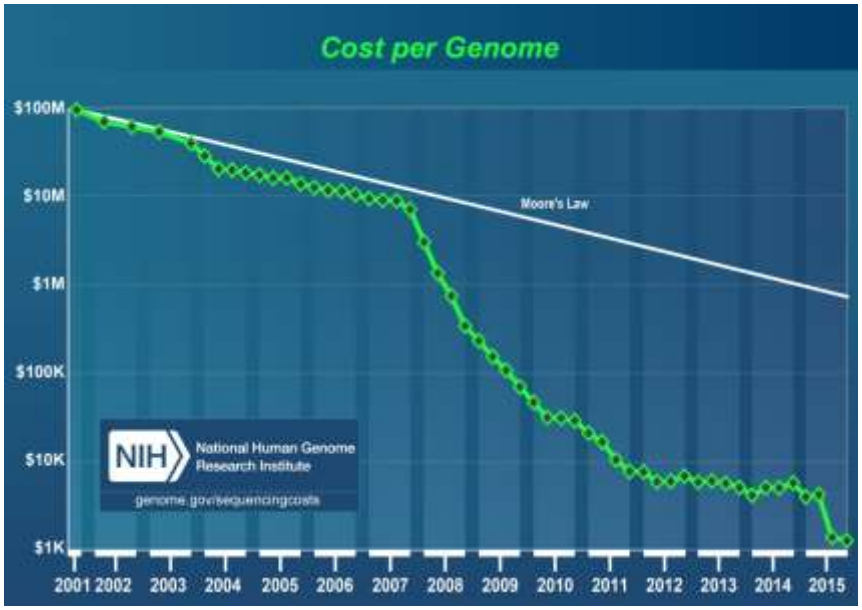
ThinkFast: Scaling Machine Learning to Modern Demands

Hristo Paskov

The Genomic Data Deluge

- Precision Medicine Initiative: sequence **1,000,000** genomes
  - $**215 Million** in 2015
  - Pilot study
  - Outputs **10-50 GB/person**



- How do we analyze all of this data to **drive progress**?

Massive Data Sources

Bioinformatics

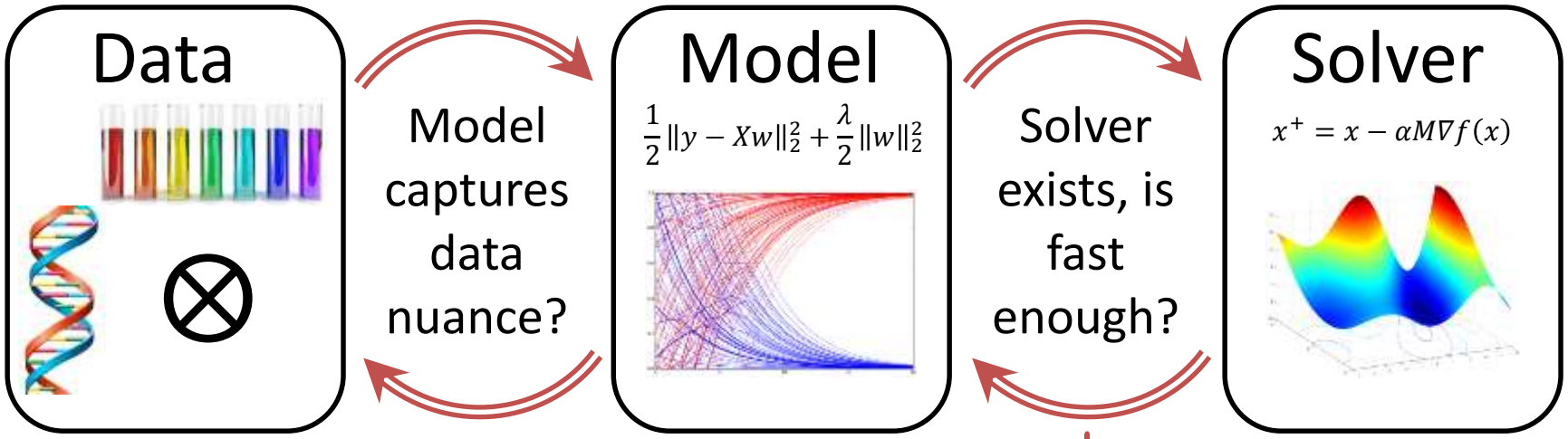
100K Genomes

News

Social Media

eCommerce

The Analysis Refinement Cycle

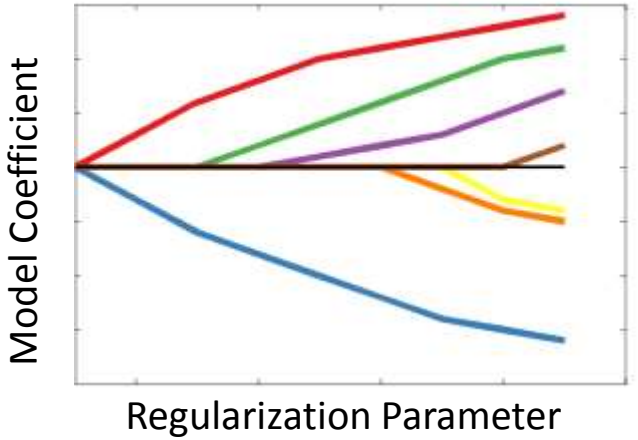


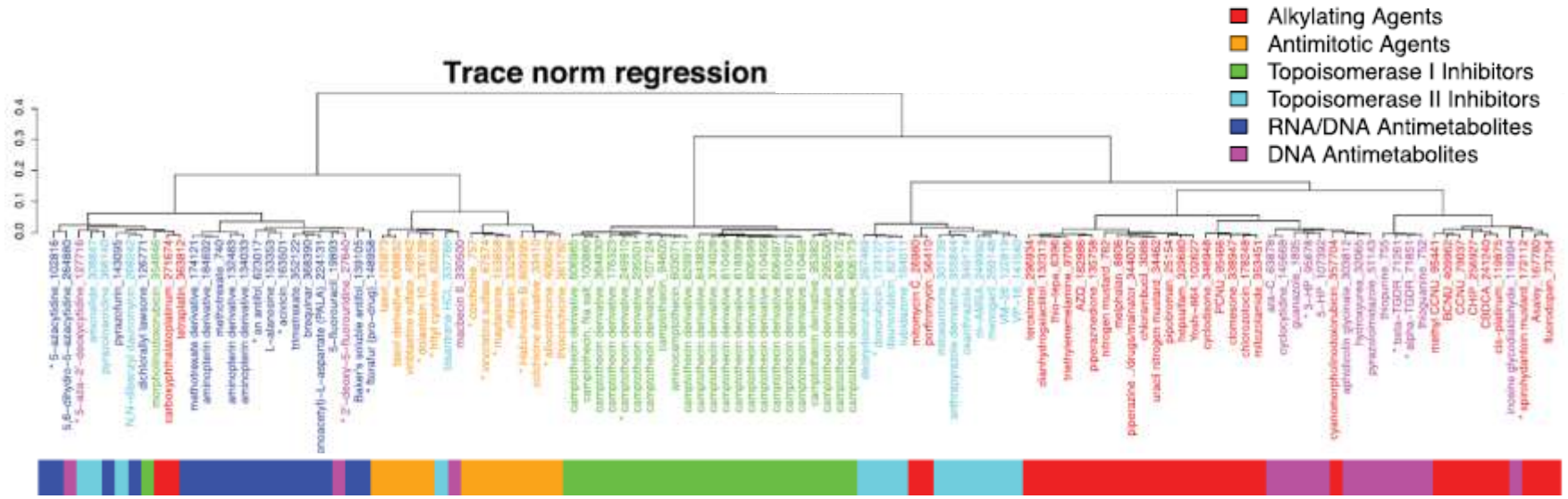
✓ Yes? — Proceed

! No? — Quit / Increase time, money, experience, resources

More Than Just Training Models

- Regularization paths
- Model risk assessment
- Interpretability

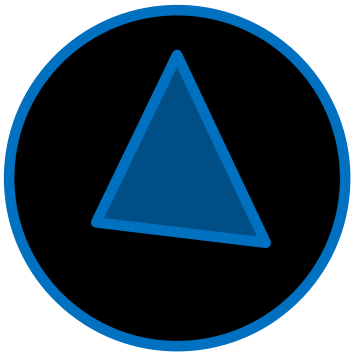




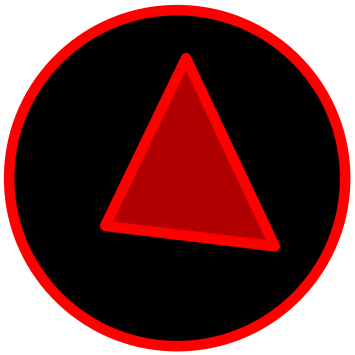
Brief History of Statistical Learning

Simple Models

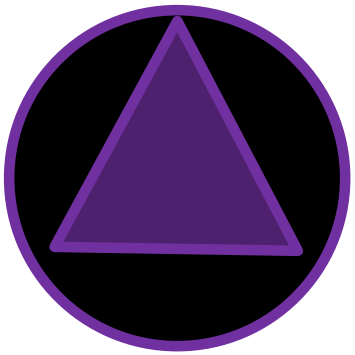
Kernel Methods



Trees & Ensembles



Structured Regularization



Interpretability & Statistical Guarantees

Scalability

Ease of Use

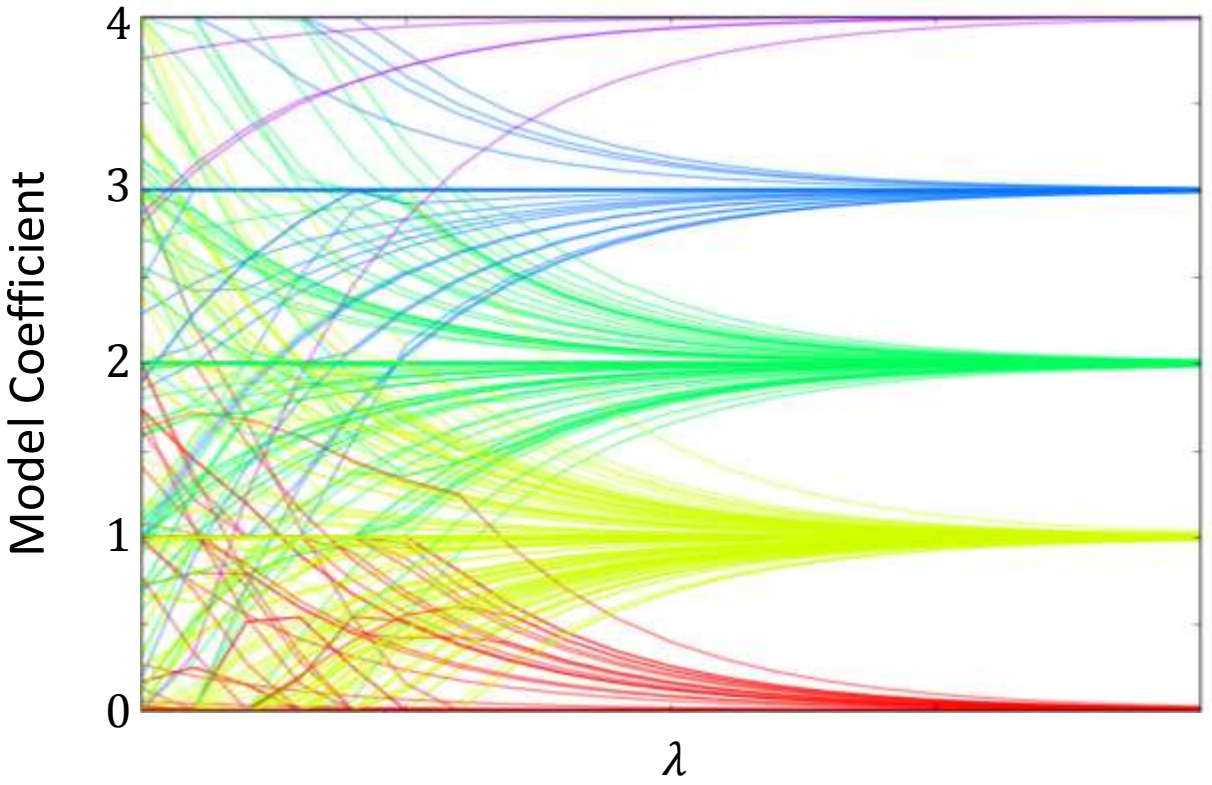
Structured Regularization

$$\min_{\beta \in \mathbb{R}^d} L(X\beta) + \lambda R(\beta)$$

Losses	**Regularizers**
Regression	Sparsity
Classification	Spatial/ Temporal /
Ranking	Manifold Structure
Motif Finding	Group Structure
Matrix Factorization	Hierarchical Structure
Feature Embedding	Structured & Unstructured
Data Imputation	Multitask Learning
…	…

The Lasso's Combinatorial Side

$$\min_{\beta \in \mathbb{R}^d} L(y - X\beta) + \lambda \|\beta\|_1$$



The Database Perspective

$$\min_{\beta \in \mathbb{R}^d} L(y - X\beta) + \lambda \|\beta\|_1$$

$$-X^T \partial_{y - X\beta} L(y - X\beta) + \lambda \partial_\beta \|\beta\|_1$$

The Database Perspective

$$-X^T \partial_{y-X\beta} L(y - X\beta) + \lambda \partial_\beta \|\beta\|_1$$

Feature & label storage

The Database Perspective

$$-X^T \partial_{y-X\beta} L(y - X\beta) + \lambda \partial_\beta \|\beta\|_1$$

Data access operations

$$u = y - X\beta$$
$$v = \partial_u L(u)$$
$$w = X^T v$$

Feature & label storage

The Database Perspective

$$-X^T \partial_{y-X\beta} L(y - X\beta) + \lambda \partial_\beta \|\beta\|_1$$

ML "Query Language"

$$\min_{\beta \in \mathbb{R}^d} L(y - X\beta) + \lambda \|\beta\|_1$$

Data access operations

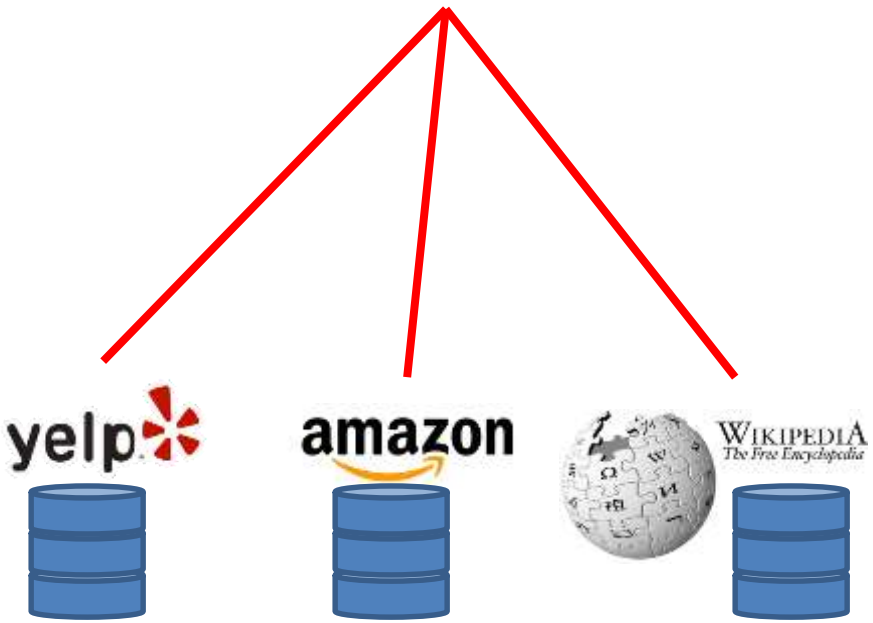
$$u = y - X\beta$$
$$v = \partial_u L(u)$$
$$w = X^T v$$

Feature & label storage

The Database Perspective

$$\min_{\beta_1,\beta_2,\beta_3 \in \mathbb{R}^d} \sum_{t=1}^{3} [L_t(y_t - X_t\beta_t) + \lambda_t R_t(\beta_t)]$$

$$+\omega \|[\beta_1 \quad \beta_2 \quad \beta_3]\|_*$$



The Database Perspective

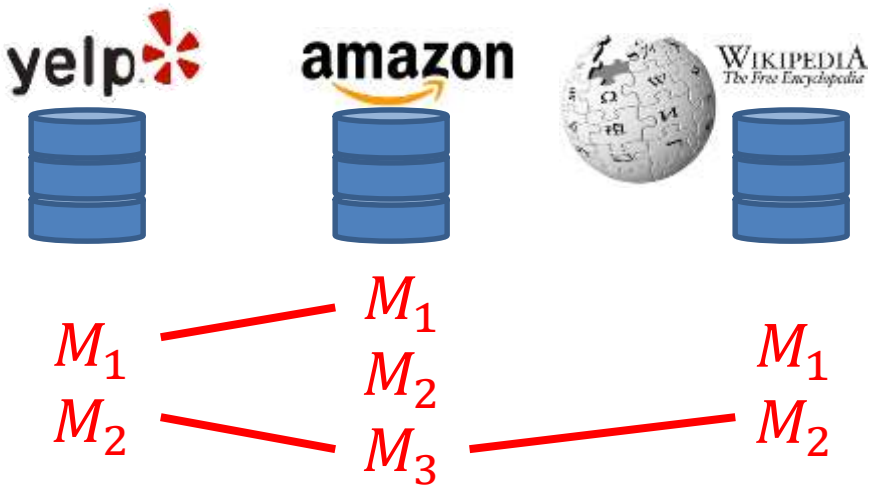
ML “Query Language”

Data access operations

Feature, label and model storage

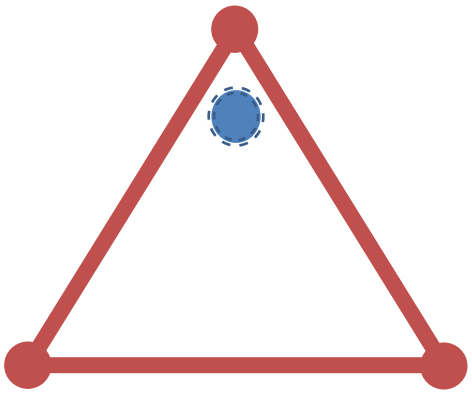
$$\min_{\beta \in \mathbb{R}^d} L(y - X\beta) + \lambda \|\beta\|_1$$

$$u = y - X\beta$$
$$v = \partial_u L(u)$$
$$w = X^T v$$



The Database Perspective

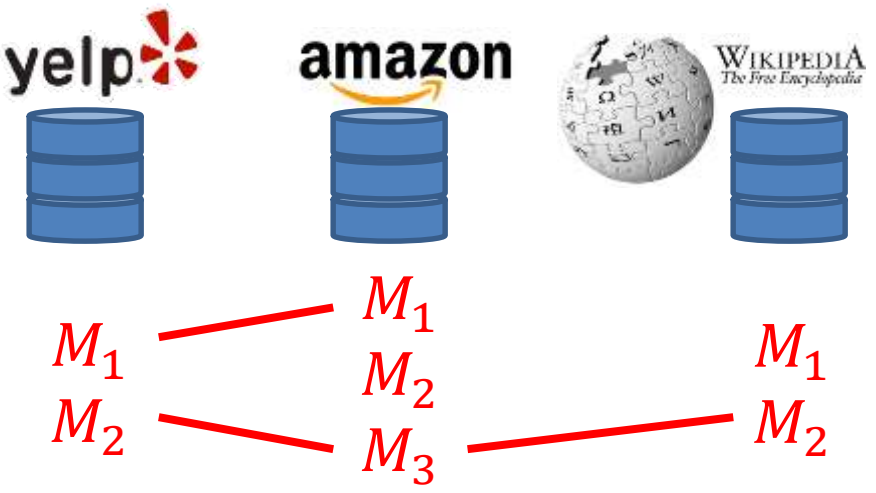
Mathematical Structure



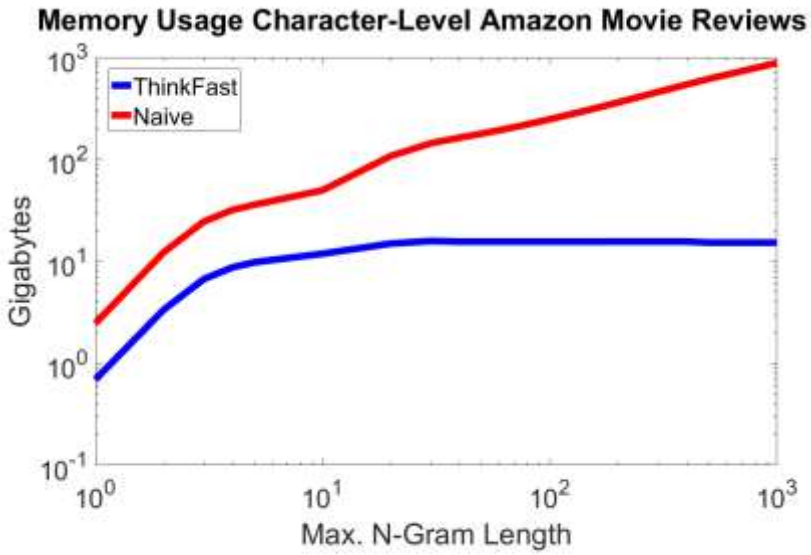
Processing Memory

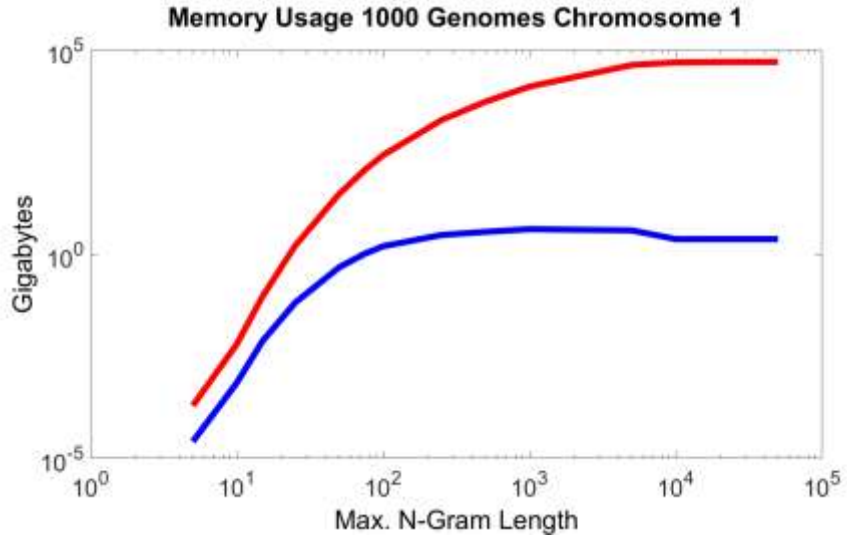
$$\min_{\beta \in \mathbb{R}^d} L(y - X\beta) + \lambda \|\beta\|_1$$

$$u = y - X\beta$$
$$v = \partial_u L(u)$$
$$w = X^T v$$



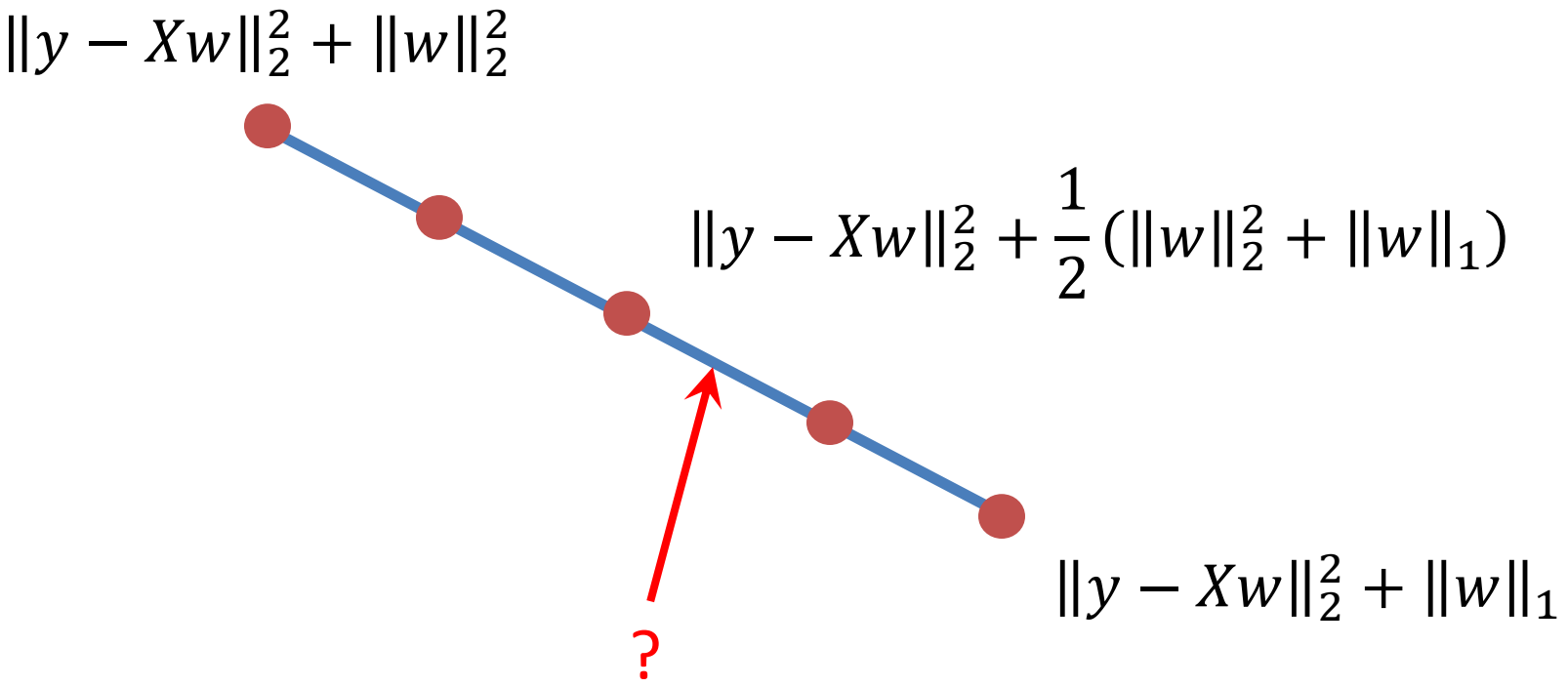
Efficient Feature Storage





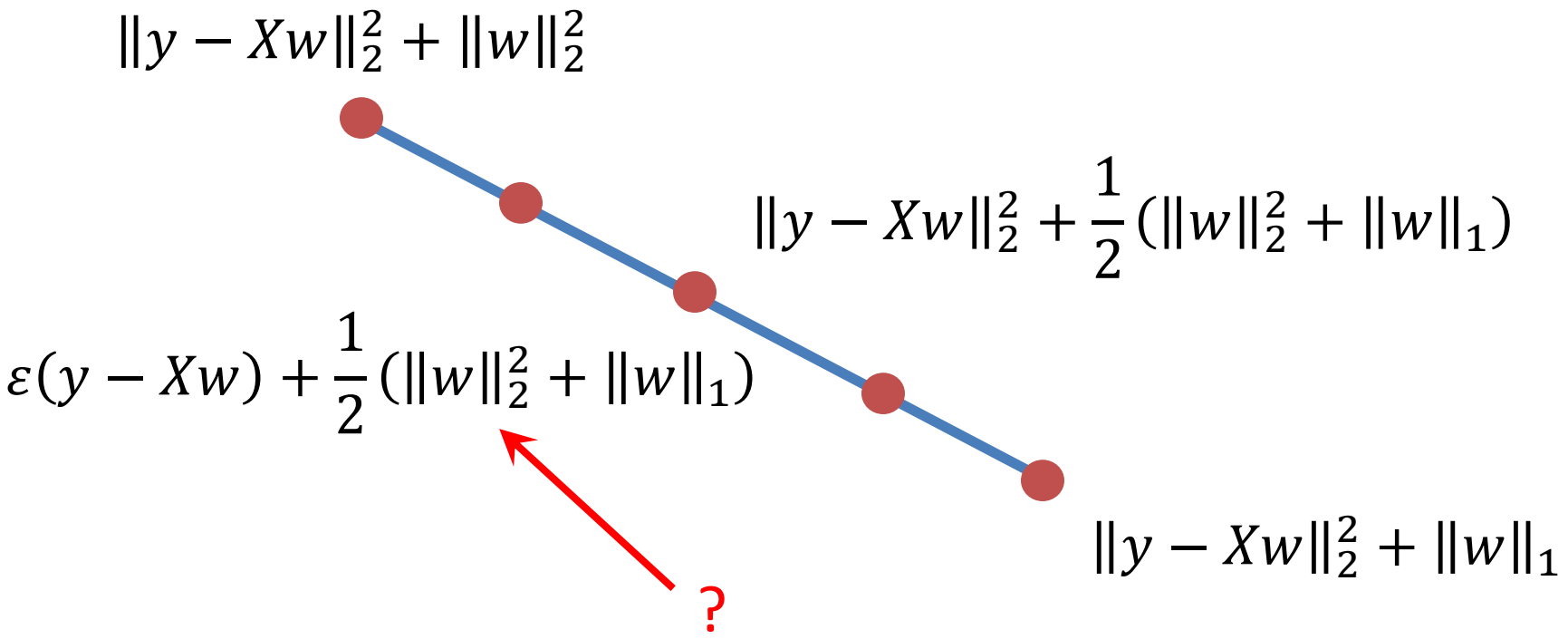
“Query Language” Optimization

- Static analysis



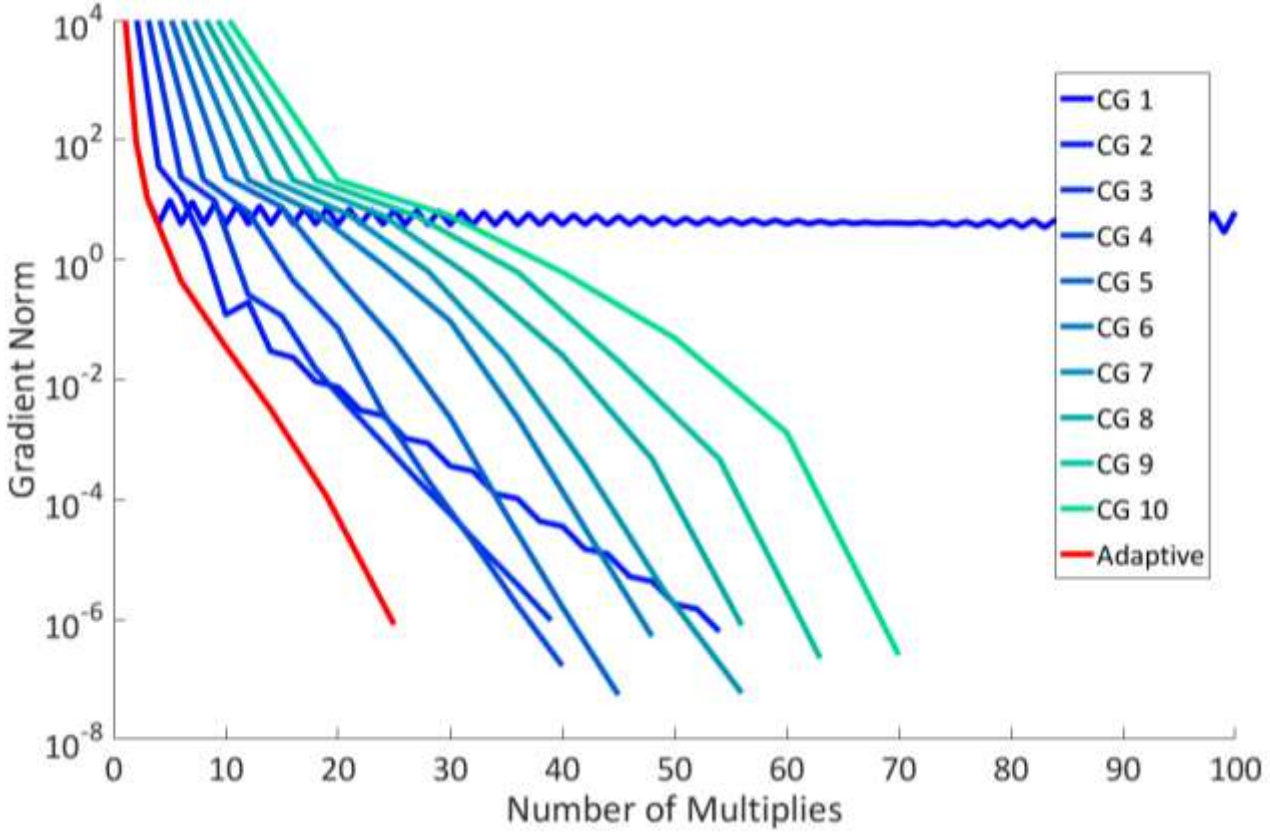
“Query Language” Optimization

- Static analysis



“Query Language” Optimization

- Static analysis
- Runtime analysis



Some Bioinformatics Applications

- Personalized medicine, Memorial Sloan Kettering Cancer Center
  - 35% accuracy improvement over state-of-the-art
- Metagenomic binning and DNA quality assessment, Stanford School of Medicine
  - Previously unsolved problem
- Toxicogenomic analysis, Stanford University
  - Improved on state-of-the-art results

Upcoming

- Massive scale character level sentiment and text analysis on Amazon data
  - Billions of features, hours to solve a model
  - Efficient multitask learning
- Characterize the global limitations of learning word structure
  - Devise provably more efficient regularizers for uncovering structure