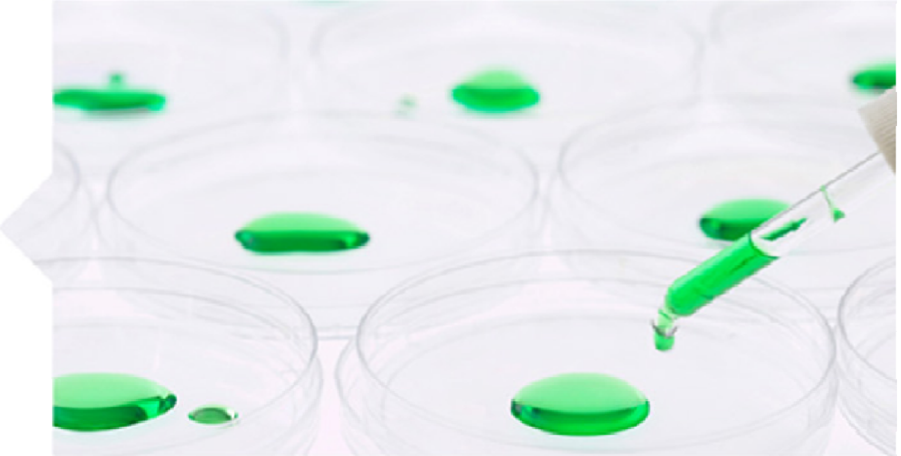
QUINTILES®

Navigating the new health

Clinical Cloud Computing, Clinical Analytics & Interchangeable Parts

Collaborative Uniform Clinical Research Across a National Network of University, Medical School, Hospital, Physician & Patient Sites



John Murphy

Vice President Clinical Informatics Innovation

PACeR Founder & Co-Director

clinical | commercial | consulting | capital

Clinical Cloud Computing, Clinical Analytics & Interchangeable Parts

Agenda

1. Background: The Partnership to Advance Clinical electronic Research (PACeR)

   Five Years of Collaborative Clinical Research

2. The Challenge of Big Data in Biology & Medicine: *Uniform Policies, Procedures, Methods, Technologies are Prerequisites for Success*
3. Automation & Interchangeable Parts: Object Libraries& Uniformity
4. PACeR Solutions in Action

   The LifeSpan Clinical Development Toolkit

   PACeR-MyHealth Locally Branded Social Networks

   PACeR- ResearchPro Collaborative Science Cloud

   LifeSpan Analytics

Appendix: Big Data Basics

Like the Weather, Everyone Talks About Big Data, but No One Knows How to Manage It

QUINTILES®

Perhaps That's Because the Data Sets That Scientists, Clinicians, Hospitals, Pharmaceutical & Medical Device Companies, CRO's….. Create for Their Own Research Lacks Uniformity and Standardization.

DILBERT
BY SCOTT ADAMS
CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
DilbertCartoonist@gmail.com
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
© 2012 Scott Adams, Inc./Dist. by Universal Uclick
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.
BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.
LET US PAY.
$
www.dilbert.com
7-29-12
IS IT TOO LATE TO SIDE WITH EVIL?
SHHHH! IT HEARS YOU.

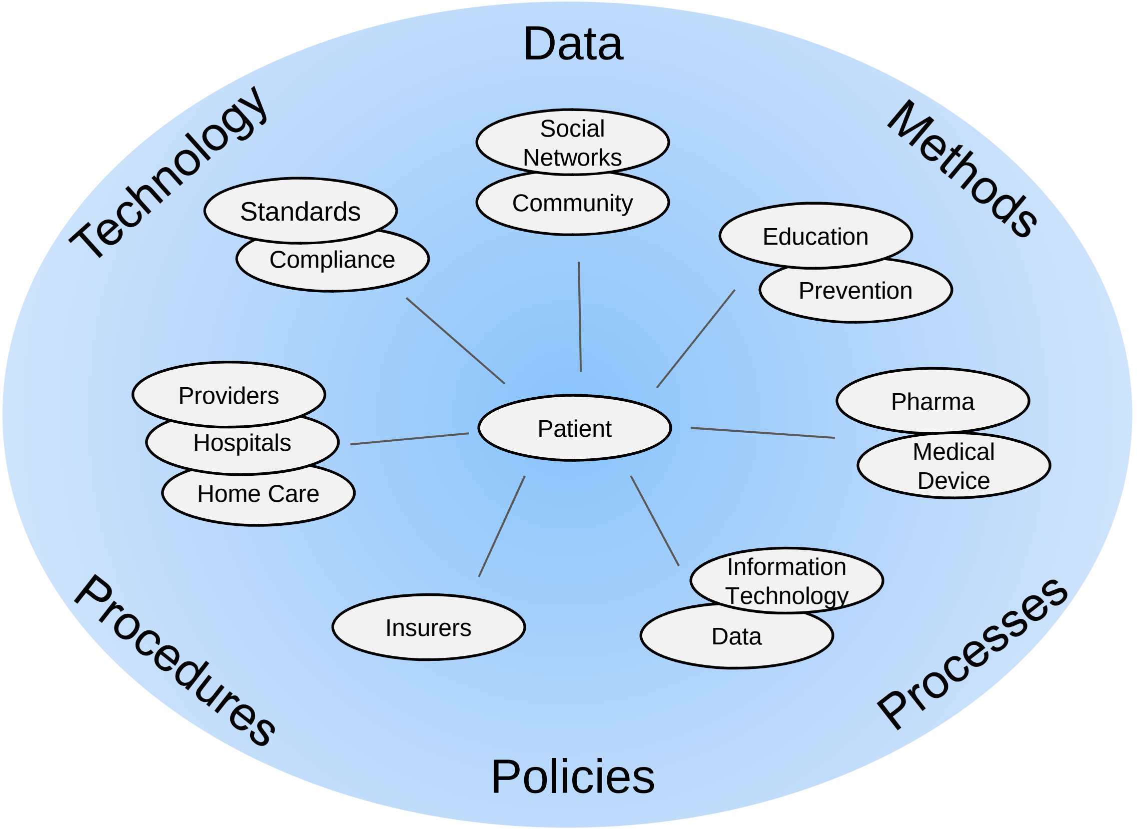
PACeR Overview

PACeR is a 501c3 not-for-profit corporation created 5 years ago by a group of pharmaceutical companies & Academic Medical Centers to create a standardized national clinical research network designed to solve problems of non-uniform policy, procedure, method, technology and data related to clinical research

PACeR captures new forms of data in a standardized way via new business models and technologies to provide solutions for clinical research.

PACeR for Free. Only commitment is to support collaborative research & utilize the network.

PACeR Clinical Research Network



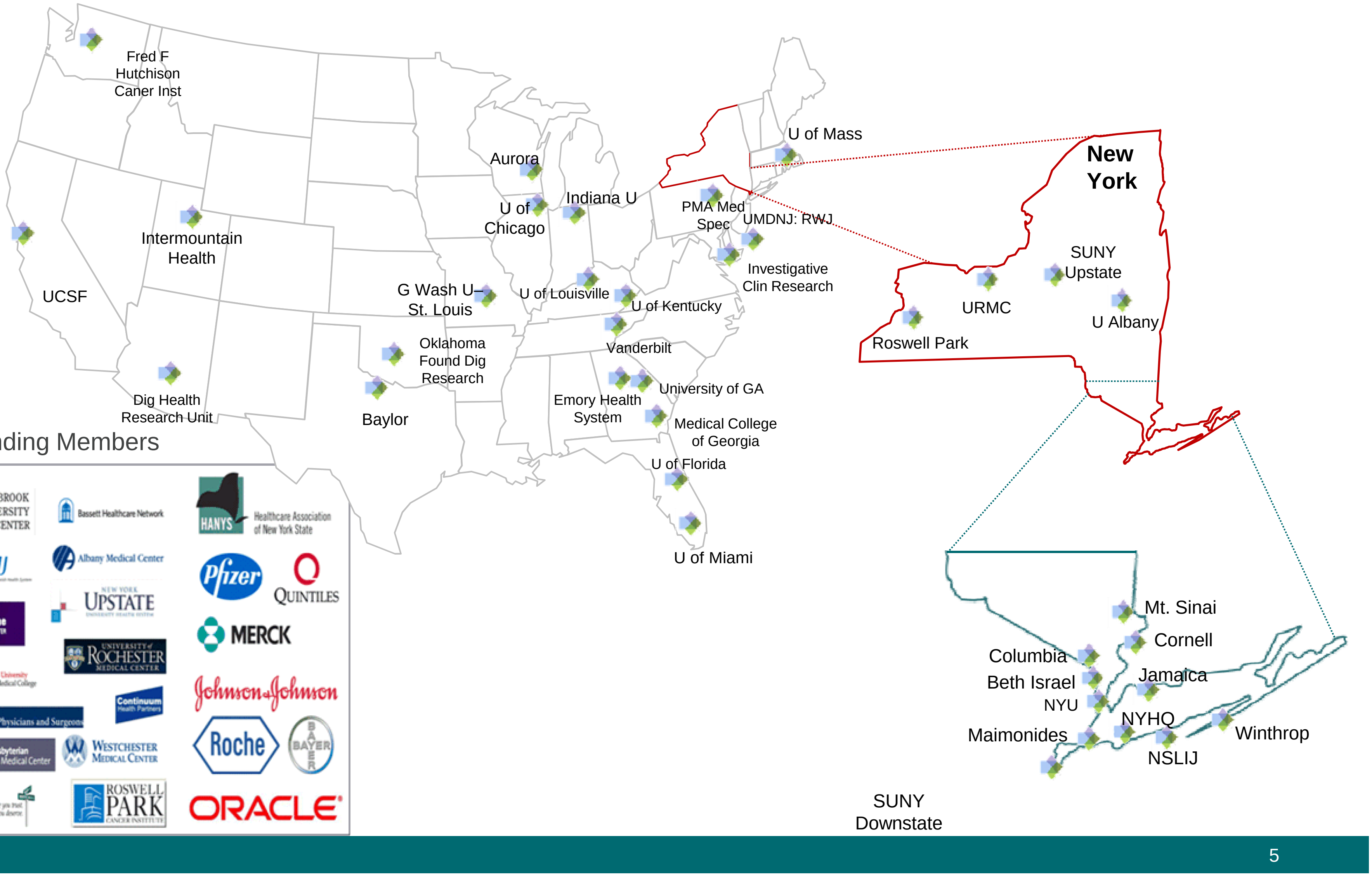
Clinical Research Solutions

- Greater understanding of patient make-up
- Data standardization
- Cross-industry standardization of business processes
- Access to patient populations
- Decreased screen failure rates
- Causal associative research
- Patient engagement through educational programs
- Collaborative research via social networks and nationwide site networks
- EMR data integration. SDV & Real-time Risk Based Monitoring

PACeR’s expanding site network

QUINTILES

Started in New York, the PACeR network has grown across the US. New sites are constantly added as member projects expand. The currently installed technical infrastructure is capable of simultaneously supporting more than 5,000 sites, 200,000 clinical users, and millions of social network users.

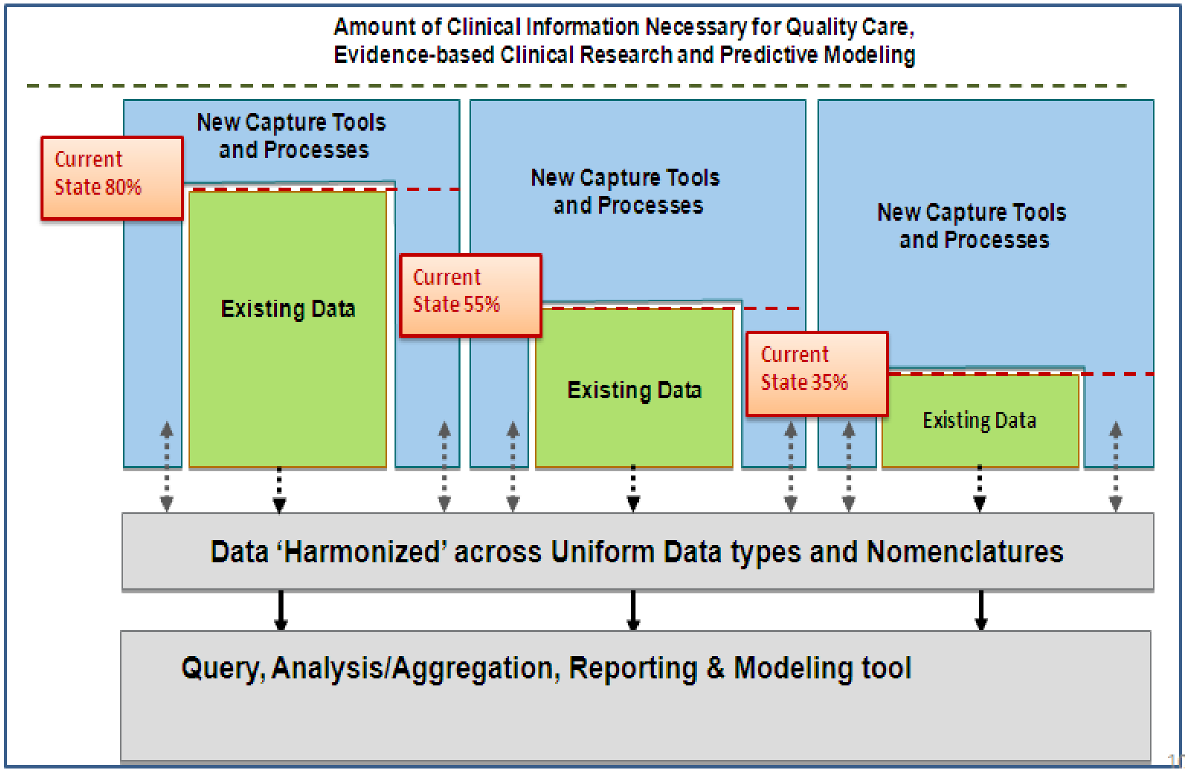


Founding Members

- STONY BROOK UNIVERSITY MEDICAL CENTER
- Bassett Healthcare Network
- HANYS Healthcare Association of New York State
- North Shore LIJ
- Albany Medical Center
- Pfizer
- QUINTILES
- NEW YORK UPSTATE
- NYU Langone MEDICAL CENTER
- MERCK
- UNIVERSITY of ROCHESTER MEDICAL CENTER
- Cornell University Weill Medical College
- Johnson & Johnson
- Continuum Health Partners
- College of Physicians and Surgeons
- Roche
- BAYER
- NewYork-Presbyterian Weill Cornell Medical Center
- WESTCHESTER MEDICAL CENTER
- Expertise you trust Service you deserve
- ROSWELL PARK CANCER INSTITUTE
- ORACLE

PACeR Research Identified Defects in Existing Clinical Databases

A detailed analysis of installed Electronic Medical Record systems across a large number of health systems revealed deficiencies of existing data within individual health systems. Not only was there a lack of data capture capabilities, there was no intra-institution and inter-institutional management of data standards. With the exception of some basic coding for billing purposes that can be used for Observational Research, there is almost no capability to perform Scientific-method's based Research and Forget About "Big Data Analytics"!



Amount of Clinical Information Necessary for Quality Care, Evidence-based Clinical Research and Predictive Modeling

New Capture Tools and Processes

Current State 80%

Existing Data

New Capture Tools and Processes

Current State 55%

Existing Data

New Capture Tools and Processes

Current State 35%

Existing Data

Data 'Harmonized' across Uniform Data types and Nomenclatures

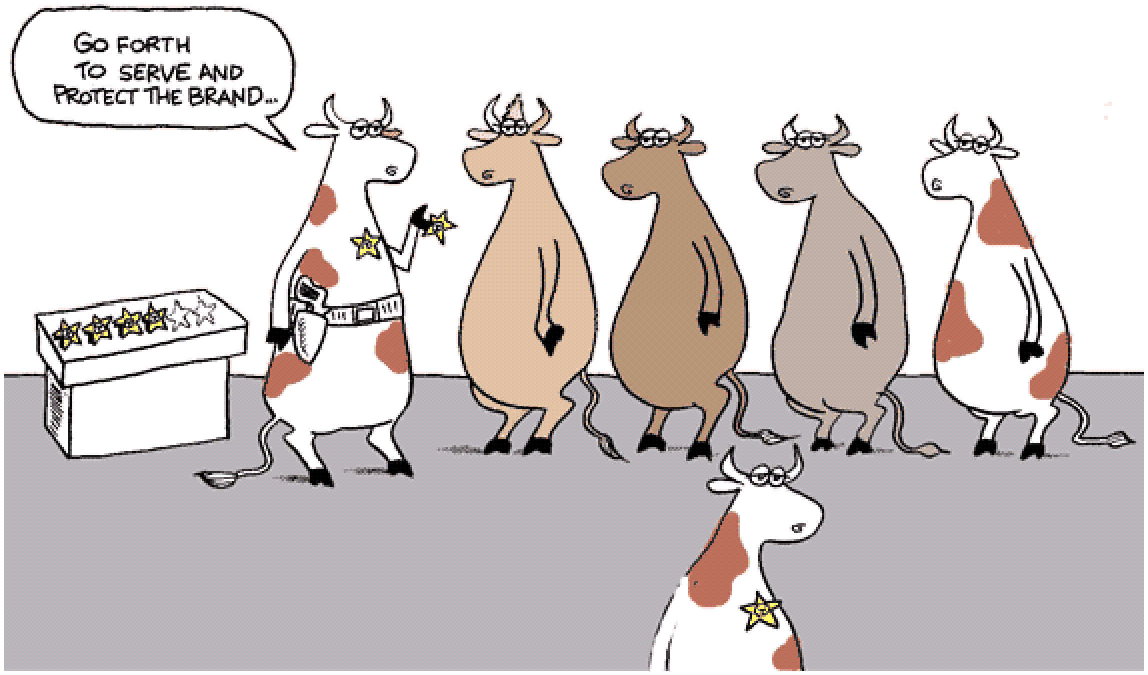
Query, Analysis/Aggregation, Reporting & Modeling tool

Institutional EMR Analysis

- Epic, Eclipsys Allscripts, GE, McKesson, Siemens… all fail to collect UNIFORM data for Scientific Methods based research. At best, provide Observational analytic capabilities.
- Much of the information that would be of use is in unstructured text and dictated notes.
- Additionally, a significant amount of data is not captured at all.
- Five Epic implementations will have five different deployments. Data is different across all five. This is true of all existing EMR vendor systems.
- A set of data related to ICD9, CPT and lab and medications is reasonably standardized as a requirement of third party billing.
- Data is not in a form that allows longitudinal tracking or analysis.

PACeR Research Identified Non-competitive Inefficiencies That Distort Efficiency, Accuracy, Quality & Cost of Clinical Research

Policies, Procedures, Contracts, Semantics, Nomenclatures, Data, Clinical Methods… can be standardized without interfering with competitive development of products, scientific discoveries, publications….. Working Collaboratively on Non-competitive aspects of scientific research saves time and money and improves everyone's research.



PACeR Identified Business Impediments the Most Striking Being Data Ownership

QUINTILES®

Data Ownership presents a barrier to collaborative science and Big Data Science

PACeR developed a Data Franchise Business Model that recognizes local ownership of basic & clinical research data, but allows each local owner to monetize their data, PACeR developed an on-line for-profit marketplace for the buying & selling of data

Some Major Components

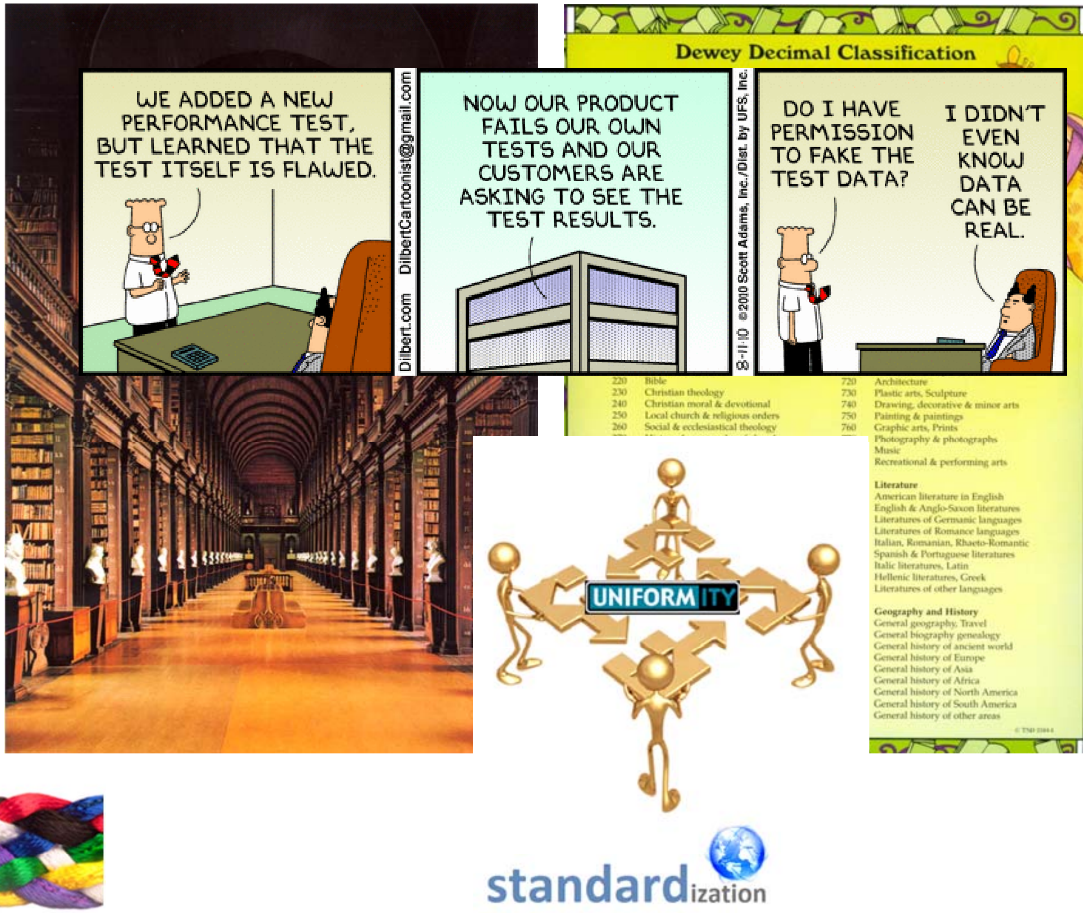
- Hospital Network-Physician-Patient social network
  - > Branded and managed by local hospital-physician network for PACeR & Institutional communication and longitudinal research—PACeRMyHealth.com
    - Study protocol deidentified survey system linked to community & national associations for real-time recruiting, educating, consenting, enrollment and management of patients
    - On-line groups and forums communication
    - Study & Clinical Content Education, Marketing, Expert Consultation…..
    - Crowd sourced survey system for study design & population interrogation
    - MyHealth Home Monitoring for longitudinal clinical outcomes research….
- Pharmaceutical & Device to Clinical Research System
  - > Components:
    - LifeSpan Developer: Decision supported eClinical Application Builder using Uniform Semantics, Nomenclatures, Data Elements & Clinical Methods Across ALL Sites & Users
    - LifeSpan Clinical Dashboard
      - » Clinical Study Manager
      - » Differential Diagnostics—A.I. Bayesian & Neural Modeling & Adaptive Learning System
      - » Uniform Cloud-based Deployment to Hospitals, Physicians, P.I.'s & Patients
    - PACeRResearchPro.com private research Network
      - » Educate, Communicate, Interrogate, Survey & Manage Clinical Research
    - PACeR participant Registries
      - » On-line workflow management

I CAN'T HELP YOU BECAUSE I'M BUSY WORKING ON A SOCIAL NETWORK STRATEGY FOR COLLABORATIVE DRUG RESEARCH

PACeR Components Operate Across the Cloud Employing Standard Interchangeable Parts

Uniform Semantics, Nomenclatures, Data Standards & Libraries of Standardized Cataloged Reusable Clinical & Research Methods are Used Across All Companies, Hospitals, P.I. & Physician Sites, Associations… and

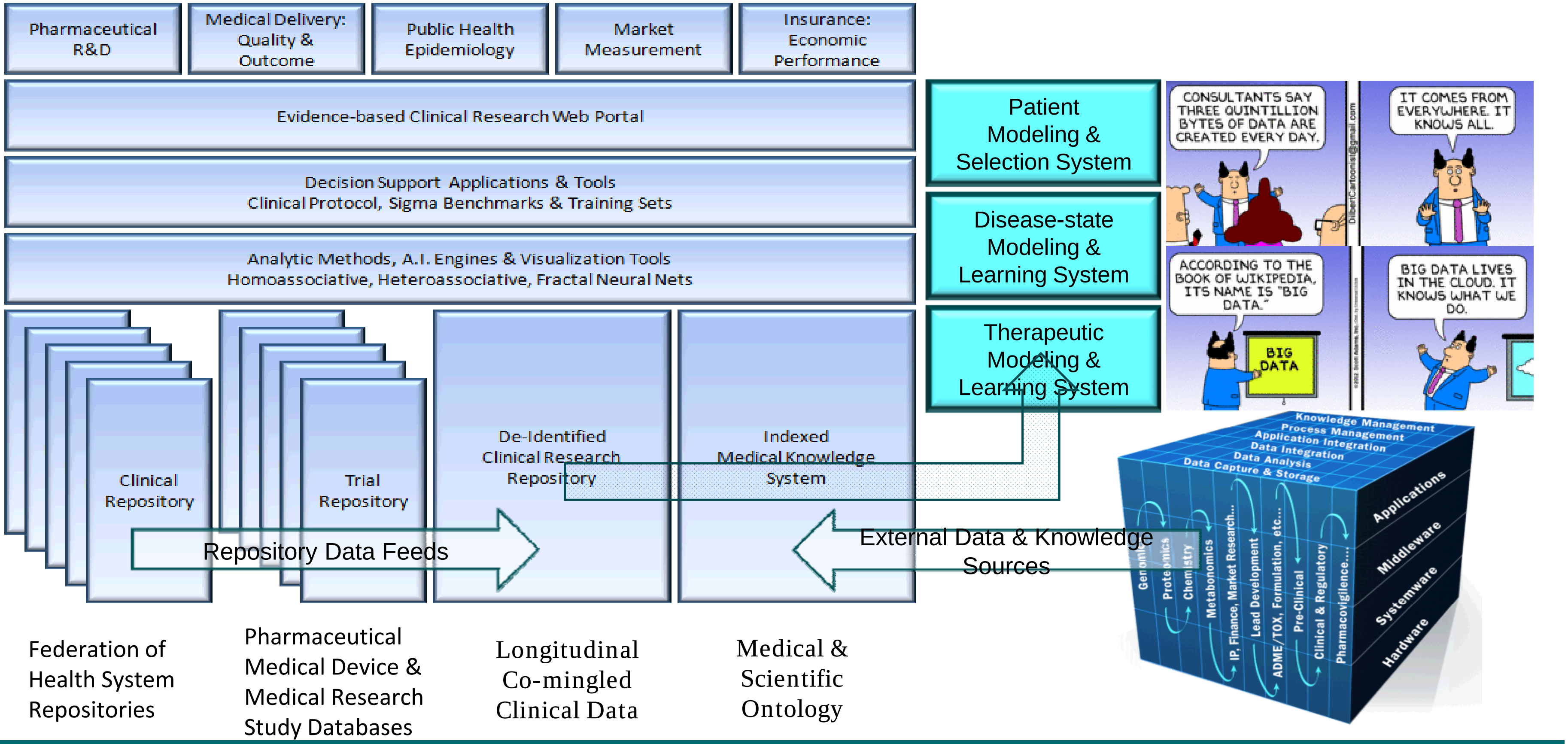
Patient Homes


Dewey Decimal Classification
WE ADDED A NEW PERFORMANCE TEST, BUT LEARNED THAT THE TEST ITSELF IS FLAWED.
Dilbert.com DilbertCartoonist@gmail.com
NOW OUR PRODUCT FAILS OUR OWN TESTS AND OUR CUSTOMERS ARE ASKING TO SEE THE TEST RESULTS.
8-11-10 ©2010 Scott Adams, Inc./Dist. by UFS, Inc.
DO I HAVE PERMISSION TO FAKE THE TEST DATA?
I DIDN'T EVEN KNOW DATA CAN BE REAL.
UNIFORMITY
standardization

Uniform Policies, Procedures, Methods, Technologies, Data & Franchised Data Ownership Business Model Creates Big Data Environment

QUINTILES®

PACeR's Federated Big Data Collaboration has enabled the development of a uniform Knowledge Cube & Advanced Analytics for Basic, Clinical, Safety, Quality, Economic & Outcome Research Across its Member Organizations



Pharmaceutical R&D

Medical Delivery: Quality & Outcome

Public Health Epidemiology

Market Measurement

Insurance: Economic Performance

Evidence-based Clinical Research Web Portal

Decision Support Applications & Tools
Clinical Protocol, Sigma Benchmarks & Training Sets

Analytic Methods, A.I. Engines & Visualization Tools
Homoassociative, Heteroassociative, Fractal Neural Nets

Patient Modeling & Selection System

Disease-state Modeling & Learning System

Therapeutic Modeling & Learning System

Clinical Repository

Trial Repository

De-Identified Clinical Research Repository

Indexed Medical Knowledge System

Repository Data Feeds

External Data & Knowledge Sources

Federation of Health System Repositories

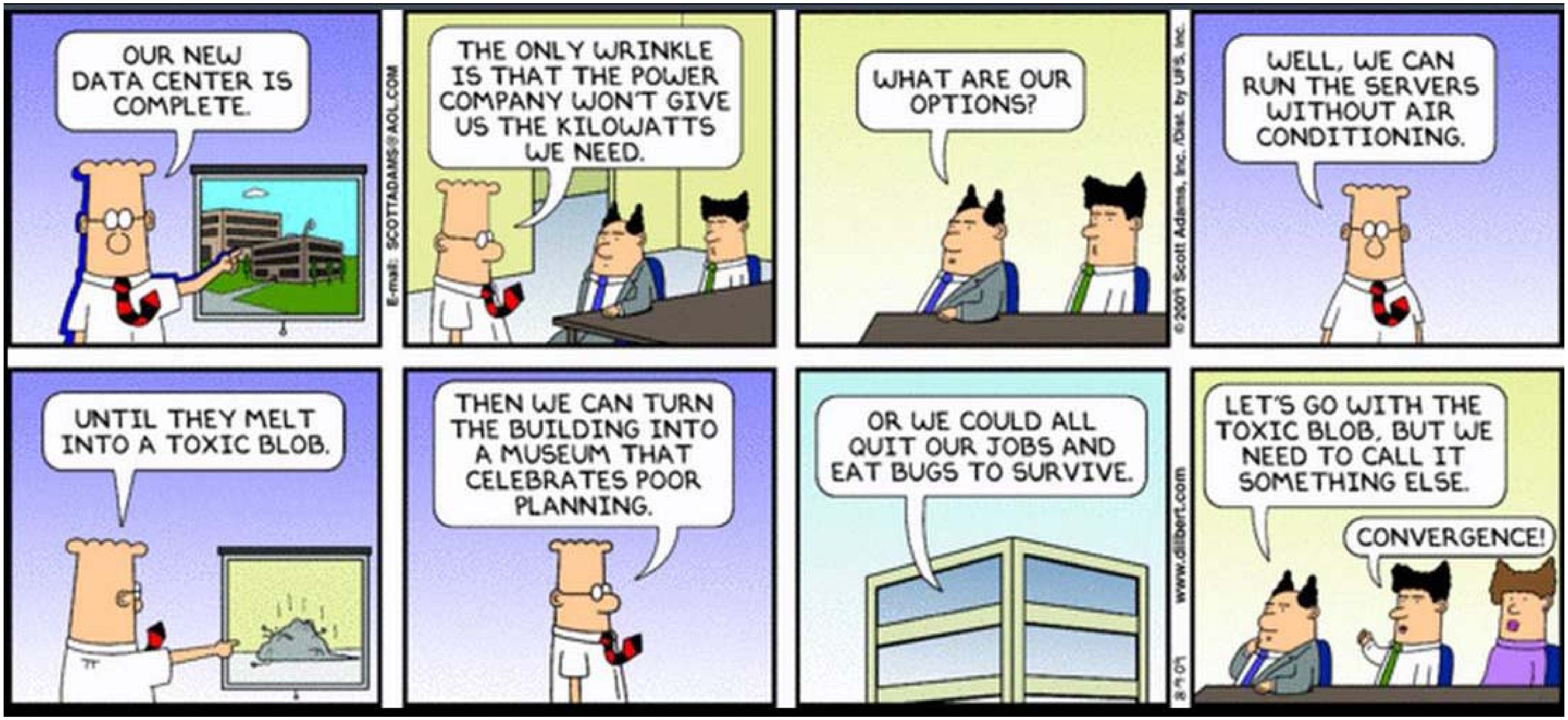
Pharmaceutical Medical Device & Medical Research Study Databases

Longitudinal Co-mingled Clinical Data

Medical & Scientific Ontology

CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
DilbertCartoonist@gmail.com
Knowledge Management
Process Management
Application Integration
Data Integration
Data Analysis
Data Capture & Storage
Genomics
Proteomics
Chemistry
Metabonomics
IP, Finance, Market Research...
Lead Development
ADME/TOX, Formulation, etc...
Pre-Clinical
Clinical & Regulatory
Pharmacovigilence....
Applications
Middleware
Systemware
Hardware

In Honor of Big Data & the NSA One Last Thought


OUR NEW DATA CENTER IS COMPLETE.
E-mail: SCOTTADAMS@AOL.COM
THE ONLY WRINKLE IS THAT THE POWER COMPANY WON'T GIVE US THE KILOWATTS WE NEED.
WHAT ARE OUR OPTIONS?
© 2001 Scott Adams, Inc./Dist. by UFS, Inc.
WELL, WE CAN RUN THE SERVERS WITHOUT AIR CONDITIONING.
UNTIL THEY MELT INTO A TOXIC BLOB.
THEN WE CAN TURN THE BUILDING INTO A MUSEUM THAT CELEBRATES POOR PLANNING.
OR WE COULD ALL QUIT OUR JOBS AND EAT BUGS TO SURVIVE.
www.dilbert.com
8-9-01
LET'S GO WITH THE TOXIC BLOB, BUT WE NEED TO CALL IT SOMETHING ELSE.
CONVERGENCE!

Appendix

Some Big Data Supporting Background

Let’s Define What We Mean by Big Data

Some Basic Facts

- **Big data**[1][2] is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage,[3] search, sharing, transfer, analysis,[4] and visualization.
- As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data.[8] Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics,[9] connectomics, complex physics simulations,[10] and biological and environmental research.[11]
- The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;[14] as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data were created.[15]
- Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources).
- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;
- Big data uses inductive statistics and concepts from nonlinear system identification [25] to infer laws (regressions, nonlinear relationships, and causal effects) from large data sets [26] to reveal relationships, dependencies, and to perform predictions of outcomes and behaviors. [25] [27]

Big Science Generates Big Data

But Not All Big Data is Equally Complex

- **Big science**
- The Large Hadron Collider experiments represent about 150 million sensors delivering data 40 million times per second. There are nearly 600 million collisions per second. After filtering and refraining from recording more than 99.999% of these streams, there are 100 collisions of interest per second.[30][31][32]
- As a result, only working with less than 0.001% of the sensor stream data, the data flow from all four LHC experiments represents 25 petabytes annual rate before replication (as of 2012). This becomes nearly 200 petabytes after replication.
- If all sensor data were to be recorded in LHC, the data flow would be extremely hard to work with. The data flow would exceed 150 million petabytes annual rate, or nearly 500 exabytes per day, before replication. To put the number in perspective, this is equivalent to 500 quintillion (5×10^{20}) bytes per day, almost 200 times higher than all the other sources combined in the world.
- **Science and research**
- When the Sloan Digital Sky Survey (SDSS) began collecting astronomical data in 2000, it amassed more in its first few weeks than all data collected in the history of astronomy. Continuing at a rate of about 200 GB per night, SDSS has amassed more than 140 terabytes of information. When the Large Synoptic Survey Telescope, successor to SDSS, comes online in 2016 it is anticipated to acquire that amount of data every five days.[5]
- Decoding the human genome originally took 10 years to process, now it can be achieved in less than a week : the DNA sequencers have divided the sequencing cost by 10,000 in the last ten years, which is 100 times faster than the reduction in cost predicted by Moore's Law.[33]
- Computational social science — Tobias Preis *et al.* used Google Trends data to demonstrate that Internet users from countries with a higher per capita gross domestic product (GDP) are more likely to search for information about the future than information about the past. The findings suggest there may be a link between online behaviour and real-world economic indicators.[34][35][36] The authors of the study examined Google queries logs made by Internet users in 45 different countries in 2010 and calculated the ratio of the volume of searches for the coming year (‘2011’) to the volume of searches for the previous year (‘2009’), which they call the ‘future orientation index’.[37] They compared the future orientation index to the per capita GDP of each country and found a strong tendency for countries in which Google users enquire more about the future to exhibit a higher GDP. The results hint that there may potentially be a relationship between the economic success of a country and the information-seeking behavior of its citizens captured in big data.
- The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.[38]
- Tobias Preis and his colleagues Helen Susannah Moat and H. Eugene Stanley introduced a method to identify online precursors for stock market moves, using trading strategies based on search volume data provided by Google Trends.[39] Their analysis of Google search volume for 98 terms of varying financial relevance, published in *Scientific Reports*,[40] suggests that increases in search volume for financially relevant search terms tend to precede large losses in financial markets.[41][42][43][44][45][46][47][48]

Big Data Outside Healthcare

By comparison, Biology & Medicine Data Sets Might Require Brontobyte Manipulation

- **Government**
- In 2012, the Obama administration announced the Big Data Research and Development Initiative, which explored how big data could be used to address important problems faced by the government.[49] The initiative was composed of 84 different big data programs spread across six departments.[50]
- The United States Federal Government owns six of the ten most powerful supercomputers in the world.[52]
- The Utah Data Center is a data center currently being constructed by the United States National Security Agency. When finished, the facility will be able to handle yottabytes of information collected by the NSA over the Internet.[53][54]
- **Private sector**
- eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Inside eBay's 90PB data warehouse
- Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.[55]
- Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress.[5]
- Facebook handles 50 billion photos from its user base.[56]
- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide.[57]
- The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates.[58]
- Sylvan Road uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.[59] We combine nationwide MLS listings, bank mortgage and foreclosure databases, county, town and school district information, Google maps and street view and a wide variety of data gathered from title and appraisal databases to automate the large-scale purchase of homes—sight unseen. Petabyte scale conversion of data into information and information into actionable knowledge.

Big Data is Useless without Big Analytics

Lessons can be learned from more advanced users of Big Data

Technologies

- DARPA's Topological Data Analysis program seeks the fundamental structure of massive data sets.
- Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. A 2011 McKinsey report[66] suggests suitable technologies include A/B testing, crowdsourcing, data fusion and integration, genetic algorithms, machine learning, natural language processing, signal processing, simulation, time series analysis and visualisation. Multidimensional big data can also be represented as tensors, which can be more efficiently handled by tensor-based computation,[67] such as multilinear subspace learning.[68] Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data-mining grids, distributed file systems, distributed databases, cloud based infrastructure (applications, storage and computing resources) and the Internet.
- Some but not all MPP relational databases have the ability to store and manage petabytes of data. Implicit is the ability to load, monitor, back up, and optimize the use of the large data tables in the RDBMS.[69]
- DARPA's Topological Data Analysis program seeks the fundamental structure of massive data sets and in 2008 the technology went public with the launch of a company called Ayasdi.
- The practitioners of big data analytics processes are generally hostile to slower shared storage,[70] preferring direct-attached storage (DAS) in its various forms from solid state drive (SSD) to high capacity SATA disk buried inside parallel processing nodes. The perception of shared storage architectures—SAN and NAS—is that they are relatively slow, complex, and expensive. These qualities are not consistent with big data analytics systems that thrive on system performance, commodity infrastructure, and low cost.
- Real or near-real time information delivery is one of the defining characteristics of big data analytics. Latency is therefore avoided whenever and wherever possible. Data in memory is good—data on spinning disk at the other end of a FC SAN connection is not. The cost of a SAN at the scale needed for analytics applications is very much higher than other storage techniques.
- There are advantages as well as disadvantages to shared storage in big data analytics, but big data analytics practitioners as of 2011 did not favour it.[71]

Lessons Learned Continued

Borrow Solutions From the Pioneers

Architecture

- In 2004, Google published a paper on a process called MapReduce that used such an architecture. MapReduce framework provides a parallel processing model and associated implementation to process huge amount of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step). The framework was incredibly successful so others wanted to replicate the algorithm. Therefore, an implementation of MapReduce framework was adopted by an Apache open source project named Hadoop.[64]
- MIKE2.0 is an open approach to information management. The methodology addresses handling big data in terms of useful permutations of data sources, complexity in interrelationships, and difficulty in deleting (or modifying) individual records.[65]