EMC Greenplum
Driving the Future of Data Warehousing and Analytics

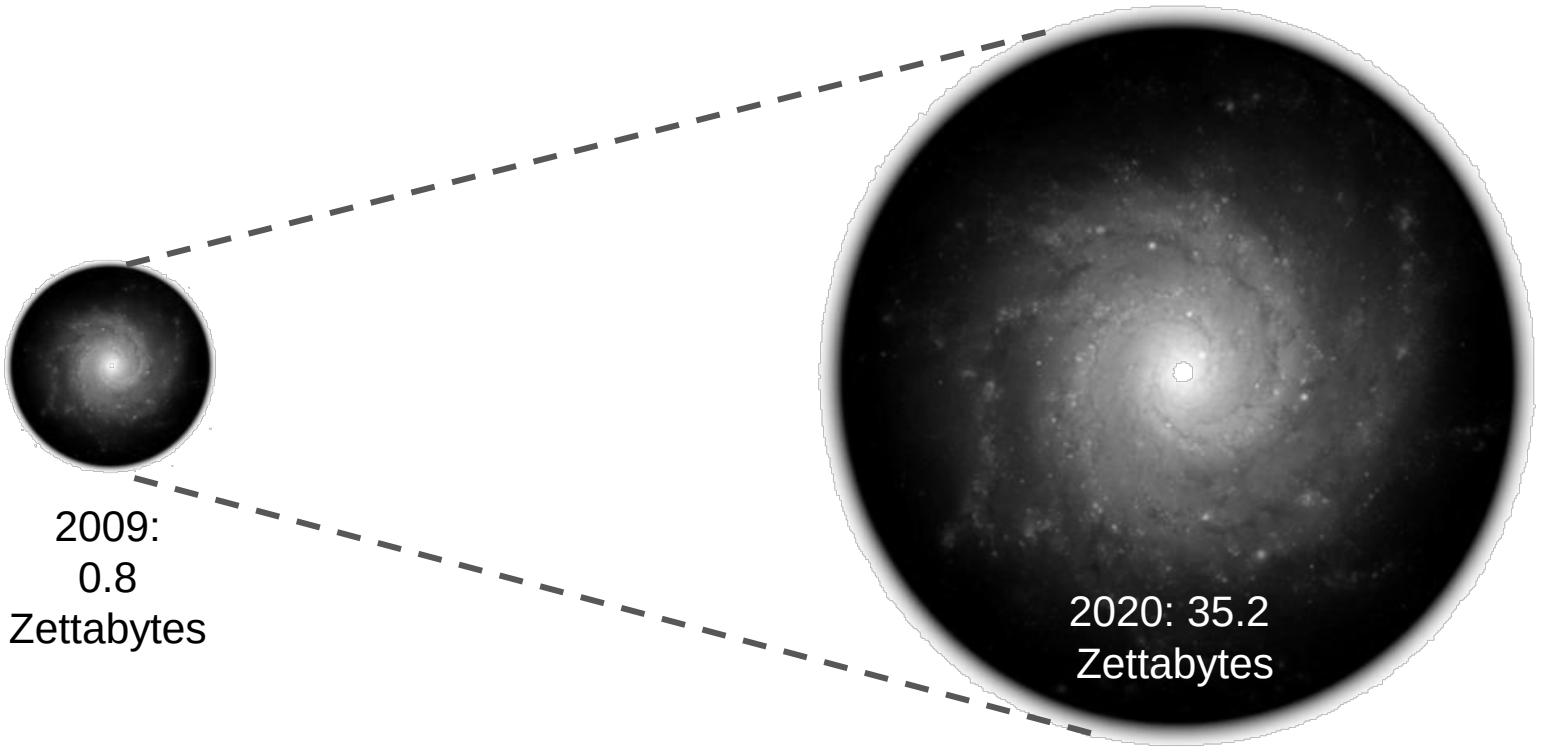
Tools and Technologies for Big Data

Steven Hillion
V.P. Analytics
EMC Data Computing Division

EMC²

Big Data **Size**: The Volume Of Data Continues To Explode

The Digital Universe 2009 - 2020



Source: IDC Digital Universe Study, sponsored by EMC, May 2010

Big Data **Significance**: Not Just For Google and Facebook…

"Just as search engines have transformed how we access information, other forms of *big data computing* can and will transform the activities of companies, scientific researchers, medical practitioners, and our nation's defense and intelligence operations."

Randal E. Bryant
Carnegie Mellon University

Randy H. Katz
UC Berkeley

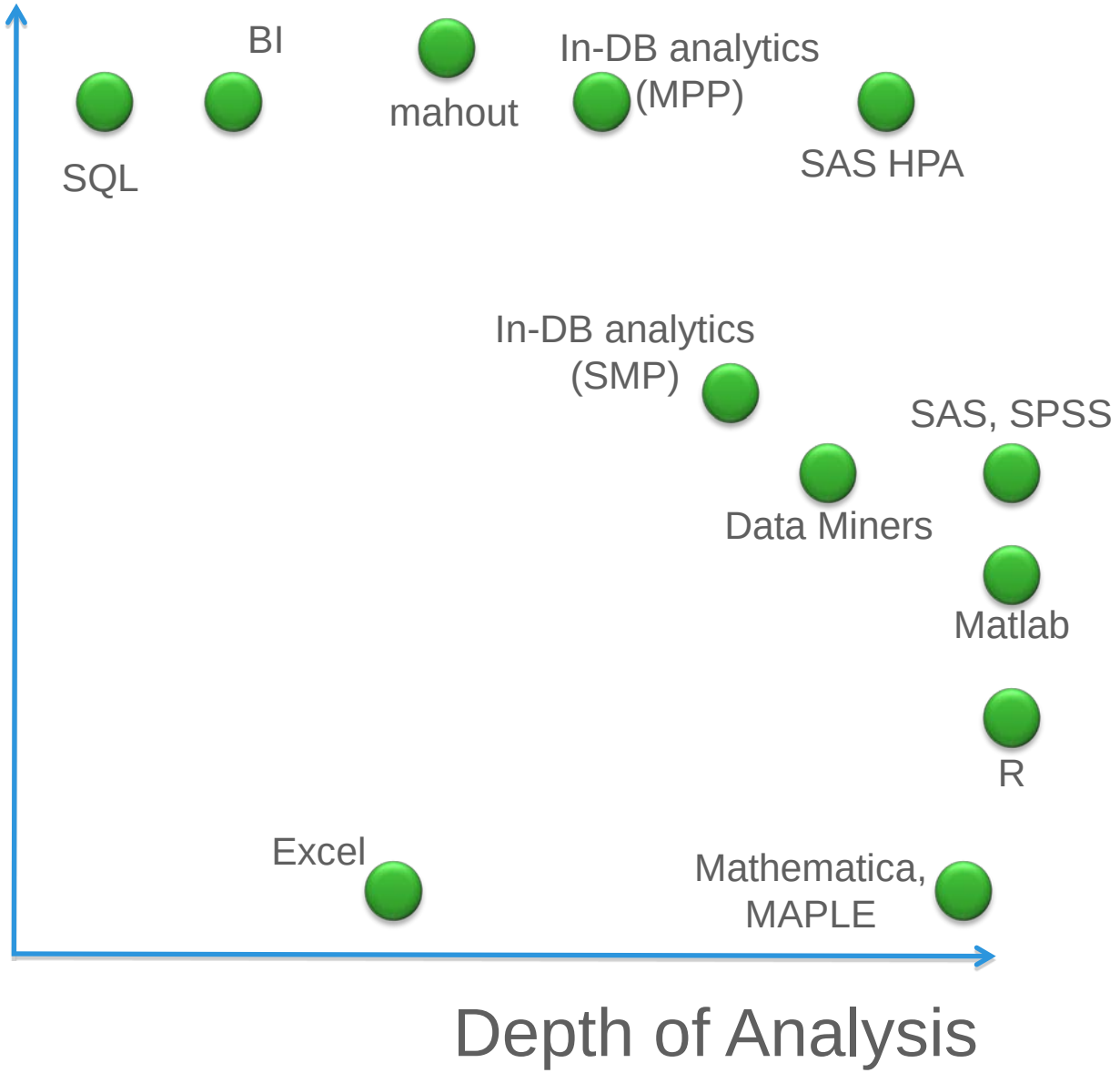
Edward D. Lazowska
University of Washington

Many Requirements, Many Technologies

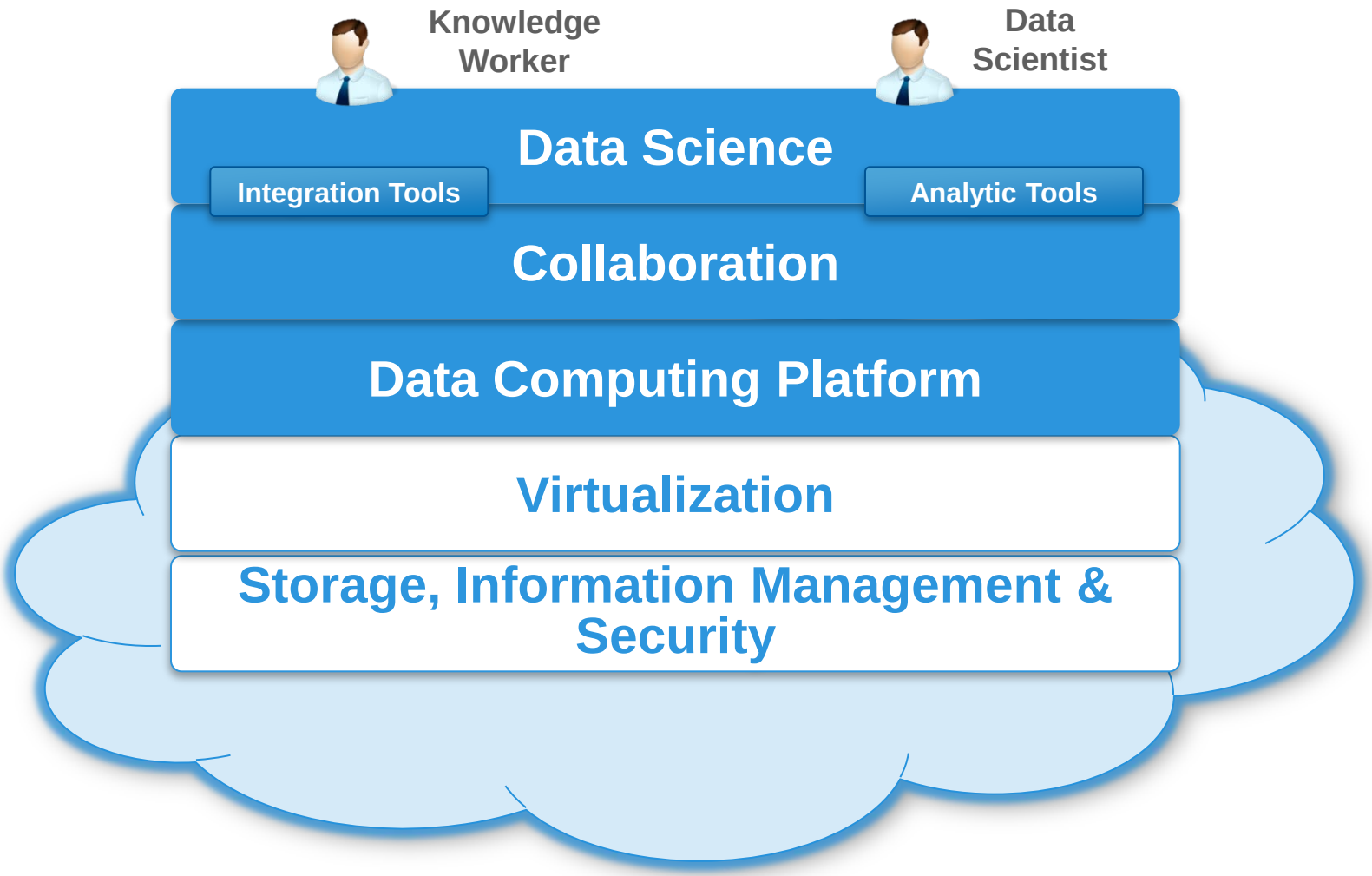
- ETL for structured and unstructured data
- e.g. MapReduce, ETL tools, SQL
- Data storage
- e.g. HDFS, RDBMS
- Computational infrastructure
- e.g. RDBMS, Hadoop, SAS Grid
- Reporting and dashboards
- e.g. Excel, BI Tools
- Data mining
- e.g. SAS, SPSS, R, MADlib
- Model deployment
- e.g. PMML, In-DB scoring
- Visualization
- e.g. Excel, Tableau
- Collaboration and Development
- e.g. Wiki, Sharepoint

EMC²

Scalability



The Big Data Analytics Stack



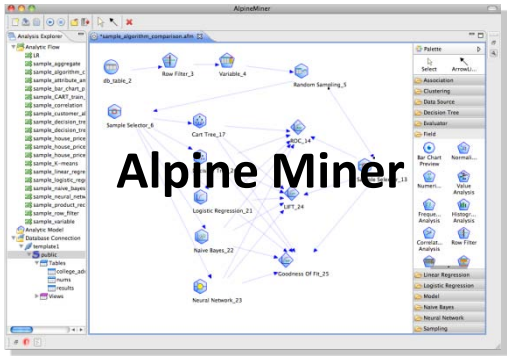
Analytics Leadership

SAS/ACCESS
SAS Scoring Accelerator
SAS High Performance Analytics

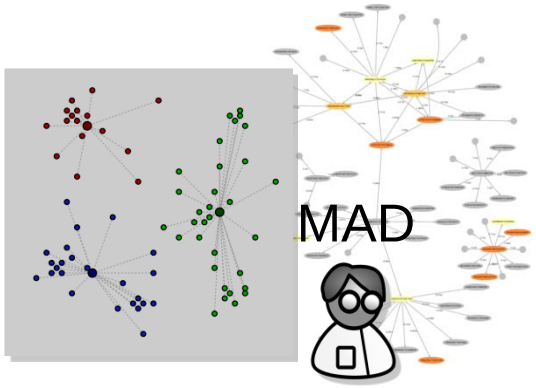
```
sql# SELECT cid, sum(tfxidf)/count(*) AS centroid
     FROM (
       SELECT id, tfxidf, cid,
         row_number() OVER (
           PARTITION BY id
           ORDER BY distance, cid) rank
       FROM blog_distance
     ) blog_rank
     WHERE rank = 1
     GROUP BY cid;
```

In-database Analytics

In-DB MapReduce
Greenplum HD



Alpine Miner

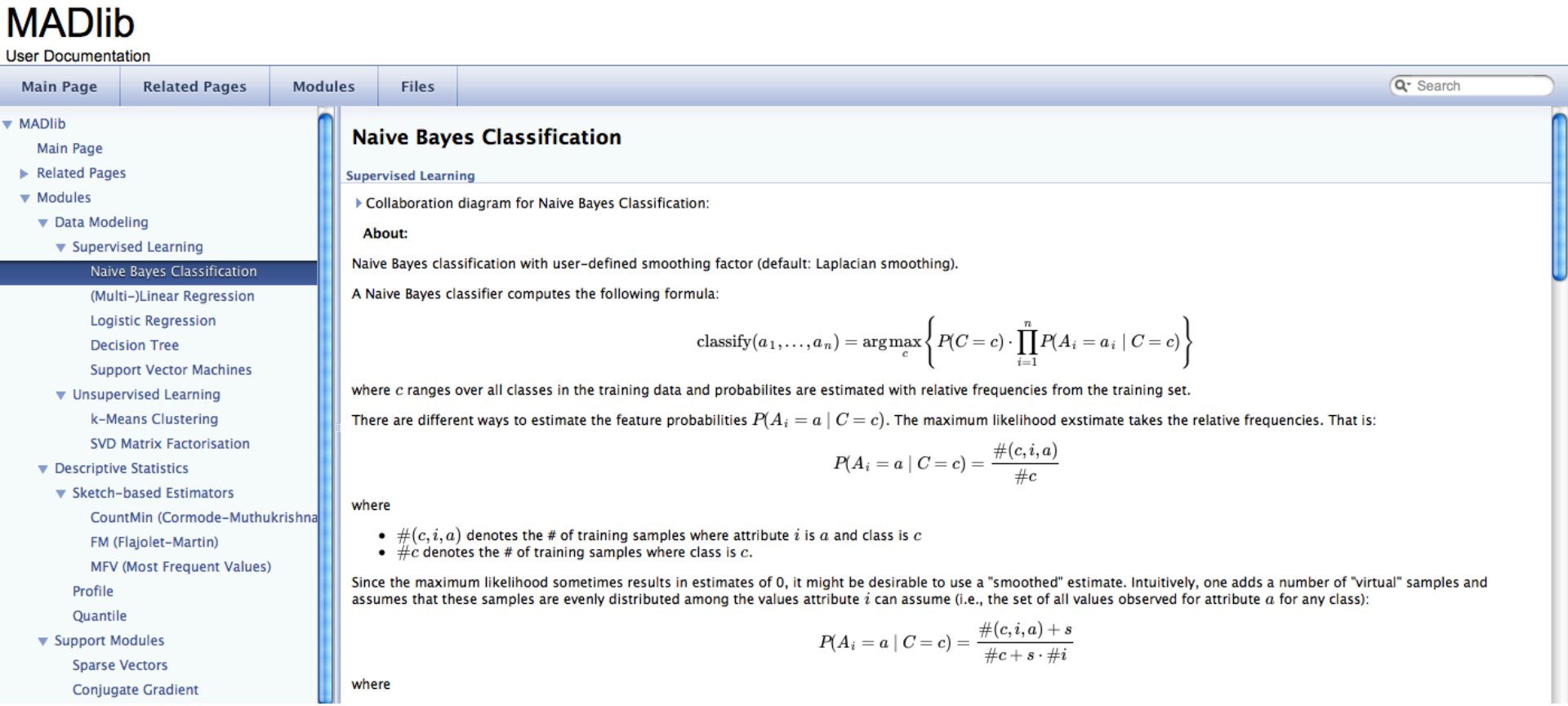


MAD lib

PMML Predictive Model Markup Language

PL/R, Tools integrations

Chorus

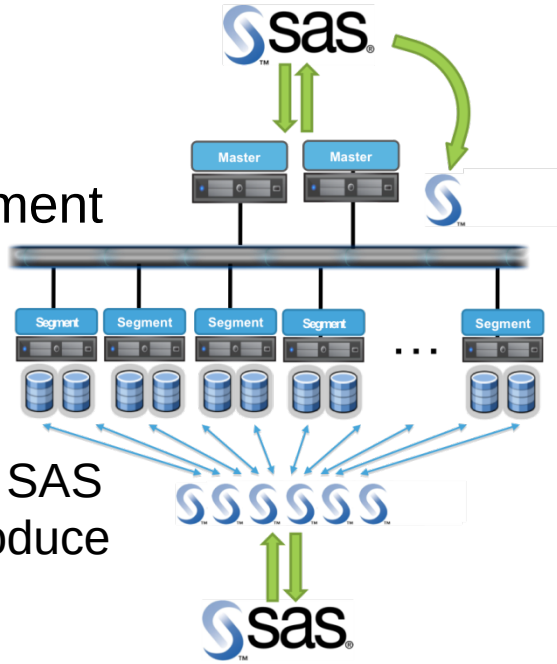


- MADlib is an open-source library for scalable in-database analytics, jointly developed by EMC and UC Berkeley. It provides data-parallel implementations of mathematical, statistical and machine learning methods for structured and unstructured data.

- The MADlib mission: to foster widespread development of scalable analytic skills, by harnessing efforts from commercial practice, academic research, and open-source development.

A Strategic Partnership for High-Performance Computing and MAD Analytics

- Access relational data-sets for agile analysis
  - **SAS/ACCESS** provides fast, transparent and secure access to Greenplum data.
- Leverage database scalability for rapid model deployment
  - **SAS Scoring Accelerator** publishes models for execution in parallel across the Greenplum cluster.
- Build complex models at massive scales
  - The **SAS/Greenplum Analytics Appliance** combines SAS In-Memory Analytics with Greenplum parallelism to produce record-breaking scalability and performance.



Large-Scale, Real-World Analytics

Question	Method
How can I identify fraudulent activity?	Variable Selection, Logistic Regression
How do I segment my customers?	K-means Clustering
Does this product appeal to some segments more than others?	Log-likelihood
Which campaign is working better?	Mann-Whitney U Test
How is product ownership distributed across customer segments?	SQL, Cumulative Distribution Functions
How do I target my marketing efforts towards customers most likely to churn?	Logistic Regression
What new products should I offer my customers?	Cosine similarity, k-Nearest Neighbors
What are my customers saying about the new product launch?	MapReduce, NLP, sparse vectors

Q&A