Supporting Large Scale Scientific and Engineering Applications Using Database Technology

José A. Blakeley
Software Architect
Database Systems Group
Microsoft Corporation

HPTS 2009, October 28, 2009

Setting Expectations

- Most computational science involving “big data” is still managed by file systems
- Programming is mainly procedural using scripting languages
- There is little data sharing independent of the programs – no data independence
- MapReduce approaches perceived as effective an easier to use

- Thus, database technology has a high-hill to climb to become the de-facto platform for engineering and scientific apps

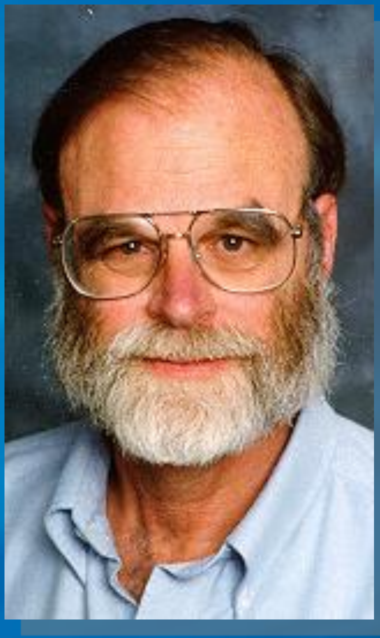
Some perceived problems

- Databases are hard to use
- Database are not sufficiently scalable
- Tools of the database field such as data modeling not widely understood by scientists
- Fear of vendor “lock-in”
- Hard to break away from established practices in a science
- ….

Our Experience

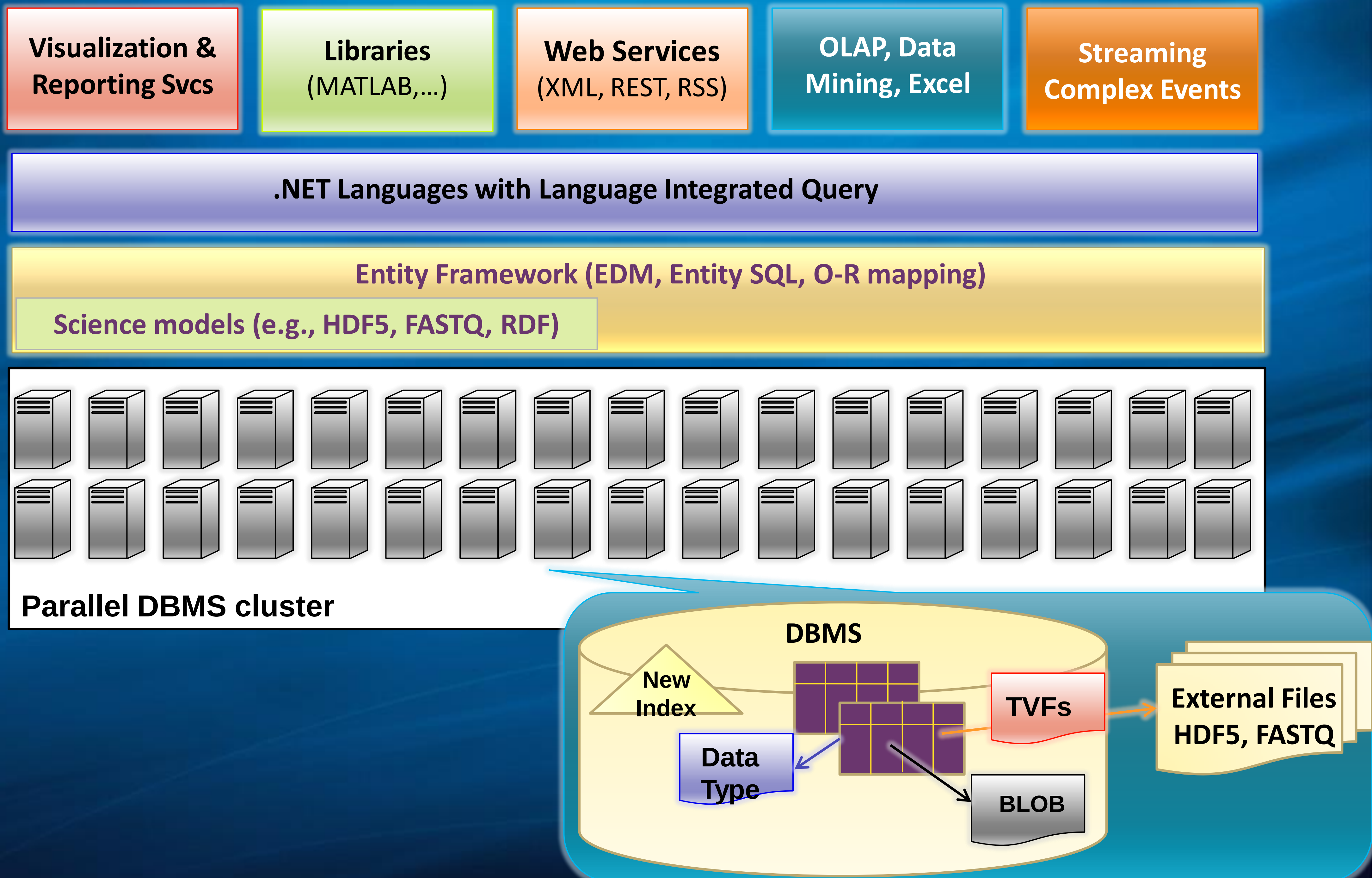
- Microsoft has been learning about scientific and engineering apps for the last 10 years
- Have had good successes working side-by-side with scientists
  - Astronomy, genomics, comparative analysis for RNA sequences, protein folding, carbon climate analysis
- Every success has been transformational to the way science is performed in each science
- We believe there is a great opportunity for databases to significantly impact the sciences

Gray's Laws for Data Engineering



- Large-scale scientific computing is data intensive
  - Database management systems can help
- The solution is in "scale-out" architectures
  - Both functional and data scale-out
- Move analysis to the data!
  - Increasingly true with larger data set sizes
- Start the design with "20 queries"
  - Engages domain and computer scientists in data modeling design on the most important queries
- Go from "working to working"
  - "Don't let the best get on the way of the better"
  - Iterative improvement works

A Data Platform for Science



Some Case Studies

- Astronomy: SkyServer, Pan-STARRS
- Global scale carbon flux – FLUXNET
- Predictive medicine – Clalit Health Services
- High-throughput genomics – 1000 genome
  - Massive sequence alignment (UT Austin)
  - Microsoft Life Sciences
- Integrated Comparative Analysis System for RNA Sequences – The Gutell’s Lab @ UTA
- … many more

Case study

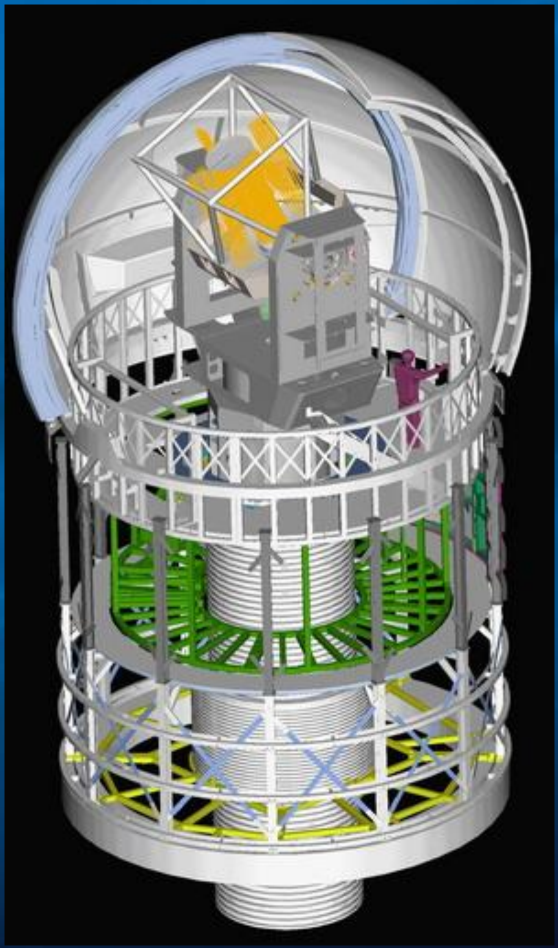
Sky-Server, PAN-STARRS

Distributed, scale-out database system, moving analysis to the data

Dr. Alex Szalay, JHU

PAN-STARRS

- Sky survey to detect ‘killer asteroids’
- Two phases
  - PS1: single telescope prototype now
  - PS4: 4 telescope array in 4 years
- Hawaii + JHU + Harvard + Edinburgh + Max Planck Society
- High data rate: 2.5 Petabytes/year
- 5B celestial objects/250B detections
- 100TB prototype database built at JHU with Microsoft help



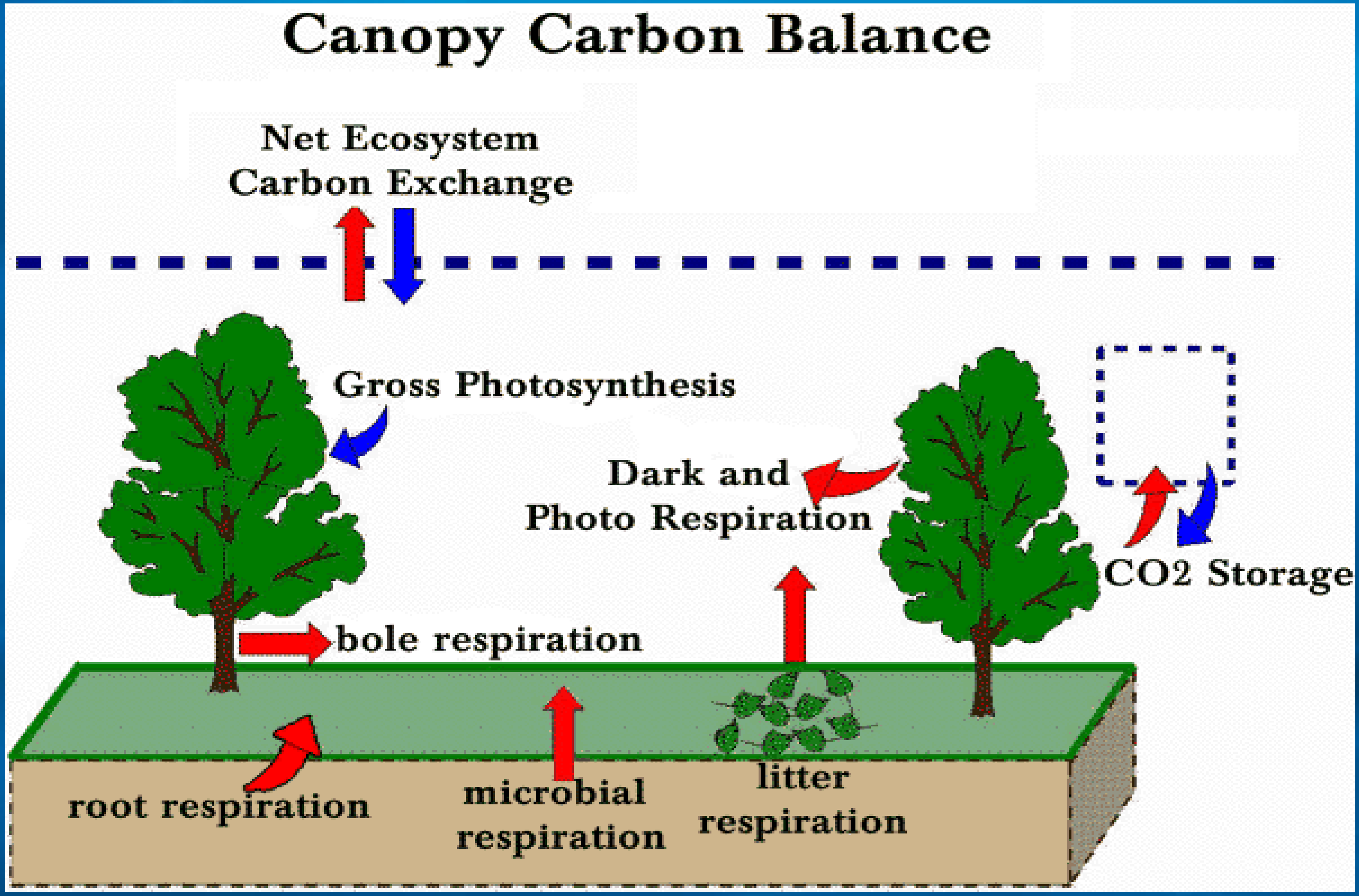
Case study

Global Scale Carbon Flux Research @ Berkeley Water Center

Leveraging Reporting and Data Analysis to increase quality of data

Dr. Deborah Agarwal and BWC tech team

Carbon Climate Analysis



$$\mathbf{F_c + F_{storage} = -NEE = P_{new} + R_{leaf} + R_{wood} + R_{roots} + R_{microbes}}$$

Applications of eddy covariance measurements, Part 1: Lecture on Analyzing and Interpreting CO2 Flux Measurements, Dennis Baldocchi, CarboEurope Summer Course, 2006, Namur, Belgium (http://nature.berkeley.edu/biometlab/lectures/)

Carbon-Climate Analysis Goals

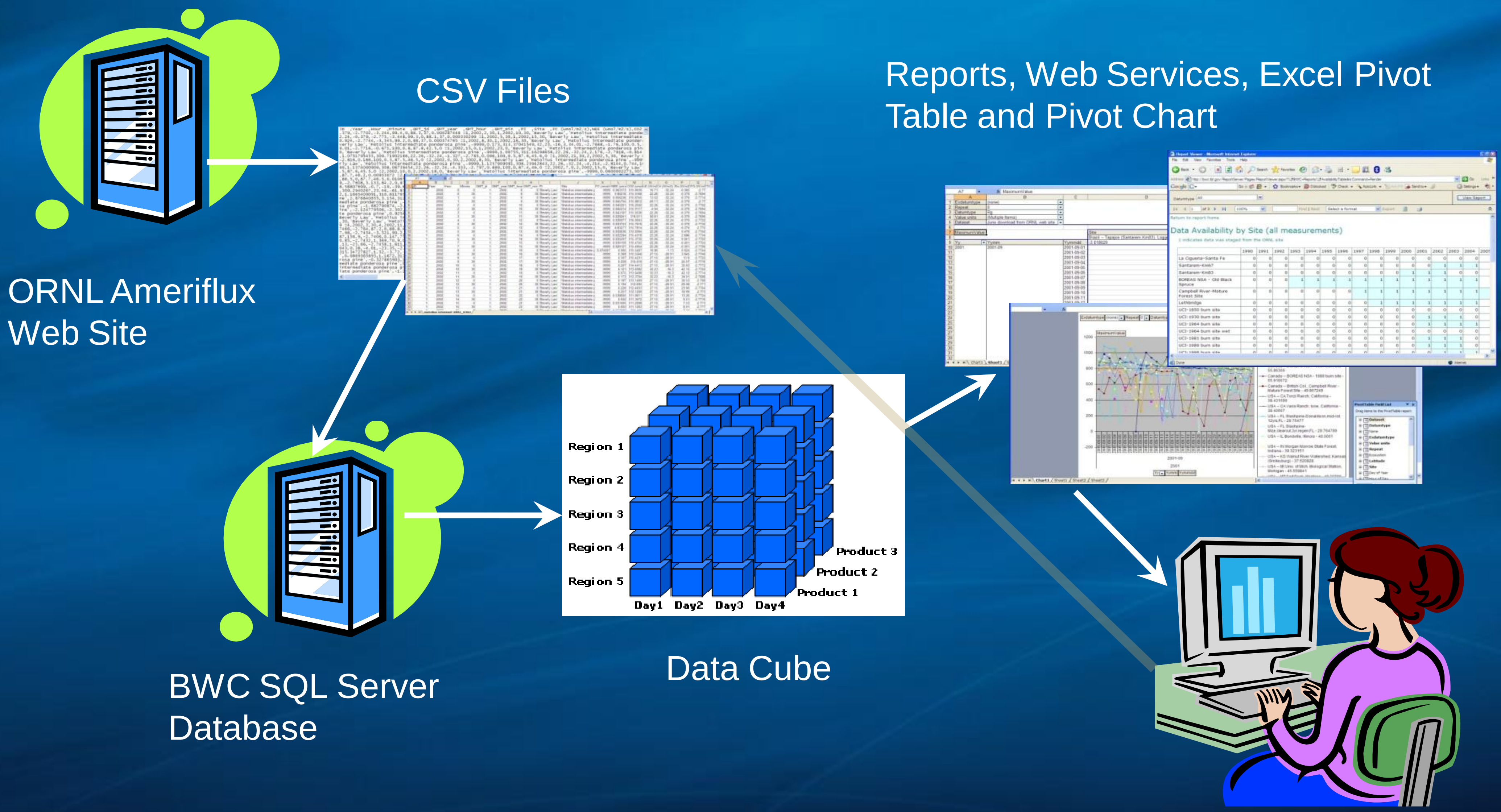
Get a handle on data collection

- Towers measure consistent carbon flux and micrometeorological parameters
- Tower researchers quality check data and then provide the data to regional archives.
- Regional and global carbon-climate analysis activities rely on data from regional archives
- Recent La Thuile workshop is gathering over 900 site-years of data available from over 200 sites around the world.

Measurements Are Often Not Simple or Complete

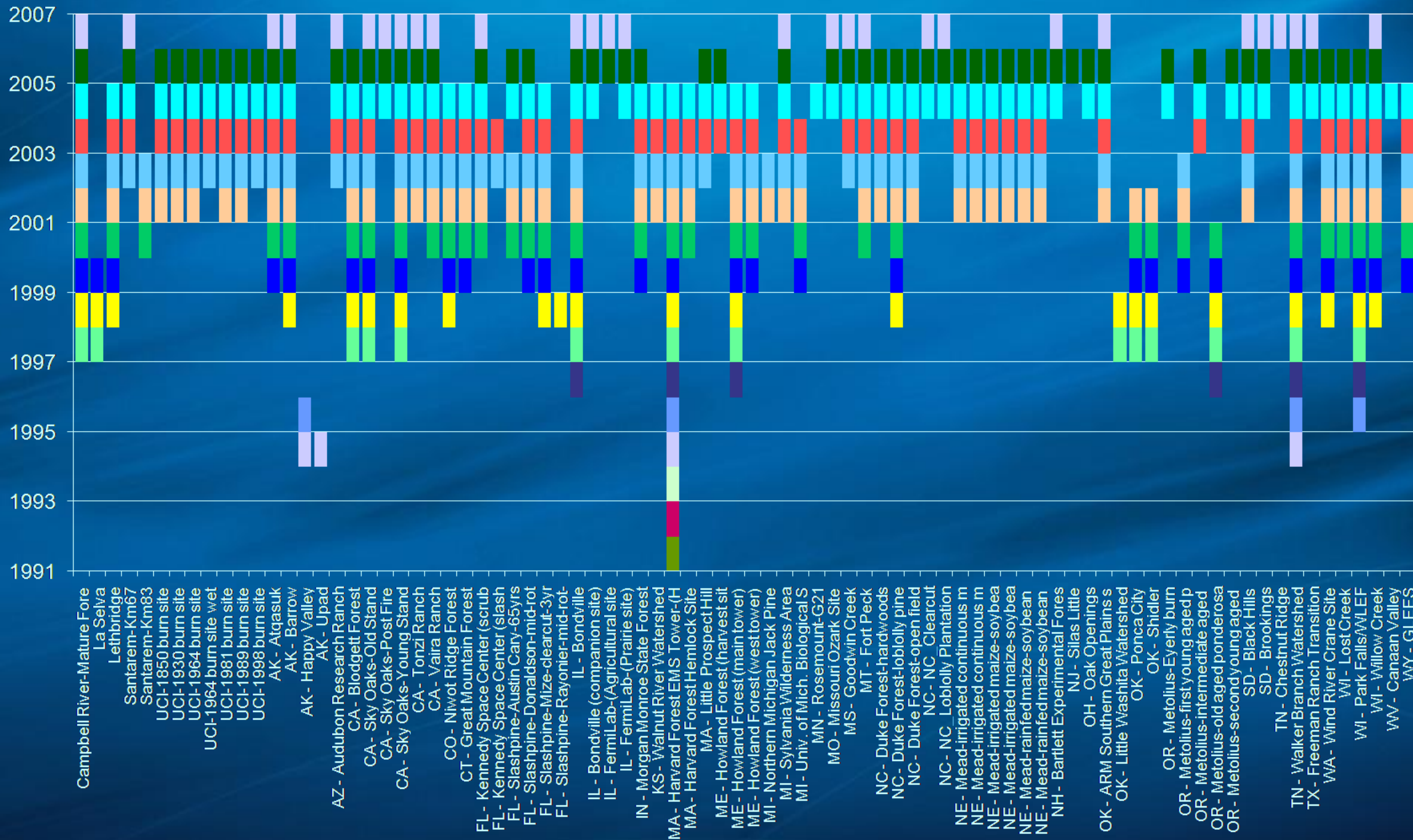
- Gaps in the data
  - E.g., quiet nights, bird poop, high winds
- Discrepancies in units of measure
- Difficult to make measurements
  - Leaf area index
  - Wood respiration
  - Soil respiration
- Localized measurements – tower footprint
- Local investigator knowledge important
- PIs' science goals are not uniform across the towers

Scientific Data Server – User Interface



Visualizing Data Availability

Ameriflux Sites Reporting Data Colored by Year



Required Variable Reporting by Site by Year

- Each row corresponds to one site-year
- Each cell corresponds to one site year of (FC, CO_2 or SCO_2, UST, PAR or Rg, TA, and Rh or H_2O).
- Color indicates:
  - Red – likely not enough for processing - % < .3 reported (roughly less than 5K of 17.5K)
  - Green – likely enough for processing .3<%<.999
  - Yellow – may not be good for processing due to gap-filling - % > .999
- Red CO2 (second column) can be ignored for cropland/grassland sites
- Sites shown are just a sample

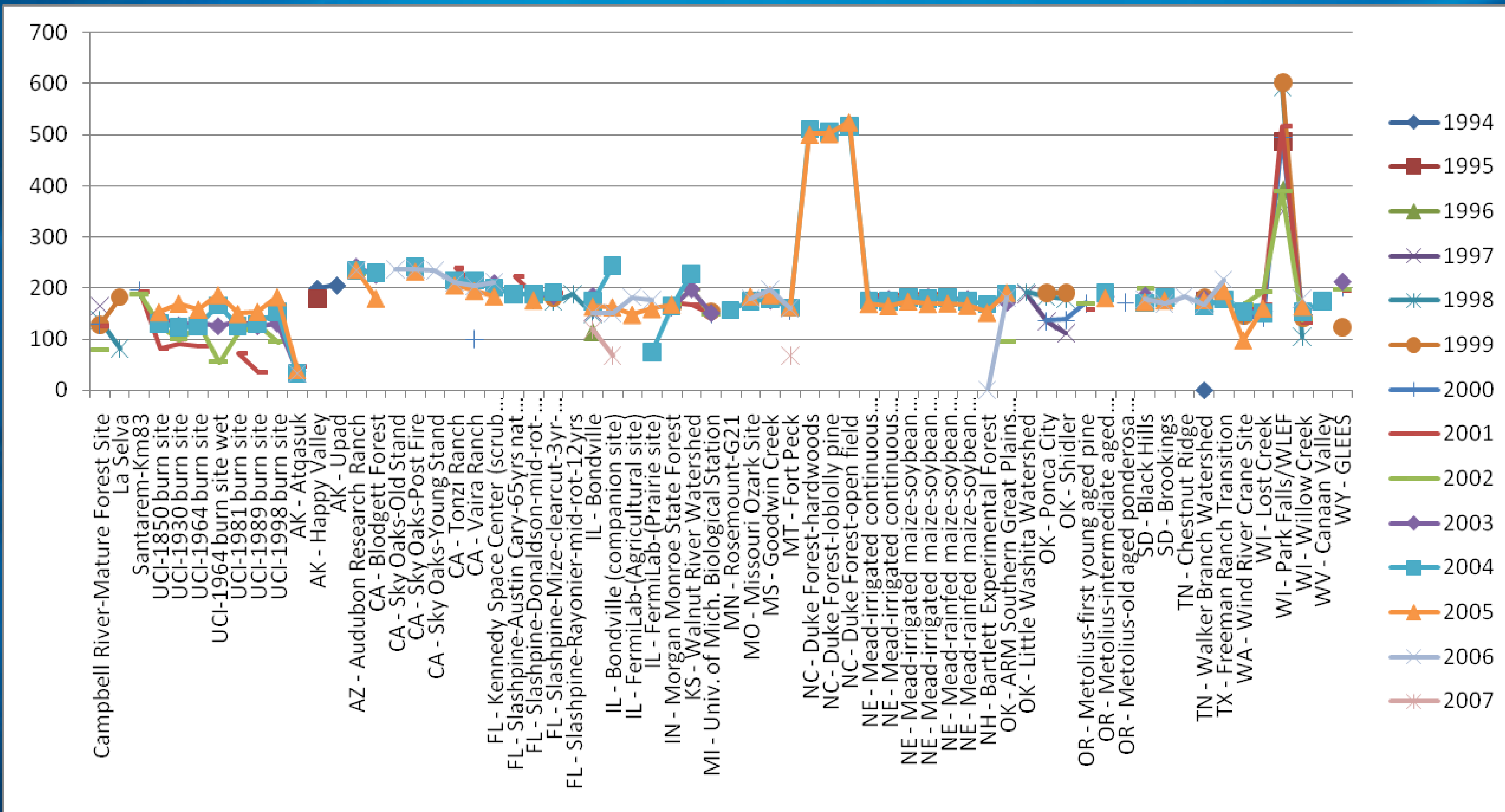
Data Availabilty	Low Threshold	0.3000
	High Threshold	0.9990

	FC	CO2/SCO2	UST	PAR/RG	TA	RH/H2O
CA - Blodgett Forest						
1997	0.0000	0.2241	0.4149	0.4937	0.2443	0.4150
1998	0.0000	0.2913	0.5228	0.6013	0.4150	0.6123
1999	0.5188	0.5124	0.8202	0.9888	0.6123	0.9207
2000	0.8081	0.8792	0.8504	0.9639	0.9207	0.9639
2001	0.8397	0.9112	0.8638	0.9845	0.9639	0.9846
2002	0.8038	0.9247	0.8523	0.9620	0.9845	0.9259
2003	0.8389	0.9116	0.8327	0.9605	0.9246	0.7778
2004	0.7437	0.8284	0.8412	0.9697	0.7778	0.8087
2005	0.6702	0.6151	0.0000	0.0000	0.8307	0.0000
CA - Tonzi Ranch						
2001	0.5132	0.5616	0.8064	0.9999	0.7080	0.9172
2002	0.7325	0.8443	0.8754	1.0000	0.9983	1.0000
2003	0.7857	0.8962	1.0000	1.0000	1.0000	1.0000
2004	1.0000	1.0000	0.8534	0.9991	1.0000	0.9999
2005	0.7490	0.8541	0.9786	0.9957	0.9999	0.9984
2006	0.7712	0.9526	0.0000	0.0000	0.9984	0.0000
CO - Niwot Ridge Forest						
1998	0.1671	0.1671	1.0000	0.9818	0.0744	0.8447
1999	1.0000	1.0000	0.9995	0.9989	0.8725	0.9831
2000	1.0000	1.0000	1.0000	0.9958	0.9926	0.9538
2001	1.0000	1.0000	0.9994	0.9890	0.9866	0.6631
2002	1.0000	0.9999	0.9822	1.0000	0.9974	0.7888
2003	1.0000	1.0000	1.0000	1.0000	0.9977	1.0000
2004	1.0000	1.0000	0.0000	0.0000	1.0000	0.0000
IL - Bondville						
1996	0.2122	0.0000	0.8388	1.0000	0.3510	1.0000
1997	0.7856	0.0000	0.9127	0.9999	1.0000	1.0000
1998	0.7289	0.0000	0.9307	1.0000	1.0000	1.0000
1999	0.8182	0.0000	0.7981	1.0000	1.0000	1.0000
2000	0.7158	0.0000	0.8720	1.0000	1.0000	1.0000
2001	0.6979	0.0000	0.8740	0.9973	1.0000	0.9973
2002	0.7216	0.0000	0.8973	0.9839	0.9973	0.9839
2003	0.7765	0.0000	0.7486	0.9356	0.9839	0.9357
2004	0.5894	0.0000	0.8522	0.9608	0.9349	0.9608
2005	0.7518	0.0000	0.7910	0.8997	0.9608	0.8997
2006	0.7748	0.0000	0.2183	0.2301	0.8997	0.2301
2007	0.2179	0.0000	0.0000	0.0000	0.2301	0.0000
NC - Duke Forest-hardwoods						
2001	1.0000	0.8523	1.0000	1.0000	1.0000	1.0000
2002	1.0000	0.9496	1.0000	1.0000	1.0000	1.0000
2003	1.0000	0.9475	1.0000	0.9402	1.0000	1.0000
2004	1.0000	0.9910	1.0000	0.9997	1.0000	1.0000
2005	1.0000	0.9792	0.0000	0.0000	1.0000	0.0000
OK - Shidler						
1997	0.2959	0.2685	0.9999	1.0000	0.2959	0.9999
1998	1.0000	0.8656	1.0000	1.0000	1.0000	0.9914
1999	1.0000	0.9660	0.2595	0.2595	1.0000	0.2563
2000	0.2595	0.2553	0.0000	0.0000	0.2595	0.0000

	FC	CO2/SCO2	UST	PAR/RG	TA	RH/H2O
OR - Metolius-intermediate aged ponderosa pine						
2003	0.4741	0.6901	0.8336	0.9493	1.0000	0.9493
2004	0.7551	0.8931	0.8678	0.9934	0.9493	0.9934
2005	0.7470	0.9664	0.0000	0.0000	0.9934	0.0000
SD - Black Hills						
2001	0.2196	0.0000	0.3359	0.3184	0.2321	0.2661
2002	0.0000	0.0000	0.4695	0.5733	0.3179	0.5733
2003	0.3329	0.0000	0.8141	0.8487	0.5733	0.8488
2004	0.6300	0.0000	0.8929	0.9305	0.8487	0.9305
2005	0.8682	0.0000	0.9522	0.9683	0.9305	0.9683
2006	0.9233	0.0000	0.0000	0.0000	0.9683	0.0000
SD - Brookings						
2004	0.4628	0.0000	0.8267	0.9918	0.6897	0.8764
2005	0.7296	0.0000	0.9072	0.9970	0.9918	0.9970
2006	0.8034	0.0000	0.0000	0.0000	0.9973	0.0000
TN - Walker Branch Watershed						
1994	0.0001	0.0000	0.7726	1.0000	0.0001	0.0000
1995	0.5992	0.9693	0.8237	0.9999	0.9972	0.0000
1996	0.6143	0.9423	0.7946	0.9999	1.0000	0.0000
1997	0.5758	0.9598	0.8038	0.9890	0.9999	0.0000
1998	0.5614	0.7850	0.9624	1.0000	0.9999	1.0000
1999	0.9052	0.9408	0.9038	0.9973	1.0000	0.9973
2000	0.8883	0.0000	0.8591	1.0000	0.9973	1.0000
2001	0.8421	0.0000	0.0665	0.0745	1.0000	0.0745
2002	0.0643	0.0000	0.7025	0.7956	0.0848	0.7977
2003	0.6162	0.0000	0.8028	0.8870	0.8326	0.9754
2004	0.7149	0.0000	0.6937	0.5789	0.8870	0.8342
2005	0.4349	0.0000	0.6907	0.8038	0.5791	0.8038
2006	0.6642	0.0000	0.0000	0.0000	0.8038	0.0000
WA - Wind River Crane Site						
1998	0.4696	0.4665	0.8691	0.8253	0.7369	0.8639
1999	0.8105	0.8529	0.7735	0.9448	0.8092	0.9814
2000	0.6081	0.6532	0.8588	0.9400	0.9814	1.0000
2001	0.8578	0.8276	0.9346	0.9163	1.0000	1.0000
2002	0.9332	0.9376	0.9518	0.9995	1.0000	1.0000
2003	0.8640	0.8848	0.9049	1.0000	1.0000	1.0000
2004	0.8796	0.8838	0.2522	0.2736	1.0000	0.2736
2005	0.2519	0.2524	0.0000	0.0000	0.2736	0.0000
WI - Park Falls/WLEF						
1995	0.0000	0.3983	0.1389	0.3899	0.2476	0.2383
1996	0.0000	0.4060	0.3042	0.3994	0.2160	0.4792
1997	0.0000	0.4603	0.2383	0.4253	0.2479	0.4349
1998	0.0000	0.4619	0.3565	0.4643	0.4349	0.3647
1999	0.0000	0.4017	0.1054	0.4912	0.3647	0.3218
2000	0.0000	0.4055	0.1388	0.4122	0.3218	0.3531
2001	0.0000	0.4317	0.0000	0.3329	0.3531	0.2796
2002	0.0000	0.4182	0.2207	0.4813	0.2796	0.4355
2003	0.0000	0.3248	0.2359	0.4570	0.4355	0.4601

Of the 285 site years with good FC, 50 site years are missing one of (UST, PAR/Rg, and TA) and 79 sites have likely gap-filled data.

Obviously bad annual averages



Data cube used to browse average yearly Rg values across all site-years

16 additional likely problematic site-years at 5 sites

Case study

Clalit Health Services

Business intelligence moves medicine from reactive to proactive

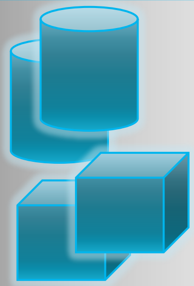
Clalit Health Services

- Largest provider of medical care in Israel
  - 3.7 million patients
  - 14 hospitals
  - 1400 clinics
- Needed to identify which members would most benefit from proactive intervention to prevent health deterioration
- Developed an integrated 1.5 TB relational database plus data mining service

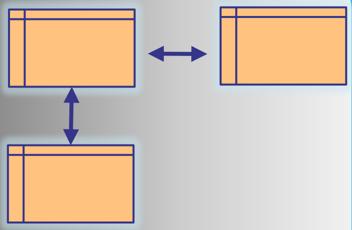
Solution

A shift from reactive to proactive medicine

Clinical Data



Data Warehouse



DM Prediction Model

Output: Deterioration Prediction Model

Medical Evaluation

Treatment Protocols

Intervention Programs

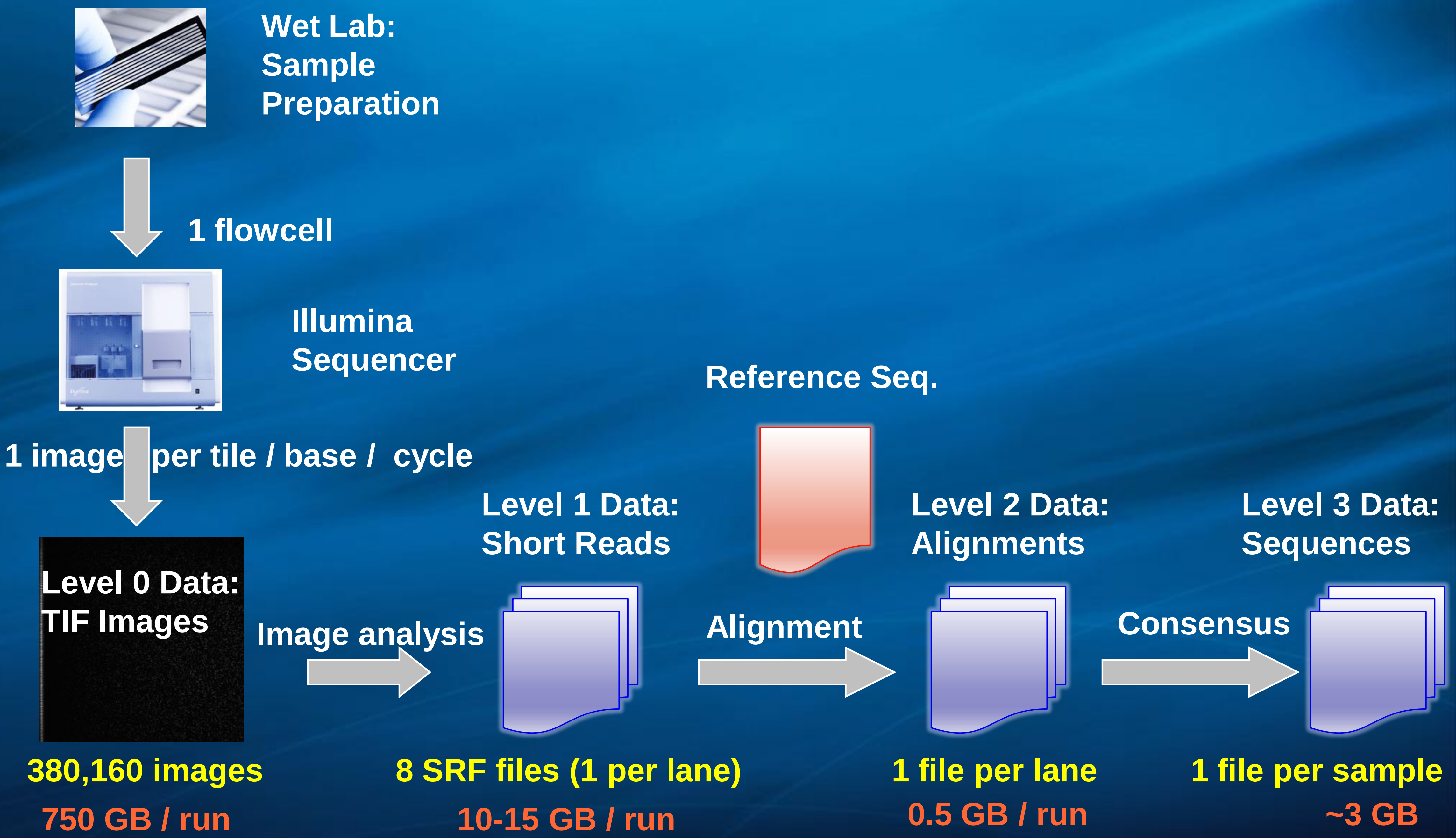
- **Preserves and enhances patients quality of life**
- **Reduced health care costs**

Case study

High-Throughput Genomics

Data Modeling, In-situ Data Management, Aggregation in database

1000 Genome Pipeline

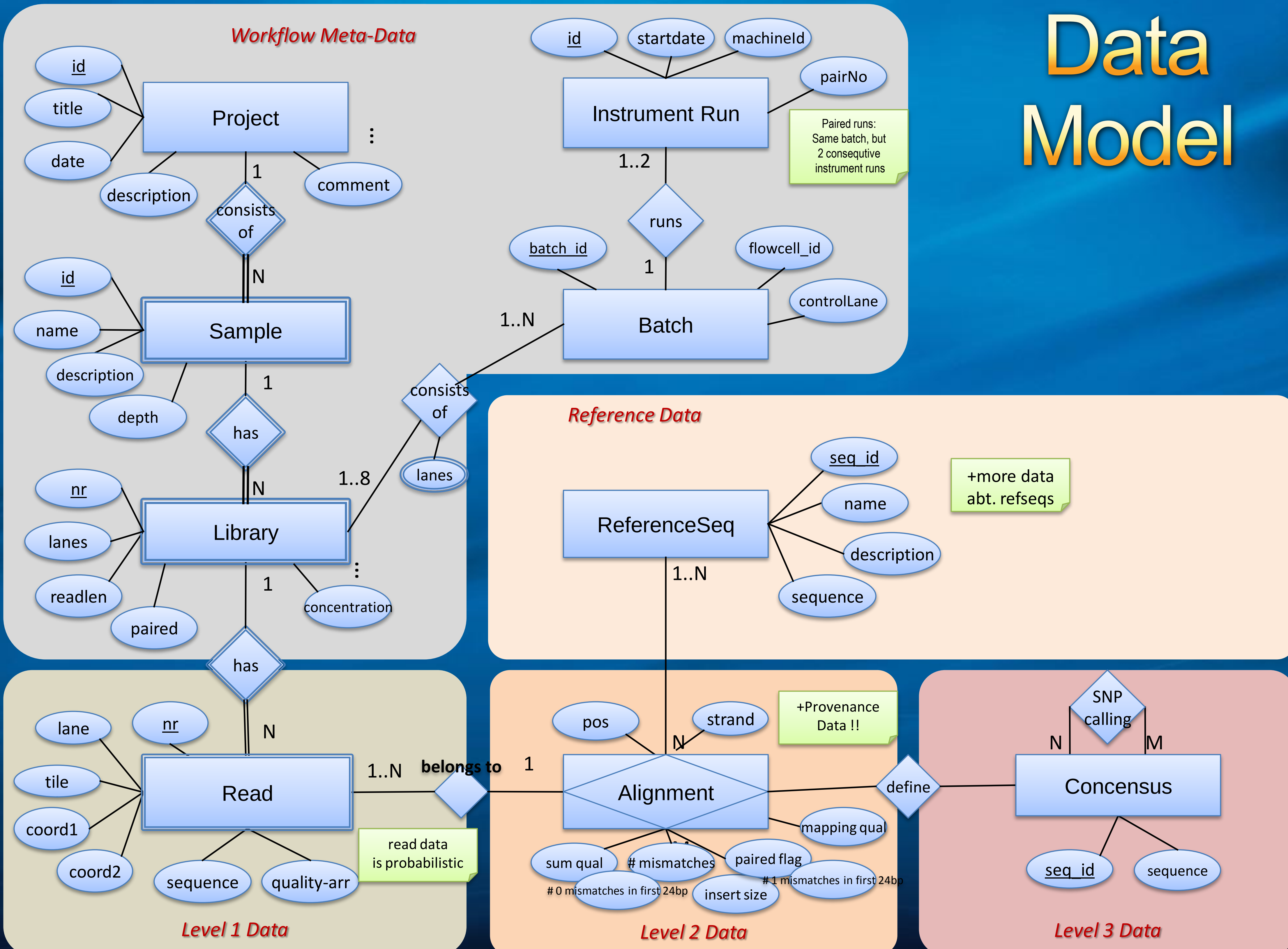


About the Scale of the Problem

- Sanger Institute has currently 28 instruments
- 24 x 7 (in avg 20 in use at any given time)
- => per week:
  ~75 TB Level 0 data (images)
  0.5 TB Level 1 data (short-reads)

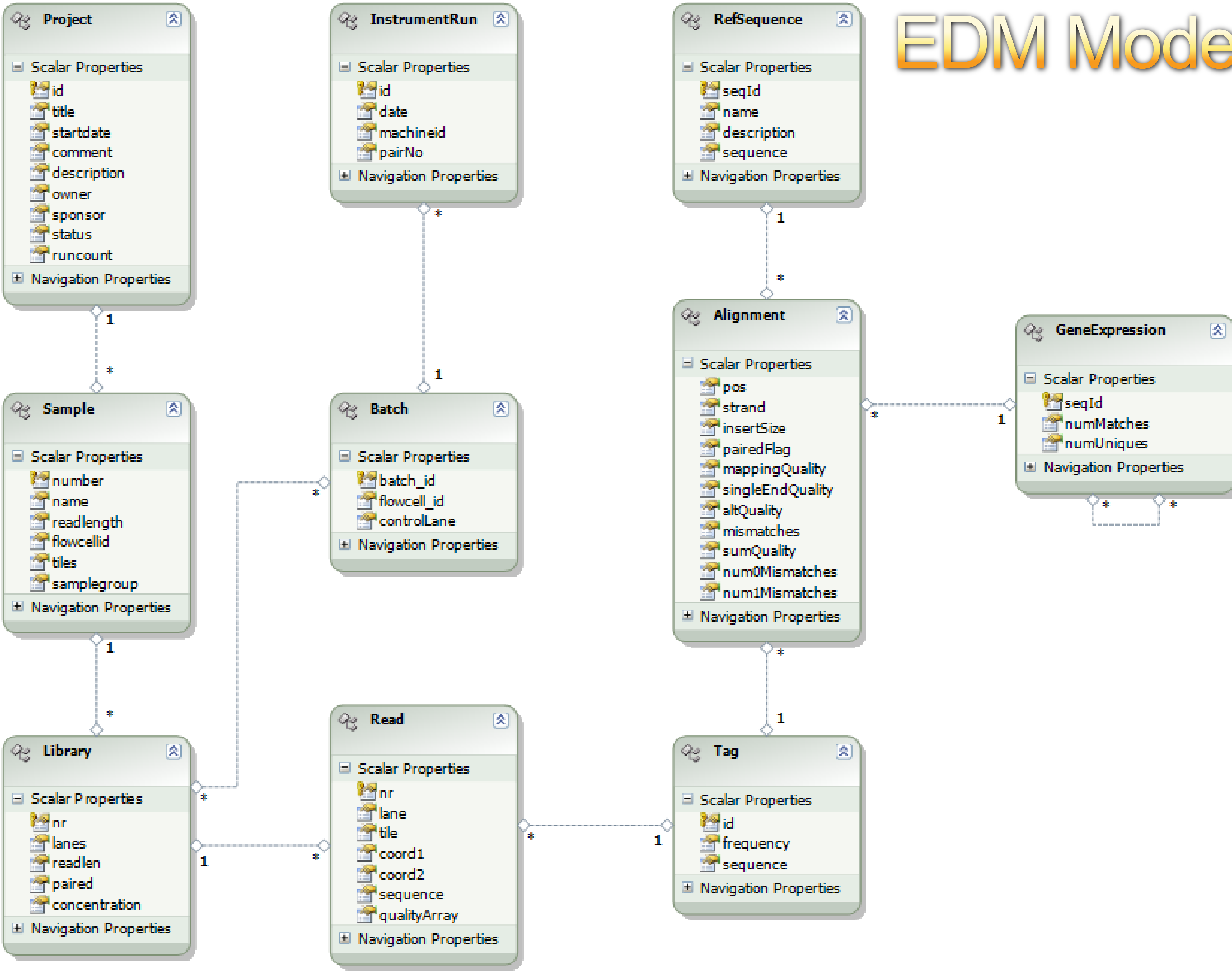
- Plan for another 10 Solexas by end of year
- Only one of 3 labs worldwide
- Technology constantly improving

Data Model



Workflow Meta-Data
Project
id
title
date
description
comment
1
consists of
N
Sample
id
name
description
depth
1
has
N
Library
nr
lanes
readlen
paired
concentration
1
has
N
Instrument Run
id
startdate
machineId
pairNo
Paired runs: Same batch, but 2 consequtive instrument runs
1..2
runs
1
Batch
batch_id
flowcell_id
controlLane
1..N
consists of
lanes
1..8
Reference Data
ReferenceSeq
seq_id
name
description
sequence
+more data abt. refseqs
1..N
Read
lane
nr
tile
coord1
coord2
sequence
quality-arr
read data is probabilistic
1..N
belongs to
1
Alignment
pos
strand
N
+Provenance Data !!
mapping qual
sum qual
mismatches
paired flag
insert size
0 mismatches in first 24bp
1 mismatches in first 24bp
define
Concensus
SNP calling
N
M
seq_id
sequence
Level 1 Data
Level 2 Data
Level 3 Data

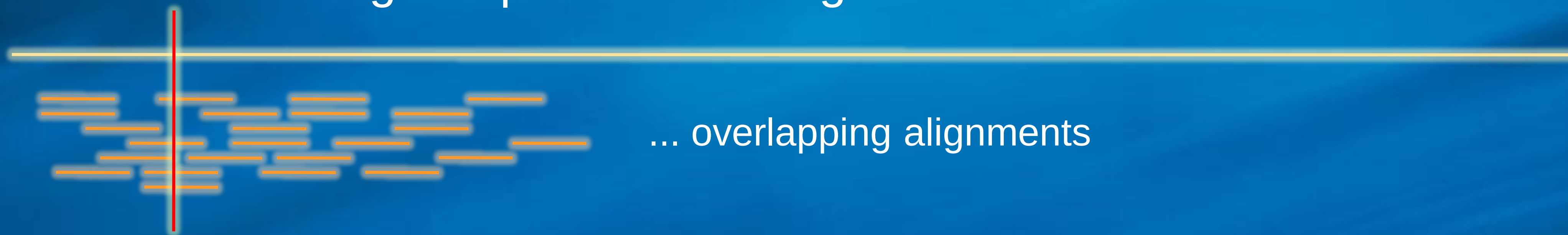
EDM Model



Project
Scalar Properties
id
title
startdate
comment
description
owner
sponsor
status
runcount
Navigation Properties
InstrumentRun
Scalar Properties
id
date
machineid
pairNo
Navigation Properties
RefSequence
Scalar Properties
seqId
name
description
sequence
Navigation Properties
Sample
Scalar Properties
number
name
readlength
flowcellid
tiles
samplegroup
Navigation Properties
Batch
Scalar Properties
batch_id
flowcell_id
controlLane
Navigation Properties
Alignment
Scalar Properties
pos
strand
insertSize
pairedFlag
mappingQuality
singleEndQuality
altQuality
mismatches
sumQuality
num0Mismatches
num1Mismatches
Navigation Properties
GeneExpression
Scalar Properties
seqId
numMatches
numUniques
Navigation Properties
Library
Scalar Properties
nr
lanes
readlen
paired
concentration
Navigation Properties
Read
Scalar Properties
nr
lane
tile
coord1
coord2
sequence
qualityArray
Navigation Properties
Tag
Scalar Properties
id
frequency
sequence
Navigation Properties
1
*

Consensus Calling

Consensus at a given position of the genome?



... overlapping alignments

- 40x coverage of an individual genome: 120B bp
- Each has an alignment position, a base and a quality (error probability)
  - Consensus: aggregation function per position
    - Take base with largest support (qualities currently not used?
  - Seq = Concat (all consensi per position)

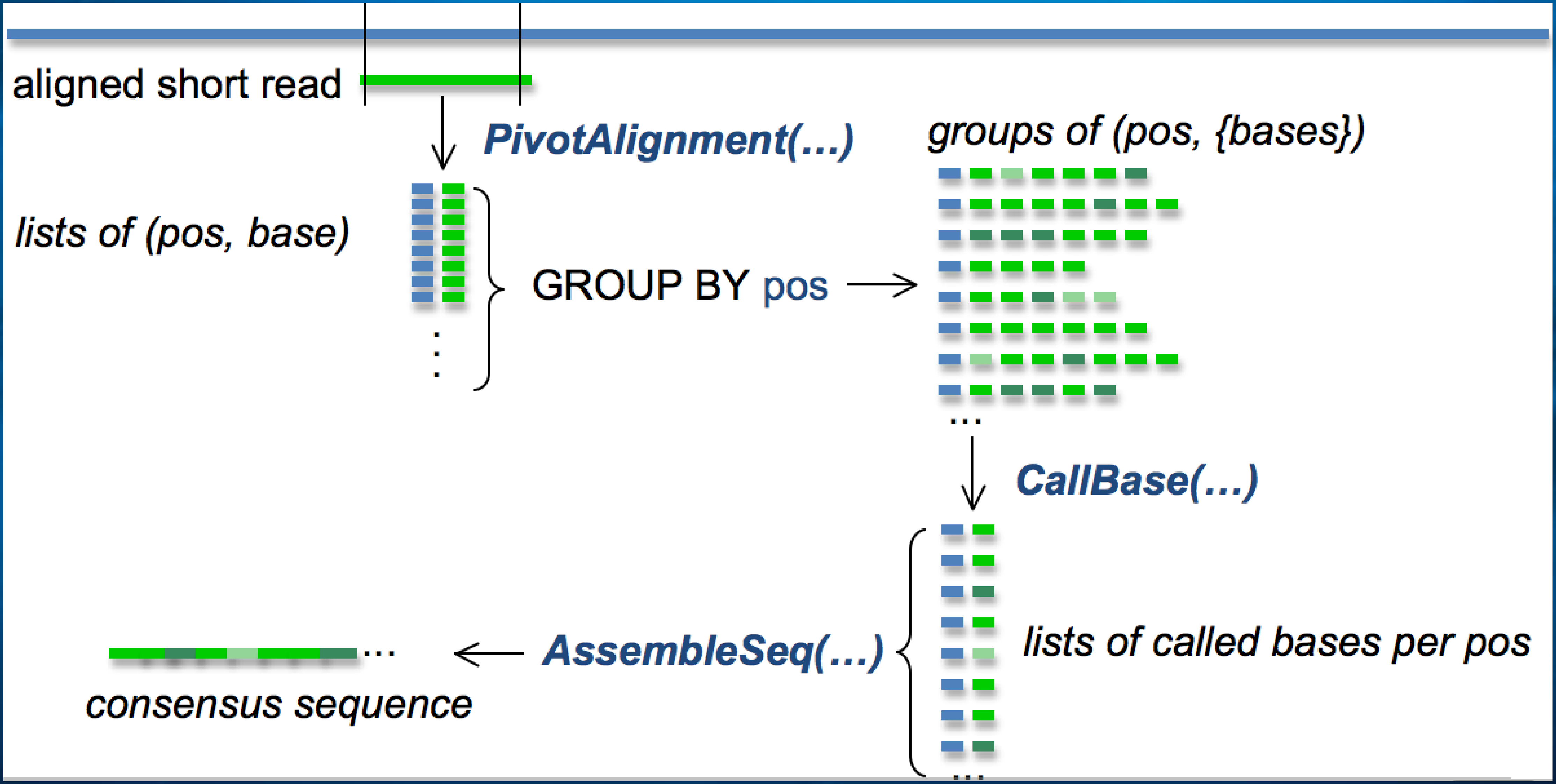
Aggregation to the Extreme

- Imagine, we could assemble a whole genome inside the database:

```
SELECT chromosome, AssembleSequence(position,base)
  FROM ( SELECT chromosome,position, CallBase(base,qual)
             FROM Alignments CROSS APPLY
                  PivotAlignment(position,strand,seq,quals)
            WHERE a_e_id=…
            GROUP BY chromosome, position )
 GROUP BY chromosome
```

- Shows benefits of SQL-CLR integration

Consensus calling



Consensus calling

```
SELECT chromosome,position, CallBase(base,qual)
  FROM Alignments CROSS APPLY
       PivotAlignment(position,strand,seq,quals)
 WHERE a_e_id=…
 GROUP BY chromosome, position
```

- PivotAlignment(…)
  - Table-valued function that pivots a short-read (with quality values) into table of the form:
    (position, base, quality)
- CallBase(b,q)
  - aggregate function that decides which base is the consensus among all alignments on a pos.

Lessons Learned

- Data modeling is key
  - Great way for learn the vocabulary of the science
  - Enables formulation of the “20 queries”
  - Separates semantics from representation
  - Enables provenance, time varying
- Extended relational DBMS solve a large portion of the problem
  - Structured, semi-structured, and “in-situ” file data
  - Powerful analysis tools (UDFs, UDTs, UDAggs)
  - Automatic parallelism
  - In-database map-reduce
- Data services important
  - Streaming, Reporting, OLAP, Data mining
  - Semantic modeling and mapping
  - Integration with scientists tools (Matlab, Lapack, R)

Microsoft®

Your potential. Our passion.™

© 2007 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.