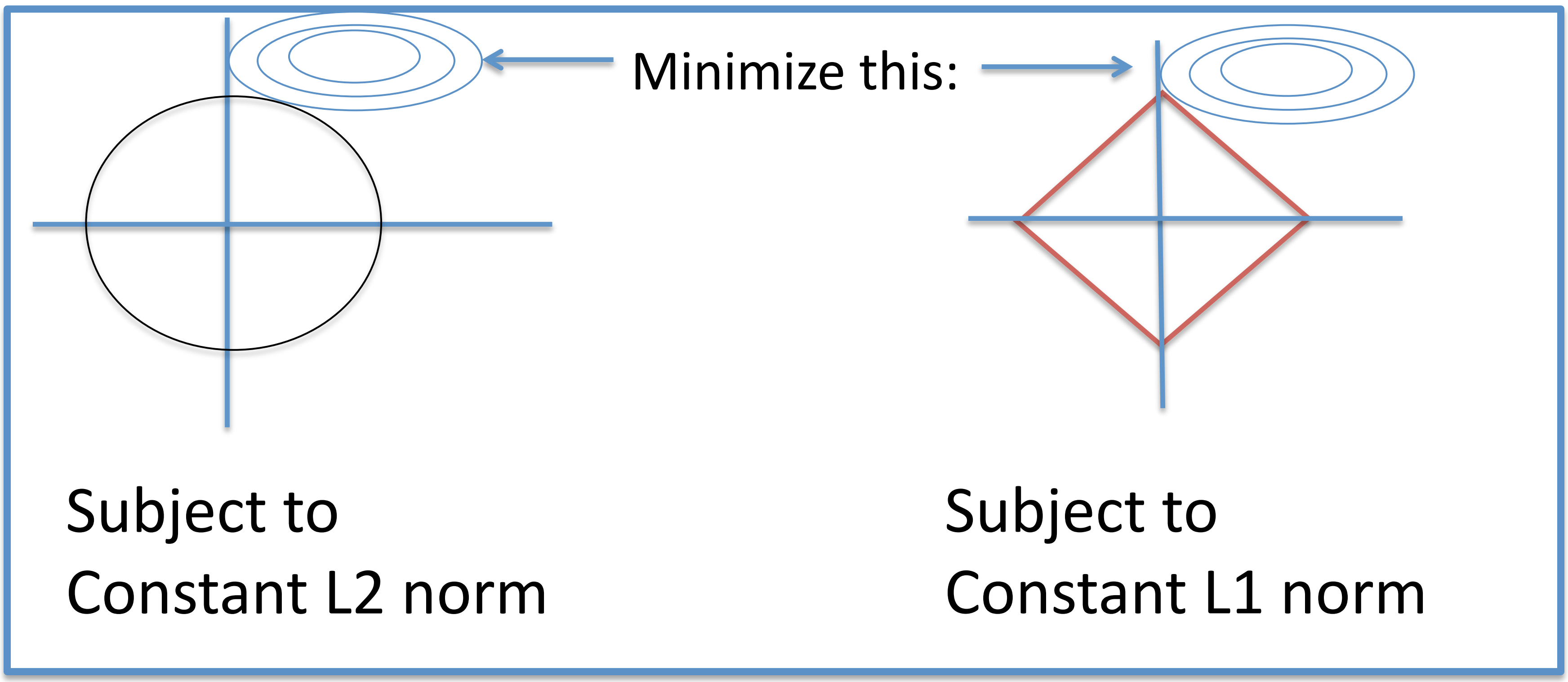
- Penalizing the L1 norm of the weight vector leads to *sparse* (read: many 0's) solutions for *w.*

$$\min_w \ell(w \cdot x, y) + \lambda |w|$$

- Why?

L1 regularization

- Penalizing the L1 norm of the weight vector leads to *sparse* (read: many 0's) solutions for *w*.

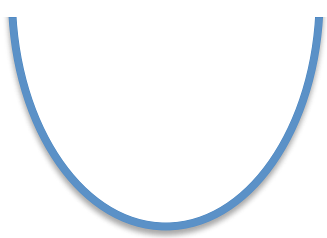


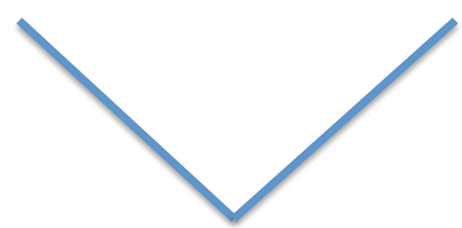
L1 regularization

- Penalizing the L1 norm of the weight vector leads to *sparse* (read: many 0's) solutions for *w.*

Intuition #2 – w.w.g.d.d
(What would gradient descent do?)

$$\frac{d}{dw_i}\lambda||w||_2 = \pm\lambda w_i \qquad \frac{d}{dw_i}\lambda|w| = \pm\lambda$$



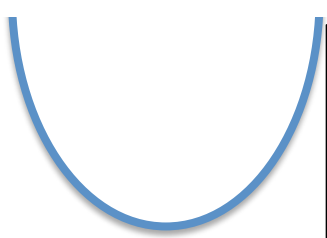


L1 regularization

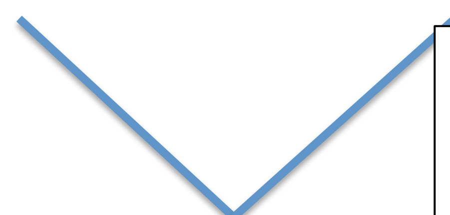
- Penalizing the L1 norm of the weight vector leads to *sparse* (read: many 0's) solutions for *w.*

Intuition #2 – w.w.g.d.d
(What would gradient descent do?)

$$\frac{d}{dw_i}\lambda||w||_2 = \pm\lambda w_i \qquad \frac{d}{dw_i}\lambda|w| = \pm\lambda$$



The push towards 0 gets weaker as wi gets smaller



Always pushes elements of wi towards 0

Example: Early Detection of Type 2 Diabetes

- Global prevalence will go from 171 million in 2000 to 366 million in 2030
- 25% of people in the US with diabetes are undiagnosed
- Leads to complications of cardiovascular, cerebrovascular, renal, and vision systems
- Early lifestyle changes shown to prevent or delay the onset of the disease better than Metformin

Traditional risk assessment

- Use small number of risk factors (e.g. ~20)
- Easy to ask/measure in the office
- Simple model: can calculate scores by hand

Finnish Diabetes Association

TYPE 2 DIABETES RISK ASSESSMENT FORM

Circle the right alternative and add up your points.

1. Age

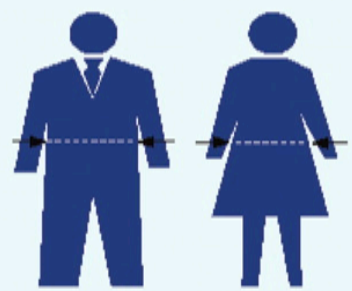
0 p. Under 45 years
2 p. 45–54 years
3 p. 55–64 years
4 p. Over 64 years

2. Body-mass index
(See reverse of form)

0 p. Lower than 25kg/m²
1 p. 25–30 kg/m²
3 p. Higher than 30 kg/m²

3. Waist circumference measured below the ribs (usually at the level of the navel)

	MEN	WOMEN
0 p.	Less than 94cm	Less than 80cm
3 p.	94–102cm	80–88cm
4 p.	More than 102cm	More than 88cm



4. Do you usually have daily at least 30 minutes of physical activity at work and/or during leisure time (including normal daily activity)?

0 p. Yes
2 p. No

5. How often do you eat vegetables, fruit'or berries?

0 p. Every day
1 p. Not every day

6. Have you ever taken anti-hypertensive medication regularly?

0 p. No
2 p. Yes

7. Have you ever been found to have high blood glucose (e.g. in a health examination, during an illness, during pregnancy)?

0 p. No
5 p. Yes

8. Have any of the members of your immediate family or other relatives been diagnosed with diabetes (type 1 or type 2)?

0 p. No
3 p. Yes: grandparent, aunt, uncle or first cousin (but no own parent, brother, sister or child)
5 p. Yes: parent, brother, sister or own child

Total risk score

The risk of developing type 2 diabetes within 10 years is

Lower than 7	**Low:** estimated 1 in 100 will develop disease
7–11	**Slightly elevated:** estimated 1 in 25 will develop disease
12–14	**Moderate:** estimated 1 in 6 will develop disease
15–20	**High:** estimated 1 in 3 will develop disease
Higher than 20	**Very high:** estimated 1 in 2 will develop disease

Please turn over

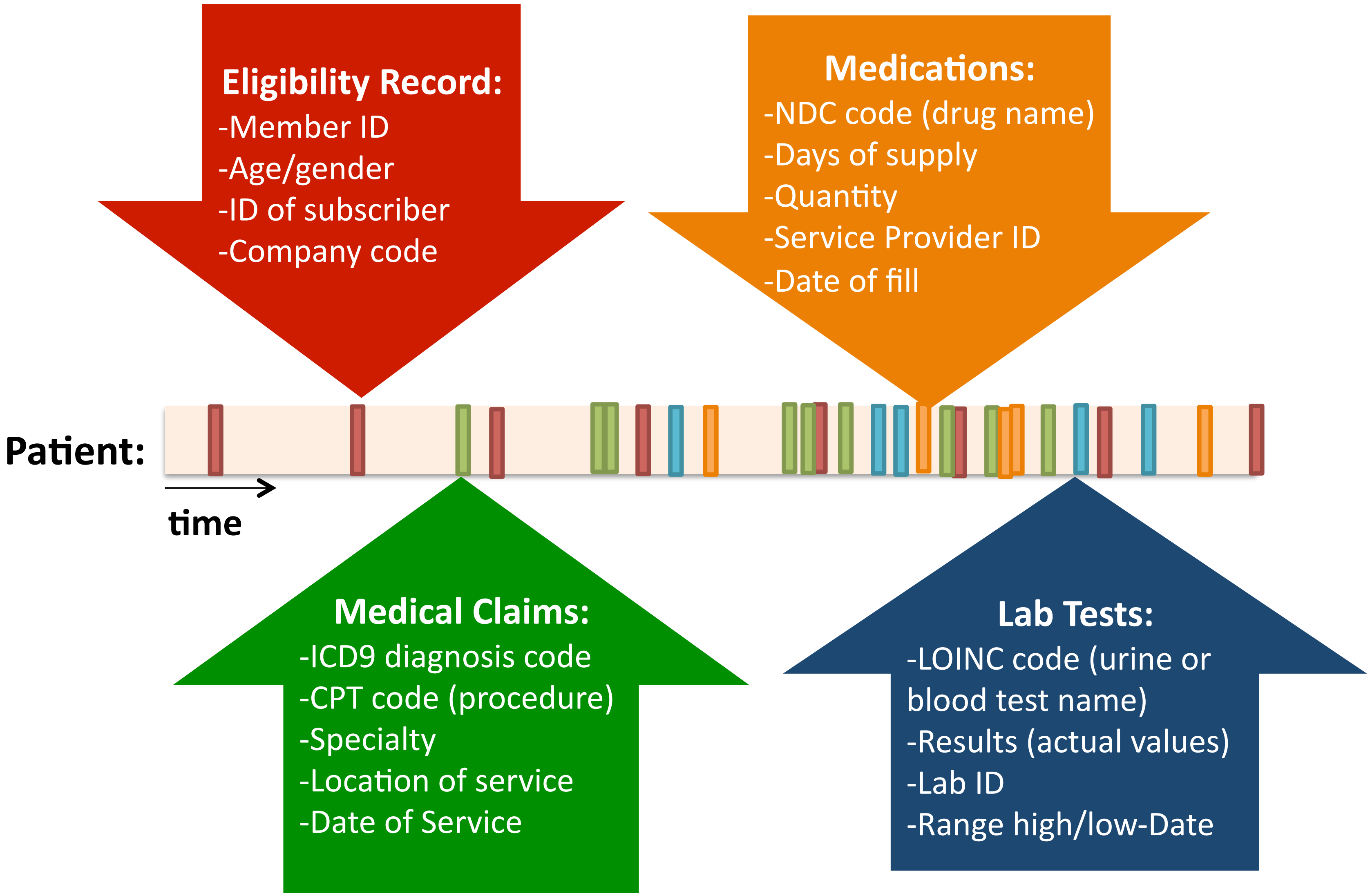
Test designed by Professor Jaakko Tuomilehto, Department of Public Health, University of Helsinki, and Jaana Lindström, MFS, National Public Health Institute.

Population-Level Risk Stratification

- Key idea: Use automatically collected administrative, utilization, and clinical data
- Machine learning will find surrogates for risk factors that would otherwise be missing
- Enables risk stratification at the population level – millions of patients

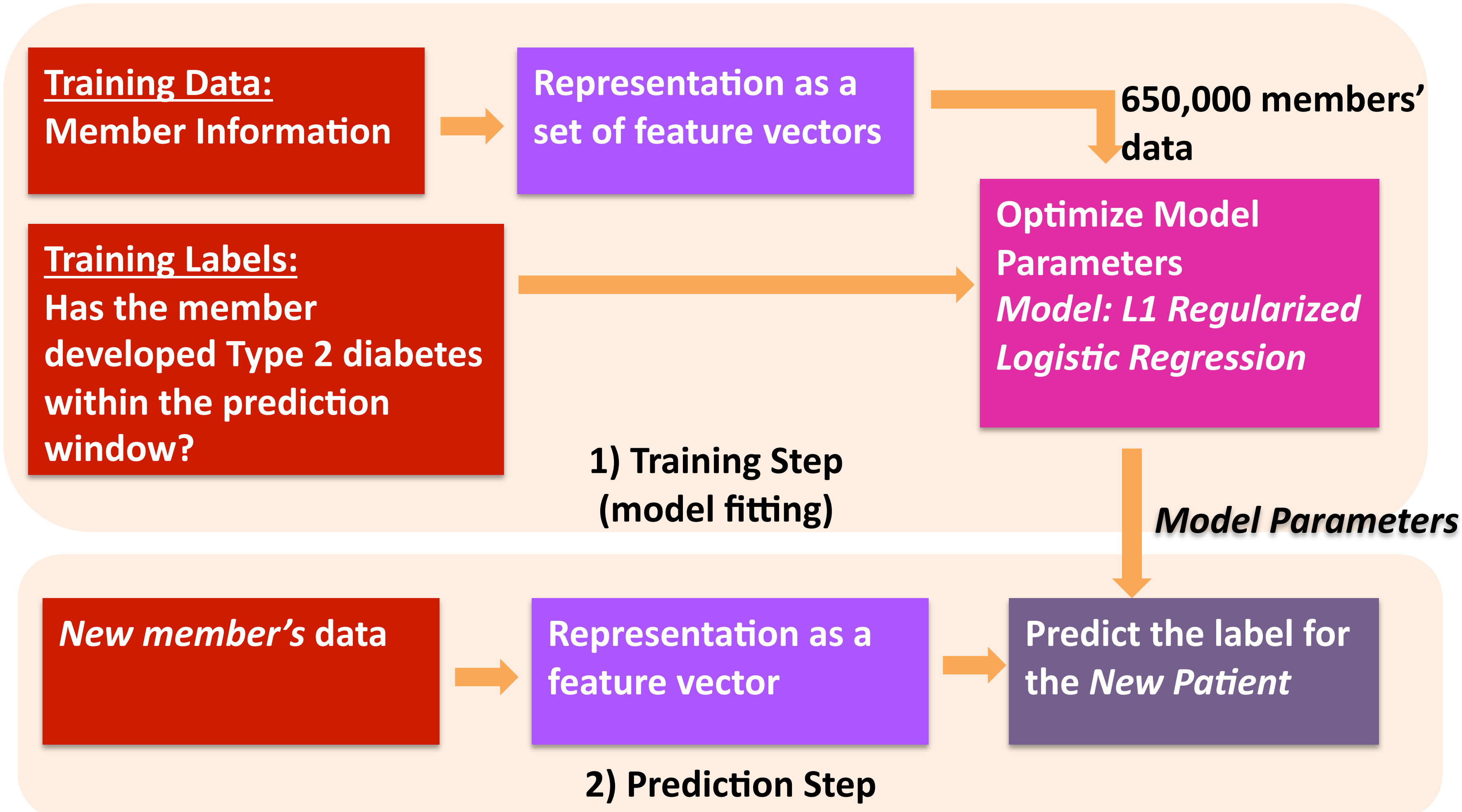
[N. Razavian, S. Blecker, A.M. Schmidt, A. Smith-McLallen, S. Nigam, D. Sontag. Population-Level Prediction of Type 2 Diabetes using Claims Data and Analysis of Risk Factors. ***Big Data***, Jan. 2016.]

Administrative & Clinical Data



Machine Learning

Task: predict the probability of a member developing diabetes

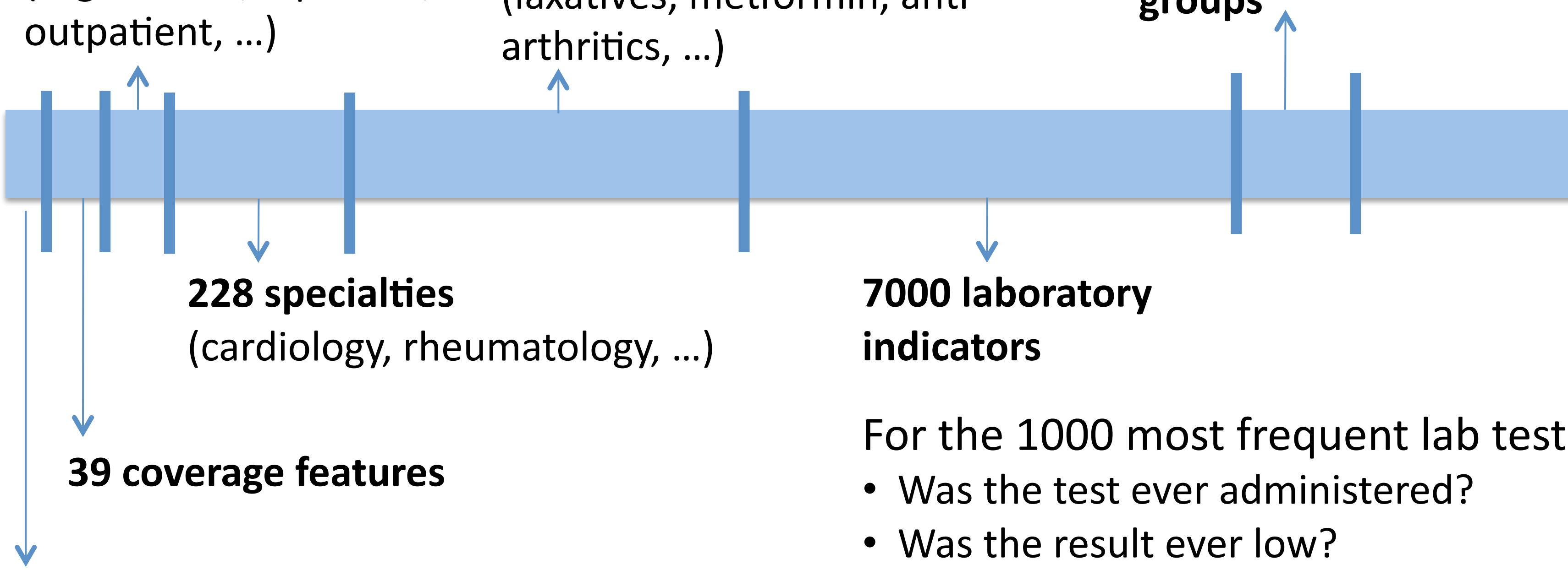


Features

32 service places
(urgent care, inpatient, outpatient, …)

999 medication groups
(laxatives, metformin, anti-arthritics, …)

457 procedure groups



228 specialties
(cardiology, rheumatology, …)

7000 laboratory indicators

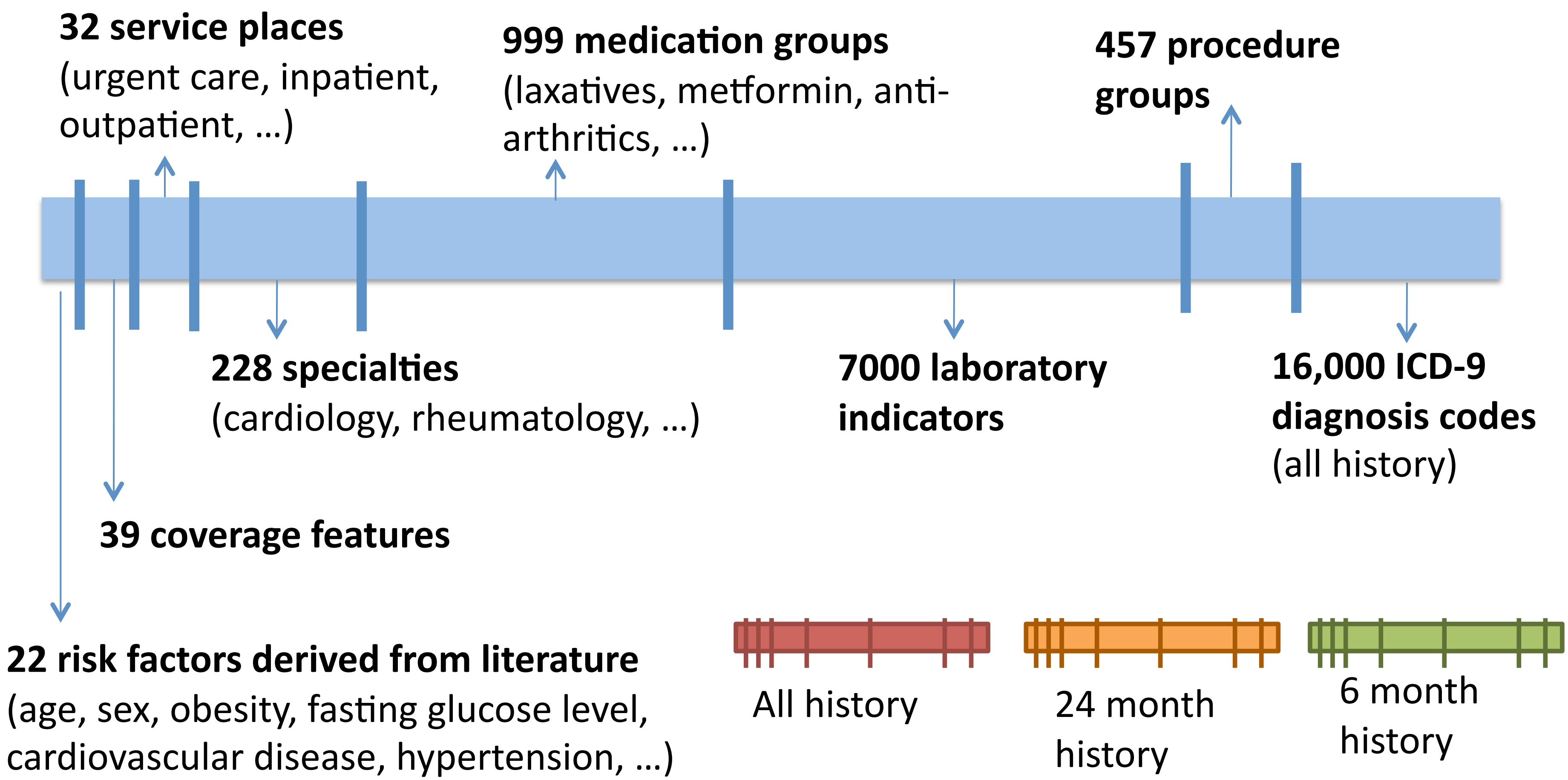
39 coverage features

22 risk factors derived from literature
(age, sex, obesity, fasting glucose level, cardiovascular disease, hypertension, …)

For the 1000 most frequent lab tests:

- Was the test ever administered?
- Was the result ever low?
- Was the result ever high?
- Was the result ever normal?
- Is the value increasing?
- Is the value decreasing?
- Is the value fluctuating?

Features



Total features per patient: 42,000

What are the discovered risk factors?

Feature Name

Impaired Fasting Glucose (790.21)
Abnormal Glucose NEC (790.29)
Hypertension (401)
Obstructive Sleep Apnea (327.23)
Obesity (278)
Abnormal Blood Chemistry (790.6)
Hyperlipidemia (272.4)
Shortness Of Breath (786.05)
Esophageal Reflux (530.81)
Acute Bronchitis (466.0)
Actinic Keratosis (702.0)

Positive weights

Additional risk factors identfied:
Impaired oral glucose tolerance, Chronic liver disease, Pituitary dwarfism, Hypersomnia with sleep apnea, Joint replaced knee, Liver disorder, Iron deficiency anemia, Mitral valve disorder…

Diagnostic groups | Procedure Group | Lab Test | Medication Group | Service Place

What are the discovered risk factors?

Feature Name

Positive weights

Hemoglobin A1c / Hemoglobin.Total - High

Glucose - High

Hemoglobin A1c / Hemoglobin.Total - Request For Test

Cholesterol.In HDL - Low

Cholesterol.Total / Cholesterol.In HDL - Hi

Cholesterol.In VLDL - Request For Test

Carbon Dioxide - Request For Test

Glomerular Filtration Rate/1.73 Sq. M. P
Black - Request For Test

Additional risk factors identfied:
Potassium (low), Erythrocyte mean corpuscular hemoglobin concentration (fluctuating), Erythrocyte distribution width (high), Alanine aminotransferase (high), Cholesterol.in LDL (increasing), Creatinine (decreasing), Albumin/Globulin (increasing)…

Diagnostic groups | Procedure Group | Lab Test | Medication Group | Service Place

What are the discovered risk factors?

Feature Name	
Routine Chest Xray	Very positive
Medication Group: Anti-arthritics	Very positive
Service Place: Emergency Room - Hospital	Very positive
Routine Medical Exam (V700)	Very negative
Routine Gynecological Examination (V7231)	Very negative
Routine Child Health Exam (V202)	Very negative

~700 risk factors selected for model

Diagnostic groups | Procedure Group | Lab Test | Medication Group | Service Place

Type 2 Diabetes Prediction Accuracy

Using patient data through Dec. 31, 2008, who will be newly diagnosed with Type 2 diabetes in the following years?

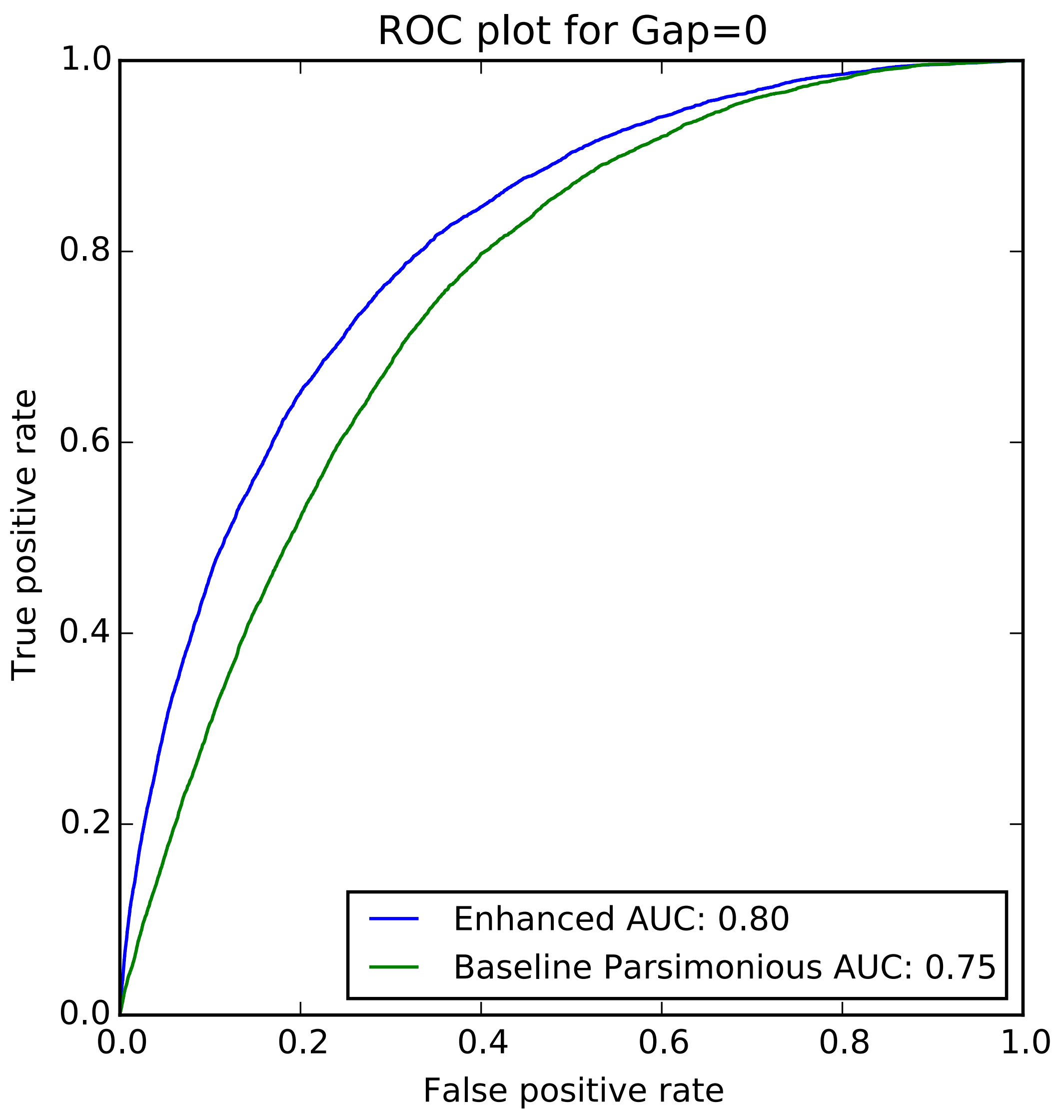
	Model	AUC
2009-2011 (incident diabetes)	Literature features only	0.75
	Overall Model	**0.8**
2011-2013 (future diabetics)	Literature features only	0.72
	Overall Model	**0.76**

Area under the ROC curve (AUC) = Randomly choosing two members, one who *did* get diabetes and one who *did not*, can we predict which is which?

← Highest risk population

← 2 years lead time for this population

Type 2 Diabetes Prediction Accuracy



Type 2 Diabetes Prediction Accuracy

Using patient data through Dec. 31, 2008, who will be newly diagnosed with Type 2 diabetes in the following years?

	Model	AUC	Top 1000 predictions		
			Sensitivity	Specificity	PPV
2009-2011 (incident diabetes)	Literature features only	0.75	0.014	0.996	0.1
	Overall Model	**0.8**	0.033	0.997	**0.24**
2011-2013 (future diabetics)	Literature features only	0.72	0.013	0.995	0.04
	Overall Model	**0.76**	0.023	0.995	**0.07**

Sensitivity = TP/P
"true positive rate" or "recall"

Specificity = TN/N
"true negative rate"

PPV = TP/(TP+FP)
"positive predictive value"

Type 2 Diabetes Prediction Accuracy

Using patient data through Dec. 31, 2008, who will be newly diagnosed with Type 2 diabetes in the following years?

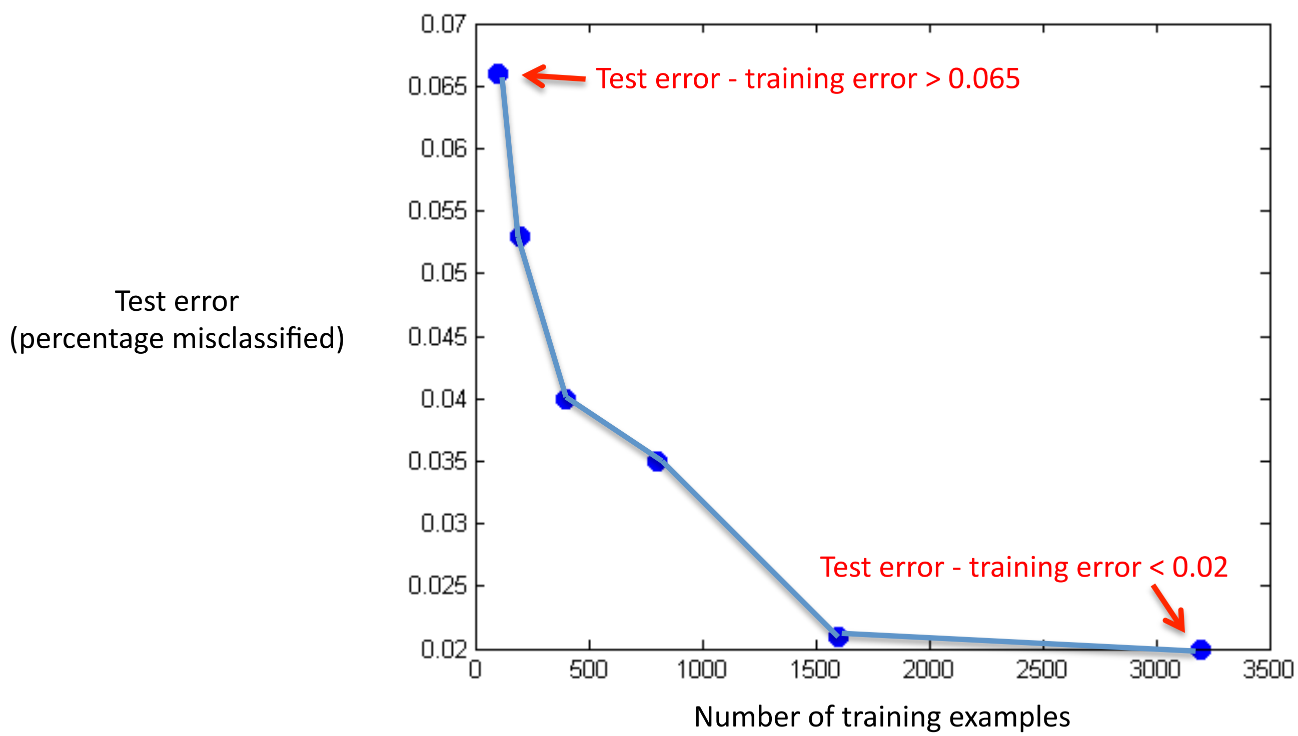
	Model	**AUC**	**Top 1000 predictions**			**Top 10000 predictions**		
			Sensitivity	Specificity	PPV	Sensitivity	Specificity	PPV
2009-2011 (incident diabetes)	Literature features only	0.75	0.014	0.996	0.1	0.114	0.967	0.08
	Overall Model	**0.8**	0.033	0.997	**0.24**	**0.212**	**0.969**	**0.14**
2011-2013 (future diabetics)	Literature features only	0.72	0.013	0.995	0.04	0.116	0.957	0.03
	Overall Model	**0.76**	0.023	0.995	**0.07**	**0.179**	**0.958**	**0.05**

What's next…

- We gave several machine learning algorithms:
  - Perceptron
  - Linear support vector machine (SVM)
  - SVM with kernels, e.g. polynomial or Gaussian
- How do we guarantee that the learned classifier will perform well on test data?
- How much training data do we need?

Example: Perceptron applied to spam classification

- In your homework 1, you trained a spam classifier using perceptron
  - **The training error was always zero**
  - With few data points, there is a big gap between training error and test error!



How much training data do you need?

- Depends on what *hypothesis class* the learning algorithm considers

- For example, consider a memorization-based learning algorithm
  - Input: training data S = { ($\mathbf{x}_i$, y_i) }
  - Output: function f($\mathbf{x}$) which, if there exists ($\mathbf{x}_i$, y_i) in S such that $\mathbf{x}=\mathbf{x}_i$, predicts y_i, and otherwise predicts the majority label
  - This learning algorithm will always obtain zero training error
  - But, it will take a ***huge*** amount of training data to obtain small test error (i.e., its generalization performance is horrible)

- Linear classifiers are powerful precisely because of their simplicity
  - Generalization is easy to guarantee

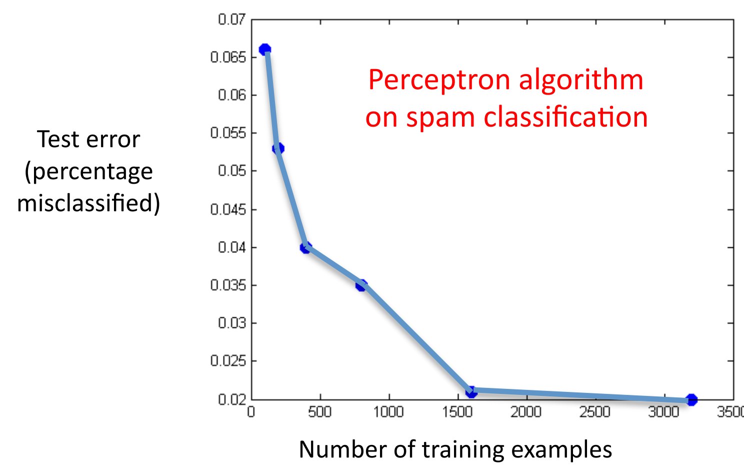
Roadmap of next lectures

1. Generalization of finite hypothesis spaces

2. VC-dimension

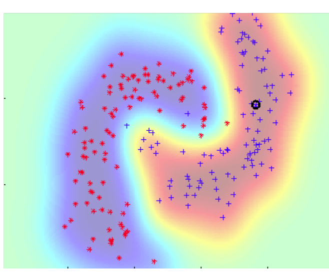
   - Will show that **linear** classifiers need to see approximately **d** training points, where **d** is the dimension of the feature vectors

   - Explains the good performance we obtained using perceptron!!!! (we had a few thousand features)



3. Margin based generalization

   - Applies to **infinite** dimensional feature vectors (e.g., Gaussian kernel)



[Figure from Cynthia Rudin]