# ACE: Classification for Information Lifecycle Management

Presenter: Li Yin (UC Berkeley/IBM Research)

Authors: Gauri Shah (IBM Research)
Kaladhar Voruganti (IBM Research)
Piyush Shivam (Duke University)
Maria Alvarez (UC Santa Barbara)

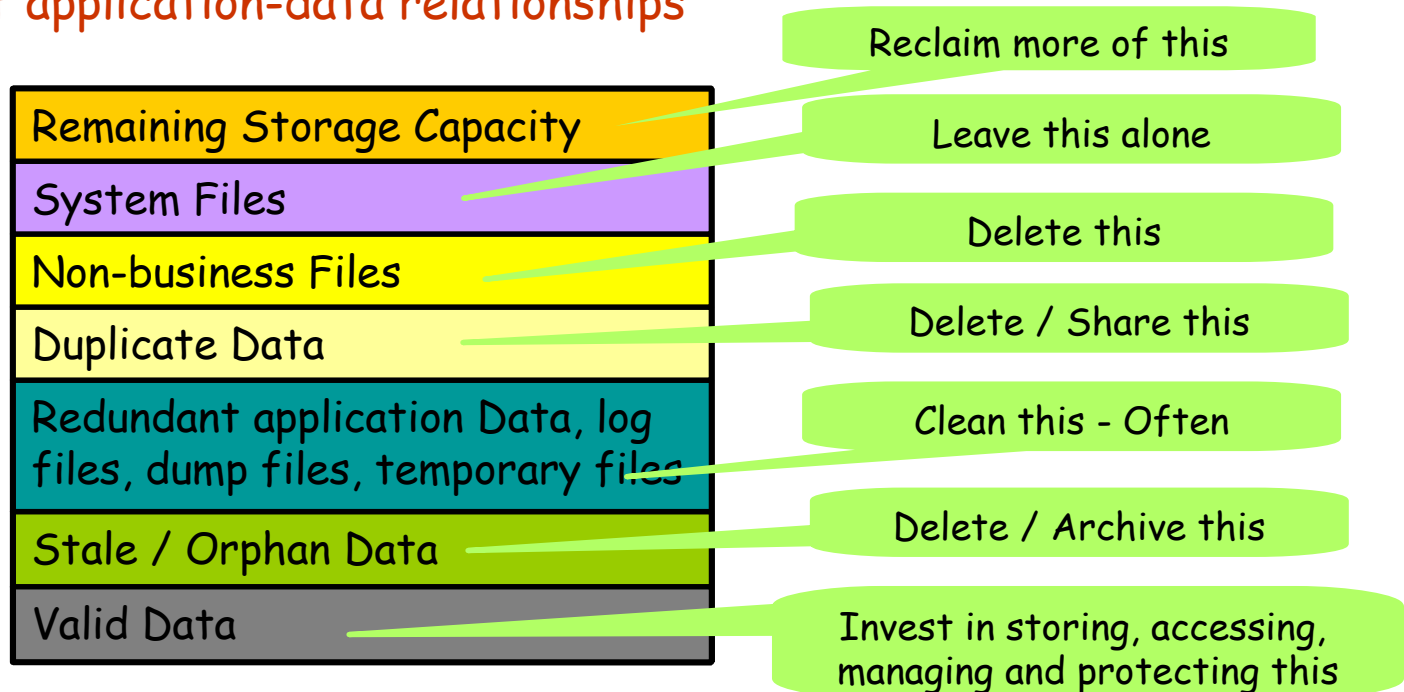# Need for Information Lifetime Management

- Storage needs of most enterprises are growing [IBM study, 2001]
  - Business automation, Compliance requirements

- Percentage of useful data on expensive storage systems is small

- Store and manage data appropriately
  - Value of data changes over time
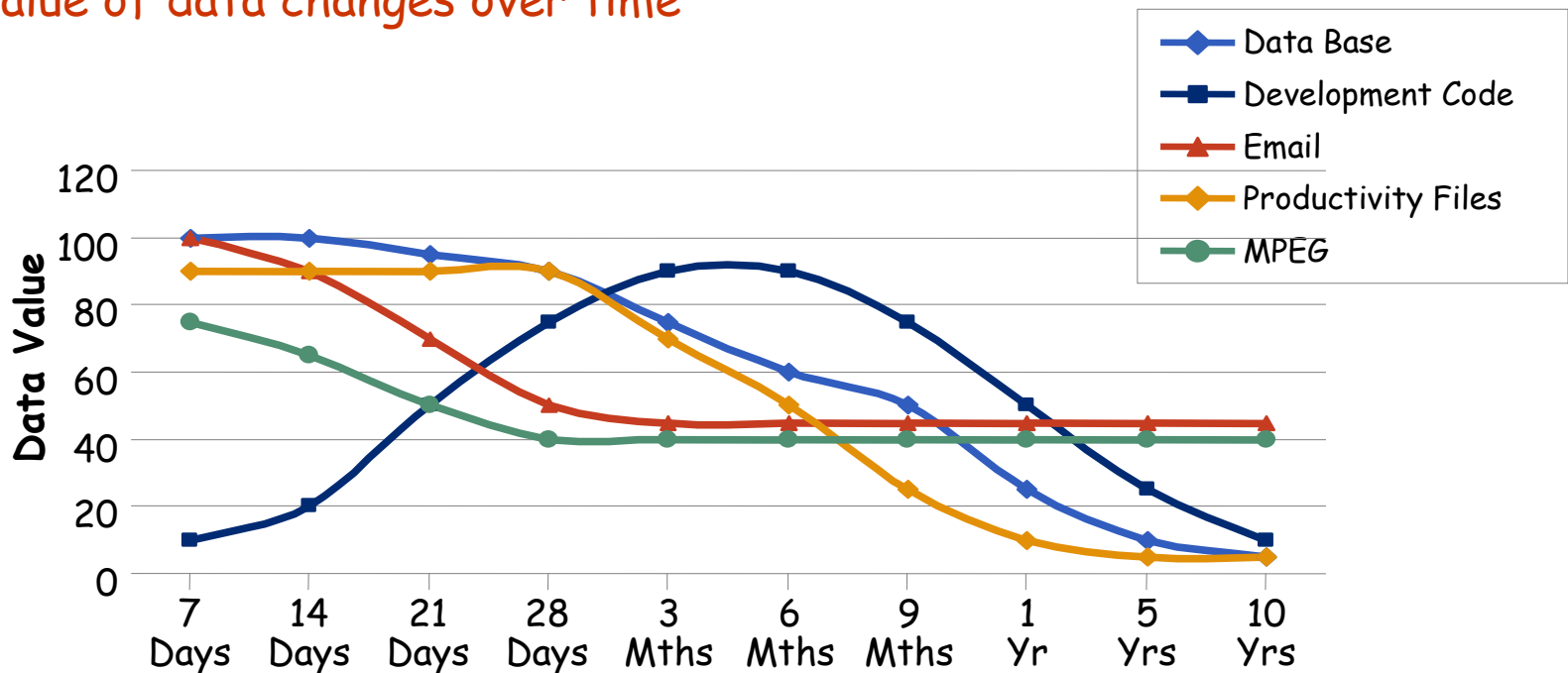
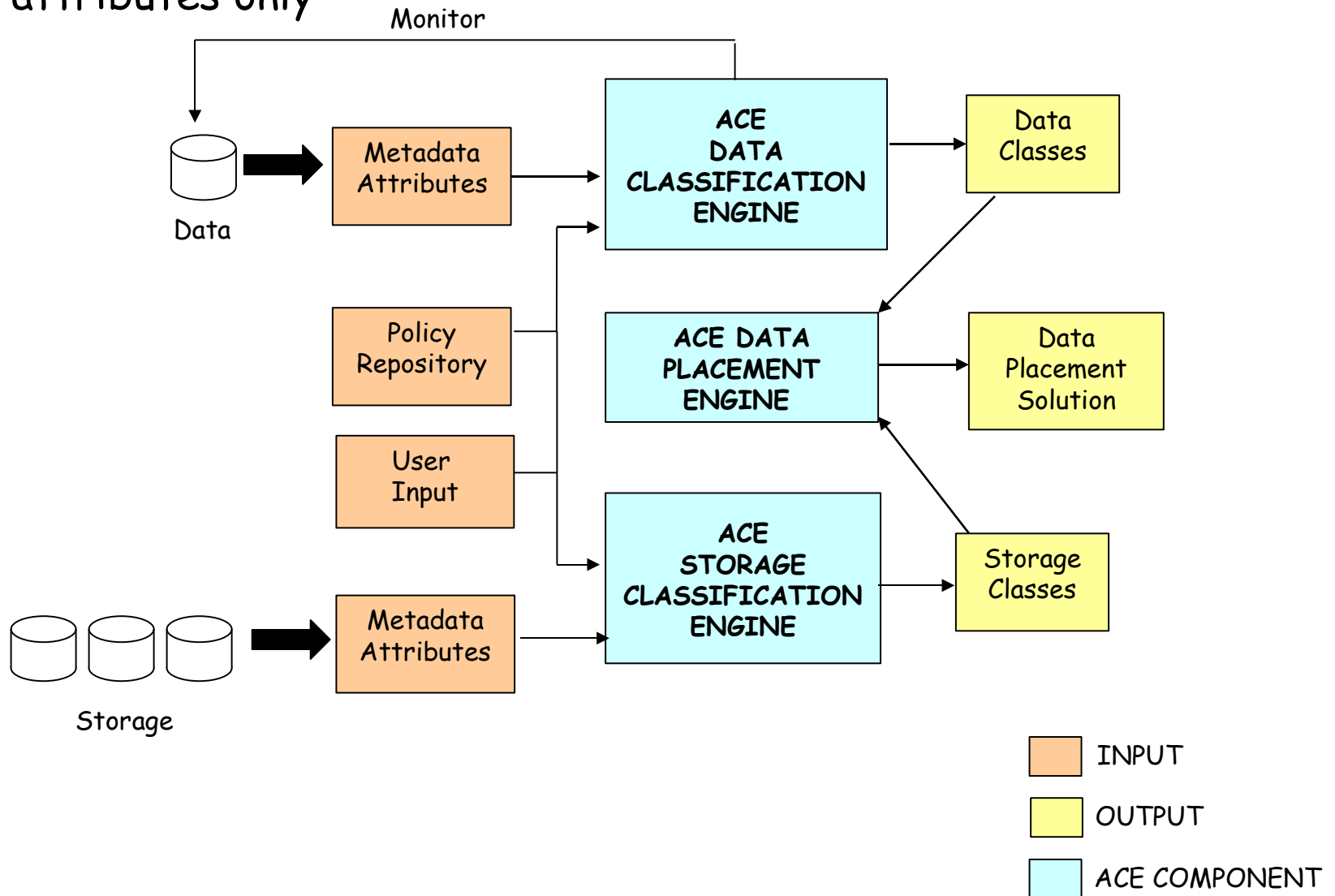Source of graph: Enterprise Storage Group

Challenges of Information Lifecycle Management

- How to manage both data and storage resources over time to make informed use of resources?
  - Right type of data residing on right type of storage at the right time

- Hard problem:
  - Non-triviality of data valuation
  - Temporal nature of business value
  - Time consuming – large volume of data and heterogeneous storage
  - Lack of application-data relationships

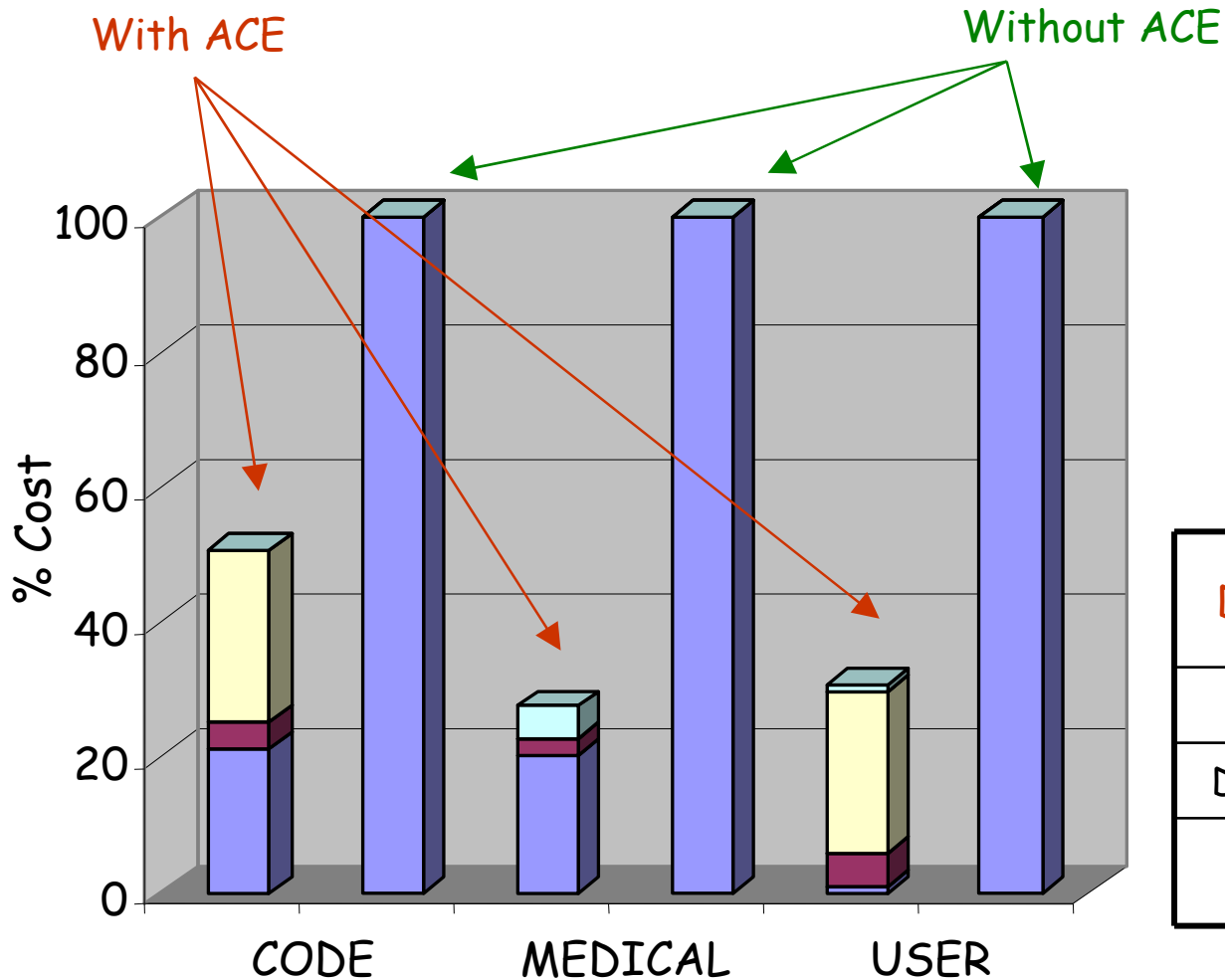|	
Remaining Storage Capacity	Reclaim more of this
System Files	Leave this alone
Non-business Files	Delete this
Duplicate Data	Delete / Share this
Redundant application Data, log files, dump files, temporary files	Clean this - Often
Stale / Orphan Data	Delete / Archive this
Valid Data	Invest in storing, accessing, managing and protecting this

ACE Architecture

- Policy Driven Classification, using data and storage metadata attributes only

Policy-driven classification

- **Knowledge-based policies:** Predefined policies, e.g., files with extension .ppt have business value 9, storage with continuous copy is assigned to the highest service class.

- **Expert-based policies:** User specified ranking of metadata attributes and their values, e.g., higher ranking to owner Administrator and lower ranking to guest.

- **Example-based policies:** Obtain policy using an example data set with known business values using machine learning techniques.

Sample Cost Benefits using ACE

With ACE

Without ACE

Cost per GB

☐ Tape $1.2

☐ Low-end disk $7

☐ Mid-range disk $13.4

☐ Enterprise $20

% Cost

100

80

60

40

20

0

CODE   MEDICAL   USER

Data Set	Total Files	Total Size
CVS files	12796	316 MB
Documents	32800	18.4GB
Medical recs.	18.4 GB	1461 GB