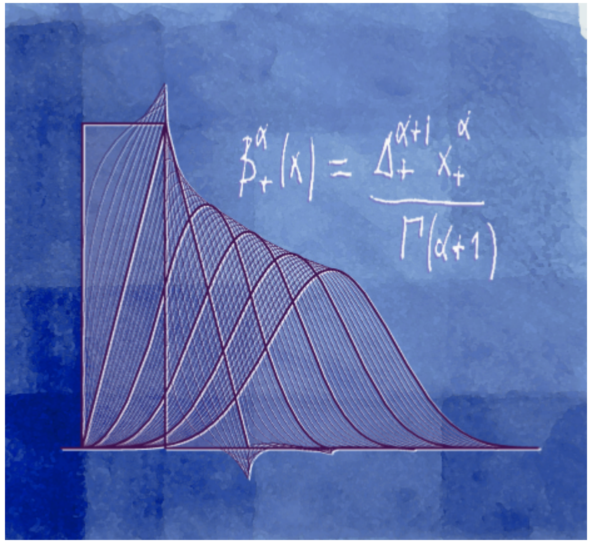
Data Science and its Evolution in CEDAR

Farzad Kamalabadi[*] and Matt Grawe

Dept. of Electrical & Computer Engineering

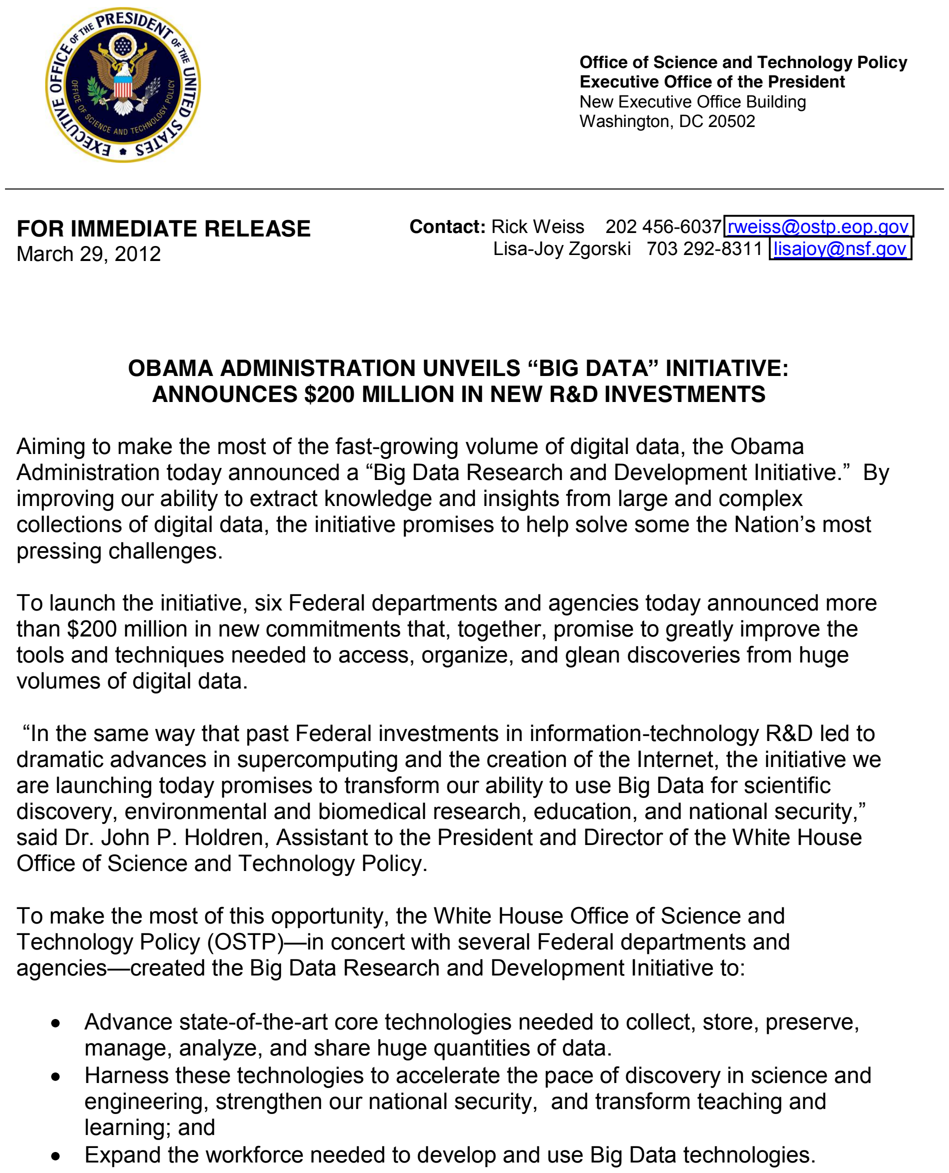
[*]Also at Dept. of Statistics

Univ. of Illinois at Urbana-Champaign


$\beta_+^\alpha(x) = \frac{\Delta_+^{\alpha+1} x_+^\alpha}{\Gamma(\alpha+1)}$

NSF CEDAR Workshop—Geospace Data Science: June 20, 2019

Advent of Data Science



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 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.

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.

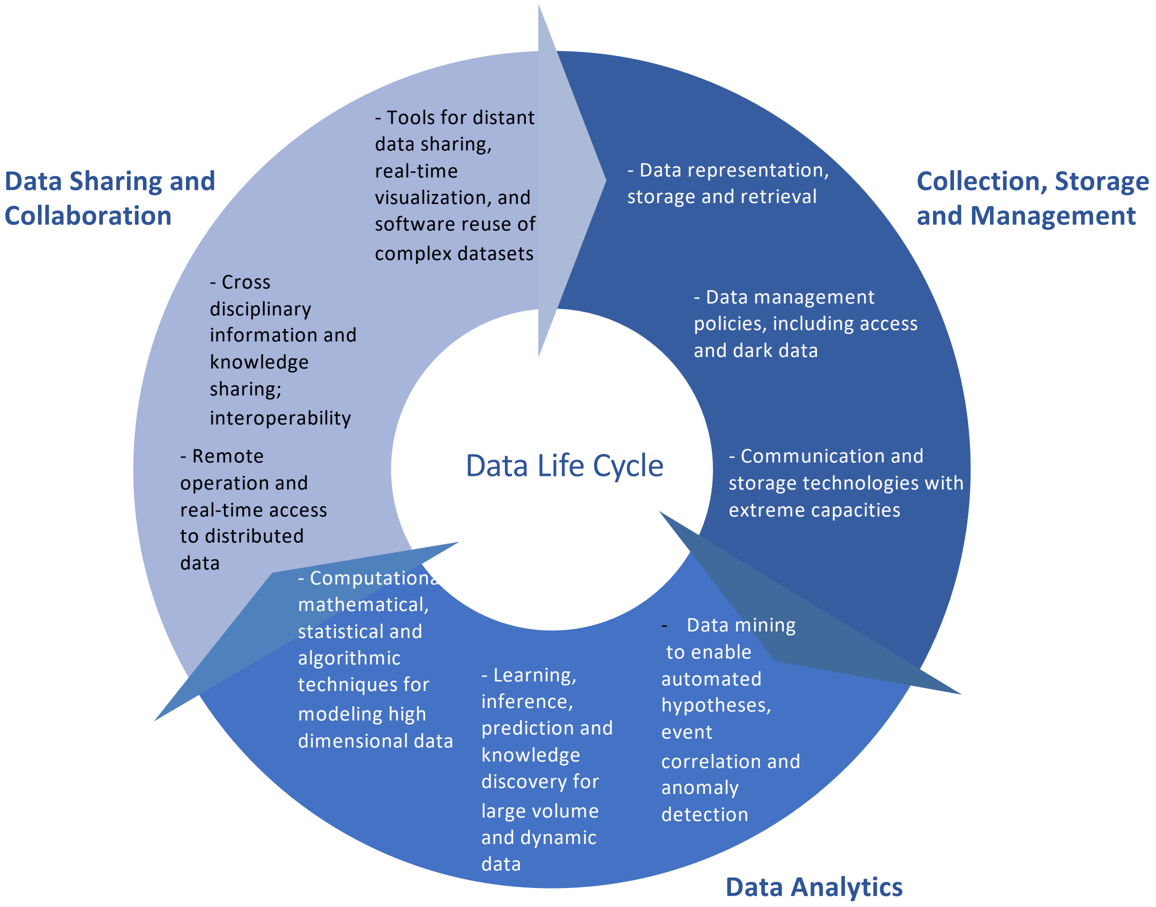
Data Analytics Elements

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



Data Life Cycle

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 Analytics

- Computationa 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

Here we focus on Data Analytics

State/Space Abstraction

Observation:

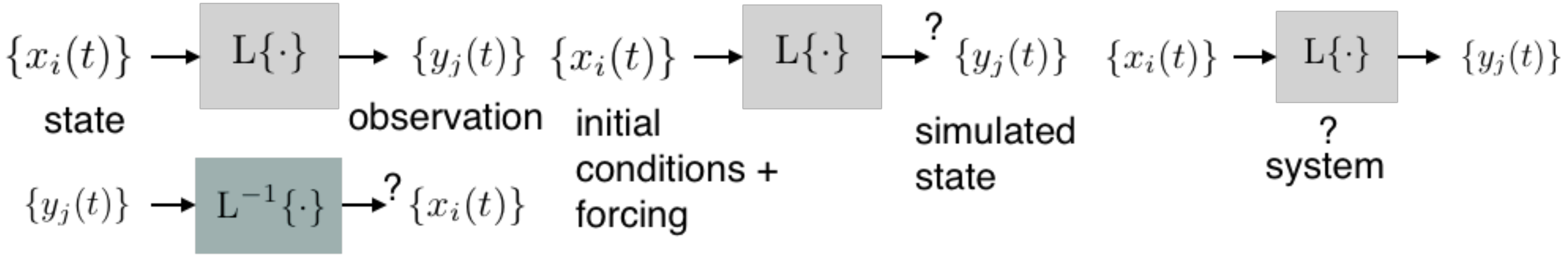
- goal: given a set of observations and a forward model relating the state, e.g., temp, density, composition, electric/mag field to the observables, determine the state parameters;
- challenges: observability, invertibility

Simulation:

- goal: given initial conditions and forcing parameters, generate simulated state parameters
- challenges: drivers often need to be (inferred) estimated from observation

Learning:

- goal: given a set of observations and the corresponding state parameters, learn the system (forward model)--system identification
- challenges: need sufficient and reliable observations



$\{x_i(t)\} \rightarrow \boxed{\mathrm{L}\{\cdot\}} \rightarrow \{y_j(t)\}$

state

observation

$\{y_j(t)\} \rightarrow \boxed{\mathrm{L}^{-1}\{\cdot\}} \xrightarrow{?} \{x_i(t)\}$

$\{x_i(t)\} \rightarrow \boxed{\mathrm{L}\{\cdot\}} \xrightarrow{?} \{y_j(t)\}$

initial conditions + forcing

simulated state

$\{x_i(t)\} \rightarrow \boxed{\mathrm{L}\{\cdot\}} \rightarrow \{y_j(t)\}$

?

system

- Task of inference: given 2 entities in this triplet, estimate (infer statistically) x the third. Techniques for accomplishing this task have been in development for a century and continue to gain sophistication.

CEDAR Examples: Data Assimilation

- Goal: For many dynamical systems, we want to estimate the (often global) physical state (e.g., magnetic field, density, temperature) from limited observations as best as we can.

- If sufficient sampling of observations are available, the task can be approached as:
  - a non-parametric time-dependent interpolation/extrapolation problem with unknown (unspecified) state evolution/dynamics, e.g., tomography.
  - a parametric (time-dependent) state estimation problem, e.g., spherical harmonics from sampled data.

If, on the other hand, observations are sparse and insufficient to produce a global specification, a forward model (time-dependent simulation) can be used as a starting point to recursively ingest the available incoming data and produce a more realistic specification.

Statistical Estimation: Dynamic Model

General State-Space Signal Model

The general hidden Markov model (HMM):

$$\begin{aligned} \text{Initial prior:} & \quad p_{\boldsymbol{x}_1}(\boldsymbol{x}_1) \qquad (1) \\ \text{Measurement/forward model:} & \quad h_i(\boldsymbol{y}_i|\boldsymbol{x_i}) \qquad (2) \\ \text{State-transition model:} & \quad f_i(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i) \qquad (3) \end{aligned}$$

$$\text{dim}(\boldsymbol{x}_i) = N \qquad \text{dim}(\boldsymbol{y}_i) = M$$

Goal: Compute minimum mean square error (MMSE) estimates of the unknown state $\boldsymbol{x}_i$ given the measurements $\boldsymbol{y}_{1:j} \triangleq \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_j\}$.

$$\widehat{\boldsymbol{x}}_{i|j} \triangleq \mathbb{E}[\boldsymbol{x}_i|\boldsymbol{y}_{1:j}] = \int \boldsymbol{x}_i \, p(\boldsymbol{x}_i|\boldsymbol{y}_{1:j}) \, d\boldsymbol{x}_i \tag{4}$$

Statistical Estimation: Dynamic Model

Linear Additive-Noise State-Space Signal Model (Linear Gaussian Model)

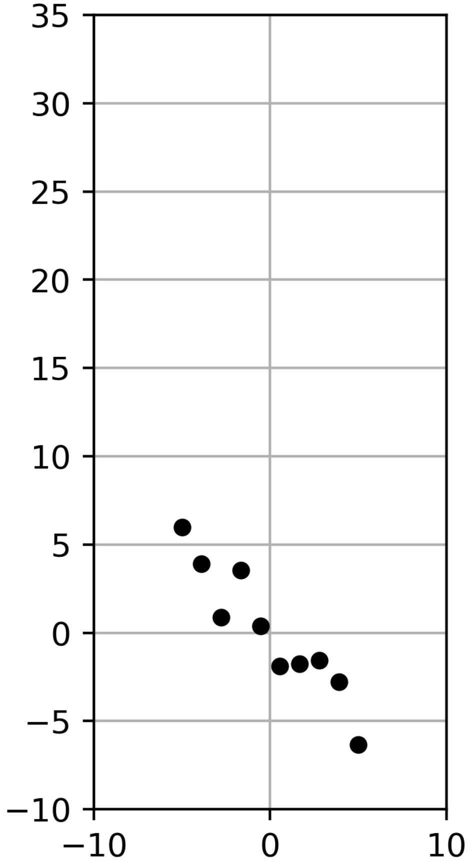
$$\begin{aligned} \text{Initial prior:} \quad & \mathbb{E}[\boldsymbol{x}_1] = \boldsymbol{\mu}_1, \ \text{Cov}(\boldsymbol{x}_1) = \boldsymbol{\Pi}_1 && (5) \\ \text{Measurement/forward model:} \quad & \boldsymbol{y}_i = \boldsymbol{H}_i \, \boldsymbol{x}_i + \boldsymbol{v}_i && (6) \\ \text{State-transition model:} \quad & \boldsymbol{x}_{i+1} = \boldsymbol{F}_i \, \boldsymbol{x}_i + \boldsymbol{u}_i && (7) \end{aligned}$$

- The first and second order statistics of the zero mean state $(\boldsymbol{u}_i)$ and measurement $(\boldsymbol{v}_i)$ noise are given: $\text{Cov}(\boldsymbol{u}_i) = \boldsymbol{Q}_i$ and $\text{Cov}(\boldsymbol{v}_i) = \boldsymbol{R}_i$.

Goal: Compute linear minimum mean square error (LMMSE) estimates of the unknown state $\boldsymbol{x}_i$ given the measurements $\boldsymbol{y}_{1:j}$.

Basic Elements of Learning Theory

Linear Regression

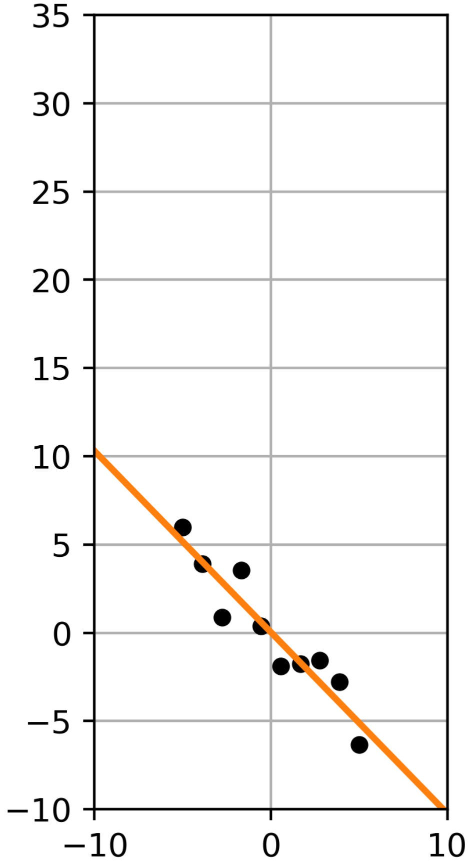


- A linear relationship clearly exists. How might this be established, mathematically?

- Among all possible lines, choose the line that is the closest to the data (in some sense).

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

Linear Regression

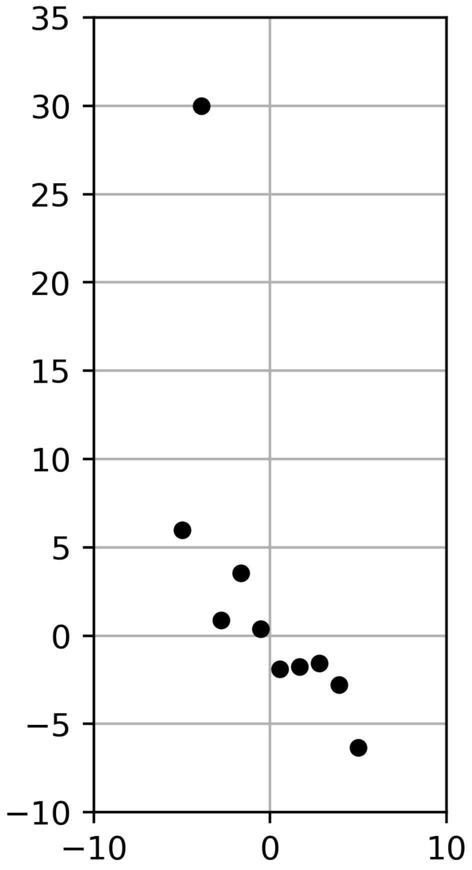


- A linear relationship clearly exists. How might this be established, mathematically?

- Among all possible lines, choose the line that is the closest to the data (in some sense).

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

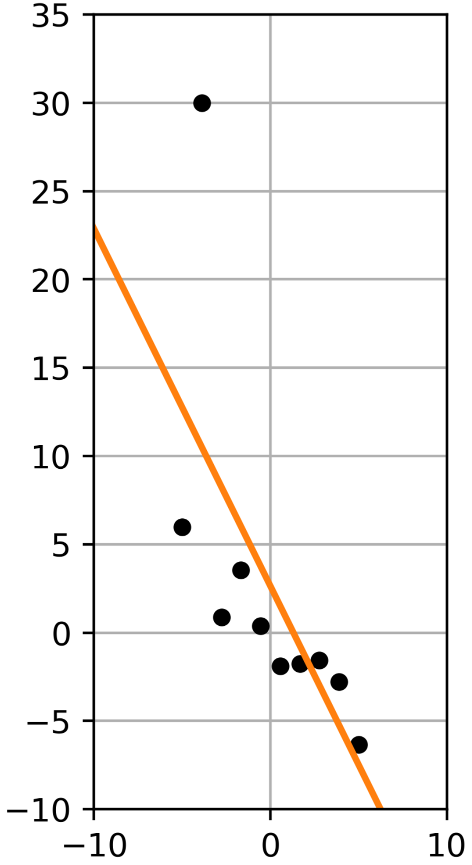
Linear Regression: Robust Statistics



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

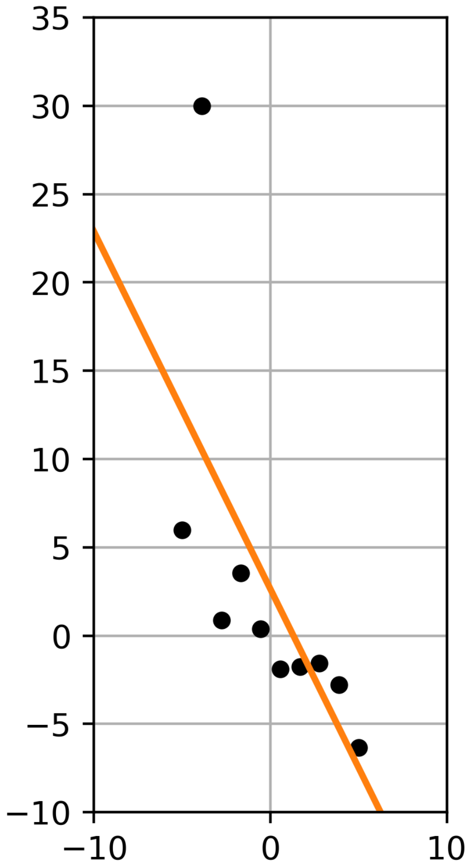
- Robust statistical estimation; Regularization
- Kamalabadi, 1999;
- Kamalabadi et al., 2002

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



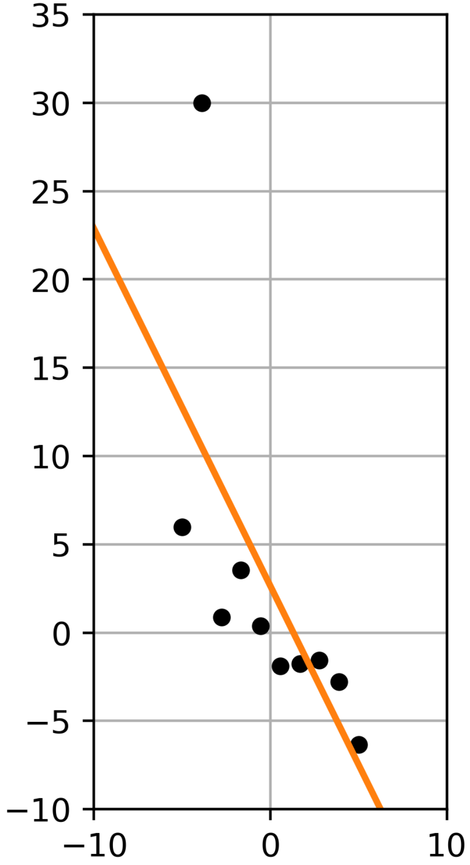
- How might we handle outliers?

We could **remove them manually.**

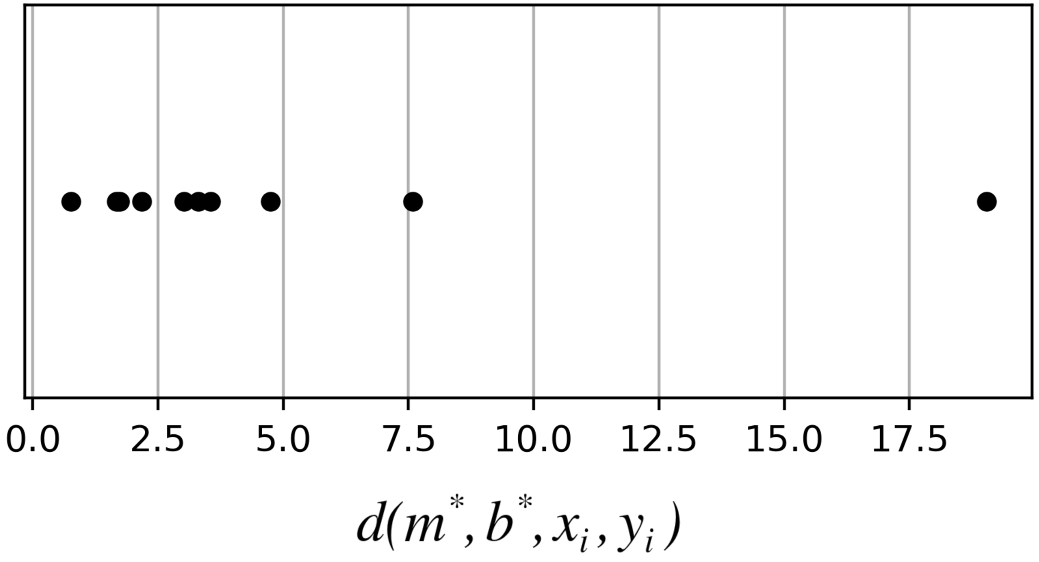
We could **explore the data for patterns** that identify an outlier boundary. (unsupervised learning)

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

Clustering

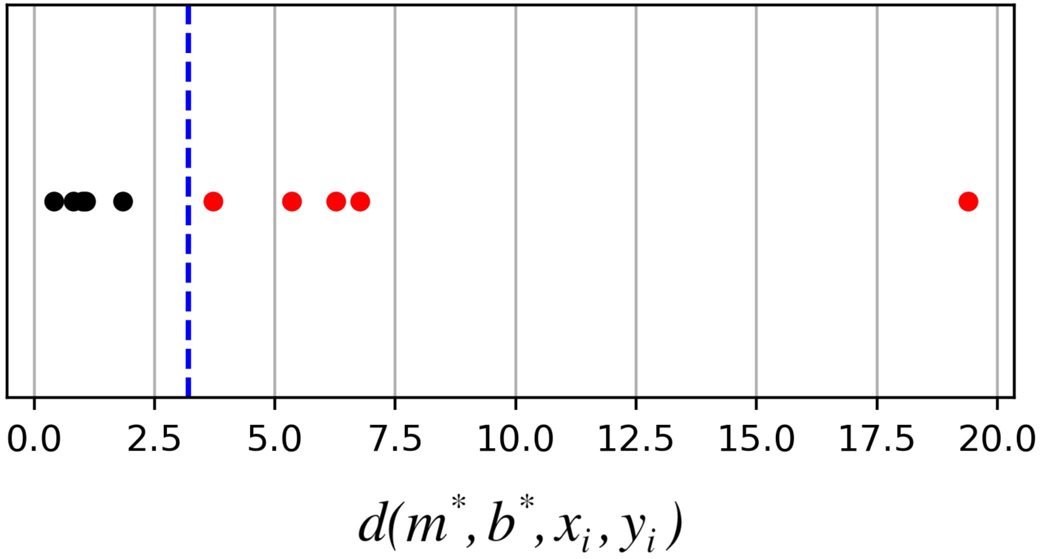


- Relative to outliers, data model errors often form a **cluster.**



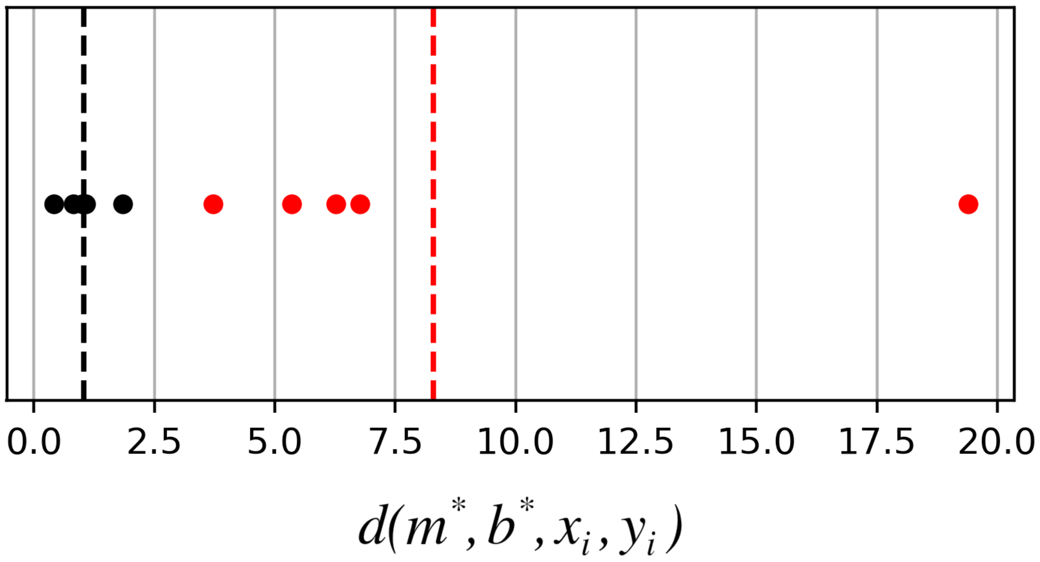
Clustering

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



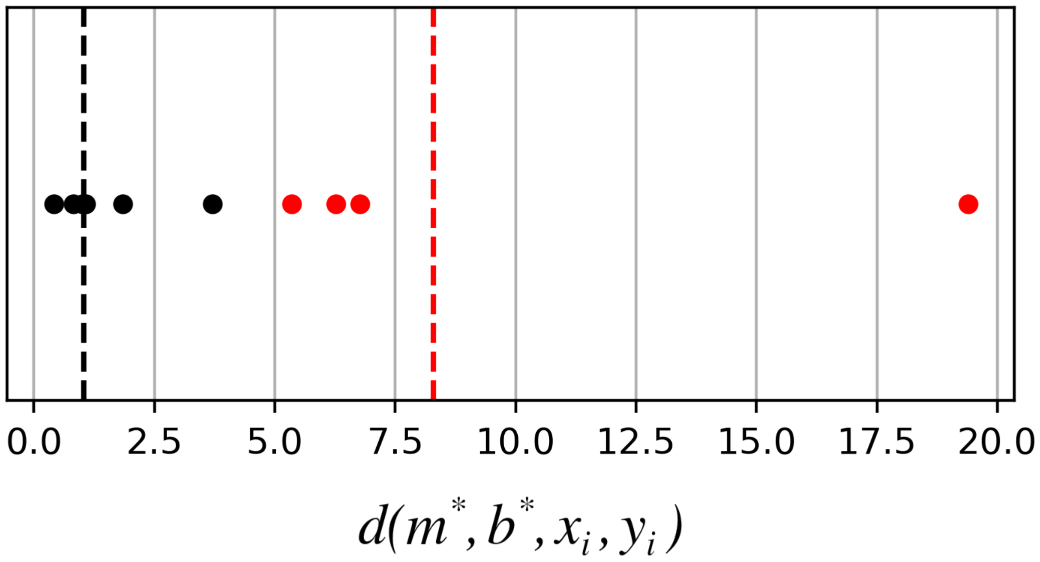
Clustering

- 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*).



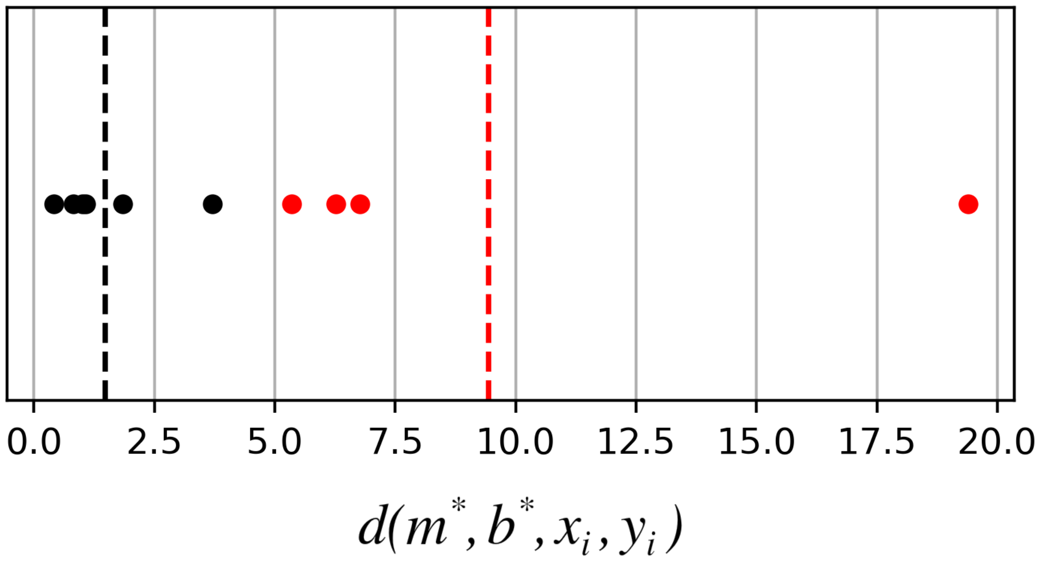
Clustering

- 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

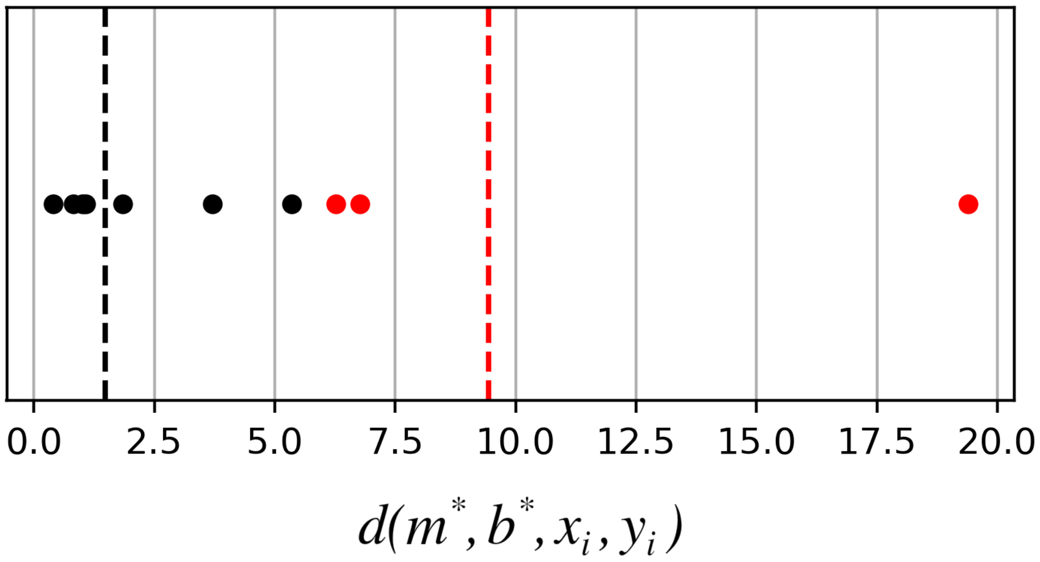
- How might we identify the cluster? One approach:
  1. Start with a random outlier boundary.
  2. Calculate the means of the two groups (call them *clusters*).
  3. Assign each data point to the nearest mean.



Repeat the process until memberships stop changing.

Clustering

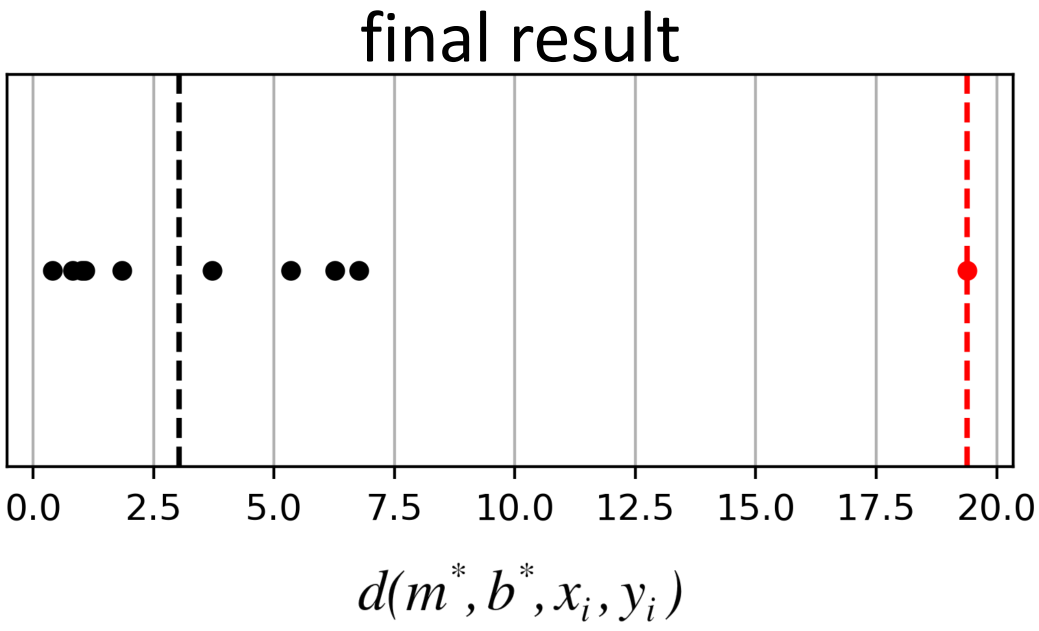
- How might we identify the cluster? One approach:
  1. Start with a random outlier boundary.
  2. Calculate the means of the two groups (call them *clusters*).
  3. Assign each data point to the nearest mean.

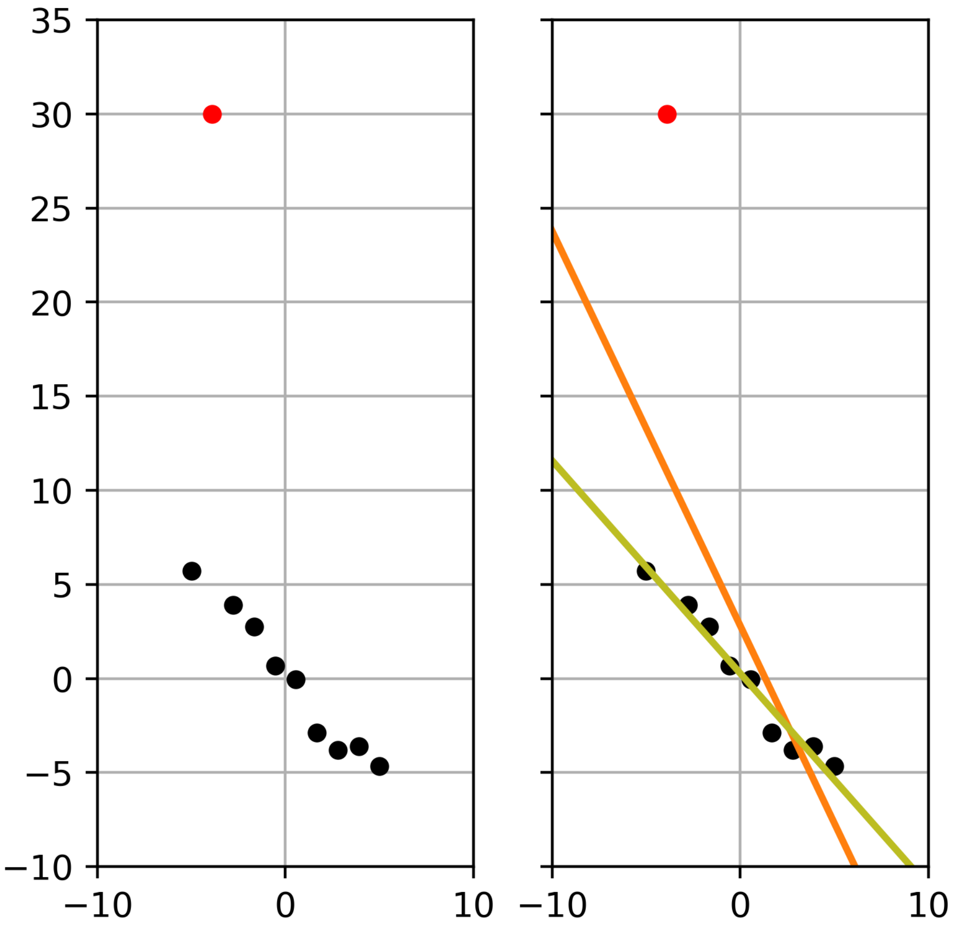


Repeat the process until memberships stop changing.

Clustering

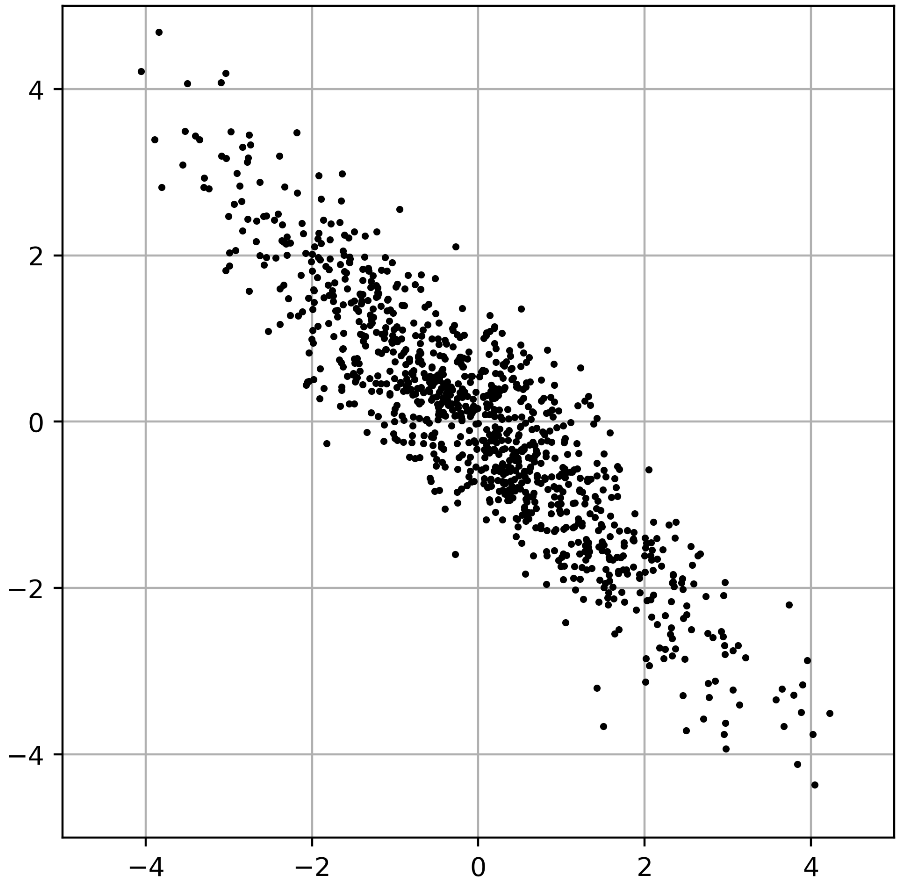
- This is known as 1D **k-means clustering**.





Dimensionality Reduction

Dimensionality Reduction—A General Ex.

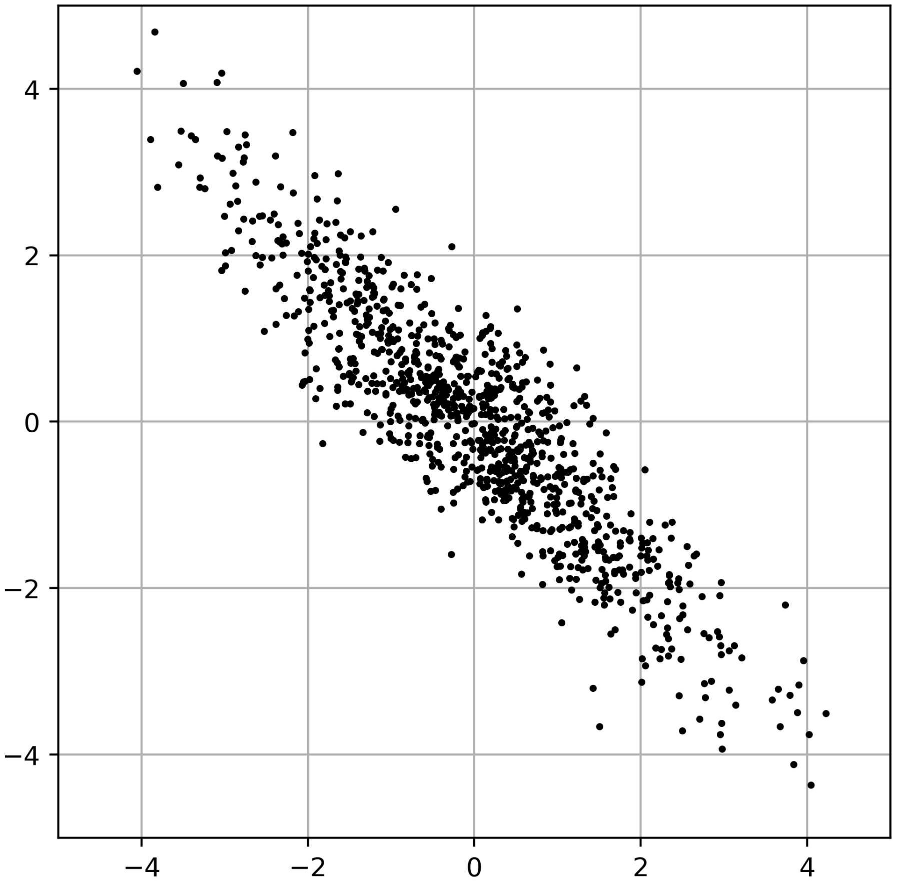


- Does this data exist across two dimensions?
  - Technically, yes.
  - Practically…?

- How might we assess the *true* dimensionality of the dataset?

- CEDAR example: dominant modes in the high-latitude ionospheric electrodynamics; Matsuo et al, 2002, 2003, 2005.

Dimensionality Reduction



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

2 x N matrix of data

$$\tilde{\Sigma} = \frac{XX^T}{N-1}$$

(sample covariance)

Dimensionality Reduction

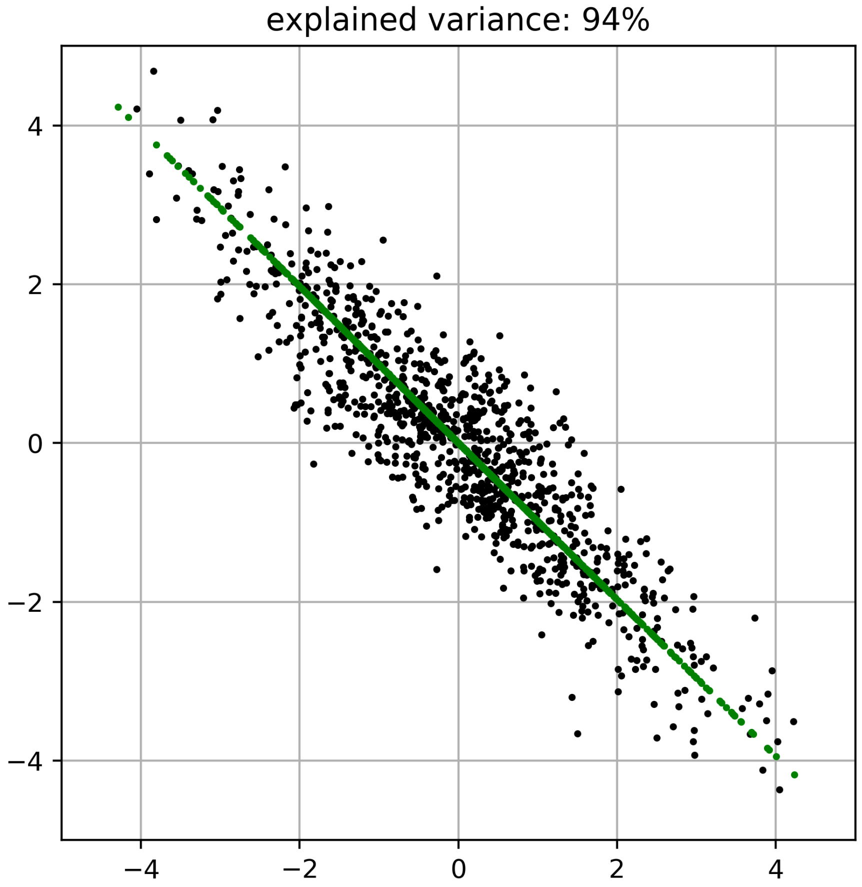
- The eigenvalues and eigenvectors of the sample covariance describe the appropriate change of basis.

$$\widetilde{\Sigma} = \underset{\substack{\uparrow \\ \text{matrix of} \\ \text{eigenvectors}}}{\boldsymbol{U}} \underset{\substack{\nwarrow \\ \text{diagonal matrix of} \\ \text{eigenvalues}}}{\Lambda} \boldsymbol{U}^T$$

(diagonalized sample covariance)

- 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

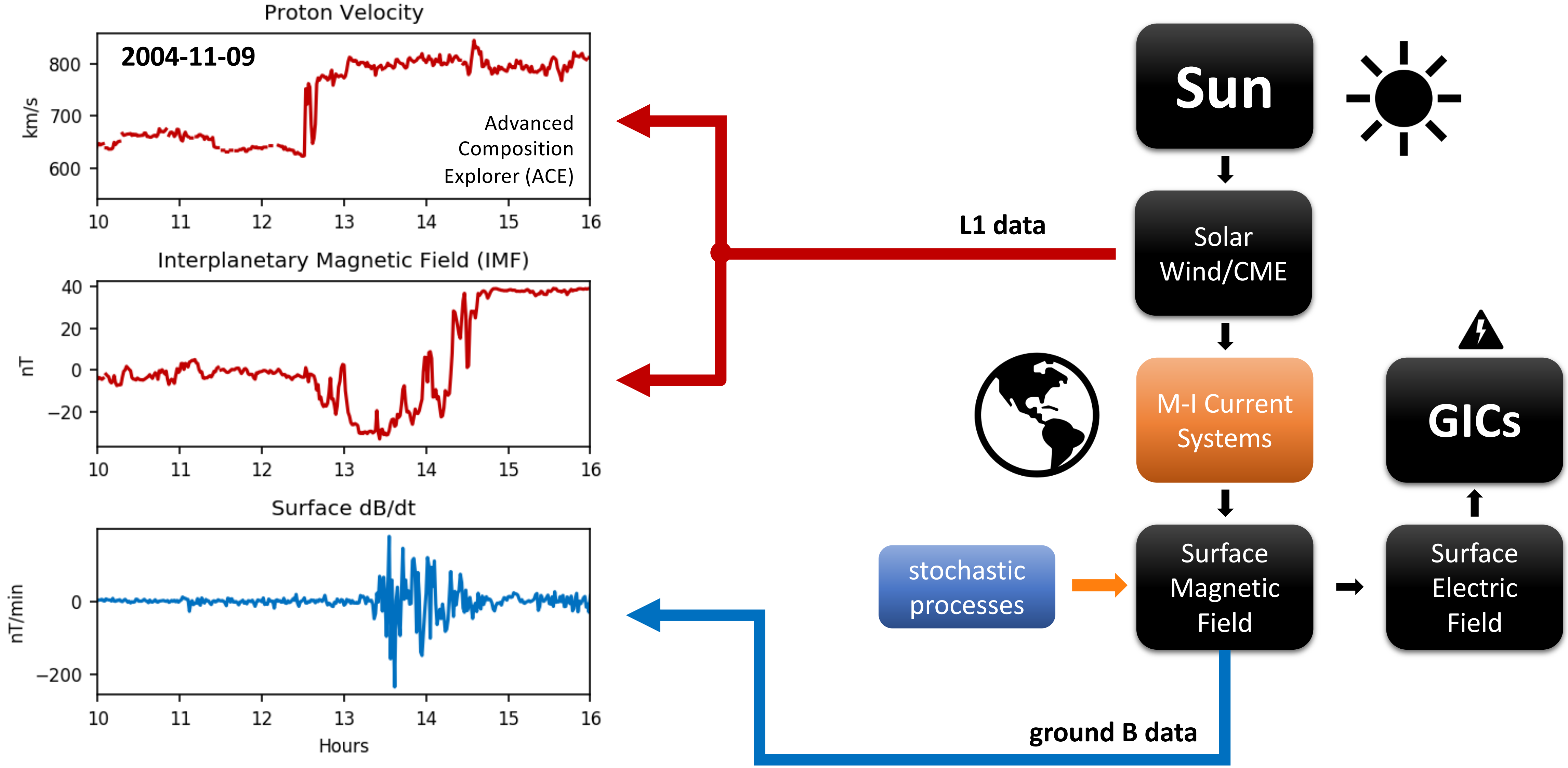
- Principal component analysis learns a basis for the data that is **adaptive**.
  - This is directly related to the singular value decomposition (SVD) of the data matrix.

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

$$\boldsymbol{X} = \sqrt{N\text{-}1} \sum_{i=1}^{\boldsymbol{d}} \sqrt{\lambda_i}\, u_i v_i^T \overset{\boldsymbol{k} \leq \boldsymbol{d}}{\approx} \sqrt{N\text{-}1} \sum_{i=1}^{\boldsymbol{k}} \sqrt{\lambda_i}\, u_i v_i^T$$

(singular value decomposition) (low-rank approximation)

A system Identification Perspective of Learning Theory



L1 data

Sun

Solar Wind/CME

M-I Current Systems

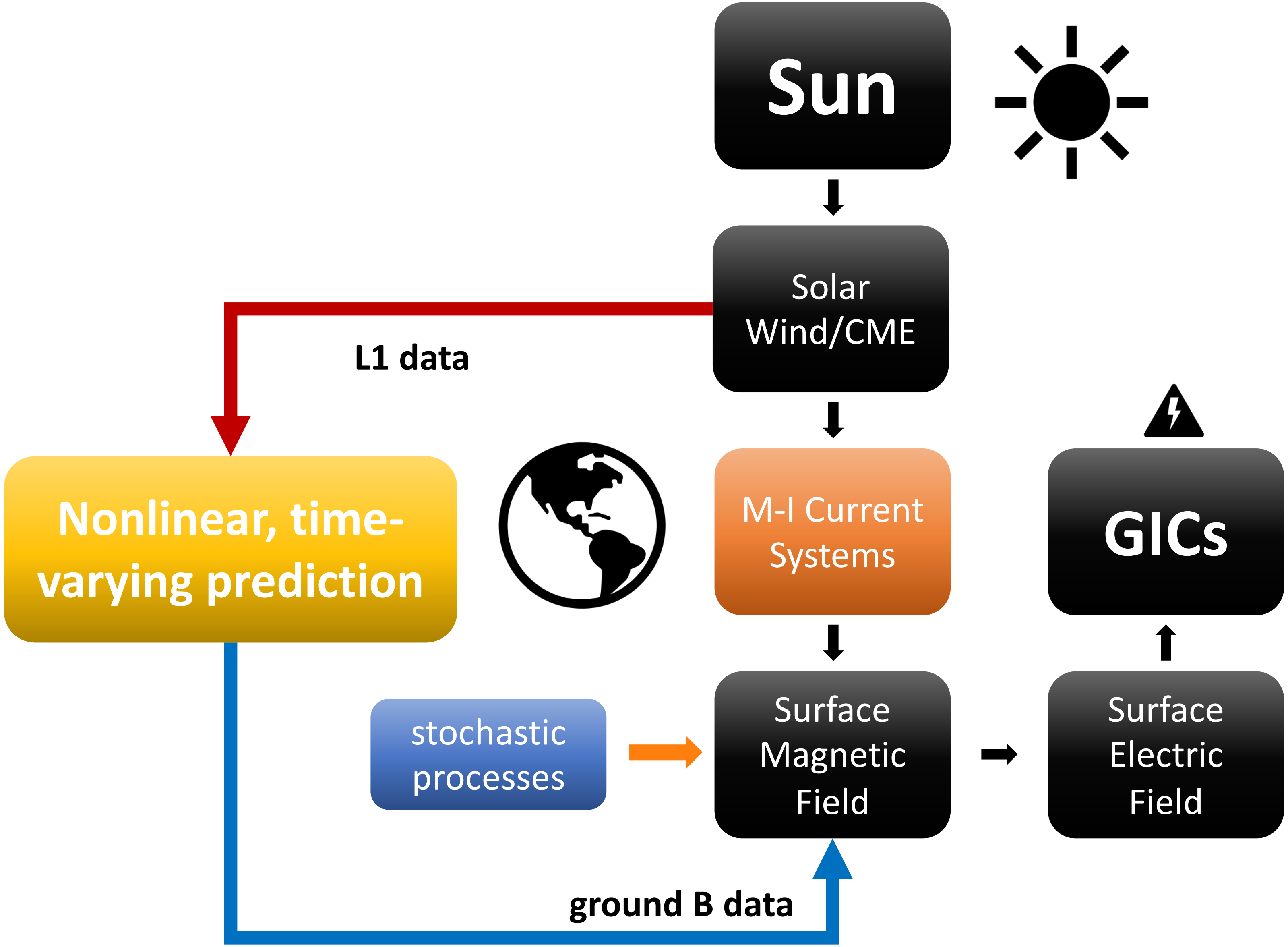
GICs

stochastic processes

Surface Magnetic Field

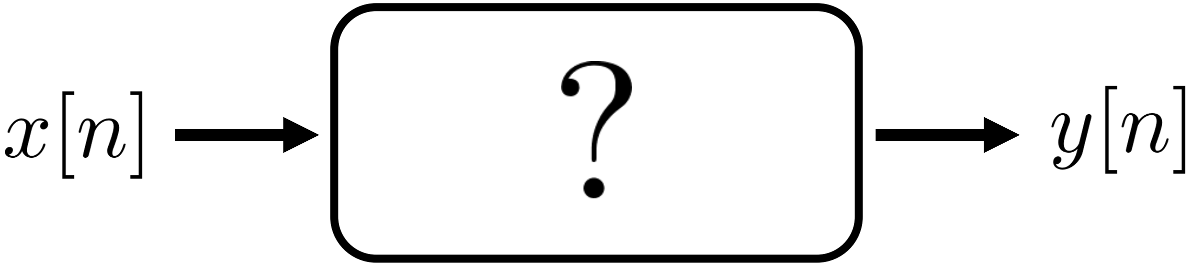
Surface Electric Field

ground B data


Sun
Solar Wind/CME
L1 data
Nonlinear, time-varying prediction
M-I Current Systems
GICs
stochastic processes
Surface Magnetic Field
Surface Electric Field
ground B data

System Identification

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



- What assumptions can be made about the nature of the system?
  - Linear and/or time-invariant?

Linear, Time-Invariant System Identification

- General linear data model:

$$y[n] = G(q, \boldsymbol{\theta})x[n] + H(q, \boldsymbol{\theta})e[n]$$

Linear, Time-Invariant System Identification

- General linear data model:

noise

$$y[n] = G(q, \boldsymbol{\theta})x[n] + H(q, \boldsymbol{\theta})e[n]$$

delay operator $(qx[n] = x[n-1])$

Linear, Time-Invariant System Identification

- General linear data model:

noise

$$y[n] = G(q, \boldsymbol{\theta})x[n] + H(q, \boldsymbol{\theta})e[n]$$

delay operator $(qx[n] = x[n-1])$

$$G(q, \boldsymbol{\theta}) = \sum_{k=1}^{\infty} g(k)q^{-k} \qquad H(q, \boldsymbol{\theta}) = \sum_{k=1}^{\infty} h(k)q^{-k}$$

Linear, Time-Invariant System Identification

- Slightly more restricted class of models: **rational transfer functions**

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

Linear, Time-Invariant System Identification

- Slightly more restricted class of models: **rational transfer functions**

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

- A, B, C, D, F are **lag polynomials**, e.g., $A(q) = 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a}$

Linear, Time-Invariant System Identification

- Slightly more restricted class of models: **rational transfer functions**

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

- A, B, C, D, F are **lag polynomials**, e.g., $A(q) = 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a}$
- The system is defined by the weights on past samples of the input, output, and noise.

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

Linear, Time-Invariant System Identification

- Given input and output data, how might we estimate the system, or, equivalently, estimate the parameter vector $\boldsymbol{\theta}$?

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**

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

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

Nonlinear, Time-Varying Systems

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

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

fixed window of past input and output data

- The general approach is the same as the LTI case, but there are far fewer restrictions on the functional form of the one-step prediction error.

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

Nonlinear, Time-Varying Systems

- Common approach: expand the mapping using a basis

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

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

- 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.**

$$g_k^{(2)}(\boldsymbol{\phi}) = \sum_l \alpha_l^{(2)} \kappa(\boldsymbol{\phi}^{(2)}, \boldsymbol{\beta}_l^{(2)}, \boldsymbol{\gamma}_l^{(2)}) \qquad \phi_k^{(2)} = g_k(\boldsymbol{\phi})$$

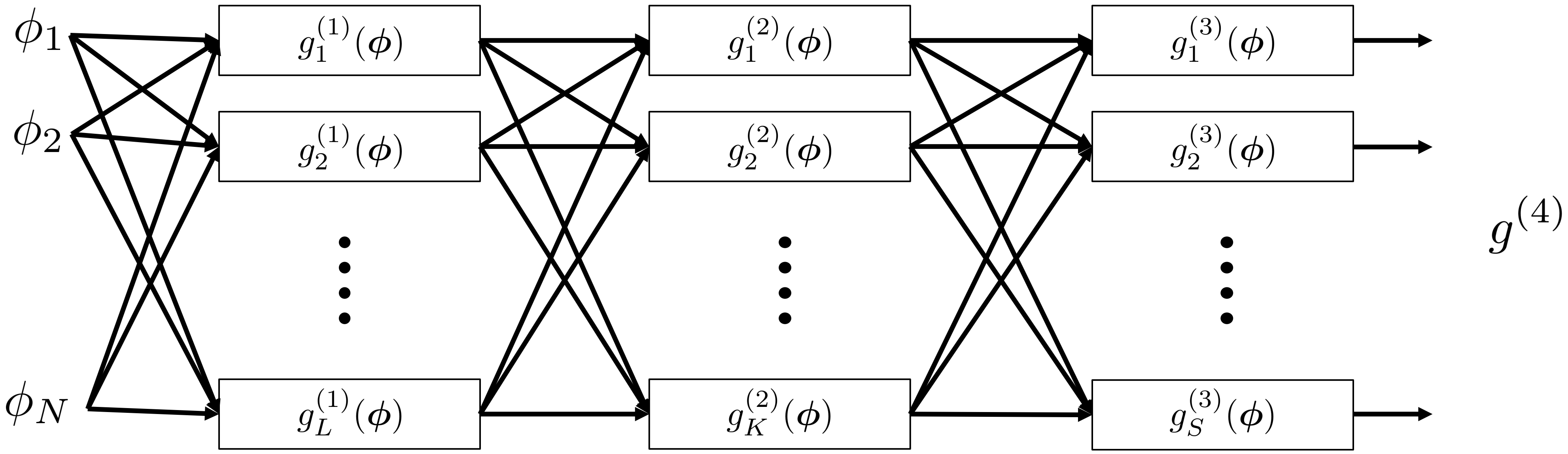
$$g_k^{(3)}(\boldsymbol{\phi}) = \sum_l \alpha_l^{(3)} \kappa(\boldsymbol{\phi}^{(3)}, \boldsymbol{\beta}_l^{(3)}, \boldsymbol{\gamma}_l^{(3)}) \qquad \phi_k^{(3)} = g_k(\boldsymbol{\phi}^{(2)})$$

$$\vdots$$

$$g_k^{(M)}(\boldsymbol{\phi}) = \sum_l \alpha_l^{(M)} \kappa(\boldsymbol{\phi}^{(M)}, \boldsymbol{\beta}_l^{(M)}, \boldsymbol{\gamma}_l^{(M)}) \qquad \phi_k^{(M)} = g_k(\boldsymbol{\phi}^{(M-1)})$$

Nonlinear, Time-Varying Systems

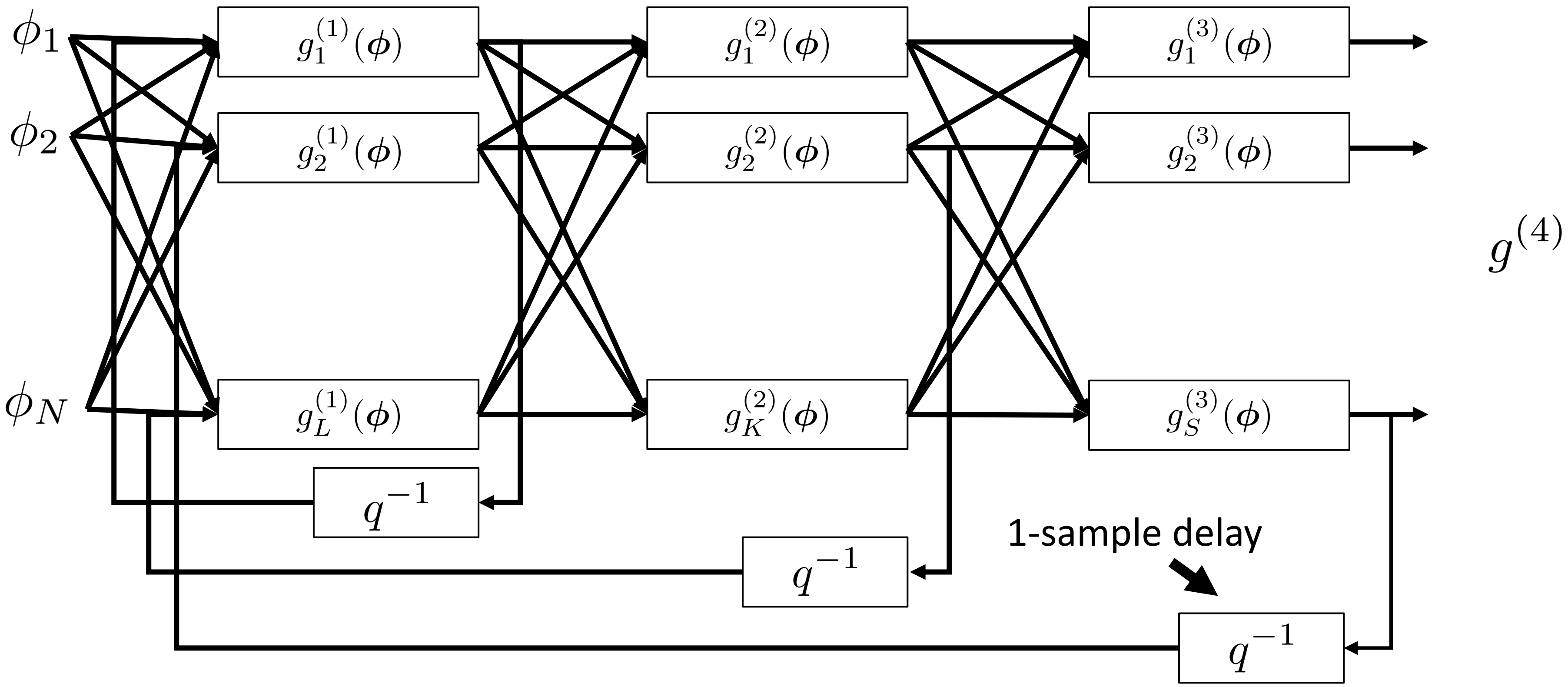
- Layered/composed expansions are **neural networks.**



ϕ_1
ϕ_2
ϕ_N
$g_1^{(1)}(\boldsymbol{\phi})$
$g_2^{(1)}(\boldsymbol{\phi})$
$g_L^{(1)}(\boldsymbol{\phi})$
$g_1^{(2)}(\boldsymbol{\phi})$
$g_2^{(2)}(\boldsymbol{\phi})$
$g_K^{(2)}(\boldsymbol{\phi})$
$g_1^{(3)}(\boldsymbol{\phi})$
$g_2^{(3)}(\boldsymbol{\phi})$
$g_S^{(3)}(\boldsymbol{\phi})$
$g^{(4)}$

Nonlinear, Time-Varying Systems

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


ϕ_1
ϕ_2
ϕ_N
$g_1^{(1)}(\boldsymbol{\phi})$
$g_2^{(1)}(\boldsymbol{\phi})$
$g_L^{(1)}(\boldsymbol{\phi})$
$g_1^{(2)}(\boldsymbol{\phi})$
$g_2^{(2)}(\boldsymbol{\phi})$
$g_K^{(2)}(\boldsymbol{\phi})$
$g_1^{(3)}(\boldsymbol{\phi})$
$g_2^{(3)}(\boldsymbol{\phi})$
$g_S^{(3)}(\boldsymbol{\phi})$
$g^{(4)}$
q^{-1}
q^{-1}
q^{-1}

1-sample delay

Learning Theory Caveats and Open Directions

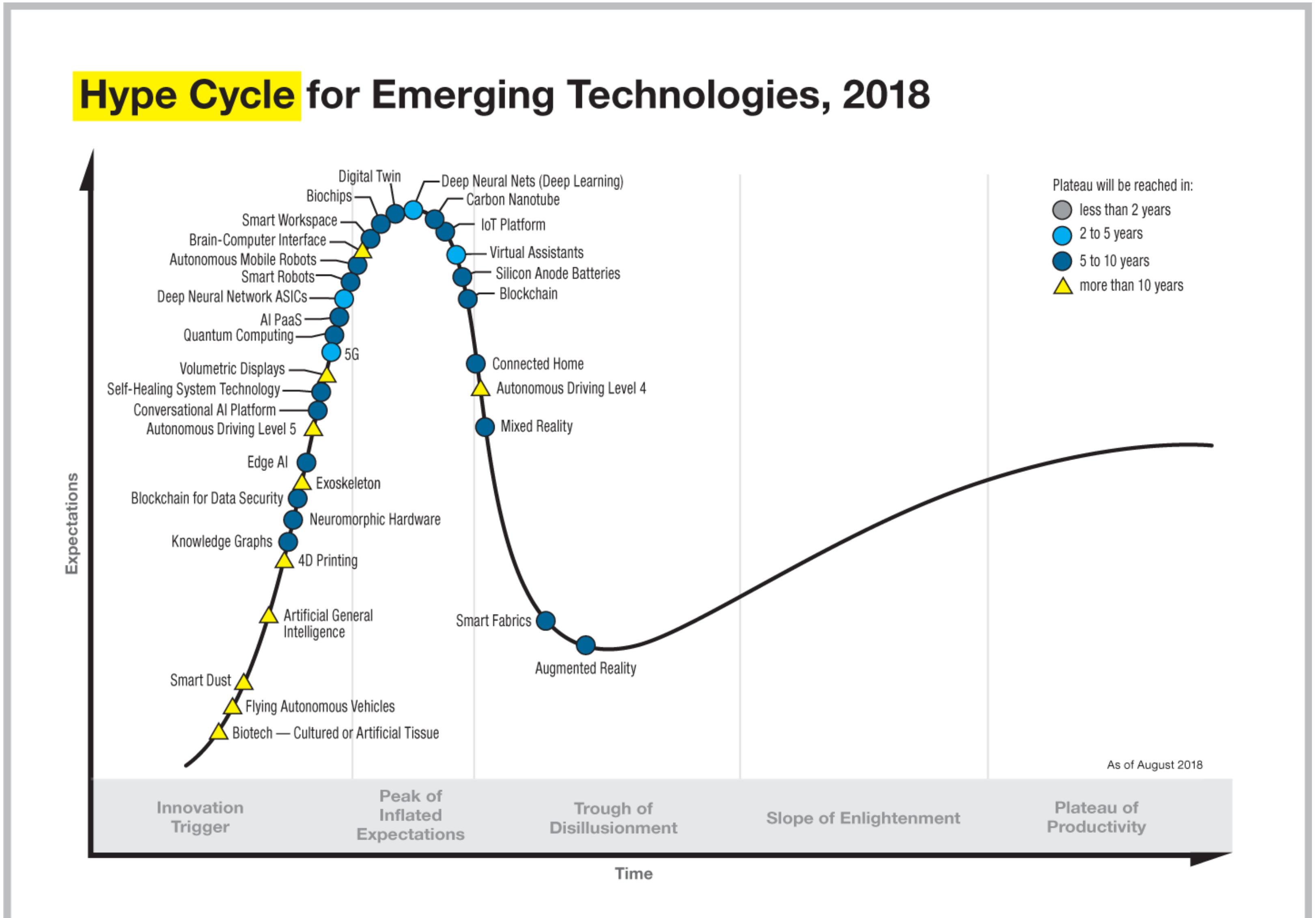
- With nonlinear systems, cost function minimization presents special challenges.
  - Nonlinear cost functions are usually non-convex, and have many local minima.

- Solutions for $\boldsymbol{\theta}$ 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.

Concluding Observations

6/15/2019 PR_490866_5_Trends_in_the_Emerging_Tech_Hype_Cycle_2018_Hype_Cycle.png (1440×1218)


Hype Cycle for Emerging Technologies, 2018
Digital Twin
Biochips
Deep Neural Nets (Deep Learning)
Carbon Nanotube
IoT Platform
Smart Workspace
Brain-Computer Interface
Autonomous Mobile Robots
Smart Robots
Deep Neural Network ASICs
AI PaaS
Quantum Computing
5G
Volumetric Displays
Self-Healing System Technology
Conversational AI Platform
Autonomous Driving Level 5
Edge AI
Exoskeleton
Blockchain for Data Security
Neuromorphic Hardware
Knowledge Graphs
4D Printing
Artificial General Intelligence
Smart Dust
Flying Autonomous Vehicles
Biotech — Cultured or Artificial Tissue
Virtual Assistants
Silicon Anode Batteries
Blockchain
Connected Home
Autonomous Driving Level 4
Mixed Reality
Smart Fabrics
Augmented Reality
Plateau will be reached in:
● less than 2 years
● 2 to 5 years
● 5 to 10 years
△ more than 10 years
Expectations
As of August 2018
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time