Big Data around UW-Madison

Brian S. Yandell,UW-Madison

www.stat.wisc.edu/~yandell

Our networks are awash in data.
A little of it is information.
A smidgen of this shows up as knowledge.
Combined with ideas, some of that is actually useful.
Mix in experience, context, compassion, discipline, humor, tolerance, and humility, and perhaps knowledge becomes wisdom.

Cliff Stoll (1995 *Silicon Snake Oil*)

PBS *NOVA* (1990) "The KGB, the Computer, and Me"

What is (are) big data?

- Old school: n > 100
- Recent history: n > 2 gigabyte
- Now: n > tera/peta/exa/zetta-byte
- Static or dynamic data?
- Single user? Shared?
- Public or private access?
- Curation and provenance (meta-data)
  - Increasingly important as data size grows

Wikipedia definition of “big data”

- datasets that grow so large that they become awkward to work with
  - capture, storage, search, sharing, analytics, visualizing
  - on-hand database management tools are indequate
    - relational databases
    - desktop statistics/visualization packages
  - requires "massively parallel software running on” 10-1000 servers[
  - current limits: terabytes, exabytes and zettabytes
- increasingly gathered by ubiquitous information-sensing mobile devices
  - aerial sensory technologies , wireless sensor networks, software logs
  - cameras, microphones, RFID readers
- benefits of working with larger and larger datasets
  - allows analysts to "spot business trends, prevent diseases, combat crime.“
- Subject areas (now)
  - meteorology, genomics, connectomics, complex physics simulations
  - biological and environmental research , Internet search
  - finance, business informatics
  - (social) media, publications, audio, video, interactive gaming

Why do we care?

- Comply with federal, private grants (audit risks)
  - security, confidentiality, access, IRB, FERPA
  - demands by NSF for data plans, NIH to publish data
- Make efficient use of scarce resources
  - save time, money, people
  - reduce errors: detection, correction
  - reduce duplication of effort in separate research projects
- Facilitate reproducible research
  - redo calculations years later
  - compare old to new data/methods
  - document data study steps in detail
- Share data within and among project groups
- Visualization: move quickly from data to insight
- Keep up with growing size: tera/peta/exa/zetta-bytes

Why is visualization important?

- Need to filter/abstract to key features
  - Picture is worth 1000 words
- Need quick views to compare ideas
  - Results using multiple methods
  - Multiple datasets from different sources
- Human mind is amazing analytic tool
  - Catalyzed by excellent visuals
- Problem: visuals need to scale up well

What types of data matter? (everything)

- Metadata: data about data
  - data description
  - study plan, experimental design
  - diagnostics and analytics: plans, tools, scripts
- Molecular
  - DNA, RNA, protein sequences and 2/3/4-D structure
  - transcriptomic, proteomic, metabolomic
  - interactomes, pathways and networks
- Spatial/temporal
  - images at all scales, static and dynamic
  - geospatial alliance, biomedical imaging, networks
  - point/line/object data
- Population-based
  - socioeconomic, cultural, political, health
  - transportation, financial, linguistic
- Methodology
  - code, algorithms, pipelines, workflows, user interfaces
  - visualization: static, dynamic, interactive
  - publication instruments: papers, graphs, audio, video, interactive
  - reproducible research tools

Data Storage is Cheap (or is it?)

- Pay Amazon (or …) to use their cloud
  - Amazon: $50-120/TB/month scalable
  - Backblaze: $4-5/month for unlimited storage
- Or build it yourself
  - Petabytes on a budget v2.0: Backblaze
    - 135 terabytes for ~$6-7K
  - But who is going to maintain it?
    - You? DoIT? *Confluence* enterprise solution?
    - Human costs are ultimately the big issue

Enterprise Storage System

(*Confluence* use at Biomedical Computing Group)

- direct access to "snapshot's" of data
- 12TB total with quick expansion to 24TB usable
- institutional cost model
- Work-spaces
  - bring some sanity and structure
  - customize specifically for user needs

Enterprise Work-Spaces

- User Work-Spaces
  - user specific, access limited
  - disk quota: 25GB for “home directory” work-spaces
- Project Work-Spaces
  - shared work-spaces for data sets/files shared among team members/co-workers
  - data retention, backups, archiving and access controls are strictly controlled
  - Example: SDAC drug study
- Computational Work-Spaces
  - shared work-spaces for high throughput computational usage
  - less strict data retention and no archiving need
  - Example: many different users all accessing and updating many shared files
- Data Warehousing Work-Spaces
  - large data sets generally written once and read many times
  - local repository for extremely large (multi Tera-byte) genetics or statistical data sets
  - no backup or retention requirements
  - can be re-fetched from another location

What are typical data process steps?

• key questions	why am I doing this study?
• experimental design	What is my data plan (***before*** you start the study)
• data inspection	What is the quality of my data?
• data management	Where is my master copy(s)? How do data it evolve? What about meta-data? (provenance)
• data analysis	What questions to ask? When (before, during, after)
• interpretation	What is the context for study?
• access/archiving	How to let others use data?

What choices do I make up front? (omic example)

- What are the key questions?
  - Why do omic studies at all? Fishing expedition or??
    - Forward genetics/genomics: discovery of disrupted pathways
  - Value of controlled experiment
    - Scientist-controlled comparison: what is effect of “insult”?
      - Genetic vs. treatment & control
  - Think of hypothesis generation, not testing
- But the cost!
  - Costs beyond massive high-throughput platforms
    - Need to prepare samples, which is costly in itself
    - Need to manage data once omic chips are run
  - Technical costs drop as use increases
    - High-throughput chip platforms: faster, better, smaller
    - Computers and storage medium in general
  - But cost of people continues to rise
    - Lab processing
    - Data management and analytics

how many computers do I need?
(omic example)

- tremendous computing resource needs
  - Multiple analyses, periodically redone
    - Algorithms improve
    - Gene annotation and sequence data evolve
  - Verification of properties of methods
    - Theory gives easy cutoff values (LOD > 3) that may not be relevant
    - Need to carefully develop re-sampling methods (permutations, etc.)
  - Storage of raw, processed and summary data (and metadata)
    - Terabyte(s) of backed-up storage (soon petabytes and more)
    - Web access tools
- high throughput computing platforms (Condor)
  - Reduce months or years to hours or days
  - Free up your mind to think about science rather than mechanics
  - Free up your desktop/laptop for more immediate tasks
  - Need local (regional) infrastructure
    - Who maintains the machines, algorithms?
    - Who can talk to you in plain language?

What do we need?

inference methods for data structures

Computer Science has historically been strong on data structures and …

Statistics has historically been … strong on inference from data.

One way to draw on the strengths of both disciplines is to pursue the study of 'inferential methods for data structures';

i.e., methods that update probability distributions on recursively-defined objects such as trees, graphs, grammars and function calls.”

Michael Jordan, UC-Berkeley (2010 UW lecture)

The translation gap between structure and inference

- Many tools emerging for data structure
  - Genomic, geospatial, ...
  - GMOD.org, .NET bio, … platforms
- Basic production inference being added
  - T-tests with FDR, enrichment
  - Glue to bind resources (GenomeSpace)
    - UCSC genome browser, Cytoscape, Galaxy, …
- But state-of-the-art collaboration tools lag
  - Translate one-off code to pipeline
  - Build, maintain, enhance new workflows

What works and what doesn’t at UW?
(the people dimension)

- What have been successful?
  - Cancer Informatics Shared Resource (30+ years)
    - Biostat & Med Info with Comp Cancer Ctr
  - other BMI collaborative research across campus (30+ years)
  - Biometry Program (30+ years)
    - Stat with CALS and later L&S (Bot,Zoo), VetMed (off and on)
  - Tech Partners (25 years)
  - Geospatial Alliance (25 years)
  - CIBM, GSTP, Biophysics training grants (10 years)
    - CS, Math, BMI with multiple collaborators
- What is missing?
  - link from Gene Expression Center to data analytics
  - “free” quantitative consulting across campus
    - experimental design, data analysis
    - informatics, workflows/pipelines

Who should be involved at UW-Madison?

- Chief Information Officers (CIOs) in all organizations
- Librarians
  - Academic and general library system
- Statisticians and biostatisticians
  - Develop methods for design and analysis
- Computer scientists
  - Design and build computers, databases, analytics
- Social scientists
- Other data analytics fields
  - Departments: Stat, BMI, CS, ECE, ISyE, SLIS, BusInfo
  - Informatics experts in general
- Subject matter scientists
  - Omics, spatial, networks, languages
- Both faculty and staff
  - Build communication to foster ideas, collaboration

Who specifically at UW-Madison?

- Stat/BMI: Brian Yandell, Zhiguang Qian, Mark Craven
- CS: Myron Livny (CHTC), Michael Gleicher, Michael Ferris (ISyE,WID)
- Libraries: Dorothea Salo (RDS,SLIS), Lee Konrad (GLS)
- DoIT: Jan Cheetham, Alan Wolf
- CIOs: Phil Barak (Soils/CALS), Umberto Tachinardi (SMPH)
- Discipline scientists
  - Sandra Splinter BonDurant (Gene Expression Ctr)
  - George Phillips (CIBM, Biochem&CS)
  - Juan de Pablo (BiolChemEngr)
  - Edgar Spalding (Botany)
  - Howard Veregin (State Cartographer, Geospatial Alliance)
  - Corinna Gries (LTER, Limnology)
  - Tom Mish (BCG/SMPH)
  - Nancy McDermott (SSCC)
- Ex-officio: Bruce Maas (UW CIO), Katrina Forest (ITC)

Data Science at UW-Madison who thinks about data for its own sake?

- Academic Programs
  - Statistics Department, L&S
  - Biostatistics & Medical Informatics Department, SMPH
  - Computer Science Department, L&S
  - Electrical & Computer Engineering Department, CoE
  - Industrial & Systems Engineering Department, CoE
  - Operations & Information Management Department, Business School
  - Library and Information Studies Department, SLIS
  - Biometry Program, CALS
  - Mathematics Department, L&S
- Research Groups
  - Cancer Informatics Shared Resource, SMPH
  - Computing & Biometry, CALS
  - Geospatial Alliance (formerly SIAC)
  - Gene Expression Center, Genome Center
  - Social Sciences (Computer Center (SSCC, formerly DACC, DPLS)
- Administrative Groups
  - General Library System (GLS)
  - Biomedical Computing Group (BCG), SMPH
  - Division of Information Technology (DoIT)
  - Research Data Services (RDS)
  - Information Technology Committee (ITC)
  - Wisconsin Institutes of Discovery (WID/MIR)