SDC 19
SNIA INDIA

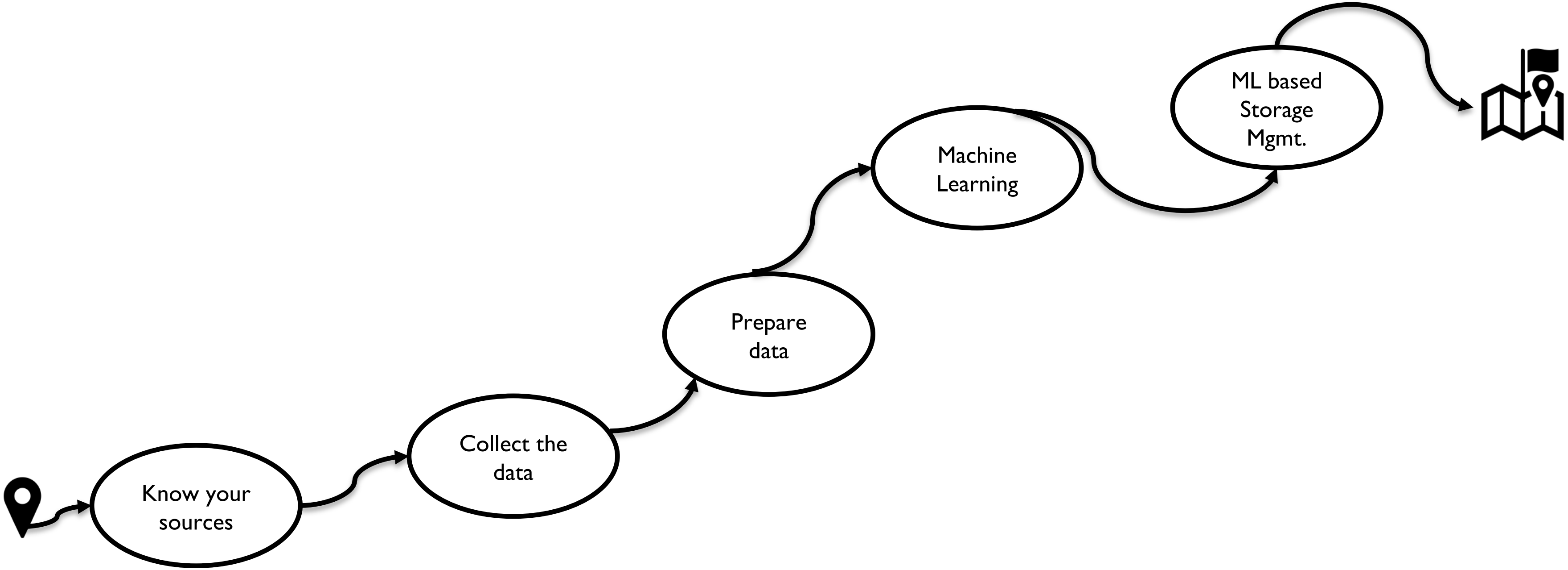
May 23-24, 2019
Bangalore, India

STORAGE DEVELOPER CONFERENCE

How IOT, Analytics and ML unfolds in Storage Fabric

Sharath T S
Microchip

Our Map!



Know your sources

Collect the data

Prepare data

Machine Learning

ML based Storage Mgmt.

Our itinerary

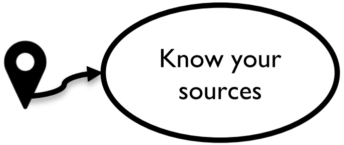
- Effect of IoT to Data Centers
- Effect of IoT in Data Centers
- Collection of data from sources
- Prepare data
- Applying ML for different uses cases
- Data visualization

Effect of IoT to Data Centers

Effect of IoT to Data Centers

- 26 B sensors by 2020 and 50 B connected devices
- 5G IoT
- Edge computing
- Detailed analysis
- IoT impact on data-center management

Effect of IoT in Data Centers



Know your sources

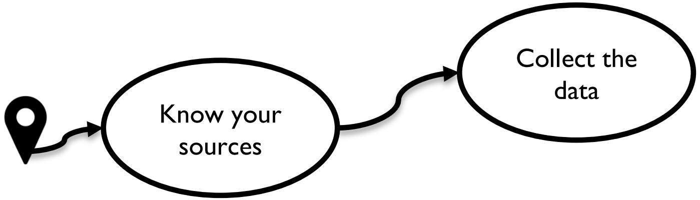
End points in Data Centers

- EMS (Environmental Monitoring Systems) & ASHRAE (American Society of Heating, Refrigeration, and Air-Conditioning Engineers)
  - Temperature (18 to 27 °C)
  - Humidity and water (RH 45% to 60%)
  - Air flow sensors
  - Static electric sensors
  - Server room and rack entry
  - Aisle conditions

End points in Data Centers

- Server
- Storage controller
- Physical Drives (S.M.A.R.T)
- Chassis

Collection of data from sources



What data to collect?

- Sensor
  - Temperature, humidity, static electric charges, intrusion
- Storage
  - System, Storage pools, Storage volumes, Drives and Chassis

Example Data Collection

- **System Information**
  - CPU Utilization
  - Network Utilization
  - Memory Utilization
  - OS details
  - Uptime
- **Storage Controller Information**
  - Status
  - Mode
  - Interface
  - Temperature
- **Storage Pool**
  - Status
  - Interface
  - Total Size
  - Unused Size
  - Spare Rebuild mode
  - Volume count
  - Drive count
  - Type
- **Storage Volume**
  - Status
  - Interface
  - Total Size
  - Unused Size
  - Block Size
  - RAID Level
  - Drive count
  - Protected by Hot-Spare
  - Write-cache
  - Read-cache
  - Acceleration method
- **Drive**
  - Manufacturer
  - Type
  - Status
  - Interface
  - Total Size
  - Unused Size
  - Reserved Size
  - Block Size
  - Transfer speed
  - SMART stats
  - Bad Blocks

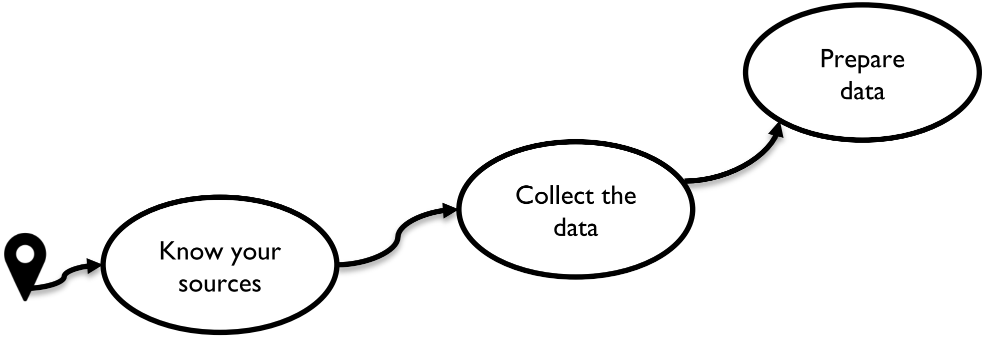
When to collect data?

- Periodic interval
  - Time Series analysis and Forecasting
- Event Based
  - User initiated, System initiated

How to collect data?

- Push mechanism
  - Source / System generated
  - Breach of any threshold values
  - Listener is required to read and store value
- Pull mechanism
  - Application / User requested
  - On demand / Periodic
  - Programmatically

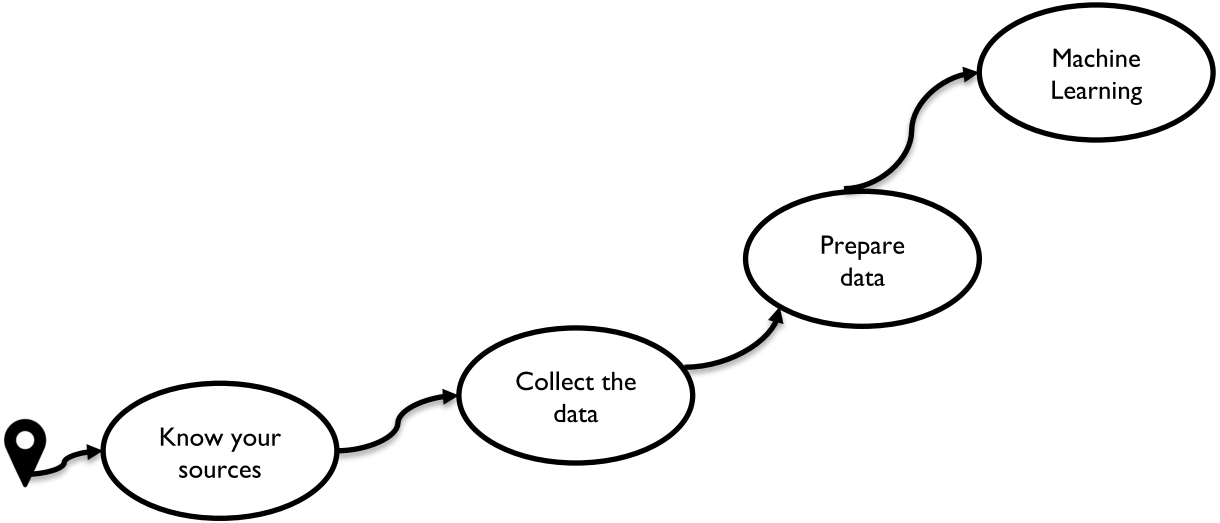
Prepare data



Data

- Types of source data
  - Unstructured
  - Semi-structured
  - Structured
- On-Line Analytical Processing (OLAP) of prepared data
  - Cubes
  - Dimensions

Introduction to Machine Learning

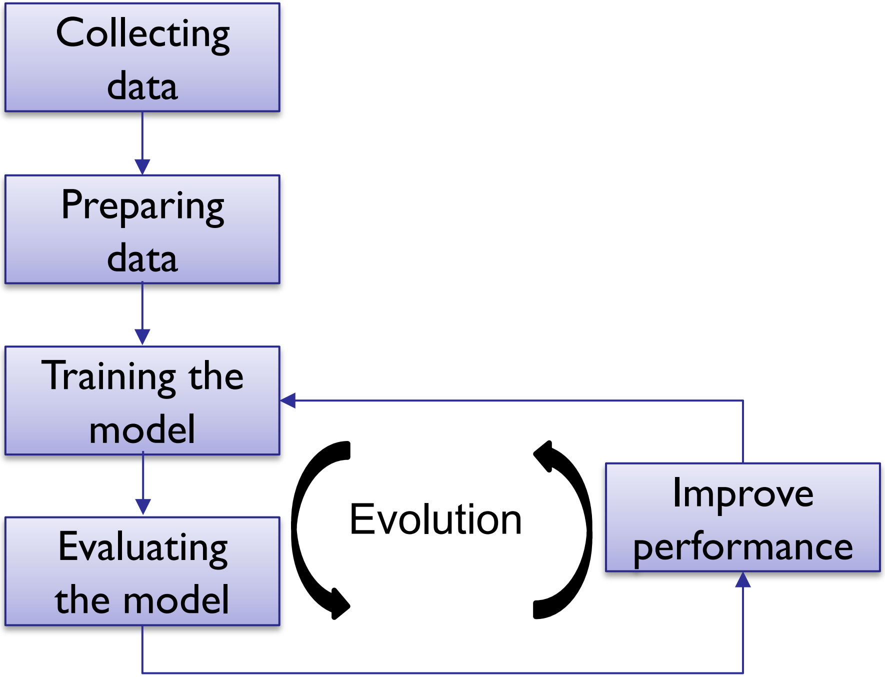


Introduction to Machine Learning

- ❒ Introduction

Digital Journal

- ❒ Flow



Collecting data

Preparing data

Training the model

Evaluating the model

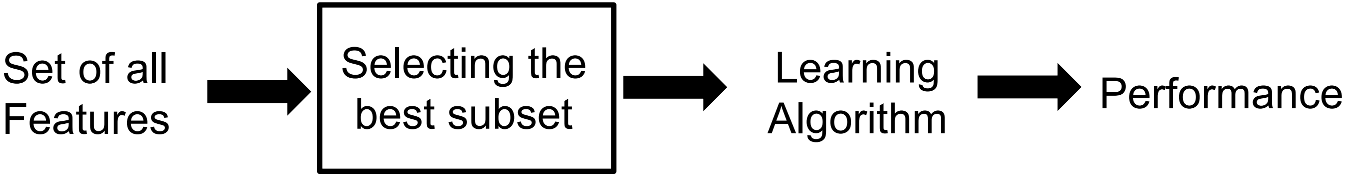
Improve performance

Evolution

Data Feature Selection

Data Feature Selection

- Filter method

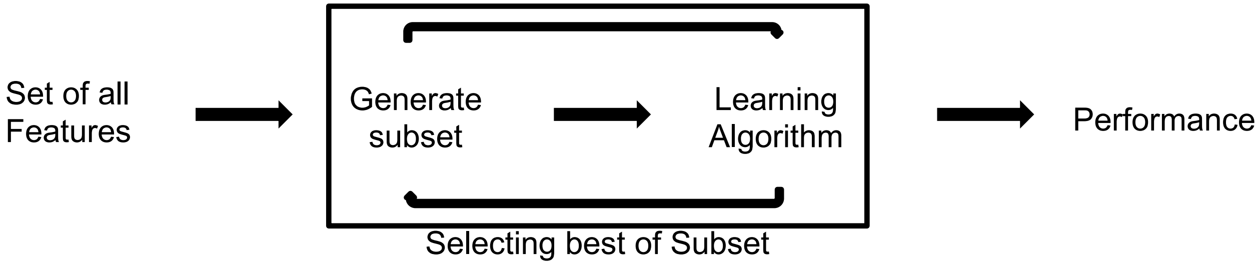


Set of all Features → Selecting the best subset → Learning Algorithm → Performance

- Ex: Chi-Square, LDA, ANOVA

Data Feature Selection…

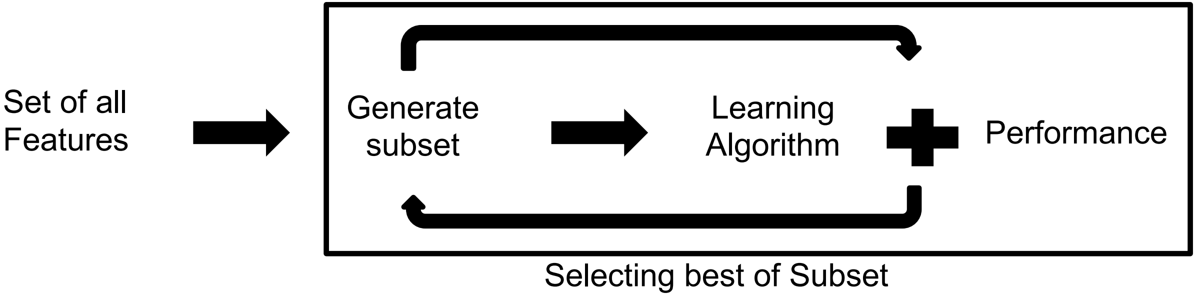
- Wrapper method



- Ex: Forward selection, Backward elimination, Recursive feature elimination, etc.

Data Feature Selection…

- Embedded method


Set of all Features

Generate subset

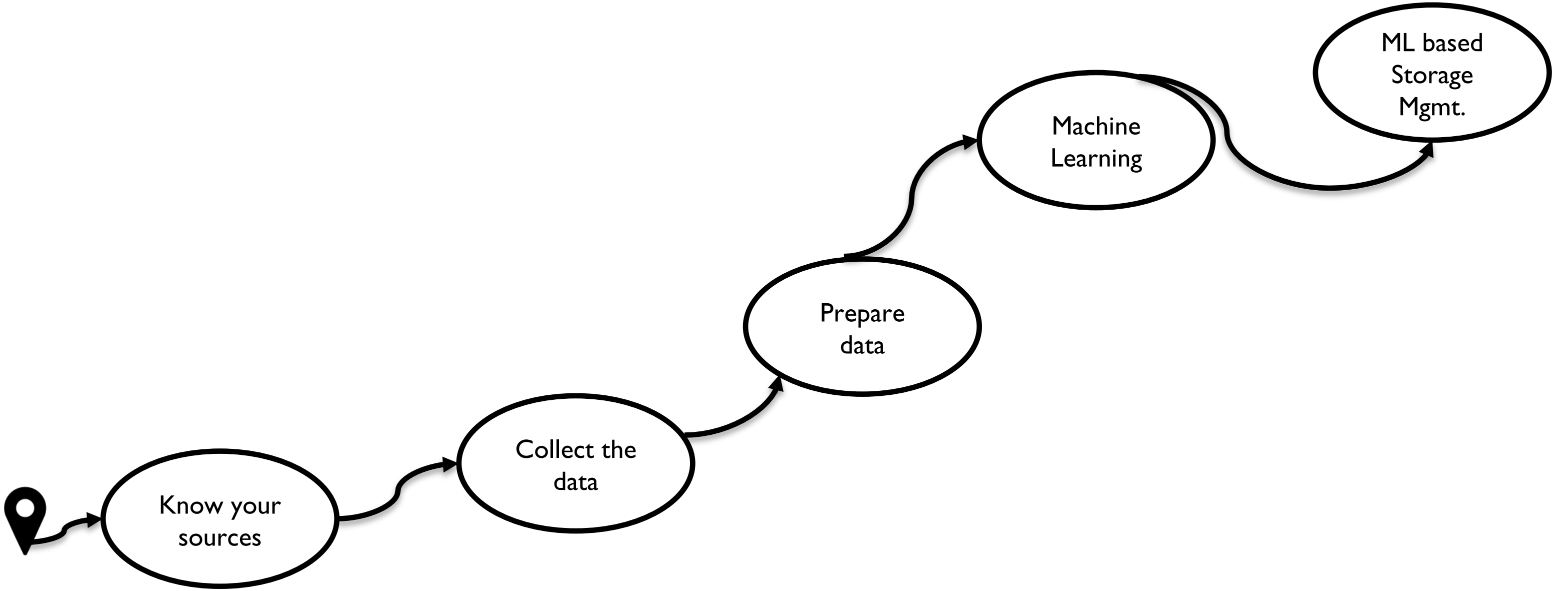
Learning Algorithm

Performance

Selecting best of Subset

  - Ex: LASSOS, Ridge Regression

Apply ML for multiple use cases



Know your sources

Collect the data

Prepare data

Machine Learning

ML based Storage Mgmt.

Apply ML for multiple use cases…

- Case study - Data center management
  - Drive failure prediction
  - Storage tiering suggestion
  - Storage usage trend
  - Storage requirement prediction

Drive Failure Prediction

Workflow

- ❒ What is S.M.A.R.T?
- ❒ Data set
- ❒ Periodic collection
- ❒ Selection features / attributes
- ❒ Applying Support-vector-machine to predict drive failure
- ❒ Outcome

Data set (Features)

- List of S.M.A.R.T attributes
  - Read Error Rate
  - Reallocated Sectors Count
  - Spin Retry Count
  - End-to-End error
  - Temperature
  - Command Timeout
  - Reallocation Event Count
  - Uncorrectable Sector Count
  - Soft ECC Correction
  - G-Sense Error Rate
  - Loaded Hours

- Sampling data (Periodic collection)

Hours	Temp1	......	ReadErr	Servo4
2633	58		6	2944
2635	57		13	2688
2637	56		36	5189
2639	57		0	4032
2641	56		0	8384
.	.		.	.
.	.		.	.
2855	58		14	3322
2857	59		20	2624

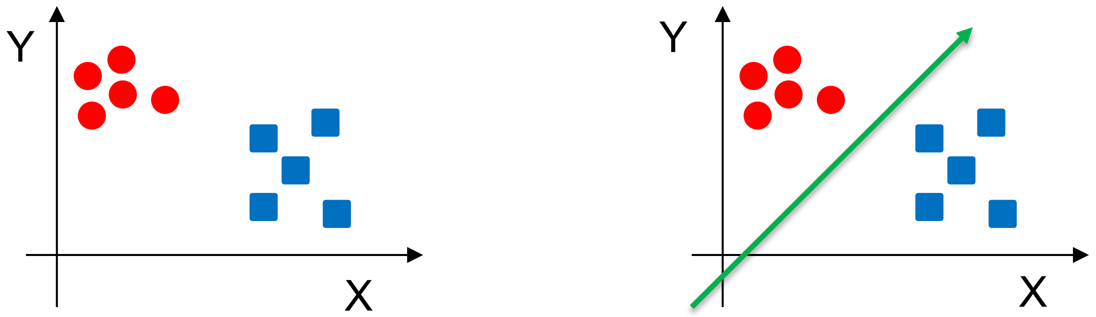
Prepared Data

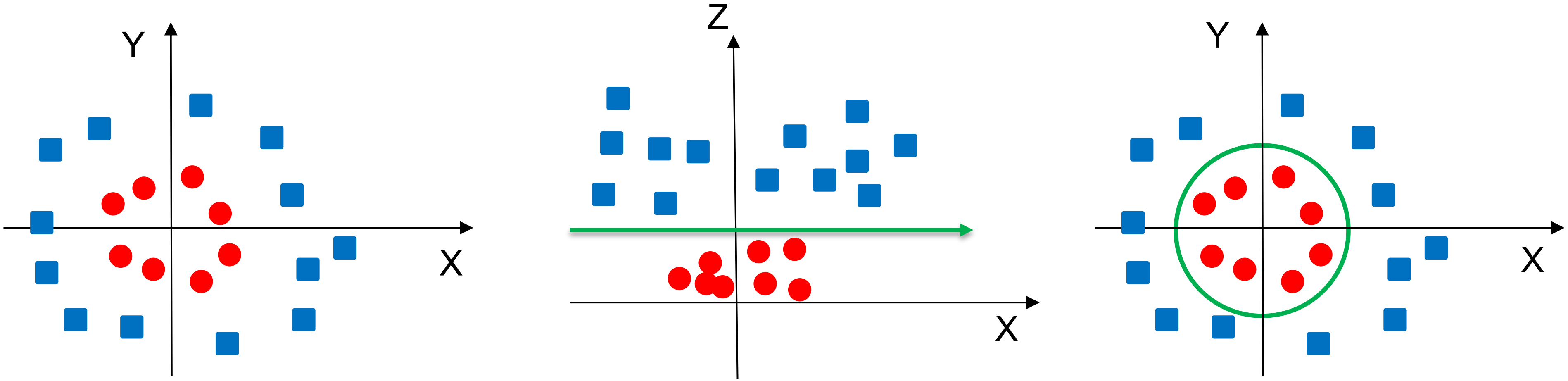
- Feature selection

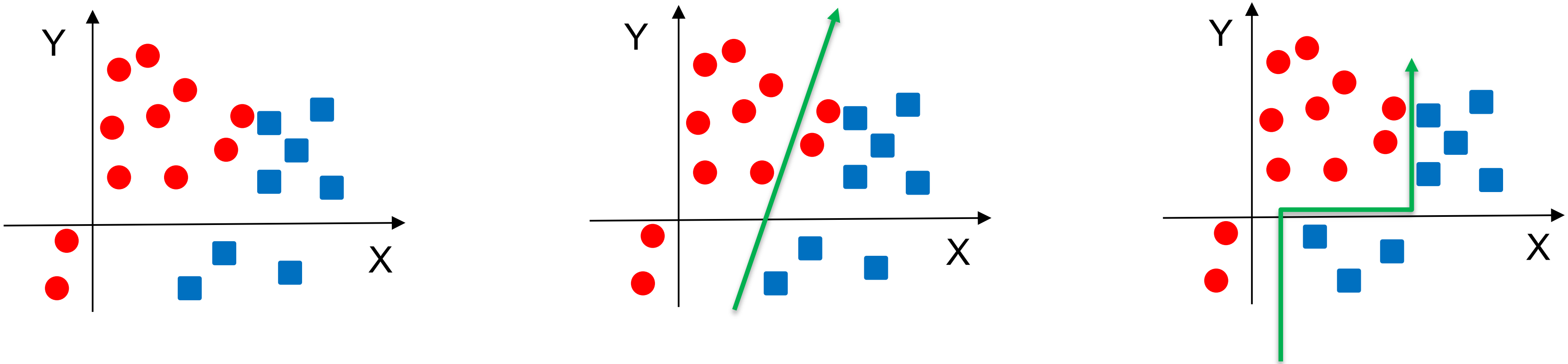
Attribute	% Good	% Failed
Temp1	11.8	48.2
Temp3	35.2	45.3
Temp4	8.8	59.2
Glist	0.5	8.8
ReadError1	0.4	0.8
WriteError	0.8	2.3
Reallocated sector	5.8	30.2
Uncorrectable sector	4.8	34.5
Spin-up time	5.2	14.2
Command timeout	6.2	29.8

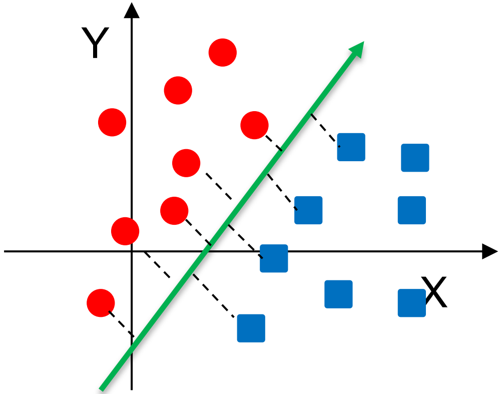
Machine Learning Model

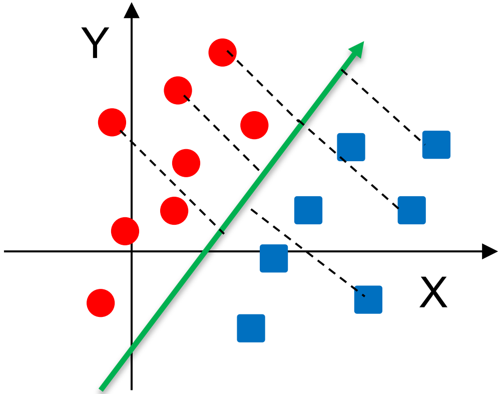
- Support-vector-machine *(supervised learning)*
  - *A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data, the algorithm outputs an optimal hyperplane which categorizes new examples.*


Y
X
Y
X


Y
X
Z
X
Y
X

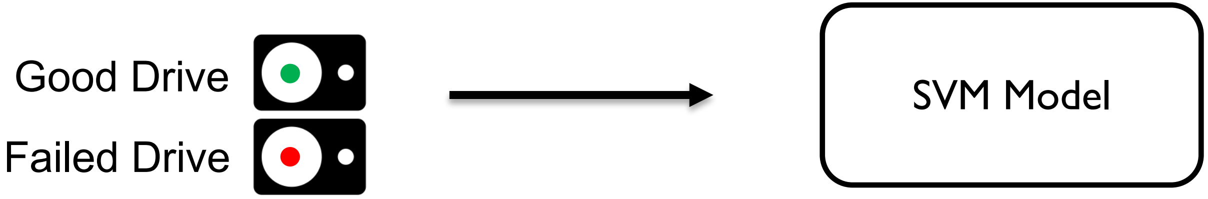

Y
X
Y
X
Y
X


Y
X

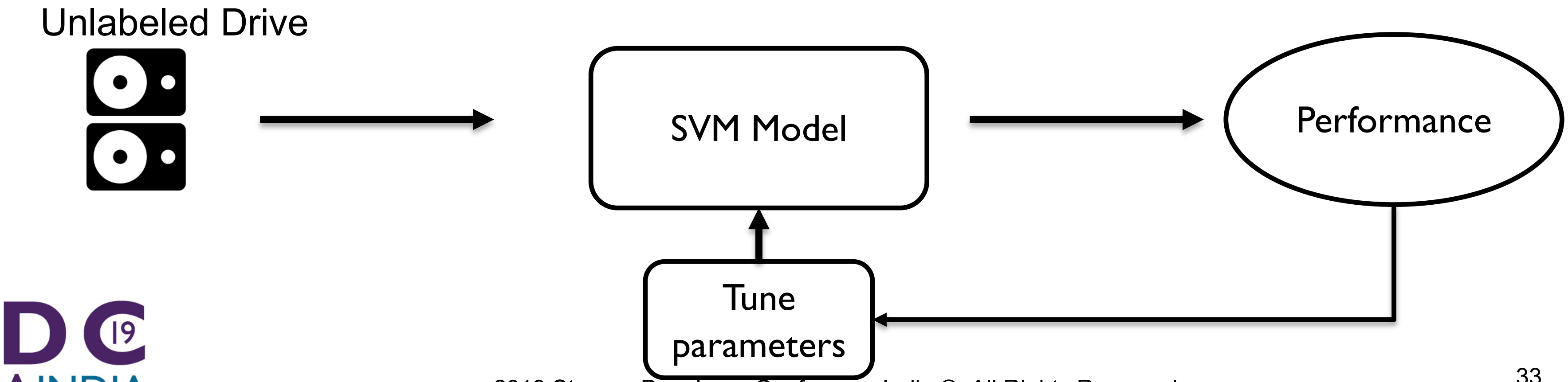

Y
X

Machine Learning

- Train the model (80%)



- Test the model (20%)



Outcome

- Reduce false alarm of failure
- Automated policy implementation

Storage Tiering Suggestion

Storage Tiering

- Physically partitioned into multiple distinct classes based on price, performance or other attributes
  - Swordfish – ClassOfService
- Data may be dynamically moved among classes within a tiered storage implementation

Storage Class (SNIA)

- Media class
  - High performance SSD/Cache
  - High performance HDD
  - High capacity HDD
  - Tape

- Data class
  - Mission critical
  - Hot
  - Cold
- Pricing class
  - Networked storage
  - DAS
  - Cloud

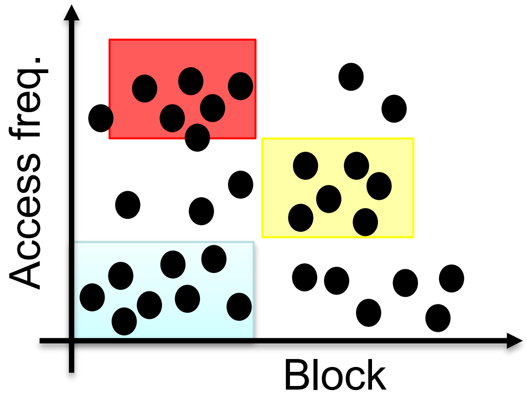
Feature Collection

- ❐ Data/Block access frequency
- ❐ Last accessed
- ❐ Last modification time
- ❐ Size of object
- ❐ Encryption
- ❐ Drive type (HDD, SSD)
- ❐ Drive interface type
- ❐ Drive temperature
- ❐ Caching information
- ❐ …

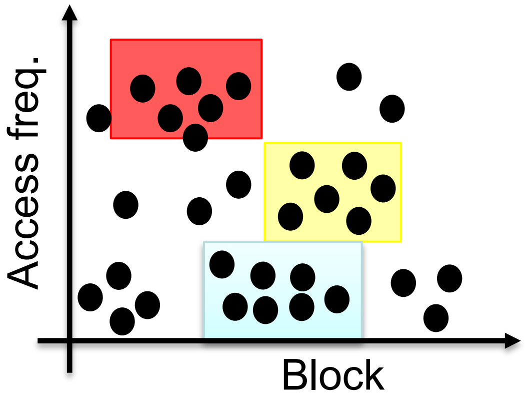
Machine Learning Model

- k-means (unsupervised learning)
  - clustering
  - centroid
  - aggregation

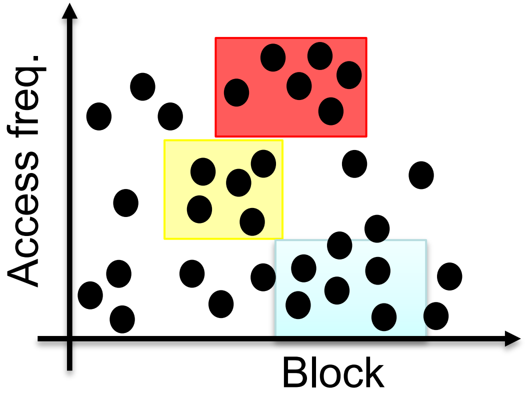
Machine Learning



On Cache



On SSD

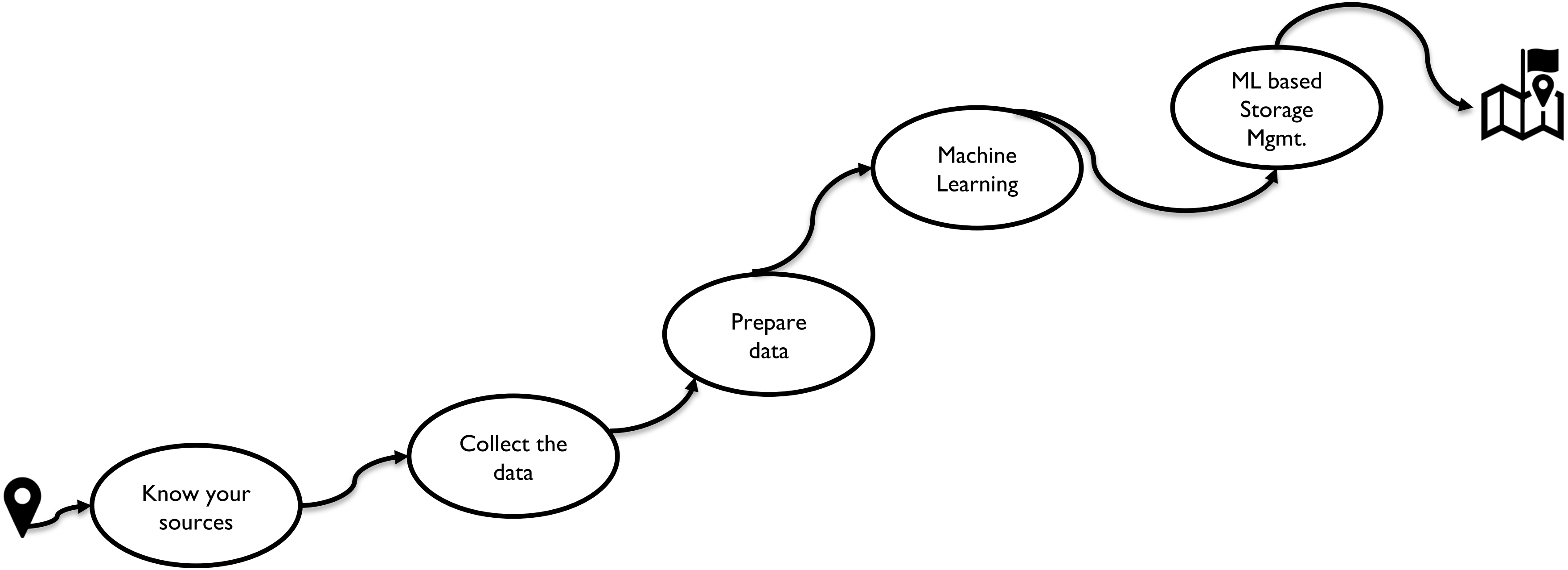


On HDD

Storage Usage Trend

- How Storage is used in a data center over a time period.
- Time Series Algorithms
- Ex of trends
  - Which class of service is more used in future
  - Which media class will be more used in future

Data Visualization



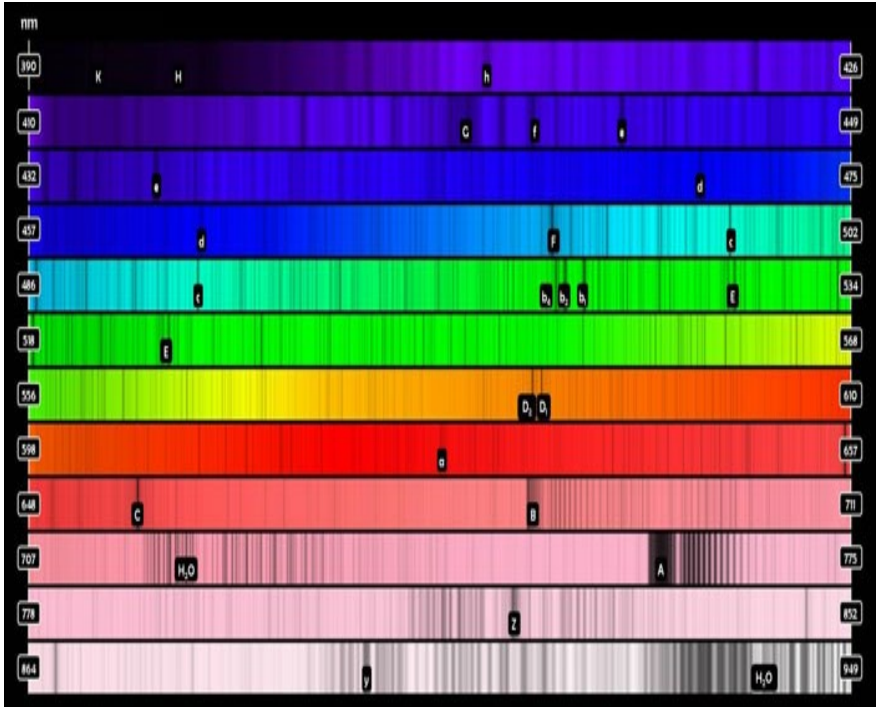
Data Visualization

- Tools available
  - Power BI, Tableau, Data Dog – commercial
  - Dash (personal favorite!) – opensource
- Dash
  - Pure python based framework
  - Abstracts away all technology and protocol
  - Ideal for building data visualization apps
  - https://plot.ly/products/dash/

SDC 19 SNIA INDIA

An image of Halley's Comet taken in 1986. (Image: © NASA)

Fraunhofer lines, in astronomical spectroscopy



Thank You!

✉ sharath.ts@microchip.com

in https://www.linkedin.com/in/sharath-ts-5720a520

SDC 19
SNIA INDIA