40 Years of VLDB

A Brief Look at History

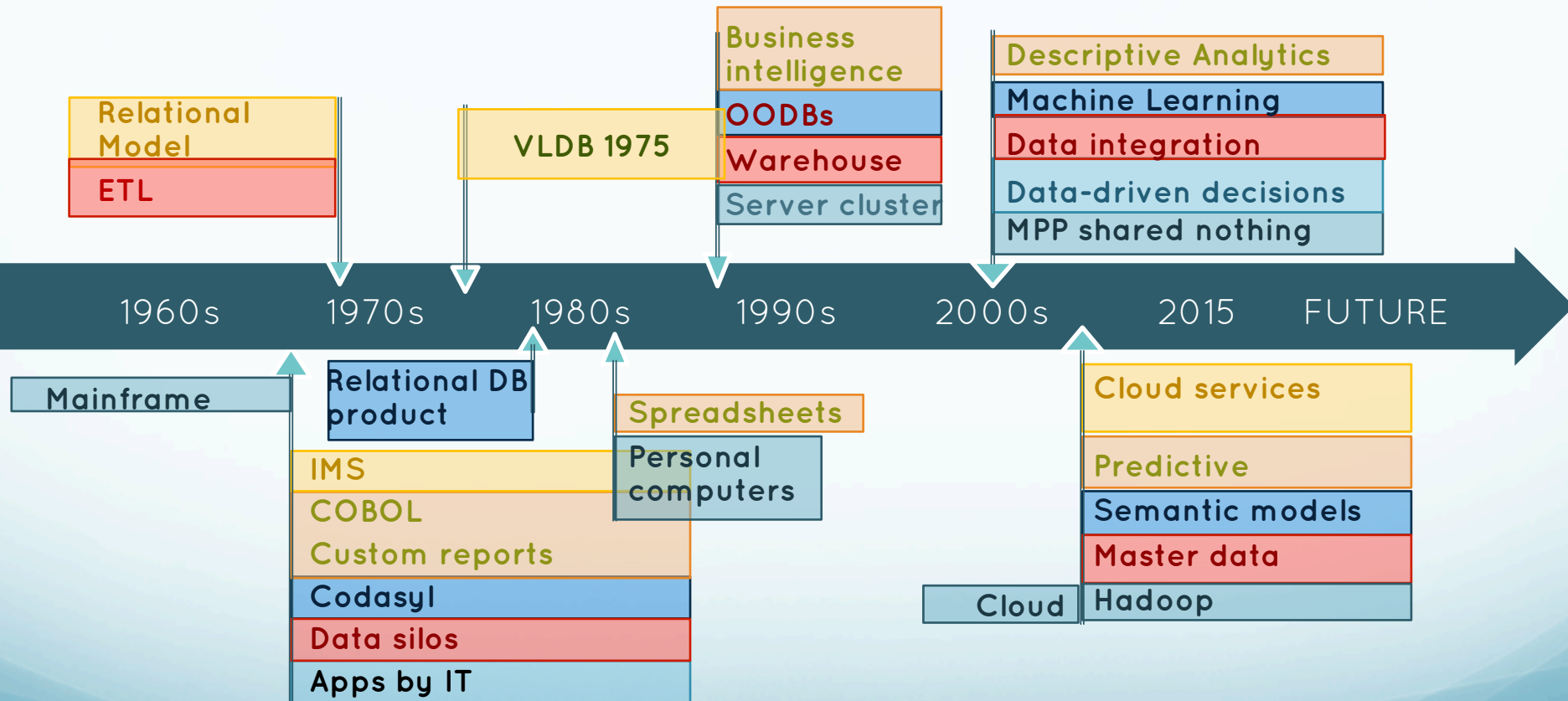| | 1960s | 1970s | 1980s | 1990s | 2000s | 2015 | FUTURE |

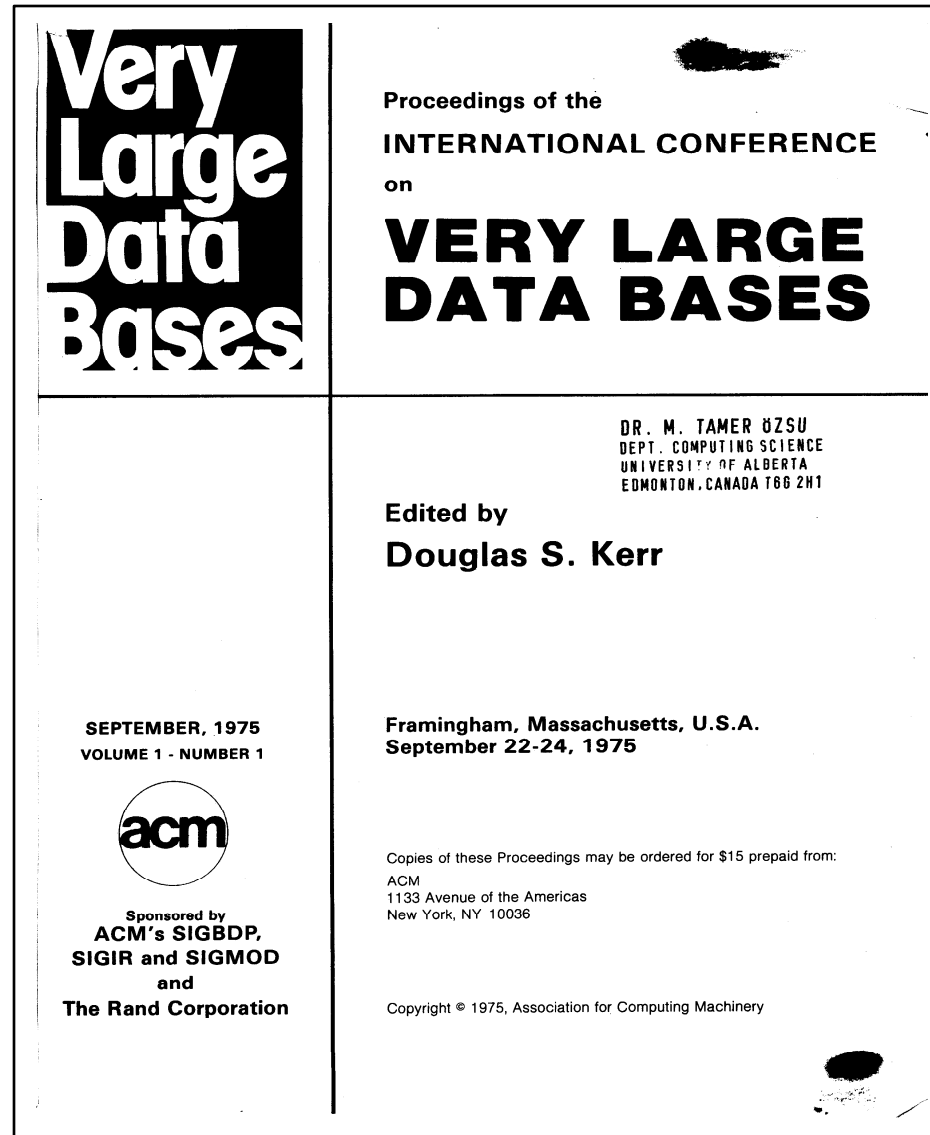# We've come a long way

IBM: 1975 – 2005, 2009 - 2014

Paradata: 2014 -

Panelists

- Don Chamberlin

- Phil Bernstein

- Michael Brodie

- Alfons Kemper

- Mike Stonebraker

- Moderator: Pat Selinger

1975

- Personal computers did not exist

- The internet did not exist

- Apple, Oracle, Google did not exist

- E-commerce did not exist

- One computer had many users

- A "minicomputer" cost tens of thousands of dollars

- Bill and Paul dropped out of college

1975

- A "mainframe" computer cost hundreds of thousands of dollars

- Main memory (magnetic cores) cost $100,000/MB (today: $0.001)

- One MB was a lot

1975

- Disk storage cost $300/MB (today: $0.00002)

- Disk space was allocated in 0.5 MB "cylinders"
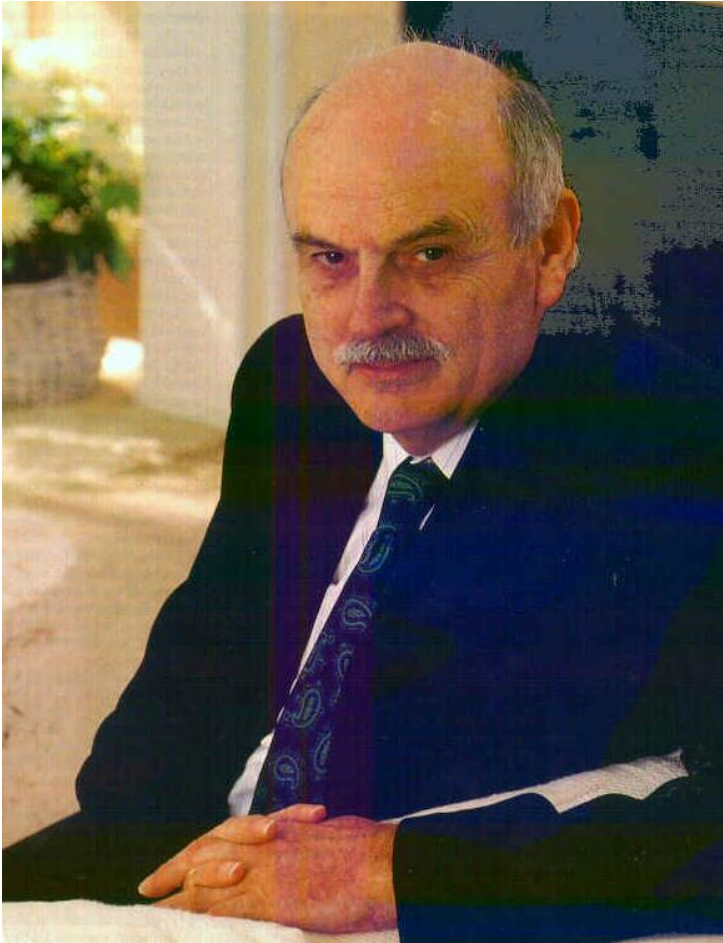
- 1 GB was a lot

Database Research in 1975

- Charles Bachman had received the 1973 ACM Turing Award

- Turing Lecture: "The Programmer As Navigator"

- The computer program "navigates" through a space of data records, following pointers.

Database Research in 1975

- E.F. Codd had published "A Relational Model of Data" in *CACM*, June 1970

- "Data independence"

- "Tell me what you want, not how to find it."

- Codd's Turing award was still in the future (1981)

- The big question: could a relational system perform?

Database Research in 1975

- Ted Codd and Charlie Bachman had just debated the merits of relational vs. network data models at ACM SIGFIDET (precursor of SIGMOD) 1974.

- Several research projects were underway to implement the relational model and answer the "big question."
  - System R at IBM Research
  - Ingres at UC Berkeley
  - Peterlee Relational Test Vehicle at IBM UK

- The '70s were an exciting time for System R
  - 45 technical papers in open literature, 1973 – 1981
  - System R and Ingres would jointly receive the ACM Software System Award in 1988

System R

- Pat Selinger designed the world's first cost-based relational query optimizer

- Along the way, she discovered and solved the "Halloween Problem"

- Pat would be named an IBM Fellow in 1994 and elected to the National Academy of Engineering in 1999

System R
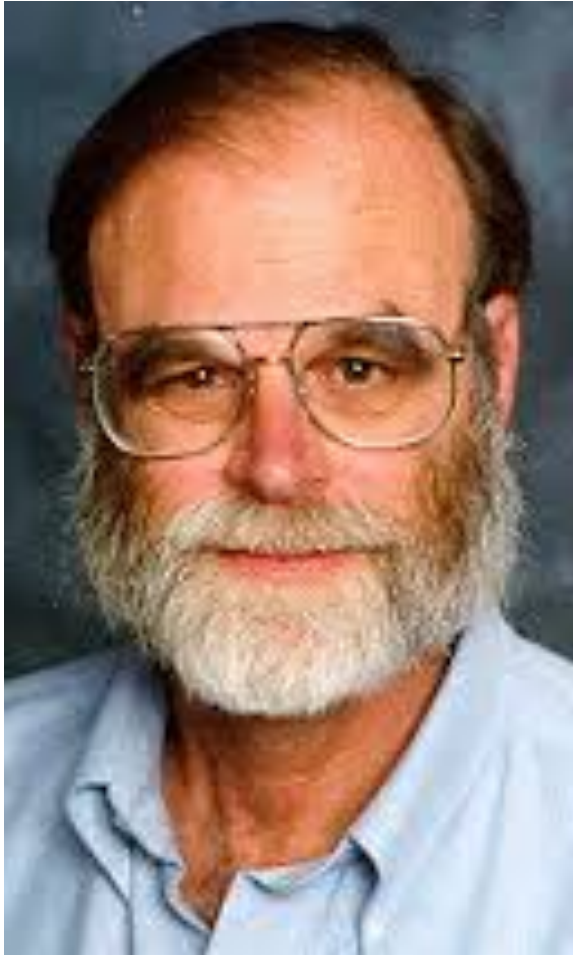
- Jim Gray defined the transactional properties (Atomicity, Consistency, Isolation, Durability) that made e-commerce possible
- "Granularity of Locks" paper in VLDB 1975
- Jim would receive a Turing Award for this work in 1999
- E-commerce retail sales exceeded $300 Billion in 2014

System R

- Don Chamberlin and Ray Boyce,
  "SEQUEL: A Structured English Query Language",
  ACM SIGFIDET Workshop, May 1974

SQL

- SQL had staying power because it had a way to evolve to meet changing requirements

- ANSI/ISO Standard adopted in 1986 and updated many times since
  - New kinds of queries (recursion, outer join)
  - Active features (assertions, triggers)
  - Online analytic processing (OLAP)
  - Object-oriented functionality (user-defined types)
  - Procedural language extensions (PL/SQL)

Relational Databases are Alive and Well

Database Systems Continue to Evolve

- The Web has eliminated the distinction between "data" and "documents"

- New data models for Web applications: XML, JSON, . . .

- New transaction models: "BASE" vs. "ACID"

- Schemaless data, or rapidly changing schemas

- Massive parallelism, imprecise search

- N. O. ("not only") SQL

40 Years of VLDB

Phil Bernstein
Microsoft Research

Papers in 1st VLDB, 1975
Which were **<u>not</u>** selected for journal?

- Moshé M. Zloof: *Query-by-Example: the Invocation and Definition of Tables and Forms*

- Jim Gray, Raymond A. Lorie, Gianfranco R. Putzolu, Irving L. Traiger:
  Granularity of Locks in a Large Shared Data Base

- Peter P. Chen: *The Entity-Relationship Model: Toward a Unified View of Data*

- Rudolf Bayer, J. K. Metzger: *On the Encipherment of Search Trees and Random Access Files*

- David W. Stemple:
  A Data Base Management Facility for Automatic Generation of Data Base Managers

- Shamkant B. Navathe, James P. Fry:
  Restructuring for Large Data Bases: Three Levels of Abstraction

Where was I in 1975?

- In May, I attended SIGMOD 1975

- In August, I received my PhD at University of Toronto
  - "Functional Dependencies and Normalization in the Relational Database Model"

- In September, I attended the 1st VLDB in Boston

- And started a post-doc Toronto, followed by 7 years at Harvard

Database field in 1975

- Wasn't totally respectable, academically

- Closely tied to business data processing

- The relational model was getting traction in the research community
  - But implementations were just getting started
  - And no one was sure it was commercially viable

Database Conferences in 1975

- Attendance was about 75.

- Single track

Session Titles at 1ˢᵗ VLDB, 1975

- Data Description Models

- Physical Structures and Implementation

- Database Design Tools

- Performance and Restructuring

- Application and Management Issues

- Distributed Databases

- Security and Integrity

- System and Memory Architecture

Session Titles at 1st VLDB, 1975

- Data Description Models

- Physical Structures and Implementation

- Database Design Tools

- Performance and Restructuring

- Application and Management Issues

- Distributed Databases

- Security and Integrity

- System and Memory Architecture

These are still the main areas of database research.

It's the sub-areas that change, driven by …

- Applications
  - Computer aided design, keyword search, e-commerce, data analytics
- Workloads
  - Data translation, ad hoc querying, transaction processing, natural language querying, data mining, data integration
- System architectures
  - Client-server, PC databases, data warehouses, column stores, MMDB
- Data models
  - Relational, Objects, RDF, XML, graphs
- Hardware
  - Logic-per-track, DB machines, parallel processors, multi-core, large main memory, SSD, sensors

Papers in 1st VLDB, 1975
Which were **<u>not</u>** selected for journal?

- Moshé M. Zloof: *Query-by-Example: the Invocation and Definition of Tables and Forms*

- Jim Gray, Raymond A. Lorie, Gianfranco R. Putzolu, Irving L. Traiger: *Granularity of Locks in a Large Shared Data Base*

- Peter P. Chen: *The Entity-Relationship Model: Toward a Unified View of Data*

- Rudolf Bayer, J. K. Metzger: *On the Encipherment of Search Trees and Random Access Files*

- David W. Stemple: *A Data Base Management Facility for Automatic Generation of Data Base Managers*

- Shamkant B. Navathe, James P. Fry: *Restructuring for Large Data Bases: Three Levels of Abstraction*

Some Papers in 1st VLDB, 1975

- **Moshé M. Zloof:** *Query-by-Example: the Invocation and Definition of Tables and Forms*

- **Jim Gray, Raymond A. Lorie, Gianfranco R. Putzolu, Irving L. Traiger:** *Granularity of Locks in a Large Shared Data Base*

- Peter P. Chen: *The Entity-Relationship Model: Toward a Unified View of Data*

- Rudolf Bayer, J. K. Metzger: *On the Encipherment of Search Trees and Random Access Files*

- David W. Stemple: *A Data Base Management Facility for Automatic Generation of Data Base Managers*

- Shamkant B. Navathe, James P. Fry: *Restructuring for Large Data Bases: Three Levels of Abstraction*

The End

40 Years of VLDB: Our Heritage and Our Future

Michael L. Brodie

VLDB 2015

Kohala Coast, Hawaiʻi

VLDB

- Heritage
  - Role
  - Contribution
  - Success
  - Trends
- Future

VLDB

"one of the most eminent venues for the dissemination of research and development results in the field of data management" Mike Carey

VLDB

- **Endowment**
- **Conference Series**

  1. September 22-24, 1975   Framingham, Massachusetts, USA

  ...

  41. August 31 – Sept 4, 2015   Kohala Coast, Hawai'i, USA
  42. September 5-9, 2016   New Delhi, India
- **VLDB Journal, since 1992**
  - Rank 1 of 45: *Comp. sc., Hardware & Architecture*
  - Rank 1 of 99: *Comp. sc., Information Systems*
- **Proceedings of the VLDB Endowment (PVLDB)**
  - Proceedings of first 23 conferences (1975-1998)
  - VLDB Conference papers 1982-2004
  - Volume 1, 2008
- **VLDB Archives**
  - VLDB Conference web sites since 1995
- **VLDB Awards**
- **Virtual VLDB VLDB.org**

Role

- Publish or Perish

- Vibrant forum

- Annual database community shindig *in exotic places*

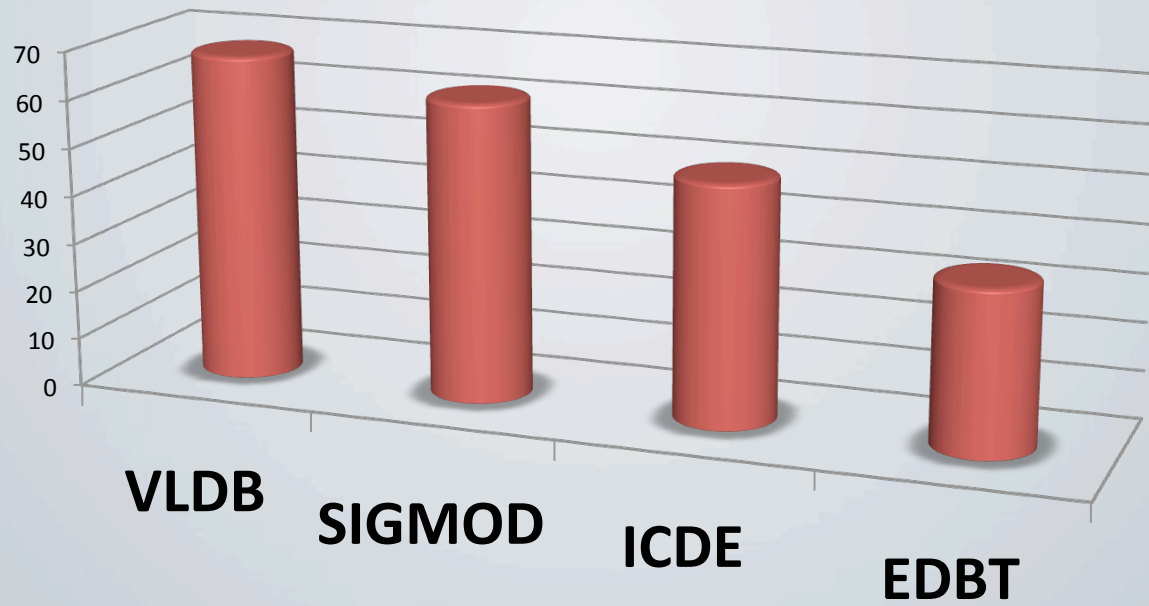# Contribution

Established Data Management as a

- Discipline

VLDB Correlations

Database Conference Ranking
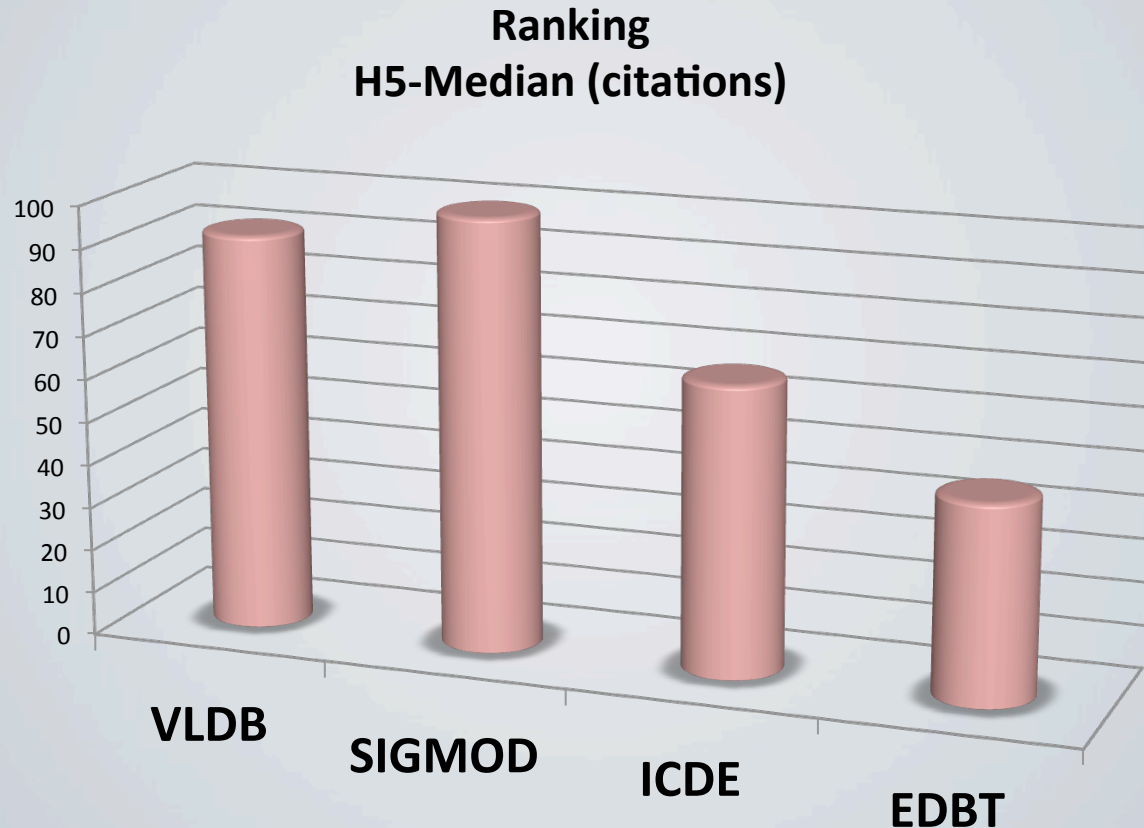
- 1st of 20* / 263**

Ranking
H5-Index (paper citations)

* Google Scholar   ** Microsoft Academic Search
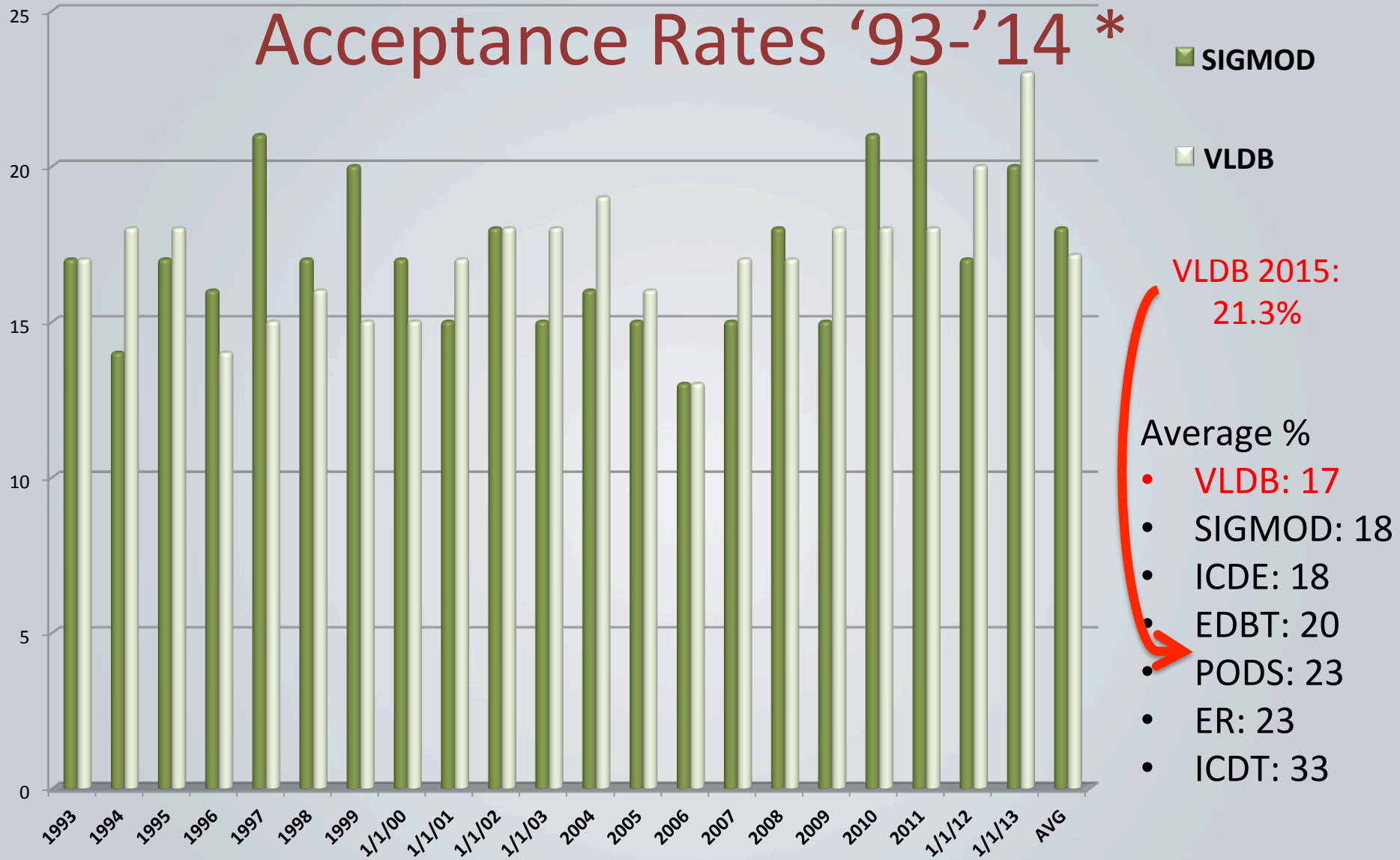
Database Conference Ranking

- 1st of 20* / 263**

Ranking
H5-Median (citations)
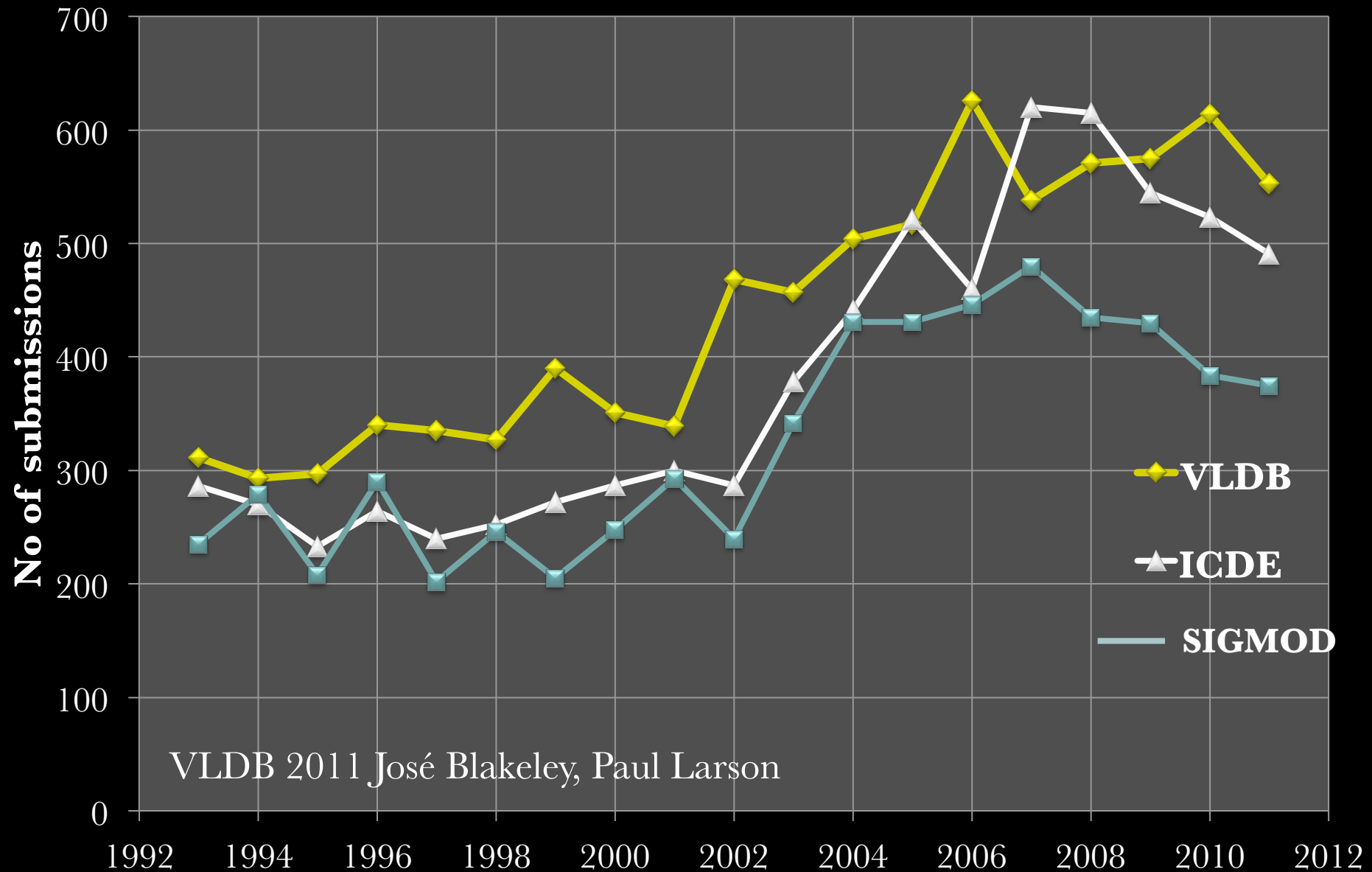
VLDB

Submit Accept Rate '93-'14
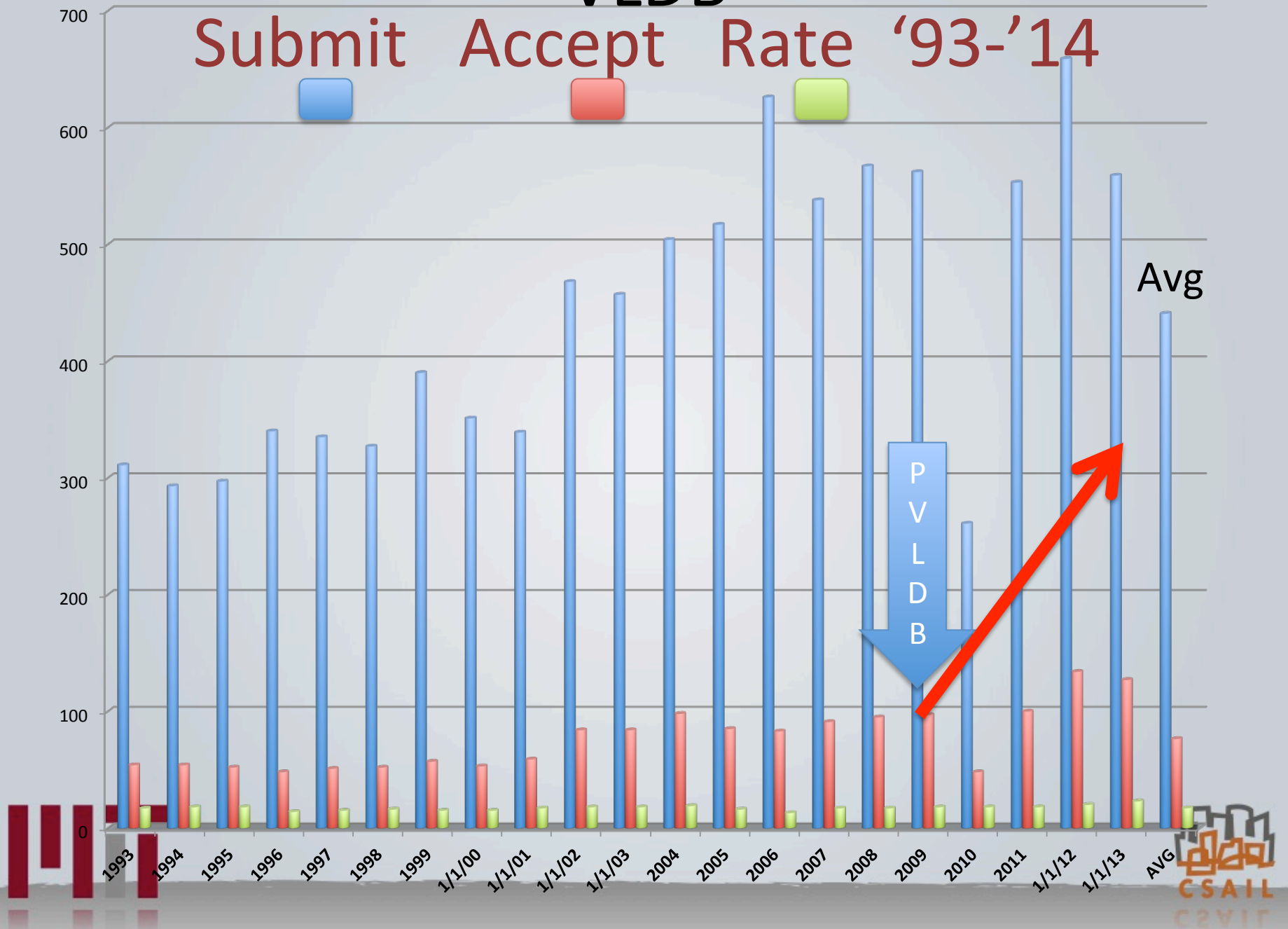
Succeed?

Scurrilous, random survey*

VLDB as a Whole
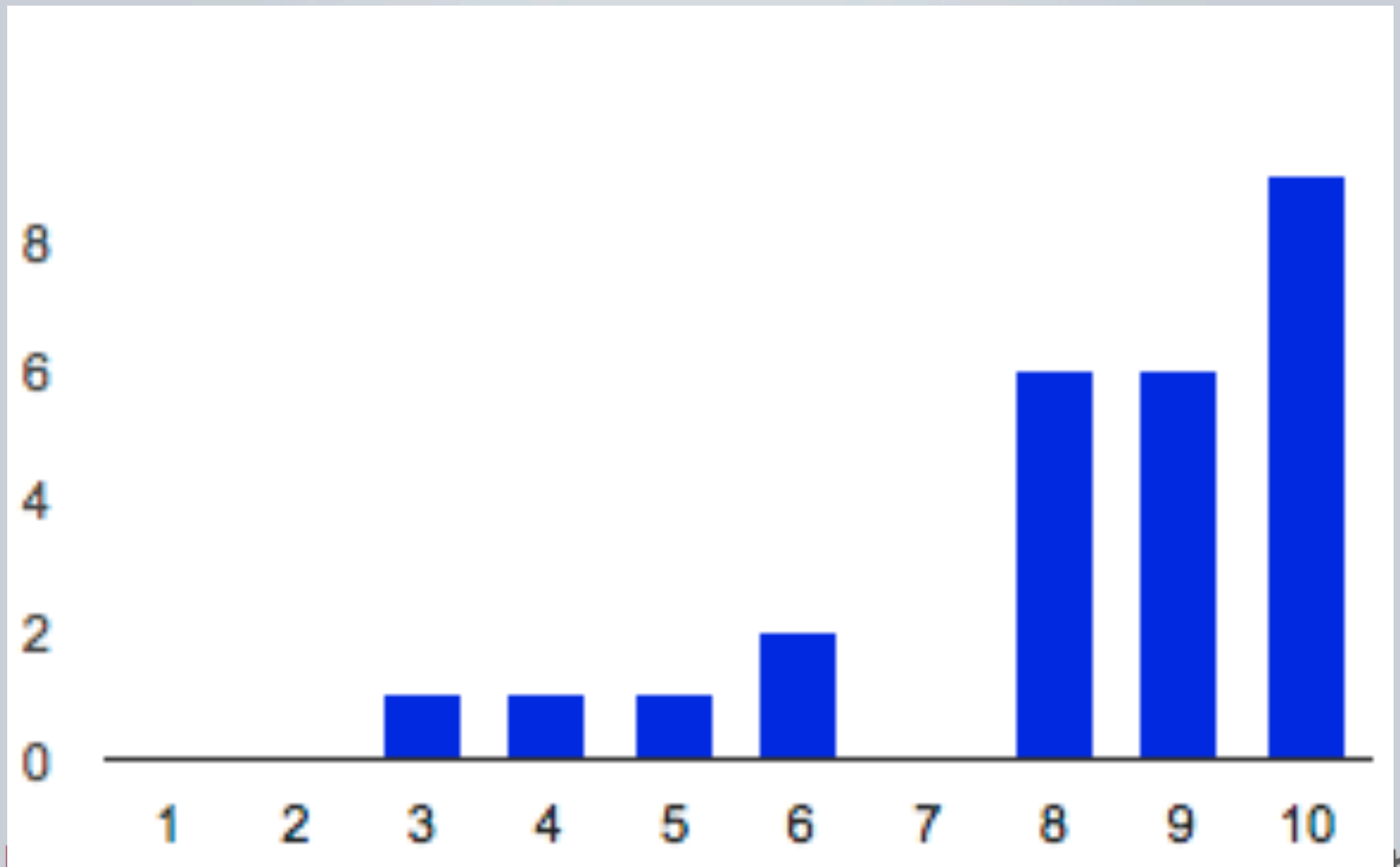
Virtual VLDB (VLDB.org)
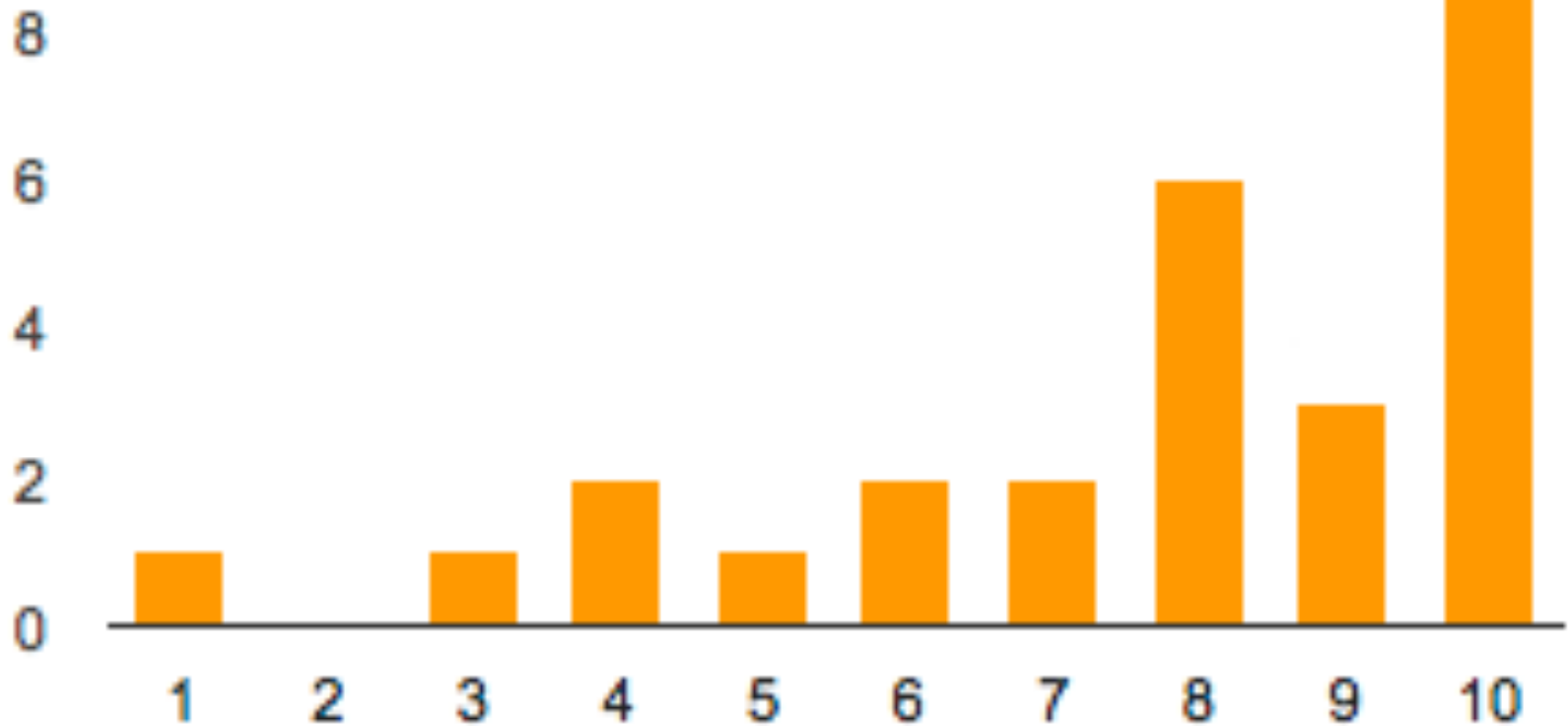
Virtual VLDB (VLDB.org)*

Unique Attendees / Year Avg = 500
Unique Visitors / Year Avg = 185,000

1 attendee ➔ 370 online average

1 attendee ➔ 232 online 2015
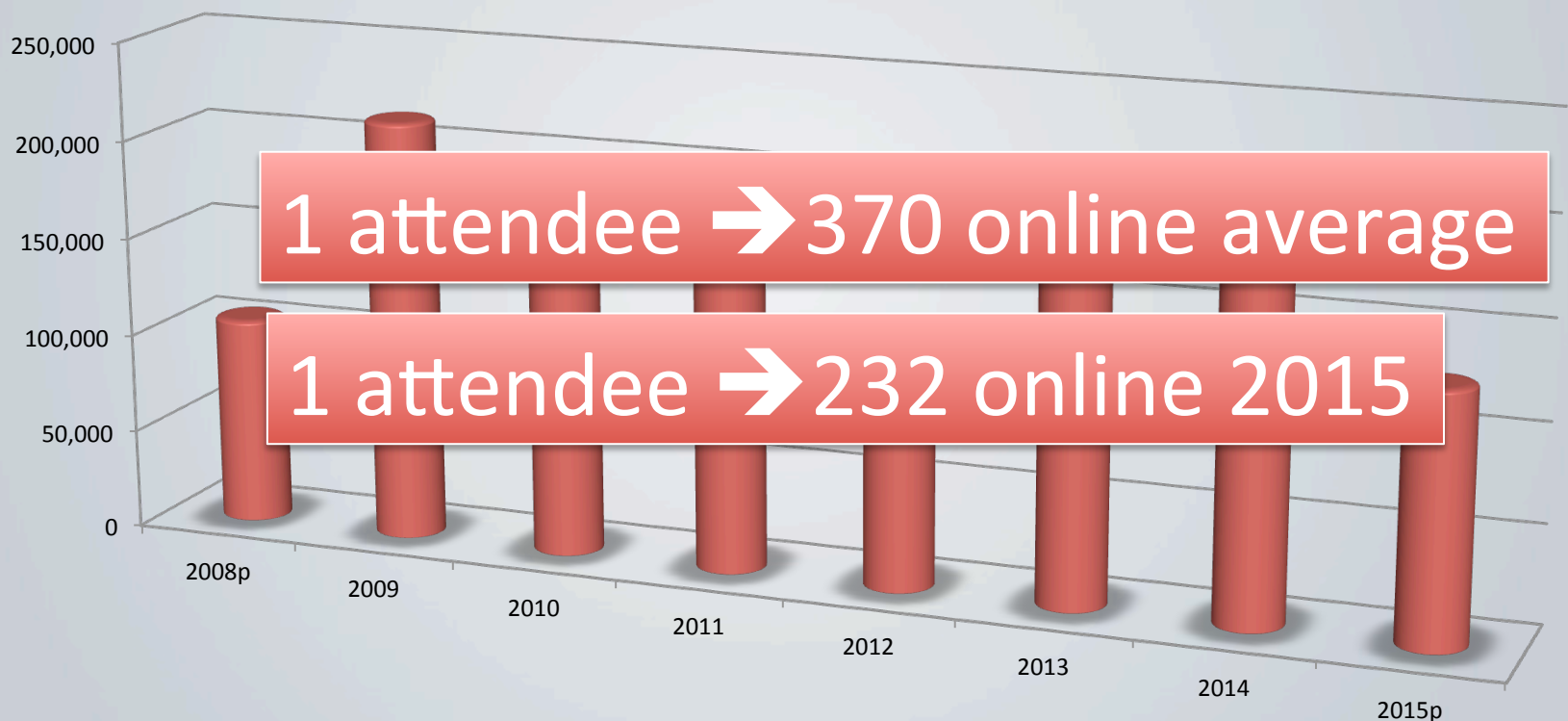
* Sheng Wang, VLDB Endowment Webmaster 2013-2015

VLDB Keynotes (invited talks) (5%)
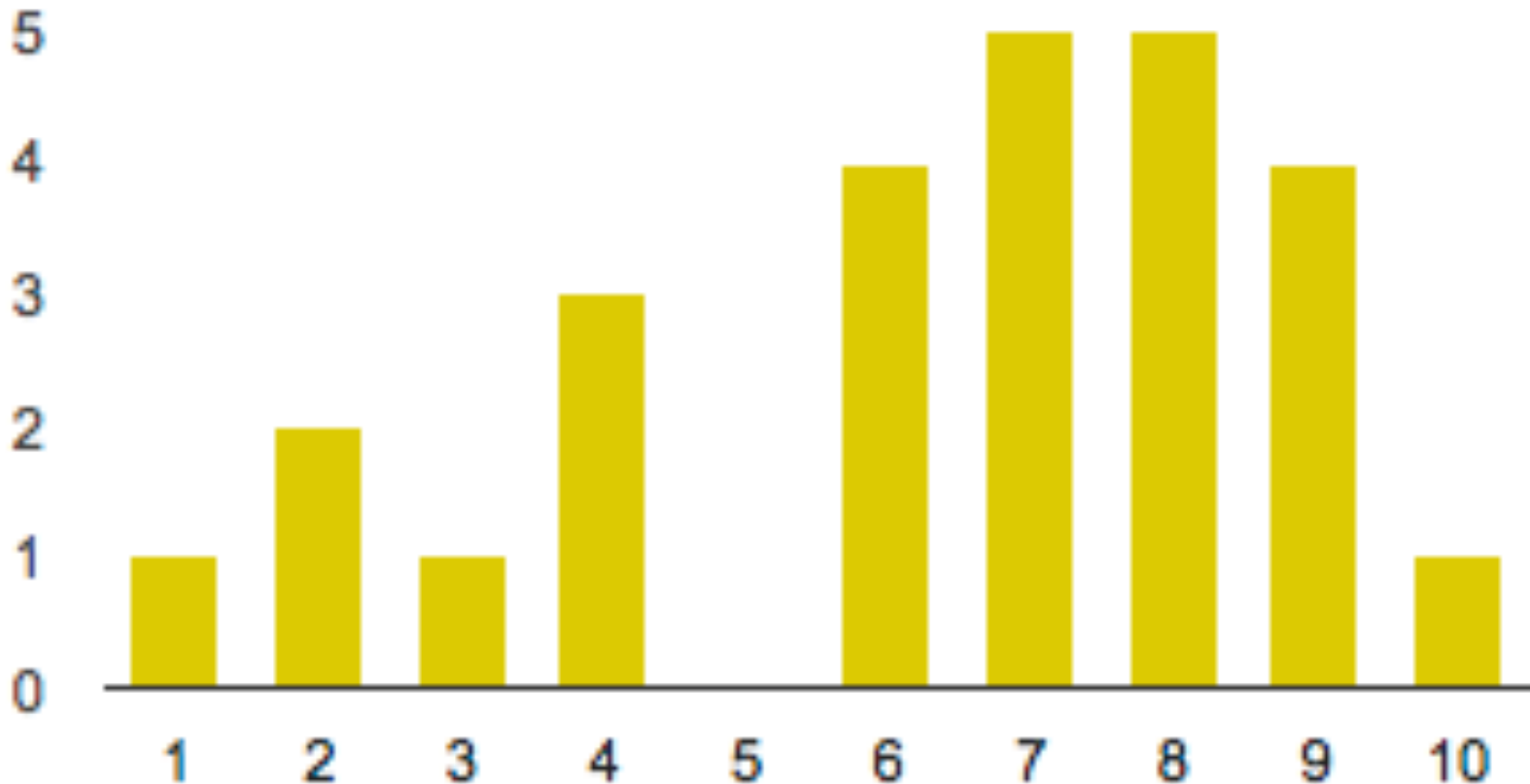
VLDB Panels (3%)
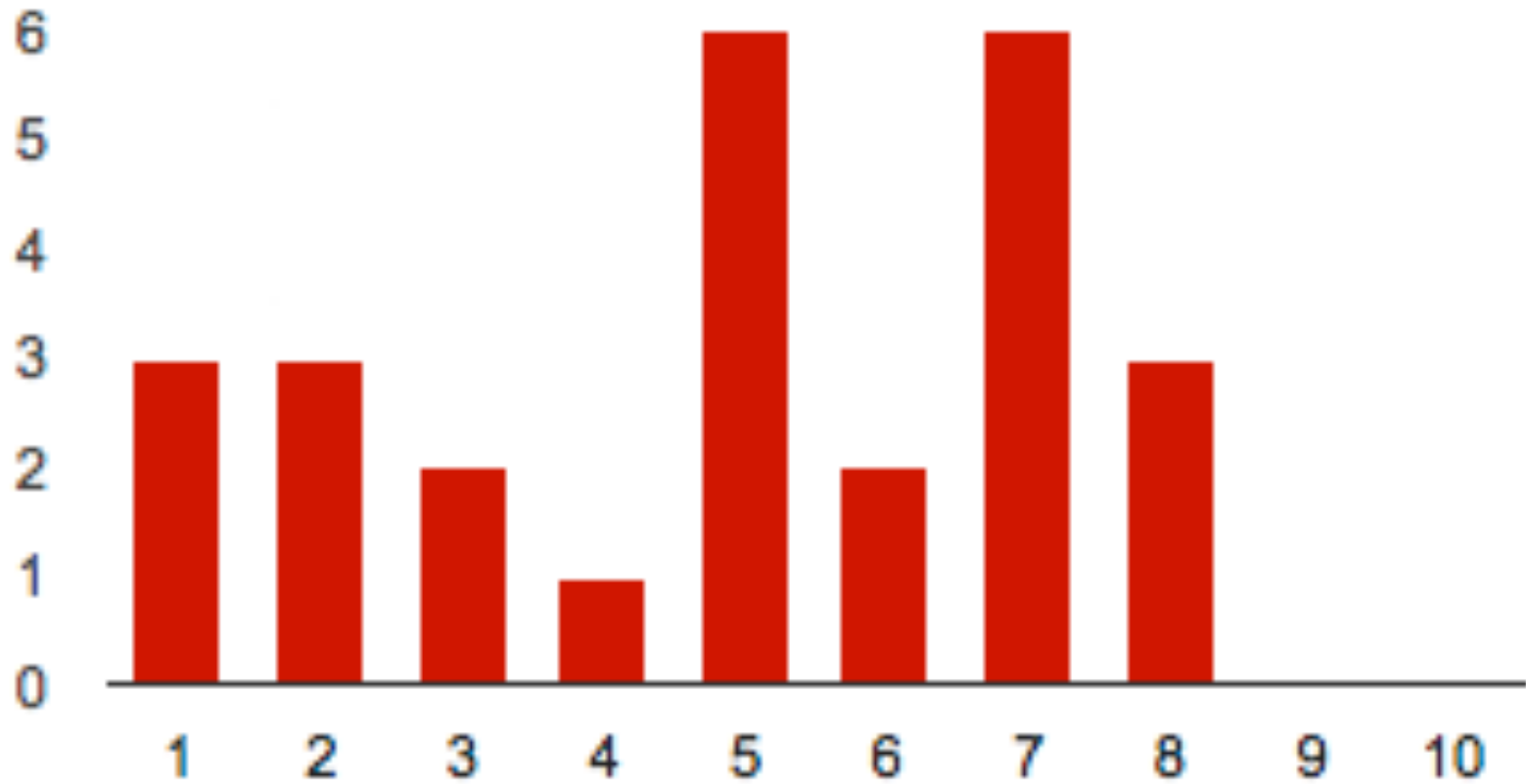
VLDB Papers (56%)
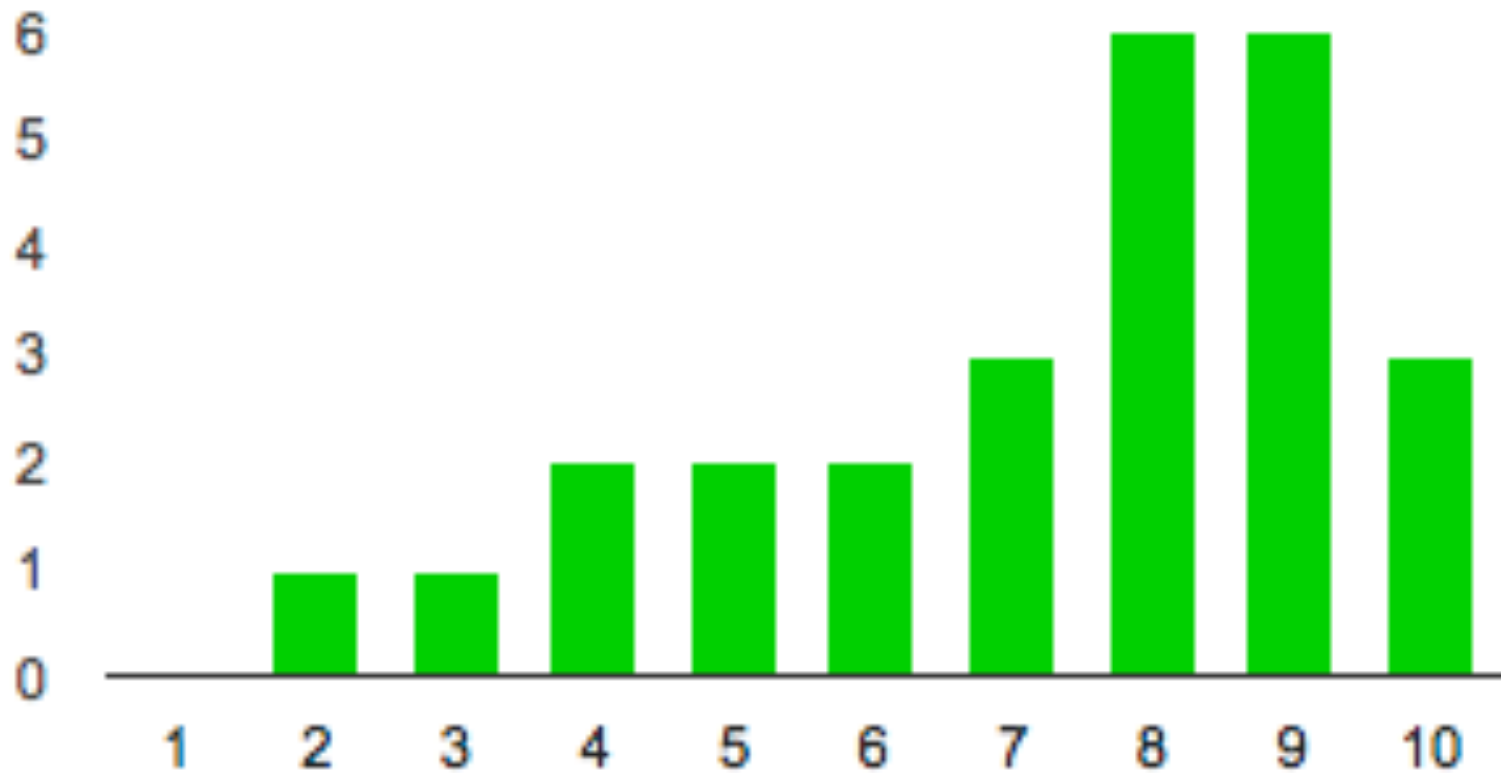
VLDB Demos (10%)
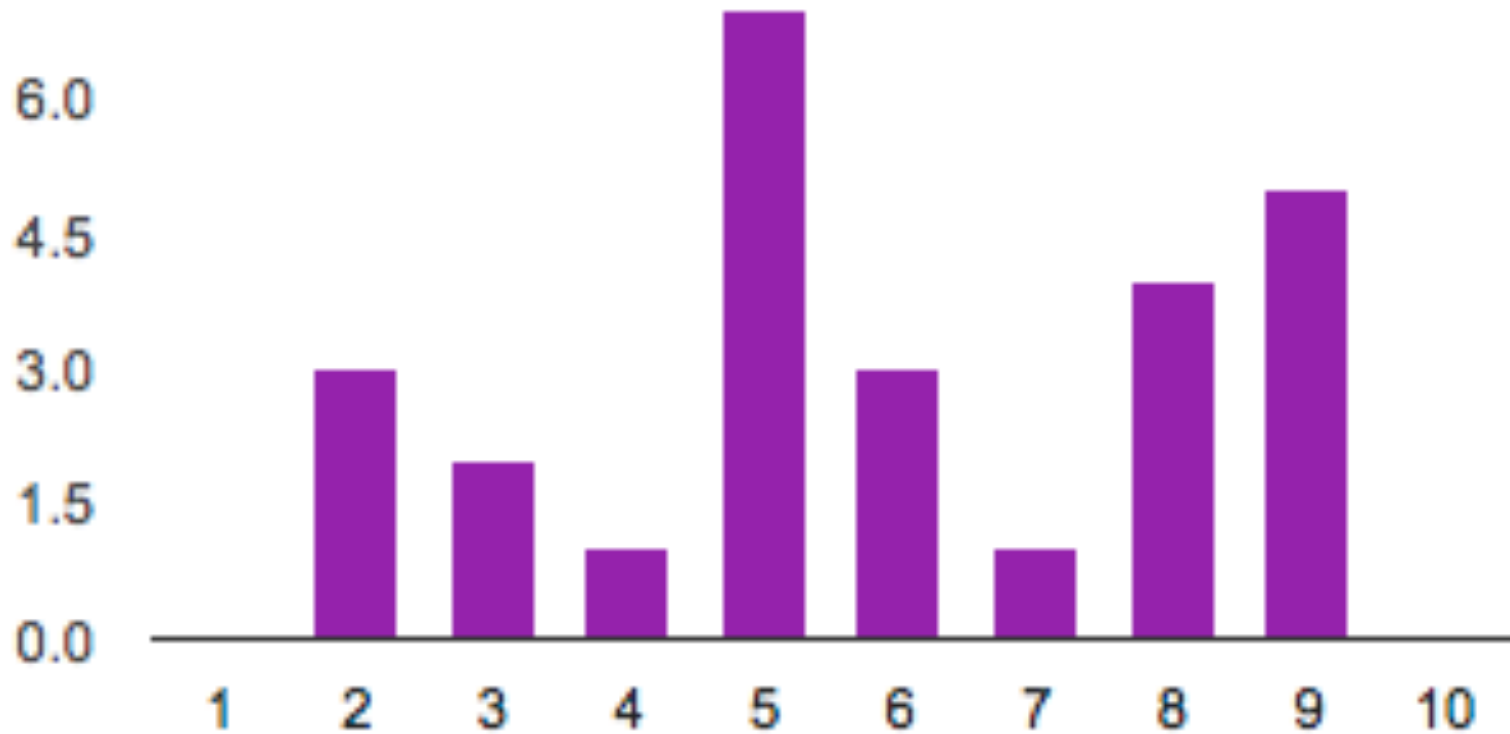
VLDB Workshops (67%)
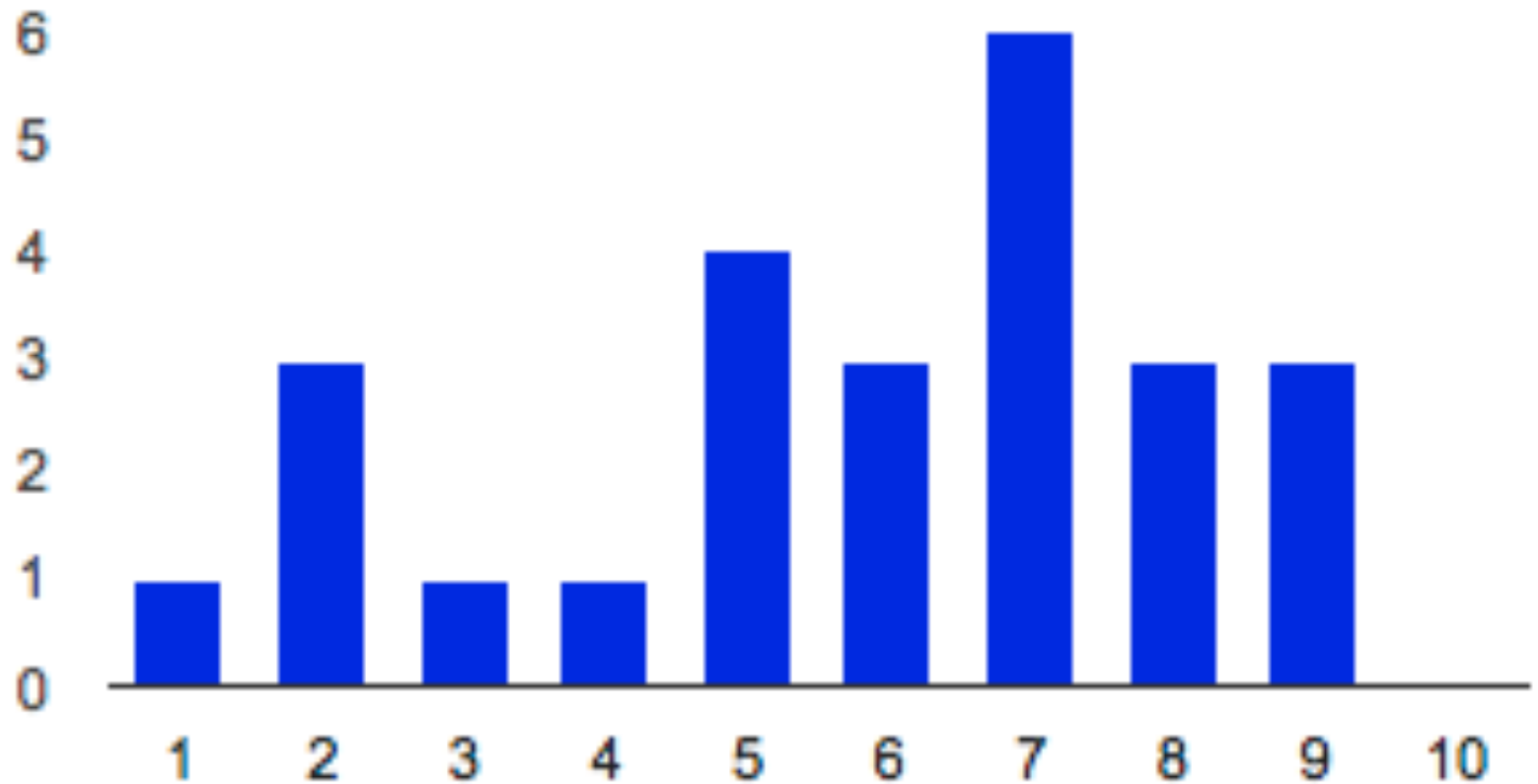
VLDB Tutorials (14%)

VLDB Benefits

Label	Category	Percentage
Career adva…	Professional	48%
Career Adva…	Community / Social	60%
Exposure: se…	Senior, Well-known people	85%
Exposure: pe…	Peers	93%
Exposure: dif…	Different Perspectives	74%
Calibration of…	Database Field	74%
Calibration of…	Your area(s)	60%

0   5   10   15   20
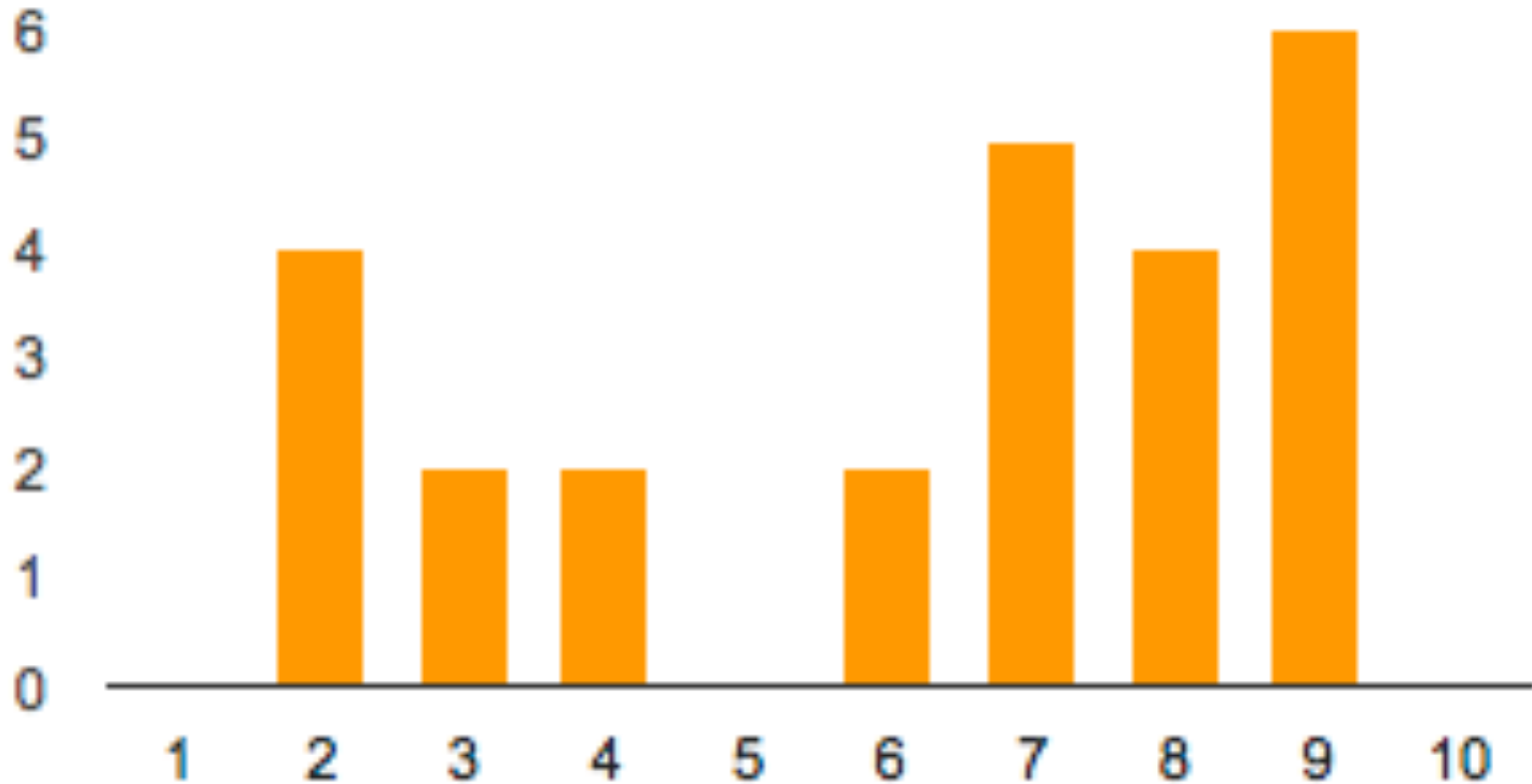
Observations

- VLDB serves us extremely well!

- Enduring professional & personal value

- Hard to get it right ➜ constant experimentation

The Peasants Are Revolting!

Data Trends

- **Paradigm Shift: Value of …**
  - Data fundamental to
    - 1975: science, engineering, business operations
    - 2015: accelerating discovery in every human endeavor
  - Databases: fundamental to
    - 1975: business operations
    - 2015: business operations

- **Data – for your next social event**
  - Brontobytes = 1,000 yotobytes = 10^{15} terebytes
  - Geopbyte = 1,000 brontobytes

*Anand Deshpande, Founder, Persistent Systems

Topic	Total
Constraints	11
Nested Relations	12
Mobile and Personal Databases	12
Clustering	12
Main Memory	13
Tools	14
Compression	15
Active Databases	15
Semantics	18
Sensor Data	19
Integration	21
Data Streams	22
Information Retrieval	23
Data Warehousing	27
Interfaces	29
Join Algorithms	29
Parallel Databases	30
Database Machines	30
Extensibility	30
User Interface	30
Knowledge based Systems	31
Views	31
Aggregates	32
Documents	34
ETL & Data Quality	37
Multimedia	38
Search	39
Integrity	42
Time Series	44
Programming	47
Web	51
OLAP	60
Invited Talks	60
Logic and Databases	62
Transaction Processing	63
Streams	63
Spatial Data	70
Statistics	74
Security and Privacy	74
Concrrency Control	76
Miscellaneous	80
Objects	82
Theory	83
Data Mining	100
Performance	105
Tutorials	109
XML	113
Access Methods	115
Distributed Systems	153
Panel	156
Data Models	160
Applications	203
Systems	265
Industry Sessions	296
Query Processing	366

Evergreen Areas* ('75-'07)

- Query processing (366 papers) … 10% of all papers
- *Systems (265 papers)*
- *Applications (203 papers)*
- Data models (160)
- Distributed systems (153)
- Indexing and access methods (115)
- Performance (105)
- Data Mining (100)

Research Waves* ('75-'07)

- Object Databases (82+12) (88-98)

- Active Databases (15) (88–94)

- DB Machines (30) (79–82)
- Parallel DB (30) (87–95)

*Anand Deshpande, Founder Persistent Systems

- Multi-media (38) (92-98)

- Logic and Databases (62) (79 – 91) peak in 87-88
- Knowledge based systems (31) (84-94)

- Transaction processing (63) (80 – 95)
- Concurrency control (76) (76 – 95)

- Theory (83) (86 – 90)

- Sensor Data (19) (03 +)
- Search (39) (84,93, 98,99,04,06,07)
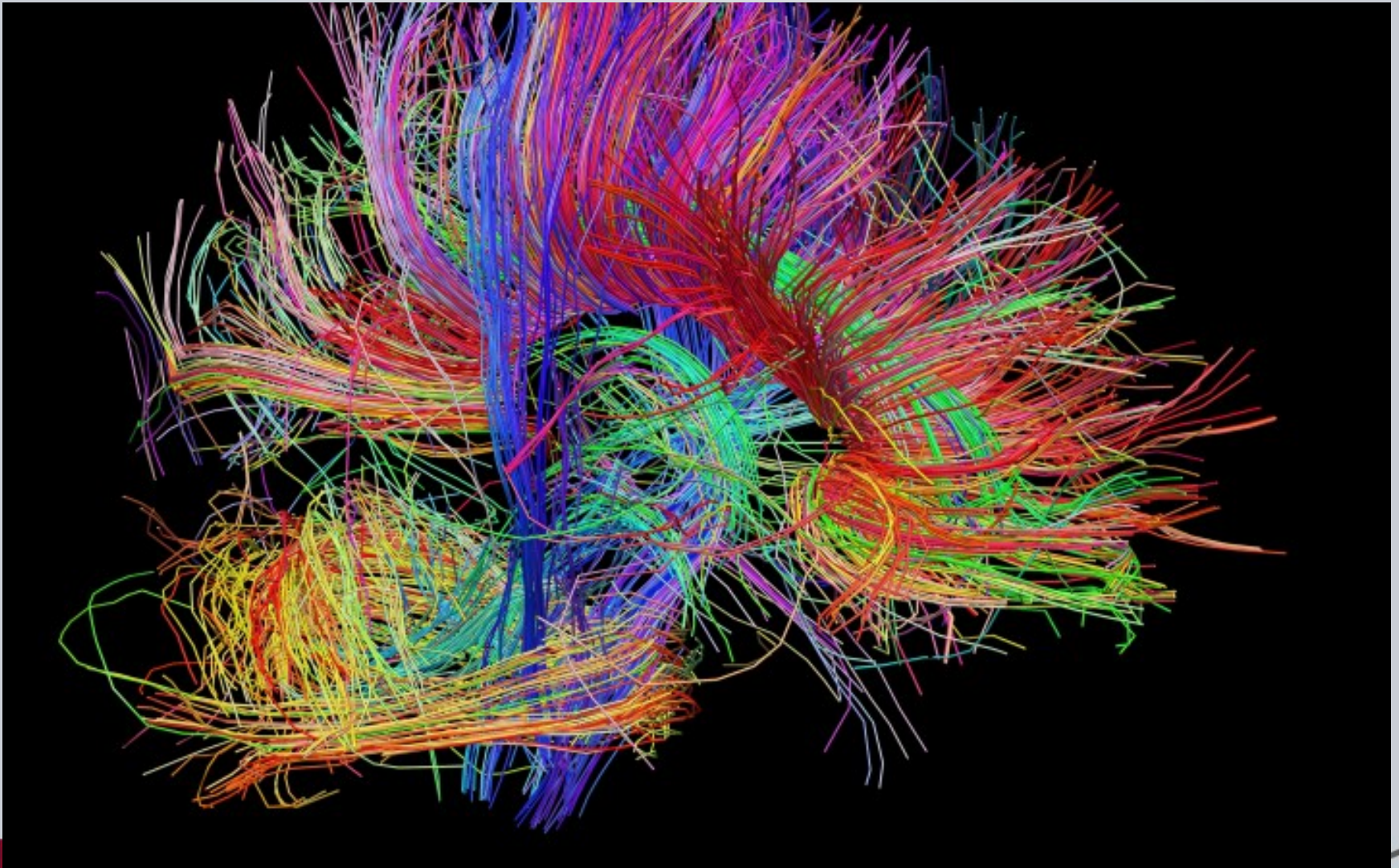
VLDB Future

More Breakthrough Ideas
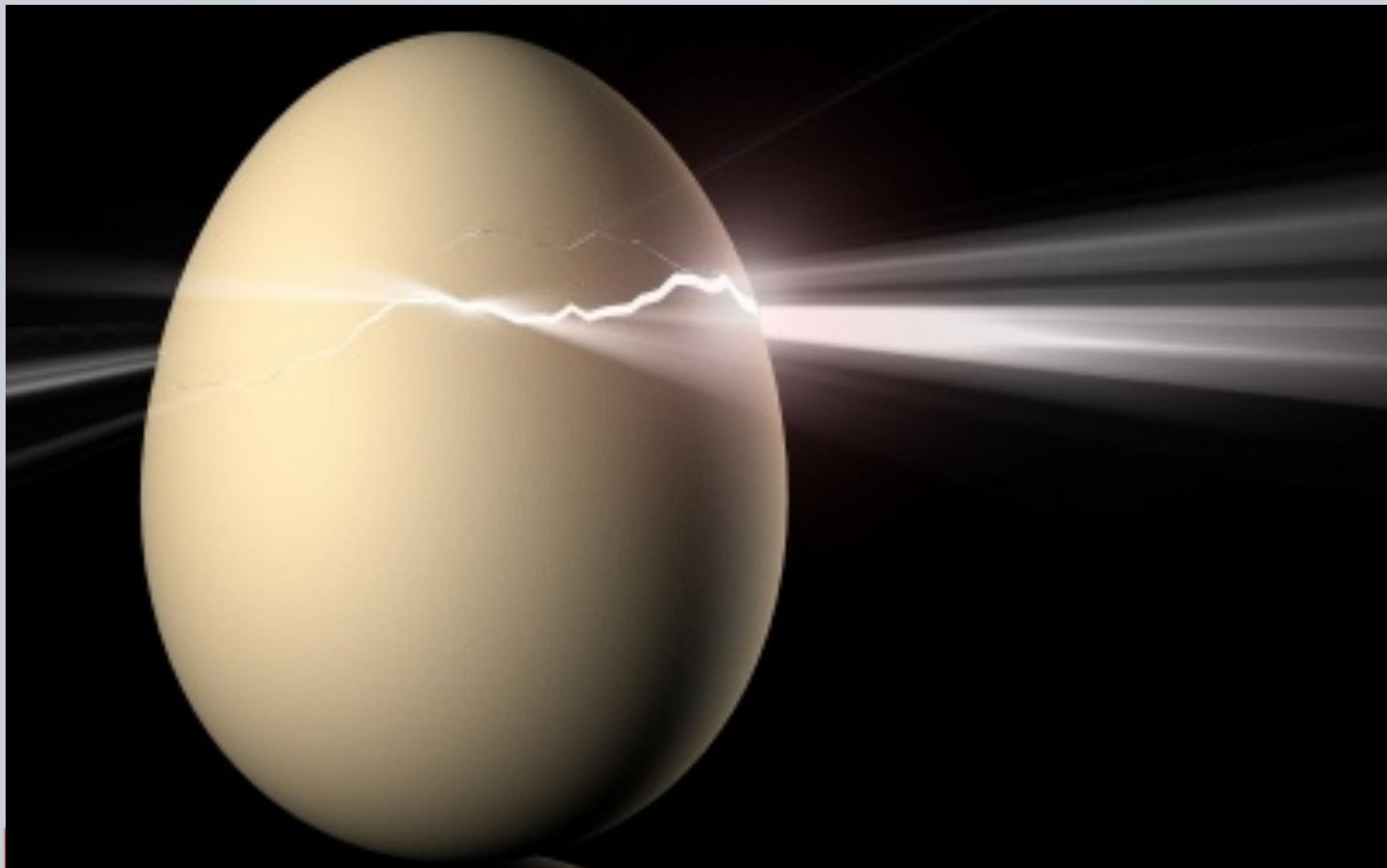
Break More Glass

More Industry & Real Applications

Human Connectome Project

CSAIL

More Industry & Real Applications

Contribution

Challenge

Established Data Management as a

- Discipline

Broadening Strategy 2000
Vision Session ~2000

Example Data Management Systems

First Class (Data Model)	Array DMS*	Probabilistic Inference DMS**
Objects	Arrays	Imprecise data
Operators	Array operators (linear algebra)	Probabilistic Inference (statistics)
Performance		
Competitor	PCA, Hierarchical Data Format 5	MCMC / lifted inference
Improvement X	10 – 1,000	10-1,000
Applications	**Complex analysis**	**Complex analysis**
	Life sciences (genomics, drug discovery), astrophysics, …	Ranking, Top-K, K-means, recognition, mining, synthesis, …

No, thanks anyway

* SciDB, tileDB,   ** ProbDMS, Trio, Deepdive, PrDB, Orion

Technology Paradigm Shift

Economist 7/25/15

"Technology is no longer a vertical industry as it has been understood by everyone for four decades.

Technology is now a horizontal, enabling force throughout the whole economy."

Michael Jordan, 2014

"for the first time in my 25-year career, a topic [Data Science] has emerged that almost everyone in academia feels is on the critical path for their sub-discipline"

Contribution

Challenge

Establish Data Management as a

- Discipline

- Sub-discipline of every discipline*

* cf. Data Science – Michael Jordan

HOPEFULLY NOT ...

The End

Panel

40 Years of VLDB

FADS, FLOPS & SUCCESS STORIES -- MY PERSONAL VIEW

Alfons Kemper
TUM

... Walked into a ready-made position...

- Unlike the other panel members ... not a „pioneer"
- I „grew up" with the relational data model
- Started Computer Science studies in 1977 and took the beauty of the relational model for granted
  - How could there ever have been anything else?
- Over the years the field has matured into one of the most fundamental corner stones of Computer Science
  - As my motivational slides for my students proves ...

How Knowledge (or Exam-Material) expands ...

1-st Edition	2-nd Edition	3-rd Edition	4-th Edition	5-th Edition	6-th Edition	7-th Edition	8-th Edition	9-th Edition	10-th Edition
1996	1997	1999	2001	2004	2006	2009	2011	2013	2015
448 pages	504 pages	504 pages	608 pages	640 pages	672 pages	718 pages	792 pages	848 pages	880 pages

Study fast ... the next (thicker) Edition is coming ☺

What was the (i.e. my) biggest disappointment of ...

- ... the 1980's

- ...the 1990's

- ... the 2000's

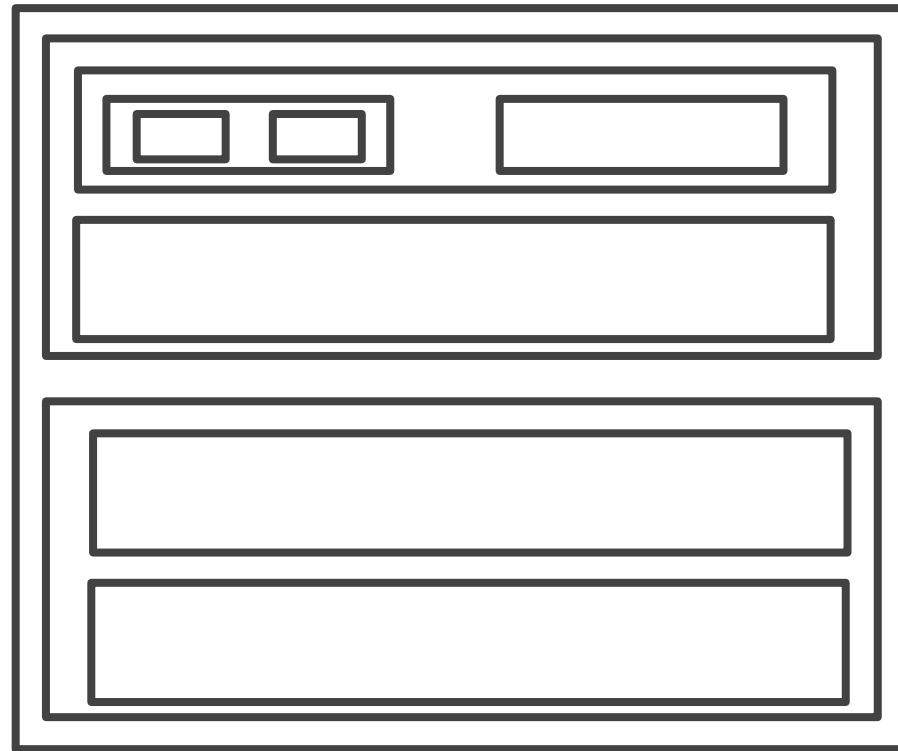# What was the (i.e. my) biggest disappointment of ...

- ... the 1980's

  – The nested relational model (NF2)

- ...the 1990's

- ... the 2000's

Fad'ed away because ...

The world is full of „shared subobjects"
only computer scientists think that a child
has a single parent ;-)

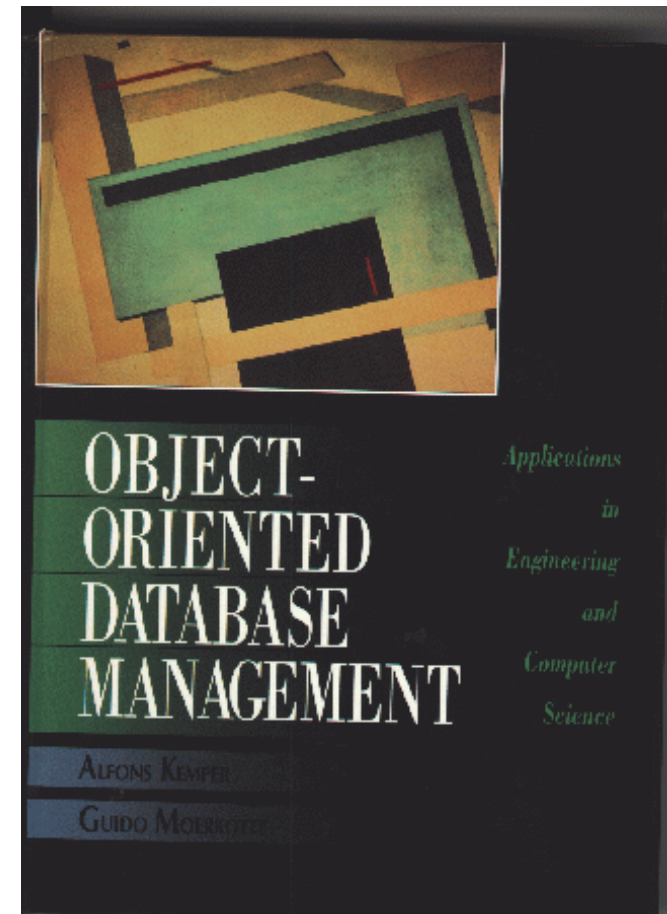# Wishful thinking ... of nested relational advocates

What was the (i.e. my) biggest disappointment of ...

- ... the 1980's

    – The nested relational model (NF2)

- ...the 1990's

    – The object-oriented model

- ... the 2000's

Alfons Kemper&Guido Moerkotte: Object-Oriented Datbase Management, 1993
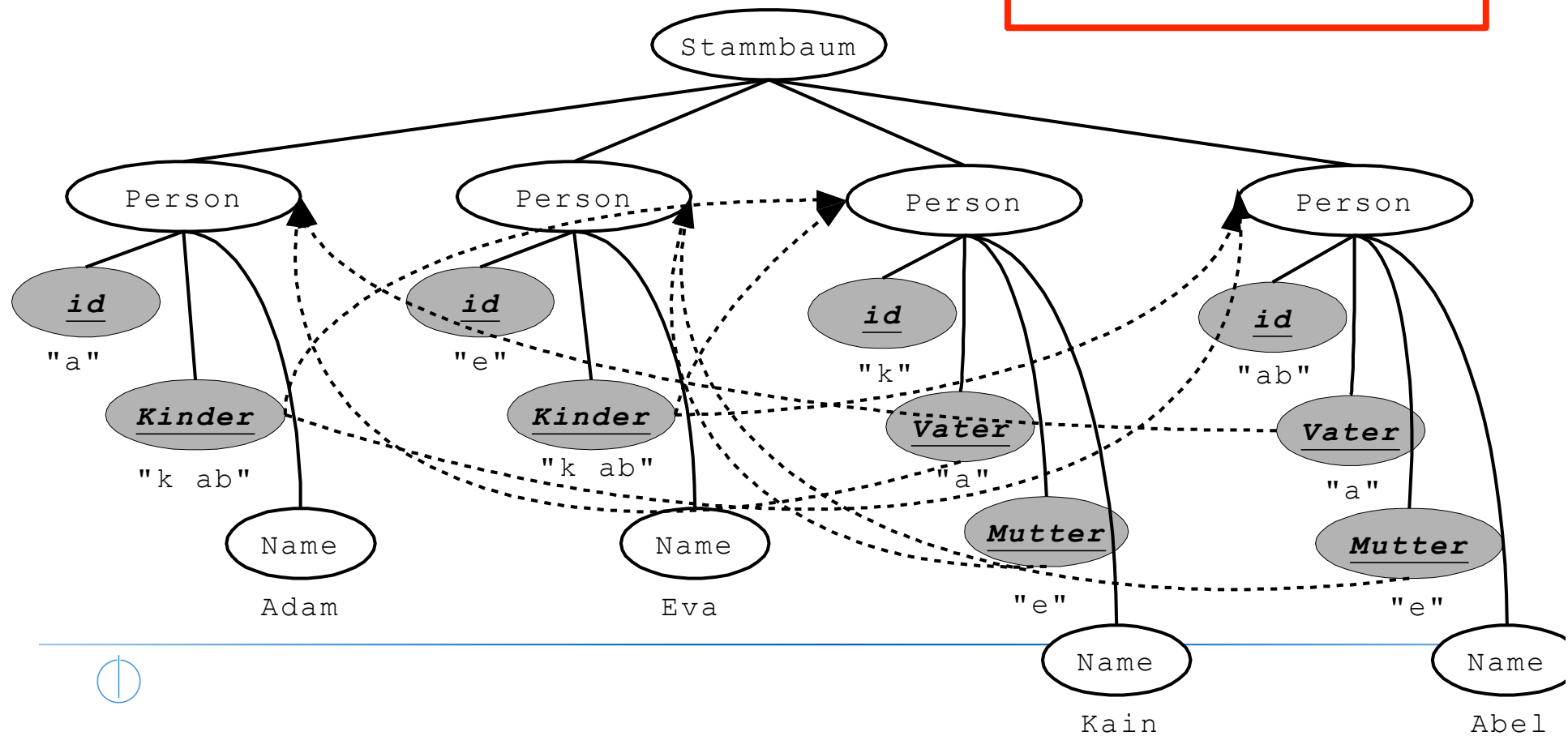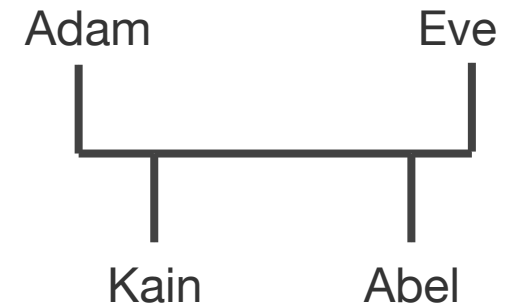
Market-leading Text Book ...

of a non-existing market ;-(

What was the (i.e. my) biggest disappointment of ...

- ... the 1980's

  – The nested relational model (NF2)

- ...the 1990's

  – The object-oriented model

- ... the 2000's

  – XML as a data model

XML Model of a Family „Tree" ...

What was the (i.e. my) biggest disappointment of ...

- ... the 1980's
  - The nested relational model (NF2)

- ...the 1990's
  - The object-oriented model

- ... the 2000's
  - XML as a data model

- ... the 2010's
 - NoSQL K/V Stores?
  - MapReduce?

We should work on problems that really matter ...

- ... in real life

- ... not only on „nice-looking" problems with LPU results

- Now is an exciting time to be a database researcher (again)

  – Big Data gives us public attention

  – Hardware vendors give us affordable „super"-computers to build systems with awesome performance

Background: Hardware Development

- Dell PowerEdge 920
  - 4 Intel Xeon E7-8880
    - 4 X 15 cores ➜60 cores ➜ 120 Hyper-Threads
    - 1.5 TB
    - $ 40,000

  - 6 TB is possible ➜ exceeds (almost) all OLTP applications

HyPer-DB: One System for Everything – OLTP and OLAP on Brawny and Wimpy Devices
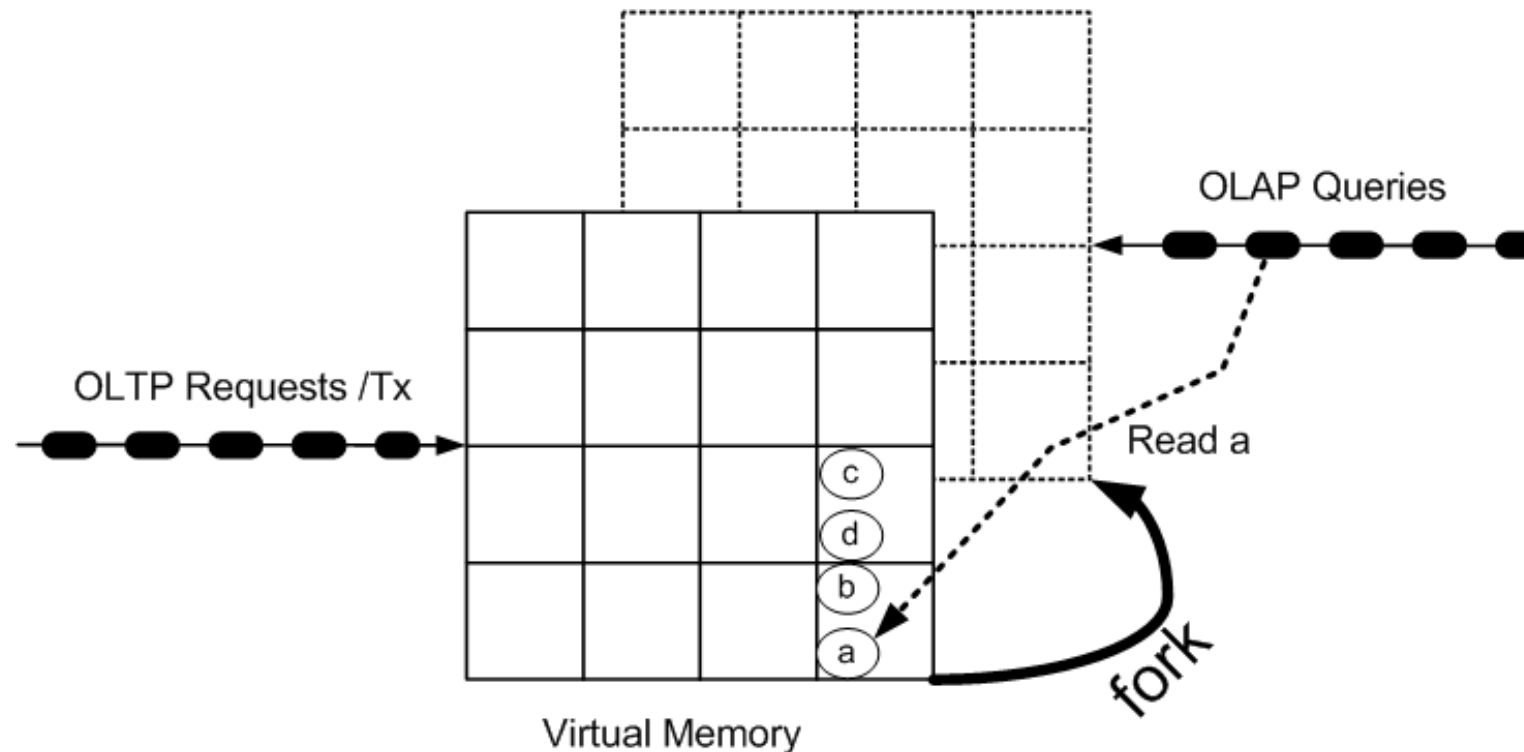
Joint Work with Thomas Neumann (who received the **VLDB Early Career Innovations Award** last year) and many PhD students

VLDB 2017 in Munich
Monday, 28.08.2017 – Friday, 01.09.2017
(plan ahead -- see you there)

Serendipity, Oracle and Whines About VLDB

by

Michael Stonebraker

DBg Database Group
MIT Computer Science and Artificial Intelligence Lab

Serendipity Created the RDBMS Market

- ## The VAX hardware market appeared
  - Department level (not data center) machines
  - A green field of opportunity
- ## Competing products (e.g. IDMS) were written in IBM assembler
  - And could not be ported to the VAX
  - Leaving the VAX market to Ingres, Oracle, …
  - Allowing these very immature products to get a toe-hold (and they were very very immature in the early 1980's)

Serendipity Created the RDBMS Market

- Codasyl did not put a relational front end on top of their proposal
  - To co-opt Codd et. al.
  - They were asleep at the switch
- Project Eagle failed(!!!!)
  - An IBM attempt to put a relational interface on top of IMS
  - If it had succeeded, SQL would just be a user-interface on top of IMS. There would be no DB2!
  - Thankfully IMS screwed up logical data bases badly

Serendipity Allowed Oracle to Beat Ingres

- Ingres (with a substantially better product) would have overtaken Oracle in 1984 except for...

- The announcement of DB2 by IBM

  - Who did not fix the known issues with SQL

- Oracle (with the same SQL) leaped ahead relative to everybody else (without SQL)

  - And then adopted the "America Cup" strategy to stay ahead

Oracle (in the 1980s)

- "Scorched earth" tactics
- Salesmen lied regularly
- Ellison confused present tense and future tense
- No QA to speak of

- Oracle (and Microsoft) are largely responsible for the dismal state of system software
  - Nobody will take early releases
  - Customers dislike the vendors
  - Crappy customer support

DBg Database Group
MIT Computer Science and Artificial Intelligence Lab

Current Whine (1)

- VLDB is largely composed of least publishable units (LPUs)
- On topics that only 3 people on the planet are interest in

- Should require every VLDB paper to have a real world use case -- i.e. somebody who cares about the solution in the paper

Current Whine (2)

- VLDB should not be a 12 hour airplane ride from everywhere
- Should probably go the way of the National Computer Conference
  - Decompose into a collection of more specialized conferences

DBg Database Group
MIT Computer Science and Artificial Intelligence Lab

Current Whine (3)

- ACM publications are totally broken

- Hiring Asst Professors should be based on a resume of not more than 3 papers
- Tenure decisions should be based on a resume of not more than 7 papers

- Current paper diarrhea has to stop