OPTIMIZATION PROBLEMS OVERVIEW

CMSC764 / AMSC607

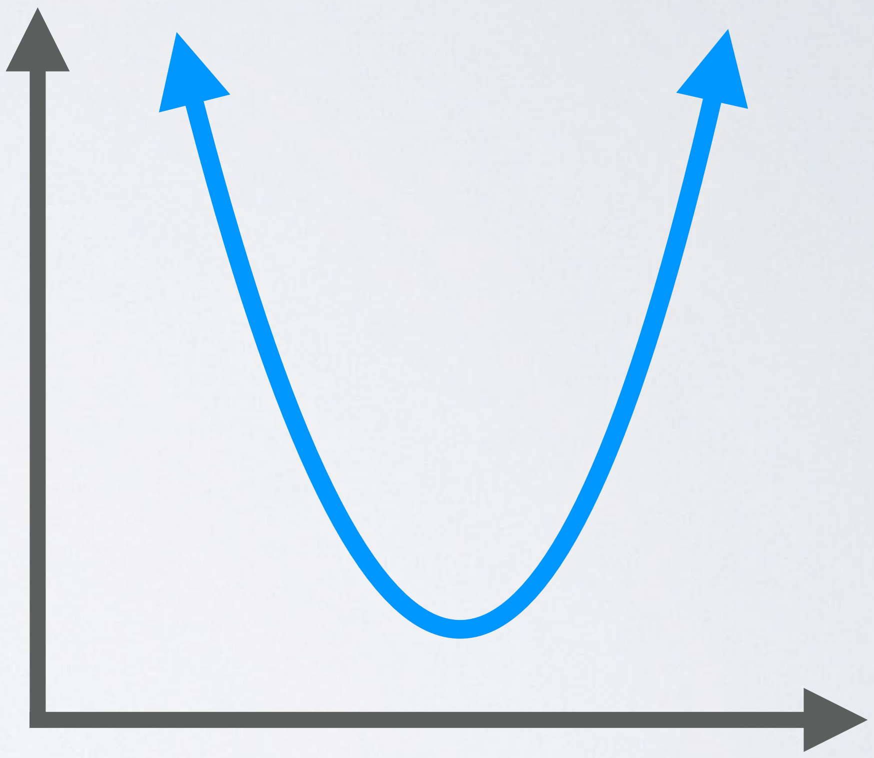
WHAT IS OPTIMIZATION?

A: Minimizing things

In college you learned:
set derivative to zero

sounds easy.

convex!



BUT THEN…

What if there's no closed-form solution?

What if the problem has 1 BILLION dimensions?

What if the problem is non-convex?

What if the function has no derivative?

What if there are constraints?

What if the objective function has a BILLION terms?

Does this ever *really* happen?

MODEL FITTING PROBLEMS

BASIC OPTIMIZATION PROBLEMS: MODEL FITTING

training data / inputs

label data / outputs

$$f(d_i, w) = y_i$$

model

parameters

$$\min \sum_i \ell(d_i, w, y_i)$$

loss function

Example: linear model

$$d_i^T w = b_i$$

least-squares

$$\min \|Dw - b\|^2$$

$$\ell(d_i, w, b_i) = (d_i^T w - b_i)^2$$

BASIC OPTIMIZATION PROBLEMS: MODEL FITTING

$$\min \sum_i \ell(d_i, w, y_i)$$

loss function

$$\min \|Dw - b\|^2$$

$$\ell(d_i, w, b_i) = (d_i^T w - b_i)^2$$

penalized regressions

$$\min J(w) + \sum_i \ell(d_i, w, y_i)$$

“prior”

$$\min \|w\|_2^2 + \|Dw - b\|^2$$

ridge penalty

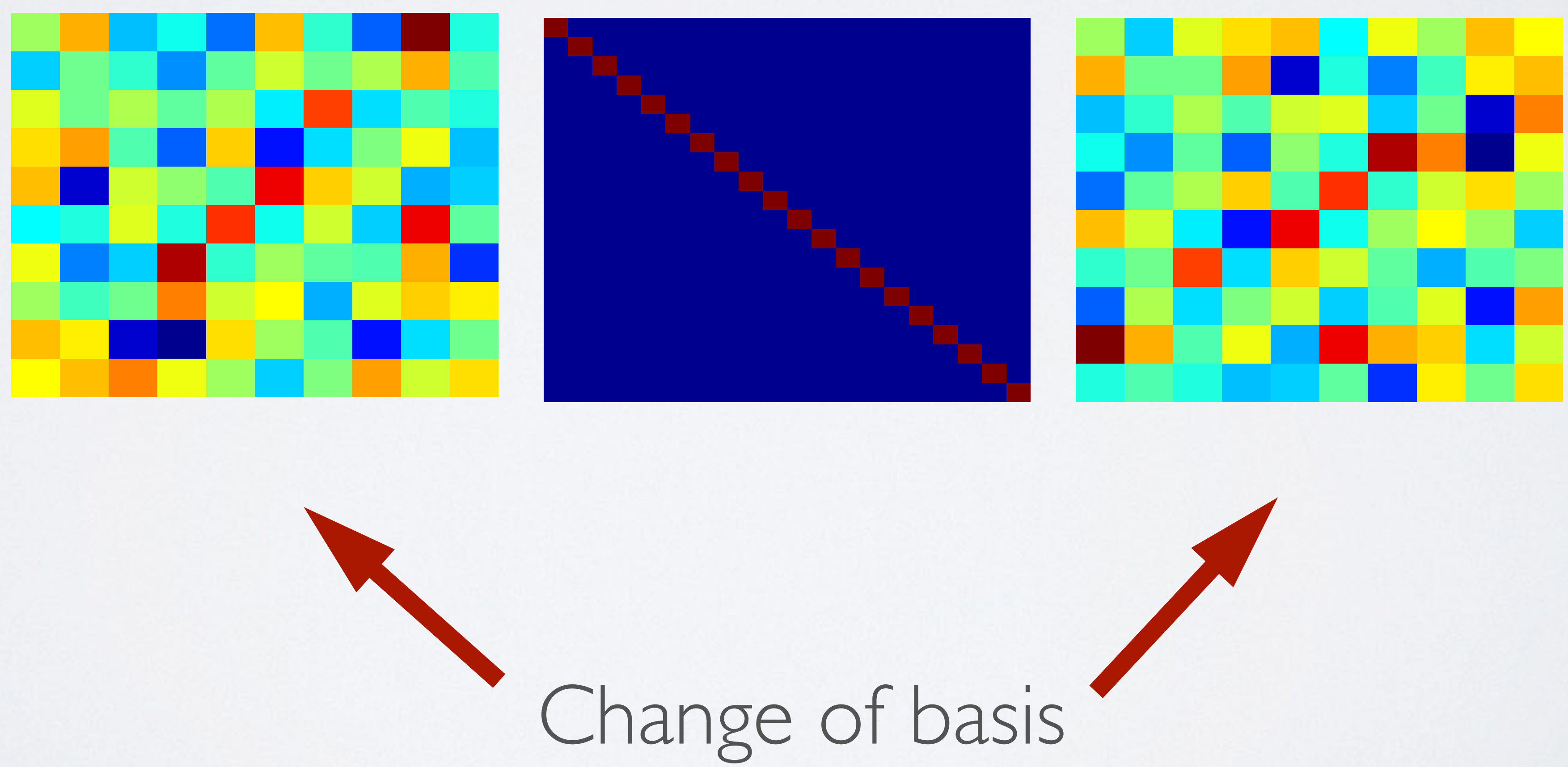
Why would you want a penalty?

Poor conditioning.

EIGENVALUE DECOMPOSITION

- Spectral theorem: symmetric matrices have a **complete**, **orthogonal** set of eigenvalues

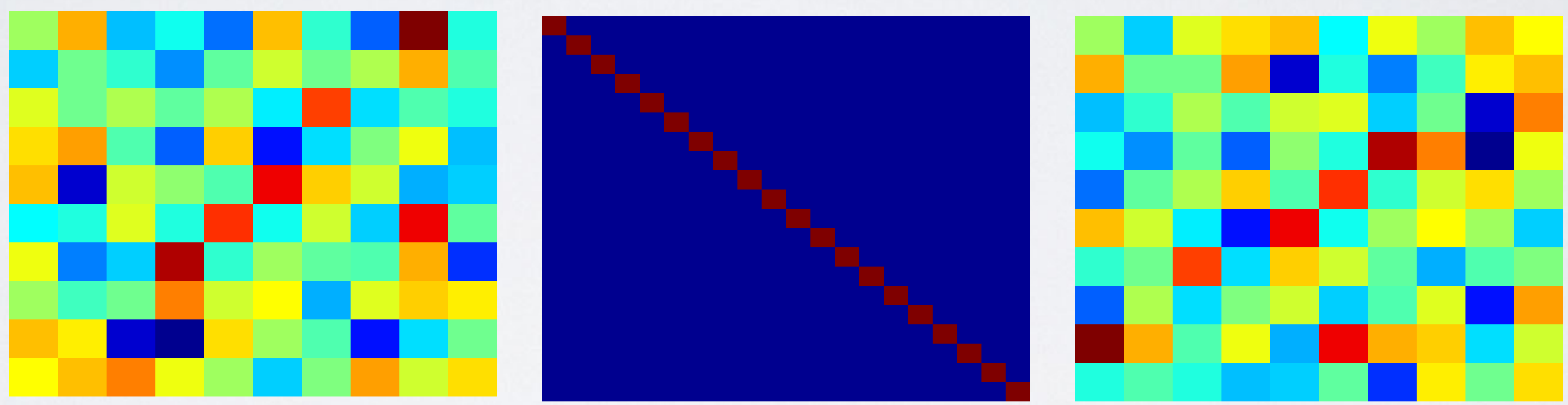
$$A =$$



EIGENVALUE DECOMPOSITION

- Action of matrix is described by eigenvalues

$$A =$$



$$x = \beta_1 \mathbf{e_1} + \beta_2 \mathbf{e_2} + \beta_3 \mathbf{e_3}$$

$$\begin{aligned} Ax &= A(\beta_1 \mathbf{e_1} + \beta_2 \mathbf{e_2} + \beta_3 \mathbf{e_3}) \\ &= \beta_1 A \mathbf{e_1} + \beta_2 A \mathbf{e_2} + \beta_3 A \mathbf{e_3} \\ &= \beta_1 \lambda_1 \mathbf{e_1} + \beta_2 \lambda_2 \mathbf{e_2} + \beta_3 \lambda_3 \mathbf{e_3} \end{aligned}$$

MATRIX INVERSE

- Action of matrix is described by eigenvalues

$$\begin{aligned} Ax &= A(\beta_1 \mathbf{e_1} + \beta_2 \mathbf{e_2} + \beta_3 \mathbf{e_3}) \\ &= \beta_1 A\mathbf{e_1} + \beta_2 A\mathbf{e_2} + \beta_3 A\mathbf{e_3} \\ &= \beta_1 \lambda_1 \mathbf{e_1} + \beta_2 \lambda_2 \mathbf{e_2} + \beta_3 \lambda_3 \mathbf{e_3} \end{aligned}$$

$$A^{-1}x = \beta_1 \lambda_1^{-1} \mathbf{e_1} + \beta_2 \lambda_2^{-1} \mathbf{e_2} + \beta_3 \lambda_3^{-1} \mathbf{e_3}$$

ESTIMATION PROBLEM

Suppose $\lambda_1 = 1,\ \lambda_2 = 0.1,\ \lambda_2 = 0.01$

$$Ax = b = \beta_1 \mathbf{e_1} + 0.1\beta_2 \mathbf{e_2} + 0.01\beta_3 \mathbf{e_3}$$

I tell you $$\hat{b} = b + \eta$$ ← noise

Can you recover x?

Does this ever *really* happen?

CONDITION NUMBER

- Ratio of largest to smallest singular value

$$\kappa = \frac{\sigma_{max}}{\sigma_{min}} \qquad \kappa = \left|\frac{\lambda_{max}}{\lambda_{min}}\right| \qquad \kappa = \|A\|\|A^{-1}\|$$

$$Ax = b$$

vs

$$Ax = \hat{b}$$

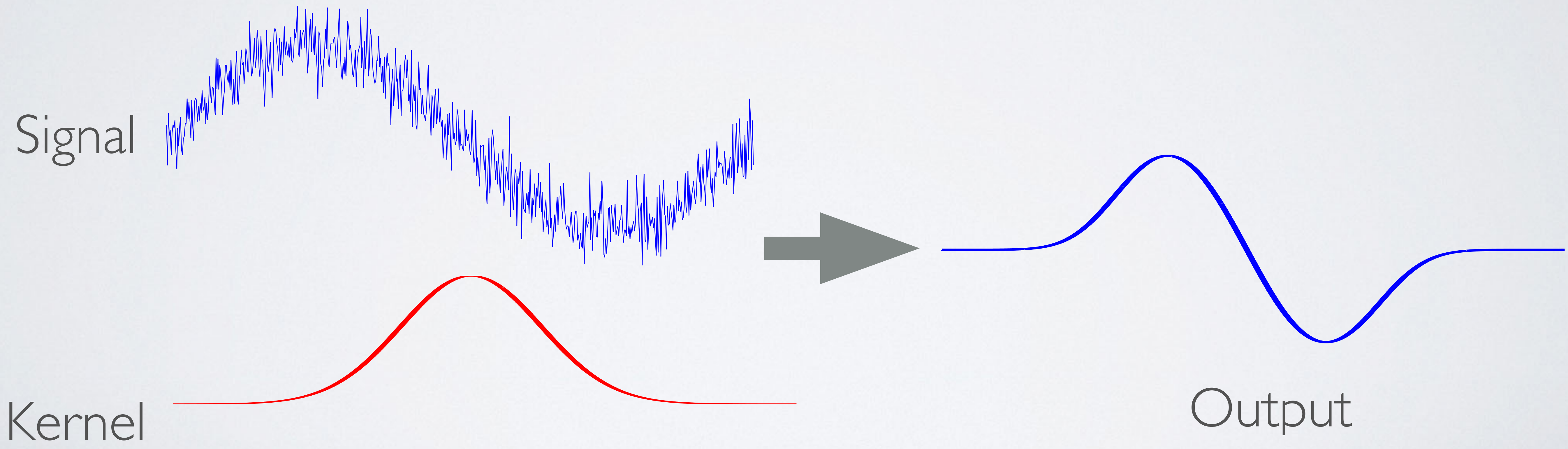
$$\longrightarrow \quad \frac{\|\hat{x} - x\|}{\|x\|} \leq \kappa \frac{\|b - \hat{b}\|}{\|b\|}$$

Why are these definitions the same for symmetric matrices?

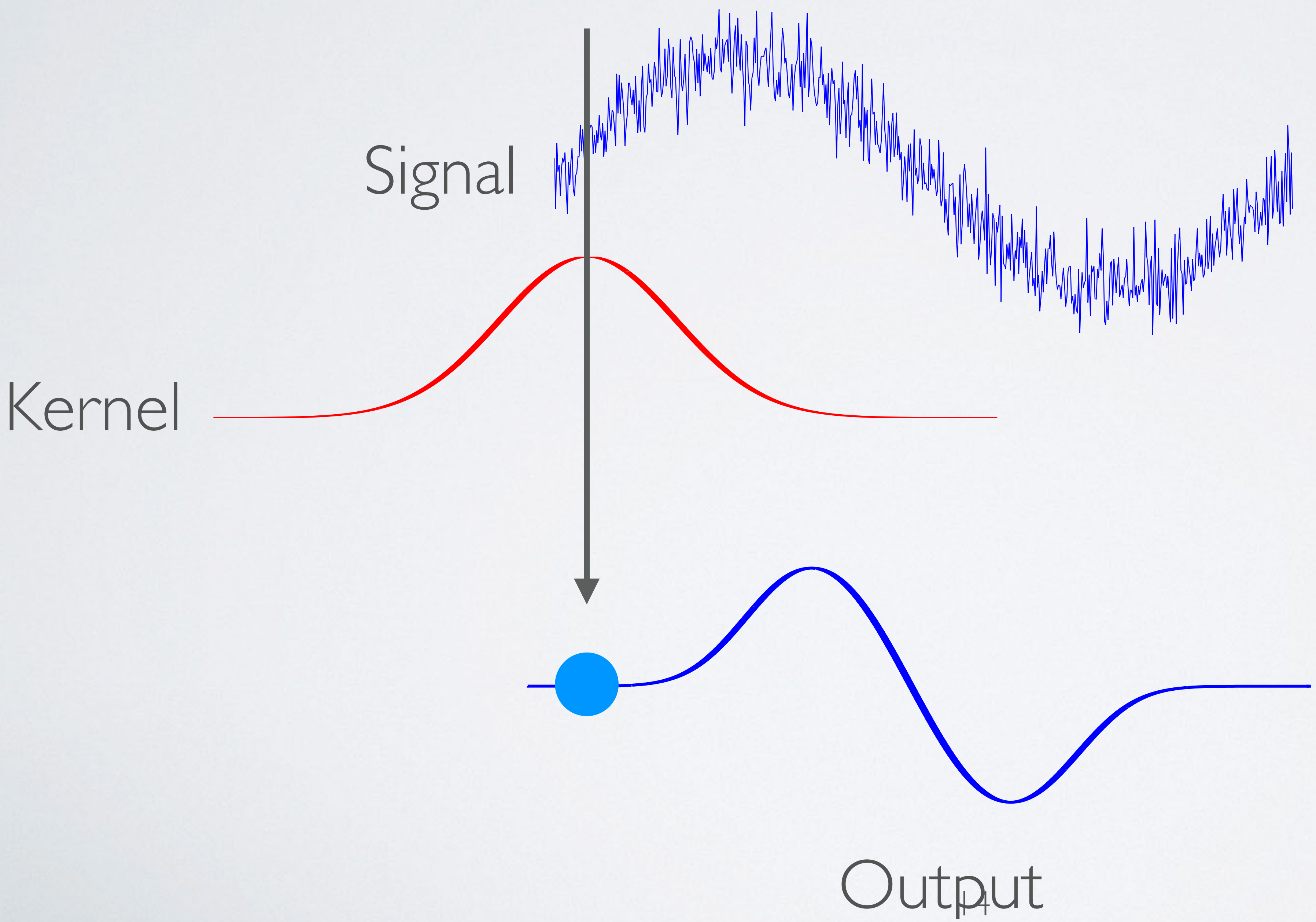
What is the condition number of our problem?

DOES THIS EVER HAPPEN IN REAL LIFE?

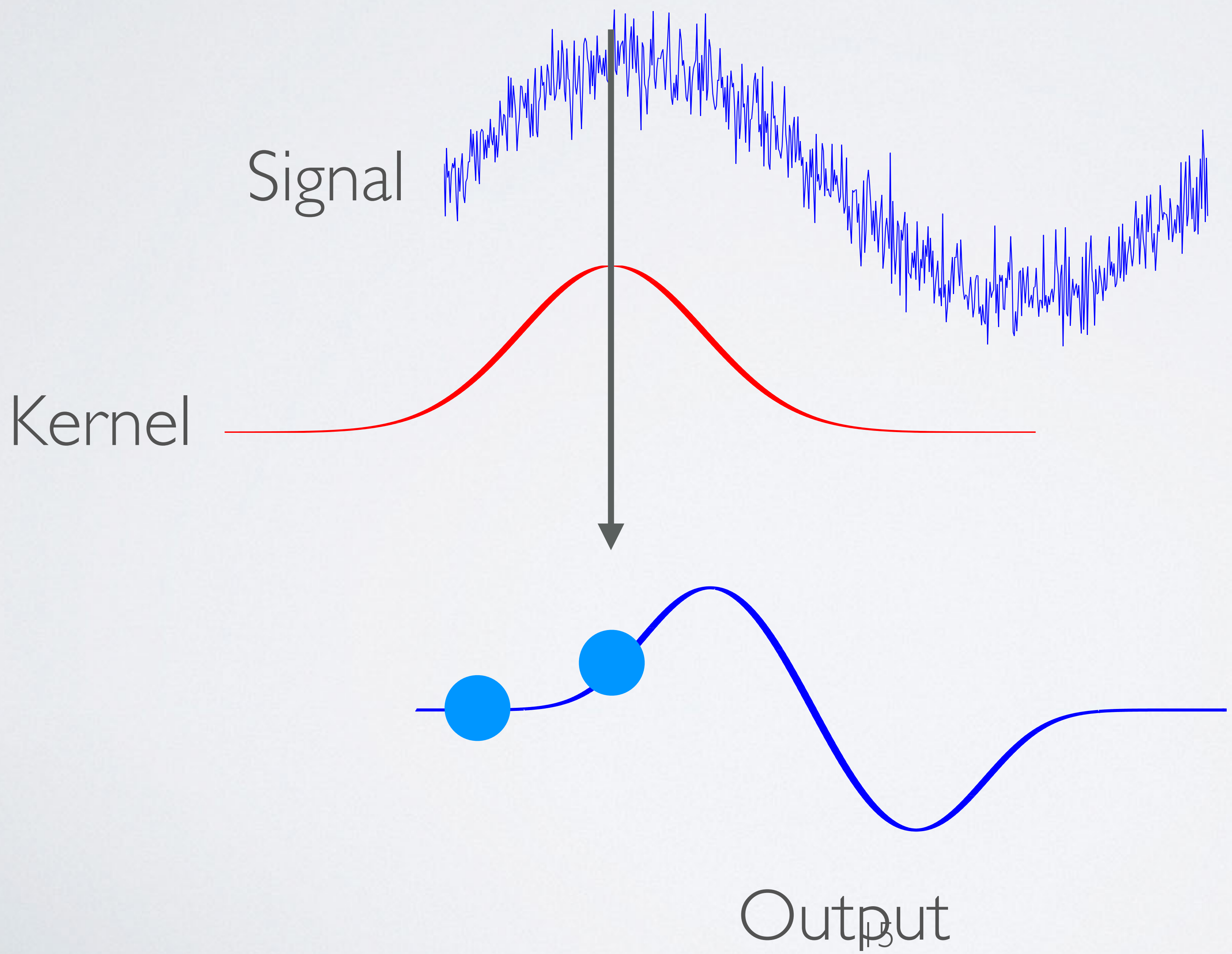
- No. The situation is never this good.
- Common in optimization: regularizations and IPM's
- Example: Convolution



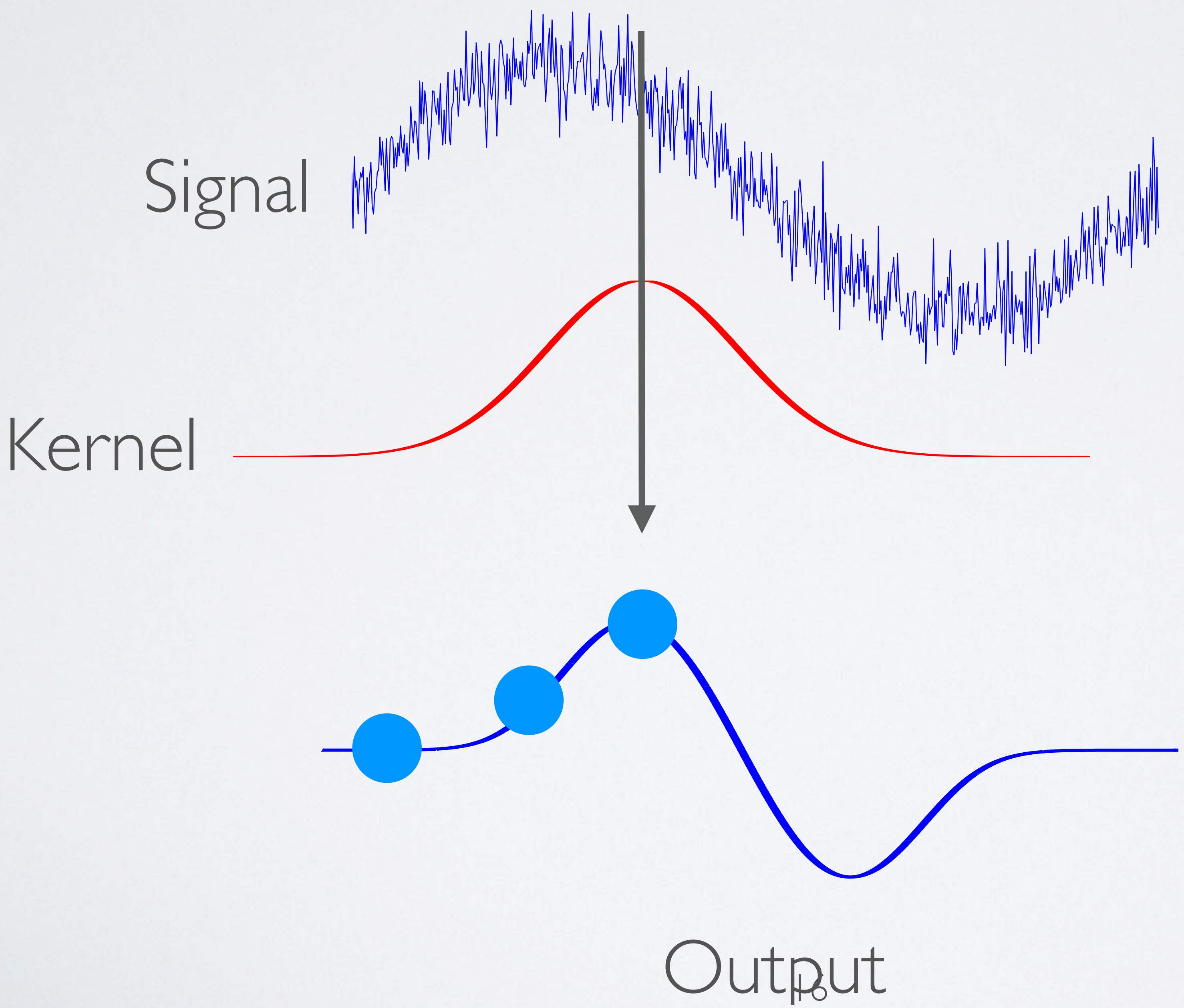
DOES THIS EVER HAPPEN IN REAL LIFE?



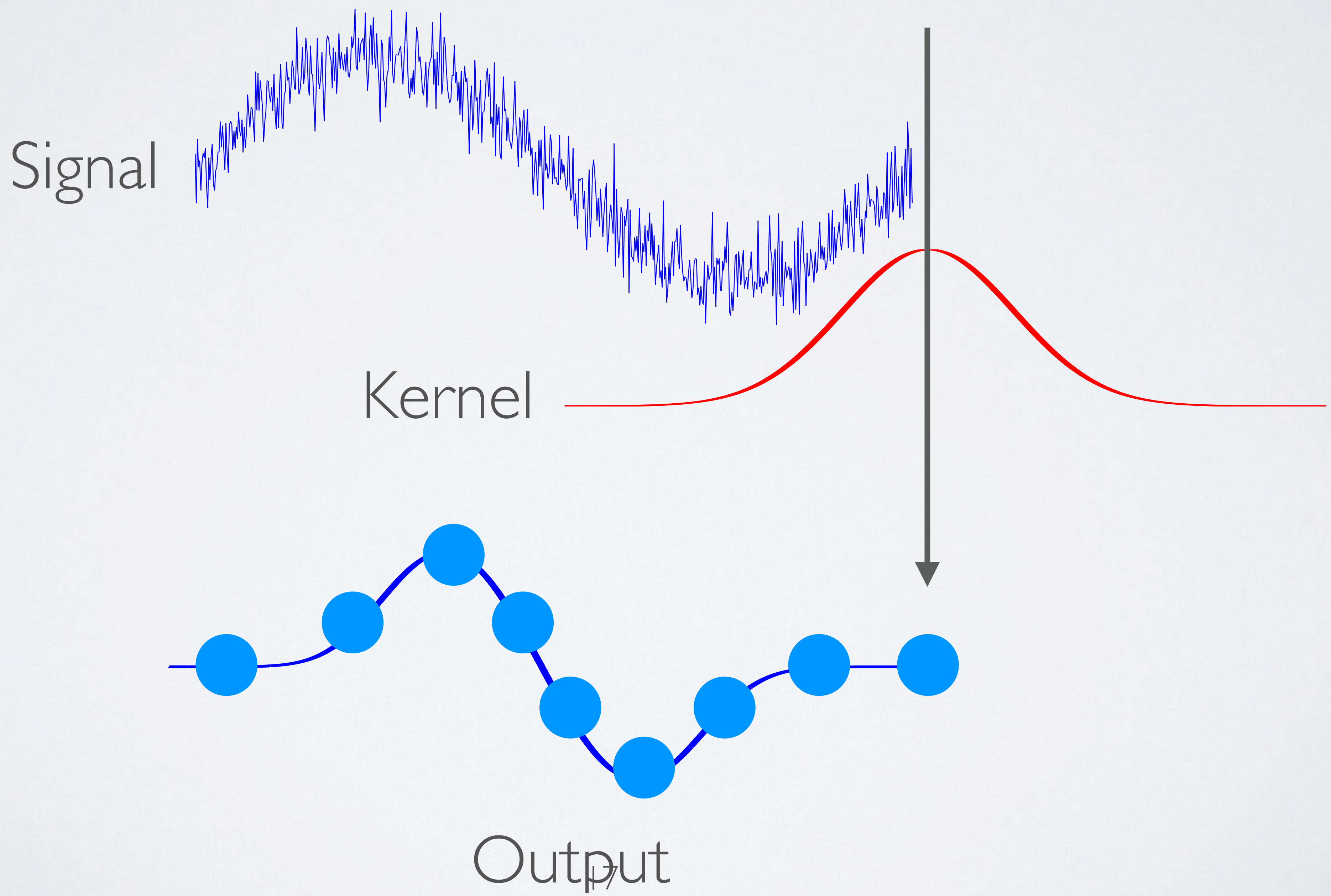
DOES THIS EVER HAPPEN IN REAL LIFE?



DOES THIS EVER HAPPEN IN REAL LIFE?

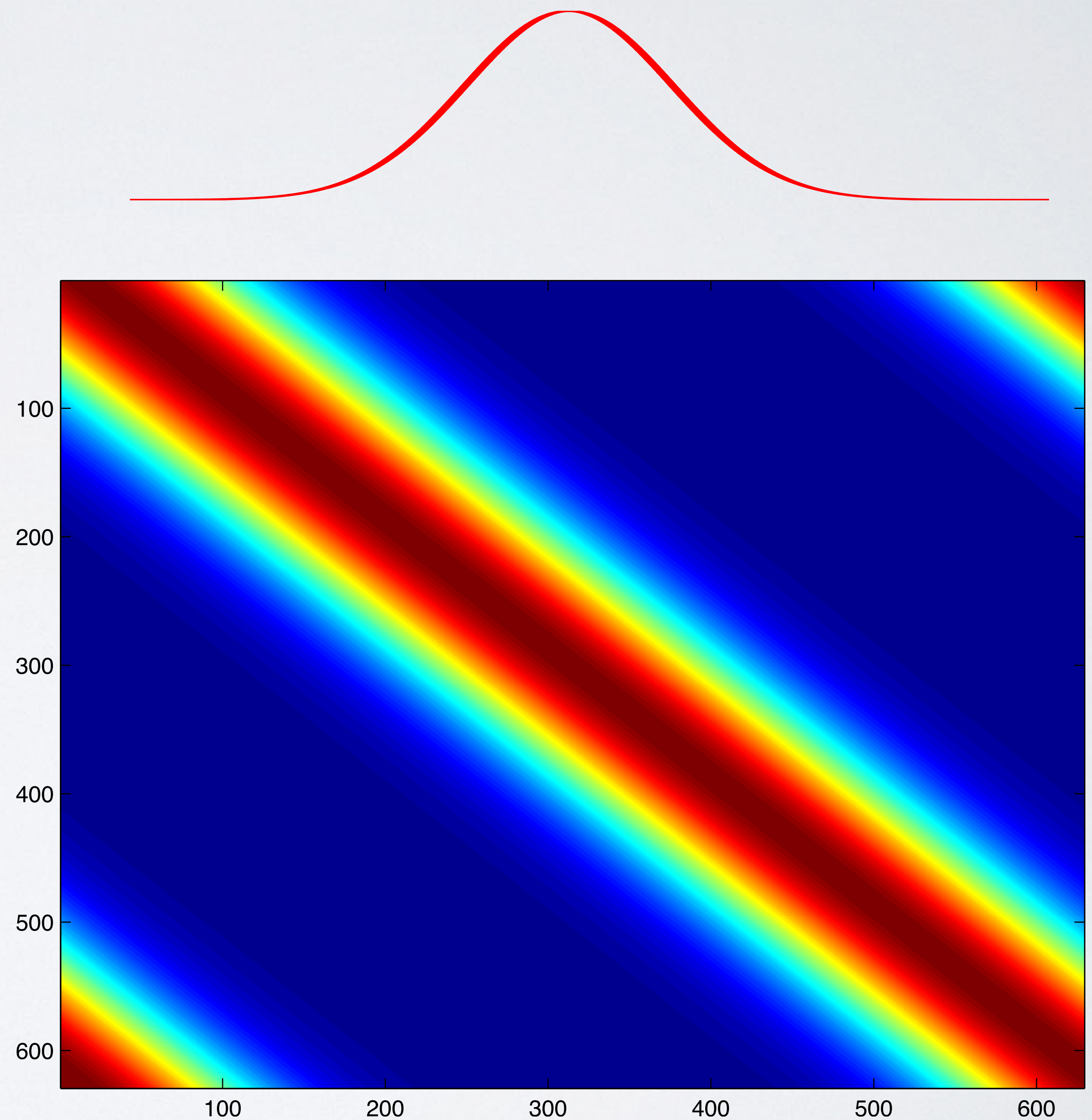


DOES THIS EVER HAPPEN IN REAL LIFE?

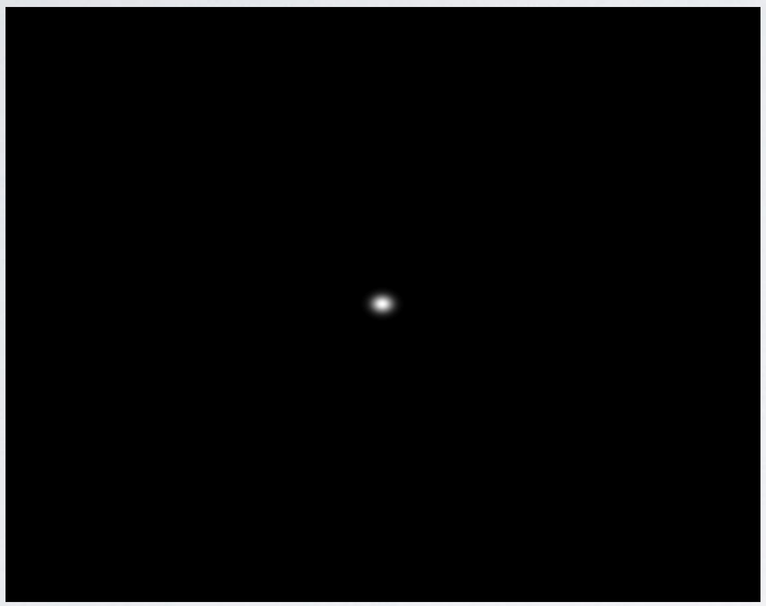


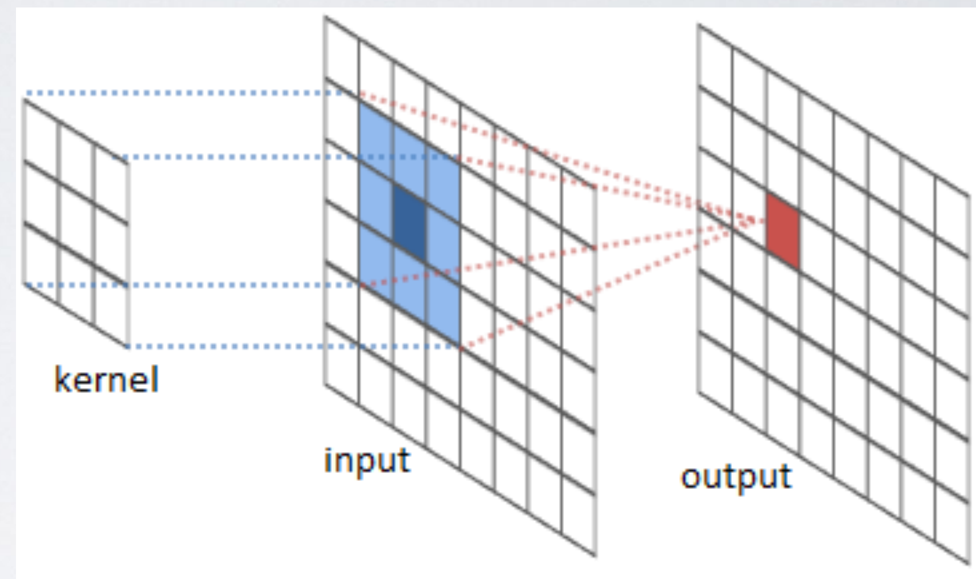
CONVOLUTION MATRIX

Condition number: 3,500,000



CONVOLUTION: 2D

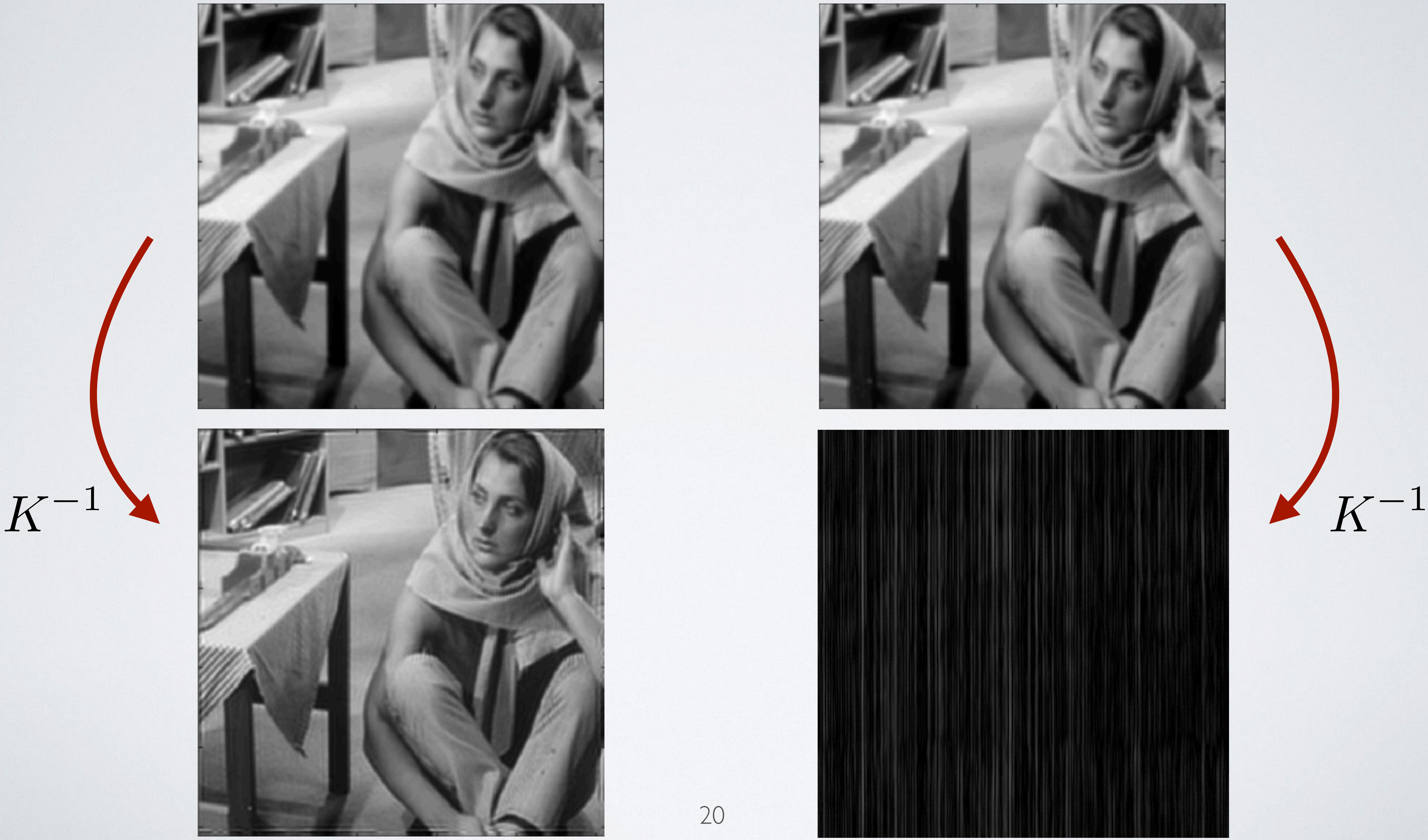




K

DEBLURRING

Relative difference 0.0092



Why would you want a penalty?

Under-determined systems.

UNDER-DETERMINED SYSTEMS

- Another problem: What if matrix isn't even full-rank?

$$A \in \mathbb{R}^{M \times N} \qquad M < N$$

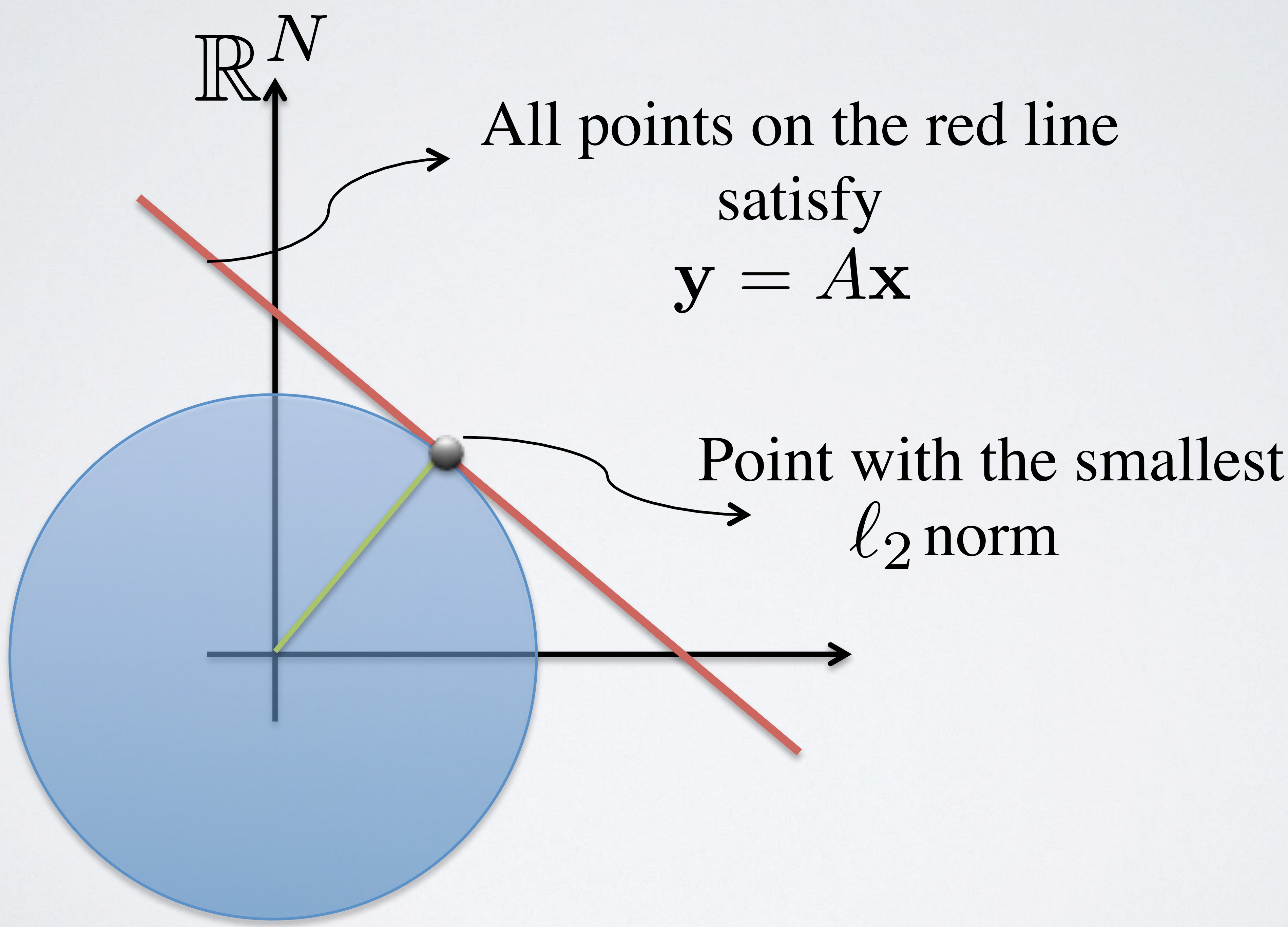
$$b = Ax + \eta$$

- If the error is bounded ($\|\eta\| \leq \epsilon$) solve

$$\text{minimize} \quad \|x\| \ \text{ subject to } \ \|Ax - b\| \leq \epsilon$$

"Occam's razor"

GEOMETRIC INTERPRETATION



RIDGE REGRESSION

- If the error is bounded ($\|\eta\| \leq \epsilon$) solve

$$b = Ax + \eta$$

$$\text{minimize} \quad \|x\| \;\; \text{subject to} \;\; \|Ax - b\| \leq \epsilon$$

- This is equivalent to

$$\text{minimize} \quad \lambda\|x\|^2 + \|b - Ax\|^2$$

for some value of λ

RIDGE REGRESSION

$$\text{minimize} \quad \lambda\|x\|^2 + \|b - Ax\|^2$$

Closed form solution!

$$(A^T A + \lambda I)^{-1} A^T b$$

What does this do to condition number?

RIDGE REGRESSION

$$\text{minimize} \quad \lambda\|x\|^2 + \|b - Ax\|^2$$

Closed form solution!

$$(A^T A + \lambda I)^{-1} A^T b$$

New condition number

$$\frac{\sigma_{max}^2 + \lambda}{\sigma_{min}^2 + \lambda}$$

TIKHONOV REGULARIZATION

$$\text{minimize} \quad \lambda\|x\|^2 + \|b - Ax\|^2$$

- Has many names (ridge regression in stats)
- Advantage: Easier to solve new problem
- Improved condition number (less noise sensitivity)
- Parameter λ can be set:
  - Empirically (e.g. cross-validation)
  - Use noise bounds + theory (BIC, etc…)

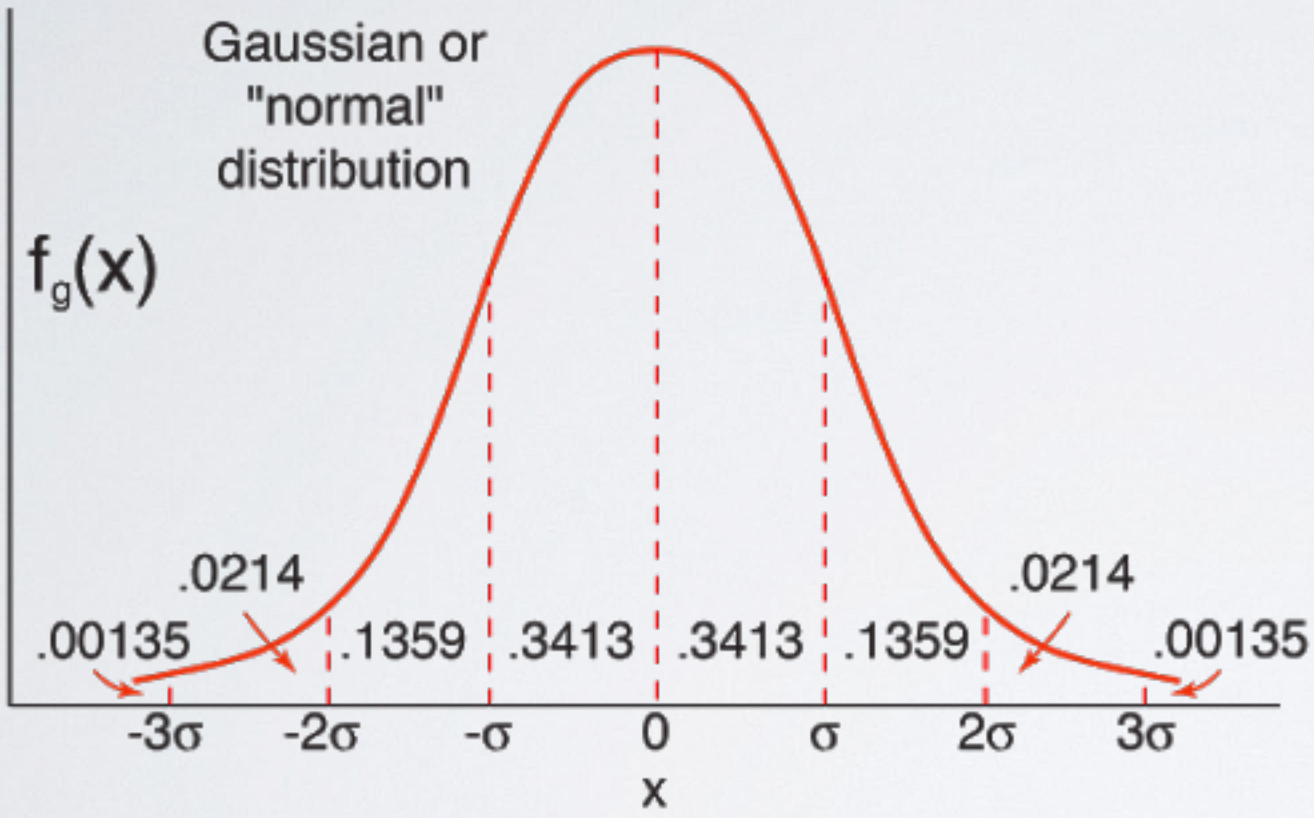
BAYESIAN INTERPRETATION

…How to cook up a Bayesian model…

- Model data formation: write distribution of data given parameters
- Observe data from random process
- Use Bayes rule: write distribution of parameters given data
- **Find "most likely" parameters given data**
- Optional: uncertainty quantification / confidence

GAUSSIAN

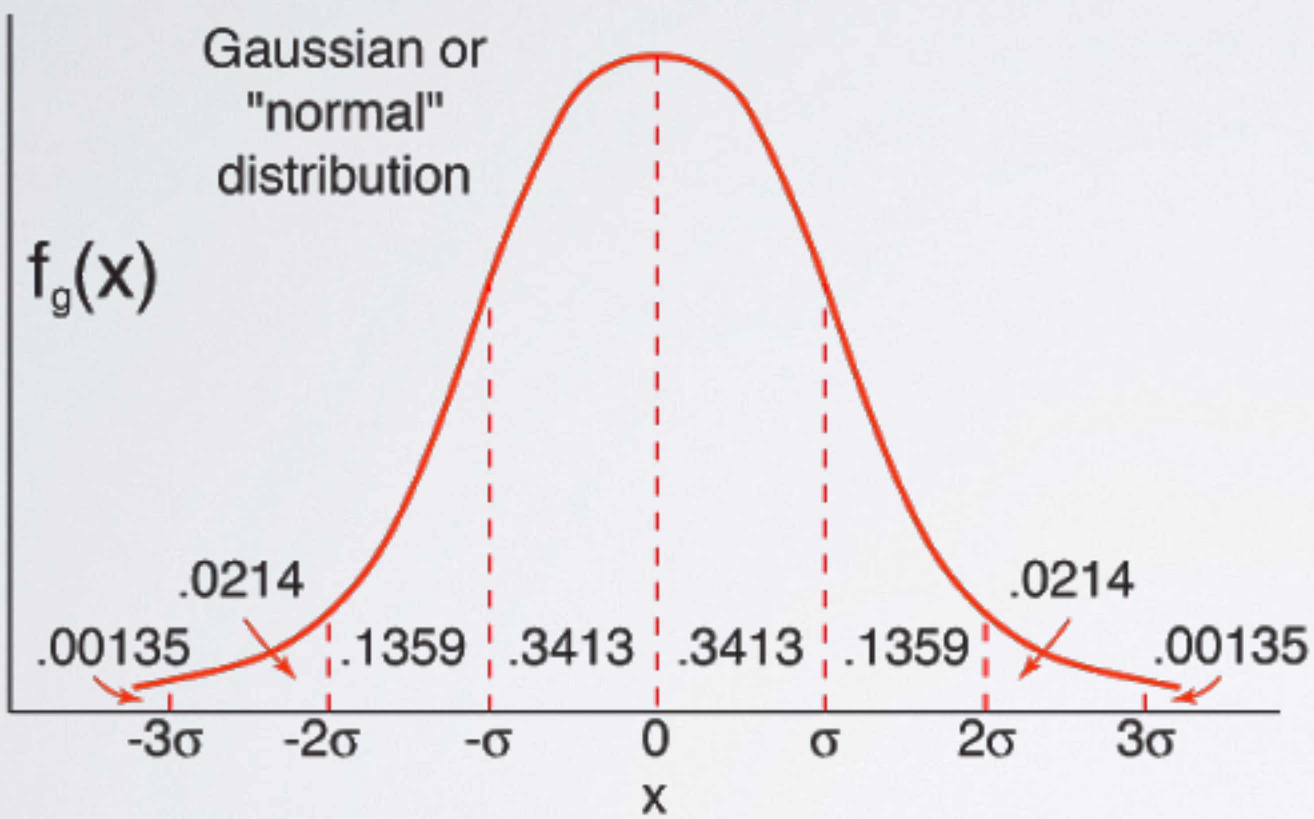
The bell curve



$$\sim e^{\frac{-x^2}{2\sigma^2}}$$

GAUSSIAN

The bell curve



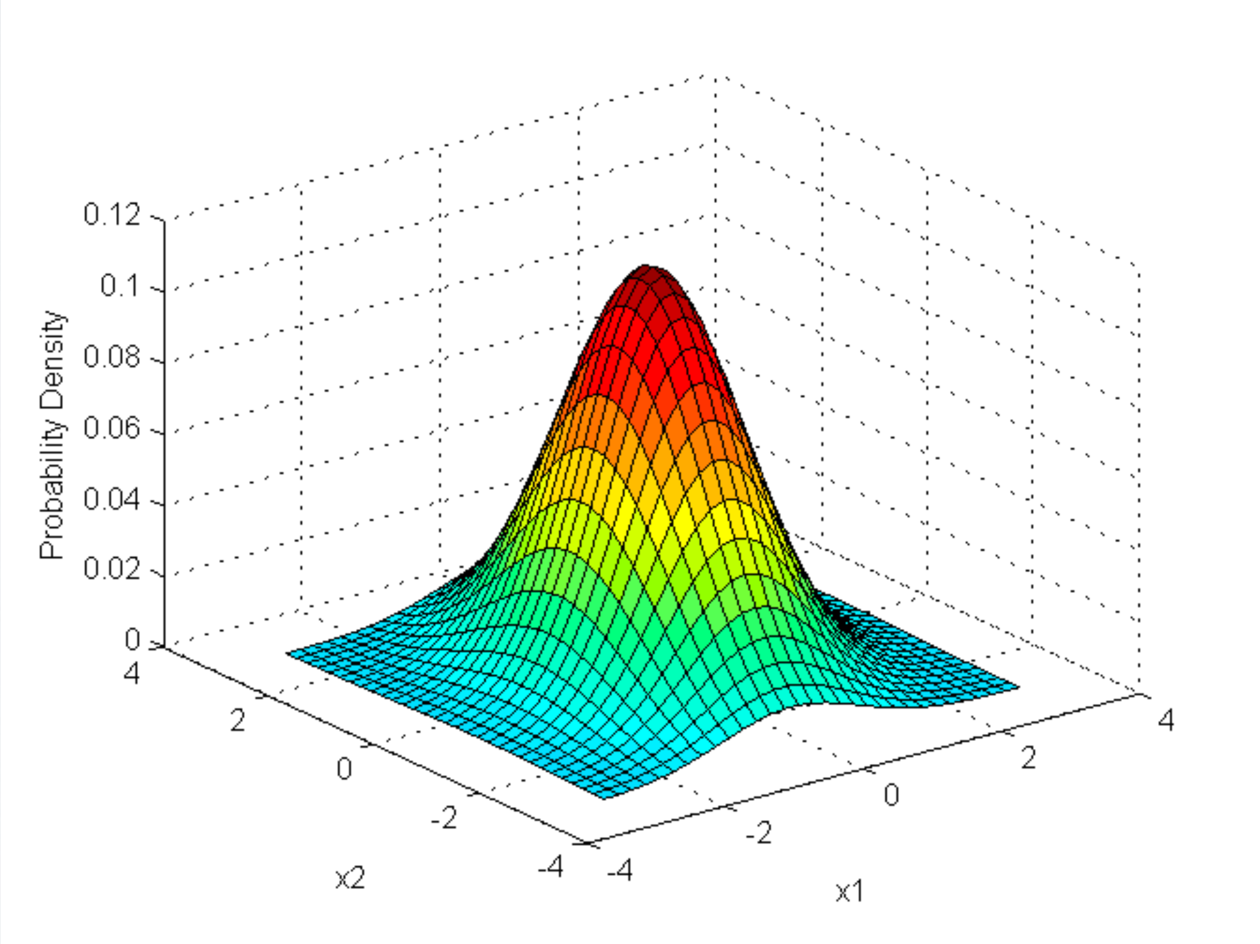
$$\sim e^{\frac{-x^2}{2\sigma^2}}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\sigma^2 = \mathbb{E}[x^2]$$

MULTIVARIATE GAUSSIAN

The bell curve



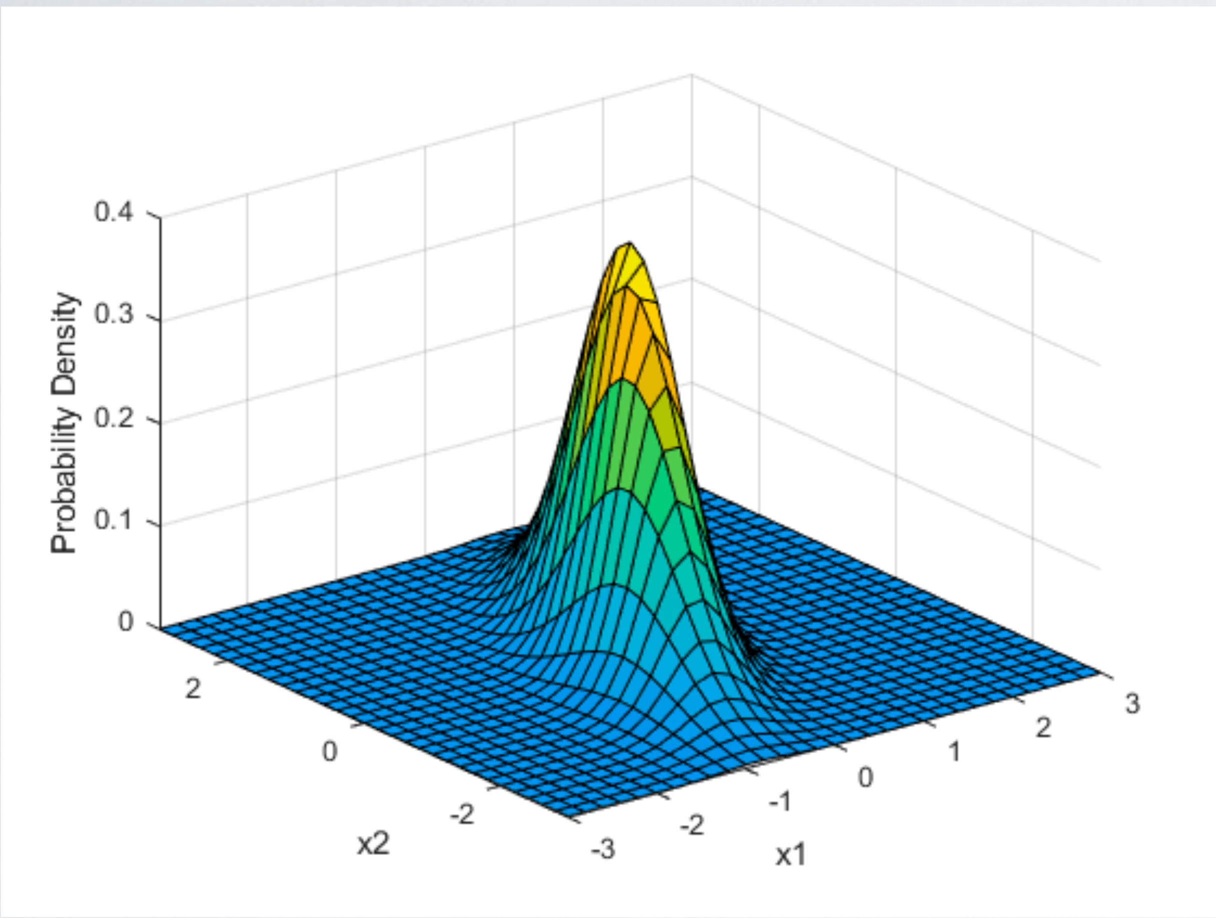
$$\sim \prod_i e^{-\frac{x_i^2}{2\sigma^2}} = e^{\frac{-1}{2\sigma^2}\sum_i x_i^2}$$

$$= e^{-\frac{x^t x}{2\sigma^2}}$$

$$= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x_i - \mu_i)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\frac{\|x-\mu\|^2}{2\sigma^2}}$$

MULTIVARIATE GAUSSIAN

The bell curve



$$\Sigma = \mathbb{E}[xx^t]$$

Covariance matrix

$$e^{-\frac{x^t x}{2\sigma^2}} \to e^{-\frac{1}{2}x^t \Sigma^{-1} x}$$

$$= \frac{1}{(2\pi|\Sigma|)^{n/2}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

BAYESIAN INTERPRETATION

$$\text{minimize} \quad \lambda\|x\|^2 + \|b - Ax\|^2$$

- Assumptions:
  - Prior: we know expected signal power $\mathbb{E}\{x_i^2\} = E^2$
  - Linear measurement model $b = Ax + \eta$
  - Noise is i.i.d. Gaussian: $\eta = N(0, \sigma)$
- MAP (maximum a-posteriori) estimate:

$$\underset{x}{\text{maximize}} \quad p(x|b)$$

BAYESIAN INTERPRETATION

$$\underset{x}{\text{maximize}} \quad p(x|b)$$

Bayes' Rule

$$p(X|Y) \propto p(Y|X)p(X)$$

$$\underset{x}{\text{maximize}} \quad p(x|b) \propto p(b|x)p(x)$$

$$b \sim N(Ax, \sigma)$$

model

$$x \sim N(0, E)$$

prior

BAYESIAN INTERPRETATION

$$\underset{x}{\text{maximize}} \quad p(x|b) = p(b|x)p(x)$$

probability of data

$$p(b|x) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(b_i-(Ax)_i)^2}$$

$$= (2\pi\sigma^2)^{-m/2} e^{-\frac{1}{2\sigma^2}\|b-Ax\|^2}$$

prior

$$p(x) = (2\pi E^2)^{-n/2} e^{-\frac{1}{2E^2}\|x\|^2}$$

$$\text{maximize} \quad \exp\left(\frac{-\|b-Ax\|^2}{2\sigma^2}\right)\exp\left(\frac{-\|x\|^2}{2E^2}\right)$$

NEGATIVE LOG-LIKELIHOOD

$$\underset{x}{\text{maximize}} \quad p(x|b) = p(b|x)p(x)$$

$$\text{maximize} \quad \exp\left(\frac{-\|b - Ax\|^2}{2\sigma^2}\right)\exp\left(\frac{-\|x\|^2}{2E^2}\right)$$

$$\text{maximize} \quad -\frac{\|b - Ax\|^2}{2\sigma^2} - \frac{\|x\|^2}{2E^2}$$

NLL

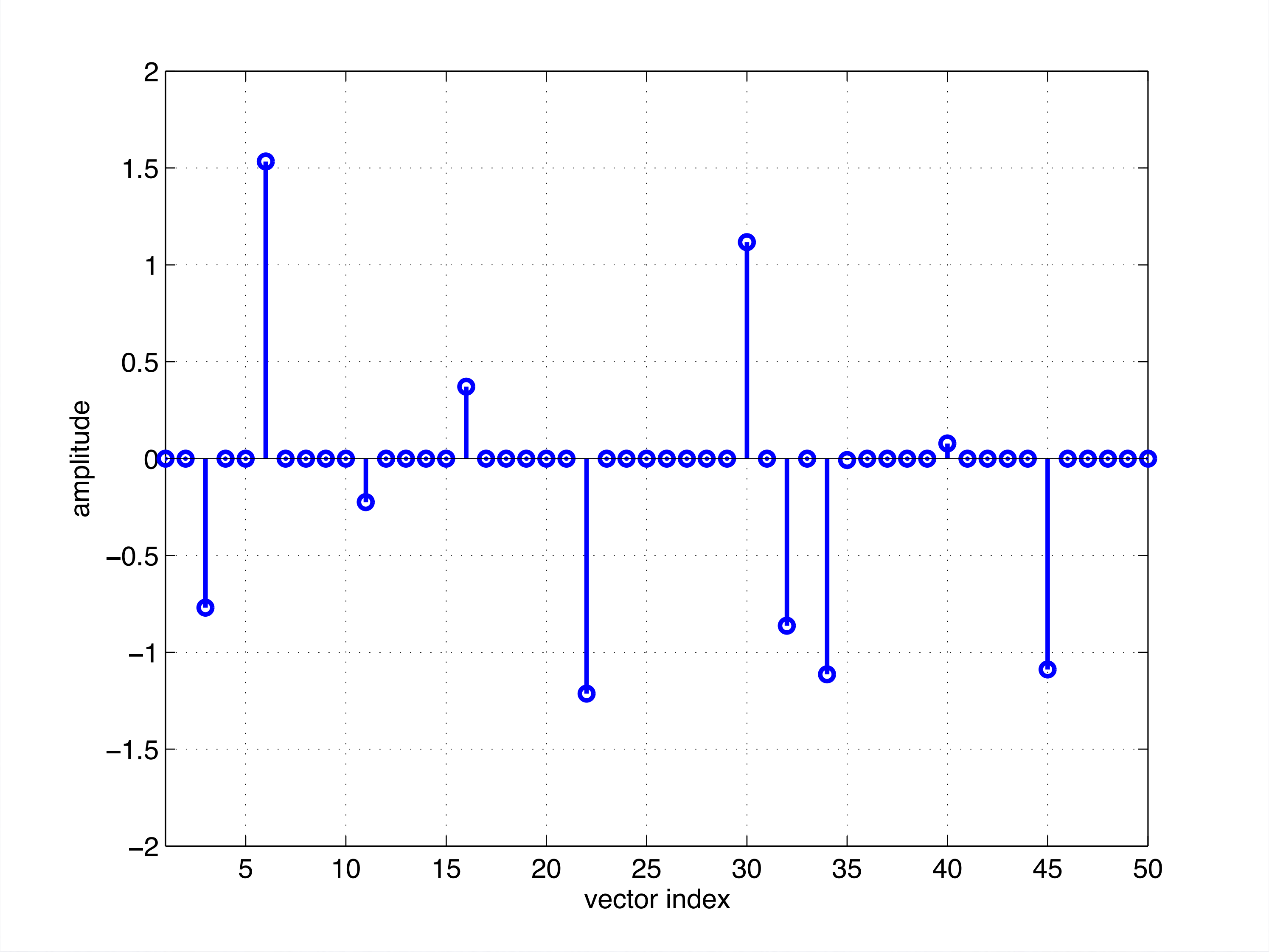
$$\text{minimize} \quad \frac{\sigma^2}{E^2}\|x\|^2 + \|b - Ax\|^2$$

SPARSE PRIORS

- Priors add information to the problem
- Ridge/Tikhonov priors require a lot of assumptions
- Prior only good when assumptions true!
- A very general prior: sparsity

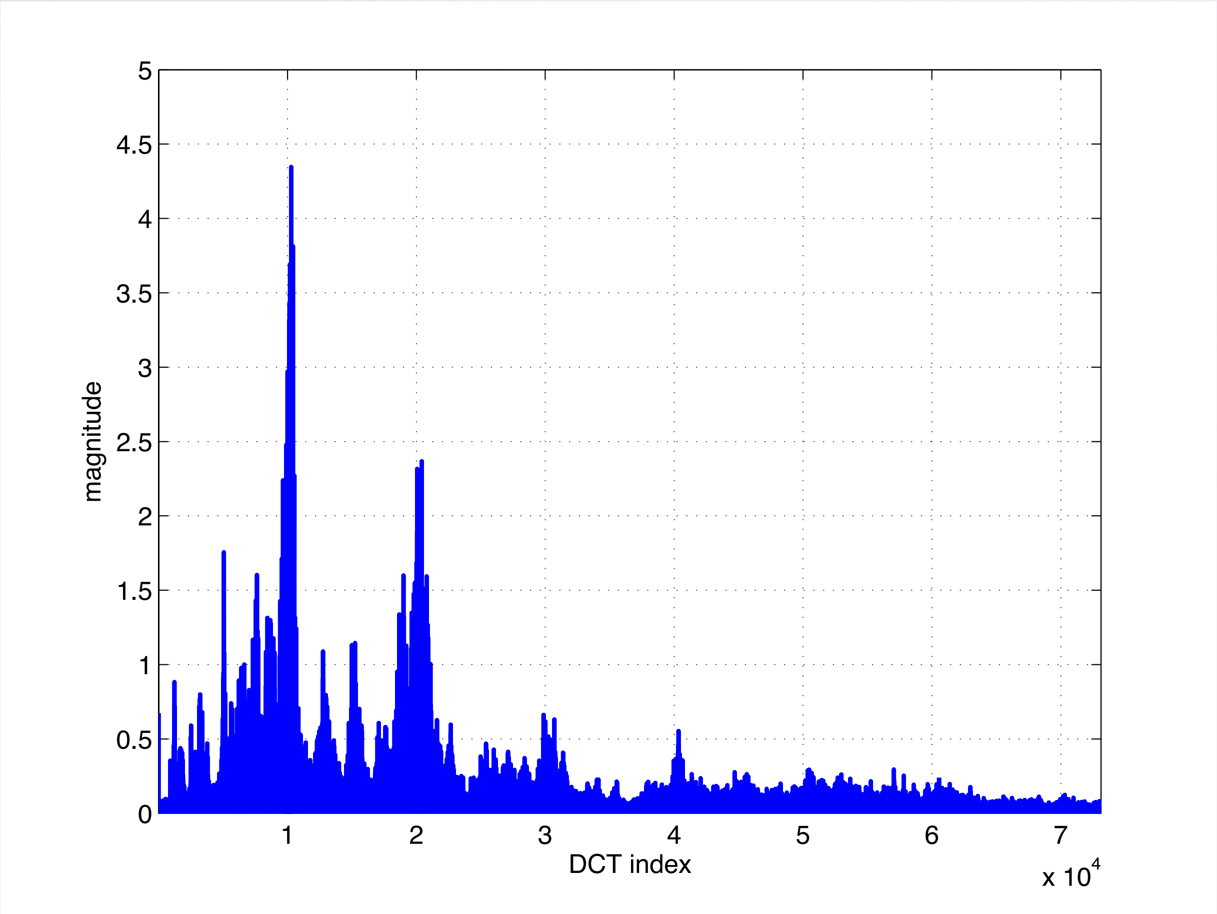
WHAT IS SPARSITY?

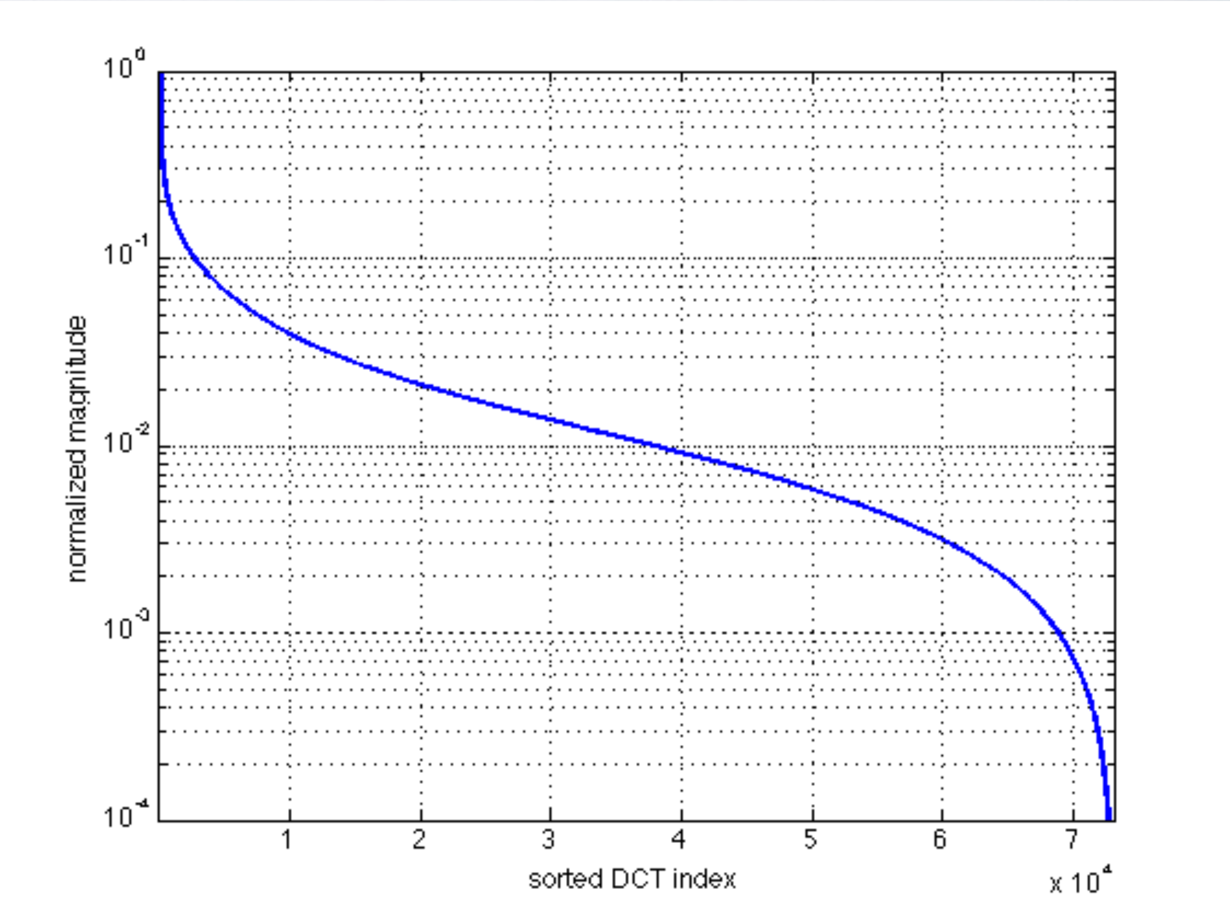
- Signal has very few non-zeros: small ℓ_0 norm



OTHER NOTIONS OF SPARSITY

- "Low density" signals - rapid decay of coefficients



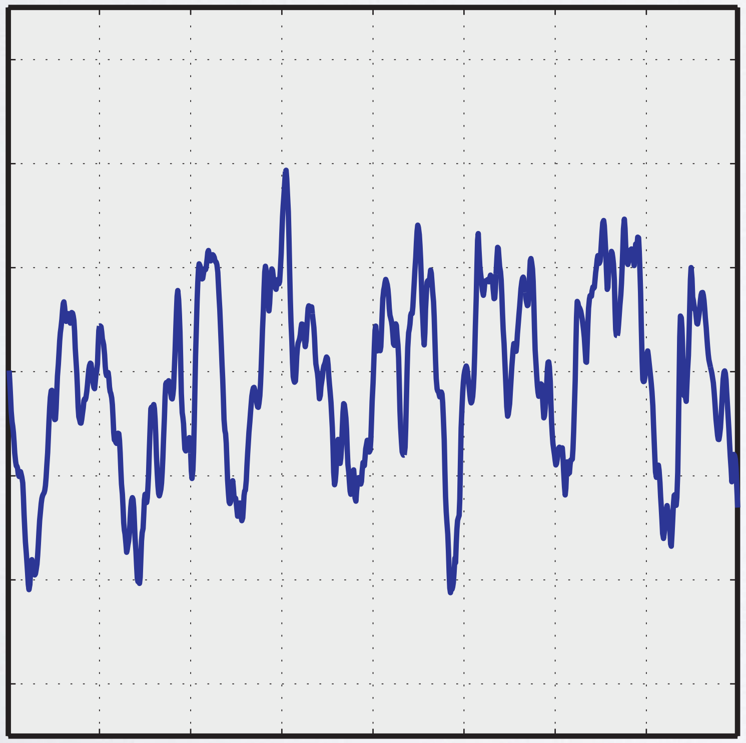


- Fast decay: = Small Weak ℓ_p norm

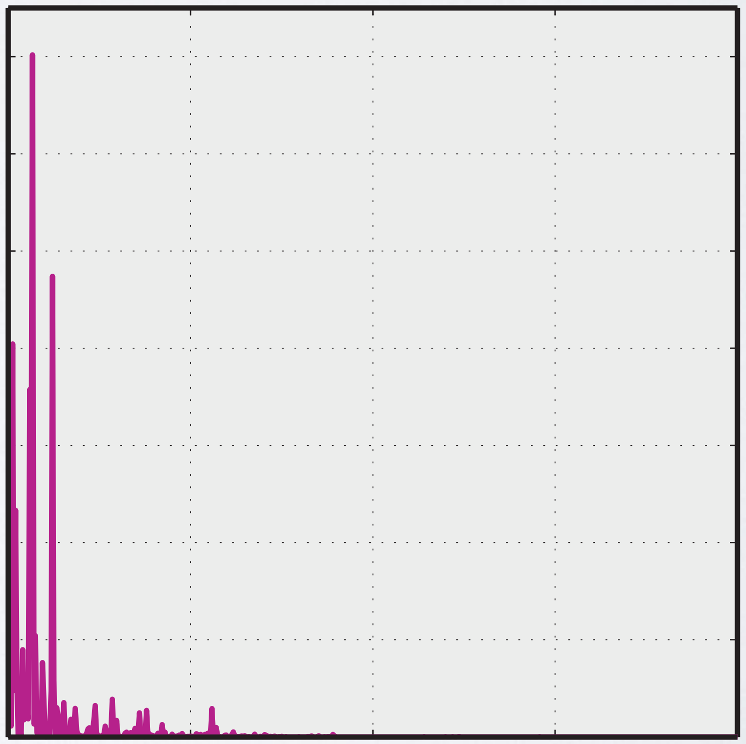
DENSE SIGNAL / SPARSE REPRESENTATION: AUDIO

- Sounds produced by vibrating objects
- Energy is concentrated at **resonance frequencies** of object
- Defined by eigenvalues of the Laplacian of vibrating surface

Audio Signal

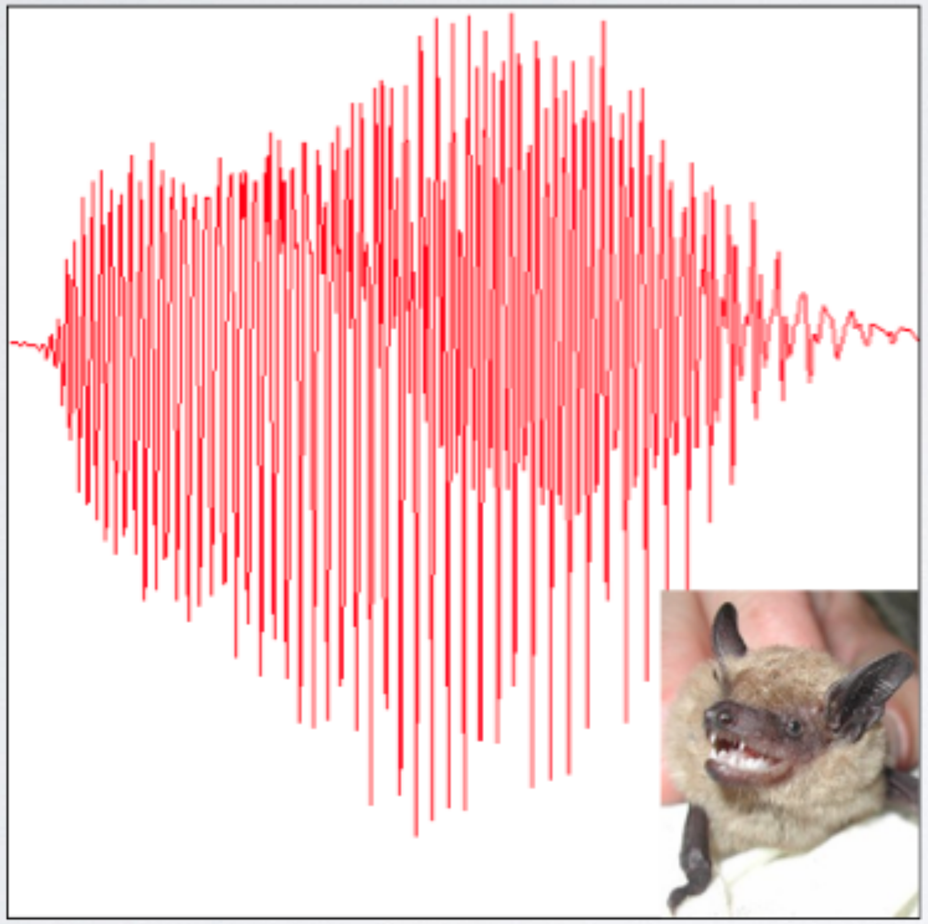


DCT

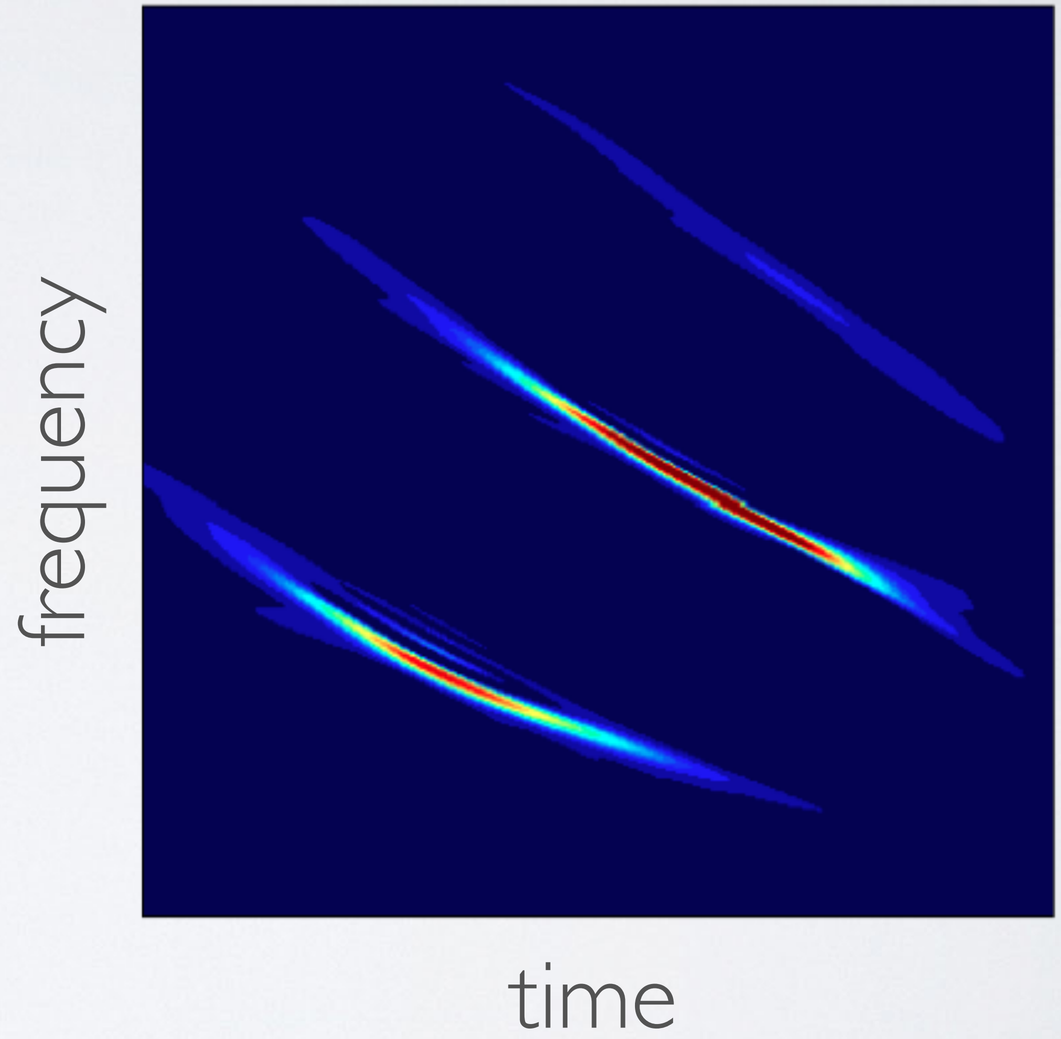


DENSE SIGNAL / SPARSE REPRESENTATION: AUDIO

Echolocation chirp: brown bat



Gabor transform (STFT)



- Bat hears convolution of signal with environment
- Chirps: generate well-conditioned convolution matrices

source: Christoph Studer

DENSE SIGNAL / SPARSE REPRESENTATION: IMAGES

- Approximately Piecewise constant
- High correlations between adjacent pixels within objects
- High variation across edges

Wavelet transform of natural image

Credit: Mark Davenport

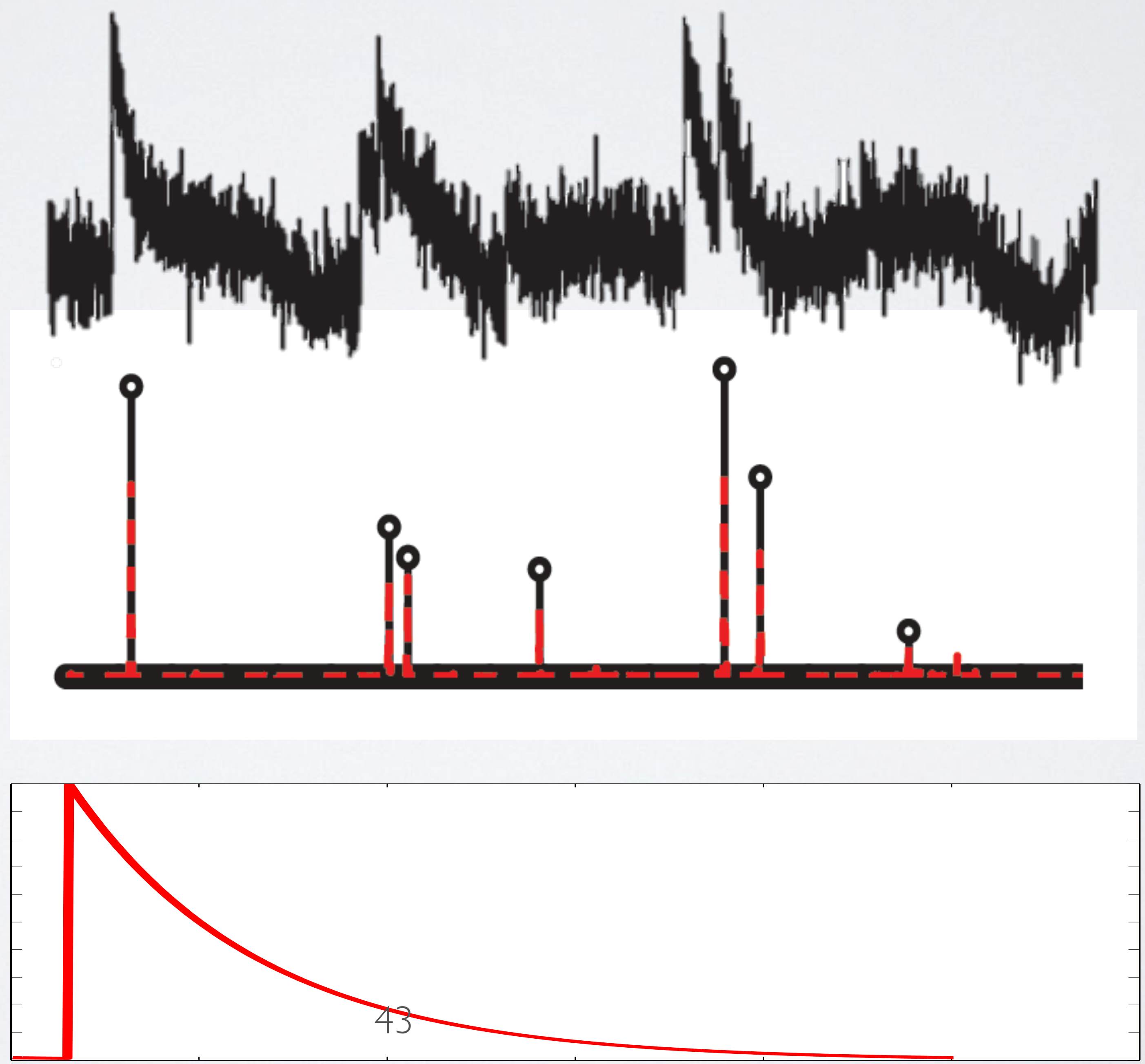
NEURAL EVENTS

- Neural potentials: convolution of spike train with kernel

Real recording

Spiking events

Kernel



MACHINE LEARNING: OVERFITTING

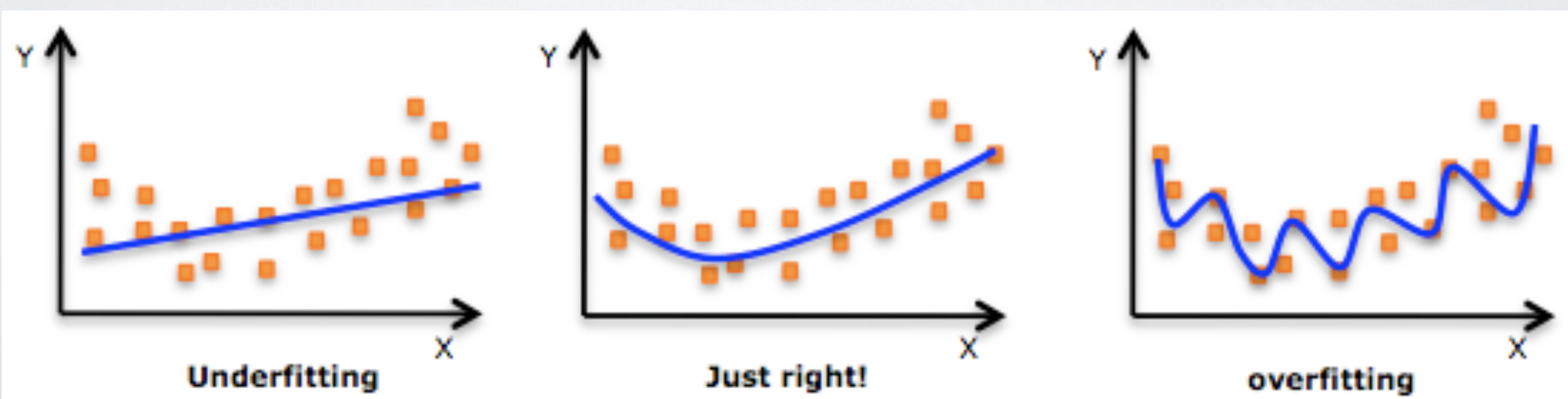
$$Ax = b$$

Features/Data → A

Model Parameters → x

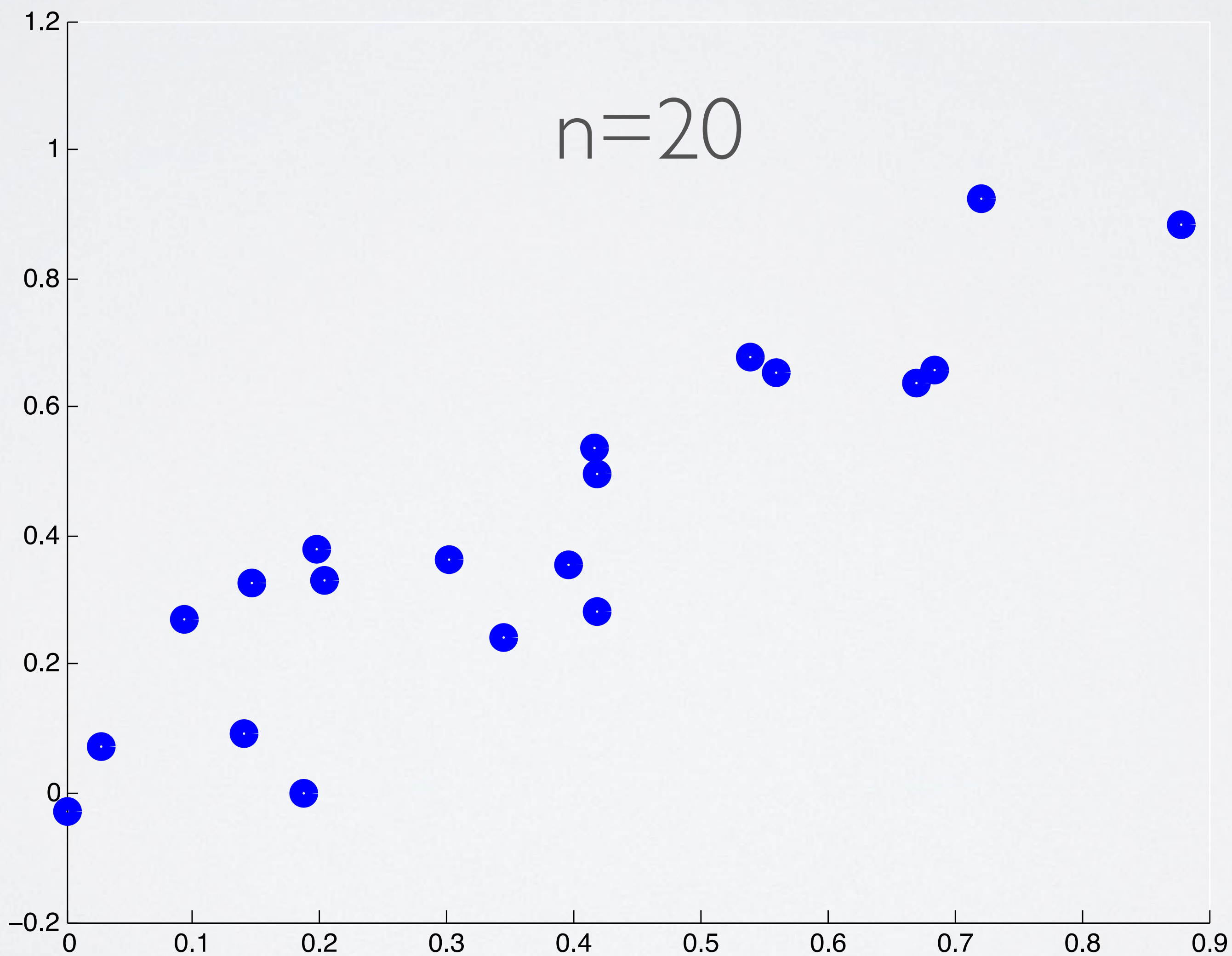
Labels → b

- Happens when you can design the measurement matrix
- More features = better fit



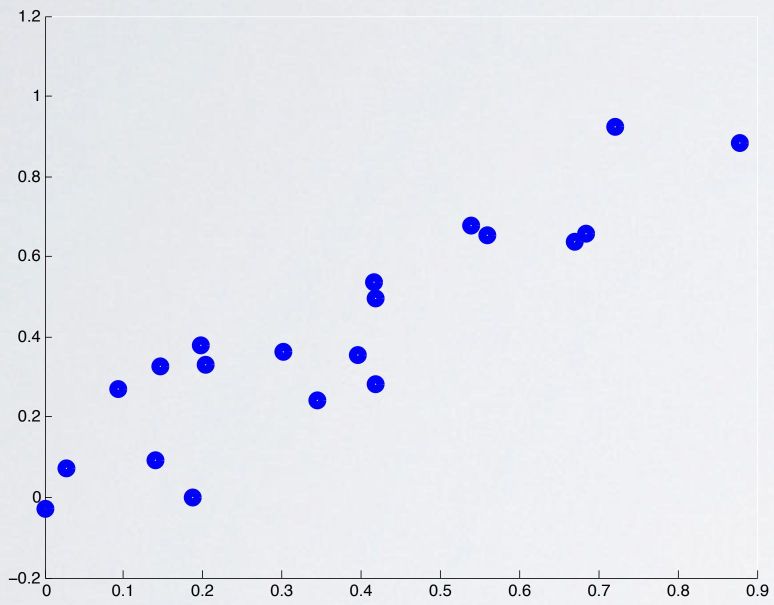
EXAMPLE: POLYNOMIAL FITTING

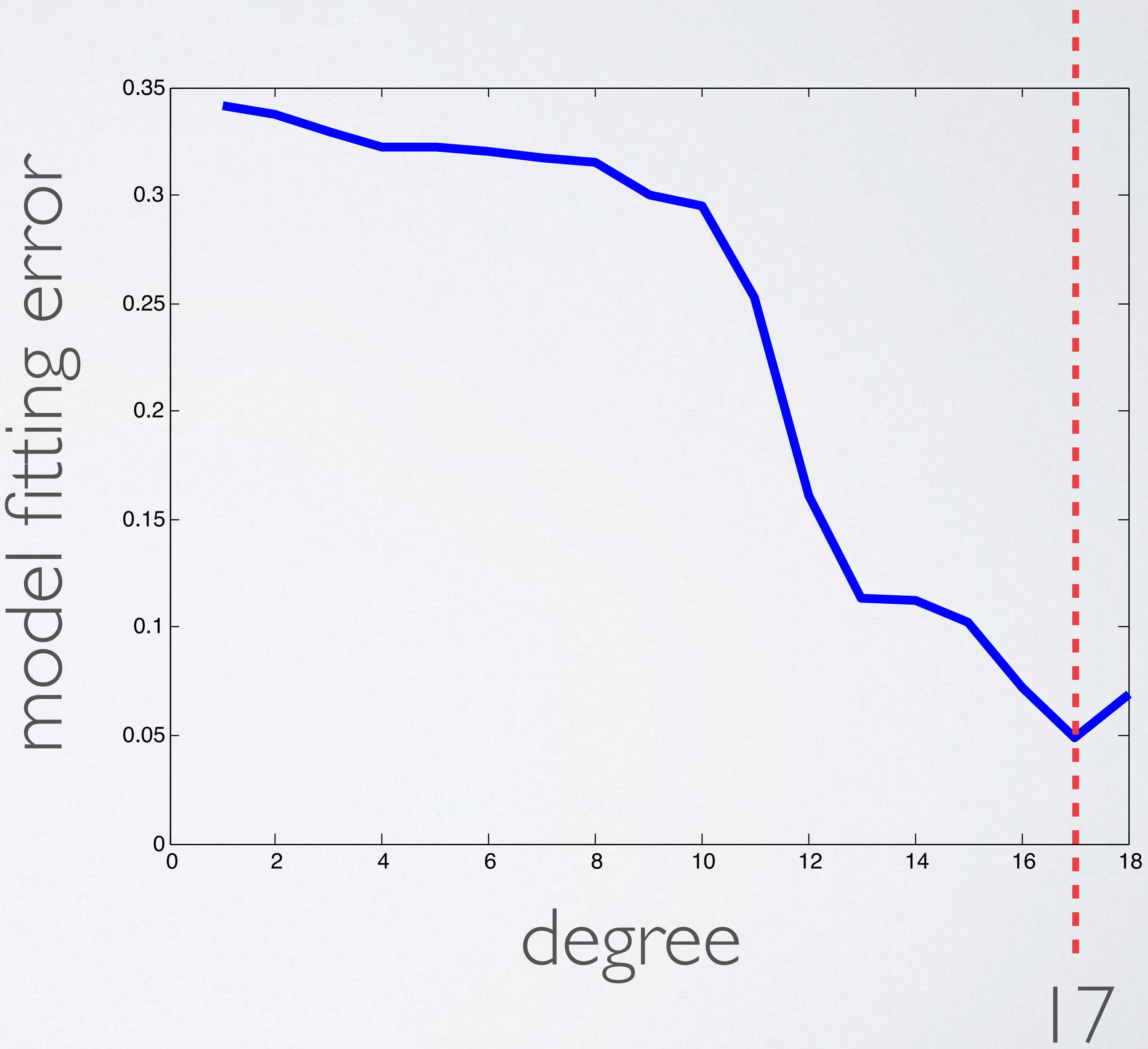
Noisy data drawn from polynomial
what degree is best?



EXAMPLE: POLYNOMIAL FITTING

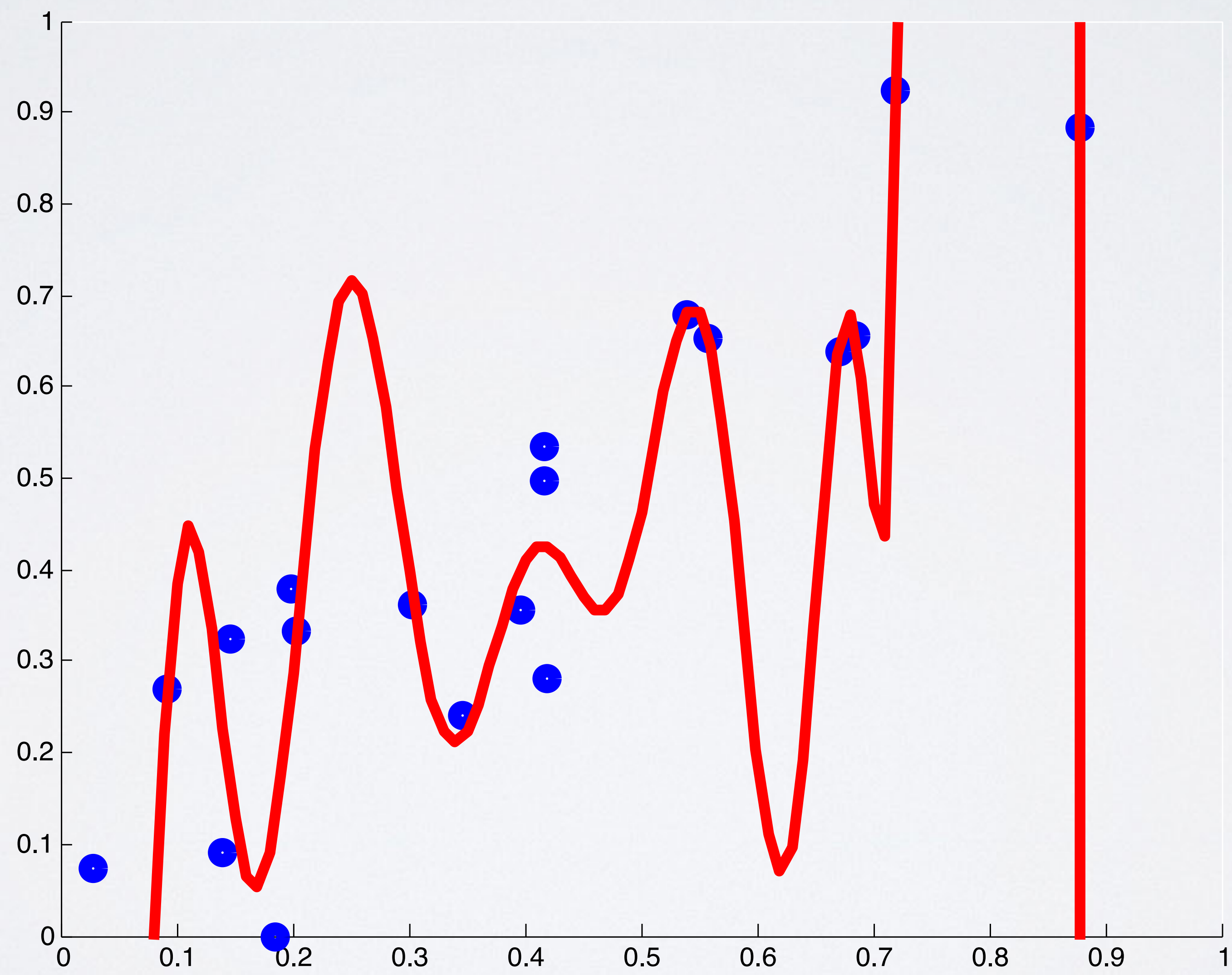
n=20





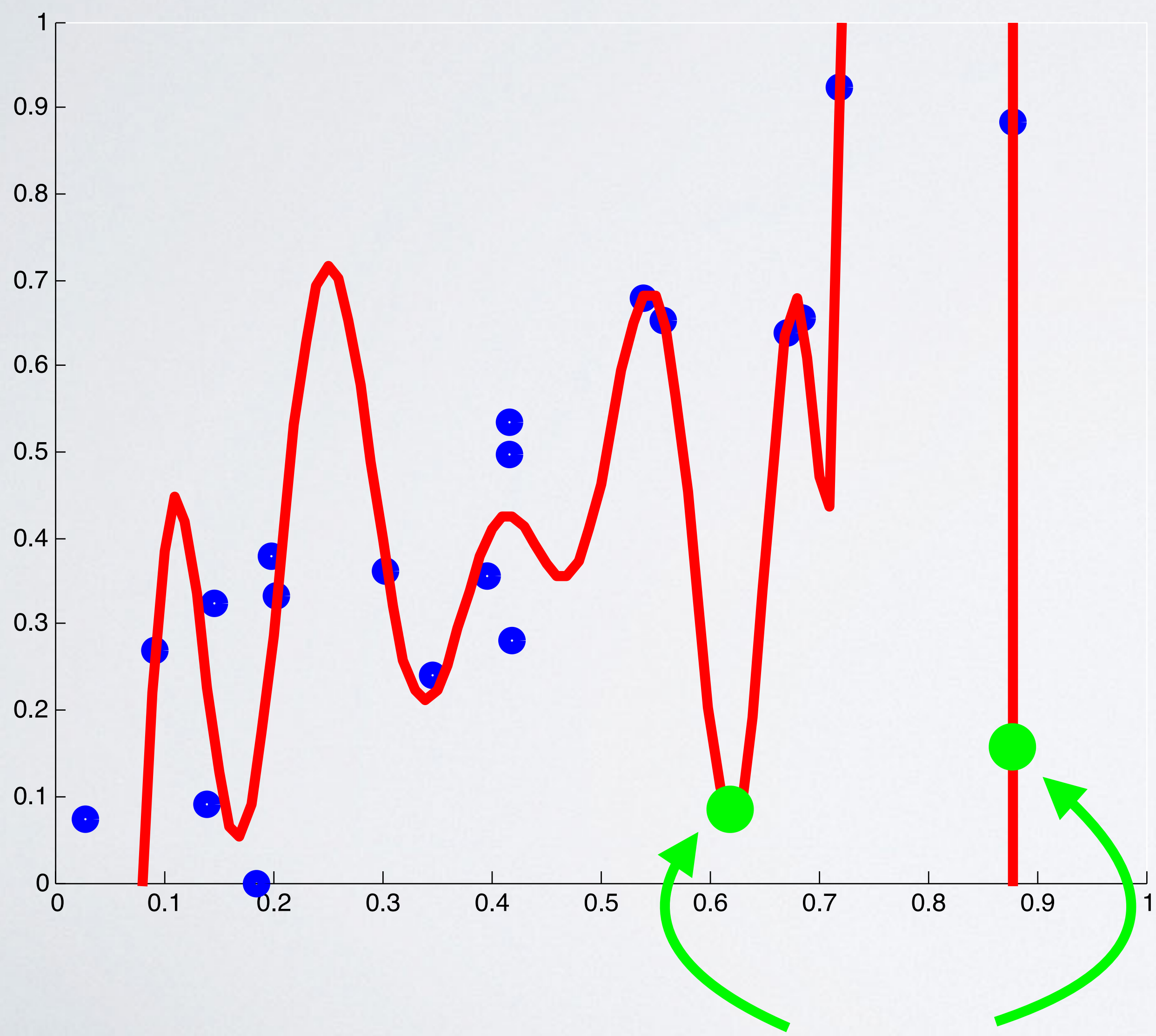
EXAMPLE: POLYNOMIAL FITTING

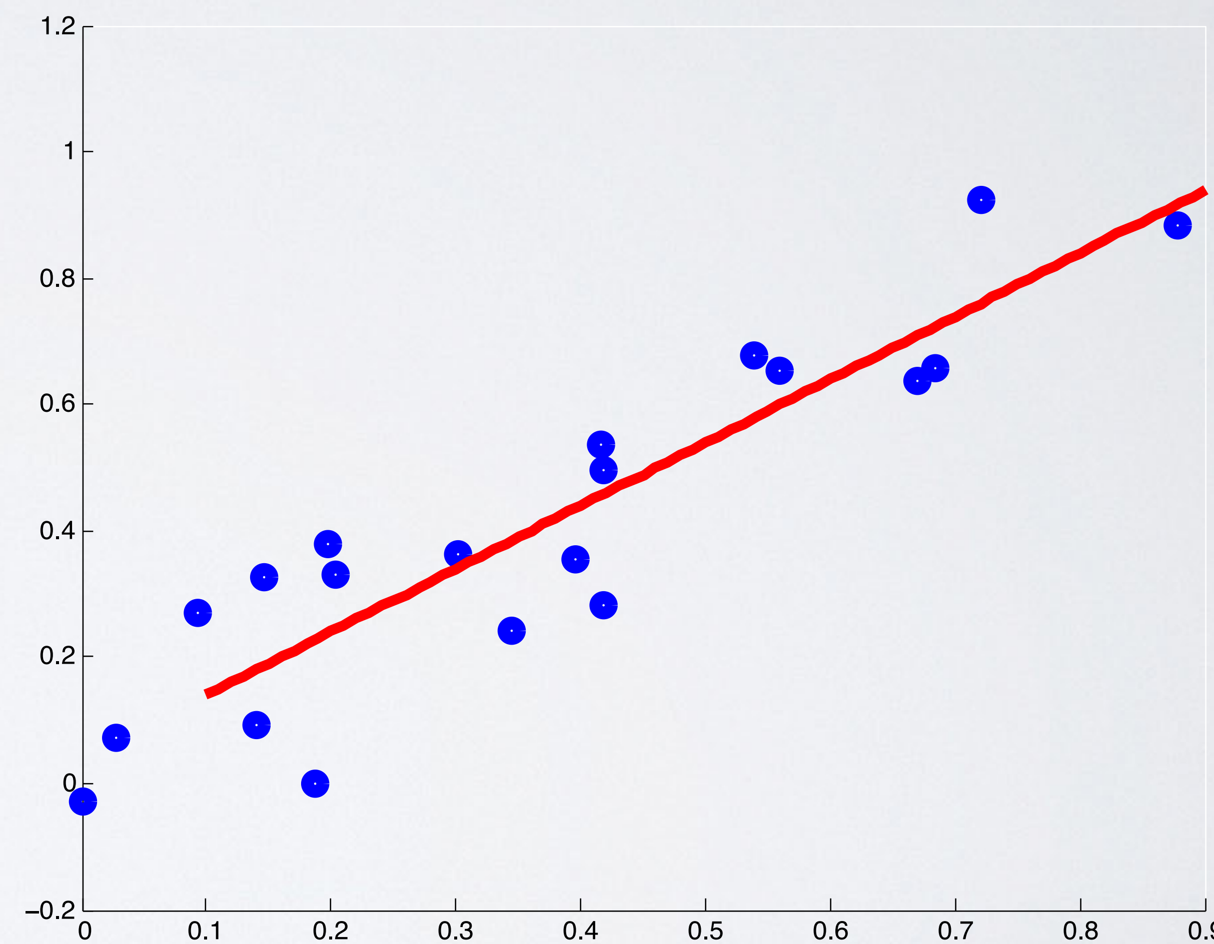
degree = 17



EXAMPLE: POLYNOMIAL FITTING

Fit **SIGNAL** not **NOISE**!

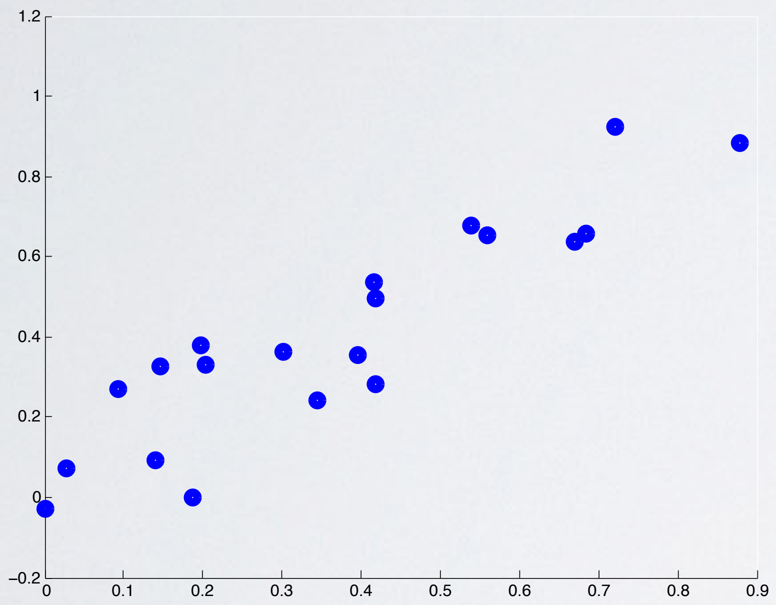


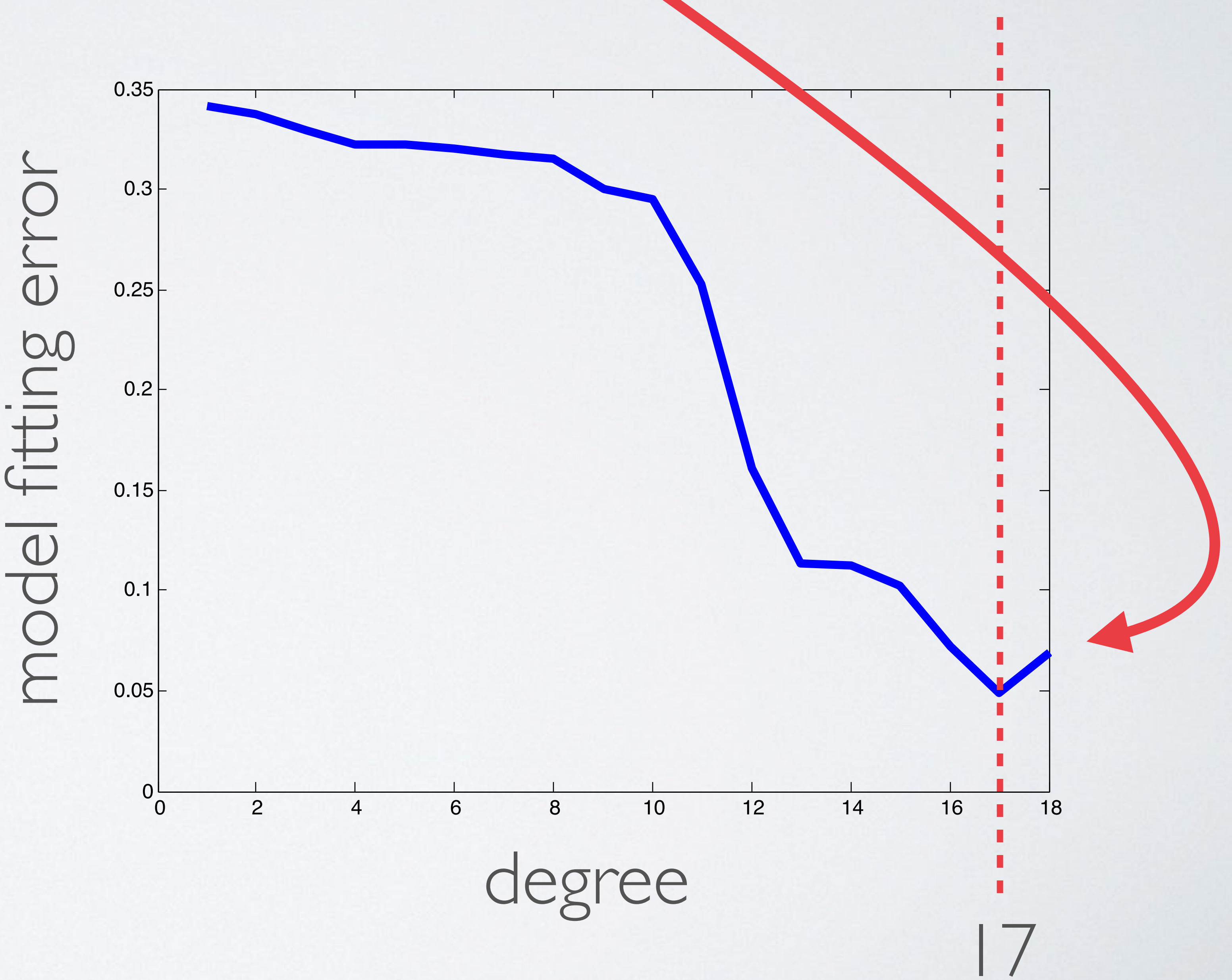


bad “out of sample” error

WHY DID THIS HAPPEN?

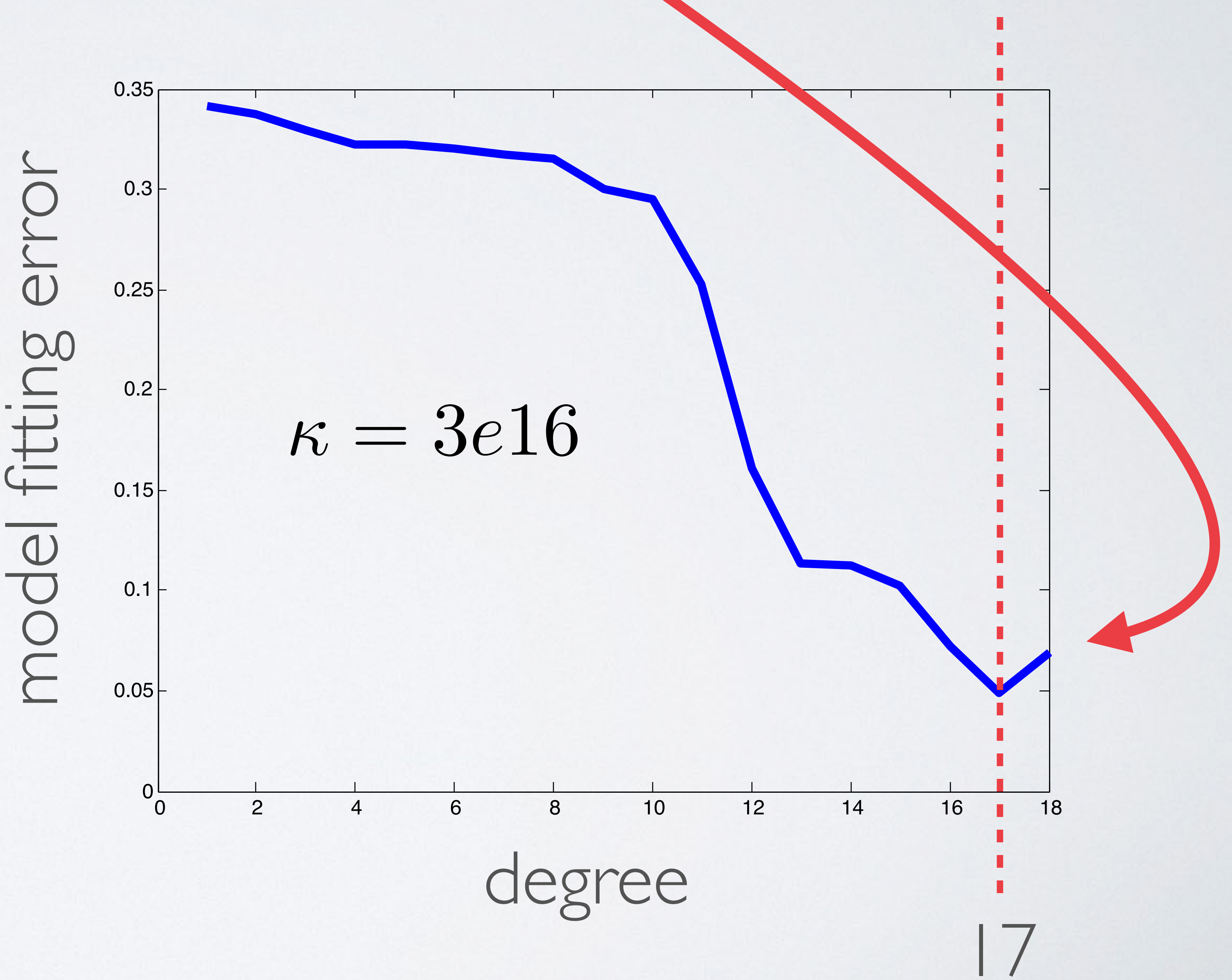
n=20





WHY DID THIS HAPPEN?

We **know** we're overfitting just from condition number. **Why**?



BIAS-VARIANCE TRADEOFF

Fitting a model with random data: $f(x; D)$

Expected model: $\bar{f}(x) = \mathbb{E}_D f(x; D)$

$$\underbrace{\mathbb{E}_{D,x,y}[y - f(x; D)]^2}_{\text{Test error}} = \underbrace{\mathbb{E}_{x,y}[y - \bar{f}(x)]^2}_{\text{Bias}} + \underbrace{\mathbb{E}_{D,x,y}[f(x, D) - \bar{f}(x)]^2}_{\text{Variance}} + \sigma^2$$

Test error

Bias

Variance

σ^2: Irreducible Error

BIAS-VARIANCE TRADEOFF

Example: linear estimation with mean-zero noise

$$Ax = b + \eta$$

Unregularized solution

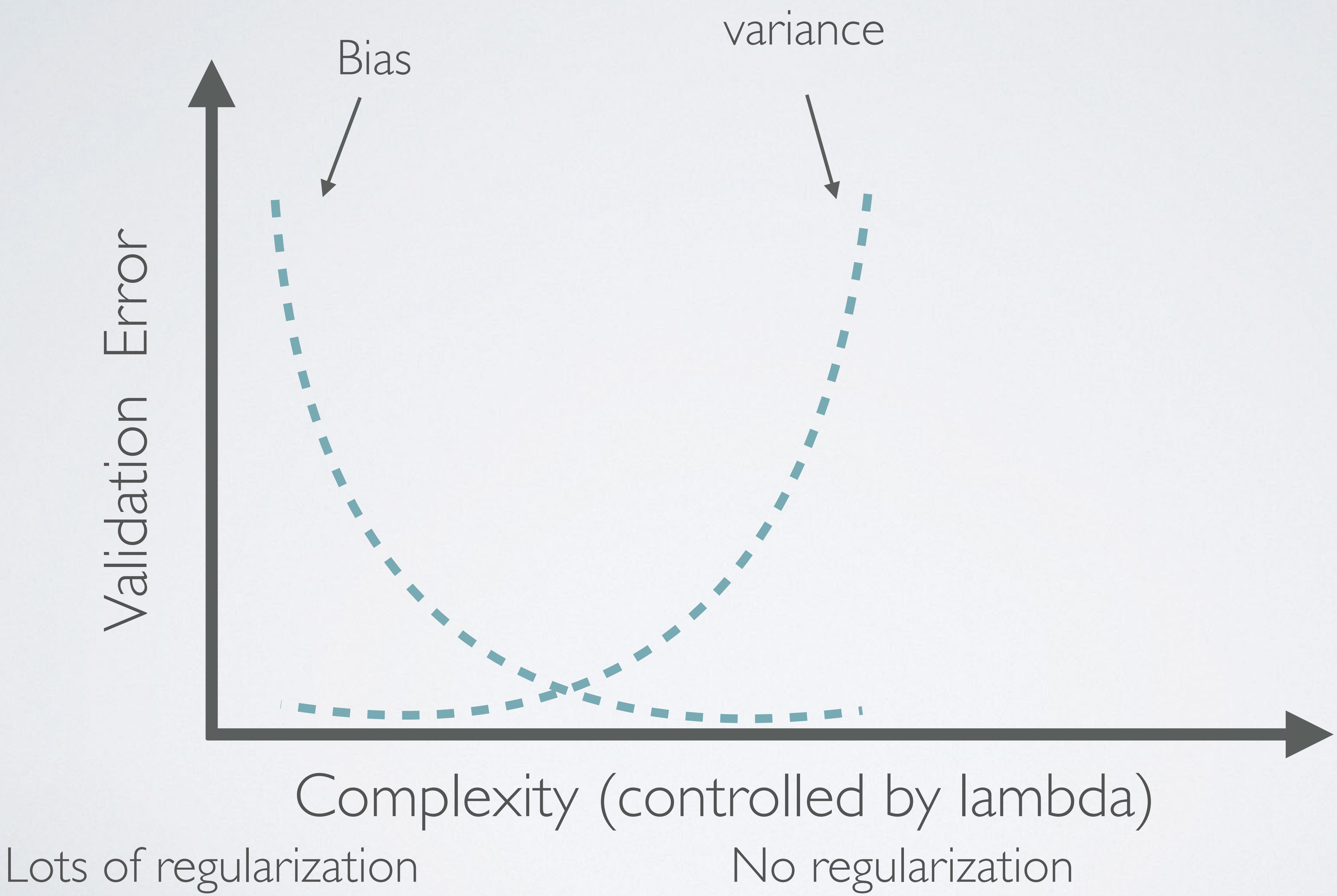
$$x = A^{-1}(b + \eta)$$

Bias

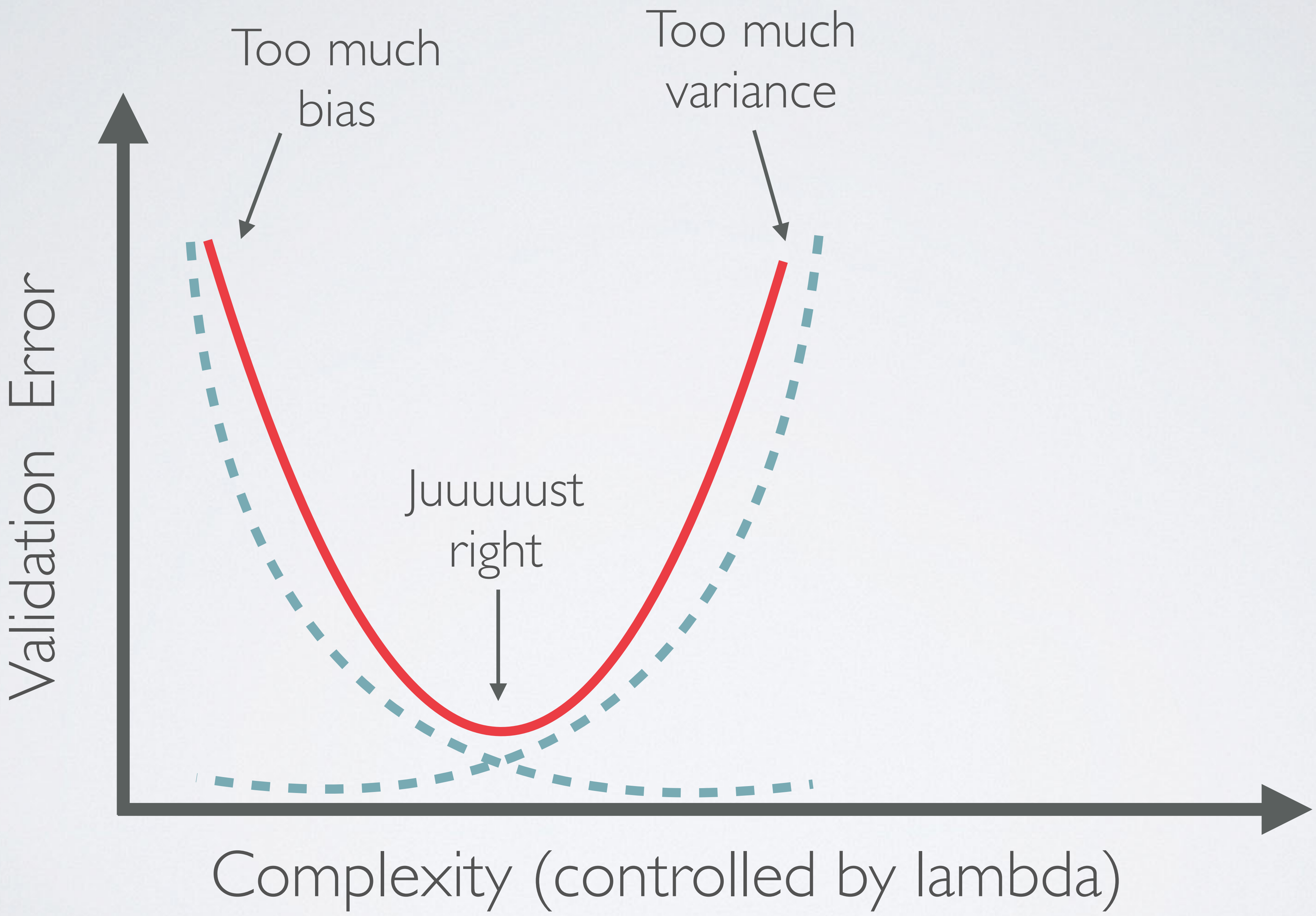
$$A^{-1}(b) - \mathbb{E}_\eta[A^{-1}(b + \eta)] = A^{-1}(b) - A^{-1}(b) + \mathbb{E}_\eta[\eta] = 0$$

Is this a good estimator? It has no bias?

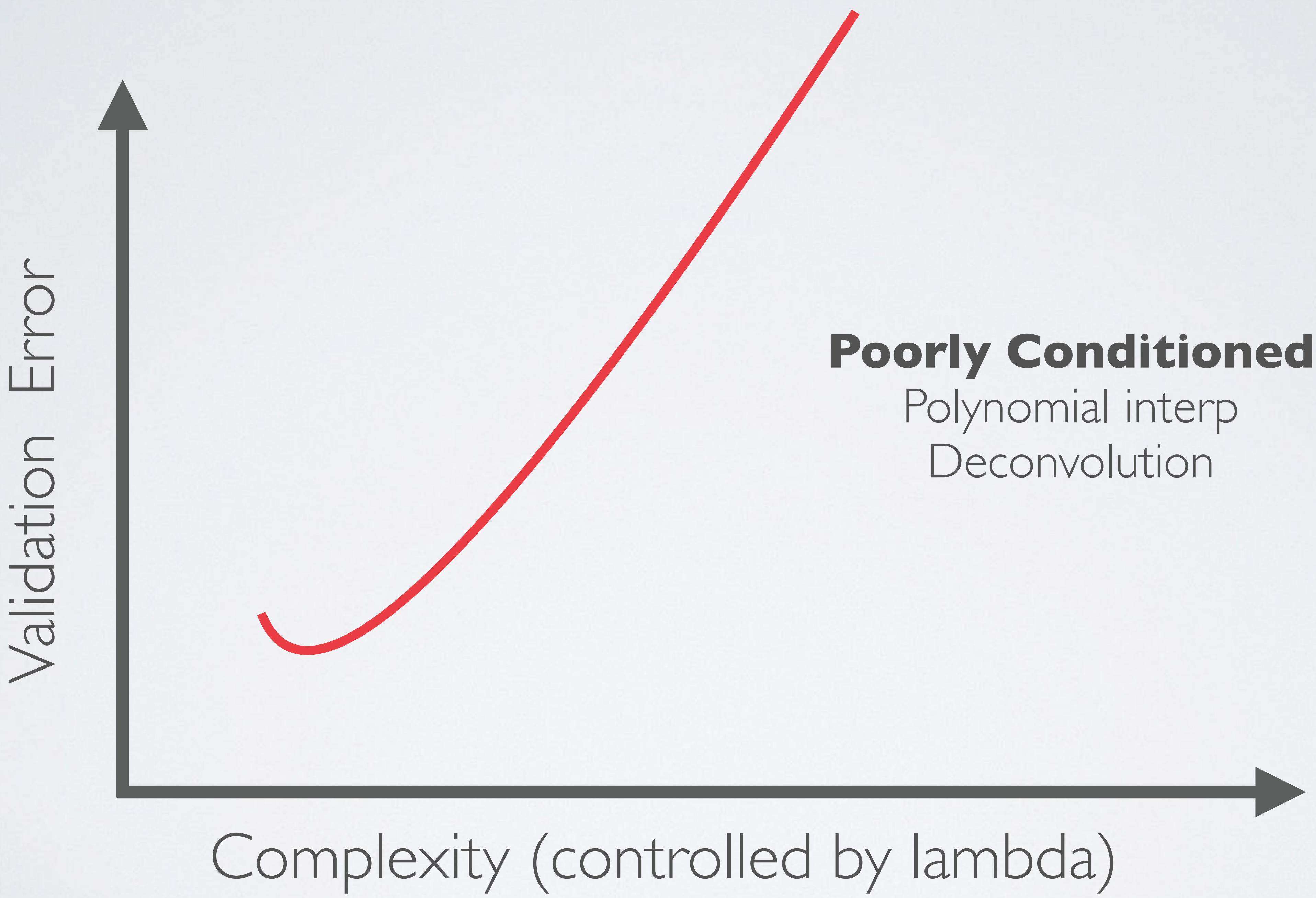
BIAS-VARIANCE TRADEOFF



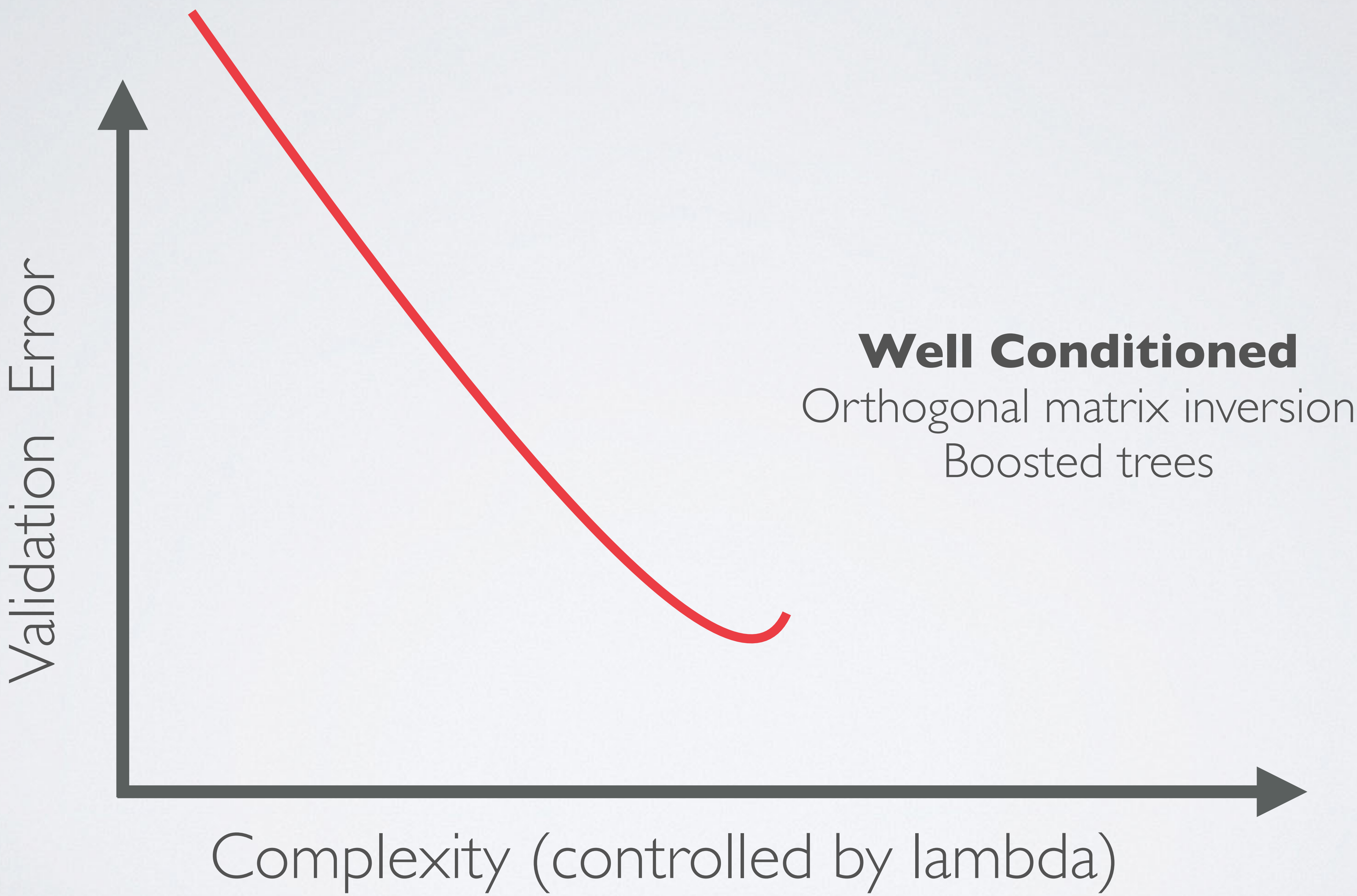
BIAS-VARIANCE TRADEOFF



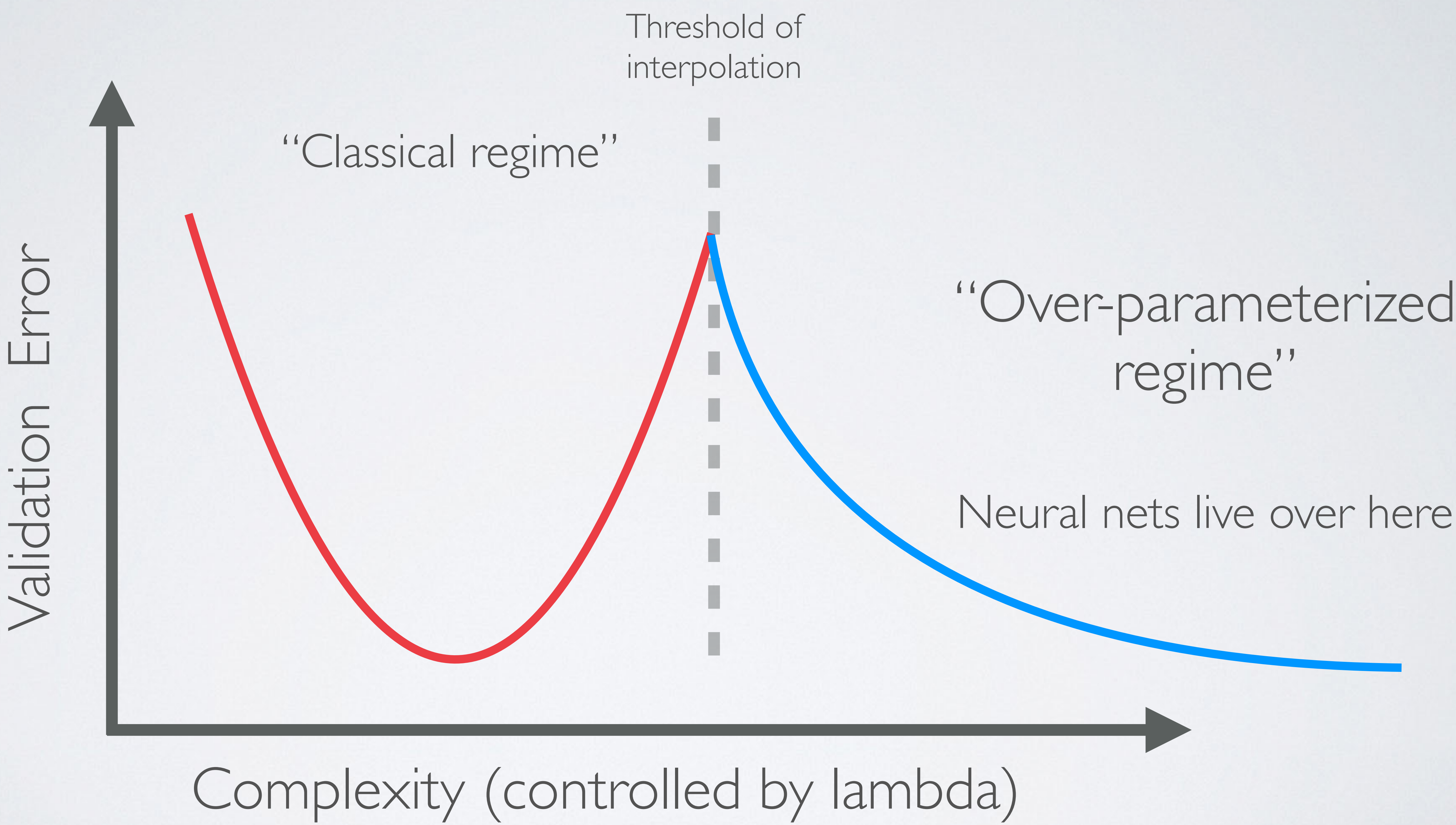
BIAS-VARIANCE TRADEOFF



BIAS-VARIANCE TRADEOFF



DOUBLE DESCENT

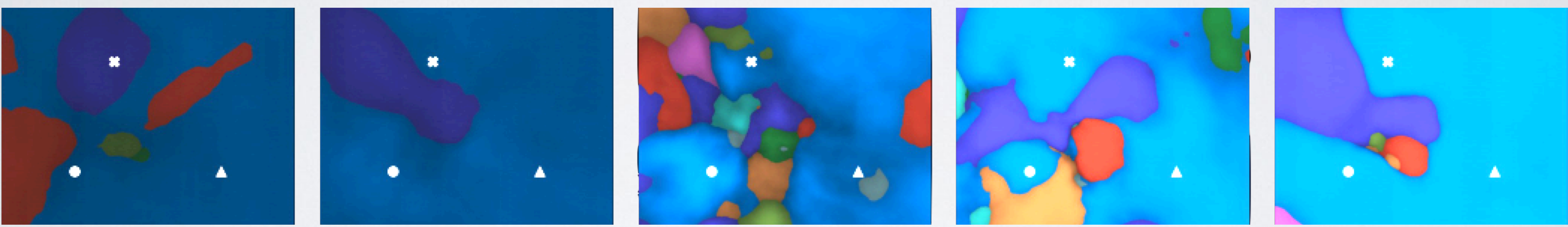


EXAMPLE

Decision boundaries in neural networks

ResNet18 on CIFAR10 with 20% label noise

Bad conditioning



1 4 10 20 64

First-layer network width

Somepalli "Can You Learn the Same Model Twice?"

SPARSE RECOVERY PROBLEMS

used to control over-fitting

$$\text{minimize} \quad \|x\|_0 \quad \text{subject to} \quad Ax = b$$

Sparse solution

"Fat" matrix (underdetermined)

Dense measurements

COMPLEXITY ALERT: NP-Complete

(Reductions to subset cover and SAT)

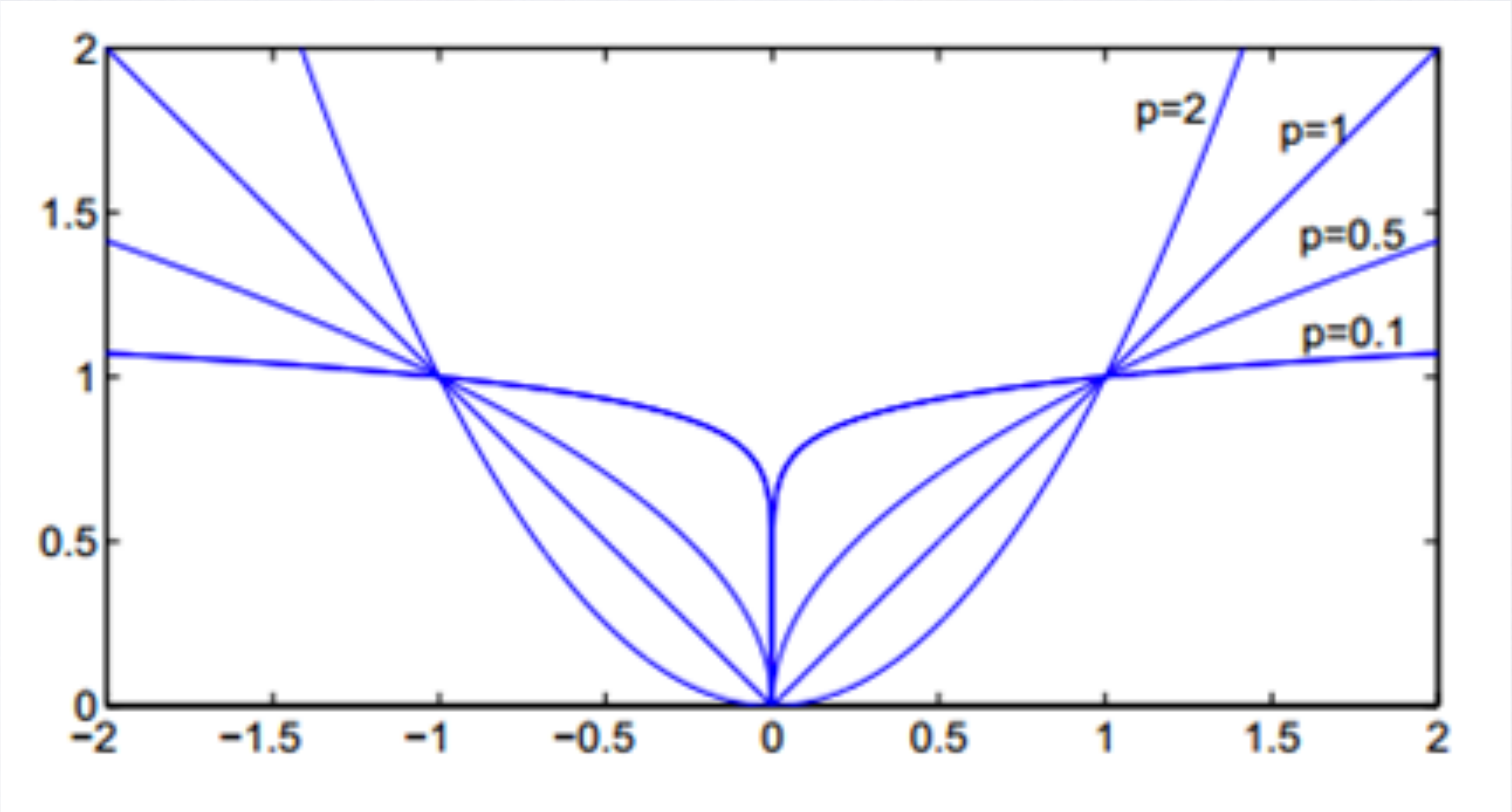
Nonetheless: can solve by greedy methods

- Orthogonal Matching Pursuit (OMP)
- Stagewise methods: StOMP, CoSAMP, etc…

L0 IS NOT CONVEX

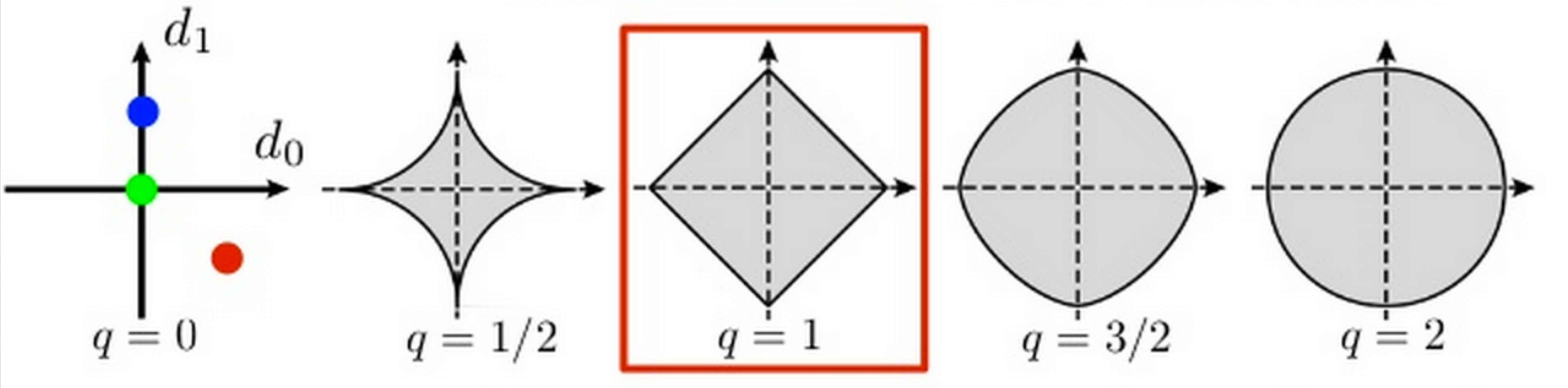
$$\text{minimize} \quad \|x\|_0 \;\; \text{subject to} \;\; Ax = b$$

$$\text{minimize} \quad |x| \;\; \text{subject to} \;\; Ax = b$$



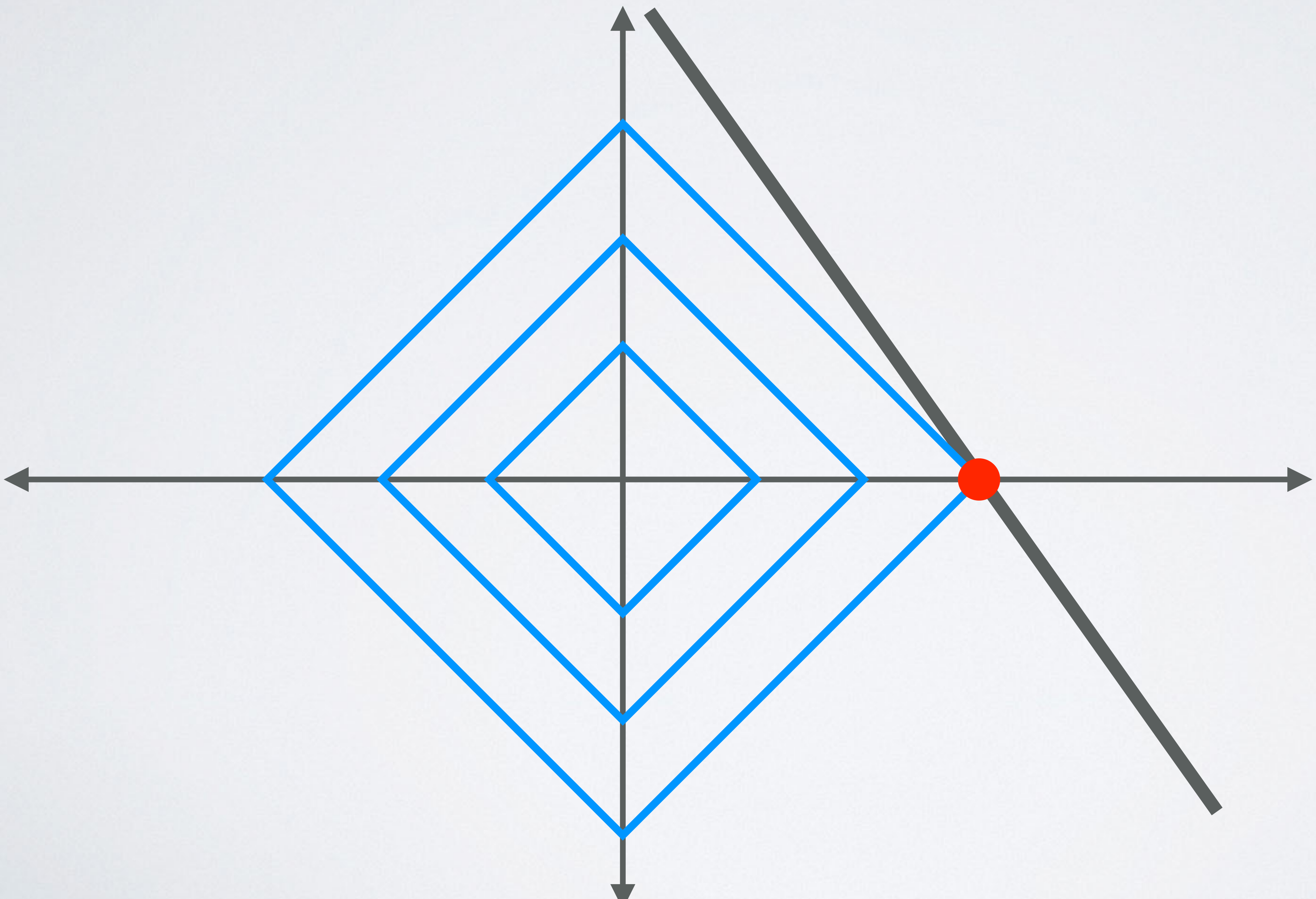
WHY USE L1?

- L1 is the "tightest" convex relaxation of L0



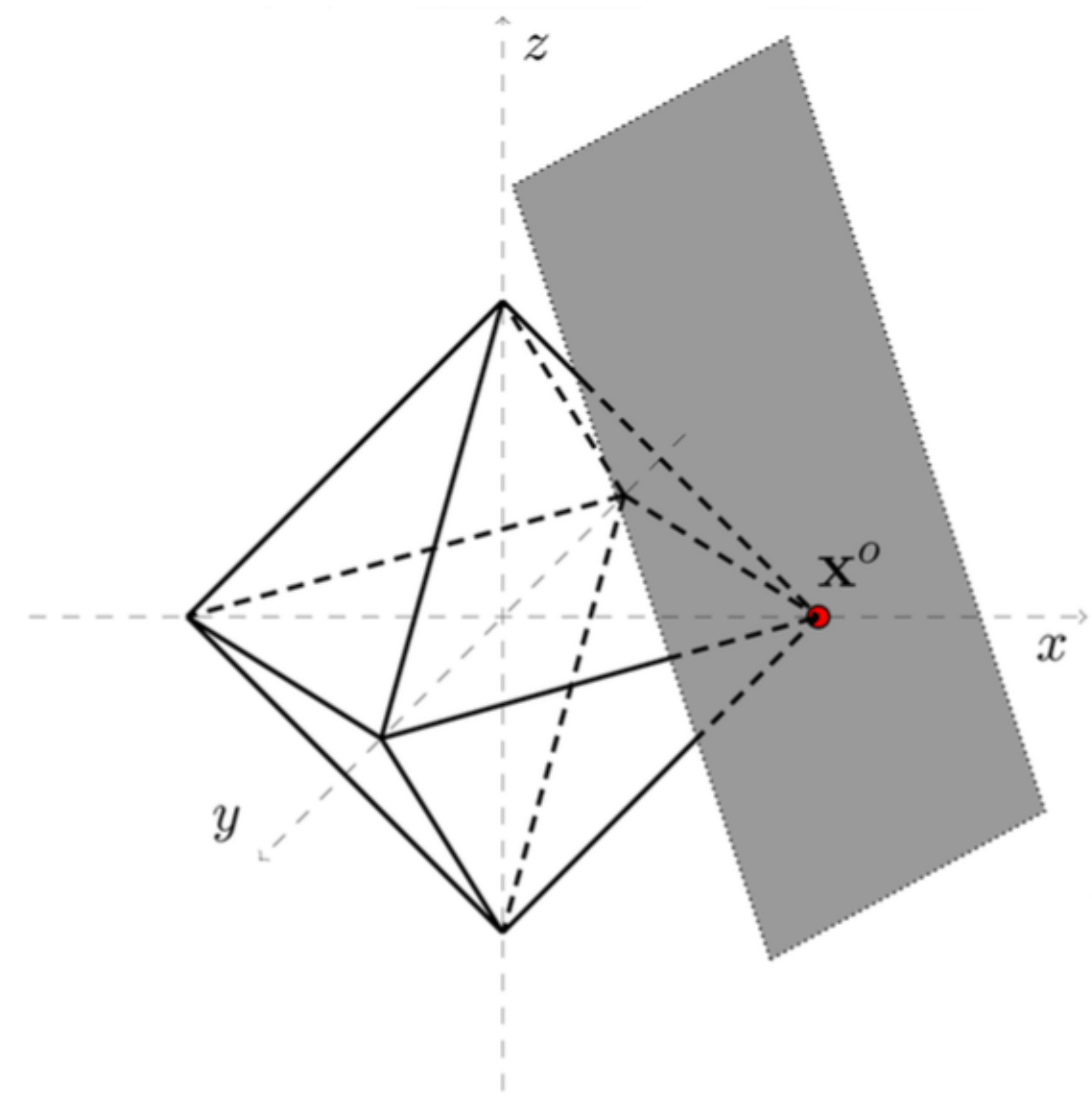
CONVEX RELAXATION

$$\text{minimize} \quad |x| \quad \text{subject to} \quad Ax = b$$



CONVEX RELAXATION

$$\text{minimize} \quad |x| \ \text{ subject to } \ Ax = b$$



ℓ_1 ball touches the affine plane

SPARSE OPTIMIZATION PROBLEMS

Basis Pursuit	$\text{minimize} \quad \lvert x \rvert \;\; \text{subject to} \;\; Ax = b$
Basis Pursuit Denoising	$\text{minimize} \quad \lambda \lvert x \rvert + \frac{1}{2}\lVert Ax - b \rVert^2$
Lasso	$\text{minimize} \quad \frac{1}{2}\lVert Ax - b \rVert^2 \;\; \text{subject to} \;\; \lvert x \rvert \leq \lambda$

BAYESIAN LAND!

Basis Pursuit Denoising

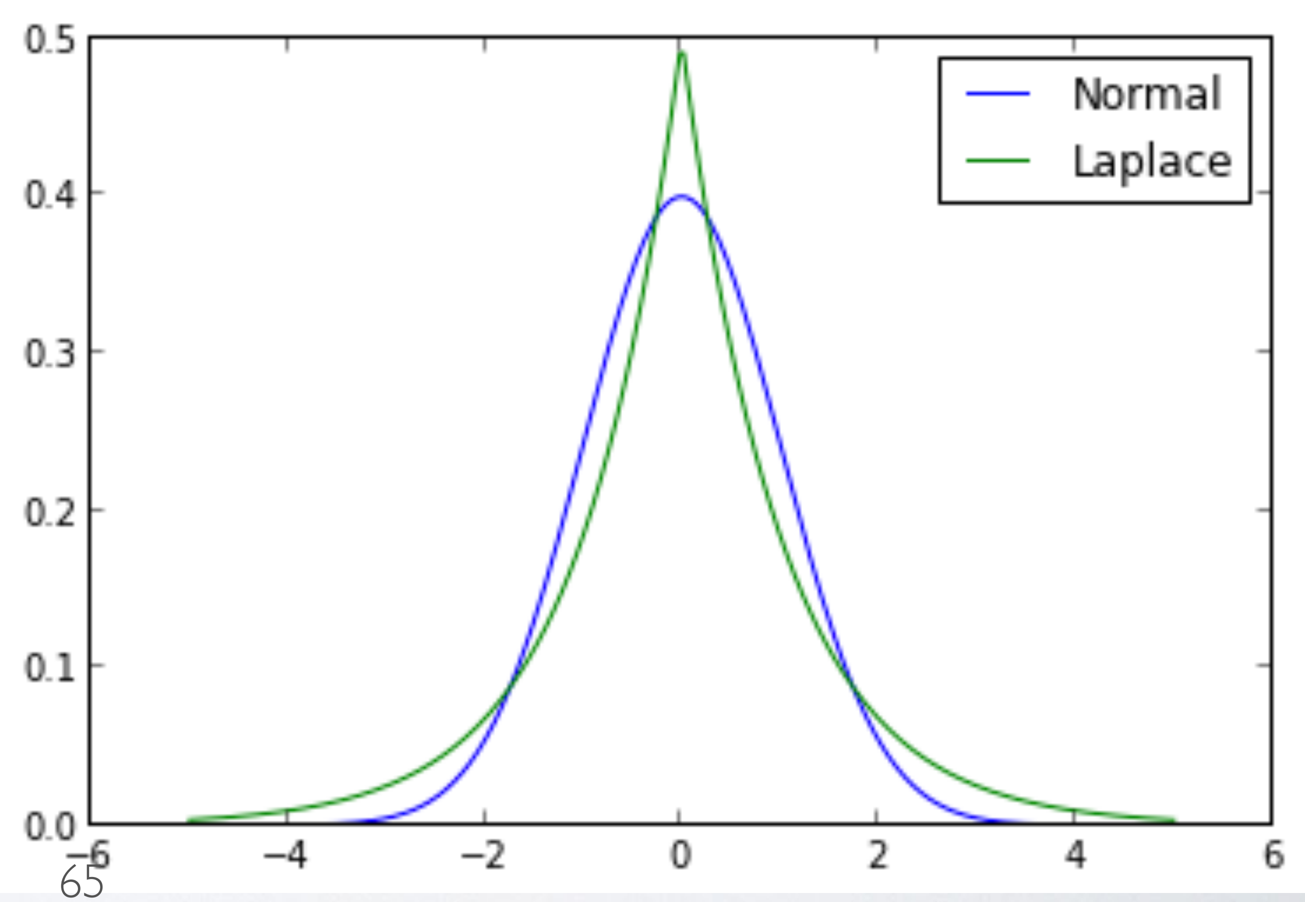
$$\text{minimize} \quad \lambda|x| + \frac{1}{2}\|Ax - b\|^2$$

What prior is this?

Laplace distribution

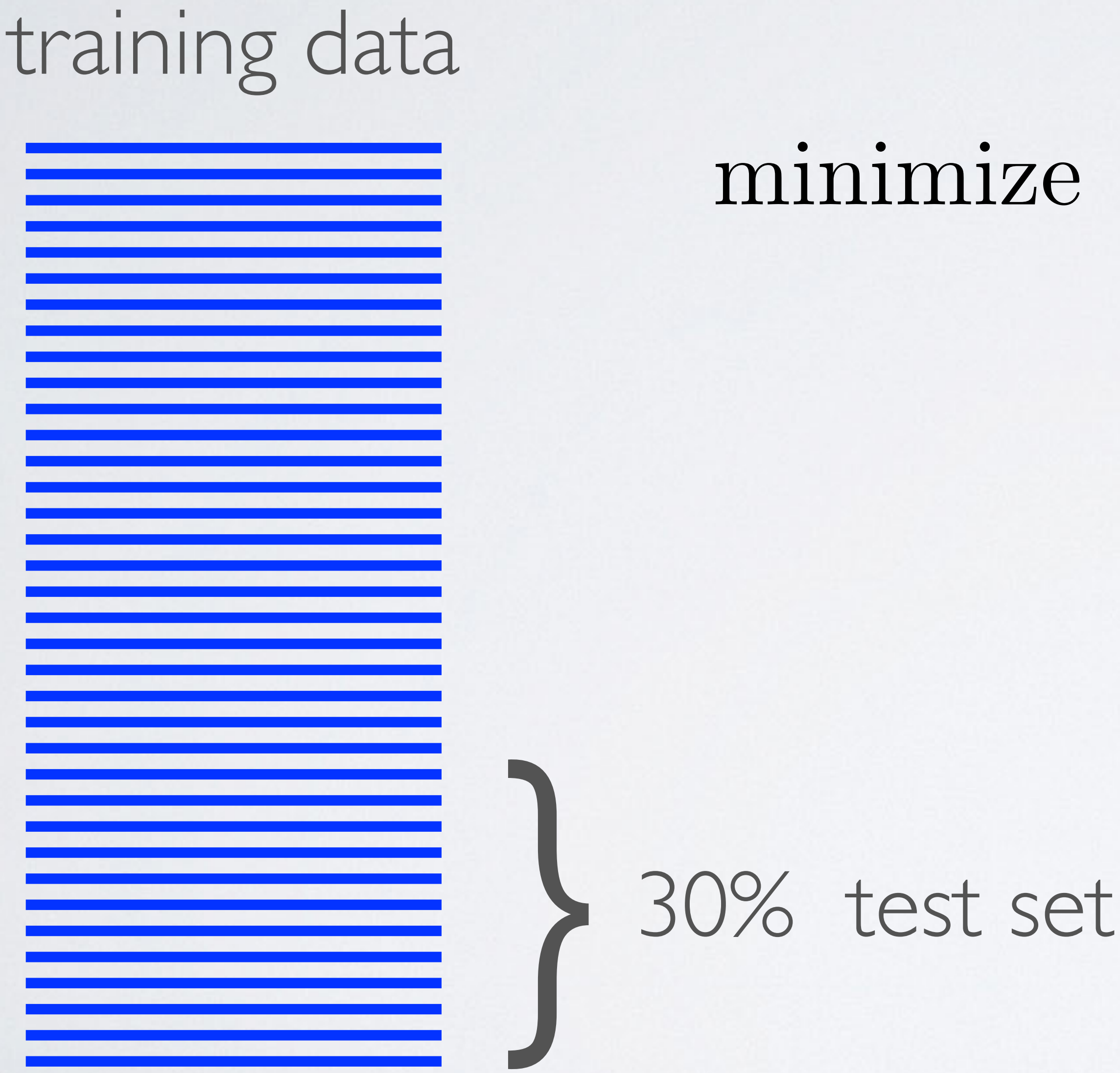
$$p(x) = e^{-|x|}$$

"robust to outliers"



HOW TO SET LAMBDA?

minimize **out-of-sample** error

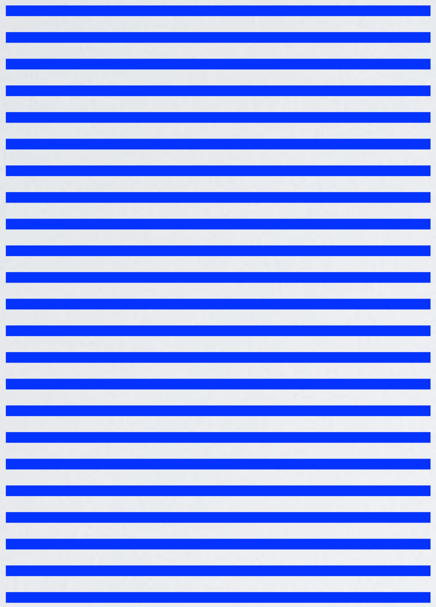


$$\text{minimize} \quad \lambda|x| + \frac{1}{2}\|Ax - b\|^2$$

CROSS VALIDATION

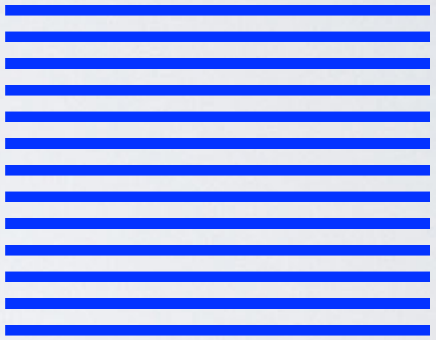
minimize **out-of-sample** error

training data



$$\text{minimize} \quad \lambda|x| + \frac{1}{2}\|Ax - b\|^2$$

test data

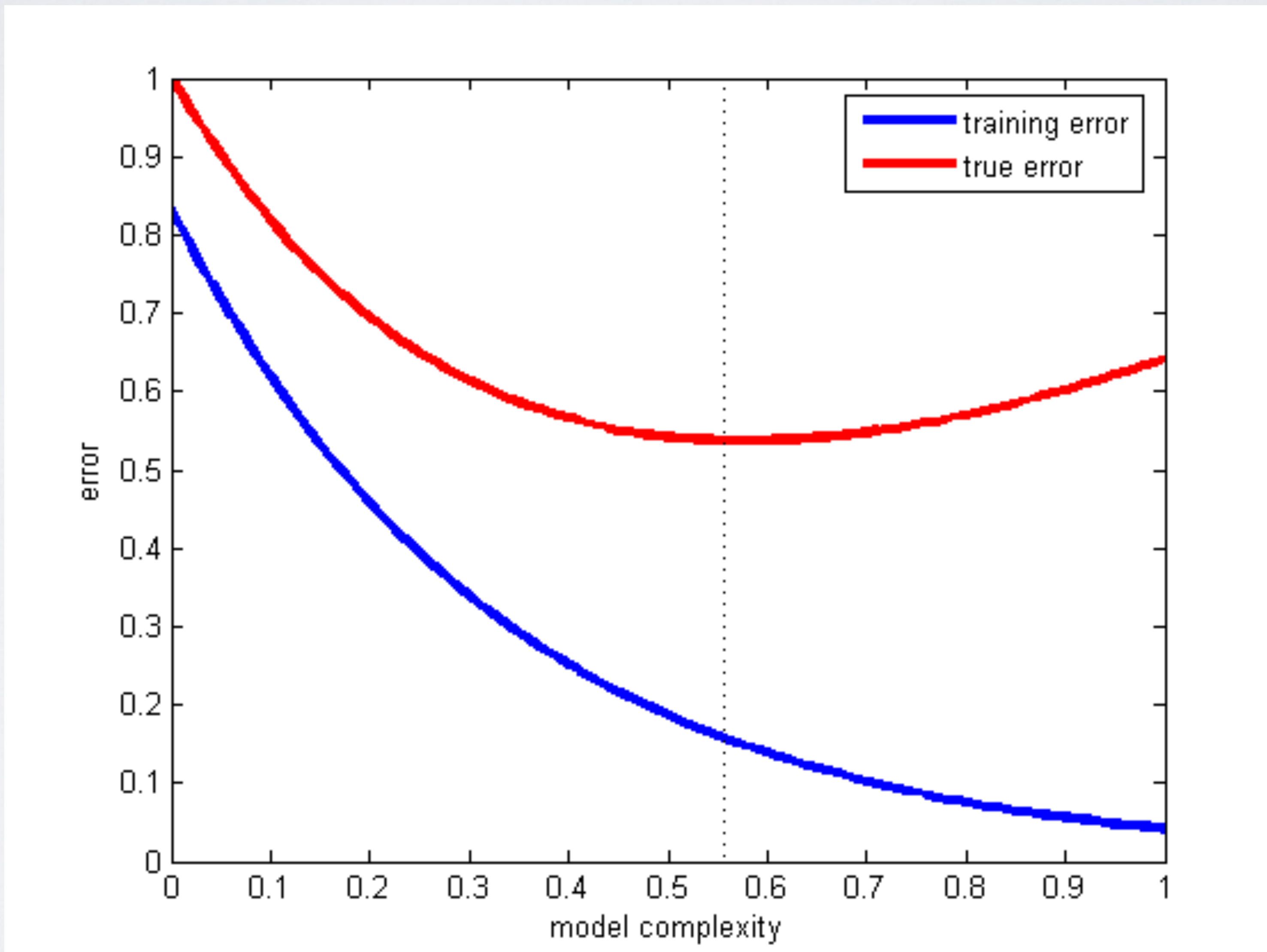


choose lambda to minimize
test error

CROSS VALIDATION

minimize **out-of-sample** error

training data



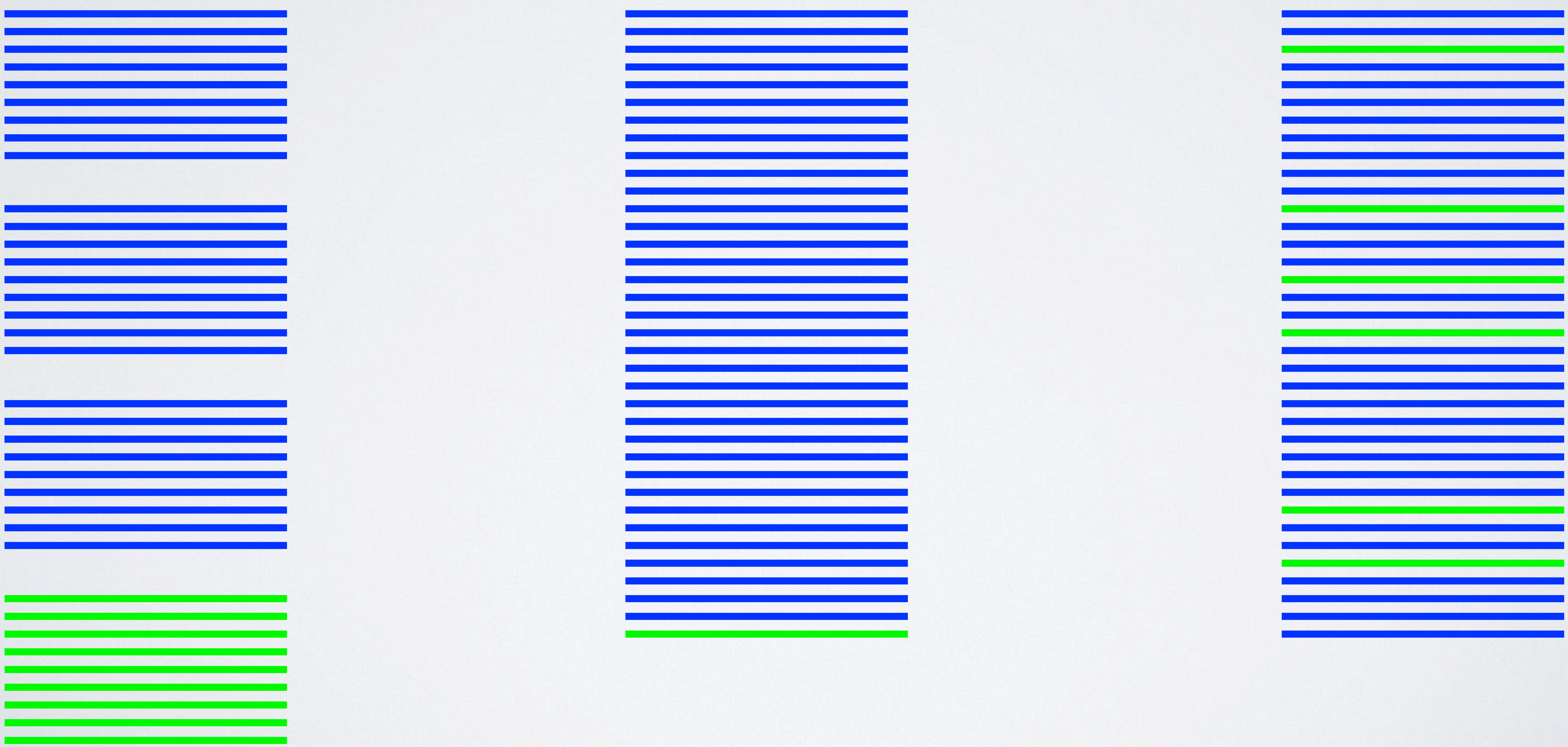
test data

idealistic curves: no sampling noise

CROSS VALIDATION

Do CV on multiple split of data to reduce noise

K-fold CV   leave-one-out CV   random sampling CV



AFTER MODEL SELECTION

$$\text{minimize} \quad \lambda|x| + \frac{1}{2}\|A_{tr}x - b_{tr}\|^2$$

de-biasing

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|A_{all}x - b_{all}\|^2 \\ &\text{subject to} \quad x \in C \end{aligned}$$

ensemble learning

$$\text{minimize} \quad \lambda|x| + \frac{1}{2}\|A_1x - b_1\|^2$$

$$\text{minimize} \quad \lambda|x| + \frac{1}{2}\|A_2x - b_2\|^2$$

$$\vdots$$

$$\text{minimize} \quad \lambda|x| + \frac{1}{2}\|A_Kx - b_K\|^2$$

These methods work best on small problems!

CO-SPARSITY

$$\text{minimize} \quad \lambda|\phi x| + \frac{1}{2}\|Ax - b\|^2$$

- Sometimes signal is sparse under a transform
- When transform is invertible, can use **synthesis**

$$\text{minimize} \quad \lambda|v| + \frac{1}{2}\|A\phi^{-1}v - b\|^2$$

- Otherwise, use the **analysis** formulation
- The thing in the L1 norm is sparse!

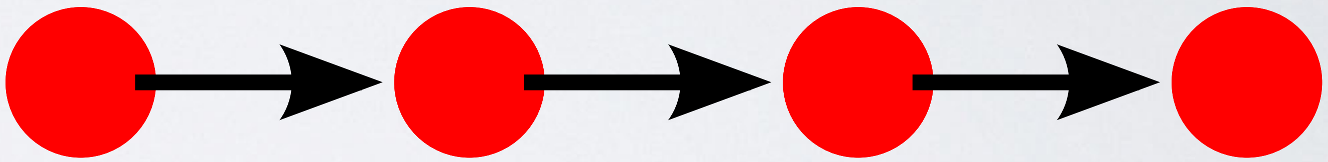
EXAMPLE: IMAGE PROCESSING

IMAGE GRADIENT

Stencil $[-1 \quad 1 \quad 0]$

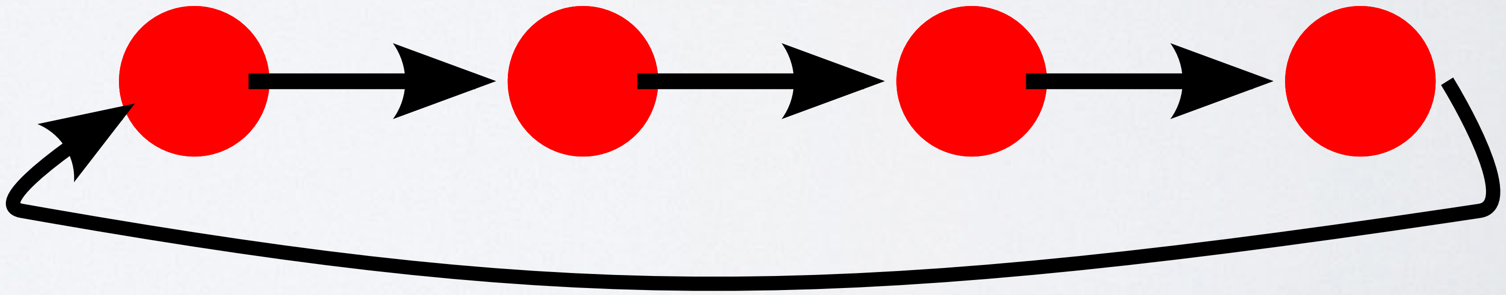
Neumann

$$\begin{pmatrix} u_1 - u_0 \\ u_2 - u_1 \\ u_3 - u_2 \end{pmatrix}$$



Circulant

$$\begin{pmatrix} u_1 - u_0 \\ u_2 - u_1 \\ u_3 - u_2 \\ \color{red}{u_0 - u_3} \end{pmatrix}$$



TOTAL VARIATION

$$TV(x) = \sum |x_{i+1} - x_i|$$

- Discrete gradient operator

$$\nabla x = (x_2 - x_1, x_3 - x_2, x_4 - x_3)$$

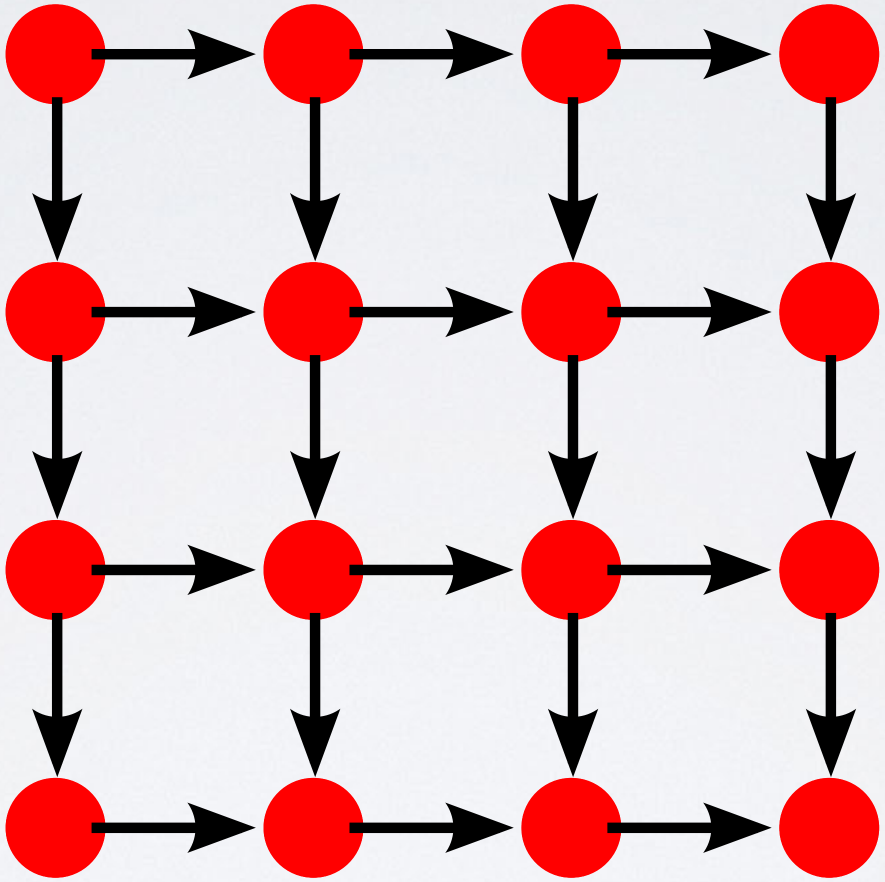
- Linear filter with "Stencil" $(-1, 1)$

$$TV(x) = |\nabla x|$$

Is this a norm?

TV IN 2D

$$(\nabla x)_{ij} = (x_{i+1,j} - x_{i,j}, x_{i,j+1} - x_{i,j})$$



Anisotropic $|(\nabla x)_{ij}| = |x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j})|$

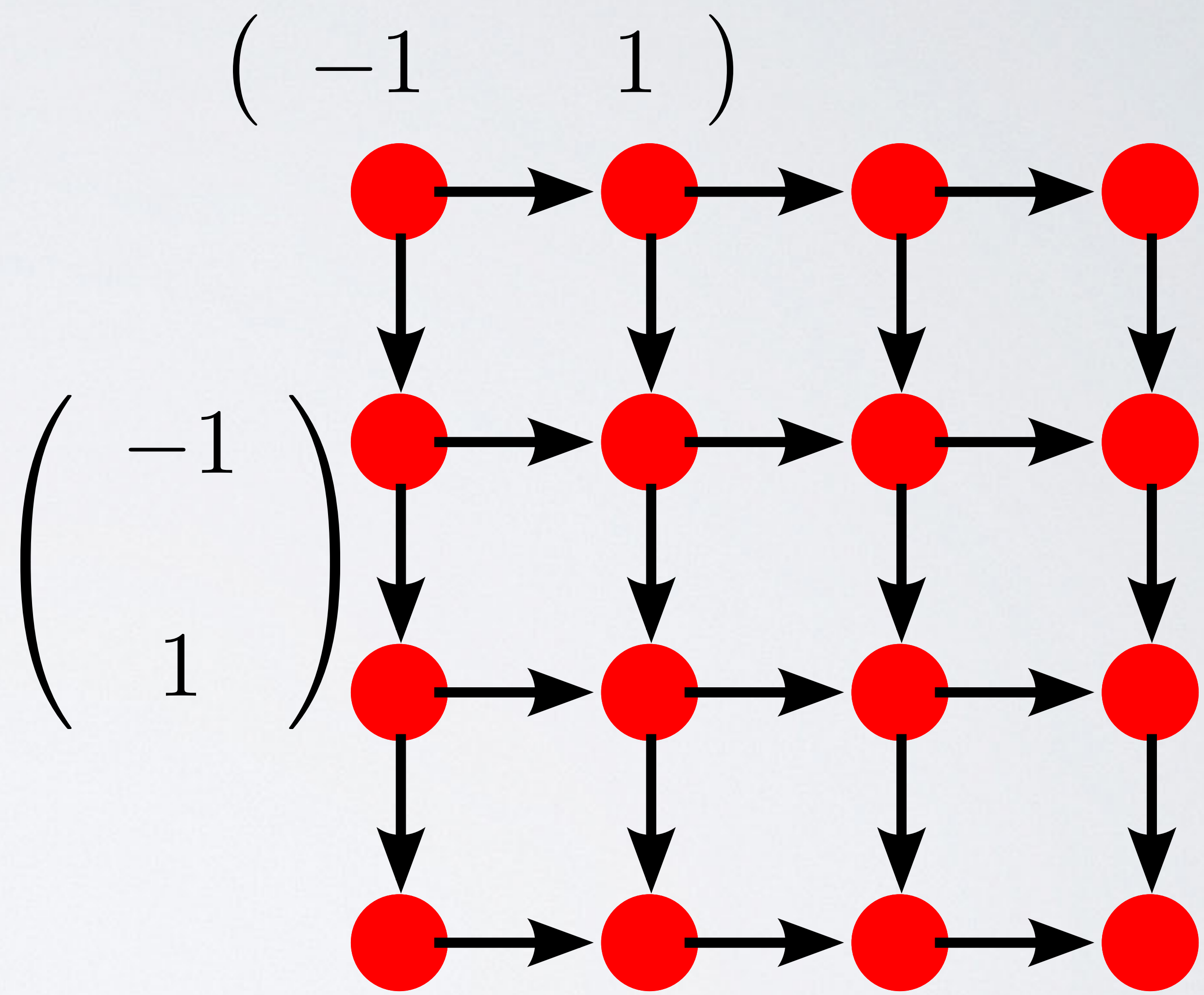
Isotropic $\|(\nabla x)_{ij}\| = \sqrt{(x_{i+1,j} - x_{ij})^2 + (x_{i,j+1} - x_{ij})^2}$

COMPUTING TV ON IMAGES

- Two linear filters
- x-stencil = (-1 1 0)
- y-stencil = (-1 1 0)'

…or…

- Two linear convolutions
- x-kernel = (0 1 -1)
- y-kernel = (0 1 -1)'

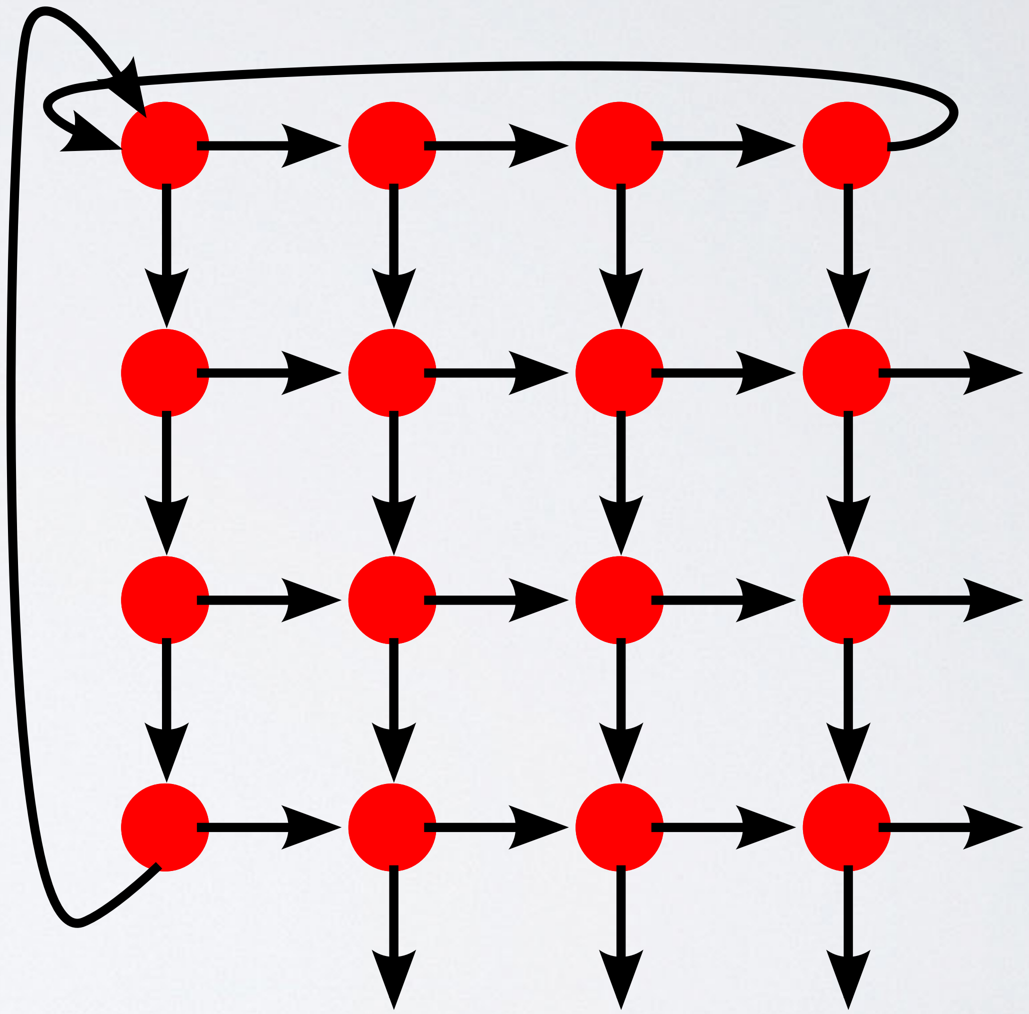


COMPUTING TV ON IMAGES

- Two linear convolutions
- x-kernel = (0 1 -1)
- y-kernel = (0 1 -1)'

Fast Transforms: use FFT

Circulant Boundary

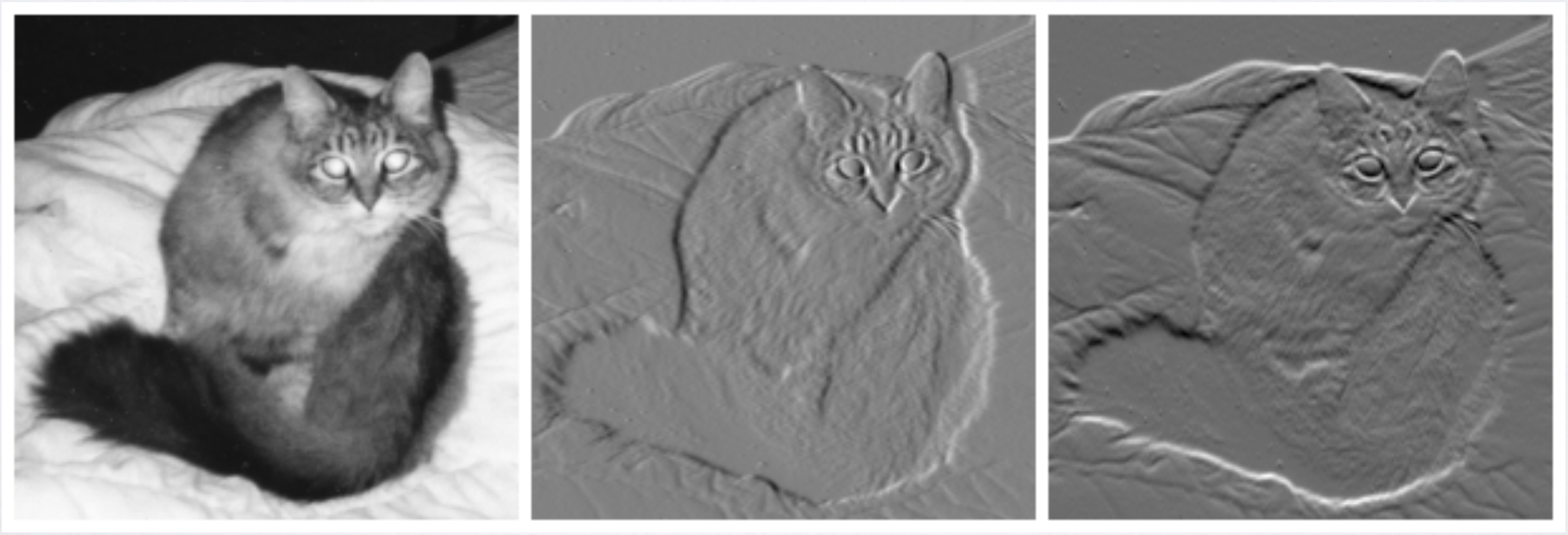


TOTAL VARIATION: 2D

$$(\nabla x)_{ij} = (x_{i+1,j} - x_{ij}, x_{i,j+1} - x_{ij})$$

$$TV_{iso}(x) = |\nabla x| = \sum_{ij} \sqrt{(x_{i+1,j} - x_{ij})^2 + (x_{i,j+1} - x_{ij})^2}$$

$$TV_{an}(x) = |\nabla x| = \sum_{ij} |x_{i+1,j} - x_{ij}| + |x_{i,j+1} - x_{ij}|$$

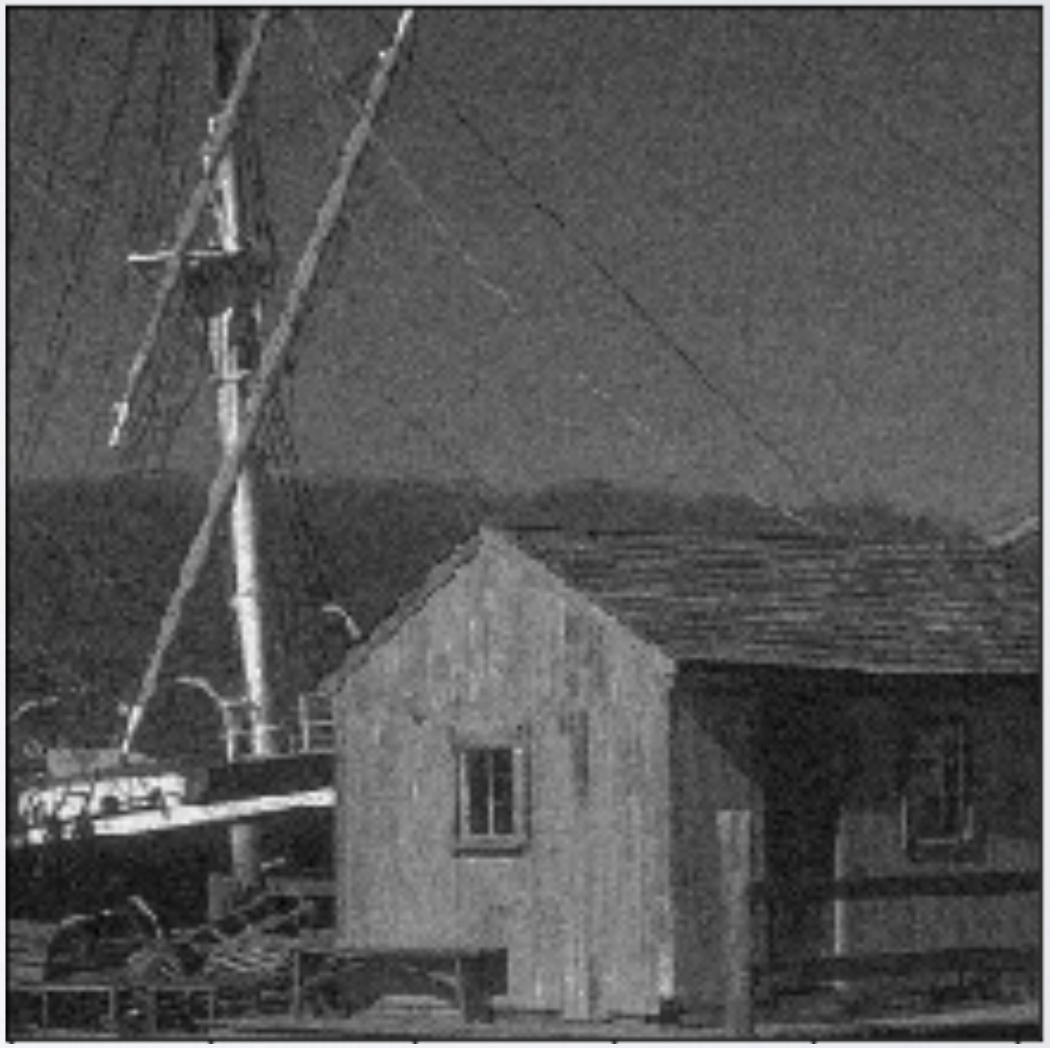


I $\quad g_x = D_x I$ $\quad g_y = D_y I$

IMAGE RESTORATION

Original

Noisy



TOTAL VARIATION DENOISING

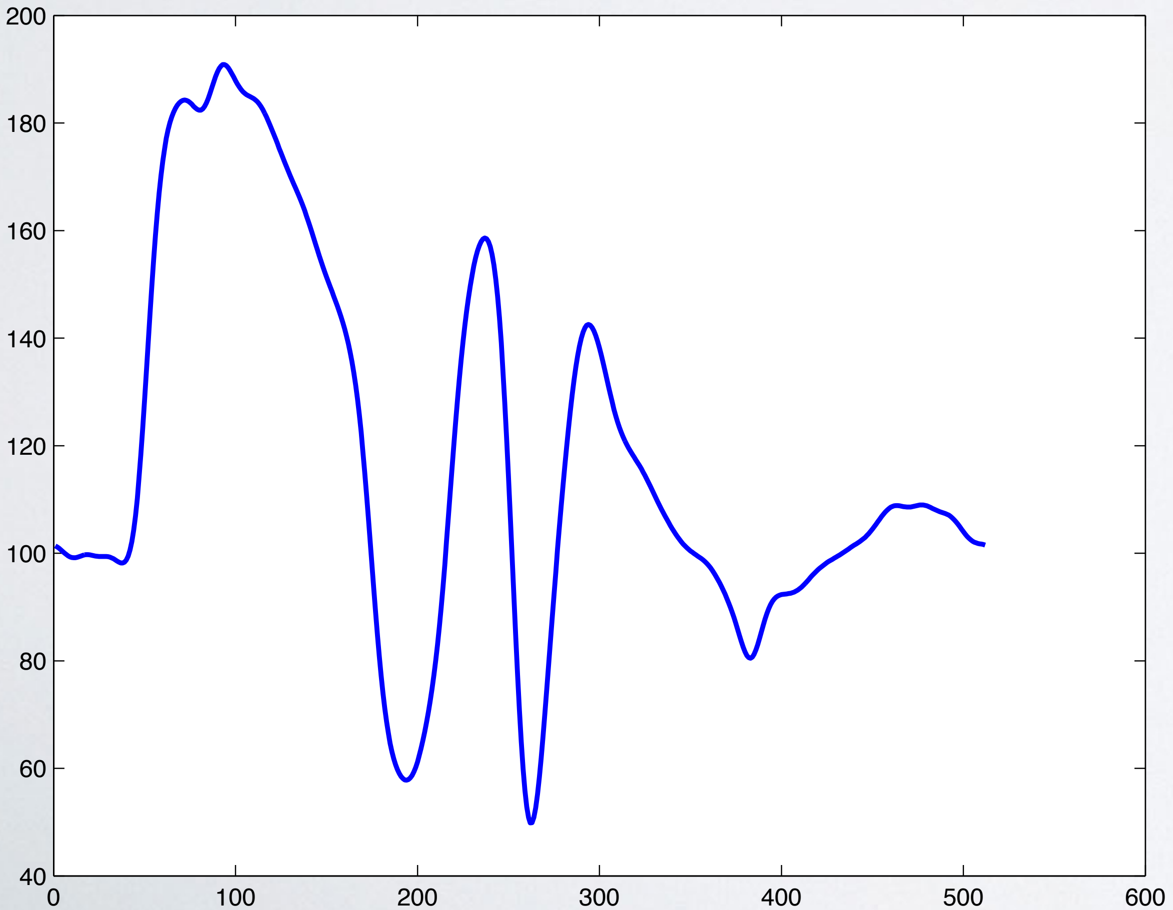
minimize $\quad \lambda\|\nabla x\|^2 + \|x - f\|^2$

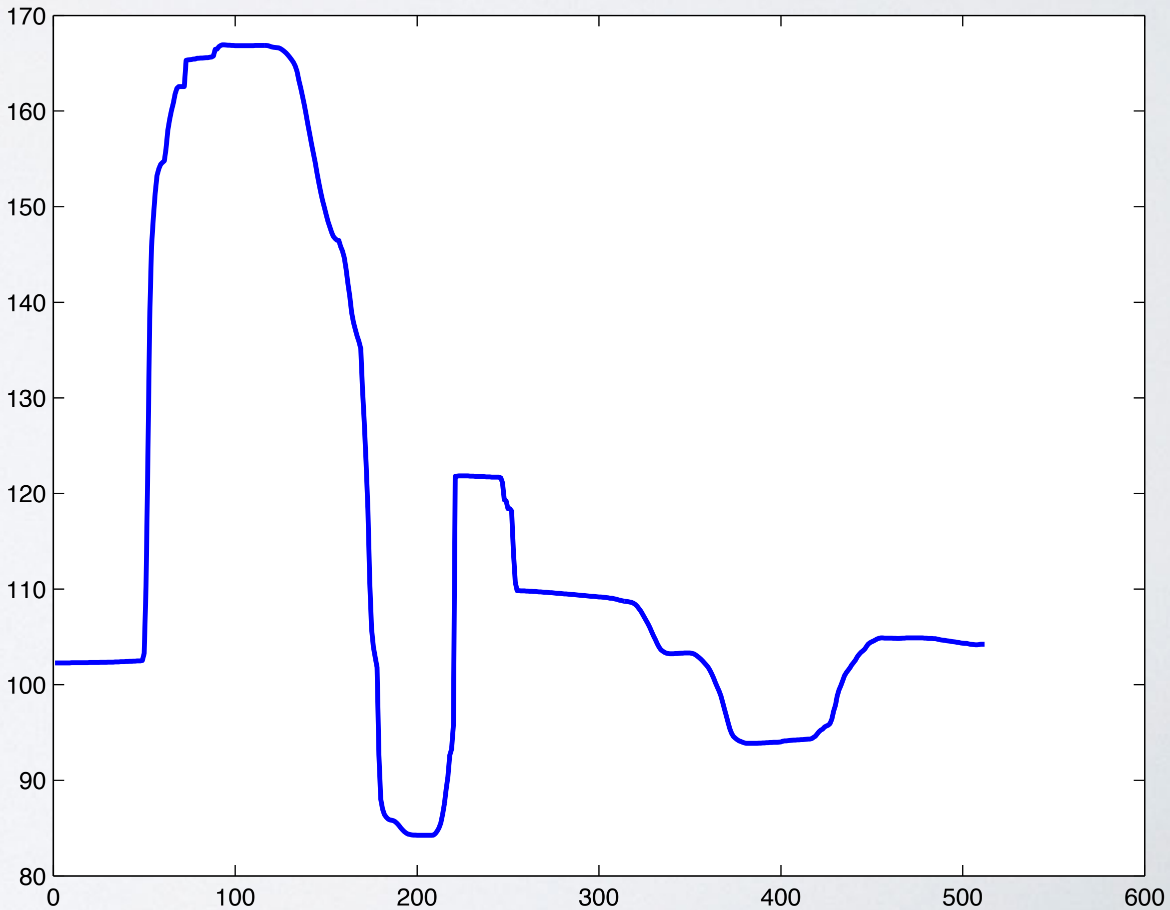
minimize $\quad \lambda|\nabla x| + \frac{1}{2}\|x - f\|^2$

slice

$$\text{minimize} \quad \lambda\|\nabla x\|^2 + \|x - f\|^2$$

$$\text{minimize} \quad \lambda|\nabla x| + \frac{1}{2}\|x - f\|^2$$





TV IMAGING PROBLEMS

Denoising (ROF) $$\text{minimize} \quad \lambda|\nabla x| + \frac{1}{2}\|x - f\|^2$$

Deblurring $$\text{minimize} \quad \lambda|\nabla x| + \frac{1}{2}\|Kx - f\|^2$$

TVL1 $$\text{minimize} \quad \lambda|\nabla x| + |x - f|$$