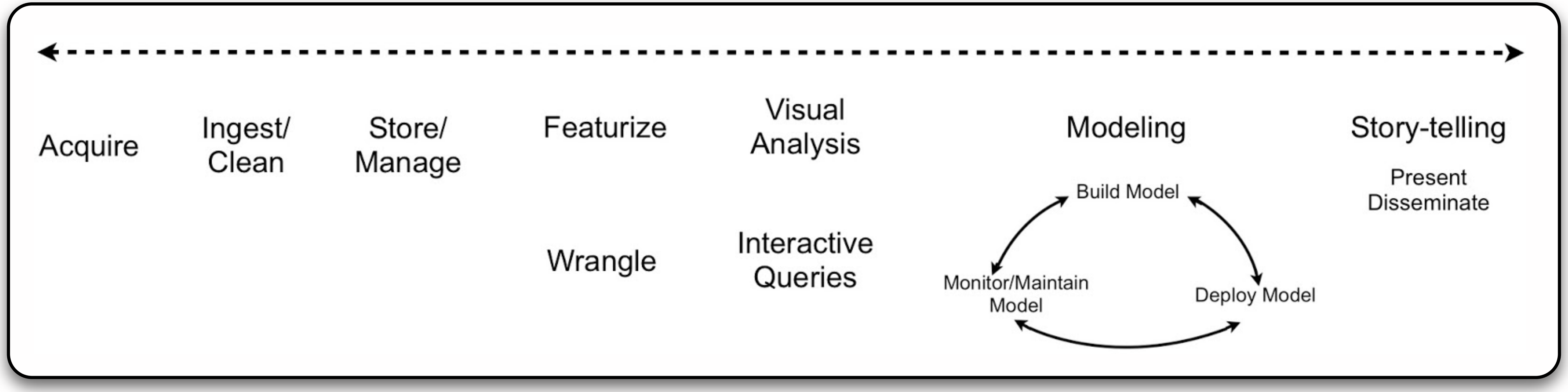
Overcoming the Barriers to Production-Ready Machine Learning Workflows

Josh Bloom Henrik Brink

@profjsb @brinkar @wiseio

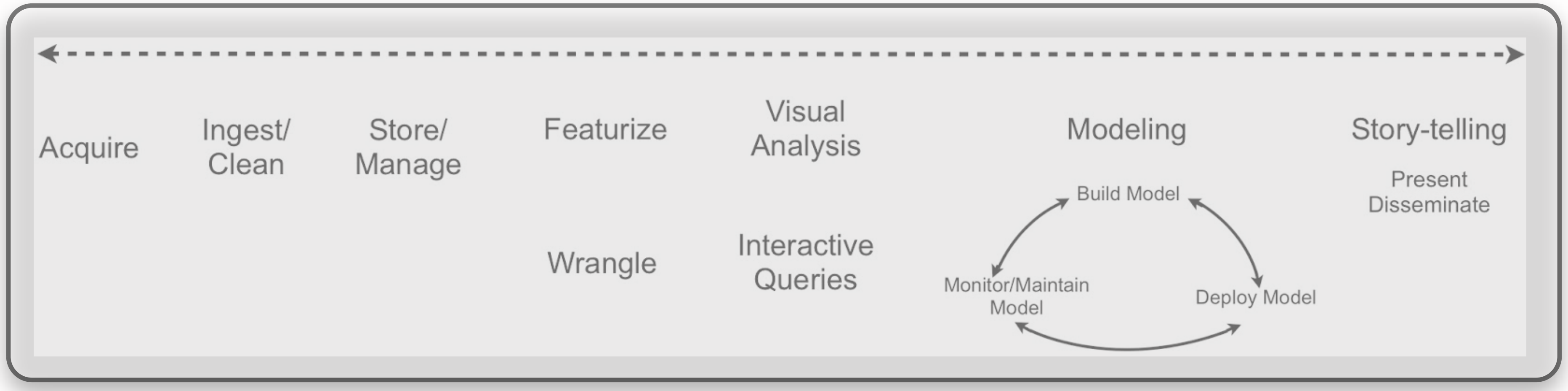
wise.io

University of California, Berkeley



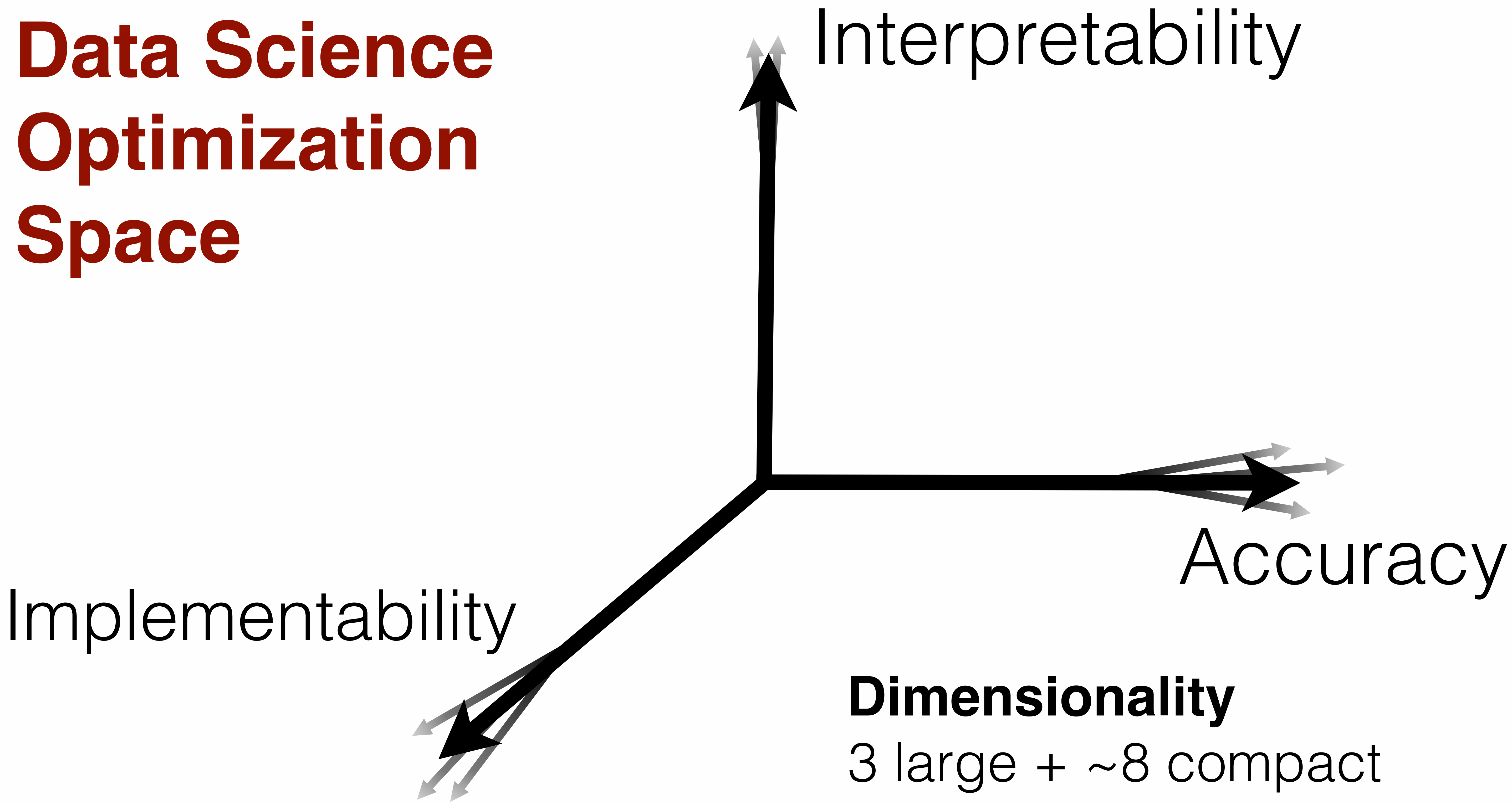
Lorica's "Data Science Workflow"

Real-World Data Science = *Optimization* over this *full* Workflow



Lorica's "Data Science Workflow"

Data Science Optimization Space



Dimensionality

3 large + ~8 compact

Our Background ...
"Data-Driven Scientists"

- Built & Deployed Real-time ML framework, discovering >10,000 events in > 10 TB of imaging
  → 50+ journal articles
- Built Probabilistic Event classification catalogs with innovative active learning
- Collective over 350 refereed journal articles including ML & timeseries analysis

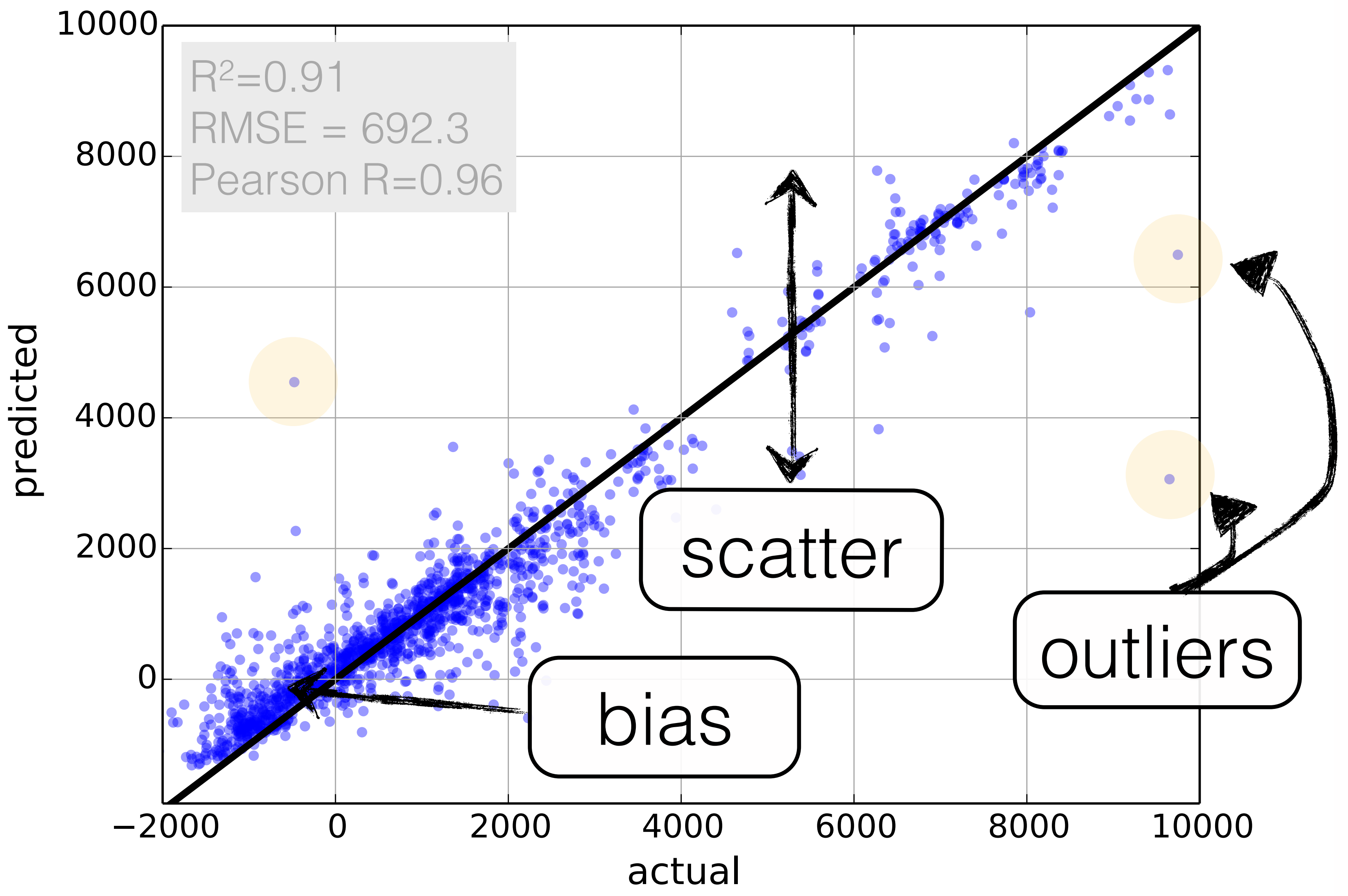
Accuracy

Evaluation Metric: What's the essence of what I care about?

Scalar proxies

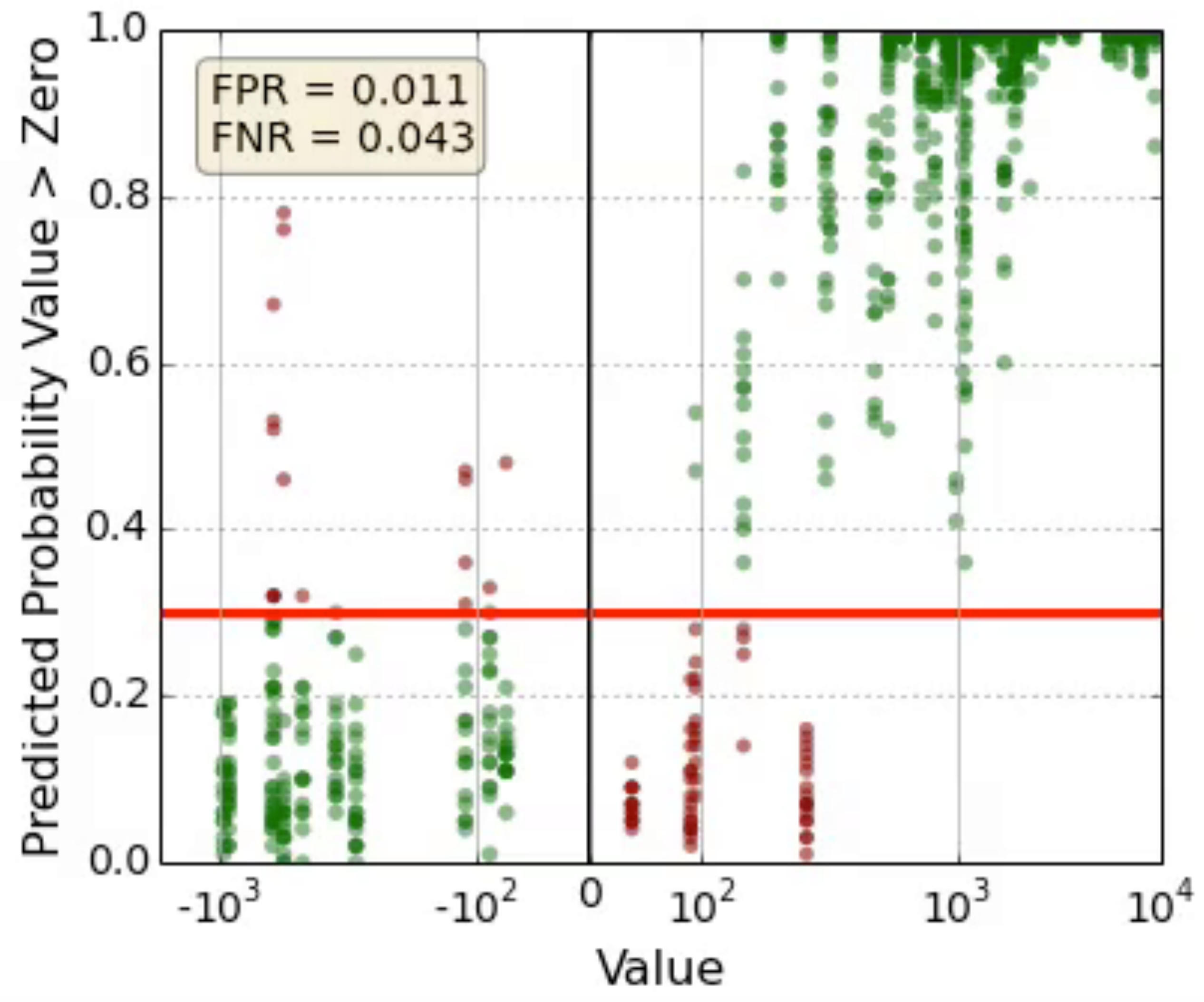
- RMSE
- RMSLE
- [adjusted] R^2
- ...

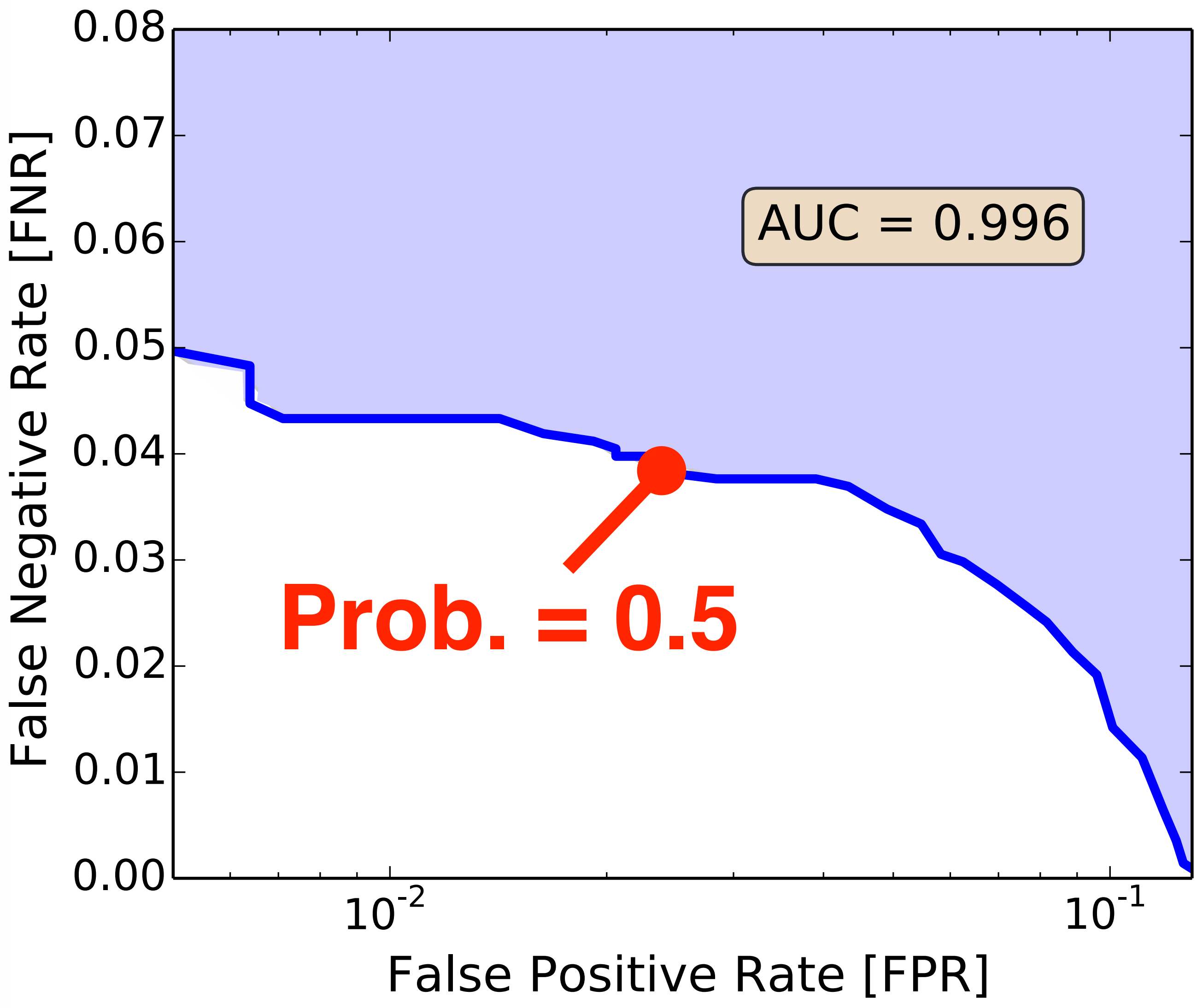
cf. sklearn.metrics



Accuracy

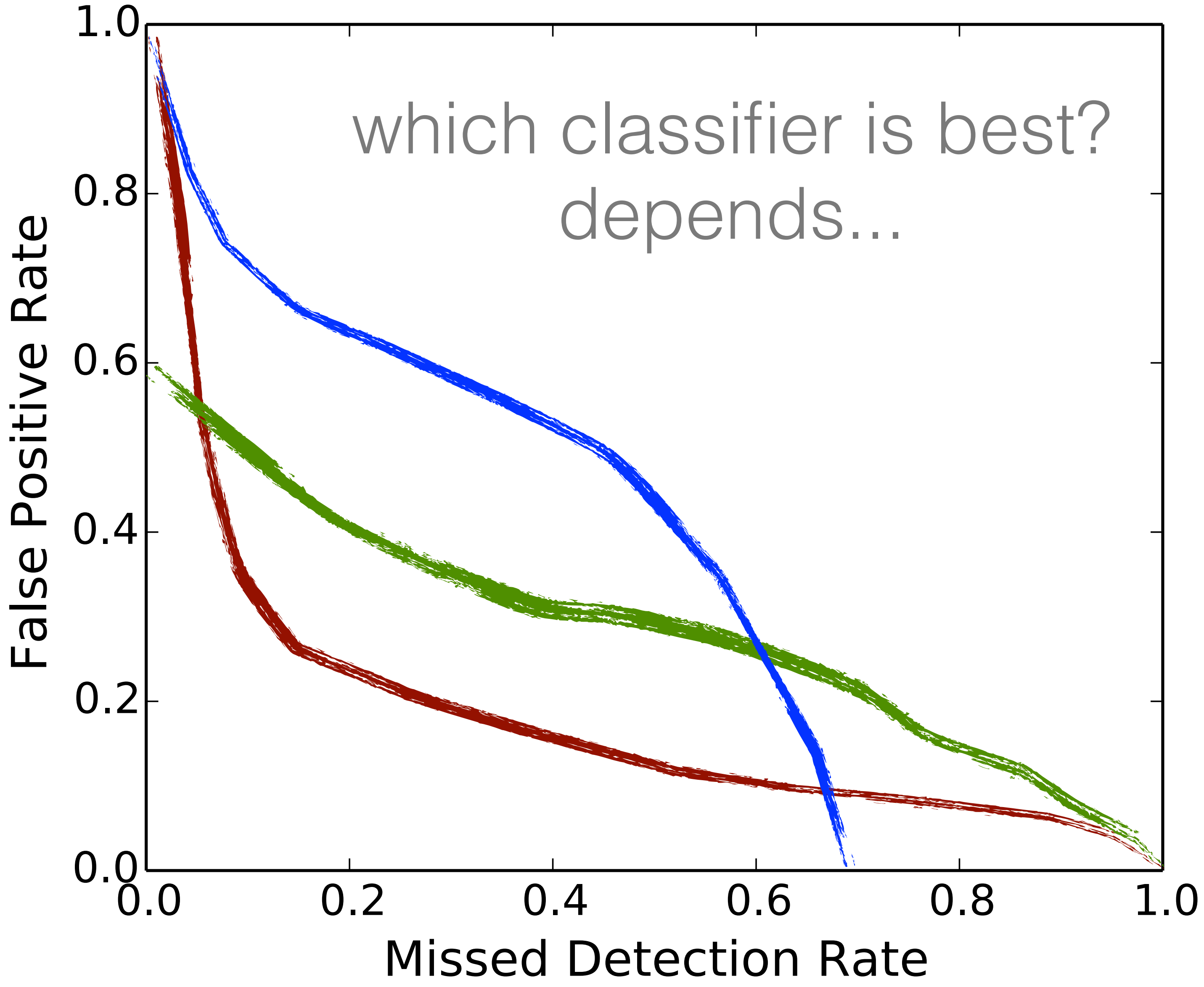
Evaluation Metric: *What's the essence of what I care about?*





Accuracy

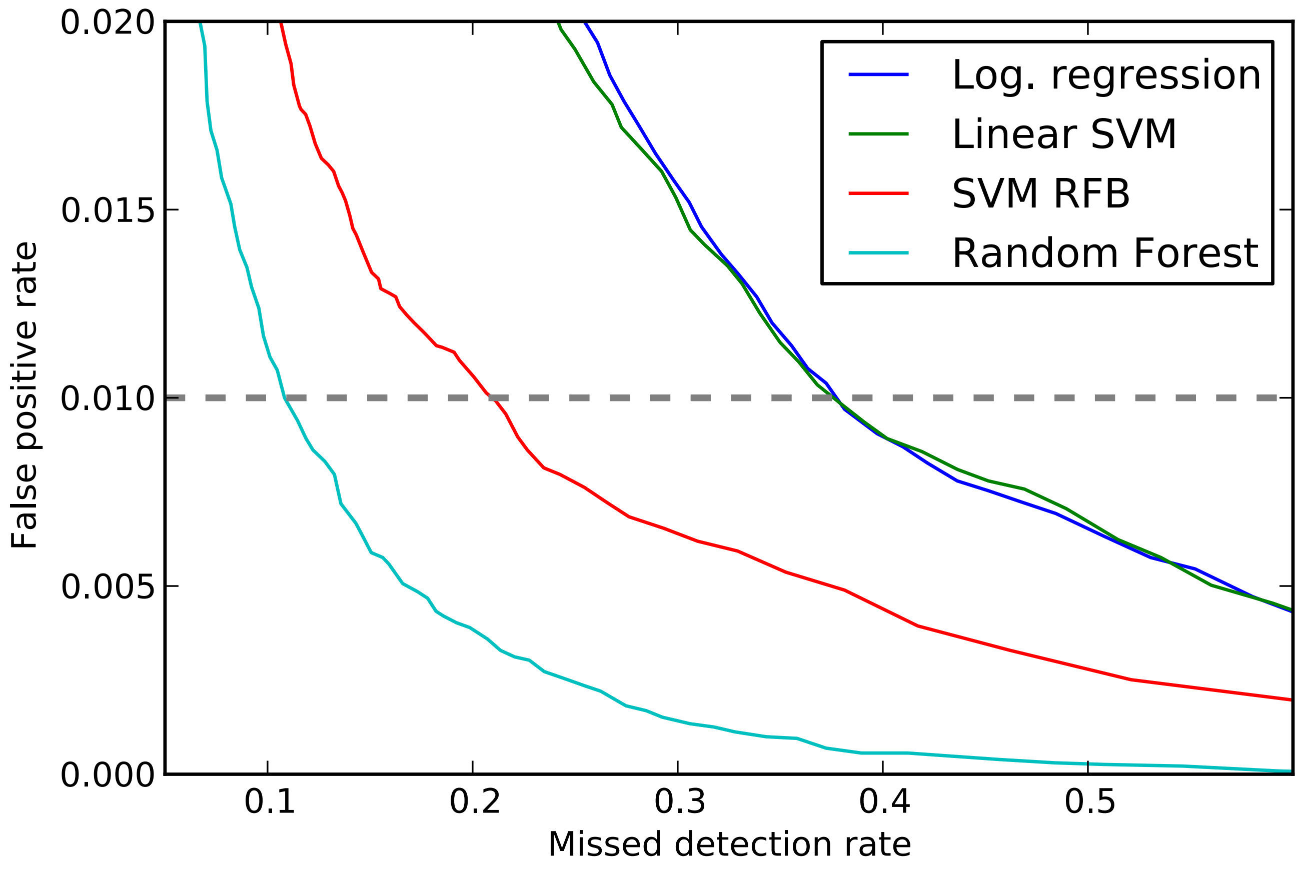
Evaluation Metric: *What's the essence of what I care about?*



Accuracy

Evaluation Metric: *What's the essence of what I care about?*

42-dimensional feature space

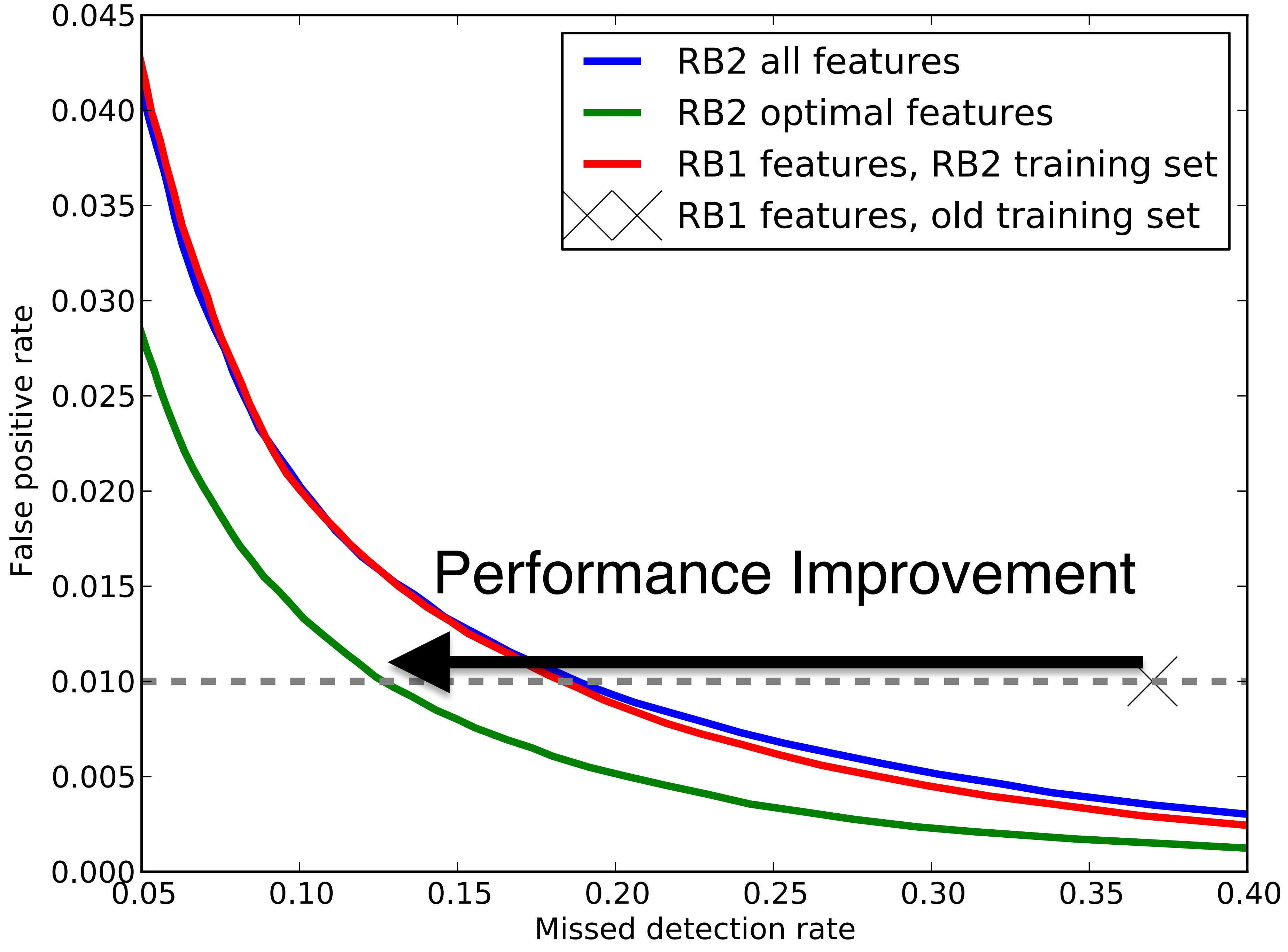


Brink, Bloom et al. 2012

Some ML algorithms just do *better*

Accuracy

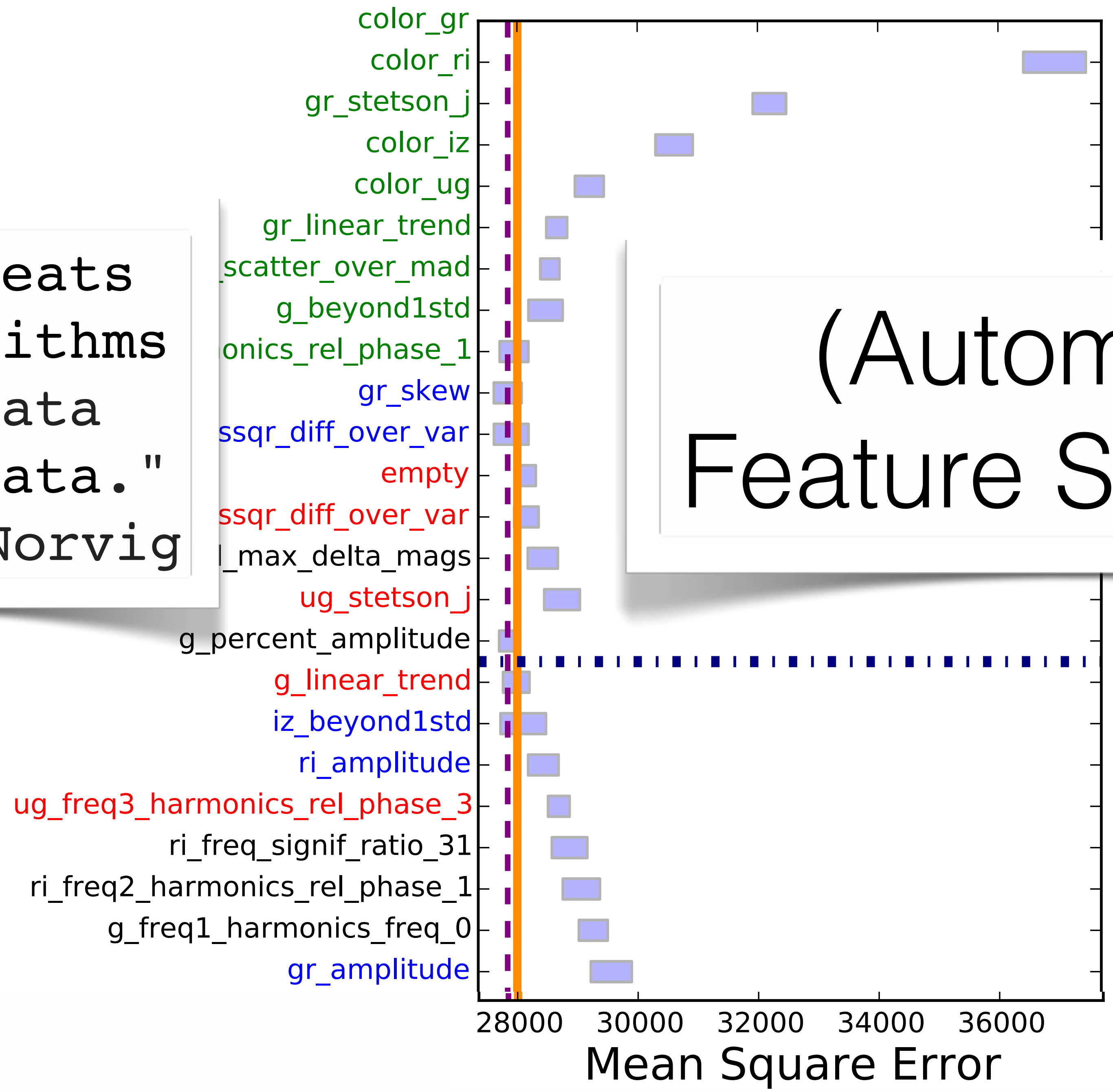
More Data (Dimensions) is better, but Protect Against Curse of Dimensionality



Accuracy

More Data (Dimensions) is better, but Protect Against Curse of Dimensionality

"More data beats clever algorithms but better data beats more data."
– Peter Norvig



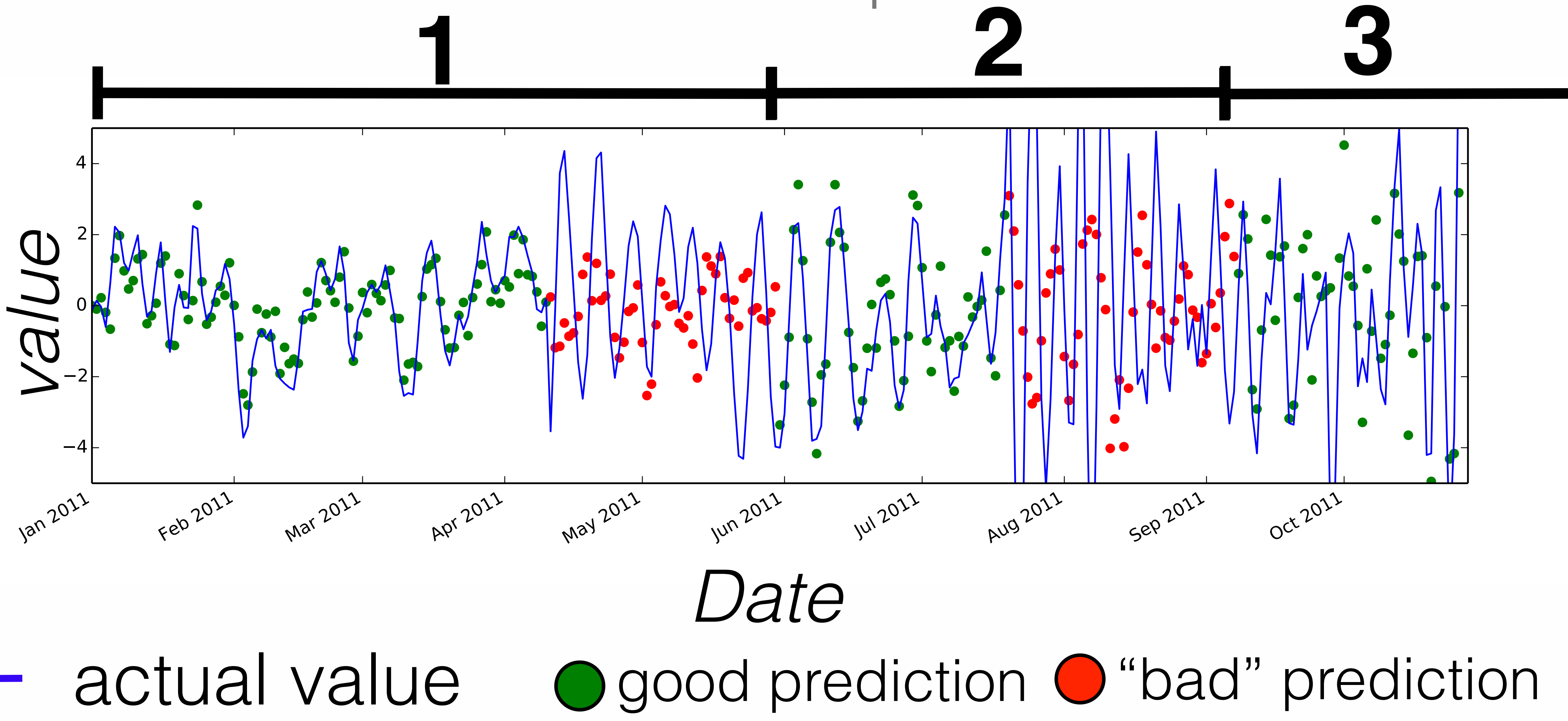
(Automatic) Feature Selection

Accuracy

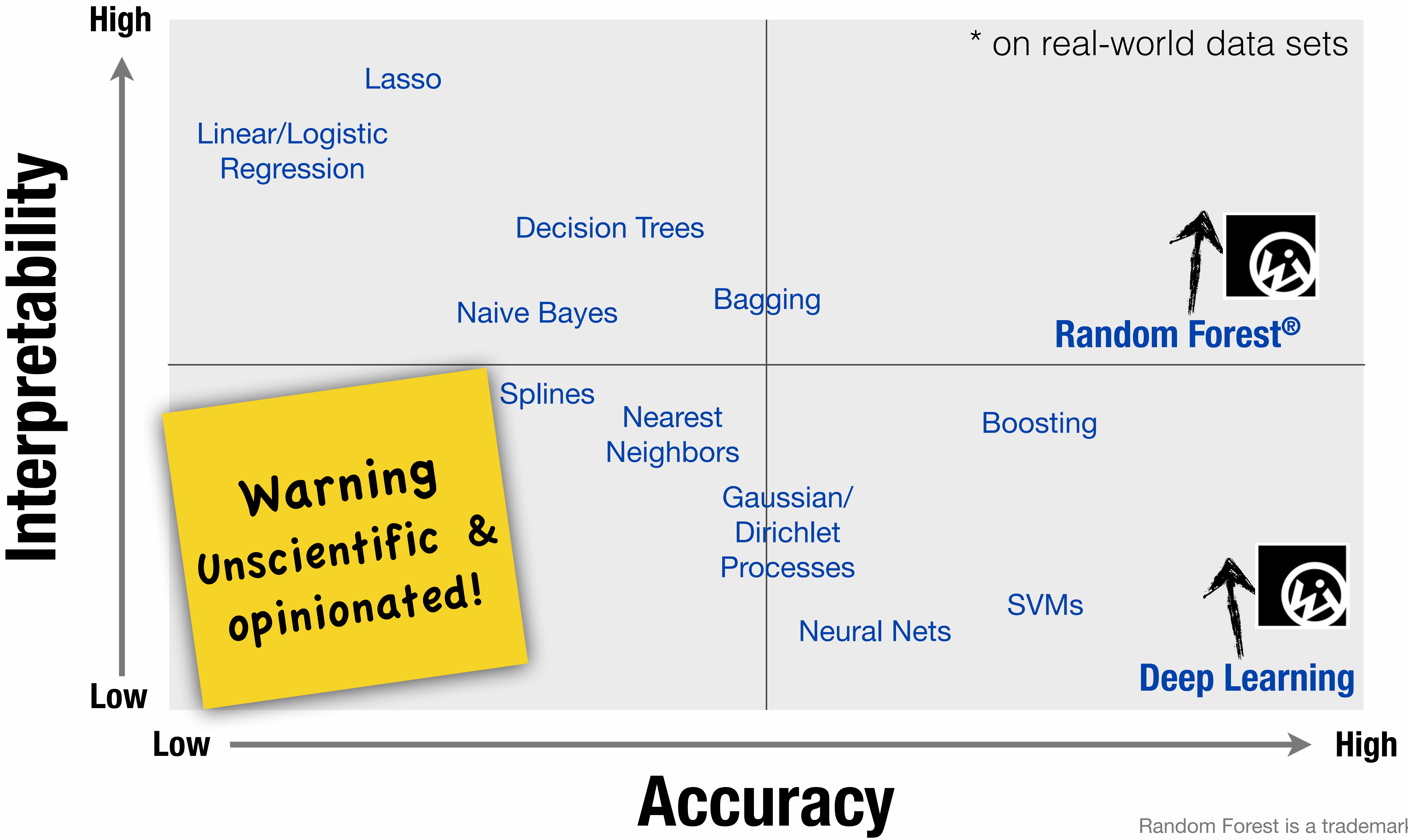
Testing Set & Continuous (Streaming) Testing & Model Updates

model 1 building + validation on historical data

Model # in production



ML Algorithmic Trade-Off



Interpretability

Interpretability

<u>How</u> does the model work?

Consider a nonlinear system of equations:

$$\begin{cases} 3x_1 - \cos(x_2x_3) - \frac{3}{2} = 0 \\ 4x_1^2 - 625x_2^2 + 2x_2 - 1 = 0 \\ \exp(-x_1x_2) + 20x_3 + \frac{10\pi-3}{3} \end{cases}$$

suppose we have the function

$$G(\mathbf{x}) = \begin{bmatrix} 3x_1 - \cos(x_2x_3) \\ 4x_1^2 - 625x_2^2 + 2 \\ \exp(-x_1x_2) + 20x_3 \end{bmatrix}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

and the objective function

$$F(\mathbf{x}) = \frac{1}{2}G^{\mathrm{T}}(\mathbf{x})G(\mathbf{x})$$
$$= \frac{1}{2}((3x_1 - \cos(x_2x_3) - \frac{3}{2})^2 +$$

With initial guess

$$\mathbf{x}^{(0)} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

We know that

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \gamma_0 \nabla F(x^{(0)})$$

where

$$\nabla F(\mathbf{x}^{(0)}) = J_G(\mathbf{x}^{(0)})^{\mathrm{T}} G(\mathbf{x}^{(0)})$$

The Jacobian matrix $J_G(\mathbf{x}^{(0)})$

$$J_G = \begin{bmatrix} 3 & \sin(x_2x_3)x_3 & \sin(x_2x_3)x_2 \\ 8x_1 & -1250x_2 + 2 & 0 \\ -x_2\exp(-x_1x_2) & -x_1\exp(-x_1x_2) & 20 \end{bmatrix}$$

Then evaluating these terms at $\mathbf{x}^{(0)}$

$$J_G\left(\mathbf{x}^{(0)}\right) = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 20 \end{bmatrix}$$

and

$$G(\mathbf{x}^{(0)}) = \begin{bmatrix} -2.5 \\ -1 \\ 10.472 \end{bmatrix}$$

So that

$$\mathbf{x}^{(1)} = 0 - \gamma_0 \begin{bmatrix} -7.5 \\ -2 \\ 209.44 \end{bmatrix}.$$

and

$$F\left(\mathbf{x}^{(0)}\right) = 0.5((-2.5)^2 + (-1)^2 + (10.472)^2) = 58.456$$

Now a suitable γ_0 must be found such that $F(\mathbf{x}^{(1)}) \le F(\mathbf{x}^{(0)})$. This can be done with algorithms. One might also simply guess $\gamma_0 = 0.001$ which gives

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0.0075 \\ 0.002 \\ -0.20944 \end{bmatrix}$$

evaluating at this value,

$$F\left(\mathbf{x}^{(1)}\right) = 0.5((-2.48)^2 + (-1.00)^2 + (6.28)^2) = 23.306$$

Interpretability

~~How does the model work?~~

Consider a nonlinear system of equations:

$$\begin{cases} 3x_1 - \cos(x_2x_3) - \frac{3}{2} = 0 \\ 4x_1^2 - 625x_2^2 + 2x_2 - 1 = 0 \\ \exp(-x_1x_2) + 20x_3 + \frac{10\pi-3}{3} \end{cases}$$

suppose we have the function

$$G(\mathbf{x}) = \begin{bmatrix} 3x_1 - \cos(x_2x_3) \\ 4x_1^2 - 625x_2^2 + 2 \\ \exp(-x_1x_2) + 20x_3 \end{bmatrix}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

and the objective function

$$F(\mathbf{x}) = \frac{1}{2}G^{\mathrm{T}}(\mathbf{x})G(\mathbf{x})$$
$$= \frac{1}{2}((3x_1 - \cos(x_2x_3) - \frac{3}{2})^2 +$$

With initial guess

$$\mathbf{x}^{(0)} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

We know that

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \gamma_0 \nabla F(x^{(0)})$$

where

$$\nabla F(\mathbf{x}^{(0)}) = J_G(\mathbf{x}^{(0)})^{\mathrm{T}} G(\mathbf{x}^{(0)})$$

The Jacobian matrix $J_G(\mathbf{x}^{(0)})$

$$J_G = \begin{bmatrix} 3 & \sin(x_2x_3)x_3 & \sin(x_2x_3)x_2 \\ 8x_1 & -1250x_2 + 2 & 0 \\ -x_2\exp(-x_1x_2) & -x_1\exp(-x_1x_2) & 20 \end{bmatrix}$$

Then evaluating these terms at $\mathbf{x}^{(0)}$

$$J_G(\mathbf{x}^{(0)}) = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 20 \end{bmatrix}$$

and

$$G(\mathbf{x}^{(0)}) = \begin{bmatrix} -2.5 \\ -1 \\ 10.472 \end{bmatrix}$$

So that

$$\mathbf{x}^{(1)} = 0 - \gamma_0 \begin{bmatrix} -7.5 \\ -2 \\ 209.44 \end{bmatrix}.$$

and

$$F(\mathbf{x}^{(0)}) = 0.5((-2.5)^2 + (-1)^2 + (10.472)^2) = 58.456$$

Now a suitable γ_0 must be found such that $F(\mathbf{x}^{(1)}) \leq F(\mathbf{x}^{(0)})$. This can be done with algorithms. One might also simply guess $\gamma_0 = 0.001$ which gives

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0.0075 \\ 0.002 \\ -0.20944 \end{bmatrix}$$

evaluating at this value,

$$F(\mathbf{x}^{(1)}) = 0.5((-2.48)^2 + (-1.00)^2 + (6.28)^2) = 23.306$$

Interpretability

Why do I get these answers?

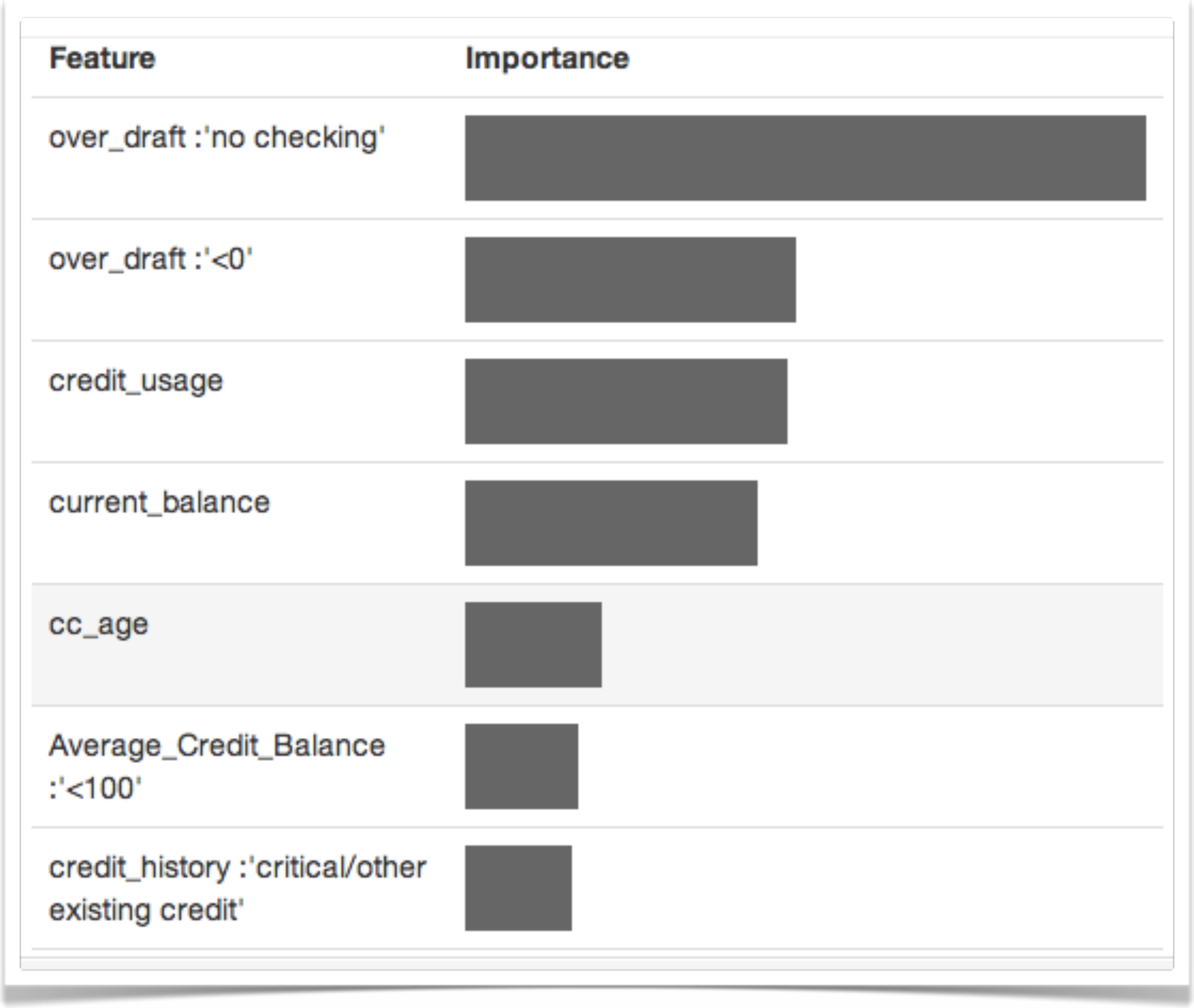
e.g., Credit score

Sample FICO® Scoring Model

Category	Characteristic	Attributes	Points
Payment History	Number of months since the most recent derogatory public record	No public record	75
		0 – 5	10
		6 – 11	15
		12 – 23	25
		24+	55
Outstanding Debt	Average balance on revolving trades	No revolving trades	30
		0	55
		1 – 99	65
		100 – 499	50
		500 – 749	40
		750 – 999	25
		1000 or more	15
Credit History Length	Number of months in file	Below 12	12
		12 – 23	35
		24 – 47	60
		48 or more	75
Pursuit of New Credit	Number of inquiries in last 6 mos.	0	70
		1	60
		2	45
		3	25
		4+	20
Credit Mix	Number of bankcard trade lines	0	15
		1	25
		2	50
		3	60
		4+	50

Interpretability

Peering Inside the Black Box

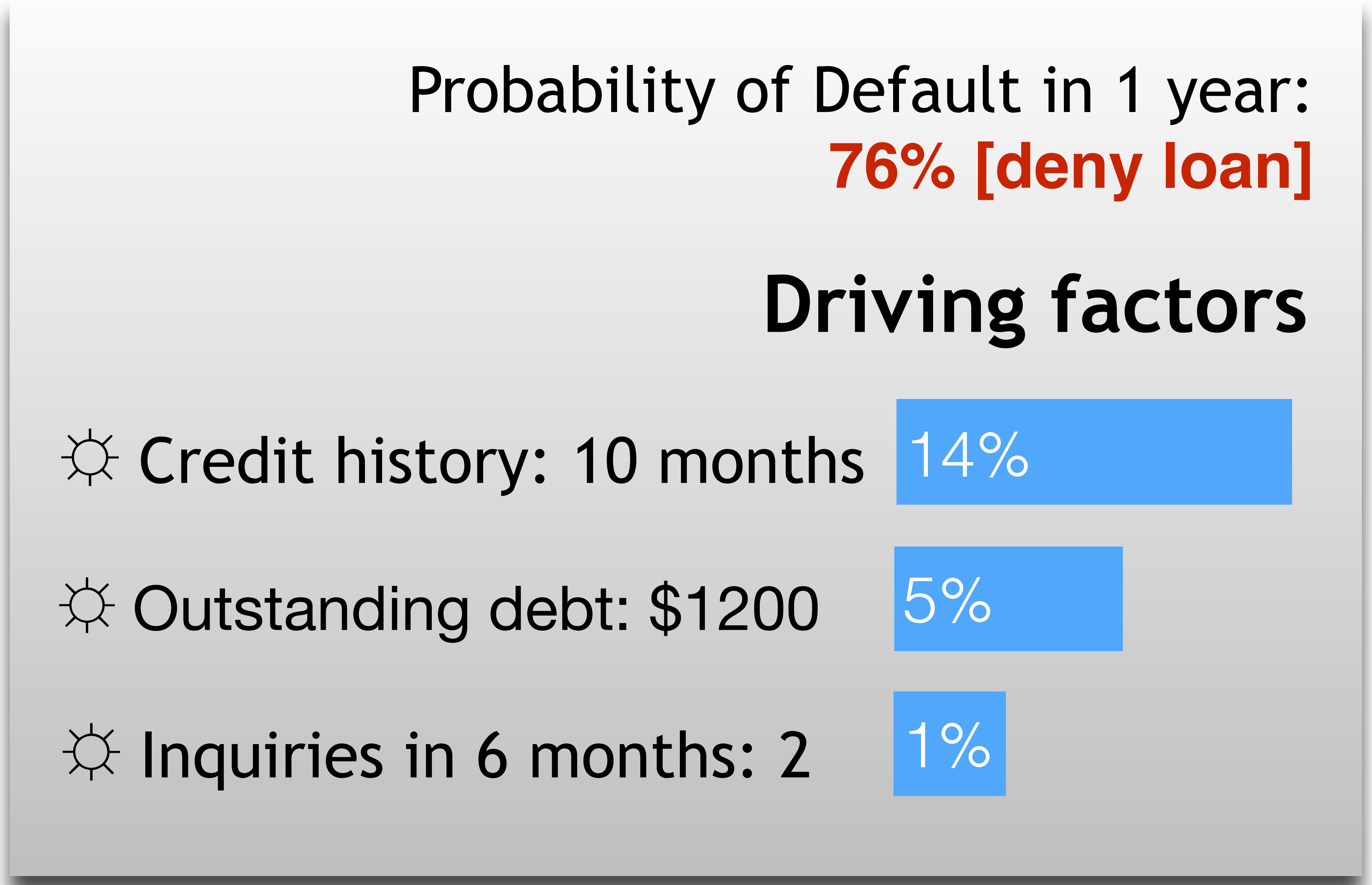


Random Forest®
model-level
feature importance

Interpretability

Peering Inside the Black Box

Individual-level prediction feature importance

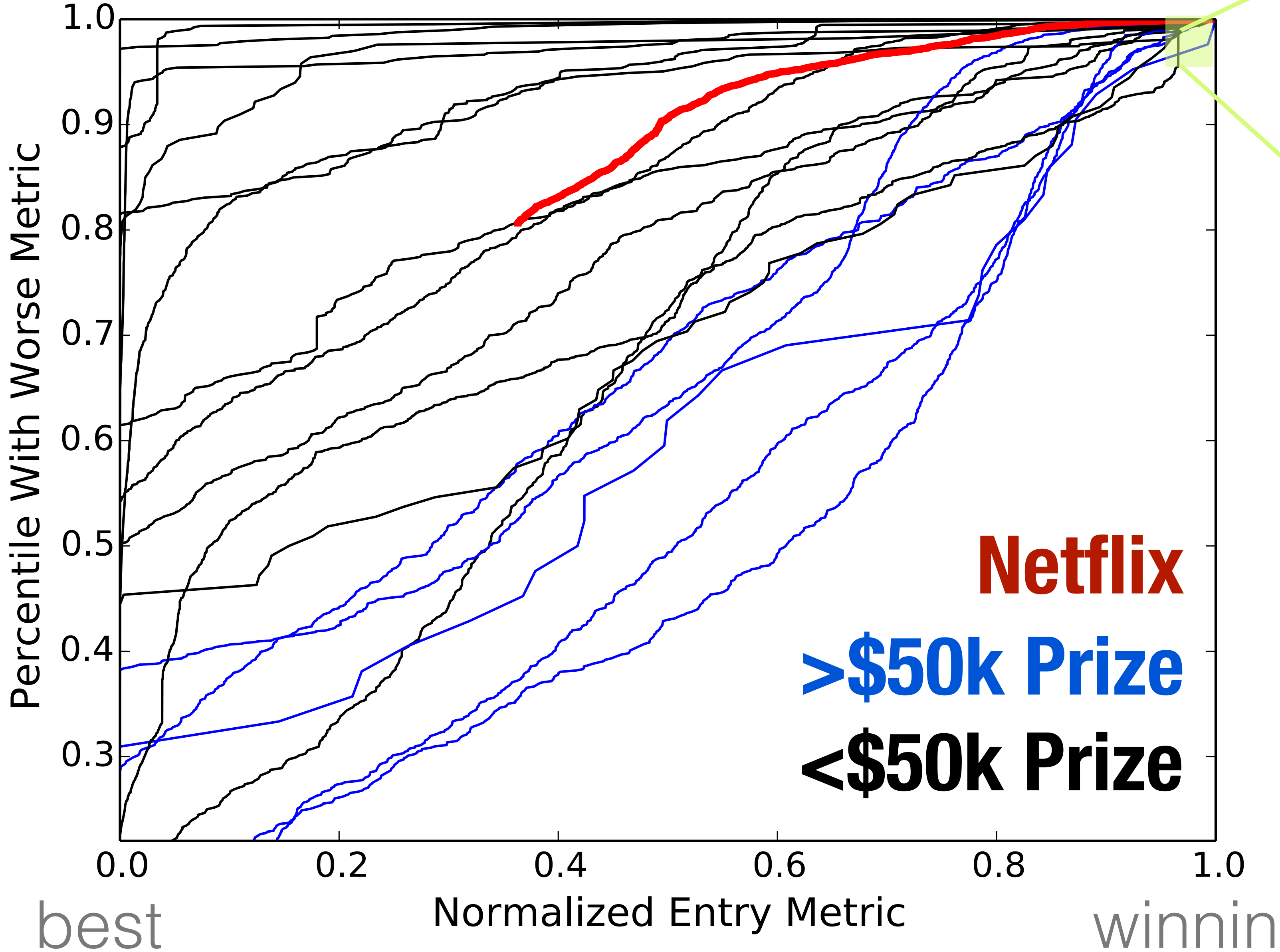


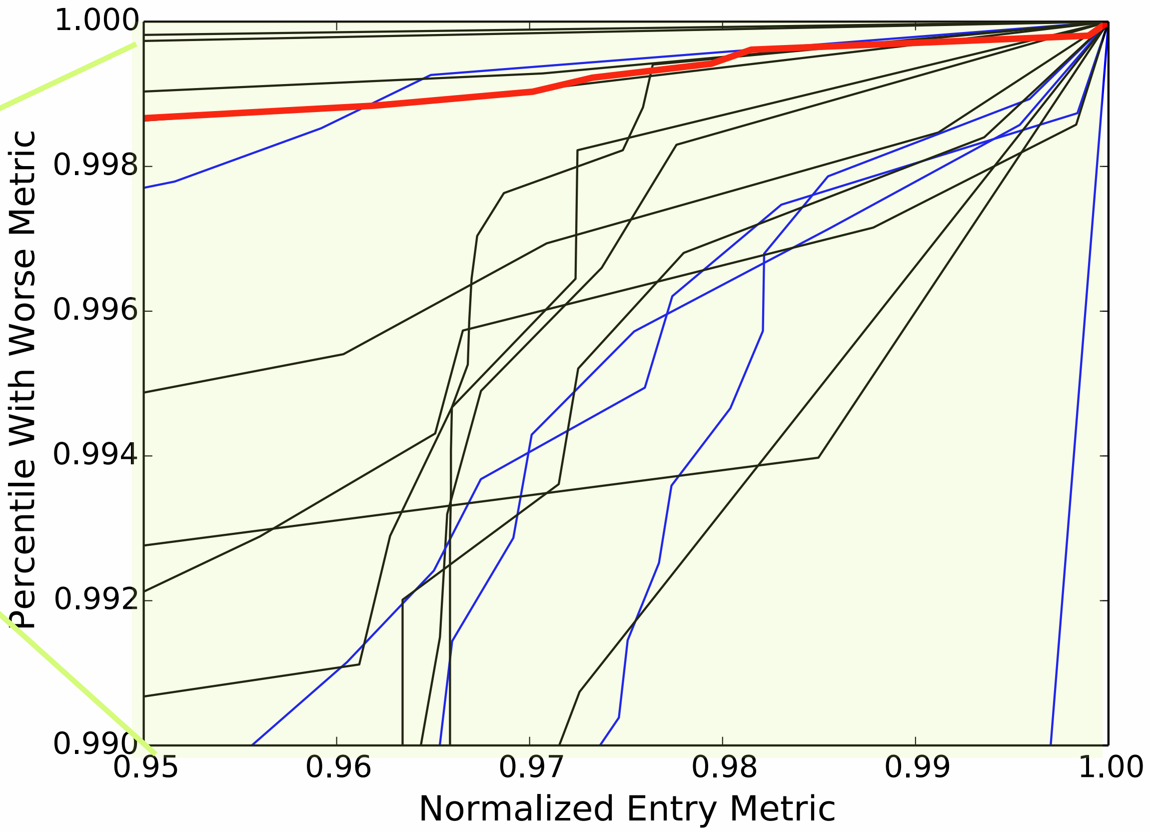
e.g. microcredit application scorecard

Implementability

How long does it take to put the model into production? At what cost?

Implementability





best benchmark

winning metric

many teams get within ~few % of optimum

so which is easier to put into production?

Leaderboard data from Kaggle & Netflix

Implementability

On the NETFLIX Prize

"We evaluated some of the new methods offline but the **additional accuracy gains** that we measured **did not seem to justify the engineering effort** needed to bring them into a **production environment.**"

Xavier Amatriain and Justin Basilico (April 2012)

Implementability

The divide between **data science** & **production**

Implementability

Treat Machine Learning Deployment as you would Software

- Continuous Deployment
- RESTful API
- Language bindings
- Security
- SLA

Integration

Connect data

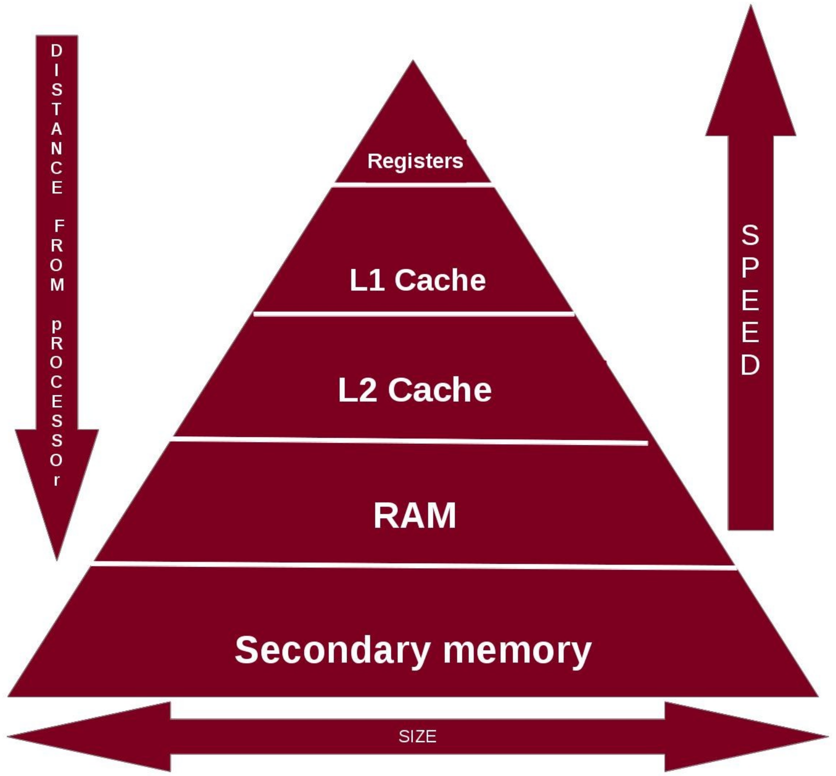
Consume predictions

Implementability

Scalability & Speed

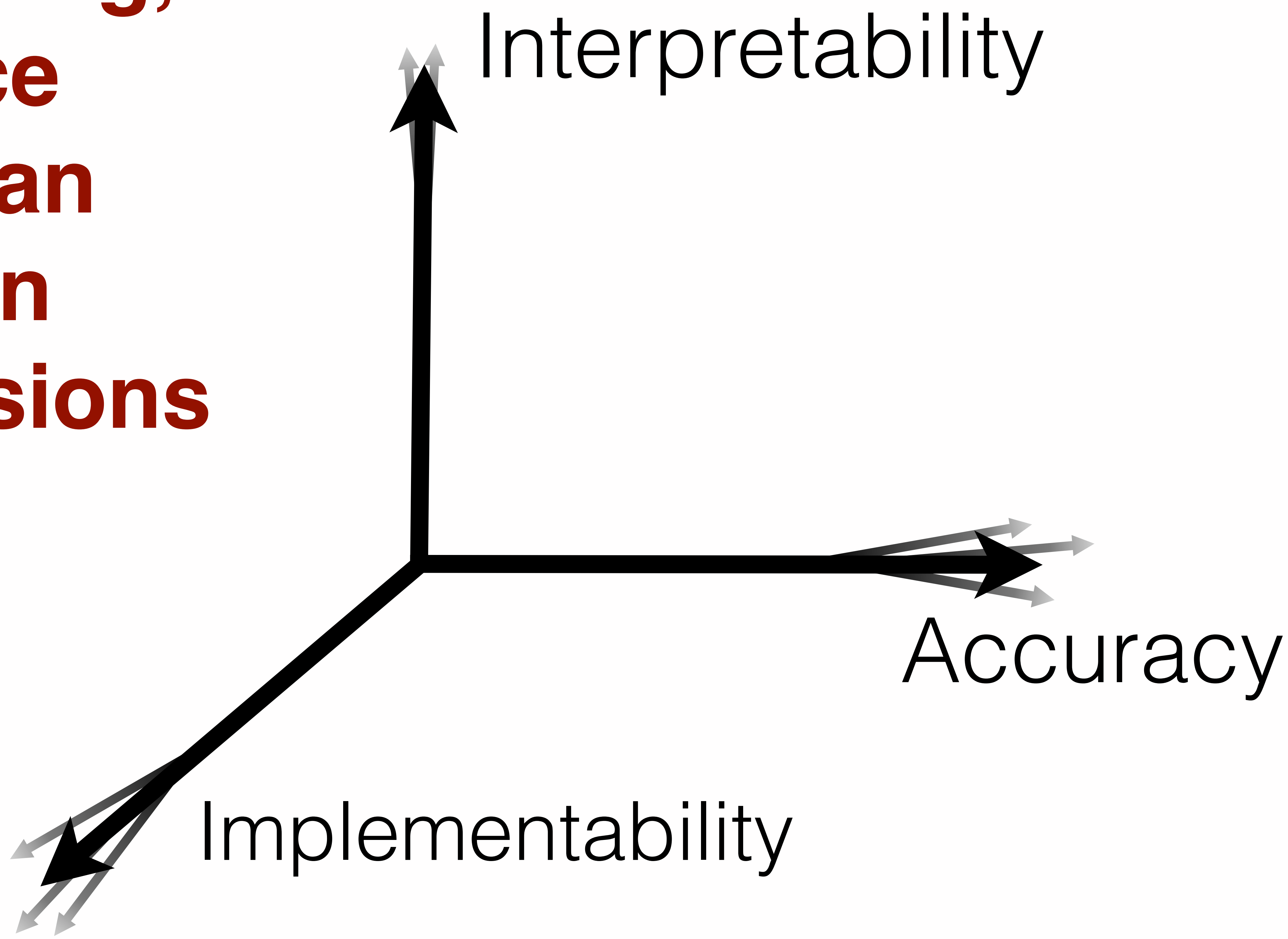
Micro-scaling

Fast, efficient use of memory hierarchy



Horizontally scalable data processing

Machine-Learning, Data Science Workflow is an Optimization in *many* dimensions



wise.io

@wiseio

We are Hiring!

- Full-stack developers
  - Javascript, Python, Spark/Shark
- Front end developers
- DevOps engineers
- C++ engineers
  - C++ template metaprogramming
- Data scientists
  - Python, Deep NN, ML expertise

wise.io

jobs@wise.io

http://wise.io/jobs/