Large Scale Feature Selection and Online Learning

Gregory Ditzler

Dept. of Electrical & Computer Engineering
ditzler@email.arizona.edu

ARIZONA

Why data science, machine learning and knowledge discovery?

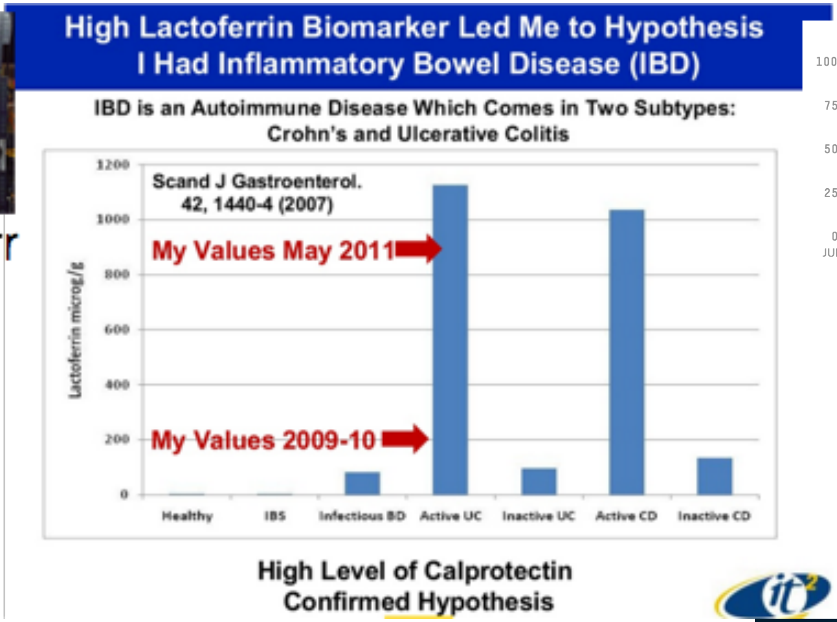
Harvard Business Review

Data Scientist: The Sexiest Job

Century

by Thomas H. Davenport and D.J. Patil

Larry Smarr



High Lactoferrin Biomarker Led Me to Hypothesis I Had Inflammatory Bowel Disease (IBD)

IBD is an Autoimmune Disease Which Comes in Two Subtypes: Crohn's and Ulcerative Colitis

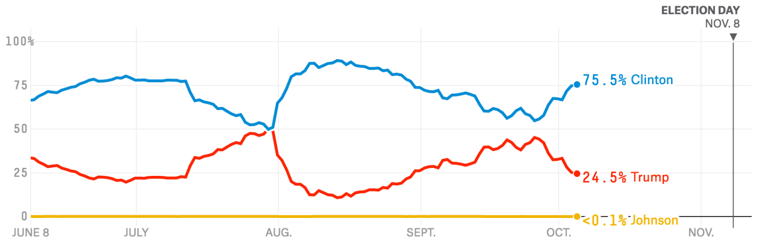
Scand J Gastroenterol. 42, 1440-4 (2007)

My Values May 2011

My Values 2009-10

High Level of Calprotectin Confirmed Hypothesis

FiveThirtyEight



http://medium.com

When A Machine Learning Algorithm Studied Fine Art Paintings, It Saw Things Art Historians Had Never Noticed

Artificial intelligence reveals previously unrecognised influences between great artists

Google DeepMind

Challenges & constraints

- There are a lot of data being generated in today’s technological climate
    - Pro: some data are useful
    - Con: some data are not useful
    - A little bit like finding the signal in the noise

- Some scenarios request that variables maintain a physical interpretation

- Life sciences and clinical tests

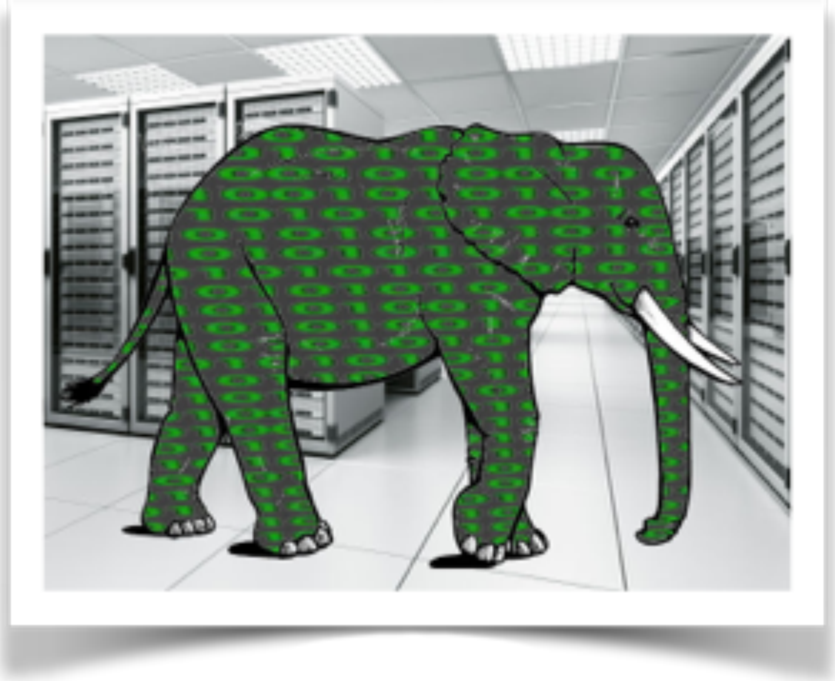
Feature Selection for Large Volumes of Data

Data aren’t small anymore

- Scale of data sets are growing rapidly in the internet era

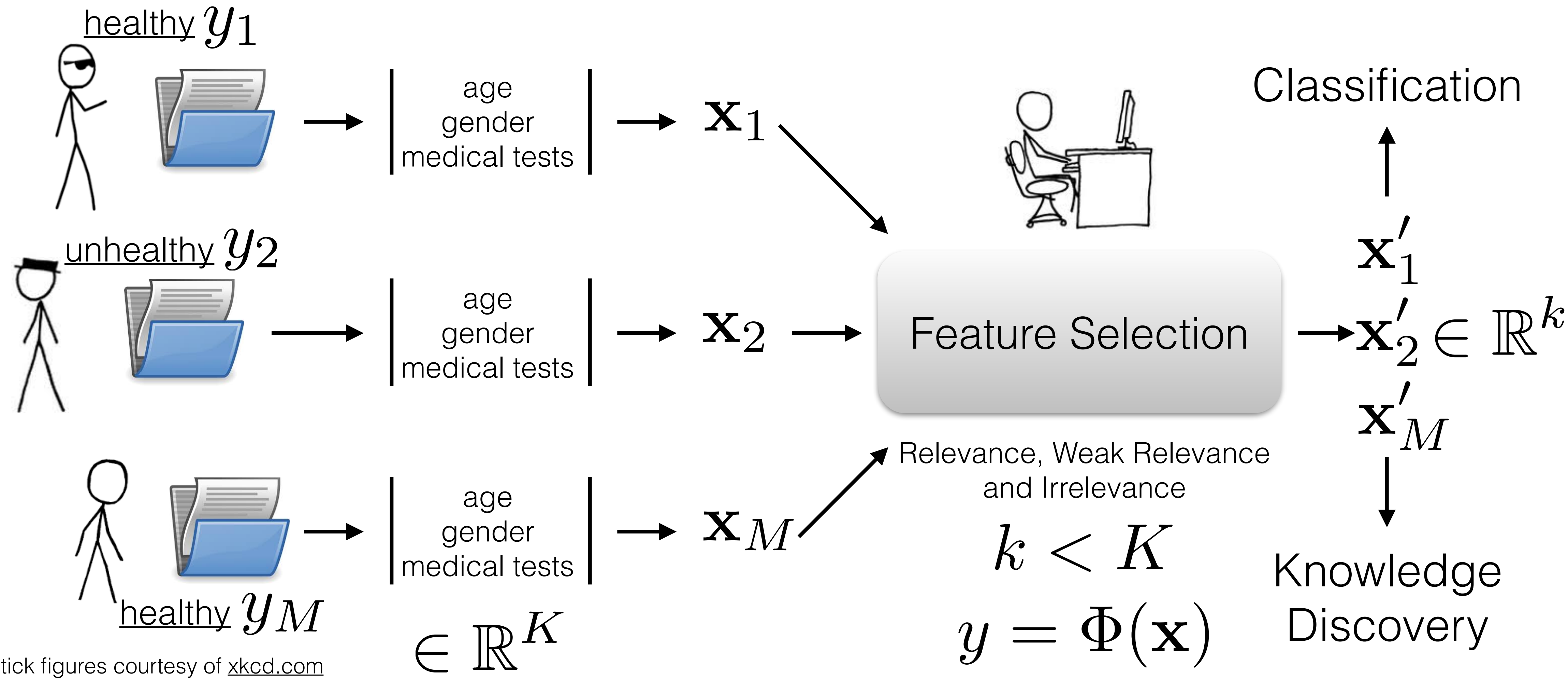
- What does it mean for data to be “large”?

- The Five V’s: volume, velocity, variety, veracity, and value
  - velocity: data arrive in a stream
  - volume: not only number of samples, but the dimensionality
  - value: not of cost, but of importance



- Twenty years of machine learning research has led to wide body of research for detecting value
  - “volume” of twenty years ago is not the volume of today
  - distributed, parallel, and statistically sound

Feature selection in a nutshell



healthy y_1

age
gender
medical tests → $\mathbf{x}_1$

unhealthy y_2

age
gender
medical tests → $\mathbf{x}_2$

healthy y_M

age
gender
medical tests → $\mathbf{x}_M$

$\in \mathbb{R}^K$

Feature Selection

Relevance, Weak Relevance and Irrelevance

$k < K$

$y = \Phi(\mathbf{x})$

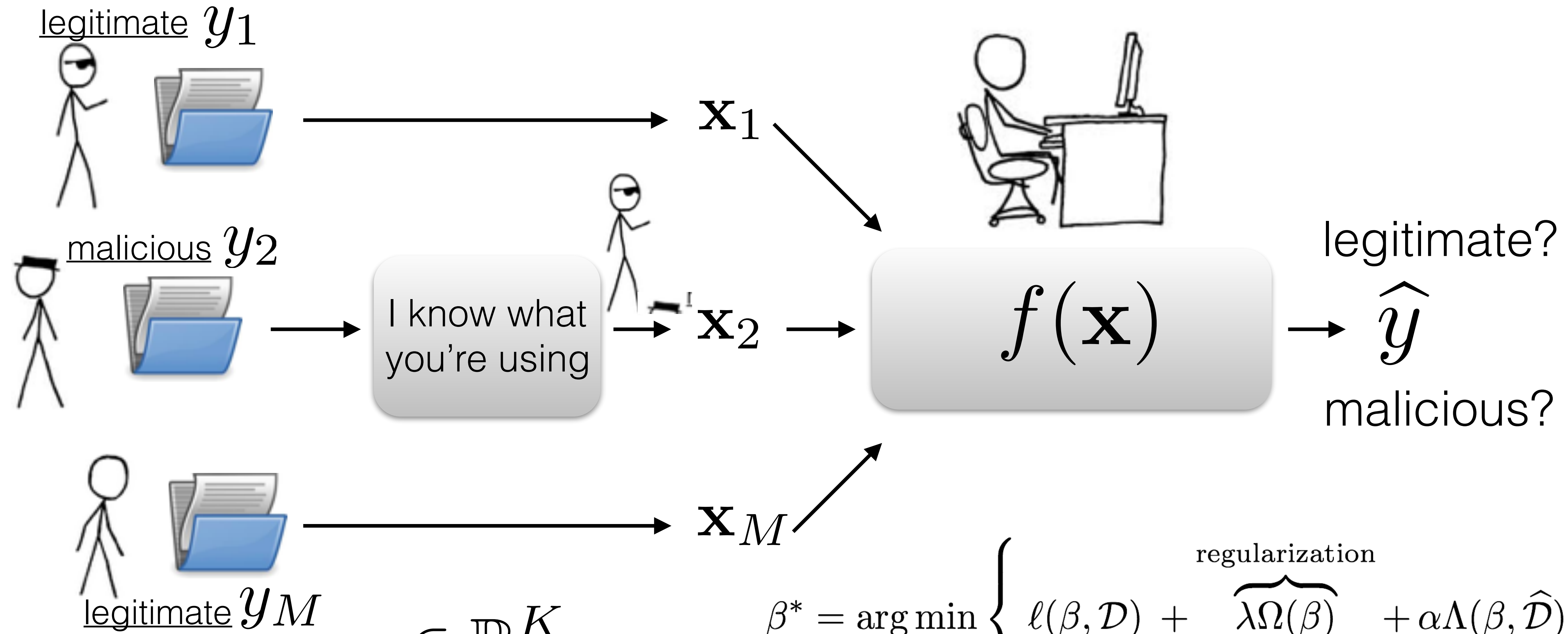
$\mathbf{x}'_1$ $\mathbf{x}'_2 \in \mathbb{R}^k$ $\mathbf{x}'_M$

Classification

Knowledge Discovery

Stick figures courtesy of xkcd.com

Feature selection with an adversary in a nutshell



$\in \mathbb{R}^K$

$$\beta^* = \arg\min_{\beta \in \Phi} \left\{ \underbrace{\ell(\beta, \mathcal{D})}_{\text{model loss}} + \overbrace{\lambda\Omega(\beta)}^{\text{regularization}} + \underbrace{\alpha\Lambda(\beta, \widehat{\mathcal{D}})}_{\text{adversary}} \right\}$$

Stick figures courtesy of xkcd.com

Approaches for feature selection

~~Wrapper Methods~~

- ~~Build a classifier, measure loss, adapt feature set, repeat..~~
- ~~Easy to over fit~~
- ~~Too computationally complex~~

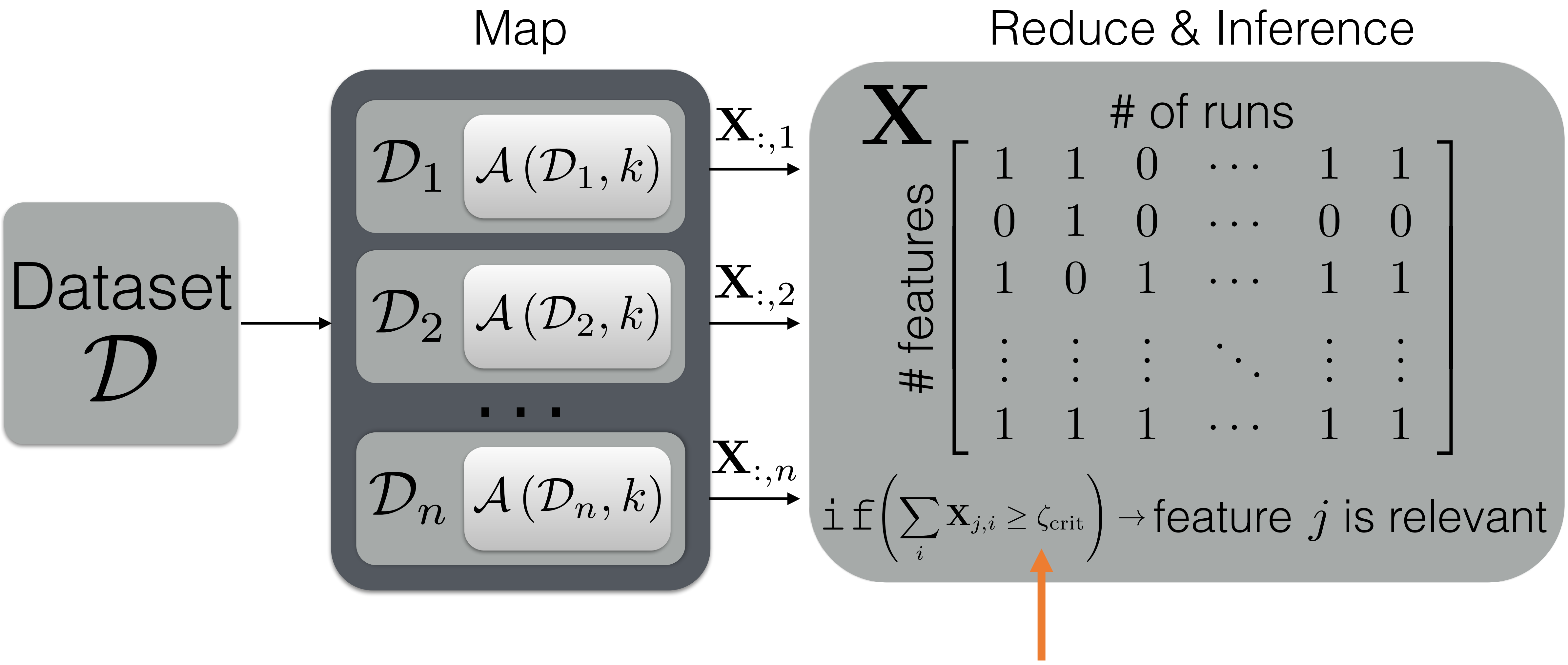
Embedded Methods

- Jointly optimize classifier and variable selector parameters
- E.g., Linear model with L1 penalization.

Filter Methods

- Optimize feature set independent from a classifier
- Fast, but need ways to scale them to large volumes data

NPFS in one picture



Map

Reduce & Inference

Dataset $\mathcal{D}$

$\mathcal{D}_1$ $\mathcal{A}(\mathcal{D}_1, k)$ $\mathbf{X}_{:,1}$

$\mathcal{D}_2$ $\mathcal{A}(\mathcal{D}_2, k)$ $\mathbf{X}_{:,2}$

$\cdots$

$\mathcal{D}_n$ $\mathcal{A}(\mathcal{D}_n, k)$ $\mathbf{X}_{:,n}$

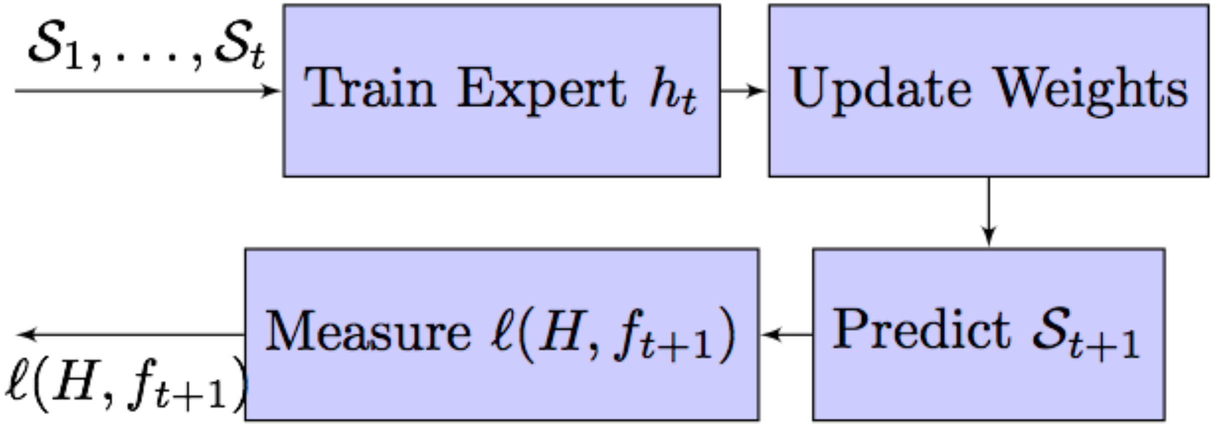
$$\mathbf{X} \quad \begin{array}{c|c} & \text{\# of runs} \\ \hline \text{\# features} & \begin{bmatrix} 1 & 1 & 0 & \cdots & 1 & 1 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix} \end{array}$$

$\texttt{if}\left(\sum_i \mathbf{x}_{j,i} \geq \zeta_{\text{crit}}\right) \rightarrow$ feature j is relevant

Neyman-Pearson hypothesis test for relevance

Streaming Data and Learning Online

Learning in Nonstationary Environments



$\mathcal{S}_1, \ldots, \mathcal{S}_t$
Train Expert h_t
Update Weights
Predict $\mathcal{S}_{t+1}$
Measure $\ell(H, f_{t+1})$
$\ell(H, f_{t+1})$

$$H(\mathbf{x}) = \sum_{k=1}^{t} w_{k,t} h_t(\mathbf{x})$$

$$\ell(\alpha y + (1-\alpha) z) \leq \alpha \ell(y) + (1-\alpha)\ell(z)$$

- The Five V's: volume, velocity, variety, veracity, and value
- Traditional Learning Paradigm: training and testing data are sampled from the same probability distribution
  - what if that is not the situation?
- What should we expect the loss to look like when predicting on data from an unknown distribution?

Discussion Points

“There are two types of machine learning practitioners: (1) those that generalized from limited data”

Challenges Moving Forward

- Advanced Frameworks for Big Data Subset Selection
    - IEEE CIM recently published a special issue on Big Data and the curse of big dimensionality
    - Millions of features & beyond while being mindful of veracity
    - Interpretability, Visualization and Real-Time
    - Migrating from a batch-based learning setting to a pure online setting
- Calibrated Prediction for Domain Adaptation in Time-Series
    - Tuning model parameters in changing domains using unlabeled data, and accessing the stability of predictions in uncertain environments
- Learning New Classes without Re-training

Thanks for Listening!

Acknowledgements

AVIRTEK
Autonomic Cyber Security

SCIENCE TO SHAPE THE FUTURE OF THE ARMY
ARMY RESEARCH OFFICE
1951

AIR FORCE RESEARCH LABORATORY