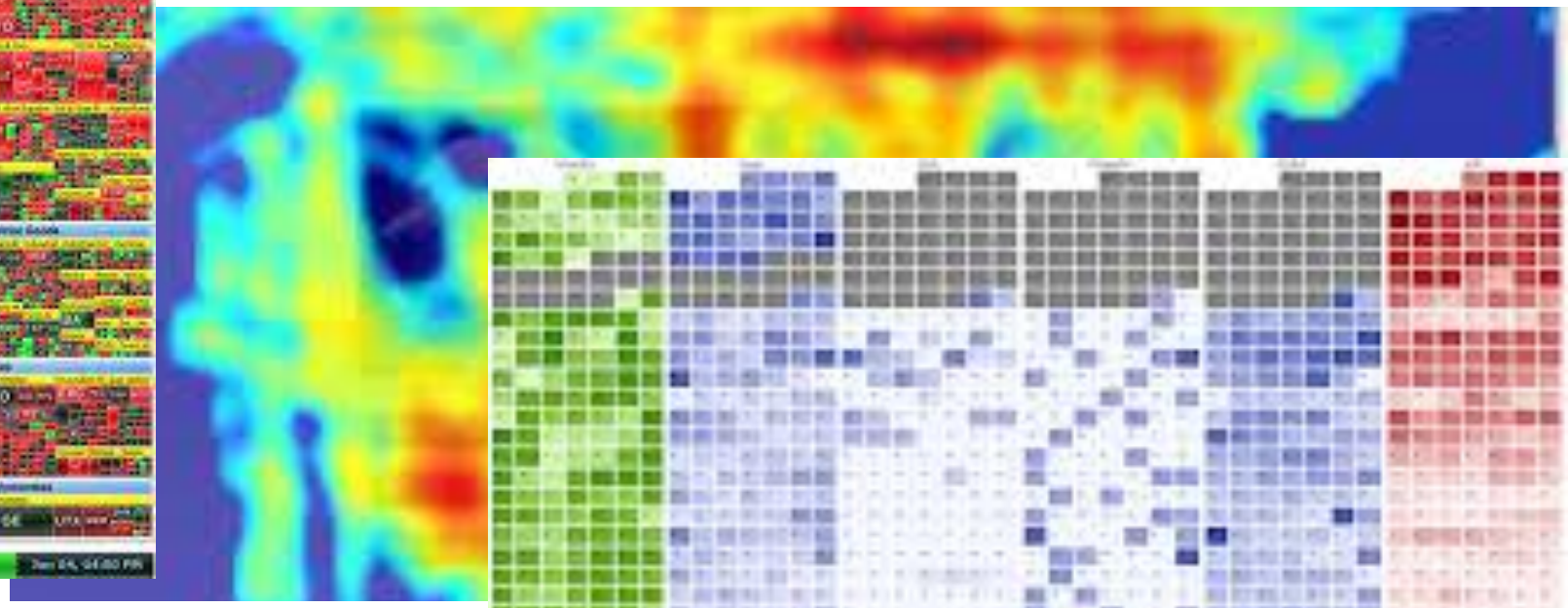
Big Data Analytics

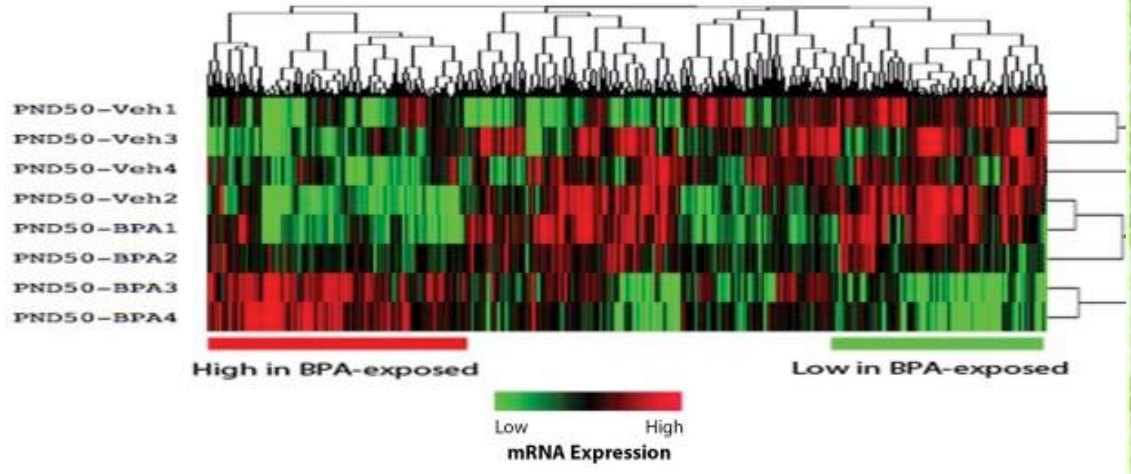
Michelle Wiest

University of Idaho

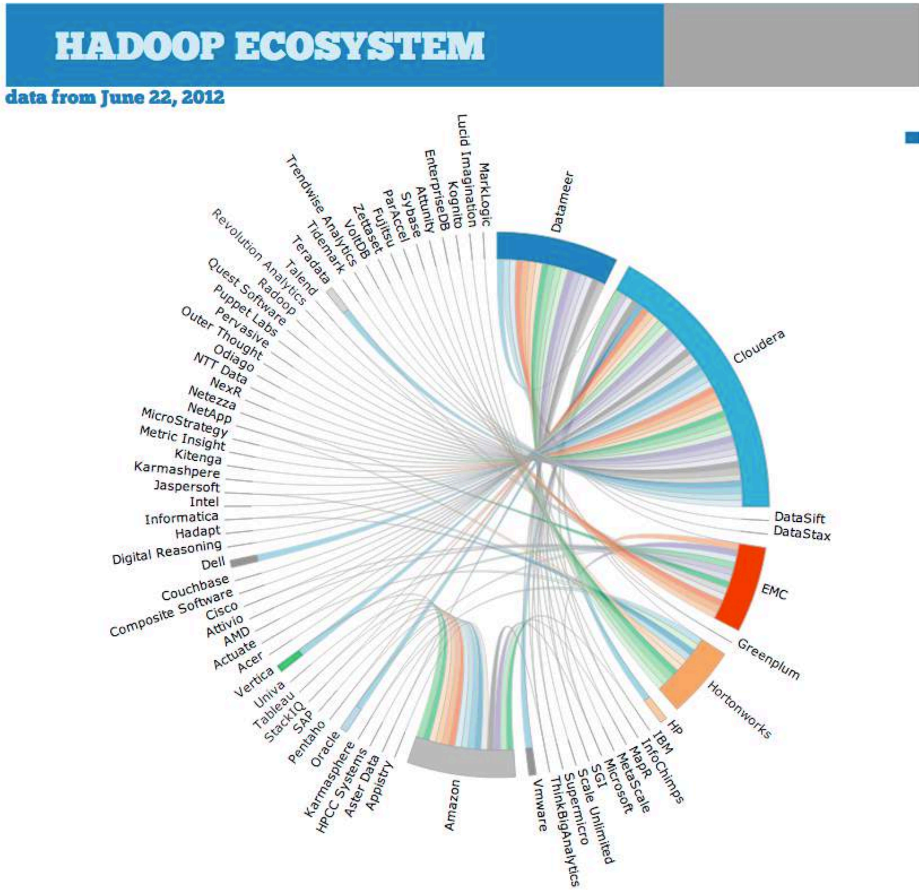
Analytics/Data Mining/Decision Support

Heatmap Trauma





Today


HADOOP ECOSYSTEM
data from June 22, 2012
Datameer
Cloudera
DataSift
DataStax
EMC
Greenplum
Hortonworks
HP
IBM
InfoChimps
MapR
MetaScale
Microsoft
SGI
Scale Unlimited
Supermicro
ThinkBigAnalytics
Vmware
Amazon
Appistry
Aster Data
HPCC Systems
Karmasphere
Oracle
Pentaho
SAP
StackIQ
Tableau
Univa
Vertica
Acer
Actuate
AMD
Attivio
Cisco
Composite Software
Couchbase
Dell
Digital Reasoning
Hadapt
Informatica
Intel
Jaspersoft
Karmashpere
Kitenga
Metric Insight
MicroStrategy
NetApp
Netezza
NexR
NTT Data
Odiago
Outer Thought
Pervasive
Puppet Labs
Quest Software
Radoop
Revolution Analytics
Talend
Teradata
Tidemark
Trendwise Analytics
VoltDB
Zettaset
Fujitsu
ParAccel
Sybase
Attunity
EnterpriseDB
Kognito
Lucid Imagination
MarkLogic


BIG
DATA
INFORMATION
STORAGE
LARGE
SETS
ANALYSIS
RESEARCH
SEARCH
PETABYTES
MANAGEMENT
TECHNOLOGIES
VOLUME
SOFTWARE
PROCESS
GOVERNMENT
DEVELOPMENT
TECHNOLOGY
MILLION
SCIENCE
FUTURE
BILLION
NEW
INTERNET
LEARNING
EVERY
WORLD
BASED
EXABYTES
DATABASES
PROCESSING
SAN
NAS
SOCIAL
SIZE
INITIATIVE
PARALLEL
TERABYTES


Derrick Harris

Fundamental flaws with “big data analytics”

- Decoupling of scientific inquiry from analysis approach
- Inappropriate methods for the data or the question
- Mislead by the idea that big data tells the truth (“big data hubris”)

An Anecdote

I just duplicated the data…

$$x\downarrow 1 - x\downarrow 2 \,/ \sqrt{2} s\uparrow 2 \,/ n$$

$$x\downarrow 1 - x\downarrow 2 \,/ \sqrt{2} s\uparrow 2 \,/ 2n$$

…so our results were significant

Avoiding pitfalls

Analysis must be appropriate for the goal

- Predictive models - models are merely algorithms and the quality of a model is assessed by its performance for predicting new observations.
  - This is where many business applications live

- Estimation models – models provide a better understanding of data and of the underlying mechanism which have produced it.
  - This is where most science lives

Analysis must be appropriate for the data

- Nature reported that Google Flu Trends was estimating more than double the percentage of doctor visits for influenza like illness than the CDC reports during the 2012 2013 flu season
- Strong evidence of autocorrelation and seasonality in the GFT errors, and the issues were likely, at least in part, due to the decision by GFT engineers not to use previous CDC reports or seasonality estimates in their models
- Overlooks considerable information that could be extracted by traditional statistical methods.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." Science 343 (14 March): 1203-1205. Copy at http://j.mp/1ii4ETo

Interpretation must be made in context of data collection

- All empirical research stands on a foundation of measurement. Is the instrumentation actually capturing the theoretical construct of interest?

  Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. “The Parable of Google Flu: Traps in Big Data Analysis.” Science 343 (14 March): 1203-1205. Copy at http://j.mp/1ii4ETo

- Many social data collection methodologies are biased in regards to the question or population of interest.

Ensuring rigour

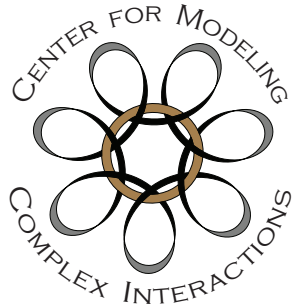
Integration

- Those performing big data analytics must have deep knowledge of the **application** which generated the data, **statistical theory**, and **computation**
- If all of these don't exist in one person, multiple people must be engaged from beginning to end.

IBEST and BCB Program at UI

- Institute for bioinformatics and evolutionary studies (IBEST)
  - Faculty members in IBEST include physicists, chemists, molecular biologists, organismal biologists, ecologists, behavioral biologists, **mathematicians**, **statisticians**, and **computer scientists**
- Bioinformatics and computation (BCB) biology graduate program
  - Requires students to engage in research and studies at the intersection of **computer science**, biology, **mathematics**, and **statistics**

Collaboratorium for Modeling Complex Problems



- Many of the big, outstanding problems facing society involve highly complex processes.
- These problems cut across traditional disciplines, solutions require a collaborative environment where researchers with different expertise can join forces.
- The Collaboratorium is designed to be a place where researchers work together using a broad range of modeling techniques, leading to innovative solutions, and where students and postdoctoral researchers are trained in multidisciplinary research.

Collaboratorium Complex Problems

- Molecular modeling
  - Specific application: Ebola virus glycoprotein
- Experimental design
  - Specific application: simulating murine and fly experiments on viral co-infection
- Model integration
  - Must we take problems down to common element?
  - Integrate at higher level?
  - Propagation of uncertainty
  - Specific applications: Social epidemiology in context of Ebola, MILES
  - Parallels database integration

Big picture areas of development: A Challenge

- Integrate statistical theory and database creation
  - Sufficency
- The structure of the data influences analysis and vica versa
- Need theoretical statisticians working with people like Hasan
  - Need to be willing to work with each other!

Training at UI – Analytics for all!

- Analytics certificate
  - Jointly offered developed by statistics and information systems management (business)

Thank you

Analysis of Systemic Metabolism Leads to Discovery of Lipokine

High dimension lipidomic data is used to identify the compartments with most substantial effects.

Then followed up with experimentation to confirm hypotheses.

Cao et al. DOI: http://dx.doi.org/10.1016/j.cell.2008.07.048