

An Overview of Data Science

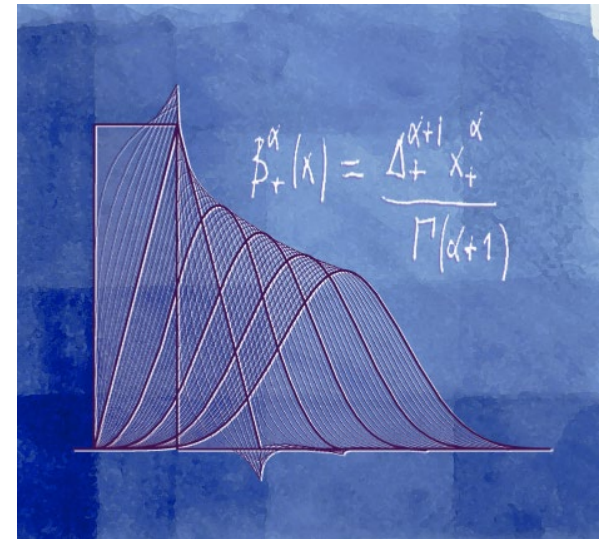
Farzad Kamalabadi^{1,2}, Matthew Grawe¹, Brian Harding³

¹Dept. of Electrical & Computer Engineering

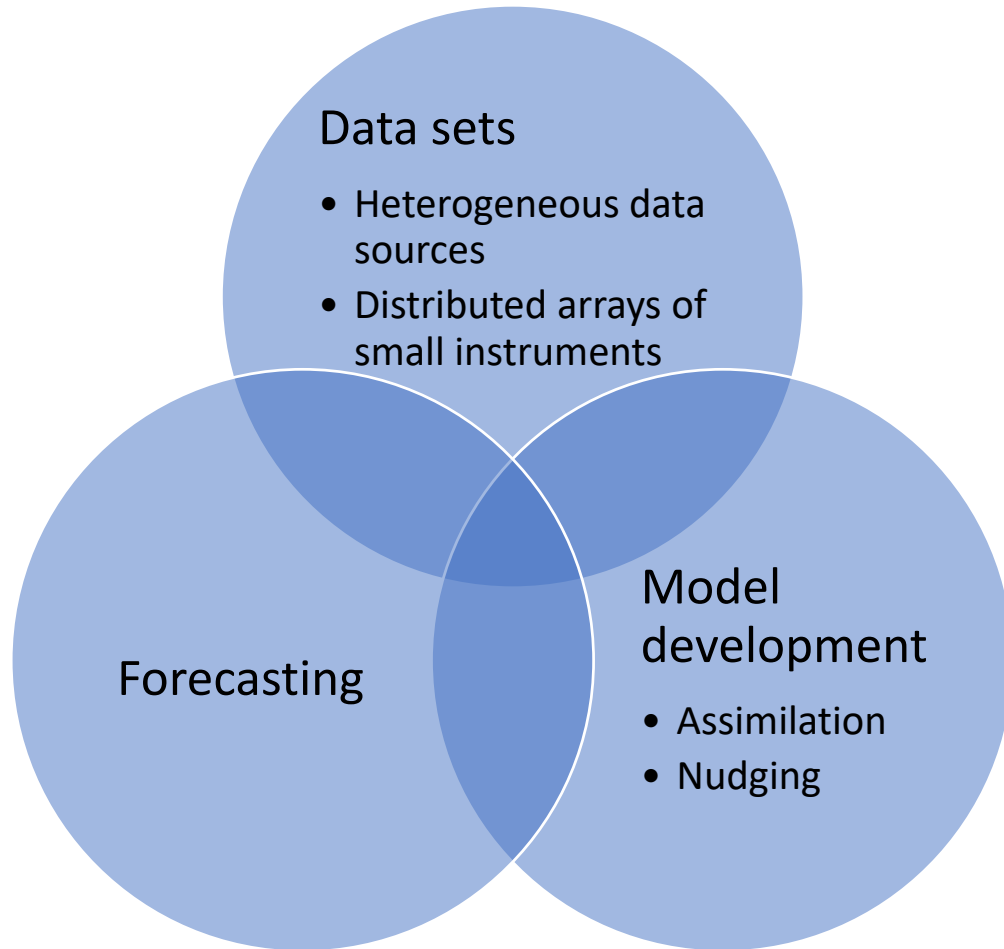
²also at Dept. of Statistics

Univ. of Illinois at Urbana-Champaign

³now at University of California, Berkeley



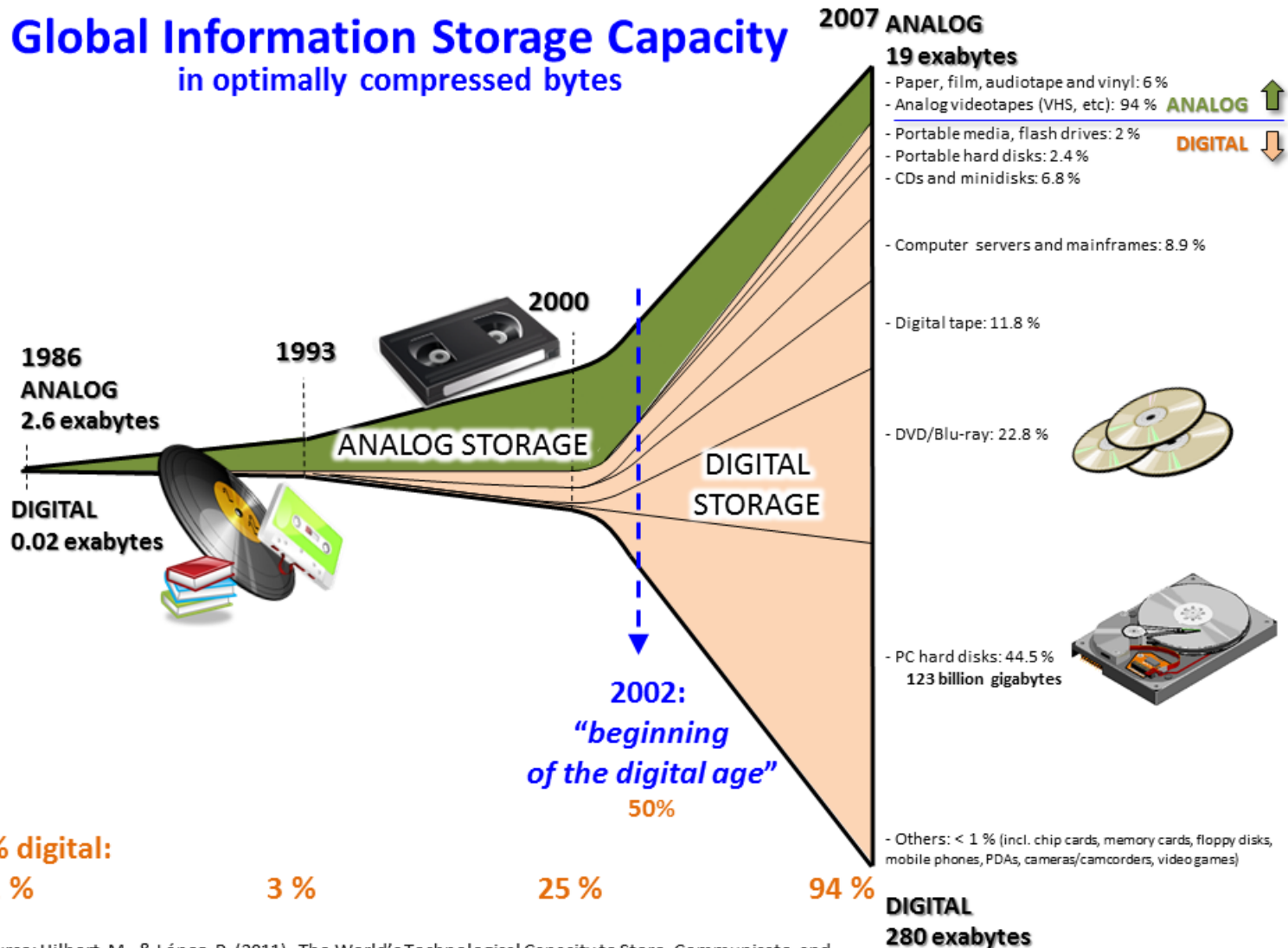
From Discovery to System Science



- CEDAR science is transitioning
- Data science fundamentals are increasingly needed
 - Enabling new science from existing data
 - Designing new sensing modalities
 - Utilizing data to make forecasts

Advent of Data Science

Global Information Storage Capacity in optimally compressed bytes



Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 | rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 | lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES \$200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some of the Nation's most pressing challenges.

To launch the initiative, six Federal departments and agencies today announced more than \$200 million in new commitments that, together, promise to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data.

"In the same way that past Federal investments in information-technology R&D led to dramatic advances in supercomputing and the creation of the Internet, the initiative we are launching today promises to transform our ability to use Big Data for scientific discovery, environmental and biomedical research, education, and national security," said Dr. John P. Holdren, Assistant to the President and Director of the White House Office of Science and Technology Policy.

To make the most of this opportunity, the White House Office of Science and Technology Policy (OSTP)—in concert with several Federal departments and agencies—created the Big Data Research and Development Initiative to:

- Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.
- Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and
- Expand the workforce needed to develop and use Big Data technologies.

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

Big Data Elements

Advance the core scientific and technological means of managing, analyzing, visualizing and extracting information from large, diverse, distributed, and heterogeneous data sets in order to accelerate progress in science and engineering research. Specifically, it includes research to develop and evaluate new algorithms, technologies, and tools for improved data management, data analytics, and e-science collaboration environments.

“In the same way that past Federal investments in information-technology R&D led to dramatic advances in supercomputing and the creation of the Internet, the initiative we are launching today promises to transform our ability to use Big Data for scientific discovery...”

Dr. John P. Holdren, Assistant to the President and Director of the White House Office of Science and Technology Policy.

Data Analytics Elements

Data to Information: powerful approaches for turning data into information – machine learning, cloud computing, and crowd sourcing.

Data to Decisions: Harness and utilize massive data in new ways and bring together sensing, perception and decision support to make truly autonomous systems that can maneuver and make decisions on their own.

Human-Computer Interaction: Developing scalable algorithms for processing imperfect data in distributed data stores; and Creating effective human-computer interaction tools for facilitating rapidly customizable visual reasoning for diverse missions.

Data Science: Data Life Cycle

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director



SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

Encourage the publication of all science products so they are discoverable and accessible, to enable reproducibility, and to ensure that they can be adapted to solve new problems.

Data Sharing and Collaboration

- Tools for distant data sharing, real-time visualization, and software reuse of complex datasets
- Cross disciplinary information and knowledge sharing; interoperability
- Remote operation and real-time access to distributed data

Collection, Storage and Management

- Data representation, storage and retrieval
- Data management policies, including access and dark data
- Communication and storage technologies with extreme capacities

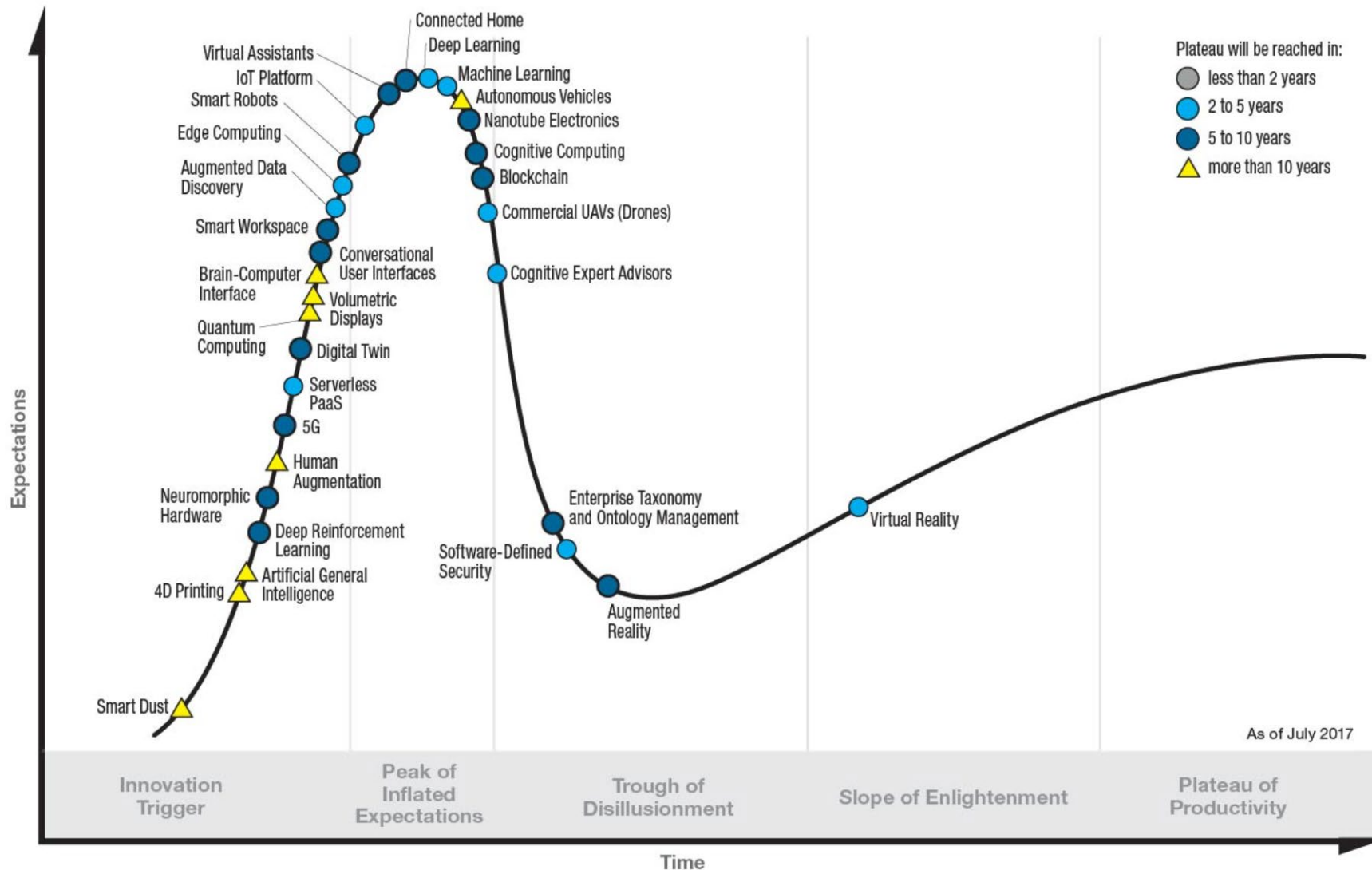
Data Life Cycle

- Computational, mathematical, statistical and algorithmic techniques for modeling high dimensional data
- Learning, inference, prediction and knowledge discovery for large volume and dynamic data
- Data mining to enable automated hypotheses, event correlation and anomaly detection

Data Analytics

Here we focus on Data Analytics

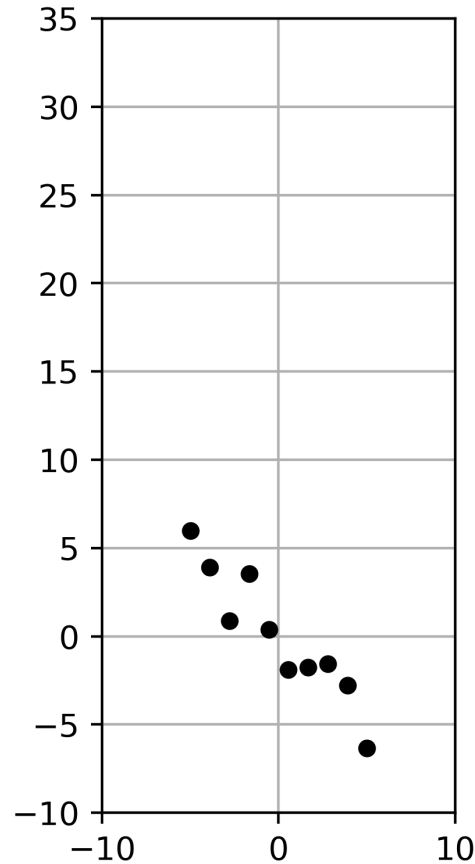
Gartner **Hype Cycle** for Emerging Technologies, 2017



Basic Elements of Learning Theory (using simple applications)

- Many core data science elements can be introduced using very simple, yet powerful, ideas.

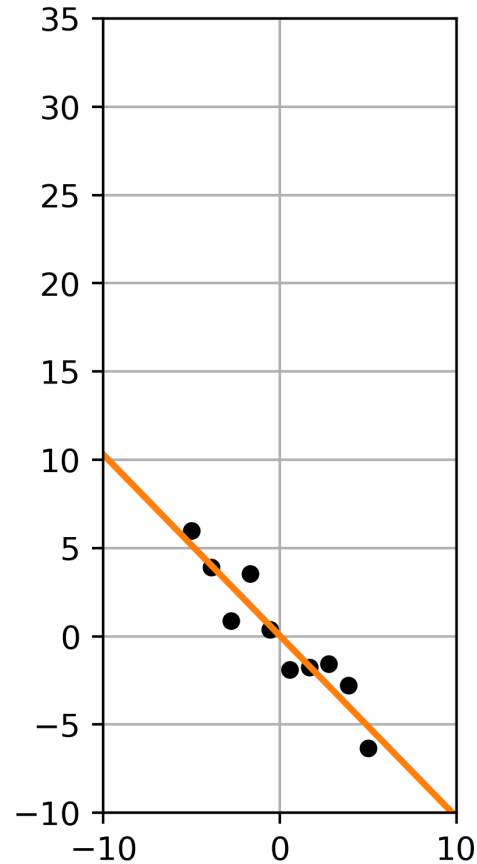
Linear Regression



- A linear relationship clearly exists. How might this be established, mathematically?
- Among all possible lines, choose the line that is the closest to the data (in some sense).

$$m^*, b^* = \operatorname{argmin}_{m,b} \sum_{i=0}^N d(m, b, x_i, y_i)$$

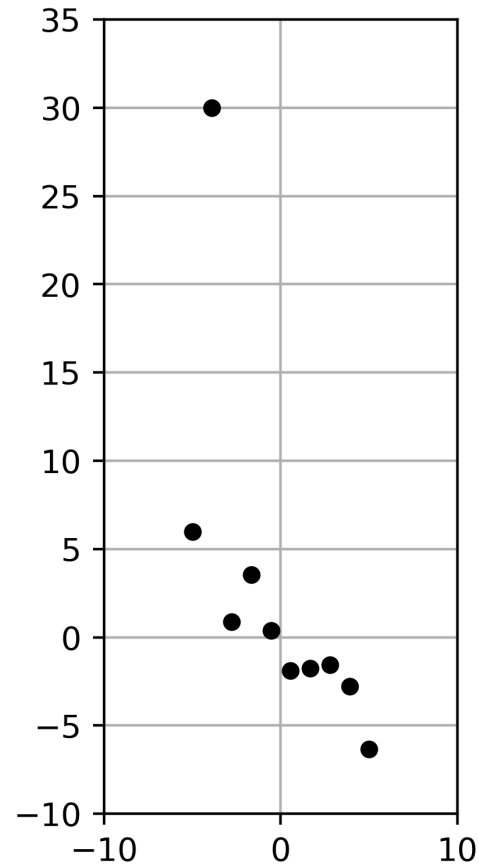
Linear Regression



- A linear relationship clearly exists. How might this be established, mathematically?
- Among all possible lines, choose the line that is the closest to the data (in some sense).

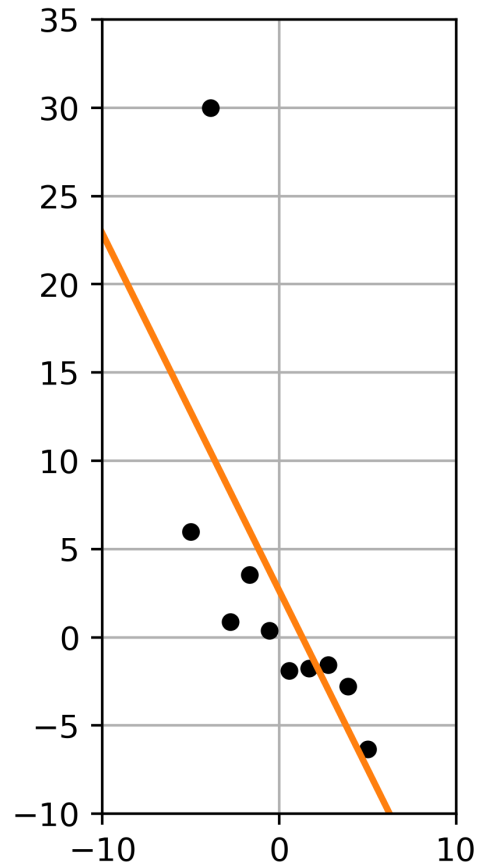
$$m^*, b^* = \operatorname{argmin}_{m,b} \sum_{i=0}^N |mx_i + b - y_i|^2$$

Linear Regression



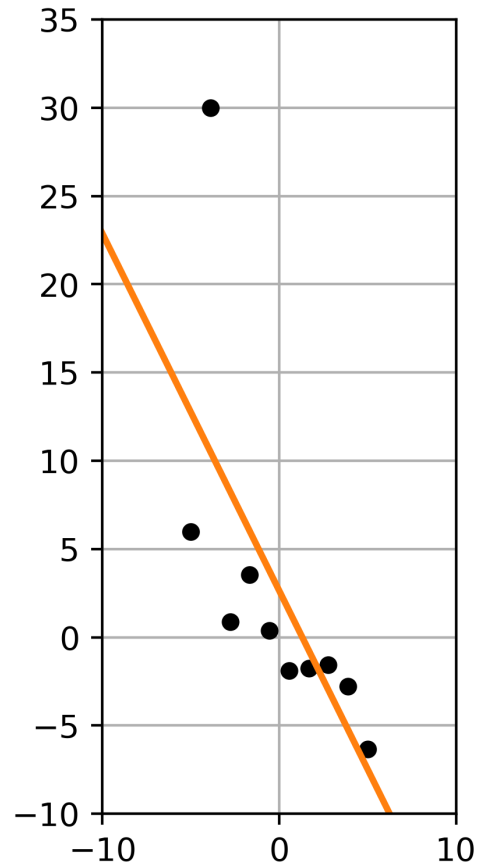
- A linear relationship clearly exists, but there is an erroneous data point (an outlier).

Linear Regression



- A linear relationship clearly exists, but there is a erroneous data point (an outlier).
- Outliers do not represent the true relationship, but change the relationship that is inferred.

Outliers



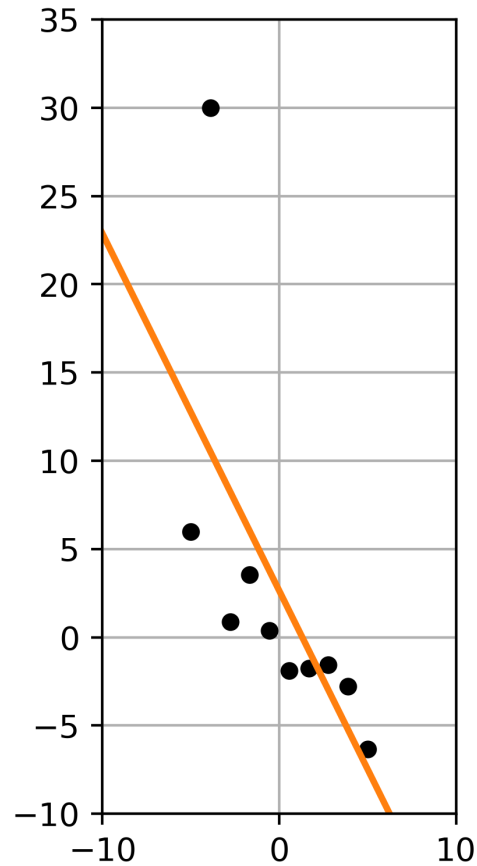
- How might we handle outliers?

We could **remove them manually**.

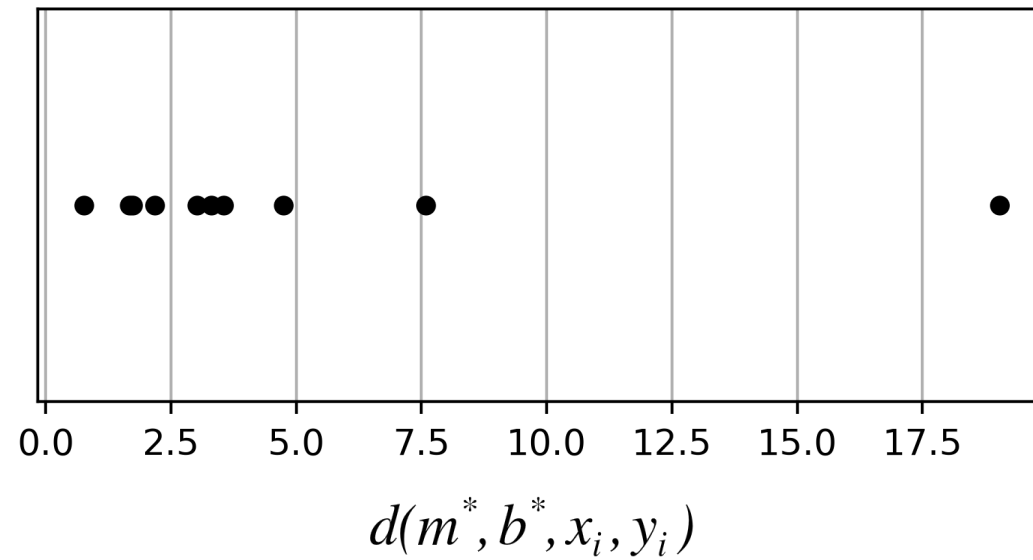
We could **explore the data for patterns** that identify an outlier boundary.
(**unsupervised learning**)

We could **train a classifier** using a set of manually-identified outliers.
(**supervised learning**)

Clustering

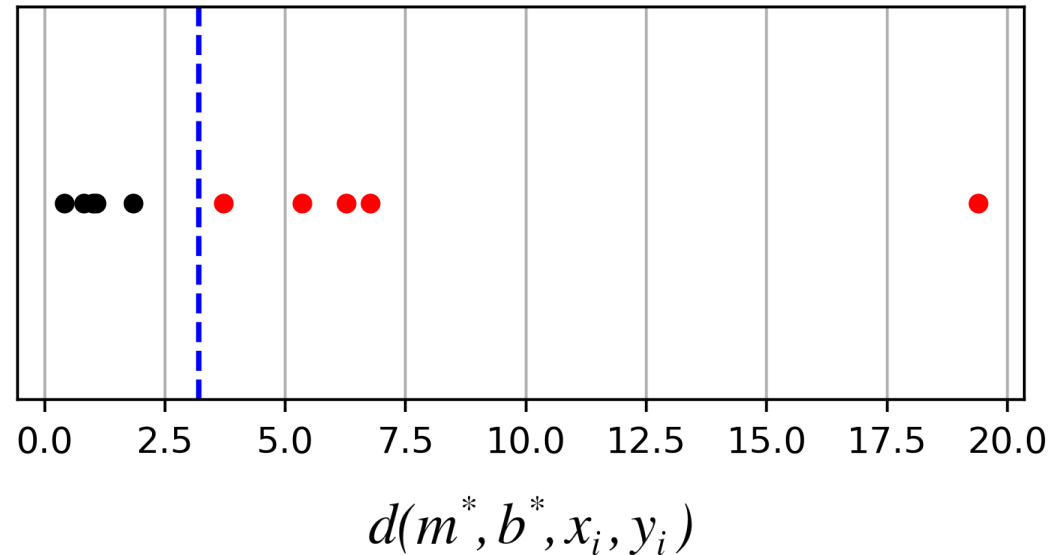


- Relative to outliers, data model errors often form a **cluster**.



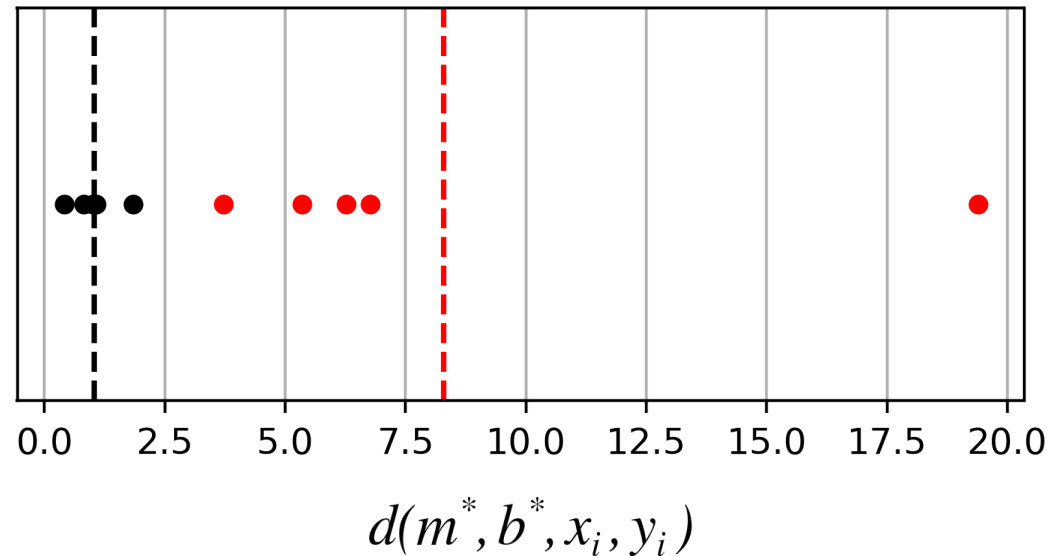
Clustering

- How might we identify the cluster? One approach:
 1. Start with a random outlier boundary.



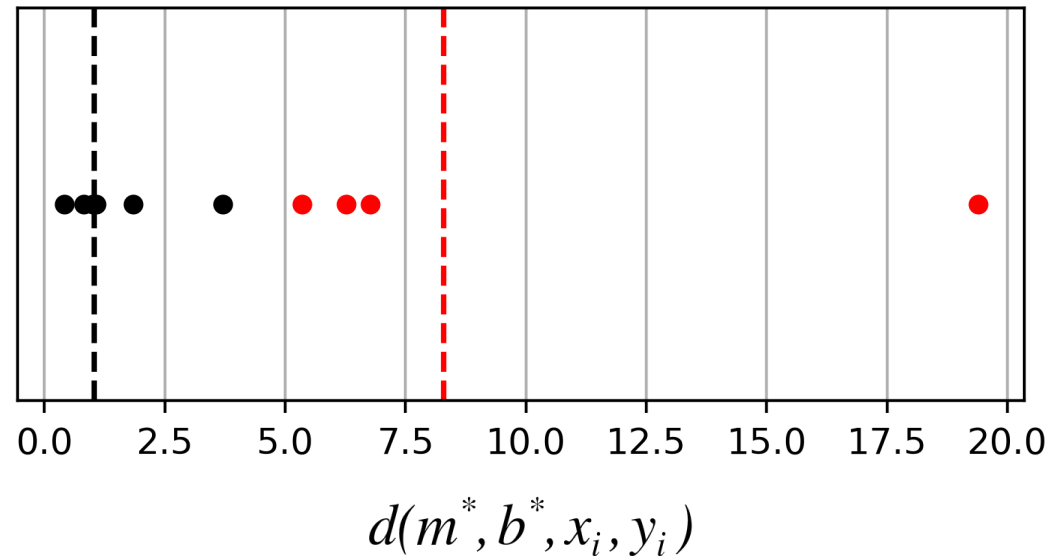
Clustering

- How might we identify the cluster? One approach:
 1. Start with a random outlier boundary.
 2. Calculate the means of the two groups (call them *clusters*).



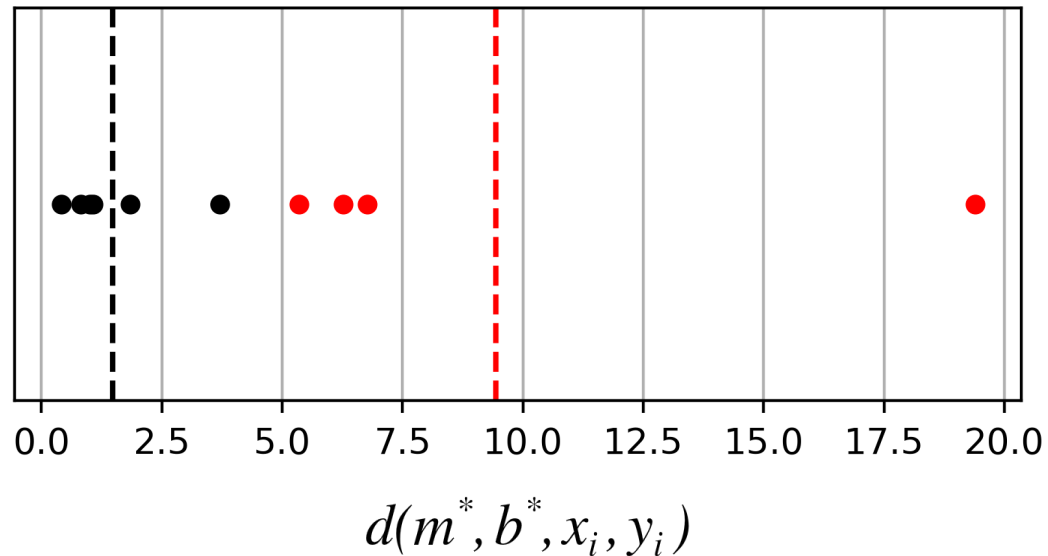
Clustering

- How might we identify the cluster? One approach:
 1. Start with a random outlier boundary.
 2. Calculate the means of the two groups (call them *clusters*).
 3. Assign each data point to the nearest mean.



Clustering

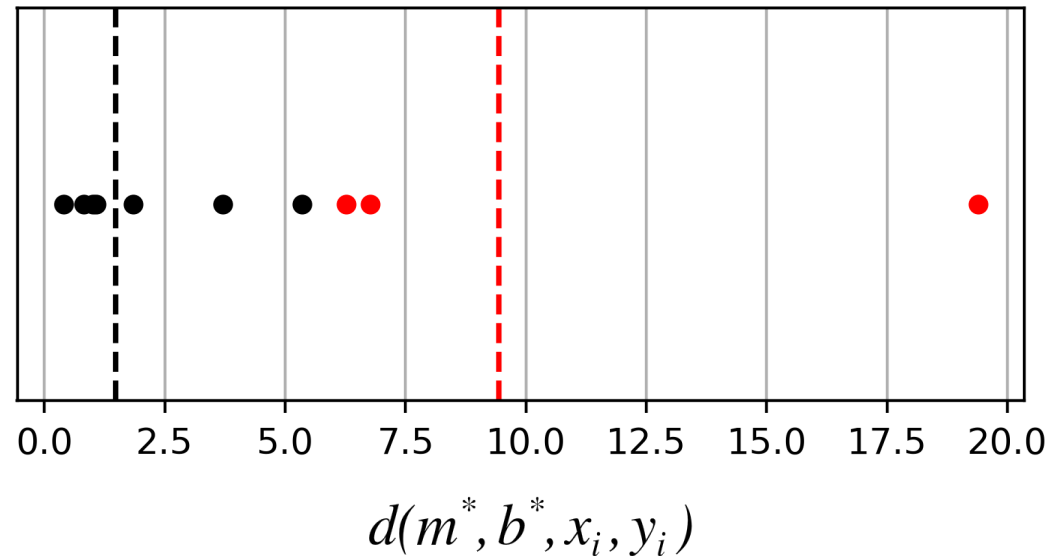
- How might we identify the cluster? One approach:
 1. Start with a random outlier boundary.
 2. Calculate the means of the two groups (call them *clusters*).
 3. Assign each data point to the nearest mean.



Repeat the process until memberships stop changing.

Clustering

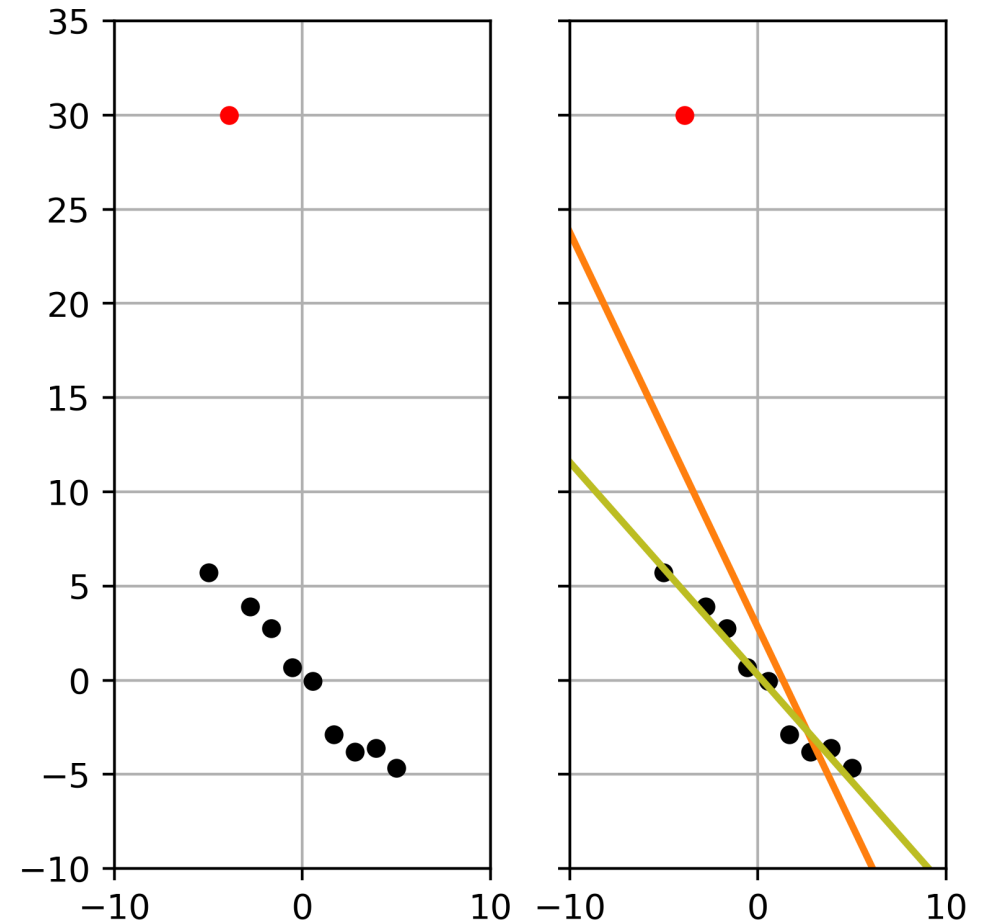
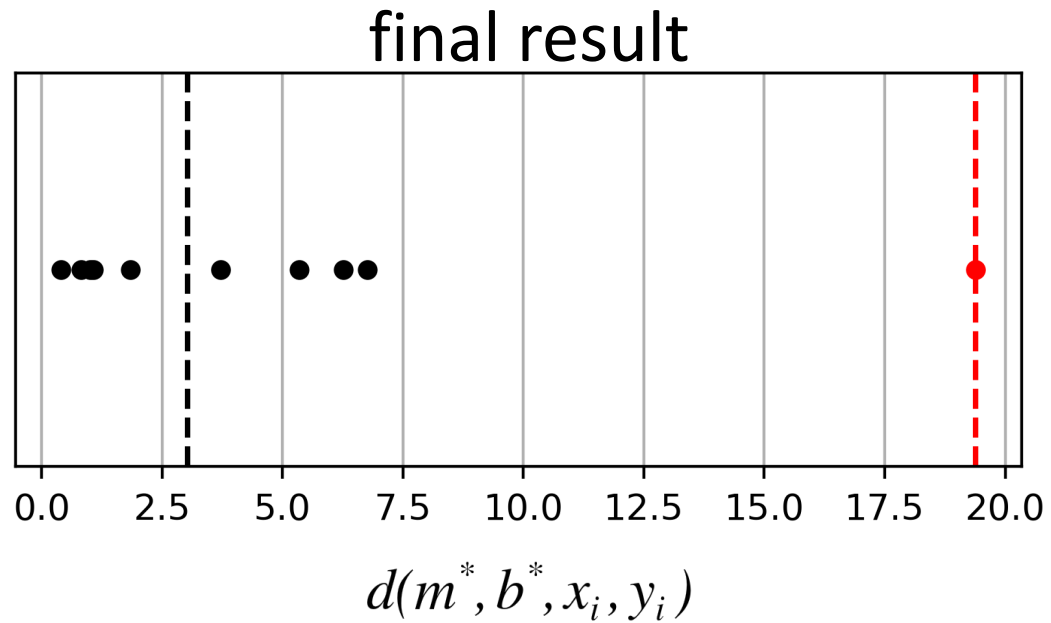
- How might we identify the cluster? One approach:
 1. Start with a random outlier boundary.
 2. Calculate the means of the two groups (call them *clusters*).
 3. Assign each data point to the nearest mean.



Repeat the process until memberships stop changing.

Clustering

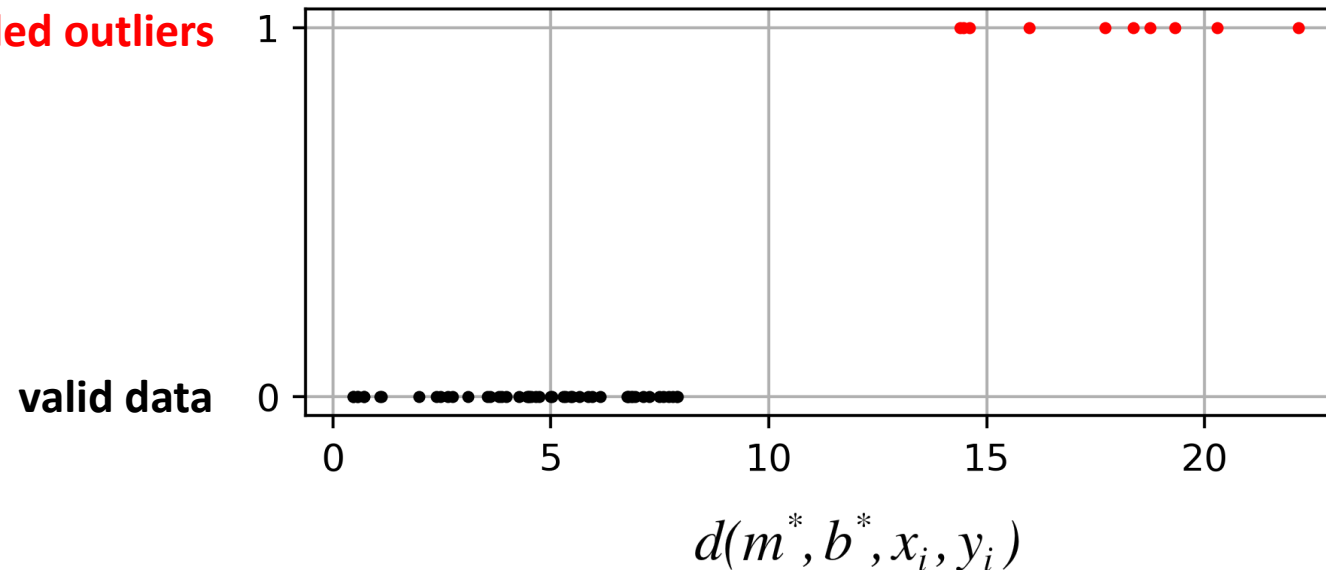
- This is known as **1D k-means clustering**.



Binary Classification

- If we have a collection of manually-identified outliers, we could **infer** an outlier boundary using the entire collection to **predict** the validity of new data.

manually-labeled outliers



Binary Classification

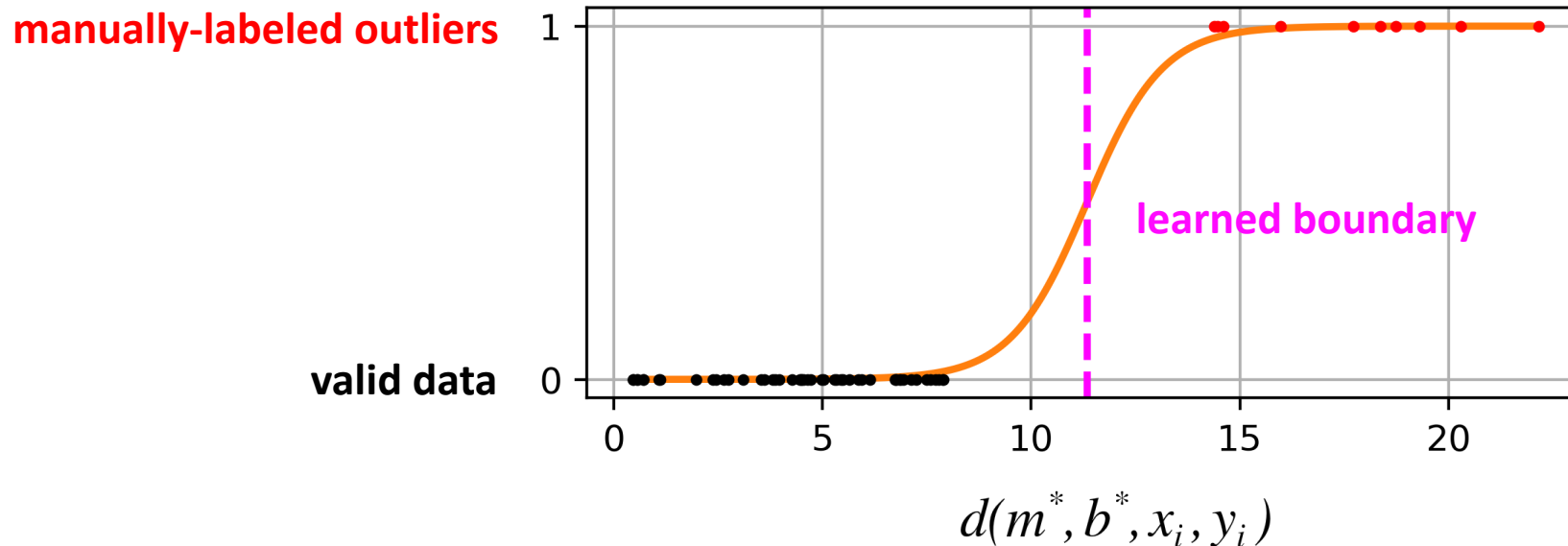
- One approach: assume that outliers occur as the result of weighted coin tosses.
 - Parameterize the coin weight and choose the boundary that maximizes the probability of the entire dataset occurring (“**maximum likelihood**”).

$$m_c^*, b_c^* = \operatorname{argmax}_{m, b} \prod_{i=0}^N p(y_i \mid d(m^*, b^*, x_i, y_i), m_c, b_c)$$

Bernoulli distribution

Binary Classification

- One approach: assume that outliers occur as the result of weighted coin tosses.
 - Parameterize the coin weight and choose the boundary that maximizes the probability of the entire dataset occurring. (“**maximum likelihood**”).

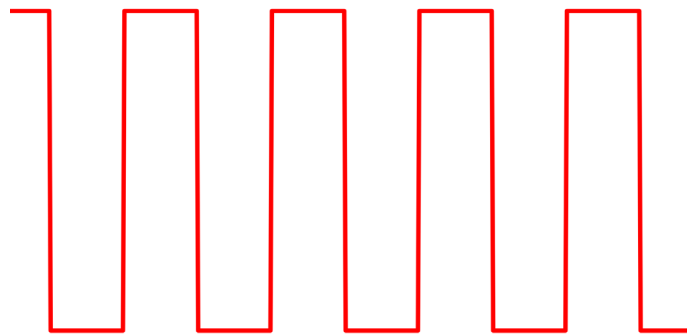


Dimensionality Reduction

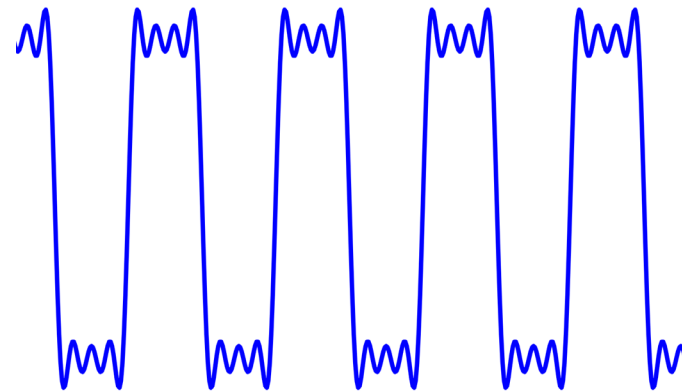
Dimensionality Reduction—Familiar Example

- Data can be expanded using a **fixed** basis (e.g., Fourier series; $\omega = n\omega_0$).

$$f(t) = \frac{A}{2} + \frac{4A}{\pi} \sum_{n=1}^{\infty} \frac{\sin((2n-1)\omega_0 t)}{2n-1} \approx \frac{A}{2} + \frac{4A}{\pi} \sum_{n=1}^k \frac{\sin((2n-1)\omega_0 t)}{2n-1}$$

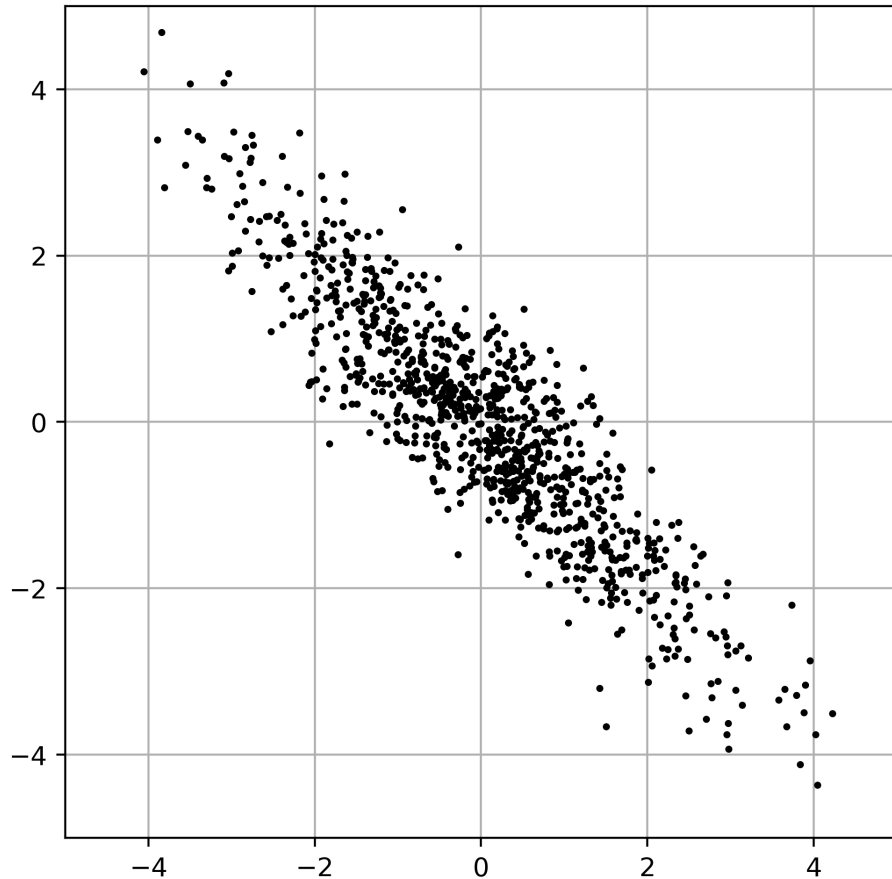


(Fourier series expansion)



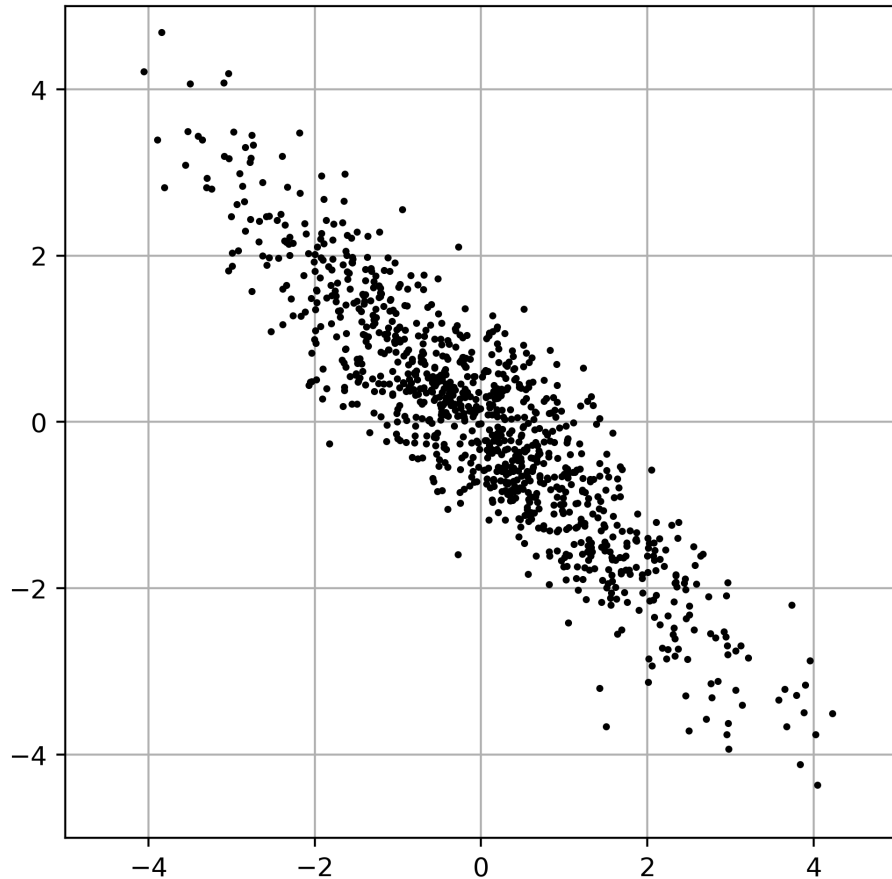
(low-dimension approximation)

Dimensionality Reduction—More General Ex.



- Does this data exist across two dimensions?
 - Technically, yes.
 - Practically...?
- How might we assess the *true* dimensionality of the dataset?

Dimensionality Reduction



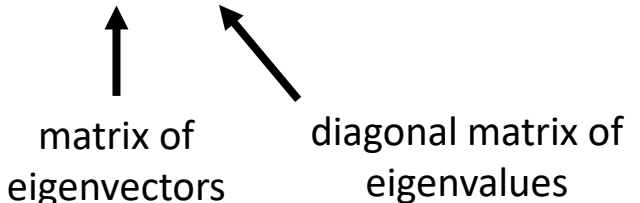
- One possible approach:
 - Find the rotational change of basis that best explains the dataset variance.

$$\tilde{\Sigma} = \frac{XX^T}{N-1} \quad (\text{sample covariance})$$

2 x N matrix of data

Dimensionality Reduction

- The eigenvalues and eigenvectors of the sample covariance describe the appropriate change of basis.

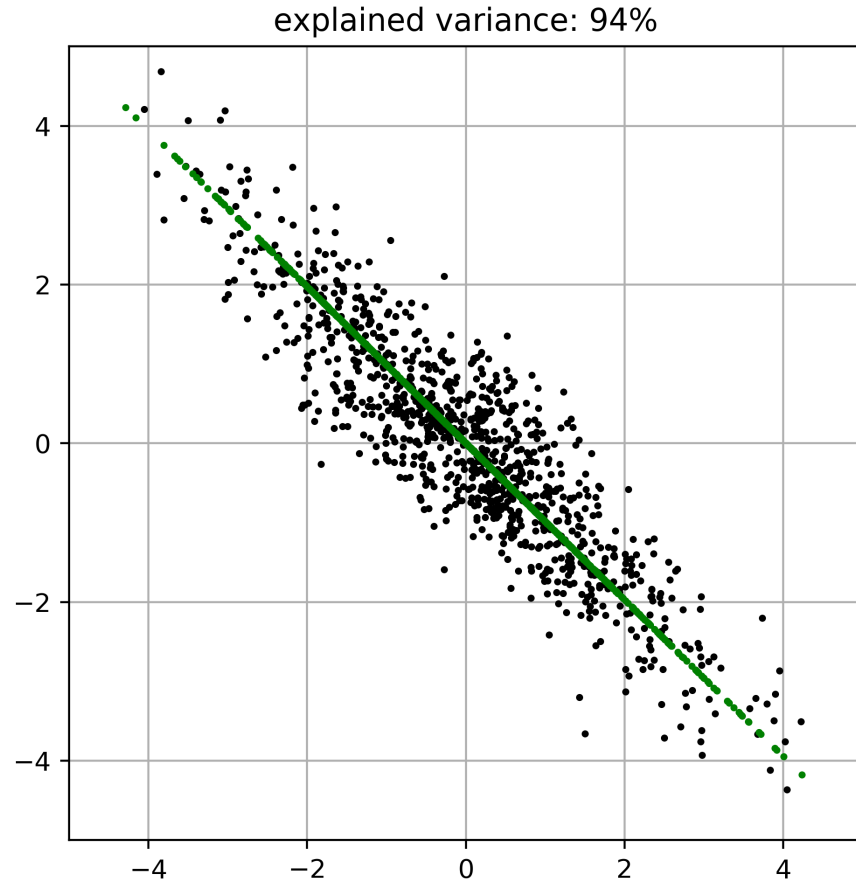
$$\tilde{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \quad (\text{diagonalized sample covariance})$$


matrix of
eigenvectors

diagonal matrix of
eigenvalues

- What if we project onto the direction of the eigenvector with the largest eigenvalue?

Dimensionality Reduction



- In this example, 94% of the dataset variance lies in a one-dimensional subspace.
- The data is “almost” one-dimensional!
- This is known as **principal component analysis**.

Dimensionality Reduction

- Principal component analysis learns a basis for the data that is **adaptive**.
 - This is directly related to the singular value decomposition (SVD) of the data matrix.

$$\tilde{\Sigma} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$$

$$\mathbf{X} = \sqrt{N-1} \sum_{i=1}^d \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^T \stackrel{k \leq d}{\approx} \sqrt{N-1} \sum_{i=1}^k \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^T$$

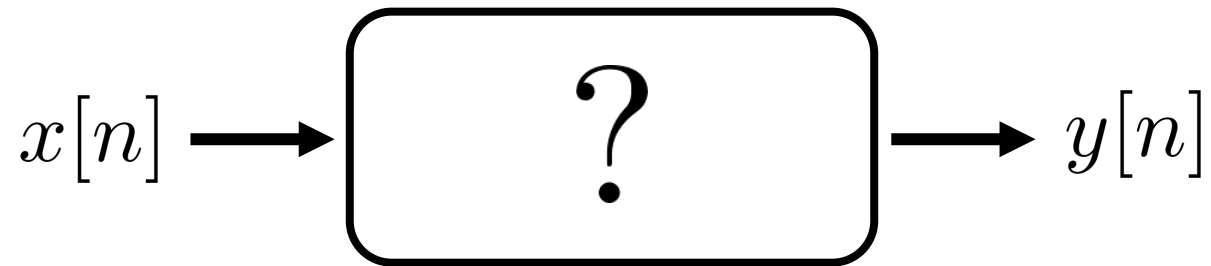
(singular value decomposition)

(low-rank approximation)

A system Identification Perspective of Learning Theory

System Identification

- If input and output data from an unknown system is available, how can we “discover” information about the system?



Linear, Time-Invariant System Identification

- Relationships in the sciences are often described by linear differential equations (e.g., Maxwell's equations).
- In discrete-time (i.e., in data space), these relationships are described using difference equations.

$$y[n] + 2y[n - 1] = 3x[n] + x[n - 2]$$

Linear, Time-Invariant System Identification

- A general version: **rational transfer function models**

$$A(q)y[n] = \frac{B(q)}{F(q)}x[n] + \frac{C(q)}{D(q)}e[n]$$

Linear, Time-Invariant System Identification

- A more workable class of models: **rational transfer function models**

$$A(q)y[n] = \frac{B(q)}{F(q)}x[n] + \frac{C(q)}{D(q)}e[n]$$

- A, B, C, D, F are **lag polynomials**, e.g., $A(q) = 1 + a_1q^{-1} + \dots + a_{n_a}q^{-n_a}$

Linear, Time-Invariant System Identification

- A more workable class of models: **rational transfer function models**

$$A(q)y[n] = \frac{B(q)}{F(q)}x[n] + \frac{C(q)}{D(q)}e[n]$$

- A, B, C, D, F are **lag polynomials**, e.g., $A(q) = 1 + a_1q^{-1} + \dots + a_{n_a}q^{-n_a}$
- The system is defined by the weights on past samples of the input, output, and noise.

$$\boldsymbol{\theta} = [a_1 \ a_2 \ \dots \ a_{n_a} \ b_1 \ b_2 \ \dots \ b_{n_b} \ f_1 \ f_2 \ \dots \ f_{n_f} \ c_1 \ c_2 \ \dots \ c_{n_c} \ d_1 \ d_2 \ \dots \ d_{n_d}]^T$$

Linear, Time-Invariant System Identification

- Many well-known linear system models fall into this category, depending on which polynomials are used.
 - $B(q)$: finite-impulse response (**FIR**)
 - $A(q)$: autoregressive (**AR**)
 - $C(q)$: moving average (**MA**)
 - $A(q), C(q)$: autoregressive moving average (**ARMA**)
 - $A(q), B(q)$: autoregressive w/ exogenous input (**ARX**)
 - $A(q), B(q), C(q)$: autoregressive moving average w/ exogenous input (**ARMAX**)
 - $B(q), F(q)$: output error (**OE**)
 - $B(q), F(q), C(q), D(q)$: Box-Jenkins (**BJ**)

Linear, Time-Invariant System Identification

- Given input and output data, how might we estimate the system, or, equivalently, estimate the parameter vector θ ?

Linear, Time-Invariant System Identification

- Given input and output data, how might we estimate the system, or, equivalently, estimate the parameter vector $\boldsymbol{\theta}$?
 - One approach: choose $\boldsymbol{\theta}$ that leads to the smallest (in some sense) **one-step prediction error**

$$\hat{y}[n|n-1, \boldsymbol{\theta}] = \left[1 - \frac{D(q)A(q)}{C(q)} \right] y[n] + \frac{D(q)B(q)}{C(q)F(q)} x[n]$$

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{n=1}^N f(y[n] - \hat{y}[n|n-1, \boldsymbol{\theta}])$$

e.g., $f = x^2$ (minimization in the least squares sense)

Nonlinear, Time-Varying Systems

- Systems often exhibit nonlinear behavior, and may change over time.

$$\hat{y}[n|n-1, \boldsymbol{\theta}] = g(\phi[n], \boldsymbol{\theta})$$

fixed window of past input and output data

- The general approach is the same as the LTI case, but the functional form of the one-step prediction error is more general.

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{n=1}^N f(y[n] - \hat{y}[n|n-1, \boldsymbol{\theta}])$$

Nonlinear, Time-Varying Systems

- Common approach: expand the mapping using a basis

$$g(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{k=1}^N \alpha_k g_k(\boldsymbol{\phi}, \mathbf{p})$$

$$\boldsymbol{\theta} = [\alpha_1 \ \alpha_2 \ \dots \ \alpha_n \ p_1 \ p_2 \ \dots \ p_n]^T$$

- Examples:
 - Wavelet expansions (g_k are then dilated and scaled versions of a “mother” basis function)
 - Sigmoid, tanh, Gaussian functions

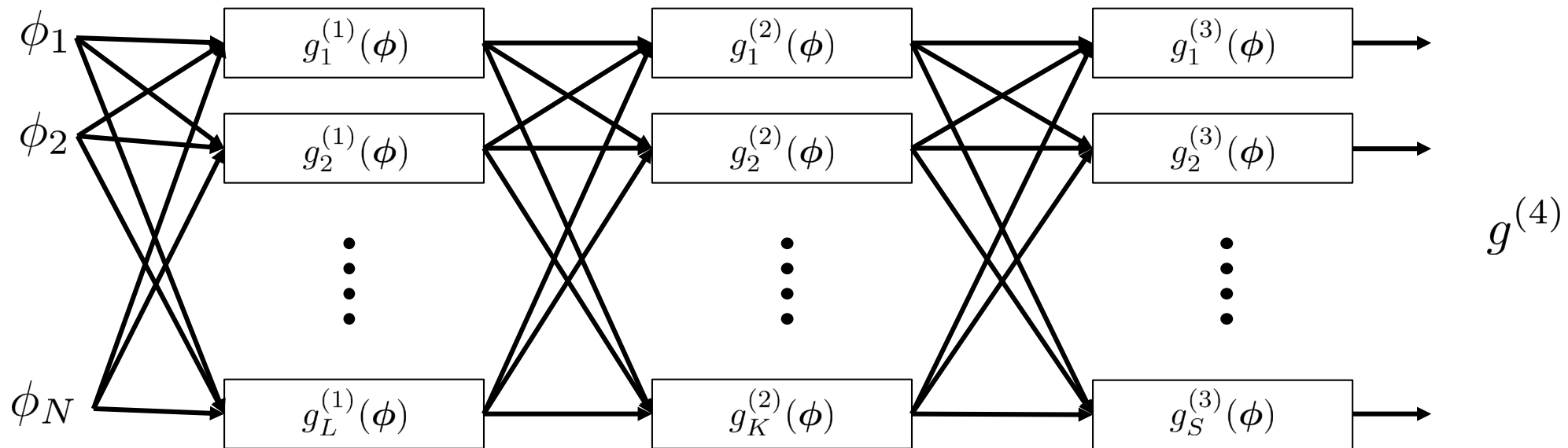
Nonlinear, Time-Varying Systems

- Layered/composed expansions are **neural networks**.

$$\begin{aligned}g_k^{(2)}(\phi) &= \sum_l \alpha_l^{(2)} \kappa(\phi^{(2)}, \beta_l^{(2)}, \gamma_l^{(2)}) & \phi_k^{(2)} &= g_k(\phi) \\g_k^{(3)}(\phi) &= \sum_l \alpha_l^{(3)} \kappa(\phi^{(3)}, \beta_l^{(3)}, \gamma_l^{(3)}) & \phi_k^{(3)} &= g_k(\phi^{(2)}) \\&\vdots & & \\g_k^{(M)}(\phi) &= \sum_l \alpha_l^{(M)} \kappa(\phi^{(M)}, \beta_l^{(M)}, \gamma_l^{(M)}) & \phi_k^{(M)} &= g_k(\phi^{(M-1)})\end{aligned}$$

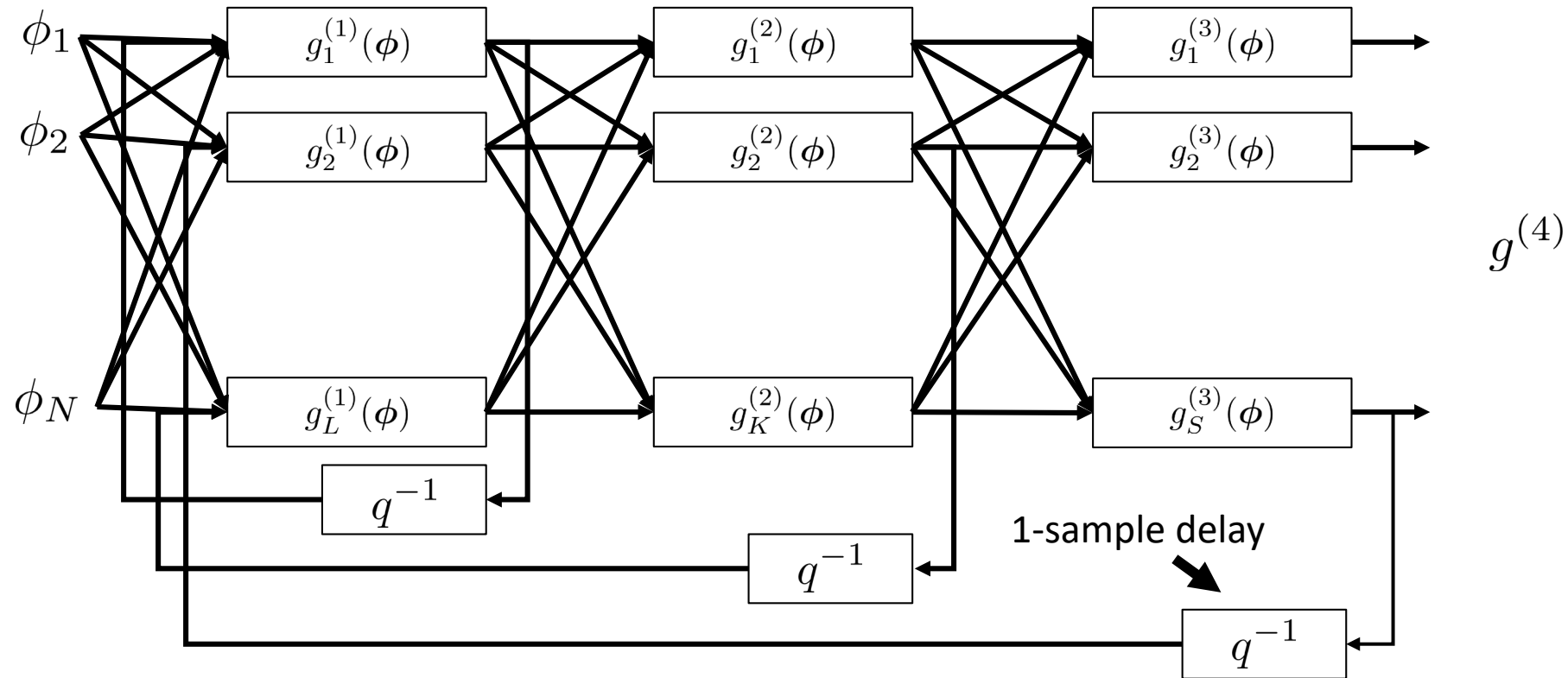
Nonlinear, Time-Varying Systems

- Layered/composed expansions are **neural networks**.



Nonlinear, Time-Varying Systems

- Time-varying systems can be described using **recurrent** networks.



Learning Theory Caveats and Open Directions

- With nonlinear systems, cost function minimization presents special challenges.
 - Nonlinear cost functions are usually non-convex, and have many local minima.
- Solutions for θ that have the lowest minimization error do not necessarily perform well on new data (poor **generalization error**).
- Understanding the behavior of generalization error in different situations is currently a very active topic of research in machine learning and data science.
 - Validation data sets are critical