ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

The Evolution of "Big Data"

Andrew Fast, Ph.D.
Chief Scientist
fast@elderresearch.com

Headquarters
300 W. Main Street, Suite 301
Charlottesville, VA 22903
434.973.7673 | fax 434.973.7875

Office Locations
Arlington, VA
Linthicum, MD
Raleigh, NC

The Basics

- This is a computer.
- A computer can run advanced algorithms for processing data
- Advanced algorithms can produce significant value for organizations.

Three Key Resources

CPU
Runs the Algorithms

Memory
Caches Data

Disk
Data Storage

OK, four…

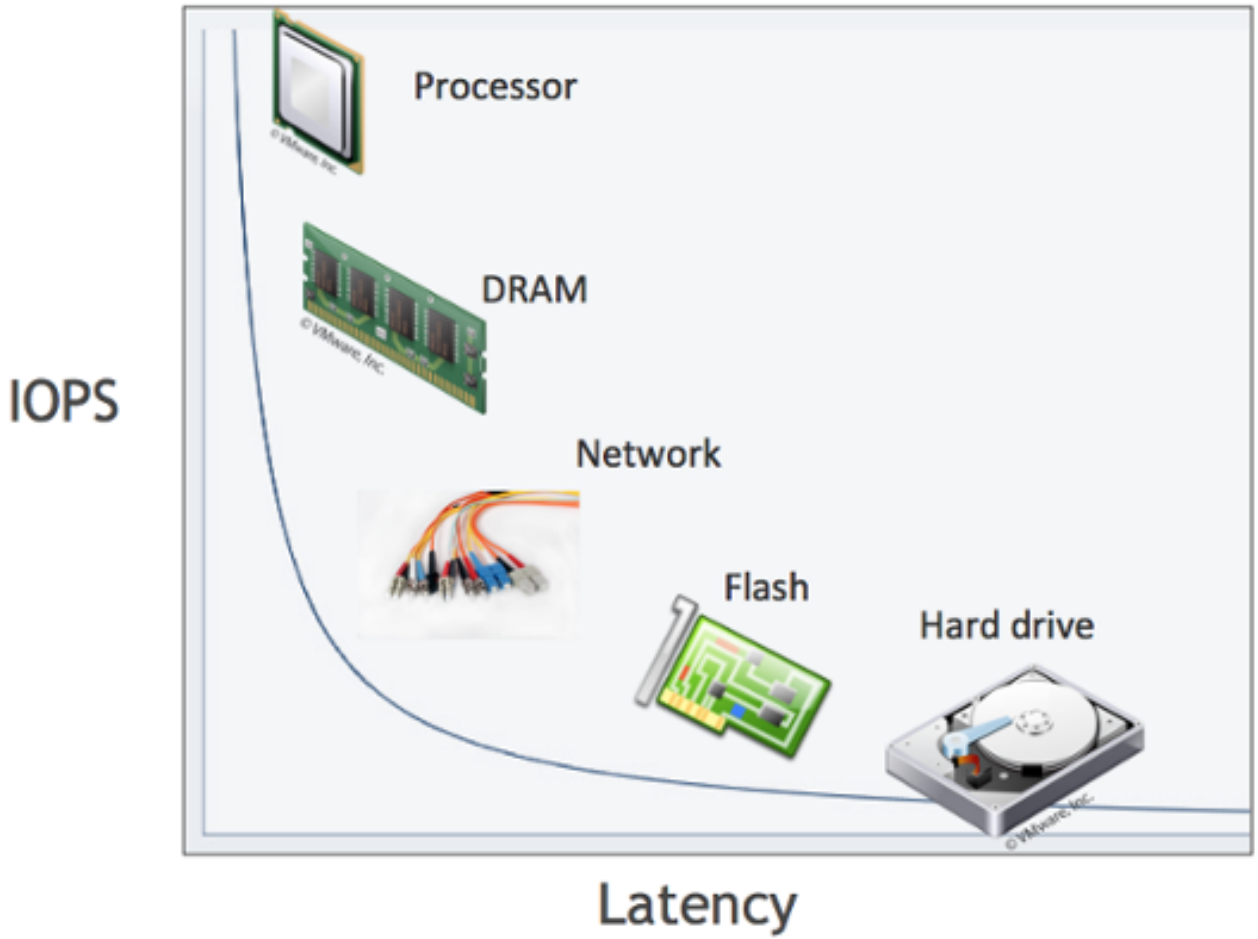
Four Vs of Big Data

Volume
Velocity
Veracity
Variety

My definition…

- *"Big Data":* data that requires more processing resources to produce value than are available within the constraints of the current business problem.

Speed Matters



http://blog.infinio.com/relative-speeds-from-ram-to-flash-to-disk

Speed Matters

If Memory = Minute

Network = Weeks

Flash = Months

Disk = Decades

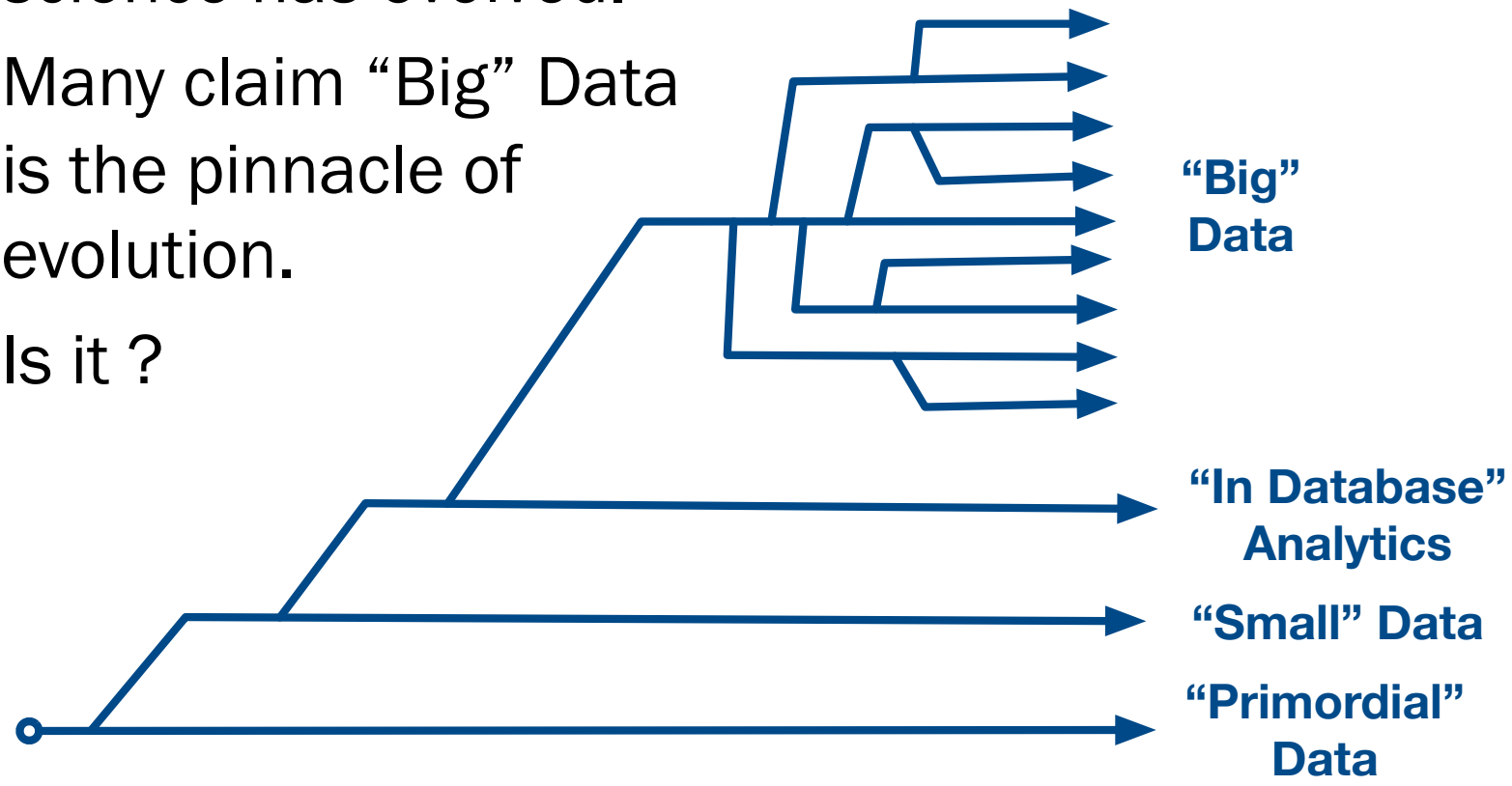
http://blog.infinio.com/relative-speeds-from-ram-to-flash-to-disk

A Word about Processors

- For the purposes of this talk, processors are the algorithms and methods used to process and summarize the data.
  - Statistical Models
  - Query Languages
  - Etc.

The Tree of Life

- The world of data science has evolved.
- Many claim “Big” Data is the pinnacle of evolution.
- Is it ?



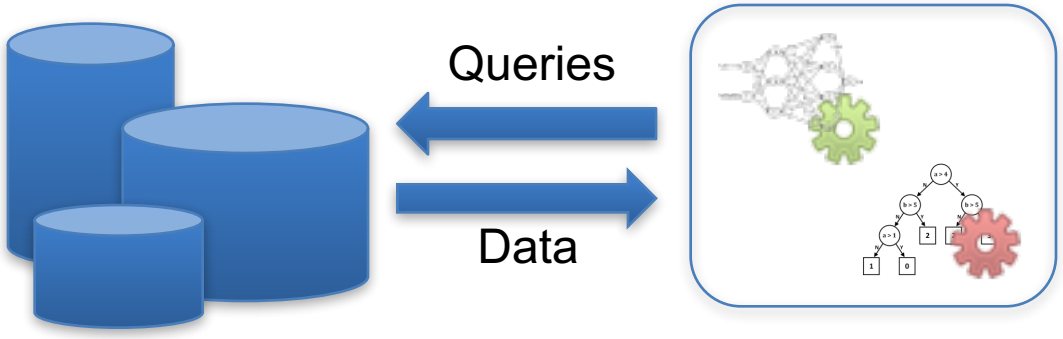
“Primordial” Data

- Characterized by data and processing all contained on a single machine.

- What could possibly go wrong?
  - Well, data grows.

“Small” Data

- Move storage off the local machine into a database
- Combines storage in RDBMS with in-core analytics
  - Limited interaction between Data Storage and Analytical Processing

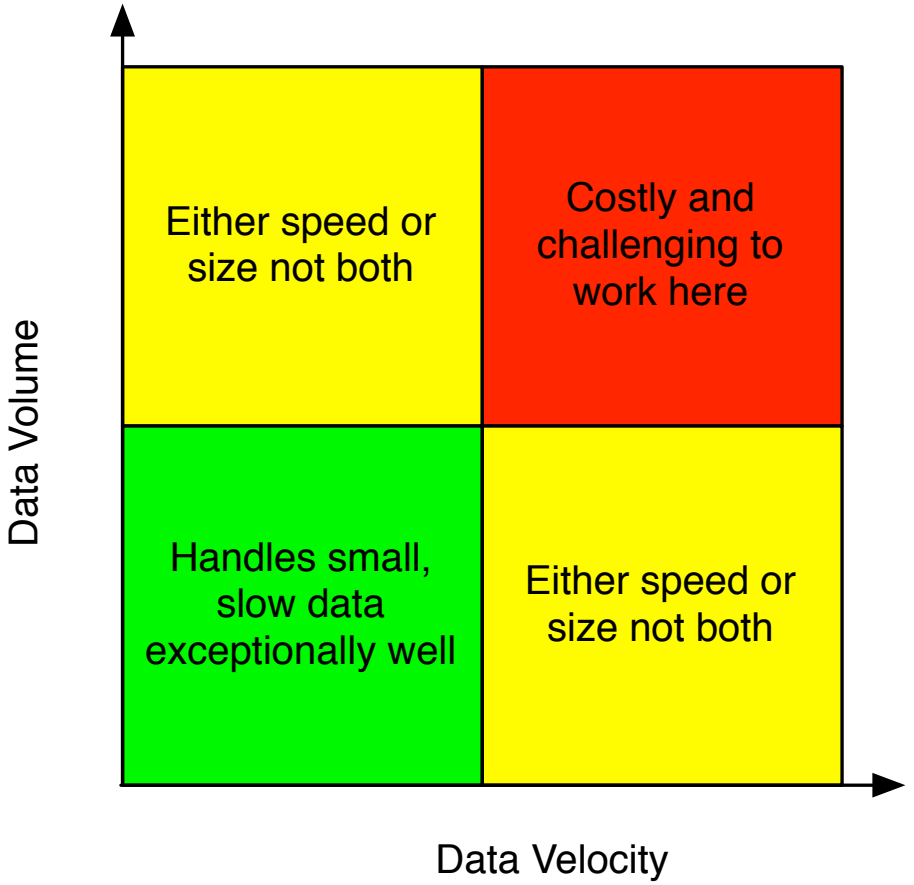


Relational Databases (Oracle, SQLServer, MySQL)

In-core Analytics (SAS, IBM, R, Python, Excel, etc.)

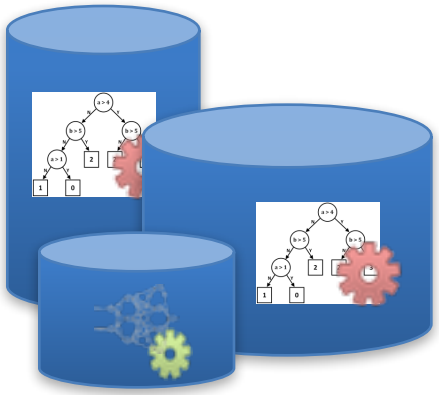
“Small Data”

- Strengths:
  - Mature Technologies
  - Defined processing standards
- Limitations:
  - Requires expensive, proprietary hardware to scale to large and/or high-velocity data
  - Algorithms limited by slow data access



“In-Database Analytics”

- Remove data transfer burdens and single-machine limitations by bringing analytics to the data
  - Database-specific implementations of analytics algorithms
  - Recognition that storage is not the only factor



Relational Databases and Data Appliances (SQLServer, PostgreSQL, Oracle ExaData Teradata, Netezza, etc.)

“In-Database Analytics”

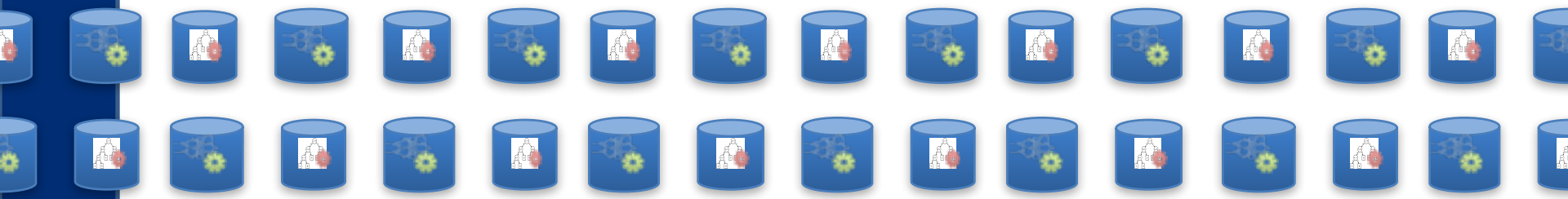
- Strengths:
  - Improves speed of analytics through reduced data transfer
  - Optimized performance
  - Mature storage technology
- Limitations:
  - Implementations are database specific
  - Not all analytic algorithms lend themselves to in-database analytics
  - Expensive, proprietary hardware and/or software usually required

What could possibly go wrong?

- Data grows (sometimes very quickly)!
  - RDBMSs and Data Appliances are still costly to upgrade when more space is needed
- Vendor lock-in due to proprietary system
  - Innovation happens in bursts, rarely at industry leader

“Big” Data

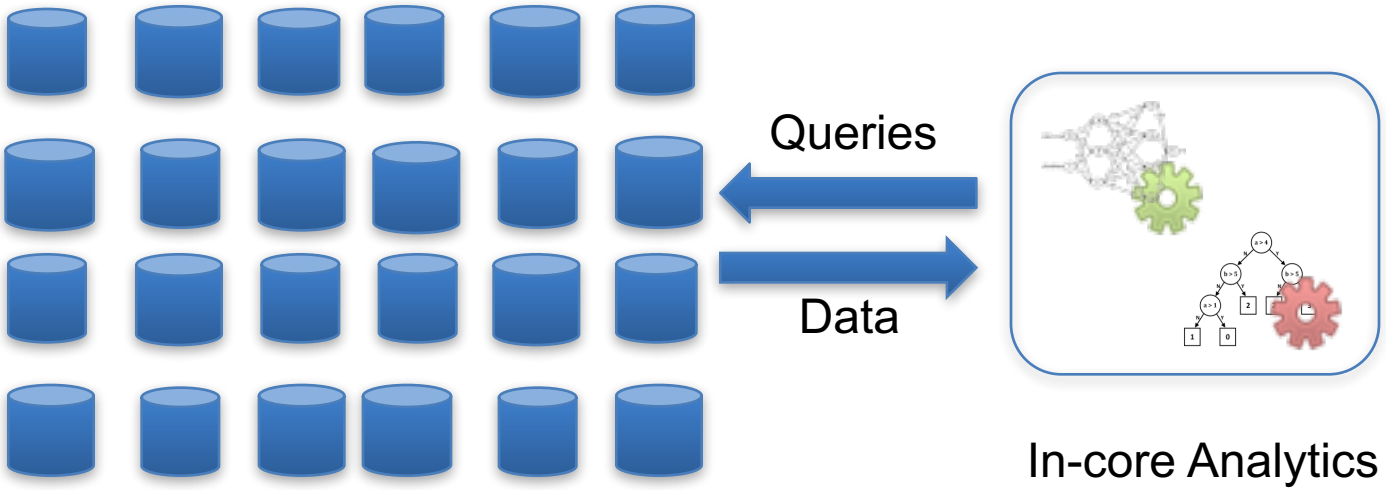
- Collection of largely open-source technologies for large-scale data storage and processing on commodity hardware
  - Massively parallel integration of storage and processing
  - Analysis of extremely large datasets now possible on commodity hardware



Massively Parallel storage and processing
on commodity hardware
(NoSQL, Hadoop, Hive, Cassandra, etc.)

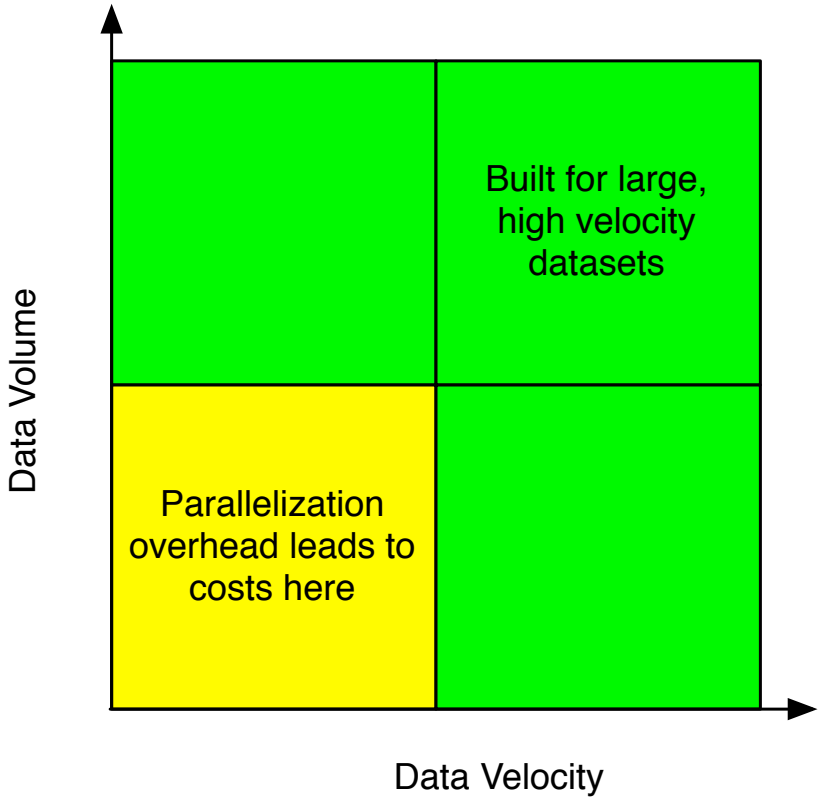
Limited "Big" Data

- It is possible to treat Big Data as only a storage mechanism, not a processing engine
  - Revert back to "Small" data paradigm



“Big” Data

- Strengths:
  - Designed for large, high-velocity data
  - Decoupling of storage and processing, but still at scale
- Limitations:
  - New tools required (not based on SQL)
  - Some analytics are slower



What could possibly go wrong?

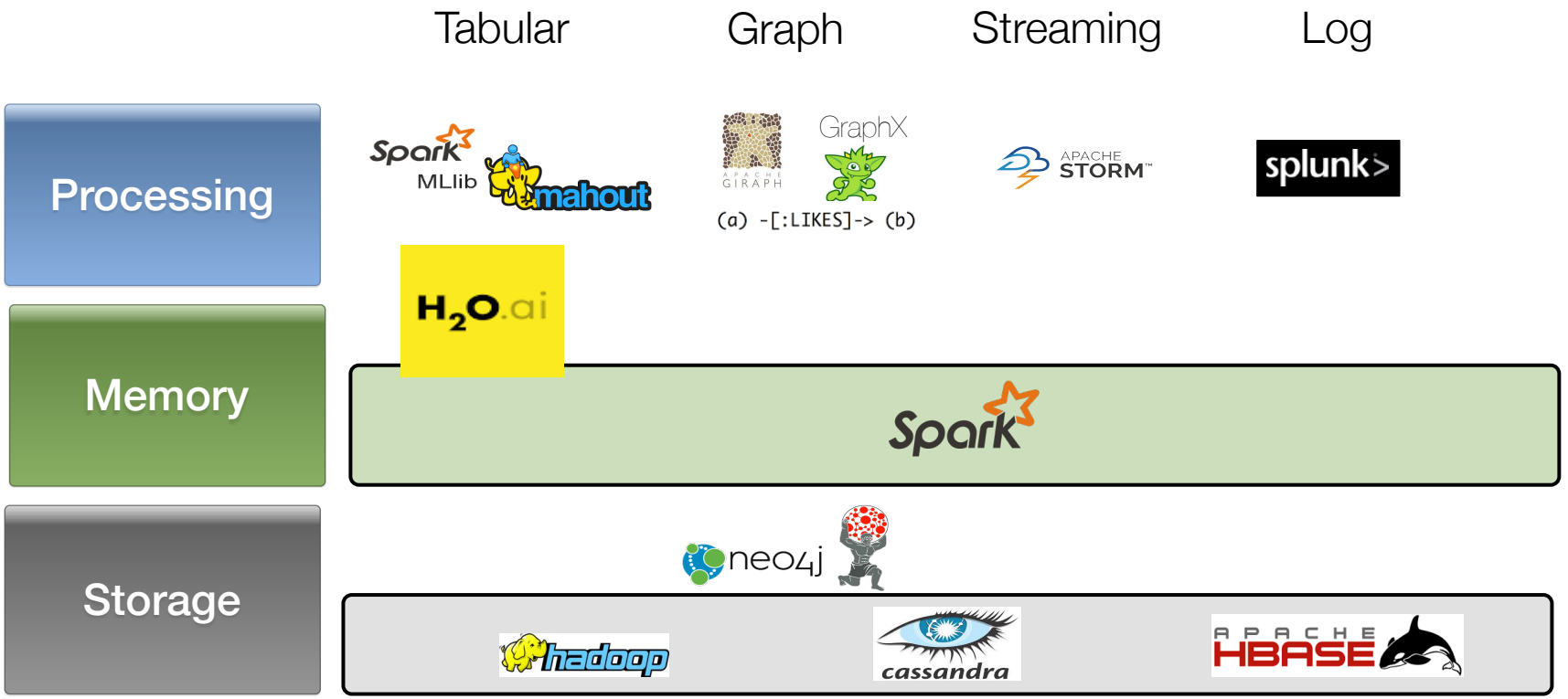
- Most advanced analytics also require multi-variable computations such as correlations or matrix inversions
  - Costly in distributed environment
- Latency can still be an issue

The Missing Link: “Big” Memory

- Big Data solves the storage problem using data distribution on commodity hardware
- Requires Big Algorithms using “in-database” strategies.
  - All analytical processing must be distributed with the data
- Now, “Big” Memory to make it all work fast

No More One-Size fits all

- Successful Big Data software solutions have very specific use cases



List of tools is not an exhaustive list

Four Vs of Big Data

~~Volume~~

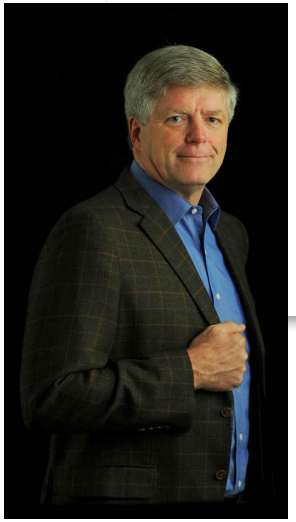
~~Velocity~~

Veracity

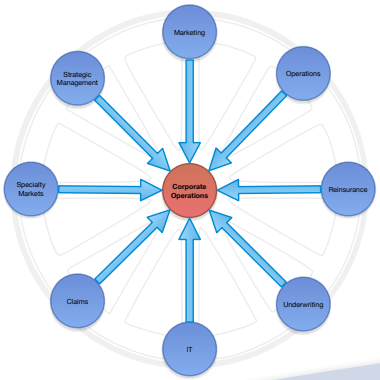
Variety

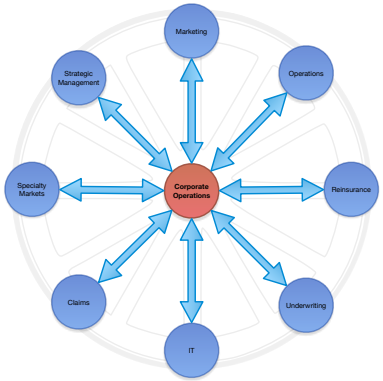
Introducing Analytics 3.0

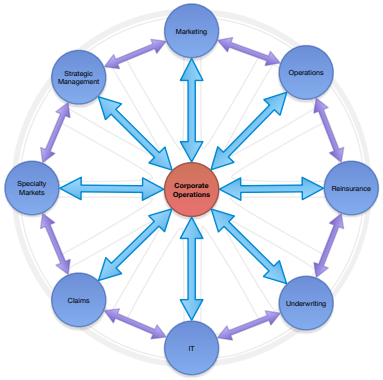
- A new era of analytics
- "A new resolve to apply powerful data-gathering and analysis methods not just to a company's operations but also to its offerings…"
- Coined by Tom Davenport,
  - Distinguished Professor of IT and Management at Babson College
  - Founder of International Institute for Analytics
  - Author of "Competing on Analytics"



Road to Pervasive Analytics







1.0

- "Small" Data
- Business Intelligence
- Reporting

2.0

- "Big" Data
- Integration of external data sources
- Visual and Descriptive Analytics

3.0

- Advanced Analytics
- "Big" and "Small" Data together
- Analytics central to strategy

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

Variety is the Key to Value

- Big Data systems "integrate information from varied sources for deeper/broader understanding"
  - Sue Feldman, CEO of Synthexsis (TAW San Francisco, 2013)

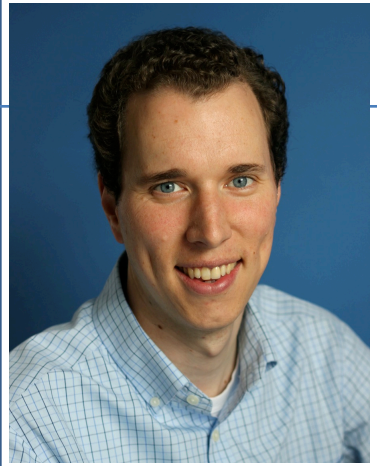
- Not just for Silicon Valley startups anymore

- Anyone can collect and acquire data to create "mash-ups" and increase value

Summary

Data Grows

- Your storage mechanism is the limiter on analytic processing
  - Evolve the storage to evolve the analytics
- Variety is the big V that drives Value. Seek it out!



Andrew Fast
Chief Scientist, Elder Research, Inc.

Dr. Andrew Fast leads research in Text Mining and Social Network Analysis at Elder Research, the nation's leading data mining consultancy. ERI was founded in 1995 and has offices in Charlottesville VA and Washington DC,(*WWW.DATAMININGLAB.COM*). ERI focuses on Federal, commercial, investment, and security applications of advanced analytics, including stock selection, image recognition, biometrics, process optimization, cross-selling, drug efficacy, credit scoring, risk management, and fraud detection.

Dr. Fast graduated Magna Cum Laude from Bethel University and earned Master's and Ph.D. degrees in Computer Science from the University of Massachusetts Amherst. There, his research focused on causal data mining and mining complex relational data such as social networks. At ERI, Andrew leads the development of new tools and algorithms for data and text mining for applications of capabilities assessment, fraud detection, and national security.

Dr. Fast has published on an array of applications including detecting securities fraud using the social network among brokers, and understanding the structure of criminal and violent groups. Other publications cover modeling peer-to-peer music file sharing networks, understanding how collective classification works, and predicting playoff success of NFL head coaches (work featured on ESPN.com). With John Elder and other co-authors, Andrew has written a book on Practical Text Mining, that was awarded the prose Award for Computing and Information Science in 2012.

About Elder Research

Elder Research Highlights

- **Experience**

  Most experienced consultancy in Data Science and Predictive Analytics (Founded in 1995)

- **Industry Leaders**

  Deep experience in predictive analytics, anomaly detection, and workload prioritization

- **Thought Leaders**

  We bridge the gap between academia and industry to bring advanced algorithms and knowledge discovery to clients across industry lines

- **Solution Neutral**

  Experts in applying COTS, open-source, and custom data mining products

- **Software Development**

  Award-winning capability in robust software application development, including commercial deployment of scientific and engineering solutions

Areas of Expertise

- **Data Mining and Predictive Analytics**

  Discovering patterns in past data that can be used to predict the outcome of future events including statistical modeling, classification & analysis, clustering, optimization & simulation, and customer segmentation

- **Text Mining**

  Understanding information stored in text documents and databases including document classification, natural language processing, information extraction and search

- **Scientific Software Engineering**

  Aiding clients in data mining, engineering and physical sciences by transforming laboratory-grade software innovations into world-class commercial-grade applications

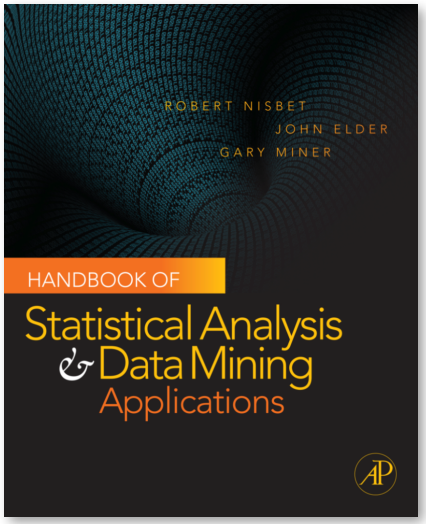
- **Data Visualization**

  Making advanced algorithms easily accessible through 2-D & 3-D, statistical and spatial visualization

- **Education & Training**

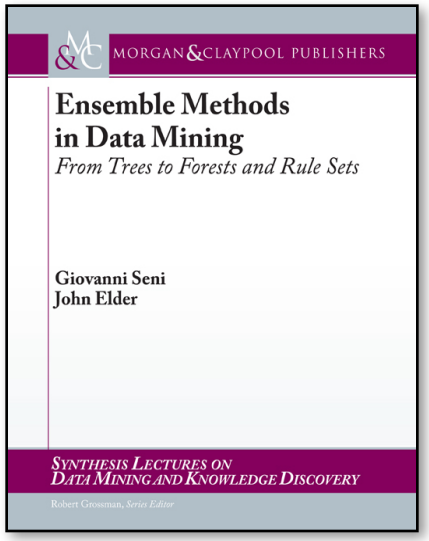
  Data mining workshops and courses (100+); keynote, plenary, and conference talks; university courses

Books Written By Elder Research

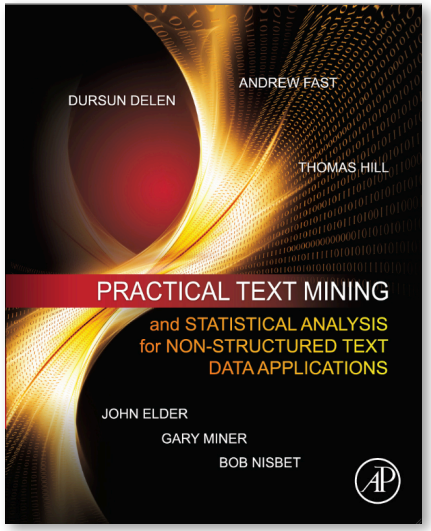


Book written *for* practitioners *by* practitioners (May 2009)

2009 Prose Award Winner

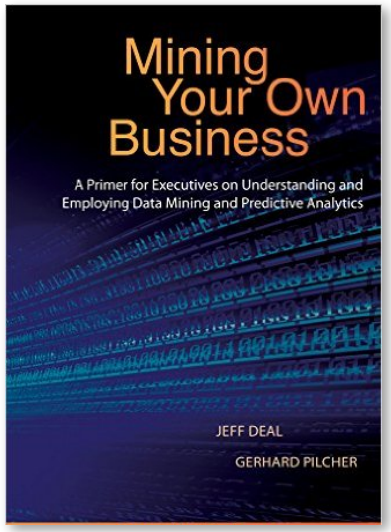


How to combine models for improved predictions (Feb 2010)



Introduction to Text Mining for practitioners (Jan 2012)

2012 Prose Award Winner



Guide for executives preparing to lead analytics initiatives (Sept 2016)

Training Services

Analytics Concepts Course

Describes leading algorithms, compares their merits, and briefly demonstrates their relative effectiveness in practical applications.

Analytics Practitioners Course

How to discover useful patterns and trends within data. Valuable practical advice on how to build reliable predictive models and interpret results with confidence.

Tools Training Course

Customized engagement based on the client's requirements. In-depth learning about common analytic software packages.

Decision Analytics Professional Course

Learn complex data access and manipulation techniques, add new tools to data crunching arsenal, and develop presentation skills. Culminates with a capstone project from the participants own department.