A BRIEF HISTORY OF THE SSDBM CONFERENCE SERIES

30TH ANNIVERSARY

Arie Shoshani

Lawrence Berkeley National Laboratory

SSDBM conference
July 9-11, 2018

Outline

- How did this conference series start
- Research topics evolution over time
- Future challenges
- Light-hearted anecdotes
- Next conference – Santa Cruz, California

30 SSDBM conferences over 37 years

PREVIOUS CONFERENCES

2018, Bozen-Bolzano, Italy
2017, Chicago, Illinois
2016, Budapest, Hungary
2015, San Diego, California
2014, Denmark
2013, Baltimore
2012, Crete, Greece
2011, Portland, Oregon
2010, Heidelberg, Germany
2009, New Orleans
2008, Hong Kong
2007, Banff, Canada
2006, Vienna, Austria
2005, Santa Barbara, California
2004, Santorini, Greece
2003, Cambridge, Massachusetts
2002, Edinburgh, Scotland
2001, Fairfax, Virginia
2000, Berlin, Germany
1999, Cleveland, Ohio
1998, Capri, Italy
1997, Olympia, Washington
1996, Stockholm, Sweden
1994, Charlottesville, Virginia
1992, Ascona, Switzerland
1990, Charlotte, North Carolina
1988, Rome, Italy
1986, Luxembourg
1983, Los Altos, California
1981, Menlo Park, California

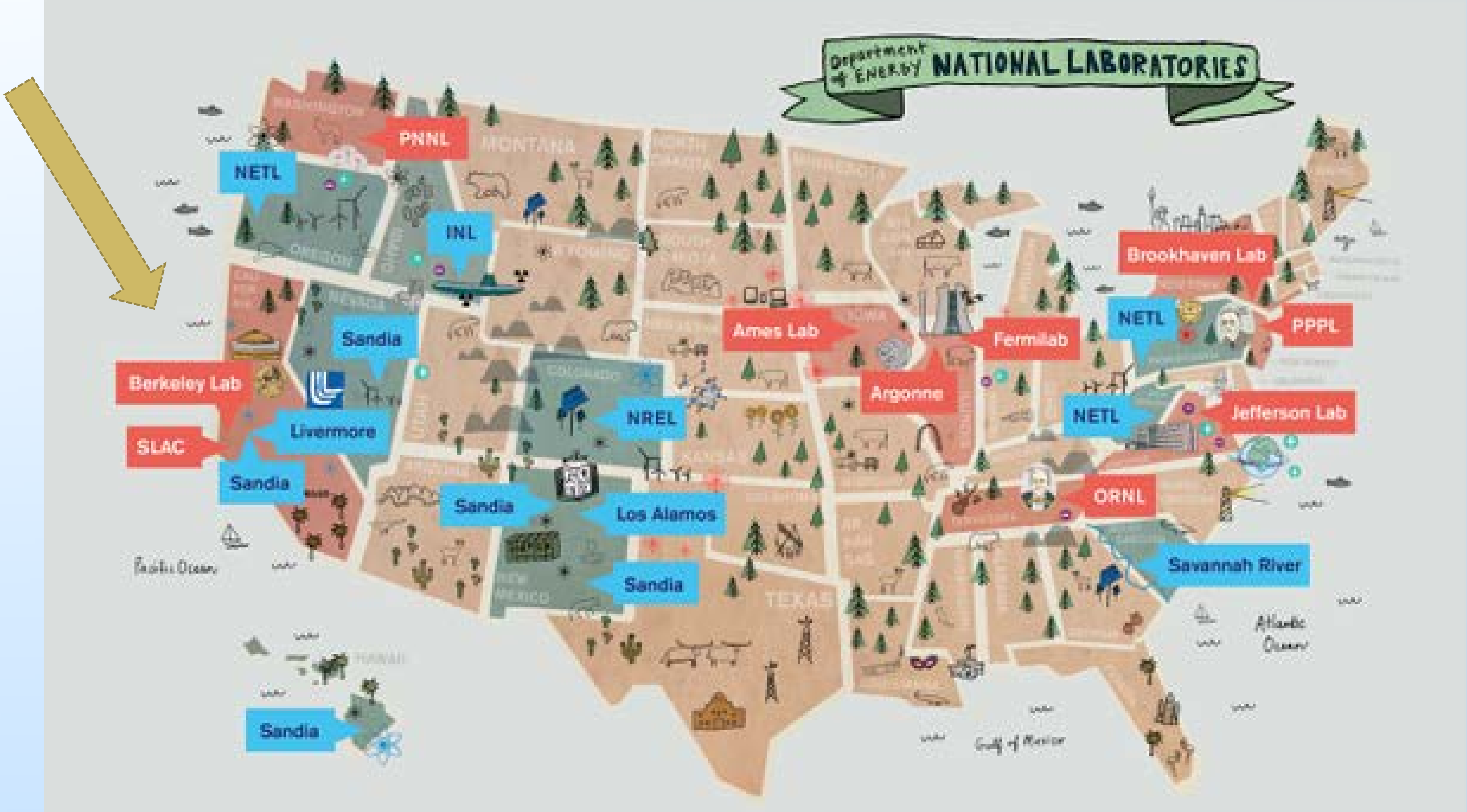
OBSERVATIONS

- Great locations
- Great social experience
- Small crowd, no parallel sessions
- All volunteer work
- Based on popular interest

- I attended all, but one
- I had papers in most

- Next: Santa Cruz, California

Department of Energy Labs



🟥 *Office of Science Labs*

🟦 *Other Offices Labs*

A. Shoshani

DOE’s Leadership Class Facilities

Oak Ridge Leadership Computing Facility

Titan

Cray XK7

20 petaflops

hybrid-architecture

18,688 AMD 16-core Opteron 6274 CPUs (a total of 299,008 processing cores)

18,688 NVIDIA Kepler GPUs

710 terabytes of memory

10 petabyte disk

Argonne Leadership Computing Facility

Mira

IBM Blue Gene/Q

10 petaflops

786,432 processors

768 terabytes of memory

7.6 petabytes disk

NERSC The National Energy Research Scientific Computing Center (NERSC) - LBNL

Hopper

Cray XE6

1.28 Petaflops/sec,

153,216 compute cores,

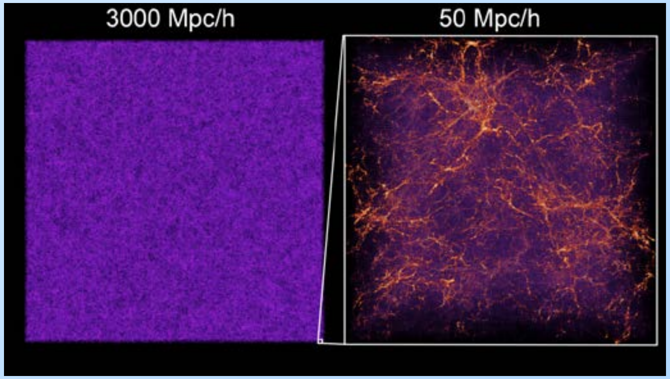
212 Terabytes of memory, and

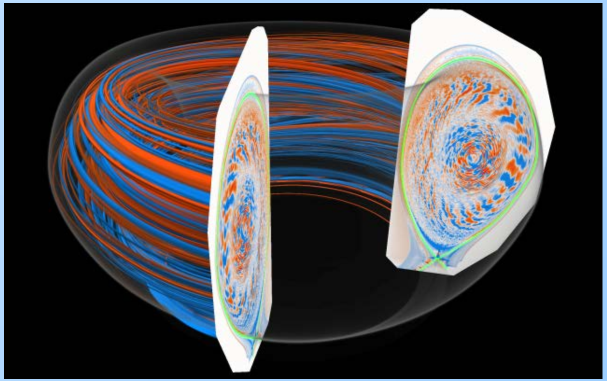
2 Petabytes of disk.

ESnet

Energy Sciences Network (ESnet)

Upgraded recently to 100 Gb/s on main connections



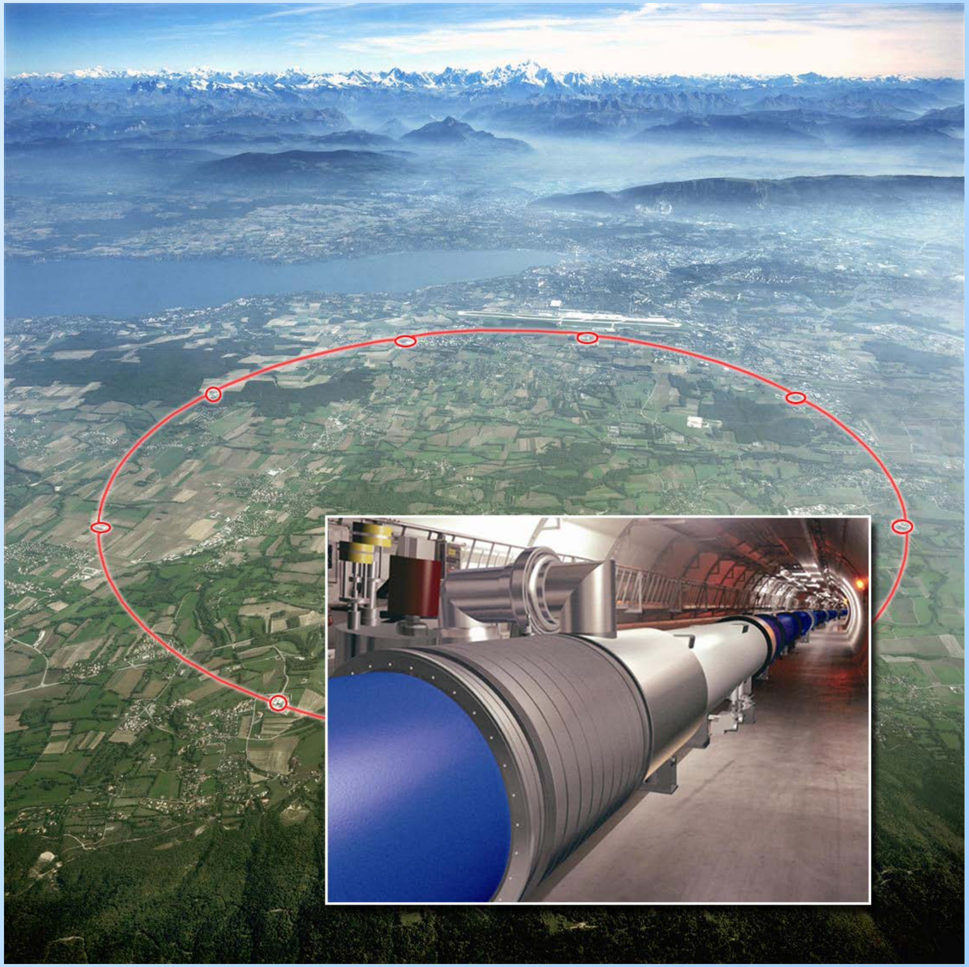


A. Shoshani

Example of Large Data Volume in Science

Large Hadron Collider: to find the God particle

- sensors capable of 140PB/s
- reduce 99.99% of data by hardware triggers
- Keep 15 PB per year
- 27 km tunnel
- ~10,000 superconducting magnets
- Operating temperature 1.9 Kelvin
- Construction cost:
  US$9Billion
- Power consumption: ~120 MW



A. Shoshani

Data models and SSDBM

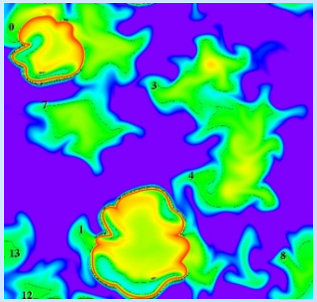
- Pre-1970
  - Hierarchical model
    - Integrated Data Store (IDS), by GE
    - Model based on efficient physical organization
      - E.g. projects → employees, employee → children
      - Specialized query interfaces (procedural: follow pointers)
    - Later: XML databases
    - Problem: data model does not capture more complex associations: projects ↔ employees
- Post-1970
  - Relational model
    - Separation of logical data model from physical data model (physical data independence)
    - Logical-level query language (SQL)
    - Mapping required query optimization, indexing, physical data layout,
    - Multiple implementation based on a standard query language

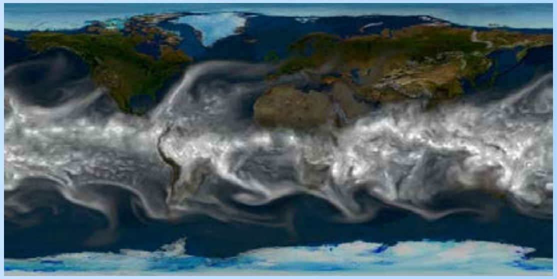
Why Scientists Don’t Use Data Management Systems?

(when I Joined LBNL in 1976)

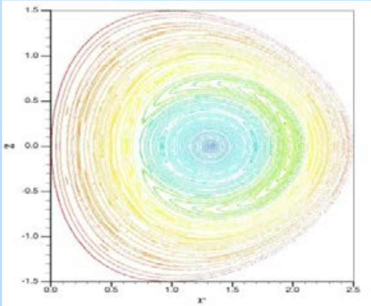
What does “Scientific Data Management” mean?

- **Target Scientific Applications**
  - Climate, Combustion, Fusion, Accelerator design, Cosmology,



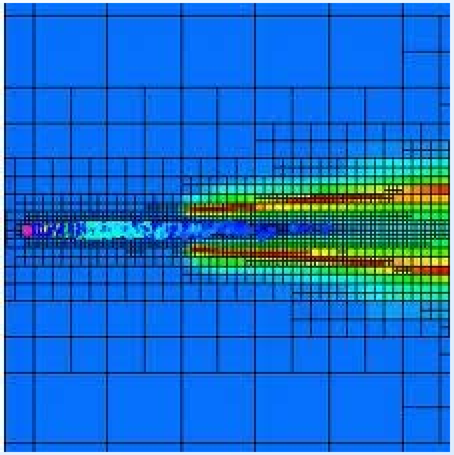


- Three pillars of science
  - Theory, Experiments, Simulations, and later Data Analysis (fourth paradigm)

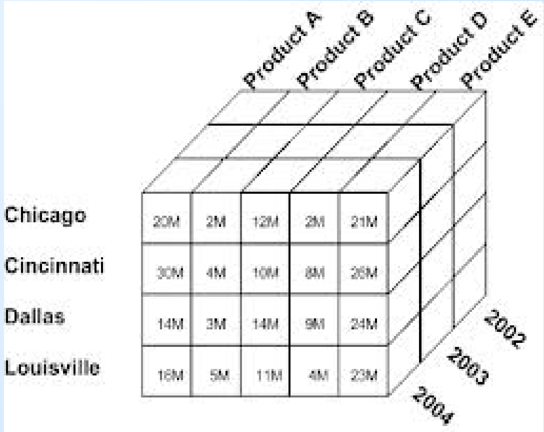


- **Algorithms, techniques, and software**
  - Representing scientific data – data models, metadata (structured/unstructured array models, geodesic models, sequence data, streaming data)
  - Managing I/O – methods for removing I/O bottleneck
  - Accelerating efficiency of access – data structures, indexing
  - Facilitating data analysis – data manipulations for finding patterns and meaning in the data
  - Support visual analytics – accelerate extraction of subsets for real-time visualization

Scientific Data Models

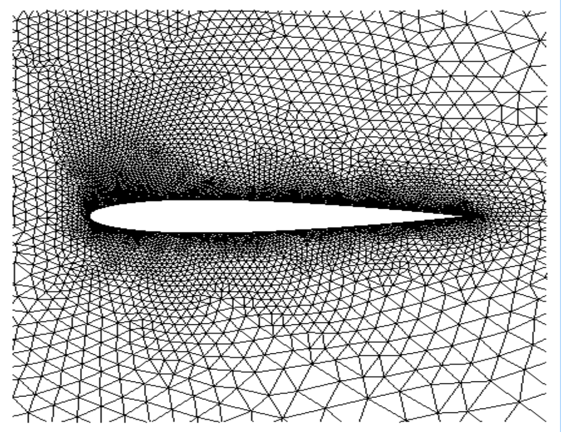


Adaptive Mesh Refinement

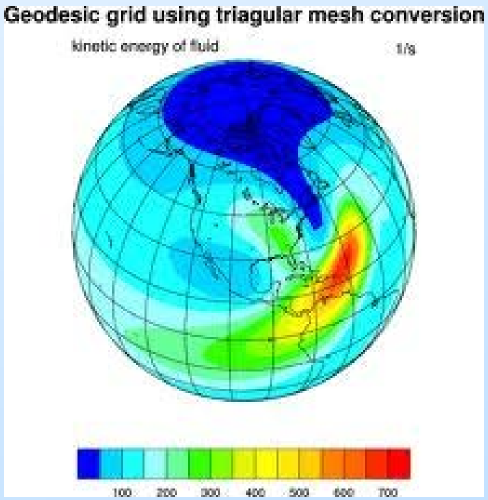


	Product A	Product B	Product C	Product D	Product E
Chicago	20M	2M	12M	2M	21M
Cincinnati	30M	4M	10M	8M	26M
Dallas	14M	3M	14M	9M	24M
Louisville	16M	5M	11M	4M	23M

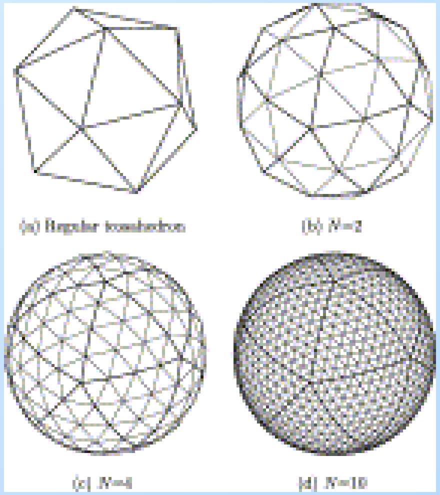
Data Cube



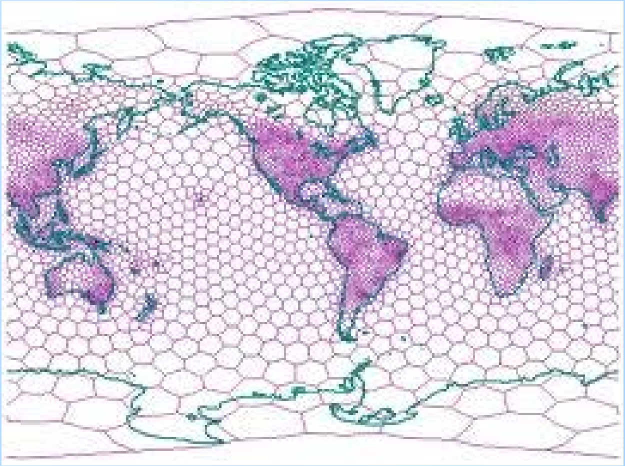
Unstructured triangular grid



Geodesic data model



Geodesic triangular data model

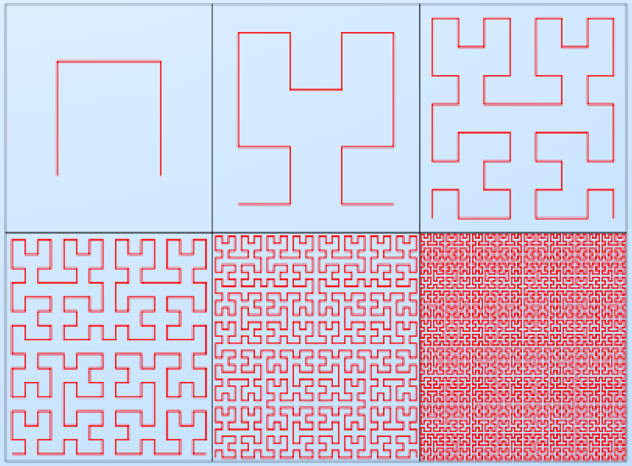


Unstructured grid: Voronoi tesselation

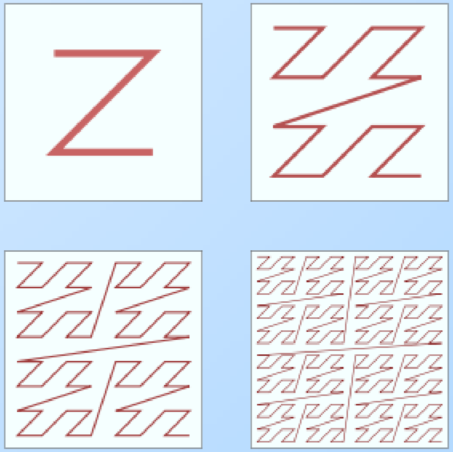
A. Shoshani

Physical Data Structure

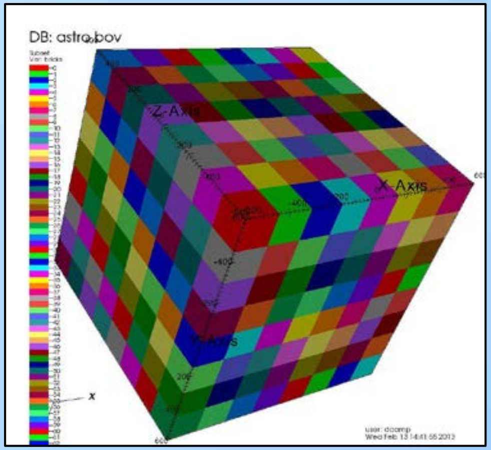
- Linearization of data based on data model
  - By coordinate order based on most prevalent access
  - Hilbert or Z-ordering to support local neighborhood access
- Partitioning data into blocks for parallel processing
  - Assigning block to different processors
  - Striping blocks on disk



Hilbert linearization order



Z-ordering



512-block dataset colored by thread ID

A. Shoshani

Scientific data models have special operators

- Spatial structures (e.g. climate, airplane wing)
  - Region operators, slices from 3D to 2D,
- Space over time structures
  - Spatial overlap over time-steps to track pattern progress
- Temporal data
  - Before/after operators, time-overlap operators
- Time-series data (e.g. sensor data)
  - Statistical operators over regular time-intervals
- Sequence data (e.g. biology)
  - Have special alphabet (4 base-pairs for DNA, 22 for protein)
- Irregular 3D structures
  - Protein folding operators
- etc., etc.

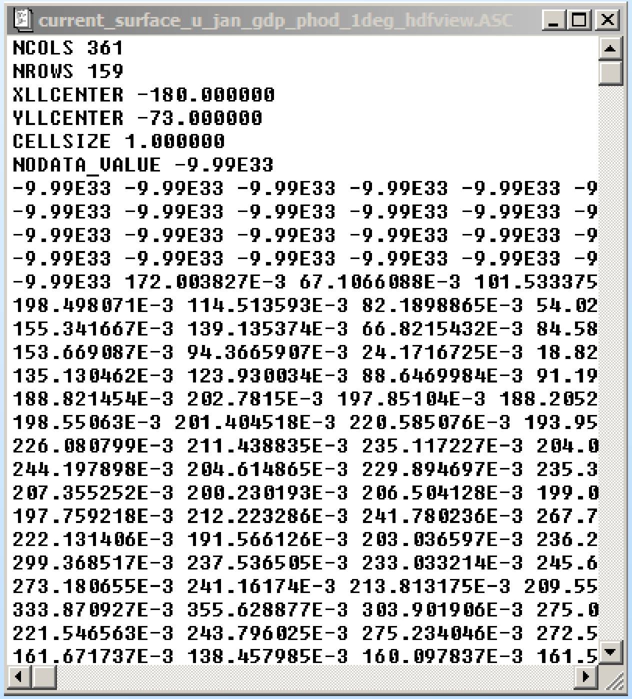
Scientific data management, analysis, and visualization

- **Data Management**
  - support of physical data structures and optimization of operations over scientific logical data structures
- **Data Analysis**
  - support for manipulations of logical data structures to enhance data understanding
- **Visualization**
  - facilitating real-time visual exploration of space-time data, as well as analysis of properties of various data structures

On Scientific Metadata

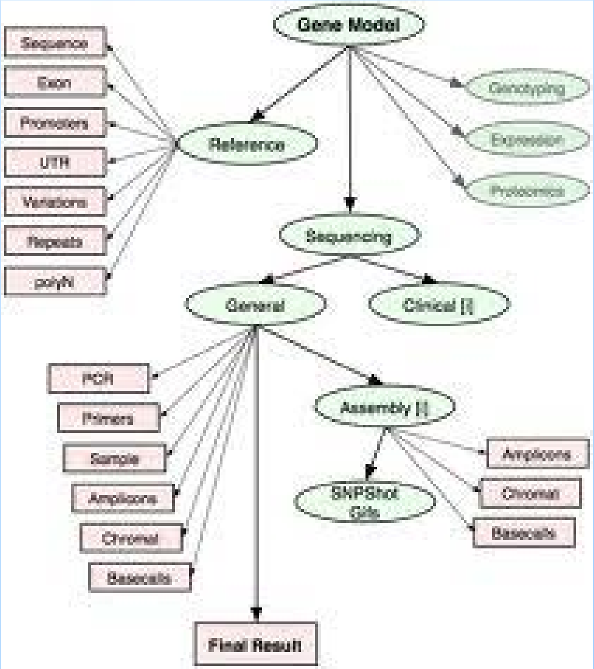
Metadata is essential to describe how the data was generated/collected

- Self-describing data formats (using headers and footers) – e.g. netCDF
- Hierarchical data formats allowing organization of data as well as annotation – e.g. HDF5
- External information: who, what, when, provenance, codes, device specifics,
- Ontologies, Controlled Vocabularies



```
NCOLS 361
NROWS 159
XLLCENTER -180.000000
YLLCENTER -73.000000
CELLSIZE 1.000000
NODATA_VALUE -9.99E33
-9.99E33 -9.99E33 -9.99E33 -9.99E33 -9.99E33 -9
-9.99E33 -9.99E33 -9.99E33 -9.99E33 -9.99E33 -9
-9.99E33 -9.99E33 -9.99E33 -9.99E33 -9.99E33 -9
-9.99E33 -9.99E33 -9.99E33 -9.99E33 -9.99E33 -9
-9.99E33 172.003827E-3 67.1066088E-3 101.533375
198.498071E-3 114.513593E-3 82.1898865E-3 54.02
155.341667E-3 139.135374E-3 66.8215432E-3 84.58
153.669087E-3 94.3665907E-3 24.1716725E-3 18.82
135.130462E-3 123.930034E-3 88.6469984E-3 91.19
188.821454E-3 202.7815E-3 197.85104E-3 188.2052
198.55063E-3 201.404518E-3 220.585076E-3 193.95
226.080799E-3 211.438835E-3 235.117227E-3 204.0
244.197898E-3 204.614865E-3 229.894697E-3 235.3
207.355252E-3 200.230193E-3 206.504128E-3 199.0
197.759218E-3 212.223286E-3 241.780236E-3 267.7
222.131406E-3 191.566126E-3 203.036597E-3 236.2
299.368517E-3 237.536505E-3 233.033214E-3 245.6
273.180655E-3 241.16174E-3 213.813175E-3 209.55
333.870927E-3 355.628877E-3 303.901906E-3 275.0
221.546563E-3 243.796025E-3 275.234046E-3 272.5
161.671737E-3 138.457985E-3 160.097837E-3 161.5
```

netCDF data structure

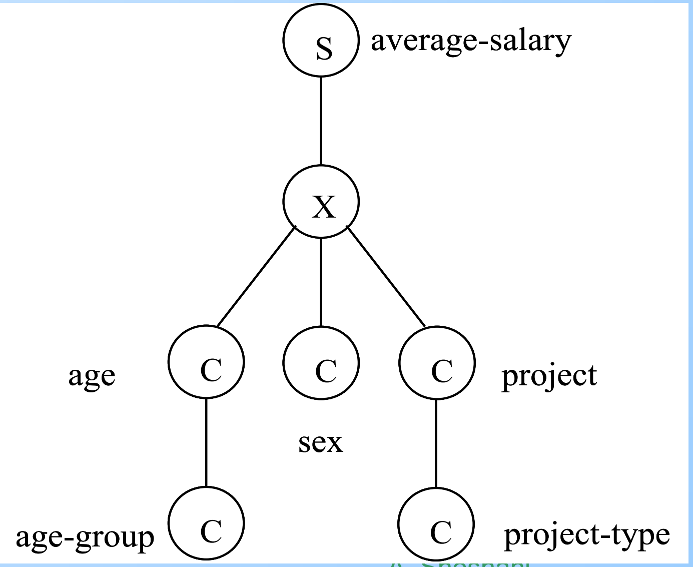


HDF5 hierarchical data format

First SSDBM (1981) – focus on statistical data

- Menlo Park, CA
- Looking at Socio-Economic data
  - Population by (state, city, race, age, sex)
  - Socio-economic scientists did not use database systems
  - Data model does not fit relational models
- Statistical data model
  - Multi-dimensional + hierarchies over dimensions
  - Became popular with SIGMOD conferences

Statistical Data Bases
Logical Model



A. Shoshani

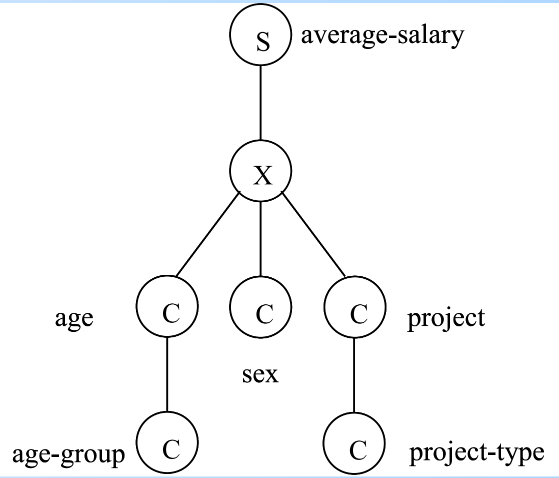
First SSDBM (1981) – focus on statistical data

- OLAP
  - Later SDBs were re-introduced as OLAP, plus operators (role-up, drill-down,   )
  - Paper on “OLAP vs. Statistical Databases” – PODS 1997
  - Later OLAP was visualized as “data cubes”, plus operators (Jim Gray)
  - Implementation of OLAP databases by Microsoft, Oracle, Sybase
  - Lesson: specialized systems developed for this type of a data model

- System S
  - 1981: Richard A. Becker: **Data Manipulation in the S System for Interactive Data Analysis.**

  R is an implementation of the S programming language

LOGICAL MODEL



S average-salary

X

age C

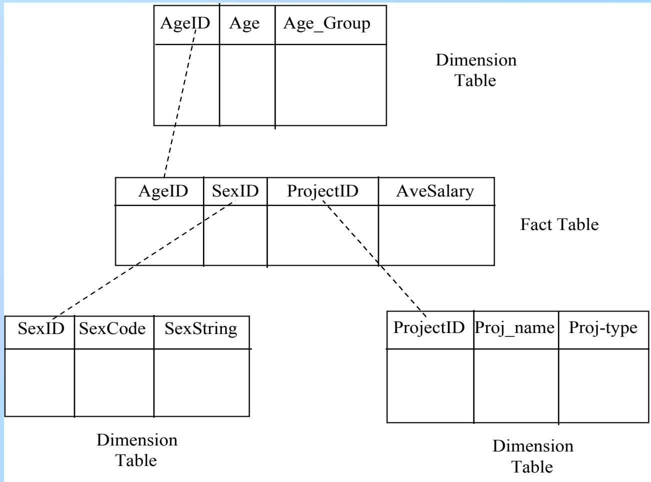
C sex

C project

age-group C

C project-type

ROLAP REPRESENTATION



AgeID	Age	Age_Group

Dimension Table

AgeID	SexID	ProjectID	AveSalary

Fact Table

SexID	SexCode	SexString

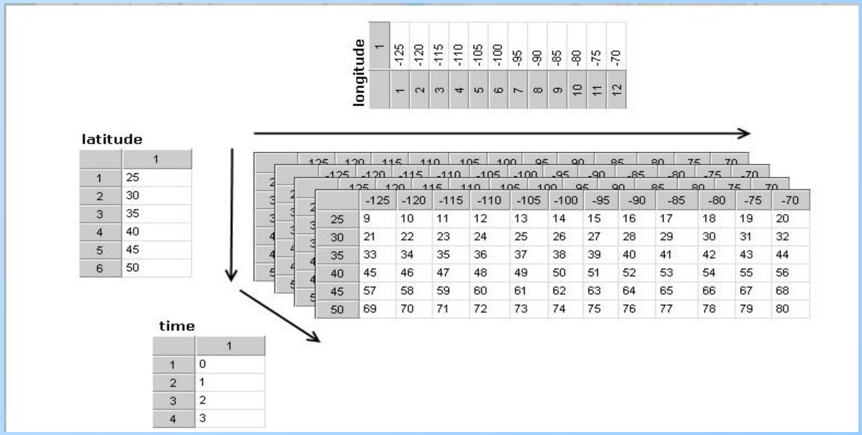
Dimension Table

ProjectID	Proj_name	Proj-type

Dimension Table

Third SSDBM (1986) – Luxemburg

- Rojer Cubbit
- Got involved in statistical office of EU
- SSDBM started alternating between US and EU
- Introducing Scientific data
- Why? Scientists in general did not use database management systems
- VLDB 1994:
  - “Characteristics of Scientific Databases” – VLDB 1984
    (Arie Shoshani, Frank Olken, Harry K. T. Wong)
  - Identified array data as an important model for scientists
  - Data kept in specialized file formats
    - NetCDF, HDF5, FITS,
    - Having their own libraries
  - This is still the case today!!!



longitude

	1
1	-125
2	-120
3	-115
4	-110
5	-105
6	-100
7	-95
8	-90
9	-85
10	-80
11	-75
12	-70

latitude

	1
1	25
2	30
3	35
4	40
5	45
6	50

time

	1
1	0
2	1
3	2
4	3

	-125	-120	-115	-110	-105	-100	-95	-90	-85	-80	-75	-70
25	9	10	11	12	13	14	15	16	17	18	19	20
30	21	22	23	24	25	26	27	28	29	30	31	32
35	33	34	35	36	37	38	39	40	41	42	43	44
40	45	46	47	48	49	50	51	52	53	54	55	56
45	57	58	59	60	61	62	63	64	65	66	67	68
50	69	70	71	72	73	74	75	76	77	78	79	80

SSDBM (1996-1998)

- NSF got interested – Maria Zemankova
  - Suggested to alternate every year between Europe and USA
  - Before that it was every other year
- 1997 – Olympia, WA
  - Interest in Environmental Data was introduced
    Francis P. Bretherton, William L. Hibbard: Metadata: A Case Study from the Environmental Sciences.
  - Also Knowledge Discovery
    Usama M. Fayyad: Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases
  - "Summarizability" of Statistical database introduced
    Hans-Joachim Lenz, Arie Shoshani: Summarizability in OLAP and Statistical Data Bases
- 1998 – Capri
  - Interest in Multidimensional Arrays was presented
    Norbert Widmann, Peter Baumann: Efficient Execution of Operations in a DBMS for Multidimensional Arrays
  - Product: Rasdaman, open-source

SSDBM (2001- 2004)

- 2001 – Fairfax, VA
  - Interest in Earth Systems was presented
    James Frew, Rajendra Bose: Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products
- 2002 – Edinburgh
  - Interest in Biology and Gene Expression was presented
    Albert Burger, Richard A. Baldock, Yiya Yang, Andrew M. Waterhouse, Derek Houghton, Nick Burton, Duncan Davidson:
    The Edinburgh Mouse Atlas and Gene-Expression Database: A Spatio-Temporal Database for Biological Research
- 2004 – Santorini
  - Interest in Scientific Workflows was presented
    Ilkay Altintas, Chad Berkley, Efrat Jaeger, Matthew B. Jones, Bertram Ludäscher, Steve Mock: Kepler: An Extensible System for Design and Execution of Scientific Workflows
  - Led to Kepler, an open-source product

SSDBM (2008- 2012)

- 2008 – Hong Kong
  - Interest in Scientific Ontology Databases was presented
    Paea LePendu, Dejing Dou, Gwen A. Frishkoff, Jiawei Rong: Ontology Database: A New Method for Semantic Modeling and an Application to Brainwave Data.
- 2011 – Portland
  - Interest in Scientific Database Systems was presented
    Michael Stonebraker, Paul Brown, Alex Poliakov, Suchi Raman: The Architecture of SciDB
  - Product: open source SciDB
- 2012 – Crete
  - Interest in Data Fusion was presented
    David Maier, V. M. Megler, António M. Baptista, Alex Jaramillo, Charles Seaton, Paul J. Turner: Navigating Oceans of Data.

SSDBM (2013- 2017)

- 2013 – Baltimore
  - Interest in Big Data was presented (keynote)
    Michael J. Franklin: Making Sense of Big Data with the Berkeley Data Analytics Stack.
  - Interest in Streaming Data was presented
    Hamid Mousavi*, Carlo Zaniolo: Fast Computation of Approximate Biased Histograms on Sliding Windows over Data Streams
- 2016 – Budapest
  - Interest in User-Defined-Functions (UDF) was presented
    Mark Raasveldt, Hannes Mühleisen: Vectorized UDFs in Column
  - Stores
  - Product: open source MonetDB/Python
- 2017 – Chicago
  - Interest in N-dimensional Arrays was presented
    Veranika Liaukevich, Dimitar Mišev, Peter Baumann, Vlad Merticariu: Location and Processing Aware Datacube Caching

Final Thoughts

- Work with domain scientists and identify their data problems
  - Their logical/abstract data model
  - Their operators on that data models, including functions on the data (UDFs)
  - Their metadata, ontology, controlled vocabularies
  - Their data constraints
- Finds out how they store their data – specialized file formats
  - Do not try to force them to reshaped their data into your system (too big of a task, they will loose interest)
- Build something useful to them, and integrate in their environment
  - That will keep their attention for continued collaboration
- Submit your paper(s) to SSDBM 🙂

LIGHT-HEARTED ANECDOTES

Fun memories

- 1988, Rome, Italy
  - Gucci bags to all + Channel perfume for woman
  - Banquet at an estate outside Rome, six course
- 1996, Stockholm, Sweden
  - River ride to forest, walk to banquet
- 1997, Olympia, Washington
  - Nature walk to ocean
- 1998, Capri, Italy
  - One afternoon free to visit blue grotto
- 2000, Berlin, Germany
  - River boat ride
- 2002, Edinburgh, Scotland
  - Yearly fireworks spectacular display
- 2004, Santorini, Greece
  - Boat ride to the islands – swimming in sea
- 2005, Santa Barbara, California
  - Banquet: barbecue on beach
- 2007, Banff, Canada
  - Spectacular nature setting in the park

- 2010, Heidelberg, Germany
  - Held at European Media Lab – beautiful gardens
- 2011, Portland, Oregon
  - Great beer at location near river
- 2012, Crete, Greece
  - Beautiful hotel with view of Mediterranean

A. Shoshani

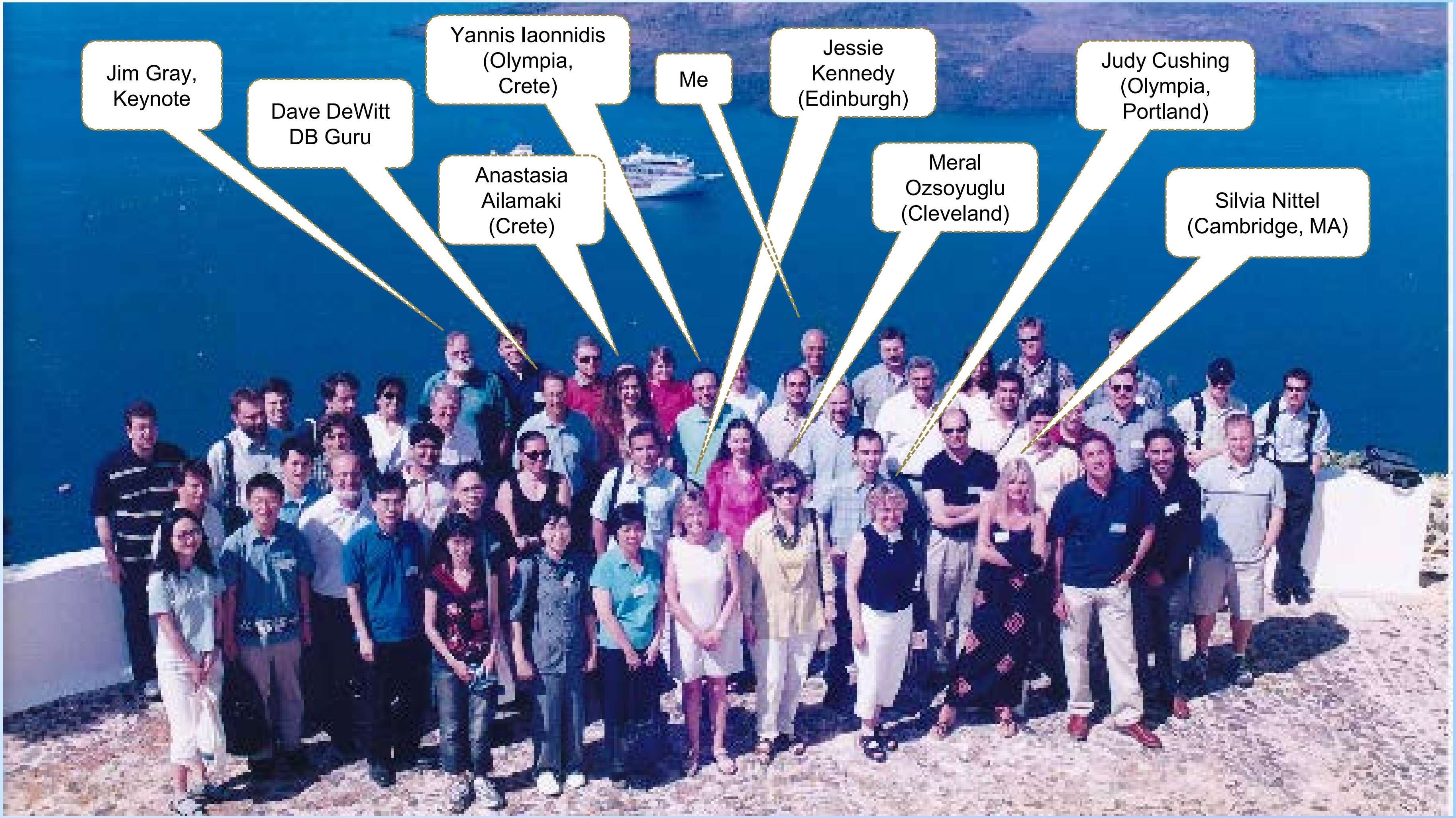
Berlin (2000)

Yannis Iaonnidis (Olympia, Crete)

Me

Jessie Kennedy (Edinburgh)

Santorini (2004)


Jim Gray, Keynote
Dave DeWitt DB Guru
Yannis Iaonnidis (Olympia, Crete)
Anastasia Ailamaki (Crete)
Me
Jessie Kennedy (Edinburgh)
Meral Ozsoyuglu (Cleveland)
Judy Cushing (Olympia, Portland)
Silvia Nittel (Cambridge, MA)

Banff (2007)

Marianne
Winslett
(New Orleans)

New Orleans (2009)

21st International Conference
on
Scientific and Statistical
Database Management
June 2–4, 2009
Room: Vieux Carré
17th Floor
Finding Structurally Different Time Series Data

A. Shoshani

Crete (2012)

A. Shoshani

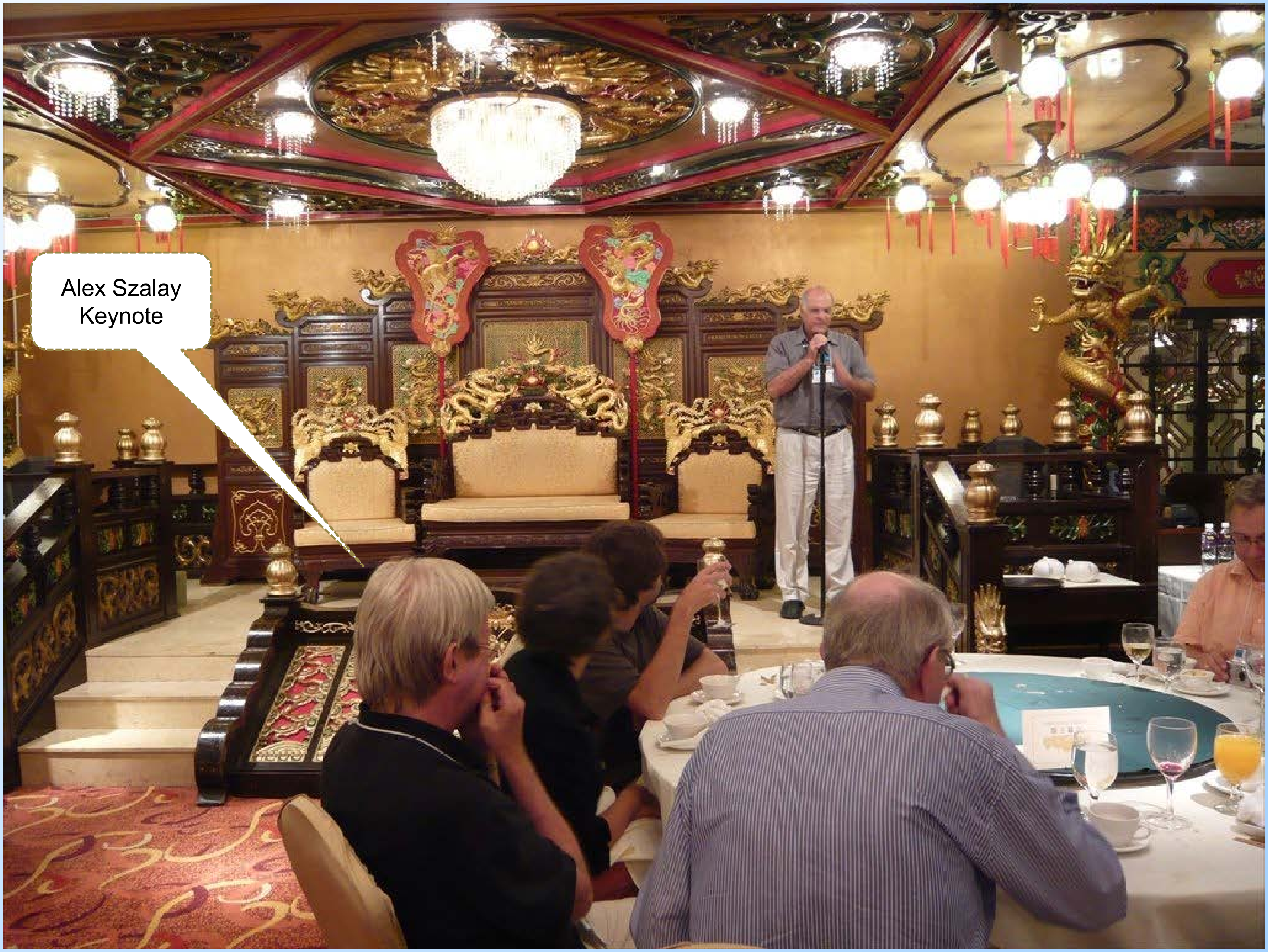
Hong Kong (2008)

Crown Chair

Hong Kong (2008)

Bertram Ludaescher (Program Chair)
Me
Nikos Mamoulis (General Chair)

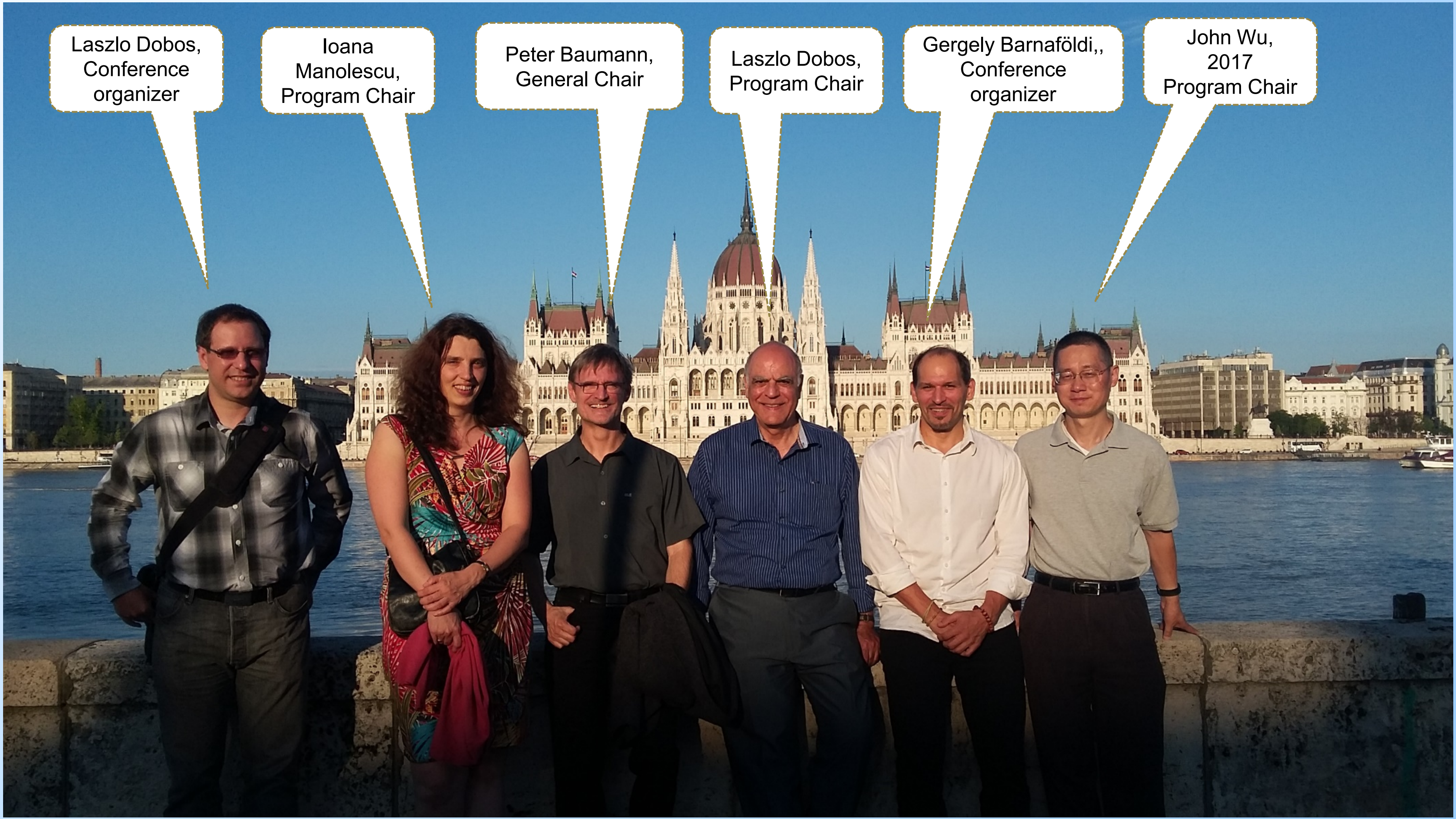
Hong Kong – rain, rain, everywhere



Aalborg (2014)

Torben Pedersen, Program Chair

Budapest (2016)


Laszlo Dobos, Conference organizer
Ioana Manolescu, Program Chair
Peter Baumann, General Chair
Laszlo Dobos, Program Chair
Gergely Barnaföldi,, Conference organizer
John Wu, 2017 Program Chair

THE END