SDS | 2015
2nd Swiss Workshop on Data Science
12 June 2015, Eulachpassage, Winterthur (ZH) Switzerland

The Emerging Discipline of Data Science

Principles and Techniques

For

Data-Intensive Analysis

MIT CSAIL

What is Big Data Analytics?

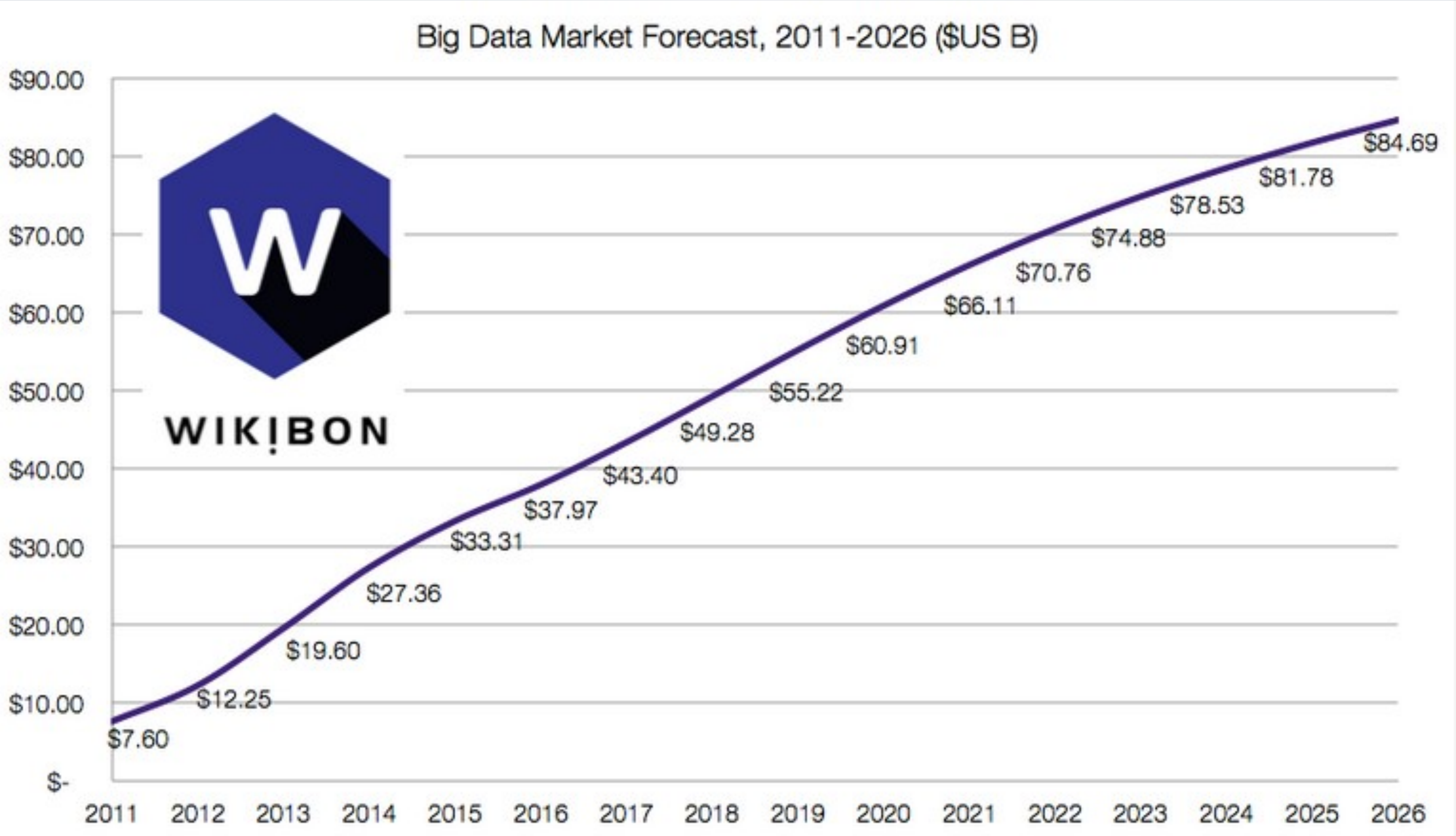
Is this a new paradigm?

What is the role of data?

What could possibly go wrong?

What is Data Science?

Big Data is Hot!



Big Data Is Important

Hot

- Market
  - Results, products, jobs
- Potential
  - 4th Paradigm
  - Accelerates discovery [urgent]
  - Better: cost, speed, specificity
  - Change 80% of processes [Gartner]
- Government Policy (45+)
  - White House; most US Govt agencies
- Adoption: Most Human Endeavors
  - All academic disciplines
  - Computational X

Cool

- Low effective adoption [EMC]
  - 60% operational
  - 20% significant change
  - < 1% effective
- Results not operational
- In its infancy ☑ lacking
  - Understanding
  - Concepts, tools, techniques (methods)
    - 21st Century Statistics
  - Theory: principles, guidelines

Healthcare Potential: Better Health; Faster, Cheaper Remedies

Lower Healthcare Costs by Utilizing Technology to Help Manage and Prevent Chronic Diseases

- In 2013, the US government spent $591 billion on Medicare. However, Medicare is projected to have insufficient funds to pay all hospital bills beginning in 2030
- Chronic disease accounts for 86% of US healthcare costs, which can be reduced by enabling the healthcare ecosystem with innovative technology

Cardiovascular
$193 Bn spent on direct medical expenditures
WATCH · fitbit · AliveCor

Diabetes
$176 Bn spent on direct medical expenditures
livongo health · omada

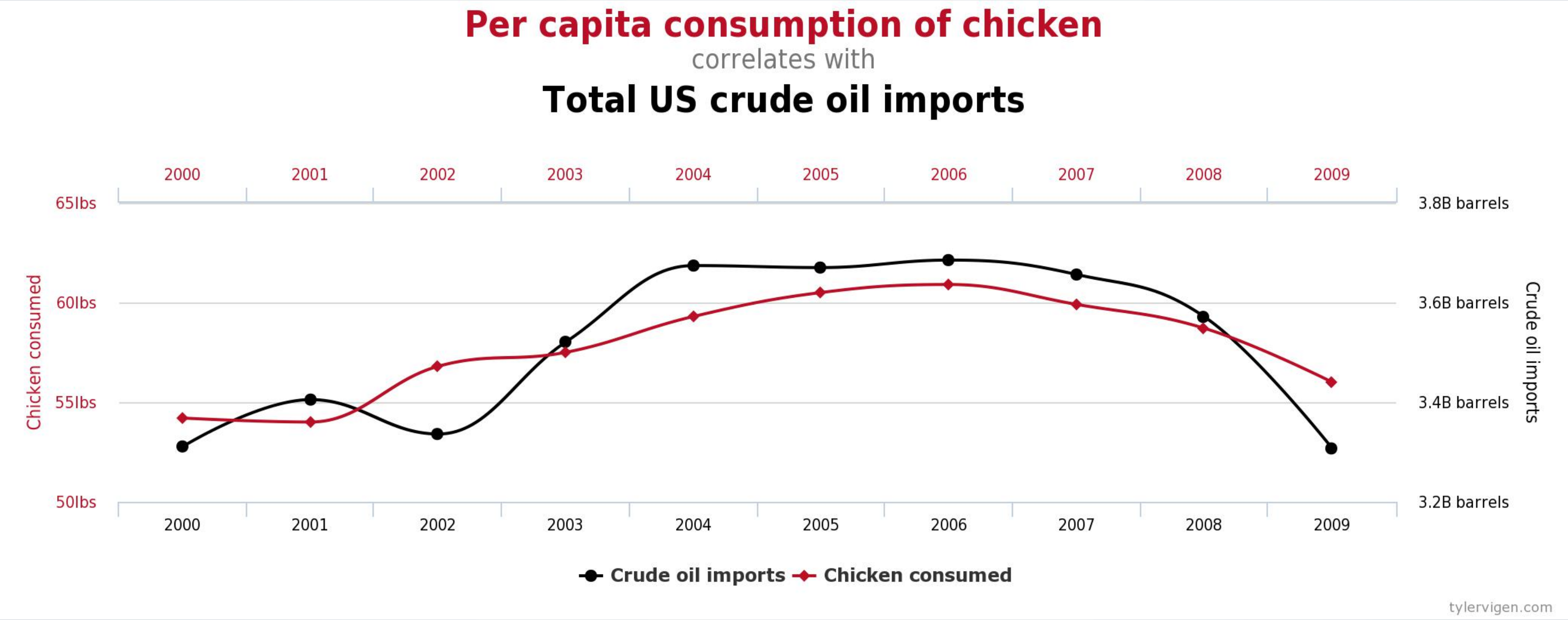
Obesity
$147 Bn linked to medical expenditures
myfitnesspal · REDBRICK HEALTH

Cancer
$157 Bn spent on direct medical expenditures
FLATIRON · FOUNDATION MEDICINE

BIG DATA AND ANALYTIC PLATFORMS
Google · ZEPHYR HEALTH · IBM · Lumeris · practice fusion

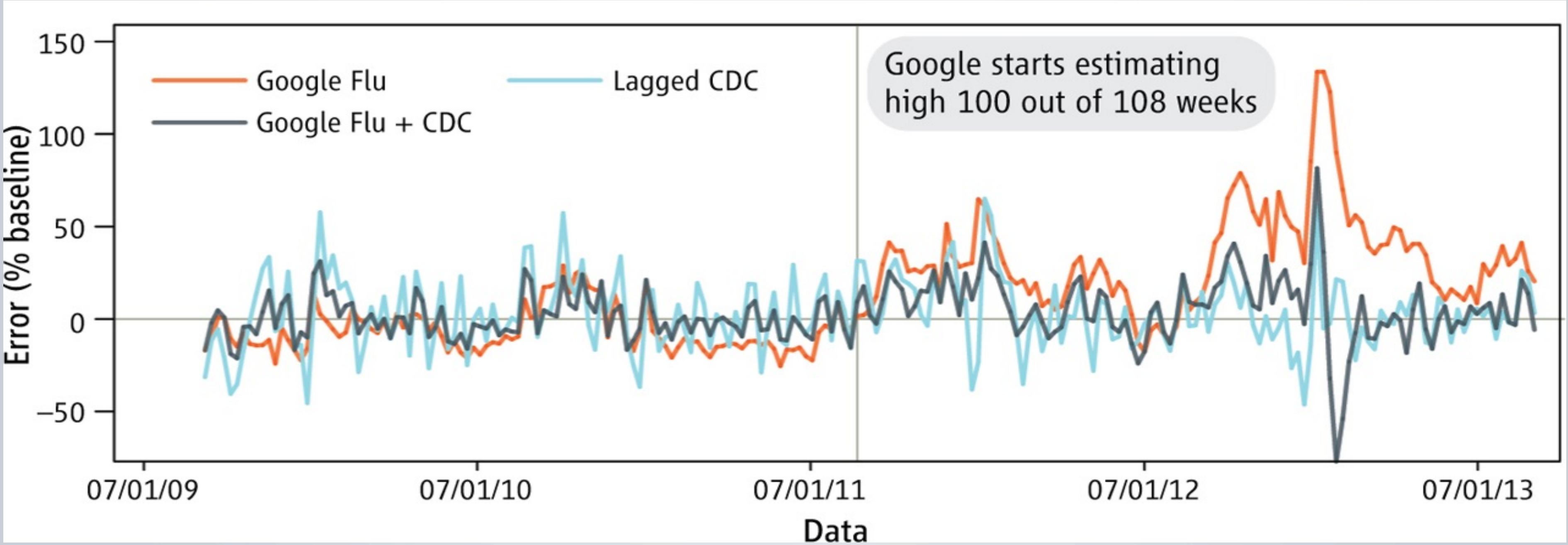
Source: Beth Seidenberg, KPCB General Partner and Lynne Chou, KPCB Partner. Sources: Kaiser Family Foundation website and CDC website http://www.cdc.gov/chronicdisease/overview/

What could go Wrong?
When are Correlations Spurious?

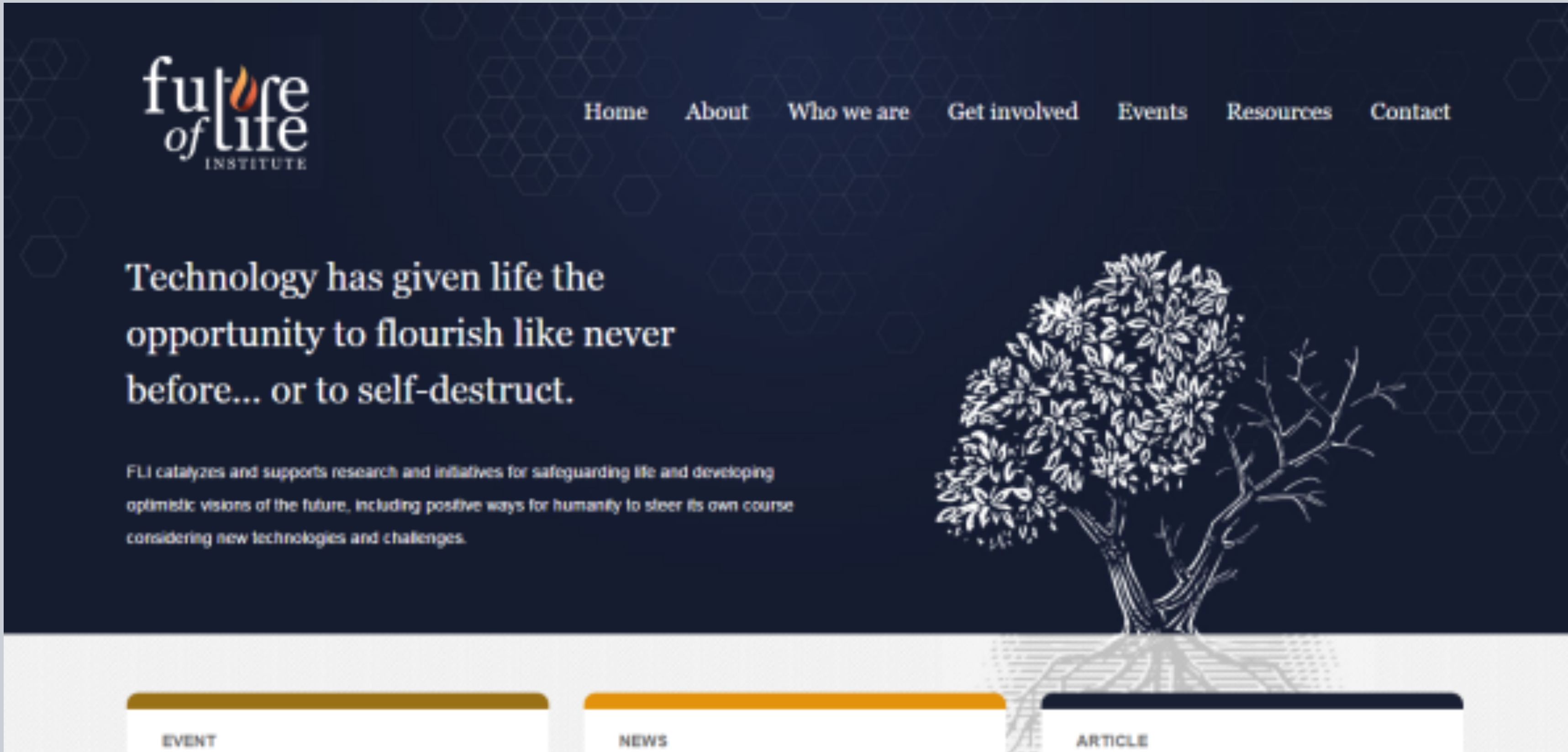


Or Just Wrong? E.g. Google Flu Trends

Allegedly Real-time, Reliable Predictions
High 100 out of 108 weeks



Future of Life: Institute to



"mitigate existential risks facing humanity"

CSAIL

US Legal Community Pursuing Algorithmic Accountability

Do We Know / Can We Prove?

- DIA Result: *correct*, *complete*, *efficient*?
- What machines / algorithms / Machine Learning / Black Boxes / DIA do?

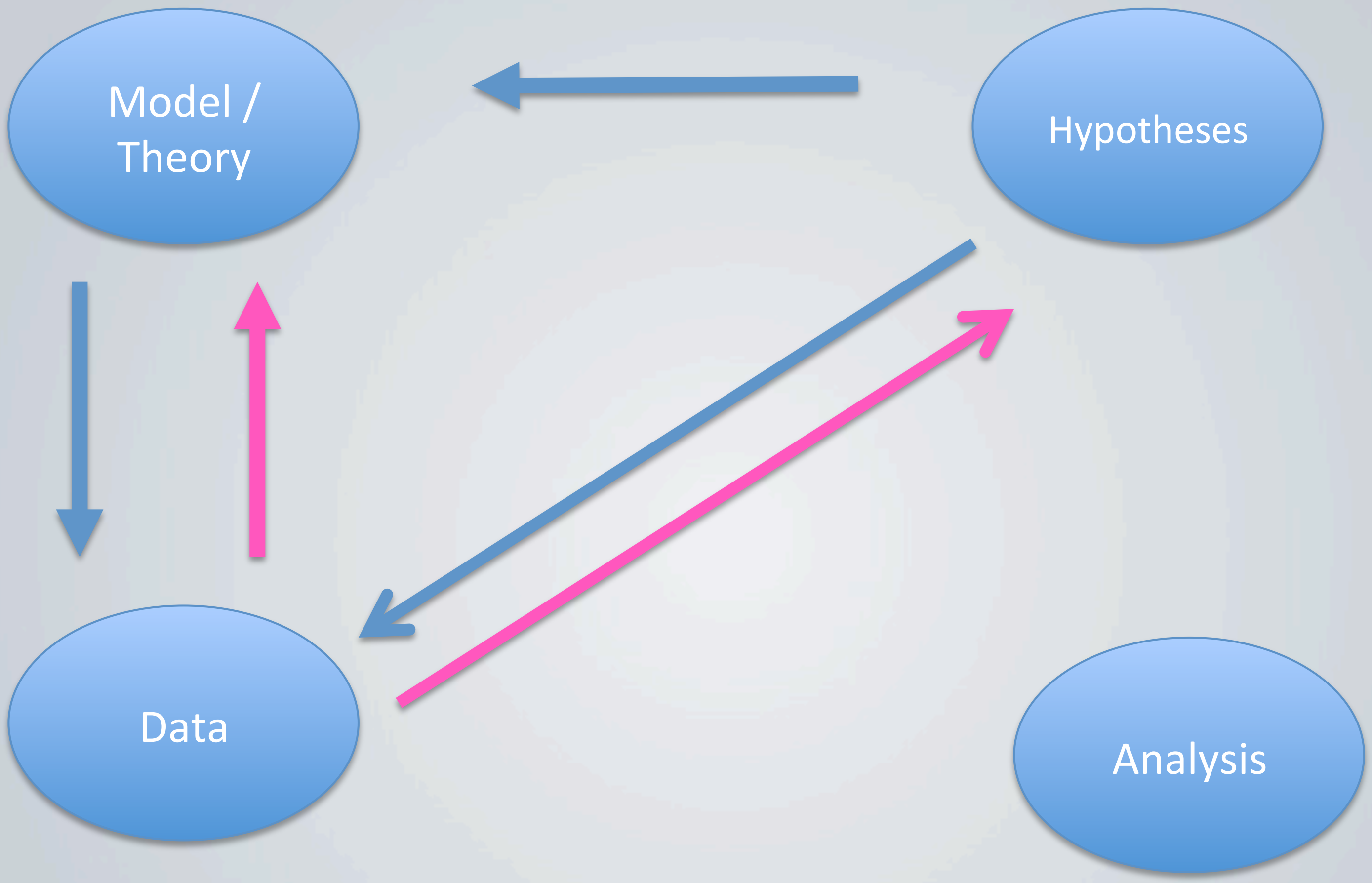
- Emergent Data-Driven Society with High
  - Reward: Cancer cures, drug discovery, personalized medicine, …
  - Risk: errors in any of the above

The search for

truth

evidence-based causality

evidence-based correlations


Model / Theory
Hypotheses
Data
Analysis

Long Illustrious Histories

Data Analysis

- Mathematics
  - Babylon (17th-12th C BCE)
  - India (12th C BCE)
- Mathematical analysis (17th C, Scientific Revolution)
- Statistics (5th C BCE, 18th C)

~4,000 years

Scientific Method

- Empiricism
  - Aristotle (384-322 BCE)
  - Ptolemy (1st C)
  - Bacons (13th, 16th C)

~2,000 years

- Scientific Discovery Paradigms
  1. Theory
  2. Experimentation
  3. Simulation
  4. eScience / Big Data

~ 1,000 years

Fourth Paradigm

Modern Computing

- Hardware: 40s-50s
- FORTRAN: 50s
- Spreadsheets: 70s
- Databases: 70s-80s
- World Wide Web: 90s

~ 60 years

Data-Intensive Analysis of Everything

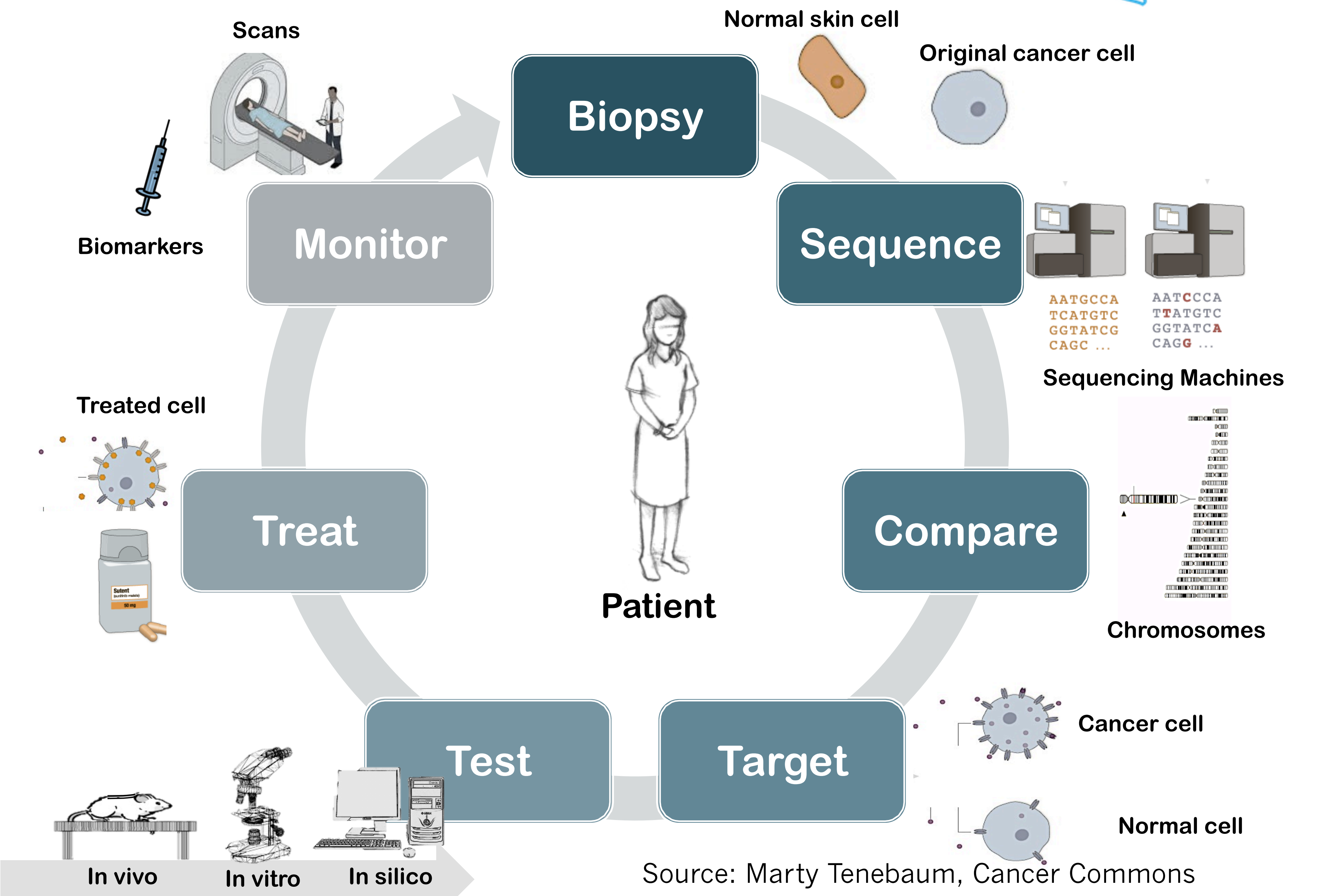
- eScience (~2000)
- Big Data (~2007)
  - Particle physics, drug discovery, …

~ 15 years

Paradigms

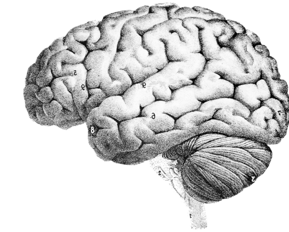
- Long developments
- Significant shifts
  - Conceptual
  - Theoretical
  - Procedural

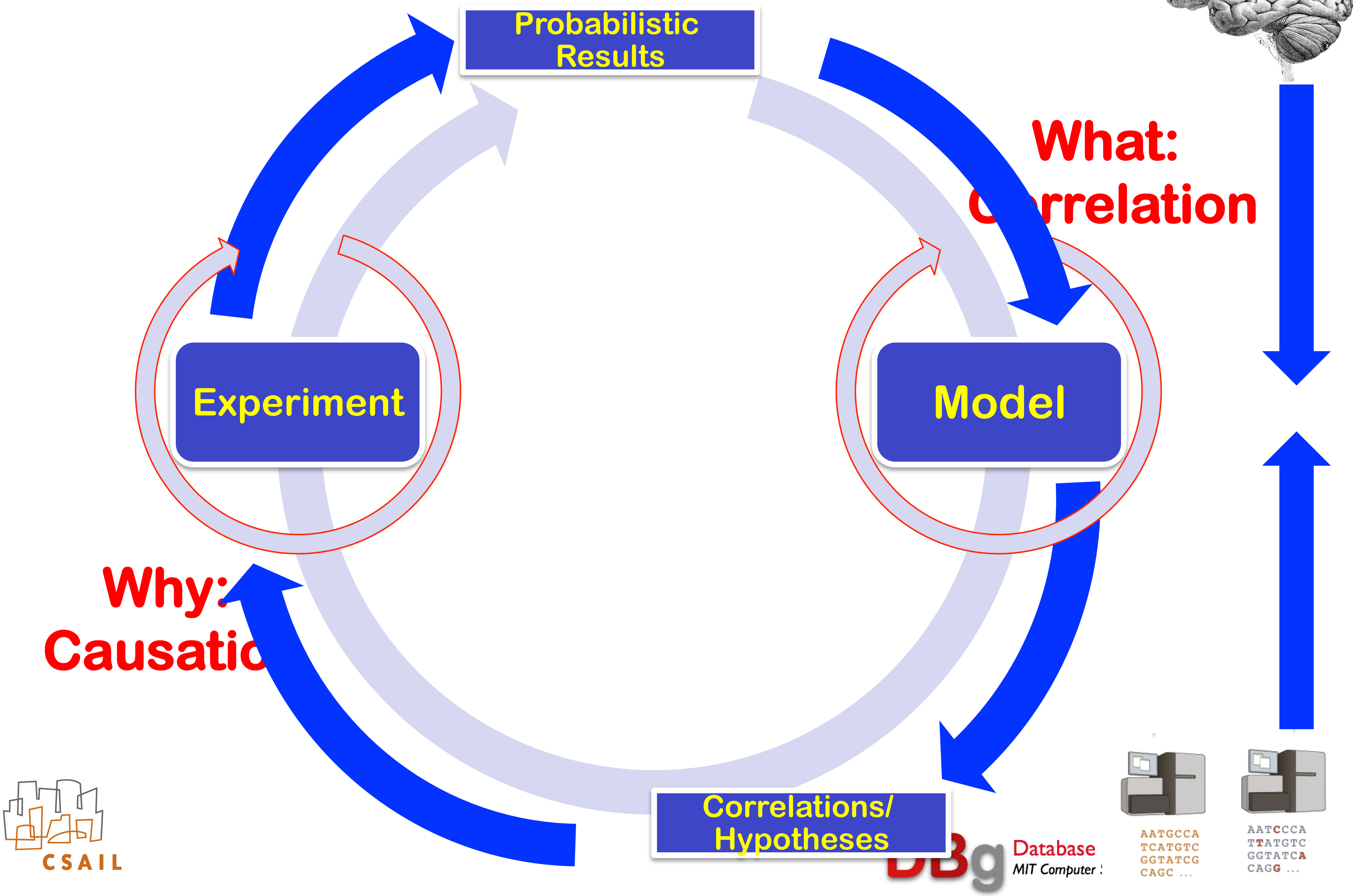
Precision Oncology



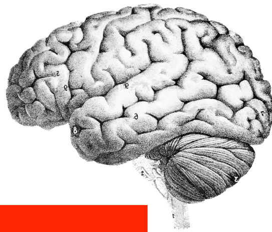
Source: Marty Tenebaum, Cancer Commons

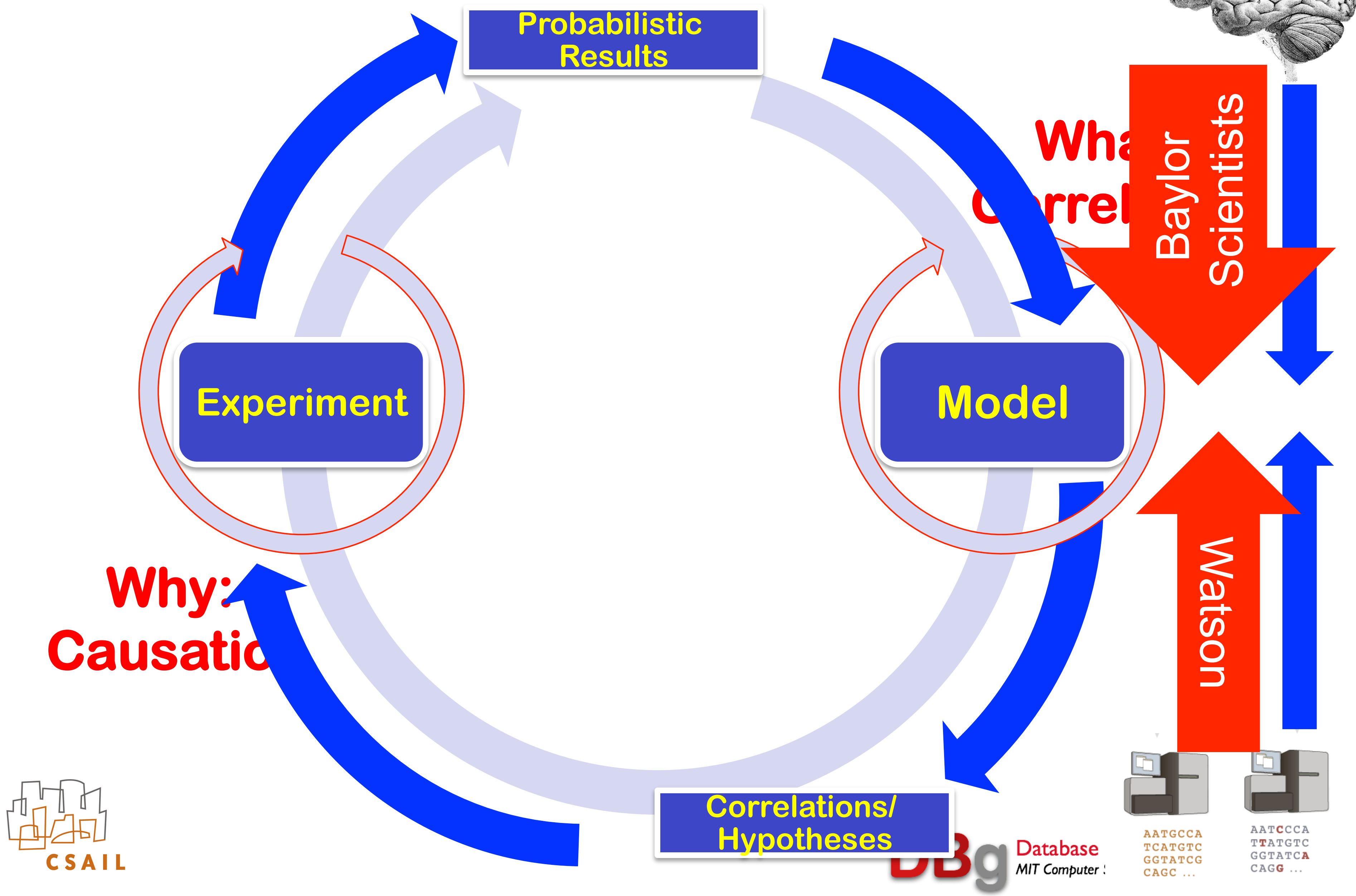
Accelerating Scientific Discovery





Accelerating Scientific Discovery





Profound Changes: Paradigm Shift [Kuhn]

- New reasoning / problem solving model
  - Data ➔ Data-Intensive (Big Data – 4 Vs)
  - Why ➔ What
  - Strategic (theory-based) ➔ Tactical (evidence-based)
  - Theory-driven (top-down) ➔ Data-driven (bottom-up)
  - Hypothesis testing ➔ Hypothesis generation
- Enabling Paradigm Shifts in most disciplines
  - Science ➔ eScience
  - Accelerating (scientific / engineering) discovery
  - Most domains
    - Personalized medicine
    - Drug interactions
    - Urban Planning
    - Social and Economic Planning
- Beyond Data-Driven: Symbiosis
  - What + Why
  - Human intelligence + machine intelligence

Big Data and Data-Intensive Analysis

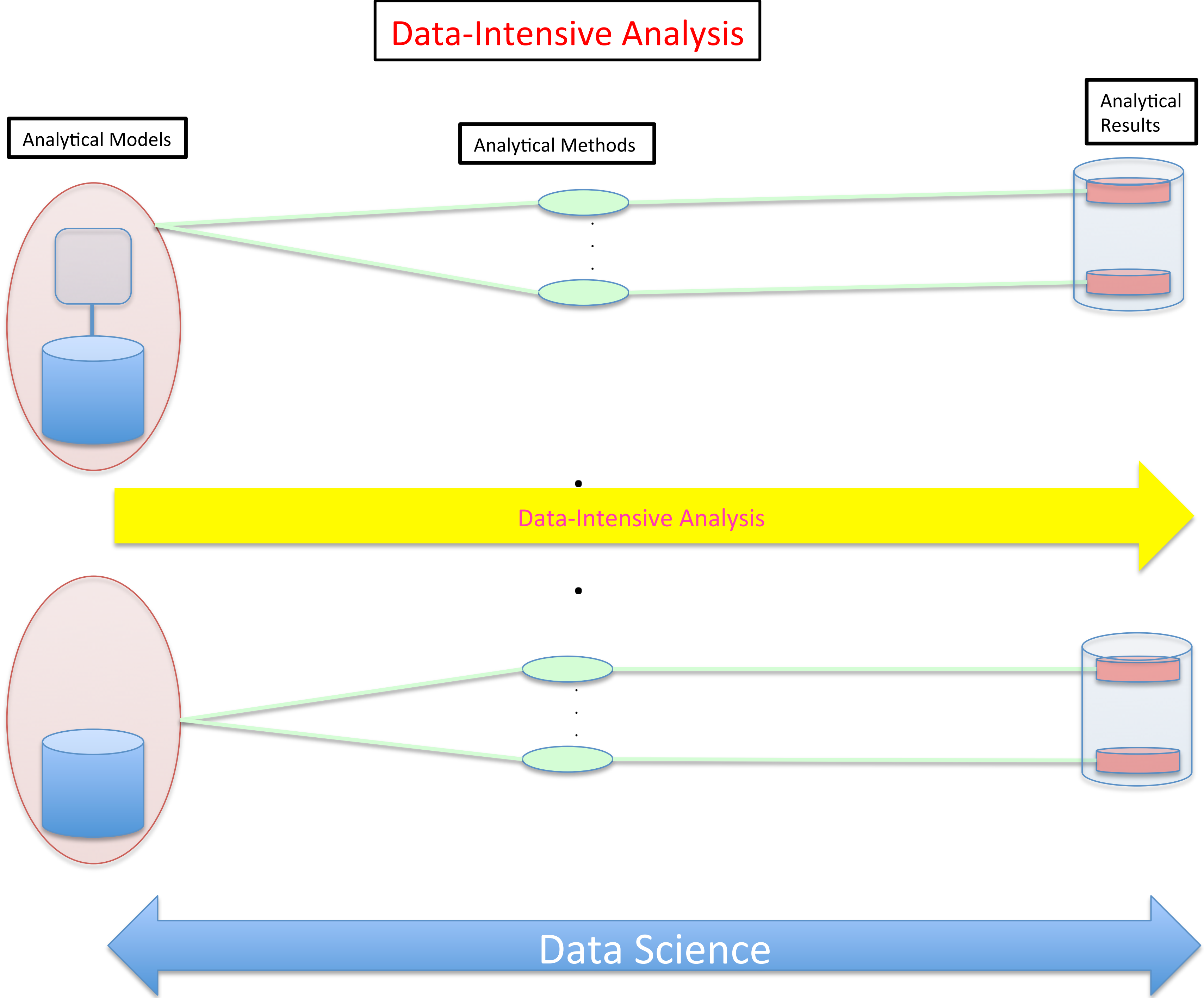
THE BIG PICTURE: MY PERSPECTIVE

DIA Pipelines / Ecosystem

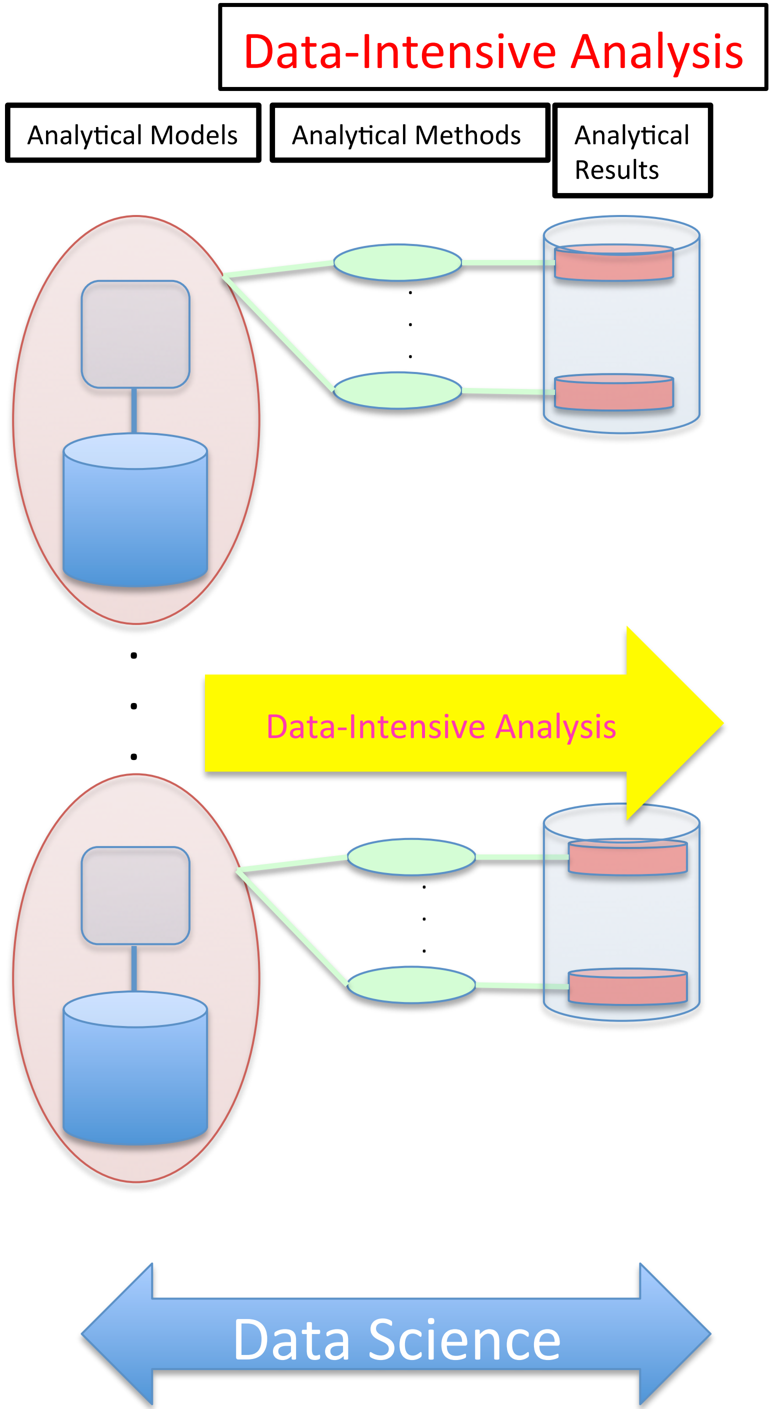
- Q: What Big Data technologies do you see becoming very popular within the next five years?

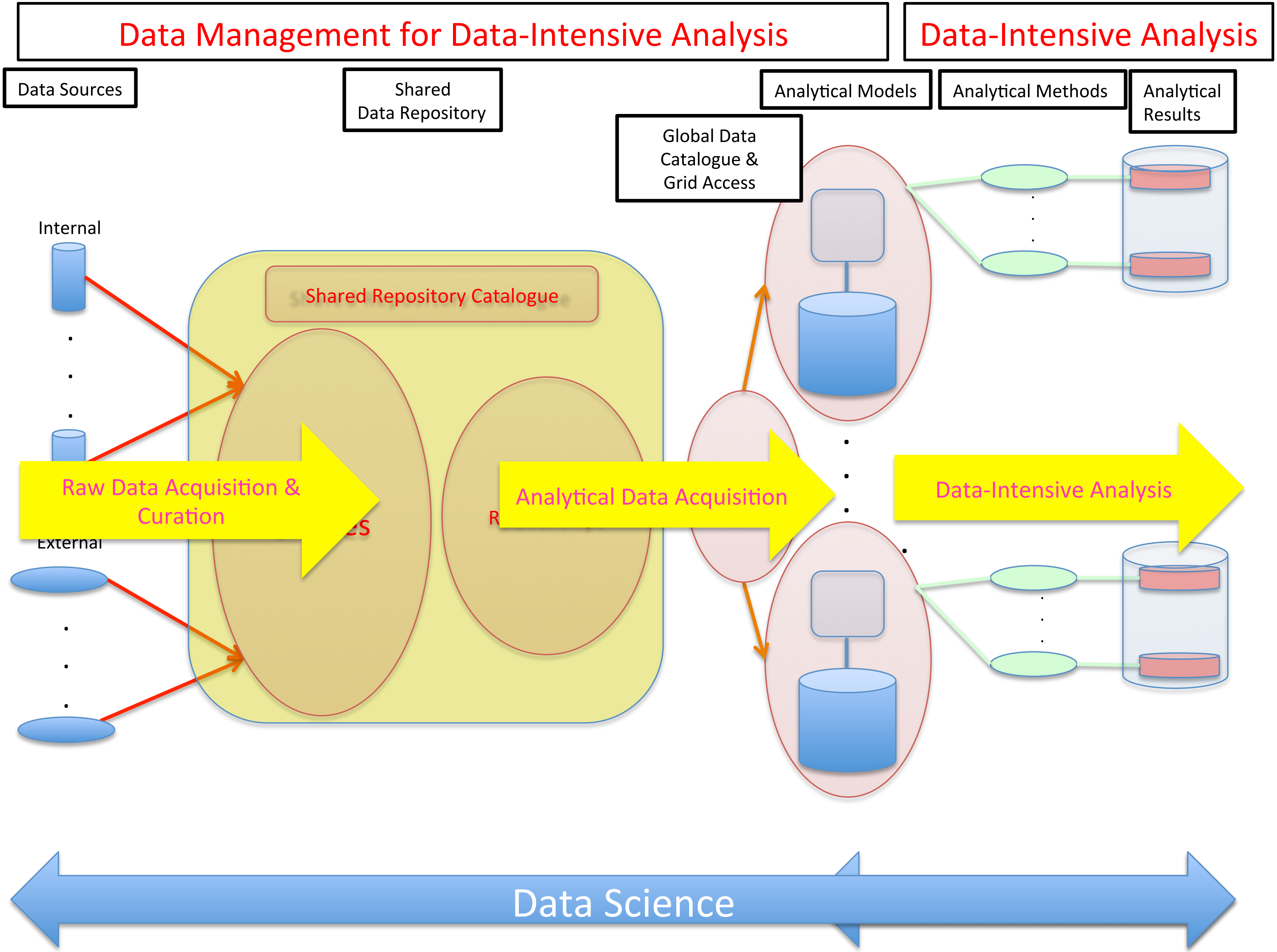
- A: I don't like to say that there's a specific technology, … there are pipelines that you would build that have pieces to them. How do you process the data, how do you represent it, how do you store it, what inferential problem are you trying to solve. There's a whole toolbox or ecosystem that you have to understand if you are going to be working in the field.

Michael Jordan, *Pehong Chen Distinguished Professor at the University of California, Berkeley*



Data-Intensive Analysis
Analytical Models
Analytical Methods
Analytical Results
Data-Intensive Analysis
Data Science


Data-Intensive Analysis
Analytical Models
Analytical Methods
Analytical Results
Data-Intensive Analysis
Data Science



Data Management for Data-Intensive Analysis
Data-Intensive Analysis
Data Sources
Shared
Data Repository
Analytical Models
Analytical Methods
Analytical
Results
Global Data
Catalogue &
Grid Access
Internal
External
Shared Repository Catalogue
Raw Data Acquisition &
Curation
Analytical Data Acquisition
Data-Intensive Analysis
Data Science

Research Method: Examine Complex, Large-Scale Use Cases that push limits

DATA-INTENSIVE ANALYSIS (DIA)
DIA PROCESS (WORKFLOW / PIPELINE)
DIA USE CASE RANGE

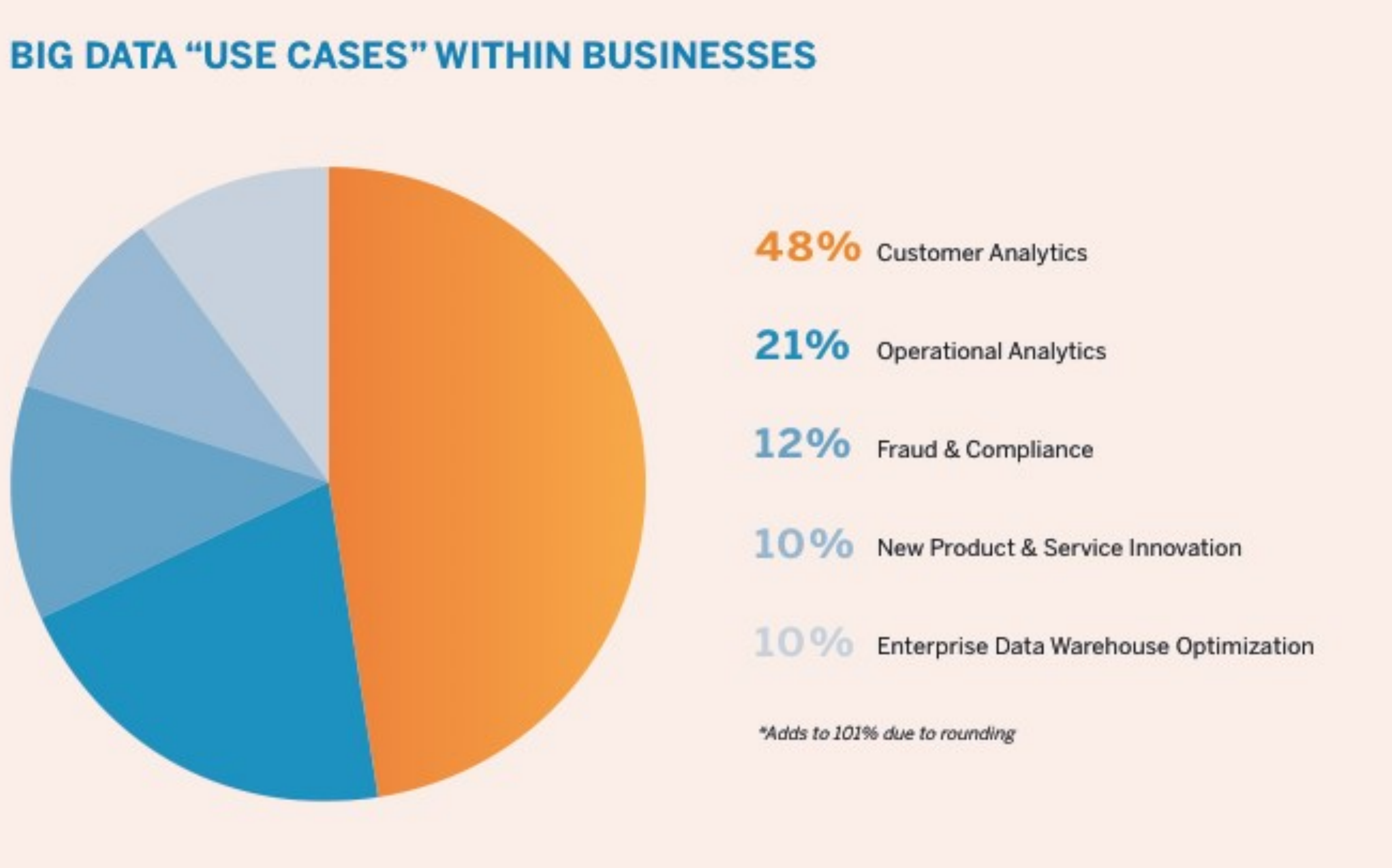
Data Analysis ➔ Data-*Intensive* Analysis

- Common definition– *far too simplistic* : *extract knowledge from data*

- **DIA:** *the activity of using data to investigate phenomena, to acquire new knowledge, and to correct and integrate previous knowledge*

- **DIA Process/Workflow/Pipeline**: *a sequence of operations that constitute an end-to-end DIA from source data to a quantified, qualified result*

My Focus is Not common DIA Use Cases



… Nor High Impact Organizational DIA

Data(-Intensive) Analysis Range *

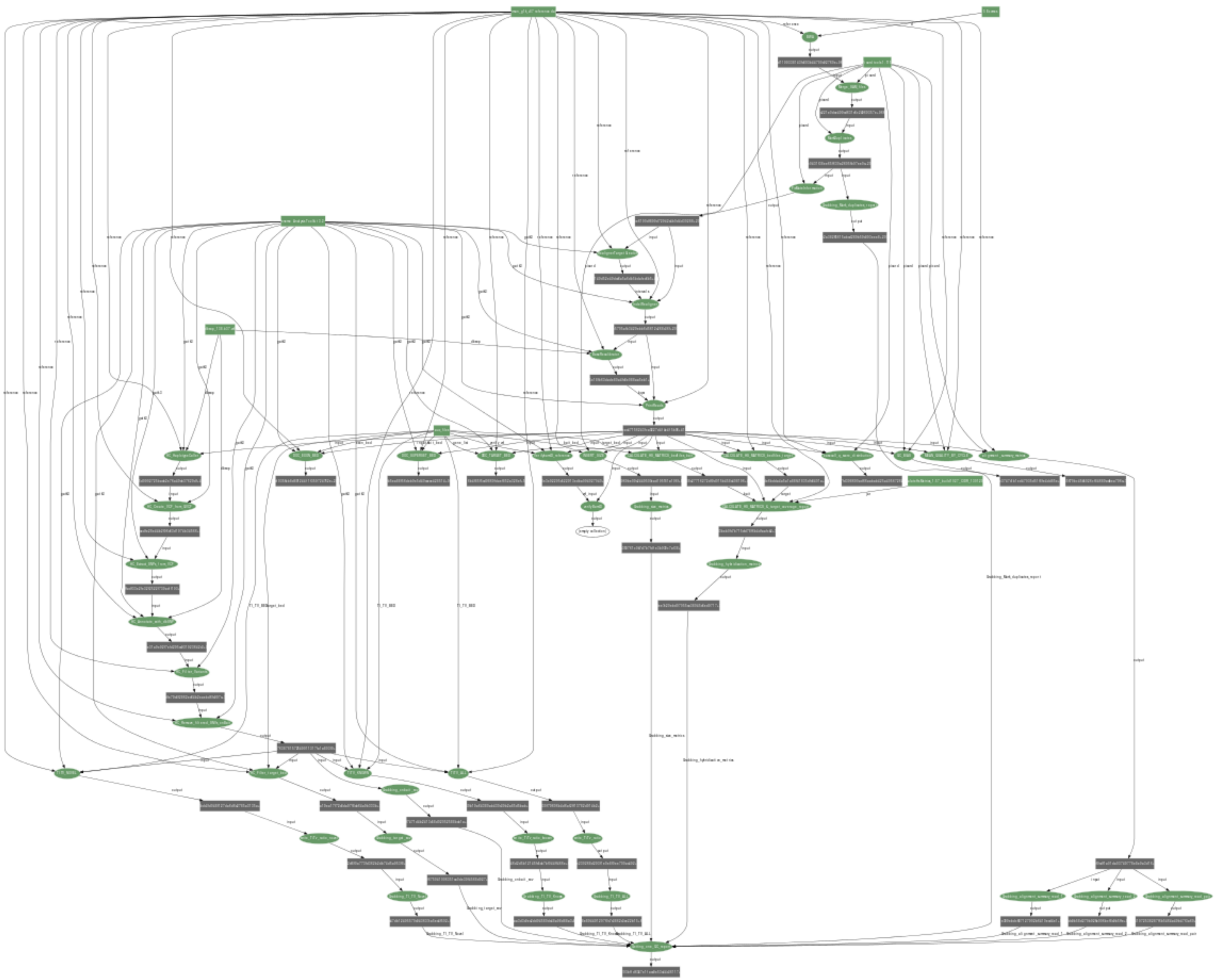
- *Small Data* ≠ (volume, velocity, variety) 98%
  - Conventional data analysis: 1 K years - statistics, spreadsheets, databases, …
- Big Data = (volume, velocity, variety) 2%
  - Simple DIA: "*most data science is simple*" Jeff Leek 96%
    - Simple models & methods, single user, short duration: 65+ self-service tools, ML, widest-usage
    - Relative simplicity: sales, marketing, & social trends, defects, …
  - Complex DIA 4%
    - Domains: particle physics, economics, stock market, genomics, drug discovery, weather, boiling water, psychology, …
    - Models & Methods: large, collaborative community, long duration, very large scale

Focus

Why?
A: This is where things *obviously* break …

* Many more factors

Example Scientific Workflow (Arvados)

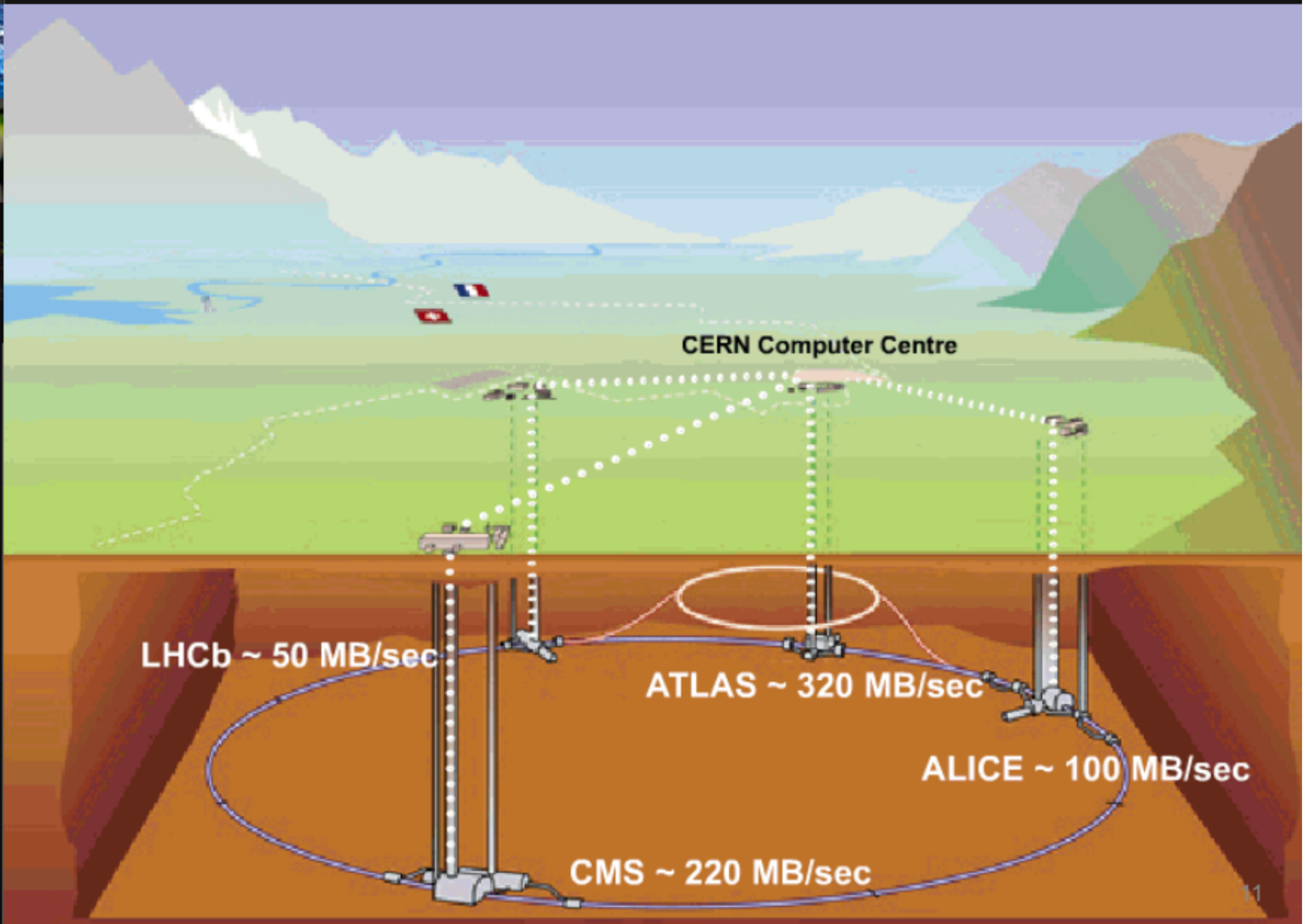


Complex DIA Use Case #1

eScience, Big Science, Networked Science, Community Computing, Science Gateway

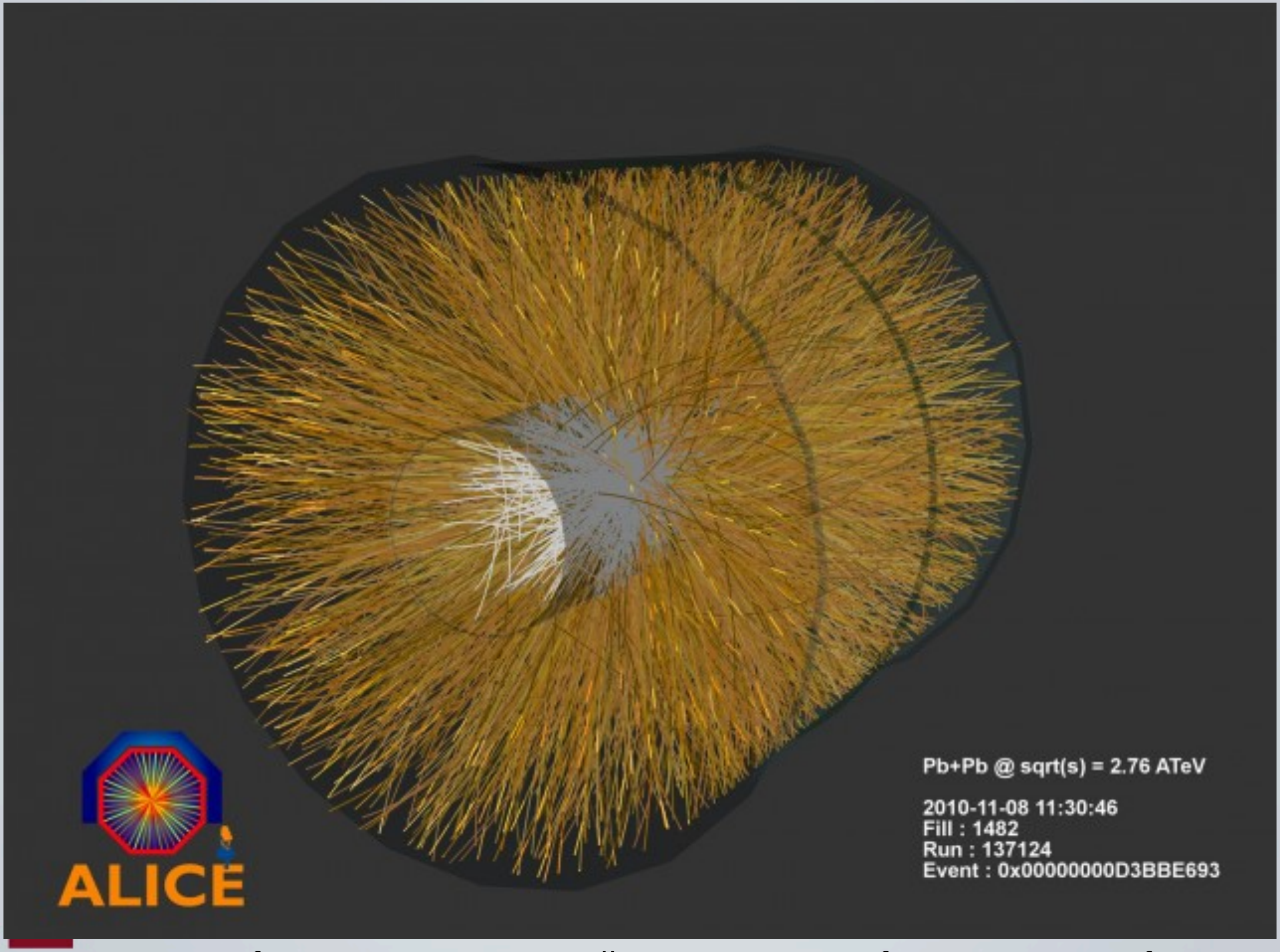
TOP-QUARK, LARGE HADRON COLLIDER, CERN, SWITZERLAND

Data acquisition and storage for LHC @ CERN



WLCG
Worldwide LHC Computing Grid

Higg's Boson: 40 Year Search

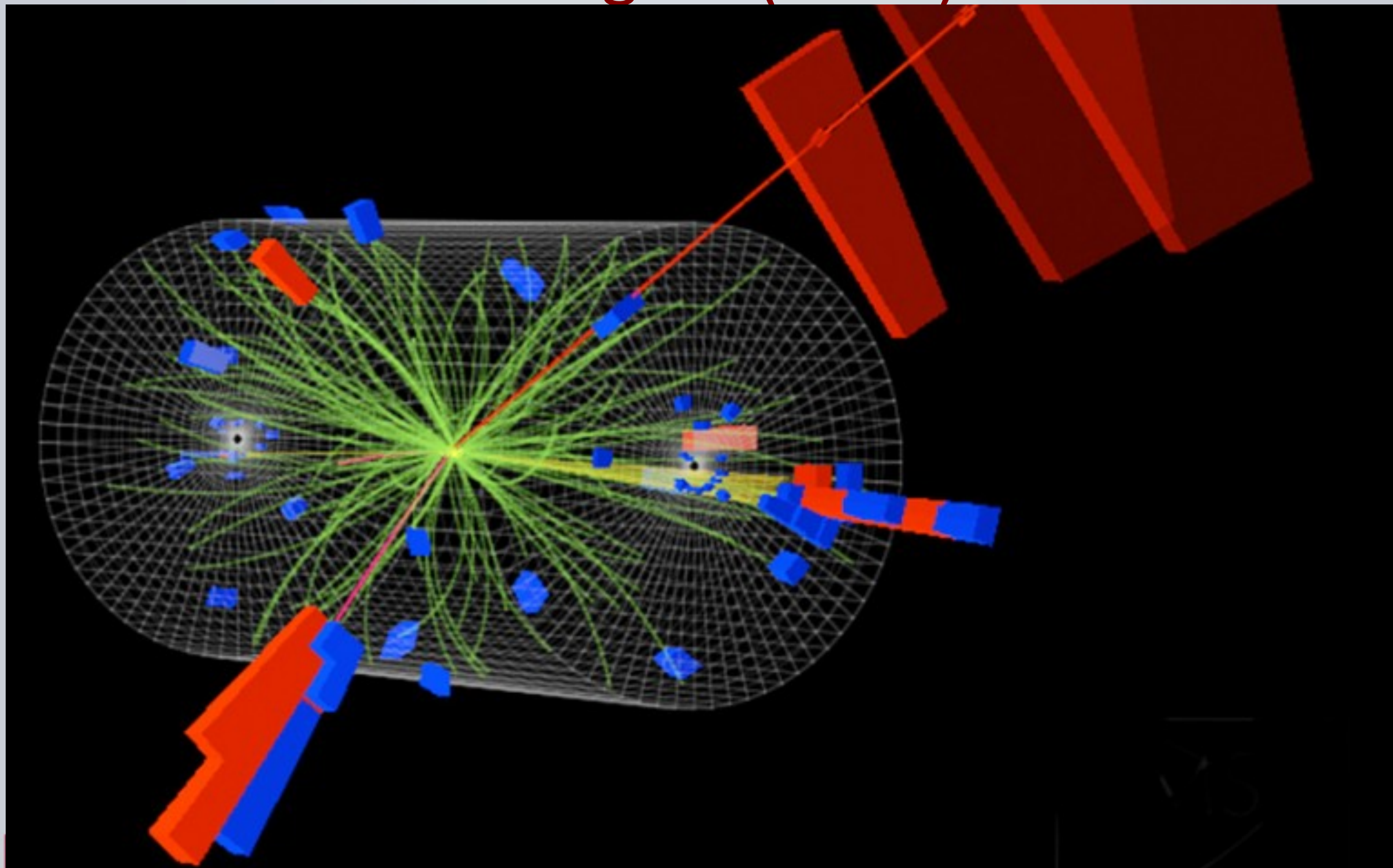


LHC Data from proton–proton collisions at centre-of-mass energies of 7 TeV (2011) and 8 TeV (2012)

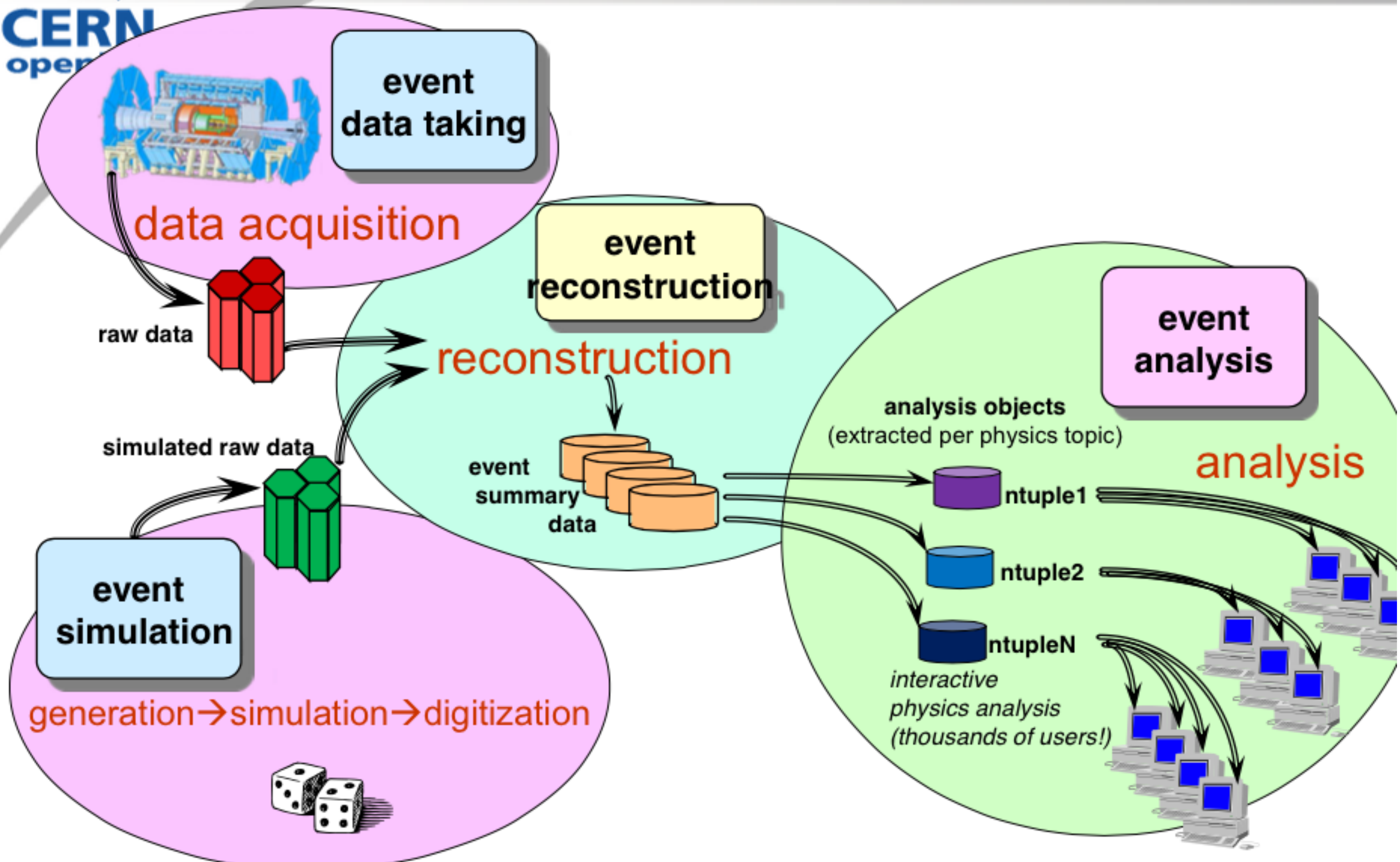
How do you Prove Higgs Boson Exists?

- Standard model of physics predicts (30 years) Higgs Boson characteristics
  - Mass ~125 GeV
  - Decays to γγ, WW and ZZ boson pairs
  - Couplings to W and Z bosons
  - Spin parity
  - Couples to up-type top-quark
  - Couples to down-type fermions?
  - Decays to bottom quarks and τ leptons

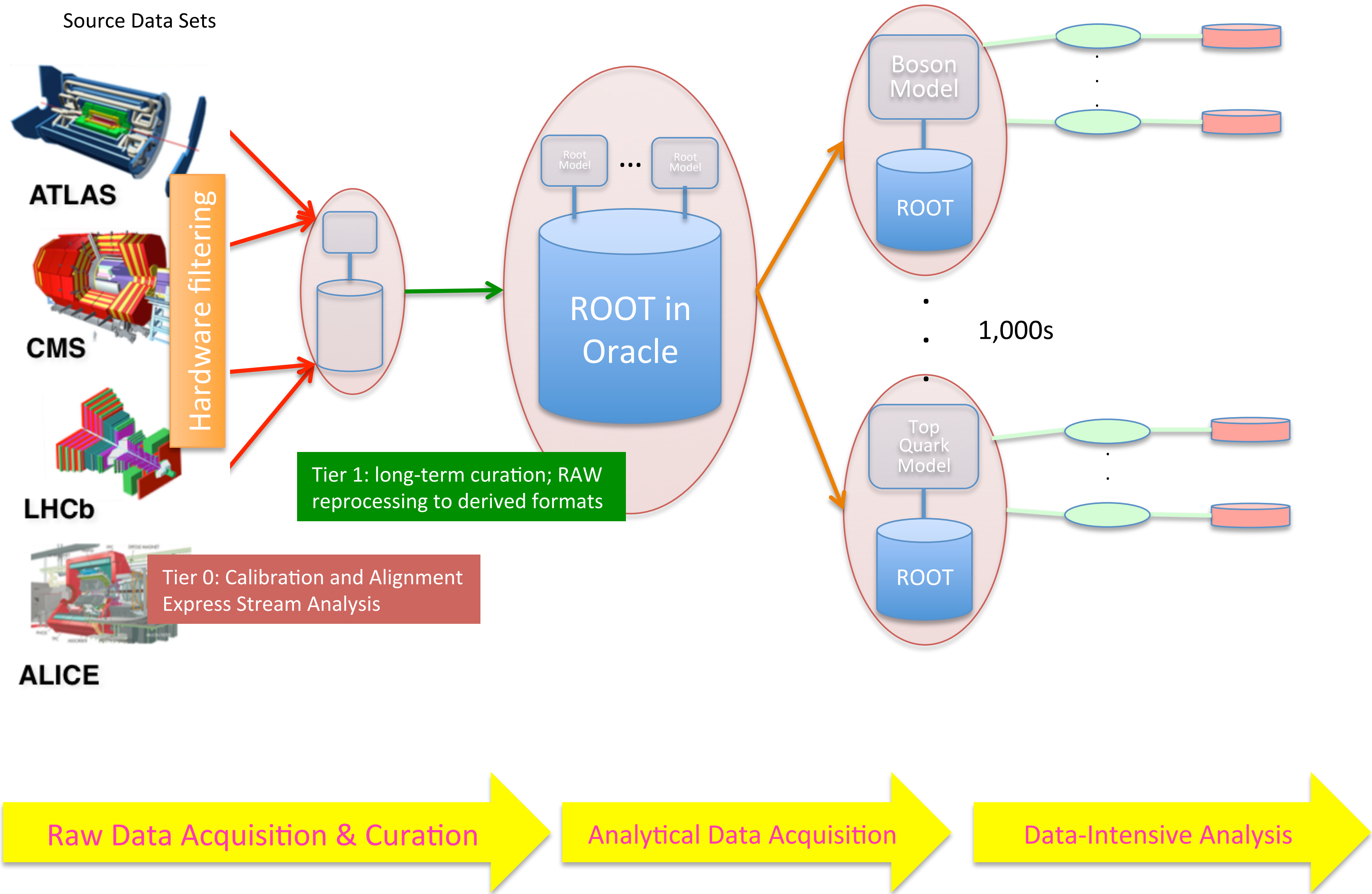
5 Sigma=0.00001% possible error (2012)
10 Sigma (2014)



LHC-scale data processing



Original Big Data Application, e.g., ATLAS high-energy physics (CERN)

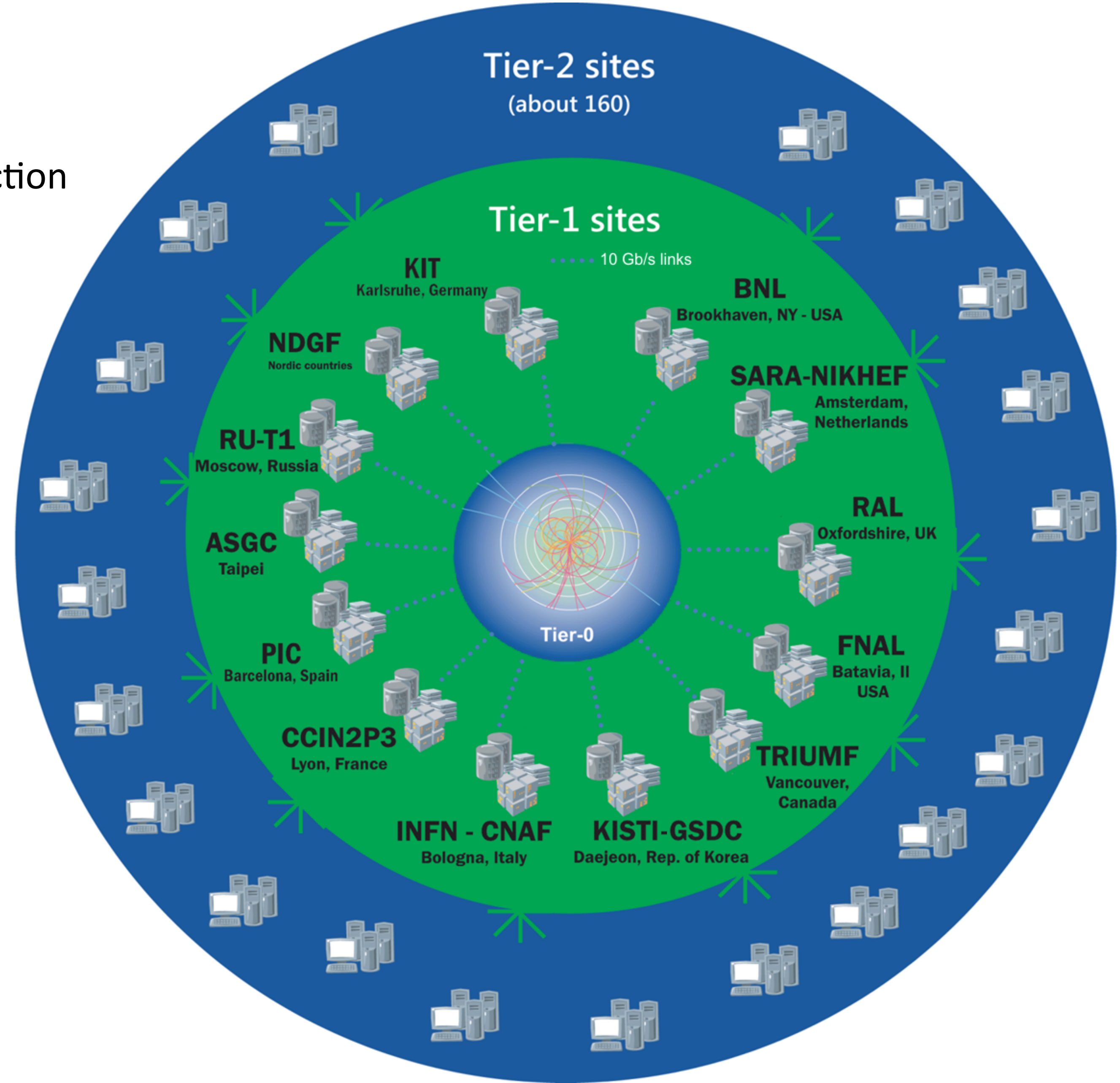


Worldwide LHC Computing Grid

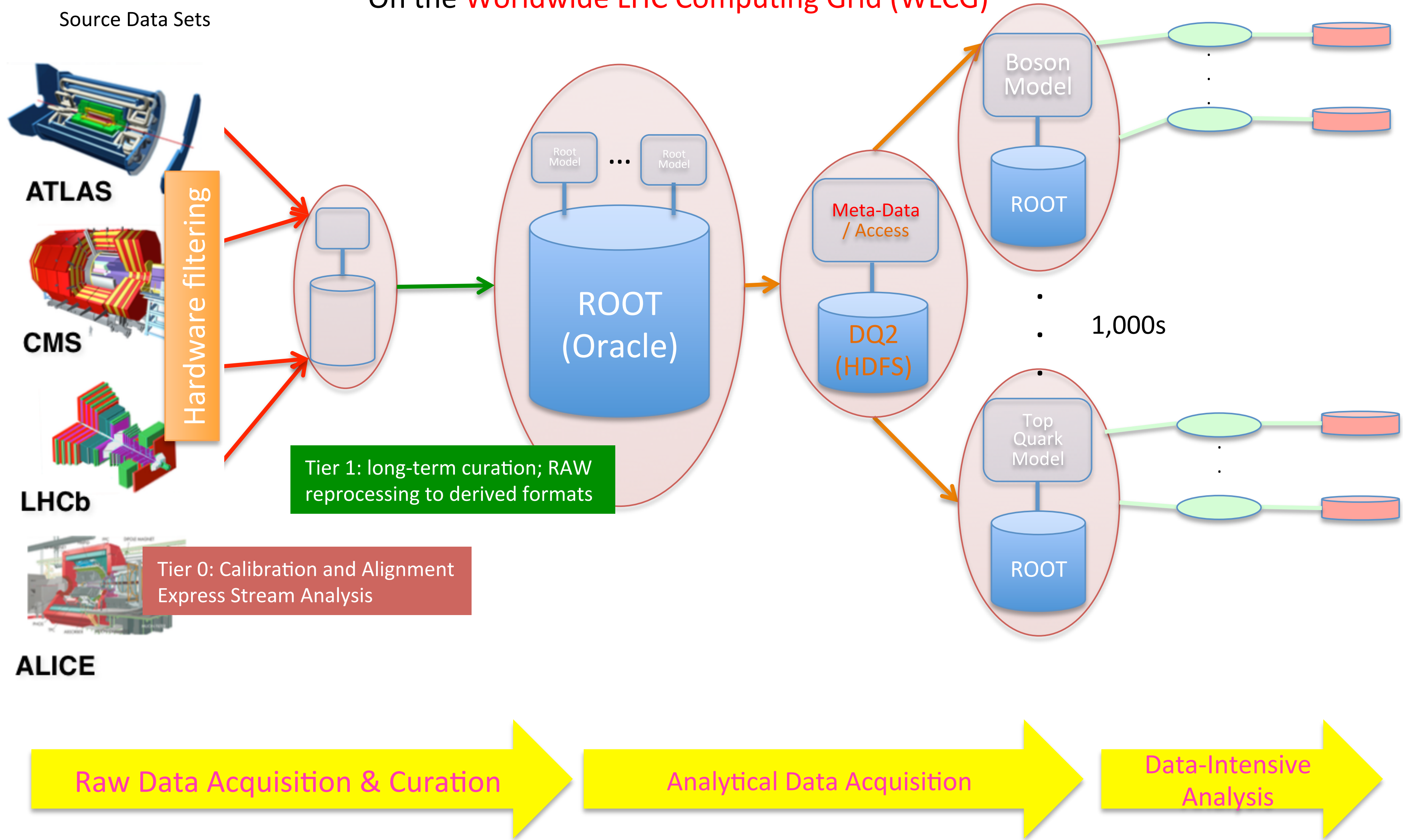
Tier 0 (CERN)
- Data recording
- Initial data reconstruction
- Data distribution

Tier 1 (13 centers)
- Permanent storage
- Re-processing
- Analysis

Tier 2 (~160 centers)
- Simulation
- End-user analysis



Original Big Data Application, e.g., ATLAS high-energy physics (CERN); Oracle + DQ2 + ROOT
ATLAS Distributed Data Management System (DQ2) (Pig, Hive, Hadoop) 2007+
On the Worldwide LHC Computing Grid (WLCG)



Based on ~30 Large-Scale DIA Use Cases

LESSONS LEARNED

DIA Lessons Learned (What)

- A Software Artifact: a workflow / pipeline
  - Data-Intensive Analysis Workflow
    - Data Management (80%)
      - (Raw) Data Acquisition and Curation
      - Analytical Data Acquisition
    - Data-Intensive Analysis (20%)
  - Objective: switch 80:20 to 20:80 ➔ *Let scientists do science*
  - Explore (DIA) vs Build (software engineering)
  - Duration: years
- Emerging Paradigm
  - New programming paradigm
  - Experiments over data
  - Convergence
    - Scientific / engineering discovery
    - ~10 programming paradigms: database, IR, BI, DM, …

DIA Lessons Learned (How)

- Result Types
  - Provable <-> Probabilistic <-> Speculative

- Nature
  - Analytical
    - Empirical: complete meta-data
    - Abstract: incomplete meta-data
  - Phases: Exploration, Analysis, Interpretation
    - Exploratory, Iterative, and Incremental
- Users
  - Individual
  - Workgroup
  - Organization / Enterprise
  - Community

DIA Lessons Learned (People)

- Machine + Human Intelligence
  - Symbiosis – optimized
  - Domain knowledge critical
- Multi-disciplinary, Collaborative, Iterative
- Community Computing: DIA Ecosystems – sharing
  - Massive resources
  - Knowledge
  - Costs
  - Many (~60): eScience, Science Gateways, Networked Science, …
    - High-energy physics (CERN: ROOT)
    - Astrophysics (Gaia)
    - Scientific Workflow Systems: ~30
    - Macroeconomics
    - Global Alliance for Genomics and Health
    - Enterprise Ecosystems, e.g., Information Services: Thomson Reuters, Bloomberg, …
    - Open-Science-Grid
    - The Cancer Genome Atlas
    - The Cancer Genomics Hub

DIA Lessons Learned (Essence)

The value and role of

What data is adequate evidence for Q?

truth

evidence-based causality

evidence-based correlations

Complex DIA Use Case #2

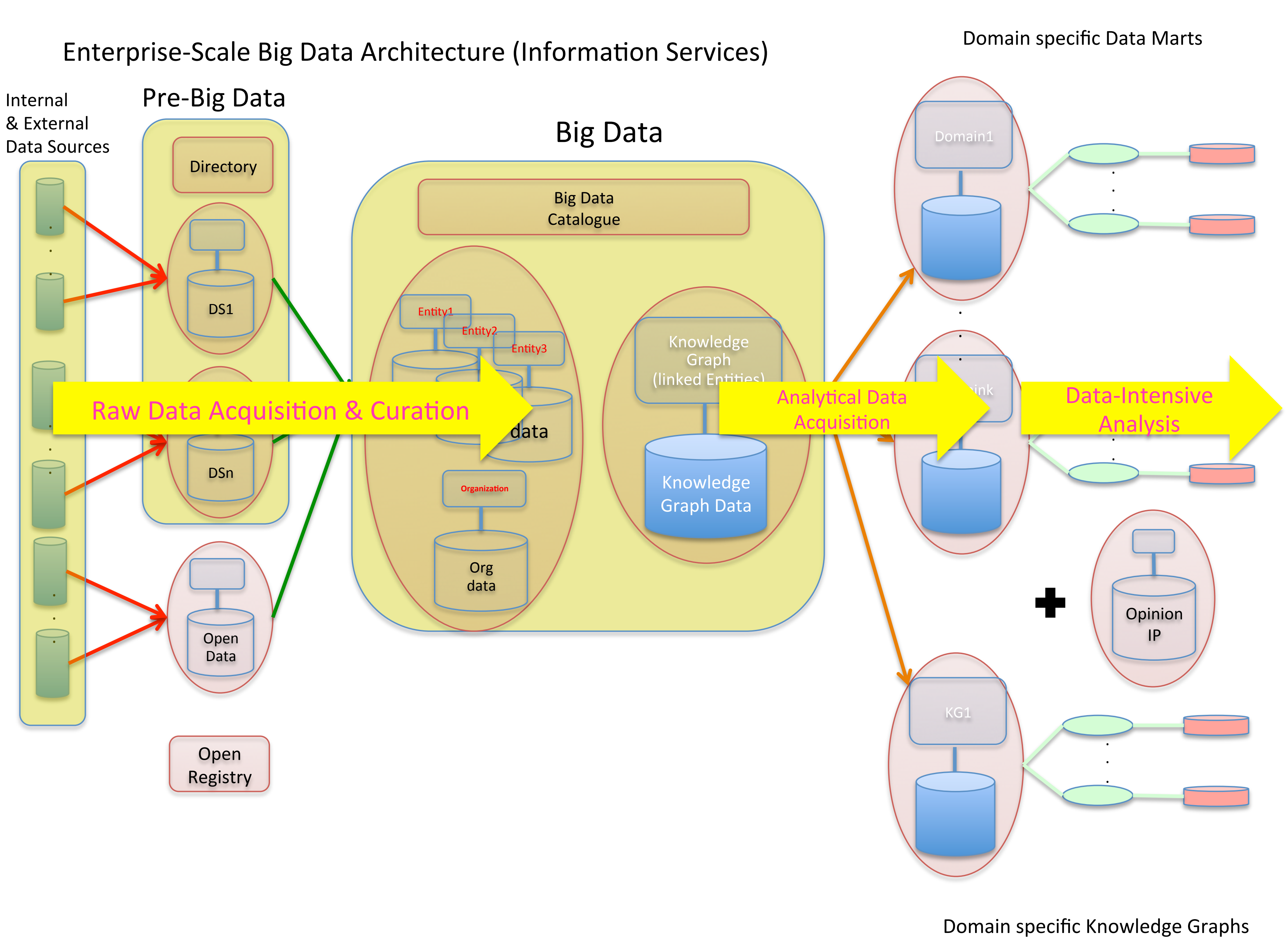
Information Services

DOW JONES, BLOOMBERG, THOMSON REUTERS, PEARSON, …

Information Services Business
Collect, curate, enrich, augment (IP) & disseminate information

- Financial & Risk
  - Investors
  - News & press releases
  - Brokerage research
  - Instruments: stocks, bonds, loans, …
- Legal
  - Dockets
  - Case Law
  - Public records
  - Law firms
  - Global businesses
- Intellectual Property & Science
  - Scientific articles
  - Patents
  - Trademarks
  - Domain names
  - Clinical trials
- Tax & Accounting
  - Corporate
  - Government
  - Solutions

Consequences of Errors



Enterprise-Scale Big Data Architecture (Information Services)
Domain specific Data Marts
Internal & External Data Sources
Pre-Big Data
Directory
DS1
DSn
Open Data
Open Registry
Big Data
Big Data Catalogue
Entity1
Entity2
Entity3
data
Organization
Org data
Knowledge Graph (linked Entities)
Knowledge Graph Data
Raw Data Acquisition & Curation
Analytical Data Acquisition
Data-Intensive Analysis
Domain1
Opinion IP
KG1
Domain specific Knowledge Graphs

DIA Lessons Learned

- Modelling
  - Analytical Models and Methods
    - Selection / creation, fitting / tuning
    - Result verification
    - Model / method management
  - Data Models
    - Entities dominate "Data Lakes"
    - Named Entities + Entity (Graph) Models
    - Ontologies (genomics), Ensembles, …
- Emerging DIA Ecosystems Technology
  - Languages (~30)
  - Analytics Suites / Platforms (~60)
  - Big Data Management (~30)

Veracity

WHAT COULD POSSIBLY GO WRONG?

Do We Know / Can We Prove

- DIA Result: *correct*, *complete*, *efficient*?
- What machines / algorithms / Machine Learning / Black Boxes / DIA do?

- High Risk / High Reward Data-Driven Society
  - Risk: drugs or medical advice that cause harm
  - Reward: faster, cheaper, more effective cancer cures, drug discovery, personalized medicine, …

Professional Cautions

- Experienced practitioners
- Medicine
  - Few data-driven results operationalized
  - Mount Sinai: no black box solutions
- Authoritative organizations
  - NIH, HHS, EOCD, National Statistical Organizations, …
- Legal*: Algorithmic Accountability*: John Zittrain Harvard Law School

Q: What could possibly go wrong? A: Every step

- **Data Sets**
  - All measurements approximate: availability, quality, requirements, sparse/dense, … ; How much can we tolerate? What is the impact on the result?
- **Models**: "All models wrong …" George Box 1974
- **Methods**
  - Select (1,000s), tune, verify
  - Different methods ➔radically different results
- **Results**: Probabilistic, error bounds, verification, …

Data Analysis is 20% of the story

Pre Big Data Challenges

- **Science**: Experimental design: hypotheses, null hypotheses, dependent and independent variables, controls, blocking, randomization, repeatability, accuracy
- **Analysis**: models, methods
- **Resources**: cost, time, precision

+Big Data Challenges …

- Pre-Big Data Challenges @ scale: volume, velocity, and variety
- **Complexity**
  - Data: sources, meta-data, 3Vs
  - Models (reflecting the domain)
  - Methods (multivariate patterns) beyond human cognition
  - Results: Massive numbers of correlations
- **Unreliability** (statistics @ scale)
  - Reliability decreases as the number of variables increases (multivariate analysis)
  - << 10 variables (science &drug discovery) ➔ 1,000s to millions (Machine Learning)
    - "In science and medical research, we've always known that"
- **Misunderstood: Self-service, Automated Data Science-in-a-Box**
  - 80% unfamiliar with statistics, error bars, causation/correlation, probabilistic reasoning, automated data curation and analysis
  - Widespread use of DIA: self-service, "democratization of analytics"
  - Like monkeys playing with loaded guns

Somewhere over there …

DIA Verification Principles & Techniques

- Conventional disciplines
- Man-machine symbiosis
- DIA Result ➔Empirical evidence
  - Flashlight analogy: DIA reduces hypothesis space
- Cross-validation
  - Validate predictive model: avoid overfitting, will model work on unseen data sets?
  - Data partitions: Training Set, Test /Validation Set, ground truth
  - K-fold cross validation
- Research Direction
  - New measures of significance, the next generation P value
  - 21st Century statistics

I Proposal

SCIENTIFIC METHOD → EMPIRICISM
DATA SCIENCE → DATA-INTENSIVE ANALYSIS

Data Science is …

A body of principles and techniques for applying data-intensive analysis for investigating phenomena, acquiring new knowledge, and correcting and integrating previous knowledge with measures of correctness, completeness, and efficiency.

DIA: an experiment over data

Conclusions
Big Data & Data-Intensive Analysis

- Value of evidence (from data)
- Emerging reasoning and problem solving paradigm
  - High risk / high reward
  - Substantial results already
  - In its infancy, not yet understood, decades to go
  - Overhyped (short term) but may change our world (long term)
- → Need for Data Science = principles & guidelines
  - "We're now at the "what are the principles?" point in time" M. Jordan
  - Decades of research and practice

MIT CSAIL

Thank You