The Age of Big Data

Prof. Mulhim Al-Doori

Contents

Contents

Introduction: Explosion in Quantity of Data

1946		2012
Eniac		LHC
X 6000000	=	1 (40 TB/S)

Air Bus A380

- 1 billion line of code
- each engine generate 10 TB every 30 min

640TB per Flight

Twitter Generate approximately 12 TB of data per day

New York Stock Exchange 1TB of data everyday

storage capacity has doubled roughly every three years since the 1980s

Our Data-driven World

- Science
  - Data bases from astronomy, genomics, environmental data, transportation data, …
- Humanities and Social Sciences
  - Scanned books, historical documents, social interactions data, new technology like GPS …
- Business & Commerce
  - Corporate sales, stock market transactions, census, airline traffic, …
- Entertainment
  - Internet images, Hollywood movies, MP3 files, …
- Medicine
  - MRI & CT scans, patient records, …

Our Data-driven World

- Fish and Oceans of Data

What we do with these amount of data?

Ignore

Big Data Characteristics

How big is the Big Data?

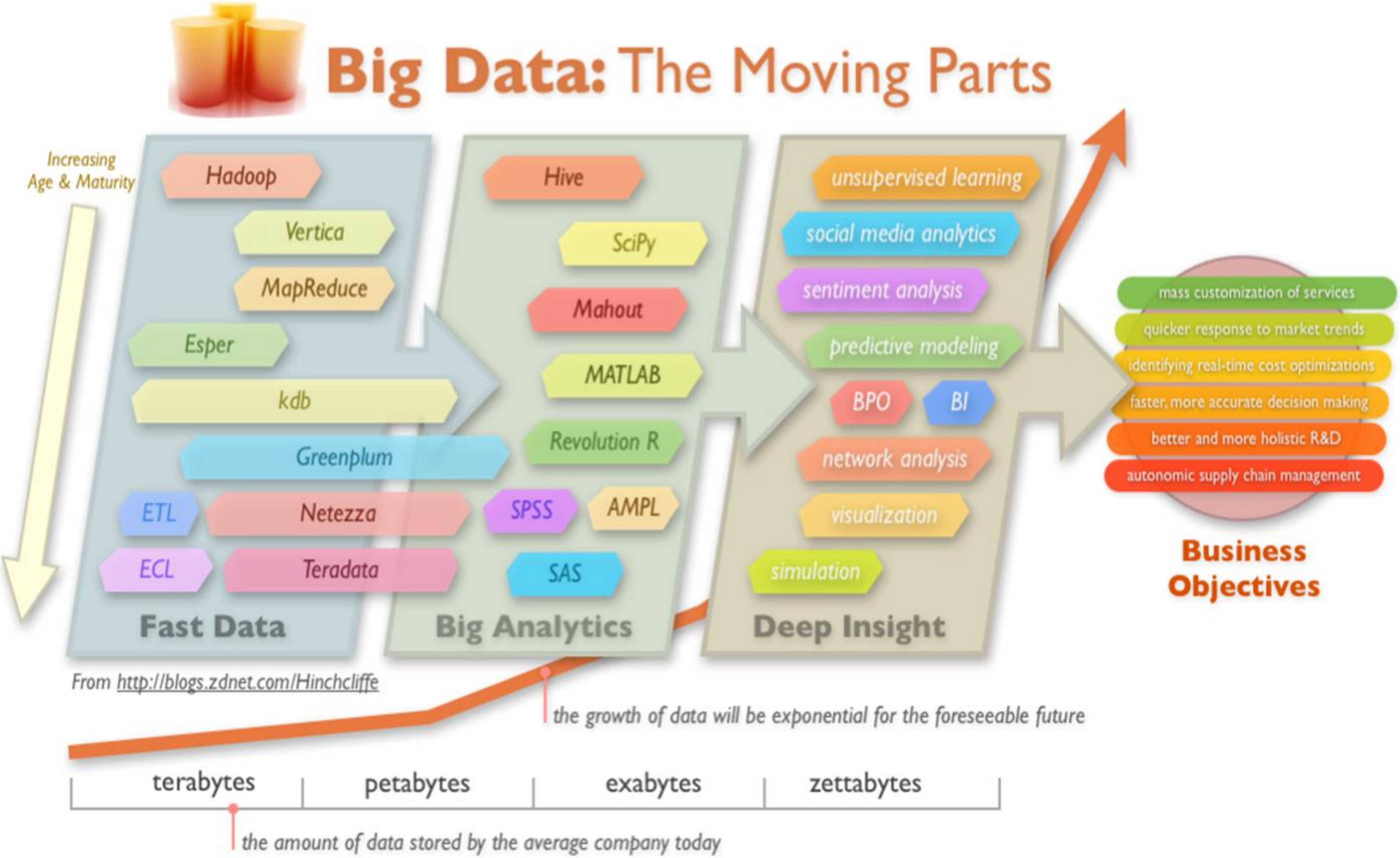
- What is big today maybe not big tomorrow

- Any data that can challenge our current technology in some manner can consider as Big Data
  - Volume
  - Communication
  - Speed of Generating
  - Meaningful Analysis

Big Data Vectors (3Vs)

"Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization”
Gartner 2012

Big Data Technology



Big Data Characteristics

Big Data Vectors (3Vs)

- high-volume

  amount of data

- high-velocity

  Speed rate in collecting or acquiring or generating or processing of data

- high-variety

  different data type such as audio, video, image data (mostly unstructured data)

Cost Problem (example)

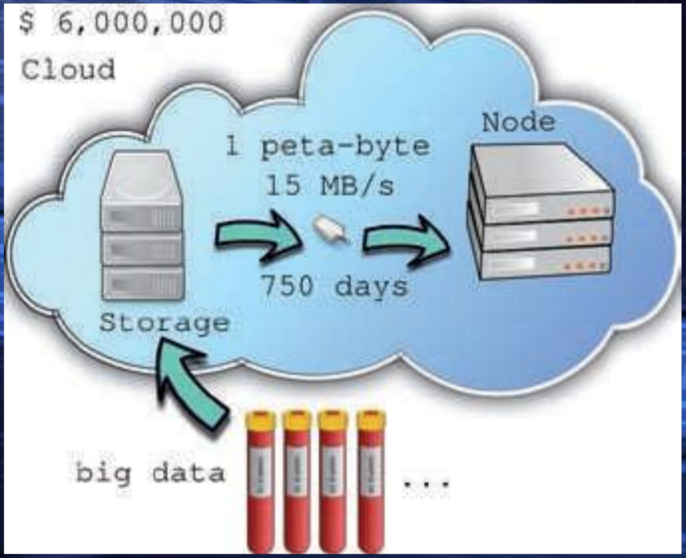
Cost of processing 1 Petabyte of data with 1000 node ?

1 PB = 10^{15} B = 1 million gigabytes = 1 thousand terabytes

- 9 hours for each node to process 500GB at rate of 15MB/S
- 15*60*60*9 = 486000MB ~ 500 GB
- 1000 * 9 * 0.34$ = 3060$ for single run

- 1 PB = 1000000 / 500 = 2000 * 9 = 18000 h /24 = 750 Day

- The cost for 1000 cloud node each processing 1PB
  2000 * 3060$ = 6,120,000$



Importance of Big Data

- Government

    In 2012, the Obama administration announced the Big Data Research and Development Initiative

    84 different big data programs spread across six departments

- Private Sector

    - Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data

    - Facebook handles 40 billion photos from its user base.

    - Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide

- Science

    - Large Synoptic Survey Telescope will generate 140 Terabyte of data every 5 days.

    - Large Hardon Colider 13 Petabyte data produced in 2010

    - Medical computation like decoding human Genome

    - Social science revolution

    - New way of science (Microscope example)

Importance of Big Data

- Job

- The U.S. could face a shortage by 2018 of 140,000 to 190,000 people with "deep analytical talent" and of 1.5 million people capable of analyzing data in ways that enable business decisions. (McKinsey & Co)
- Big Data industry is worth more than $100 billion

growing at almost 10% a year (roughly twice as fast as the software business)

- Technology Player in this field
- Oracle
  - Exadata
- Microsoft
  - HDInsight Server
- IBM
  - Netezza

Usage Example in Big Data

- Moneyball: The Art of Winning an Unfair Game

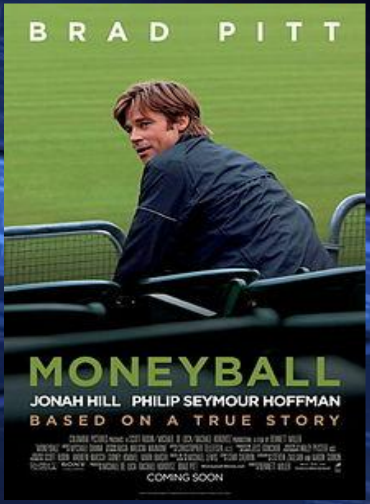
Oakland Athletics baseball team and its general manager Billy Beane

- Oakland A's' front office took advantage of more analytical gauges of player performance to field a team that could compete successfully against richer competitors in MLB

- Oakland approximately $41 million in salary,
New York Yankees, $125 million in payroll that same season.
Oakland is forced to find players undervalued by the market,

- Moneyball had a huge impact in other teams in MLB

And there is a moneyball movie!!!!!



Usage Example of Big Data

US 2012 Election

- predictive modeling
- mybarackobama.com
- drive traffic to other campaign sites
 Facebook page (33 million "likes")
 YouTube channel (240,000 subscribers
 and 246 million page views).
- a contest to dine with Sarah Jessica Parker
- Every single night, the team ran 66,000
computer simulations, Reddit!!!
- Amazon web services

- data mining for
individualized ad targeting

- Orca big-data app

- YouTube channel(23,700 subscribers
 and 26 million page views)

- Ace of Spades HQ

Usage Example in Big Data

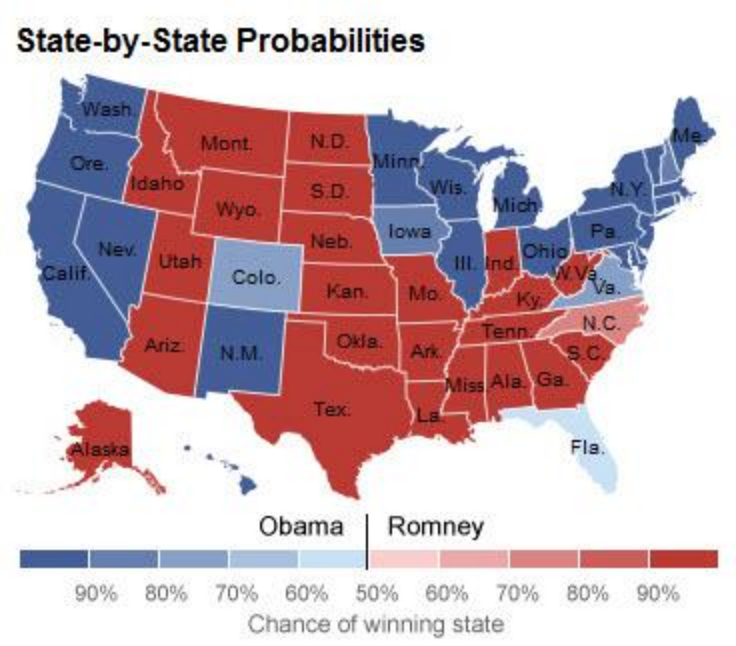
Data Analysis prediction for US 2012 Election

Drew Linzer, June 2012
332 for Obama,
206 for Romney

media continue reporting the race as very tight

Nate Silver's, Five thirty Eight blog
Predict Obama had a 86% chance of winning
Predicted all 50 state correctly

Sam Wang, the Princeton Election Consortium
The probability of Obama's re-election
at more than 98%



Some Challenges in Big Data

- Big Data Integration is Multidisciplinary
  - Less than 10% of Big Data world are genuinely relational
  - Meaningful data integration in the real, messy, schema-less and complex Big Data world of database and semantic web using multidisciplinary and multi-technology methods

- The Billion Triple Challenge
  - Web of data contain 31 billion RDf triples, that 446million of them are RDF links, 13 Billion government data, 6 Billion geographic data, 4.6 Billion Publication and Media data, 3 Billion life science data
  - BTC 2011, Sindice 2011

- The Linked Open Data Ripper
  - Mapping, Ranking, Visualization, Key Matching, Snappiness

- Demonstrate the Value of Semantics: let data integration drive DBMS technology
  - Large volumes of heterogeneous data, like link data and RDF

Other Aspects of Big Data

Six Provocations for Big Data

1- Automating Research Changes the Definition of Knowledge

2- Claim to Objectively and Accuracy are Misleading

3- Bigger Data are not always Better data

4- Not all Data are equivalent

5- Just because it is accessible doesn’t make it ethical

6- Limited access to big data creatrs new digital divides

Other Aspects of Big Data

- Five Big Question about big Data:

1- What happens in a world of radical transparency, with data widely available?

2- If you could test all your decisions, how would that change the way you compete?

3- How would your business change if you used big data for widespread, real time customization?

4- How can big data augment or even replace Management?

5-Could you create a new business model based on data?

Platforms for Large-scale Data Analysis

- **Parallel DBMS technologies**
  - Proposed in late eighties
  - Matured over the last two decades
  - Multi-billion dollar industry: Proprietary DBMS Engines intended as Data Warehousing solutions for very large enterprises
- **Map Reduce**
  - pioneered by Google
  - popularized by Yahoo! (Hadoop)

Implementation of Big Data

MapReduce

- Overview:
  - Data-parallel programming model
  - An associated parallel and distributed implementation for commodity clusters
- Pioneered by Google
  - Processes 20 PB of data per day
- Popularized by open-source Hadoop
  - Used by Yahoo!, Facebook, Amazon, and the list is growing …

Parallel DBMS technologies

- Popularly used for more than two decades
  - Research Projects: Gamma, Grace, …
  - Commercial: Multi-billion dollar industry but access to only a privileged few
- Relational Data Model
- Indexing
- Familiar SQL interface
- Advanced query optimization
- *Well understood and studied*

MapReduce Advantages

- Automatic Parallelization:
  - Depending on the size of RAW INPUT DATA → instantiate multiple MAP tasks
  - Similarly, depending upon the number of intermediate <key, value> partitions → instantiate multiple REDUCE tasks
- Run-time:
  - Data partitioning
  - Task scheduling
  - Handling machine failures
  - Managing inter-machine communication
- Completely transparent to the programmer/analyst/user

Map Reduce vs Parallel DBMS

	Parallel DBMS	MapReduce
Schema Support	✓	Not out of the box
Indexing	✓	Not out of the box
Programming Model	Declarative (SQL)	Imperative (C/C++, Java, …) **Extensions through Pig and Hive**
Optimizations (Compression, Query Optimization)	✓	Not out of the box
Flexibility	Not out of the box	✓
Fault Tolerance	Coarse grained techniques	✓

Zeta-Byte Horizon

- As of 2009, the entire World Wide Web was estimated to contain close to 500 exabytes. This is a half zettabyte

- the total amount of global data is expected to grow by 48% annually to 7.5 zettabytes during 2015.

x50

2012 ⟷ 2020

Wrap Up

1 What is a Spatial Database System?

Requirement: Manage data related to some *space*.

Spaces: 2D or"2.5D“ or 3D

geographic space (surface of the earth, at large or small scales)

→ GIS, LIS, urban planning, ...

- the universe

→ astronomy

a VLSI design

a model of the brain (or someone's brain)

→ medicine

a molecule structure

→ biological research

Characteristic for the supporting technology: capability of managing large collections of relatively simple geometric objects

Terms:

pictorial database system

image

geometric

geographic

spatial

collections of *objects* in some space

clear identity, location, extent

raster images of some space

A database may contain

spatial database system

image database system

- analysis,
- feature extraction

- a spatial DBMS:
  (1) A spatial database system is a database system
  (2) It offers *spatial data types* in its data model and query language
  (3) It supports spatial data types in its implementation, providing at least *spatial indexing* and efficient algorithms for *spatial join*.

- Focus : describe fundamental problems and known solutions in a coherent manner.
  2 Modeling
  3 Querying
  4 Tools for Implementation: Data Structures and Algorithms
  5 System Architecture

Modeling

1. What needs to be represented?
2. Discrete Geometric Bases
3. Spatial Data Types / Algebras
4. Spatial Relationships
5. Integrating Geometry into the DBMS Data Model

1. What needs to be represented?

- Two views:

  (i) objects in space

  (ii) space itself

- (i) Objects in space
  - city Berlin, …, population: 3 500 000, city area:
    river Rhine, …, route:

(ii) Space

Statement about every point in space (↔ raster images)

- land use maps (“thematic maps”)
- partitions into states, counties, municipalities, …

We consider:

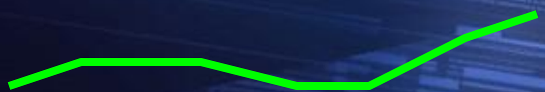
1. modeling single objects
2. modeling spatially related collections of objects

1. Basic abstractions for modeling single objects:

- *point* *city*

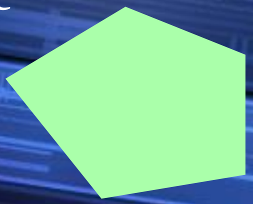
geometric aspect of an object, for which only its *location* in space, but not the *extent*, is relevant

- *line* (polyline)



river cable highway

moving through space, connections in space

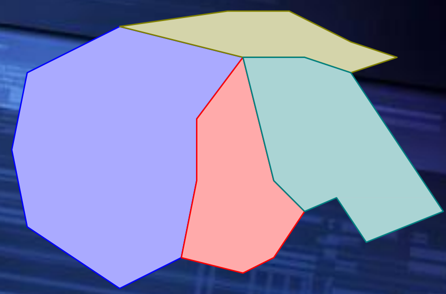


- *Region*

forest lake city

- abstraction of an object with extent

- *Partition*



- *land use*
- *Districts*
- *land ownership*
- *“environments” of points Voronoi diagram*

- Spatially embedded *network* (graph)

- *highways, streets*
- *railways, public*
- *Transport*
- *Rivers*
- *electricity, phone*

Others:

- nested partitions
- digital terrain models

Organizing the Underlying Space: Discrete Geometric Bases

Is Euclidean geometry a suitable base for modeling?

Problem: space is continuous

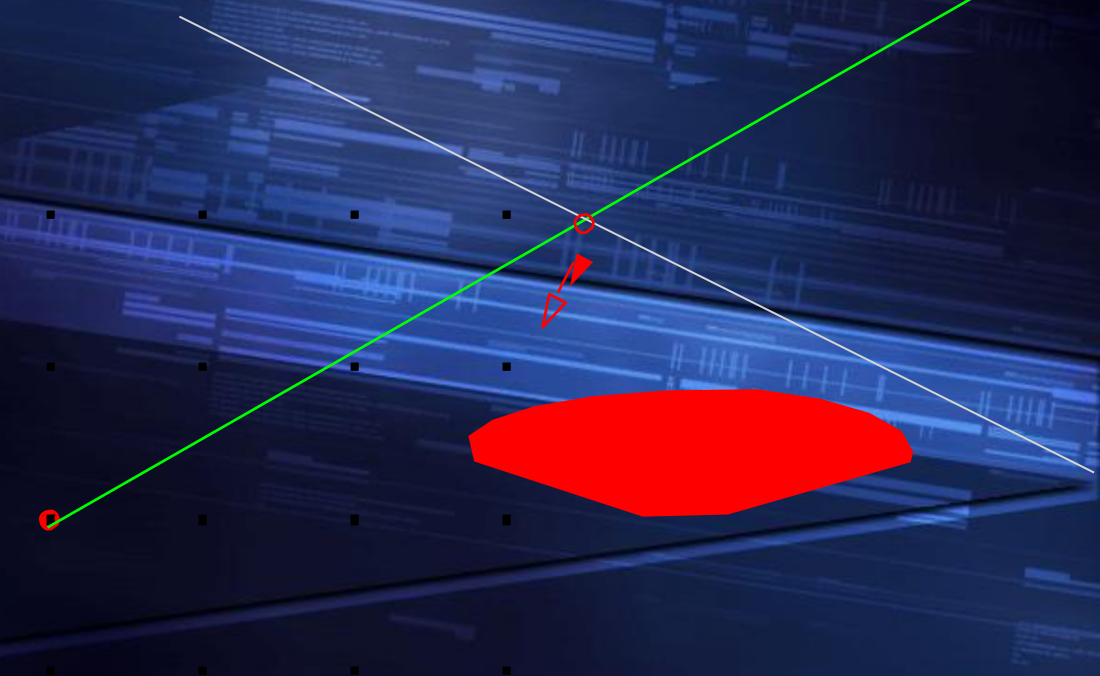
computer numbers are discrete

$p = (x, y) \in |R^2$

$p = (x, y) \in$ real $\times$ real

Is D on A

Is D Properly contained in the specified area



- Goal: Avoid computation of any new intersection points within geometric operations

Definition of geometric types and operations

____Geometric Bases

Treatment of numeric problems upon updates of the geometric basis

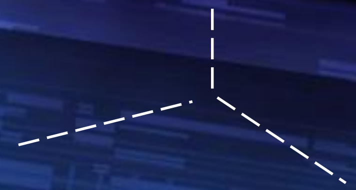
Two approaches:

- *Simplicial complexes*
- *Realms*

Simplicial Complexes

d-simplex: minimal object of dimension *d*

0-simplex



3-simplex

1-simplex

2-simplex

d-simplex consists of *d*+1 simplices of dimension *d*-1.

Components of a simplex are called *faces*.

Simplicial complex: finite set of simplices such that the intersection of any two sim- plices is a face.

Realms

- *Realm* (intuitive notion): Complete description of the geometry (all points and lines) of an application.
- *Realm* (formally): A finite set of points and line segments defined over a grid such that:

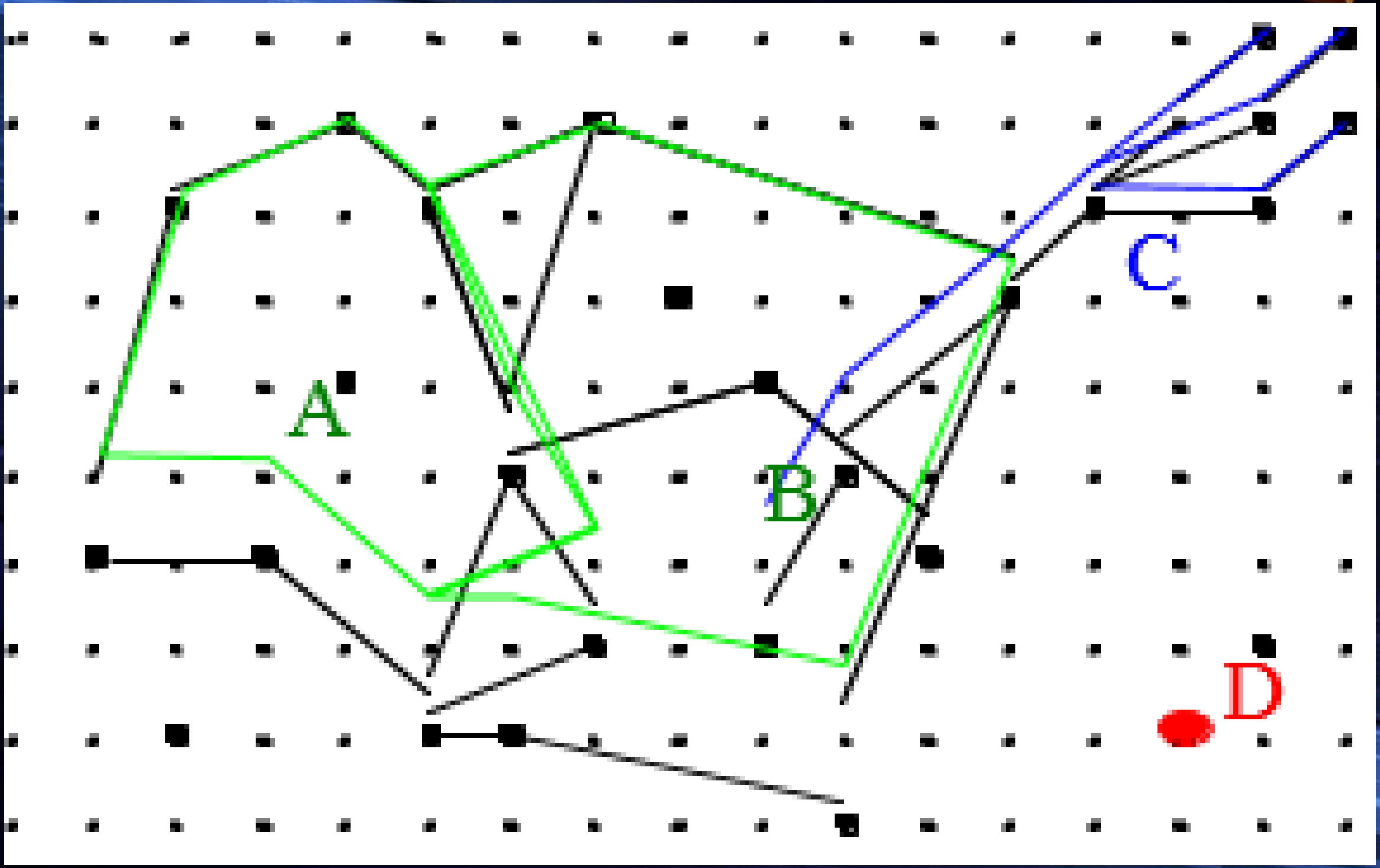
1-each point or end point of a segment is a grid point

2-each end point of a segment is also a point of the realm

3- no realm point lies *within* a segment

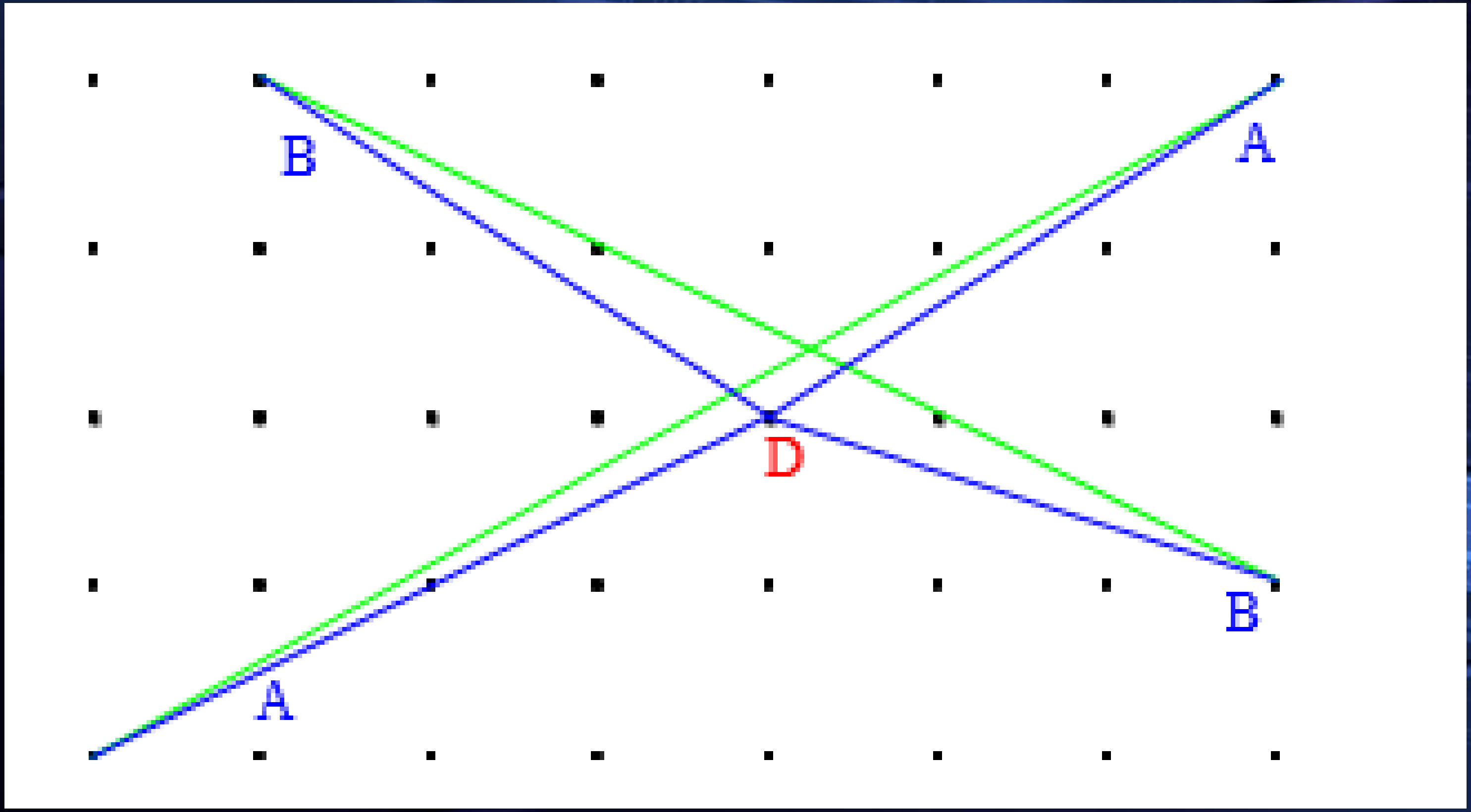
4-any two distinct segments do neither intersect nor over- lap

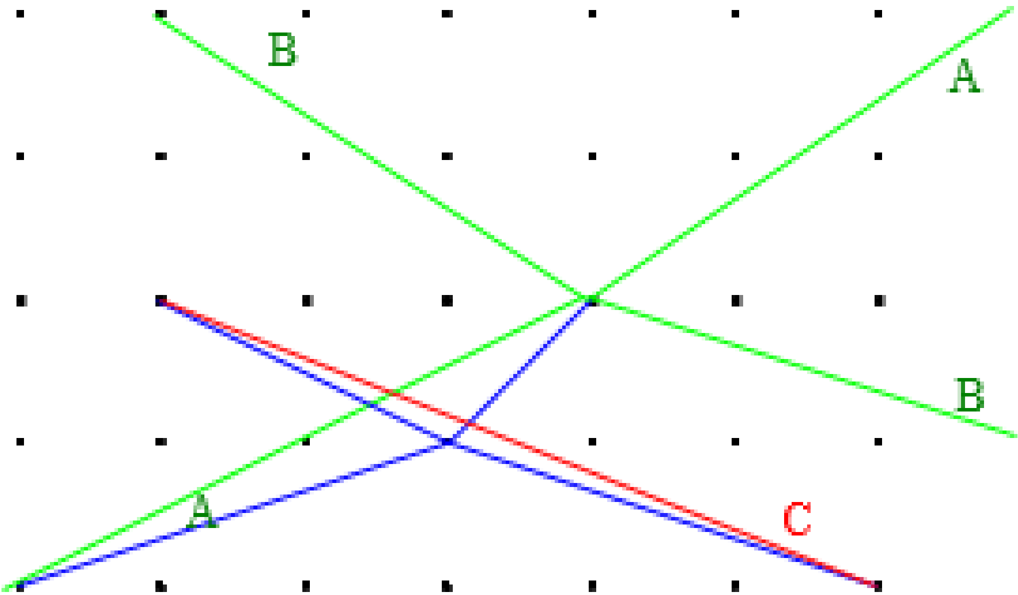
Realms

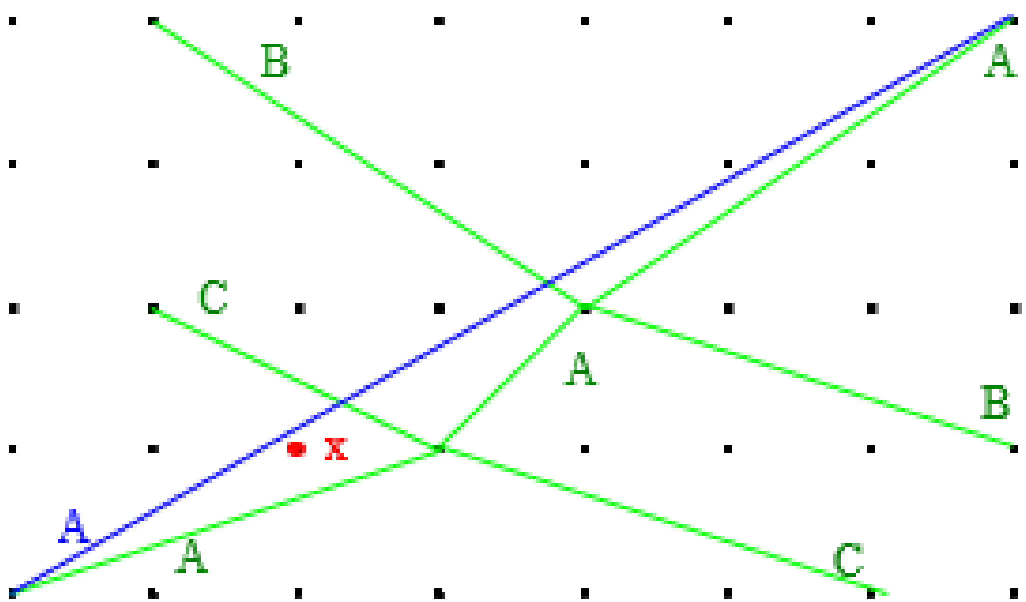


Numeric problems are treated *below* the realm layer:

Application data are sets of points and *intersecting* line seg- ments. Need to insert a segment intersecting other segments. Basic idea: slightly distort both segments.

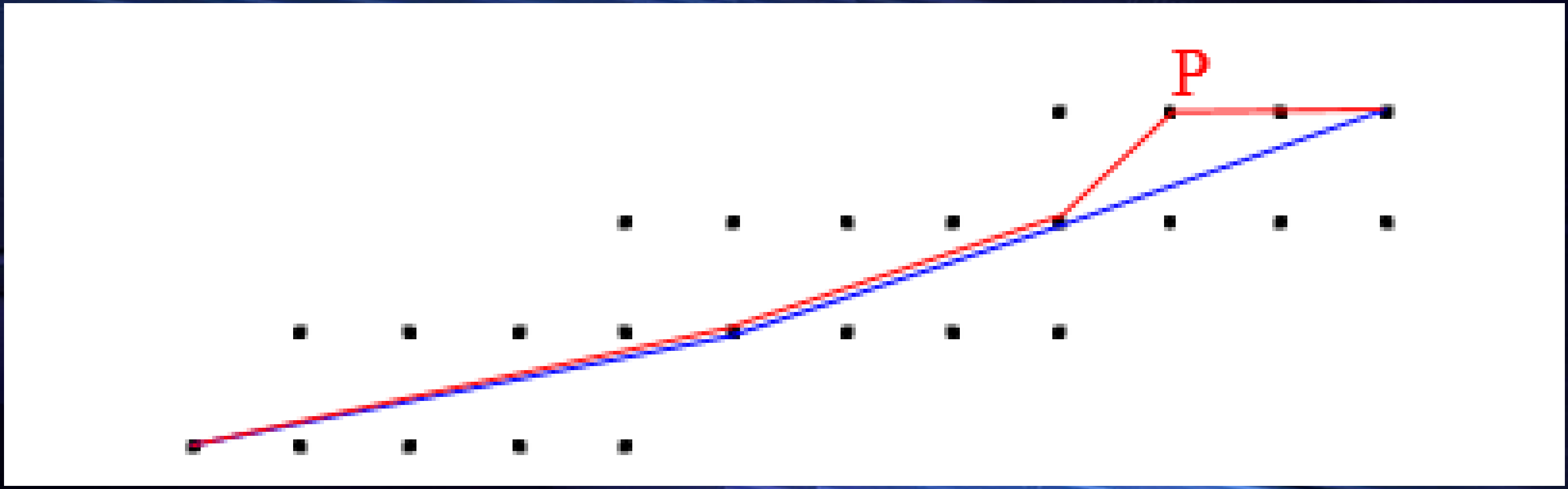






Segments can move! Point *x* is now on the wrong side of *A*!

Concept of Greene & Yao (1986): *Redraw* segments within their *envelope*.

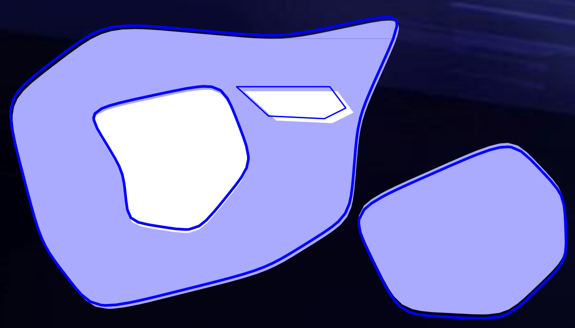


Segments are “captured” within their envelope; can never cross a grid point.

3. Spatial Data Types / Algebras SDT

- "general structure" of values ↔ closed under set operations on the underlying point sets

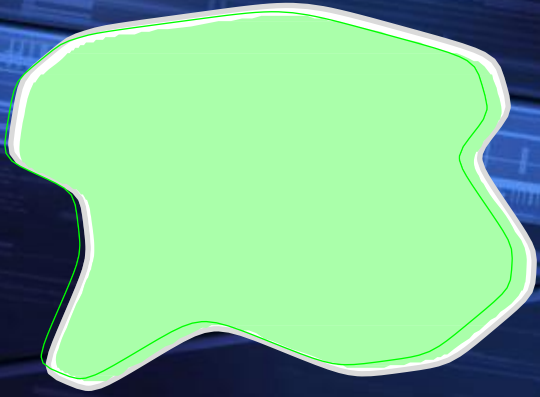
- precise formal definition of SDT values and functions
- definition in terms of finite precision arithmetics
- support for geometric consistency of spatially related objects



4. Spatial Relationship

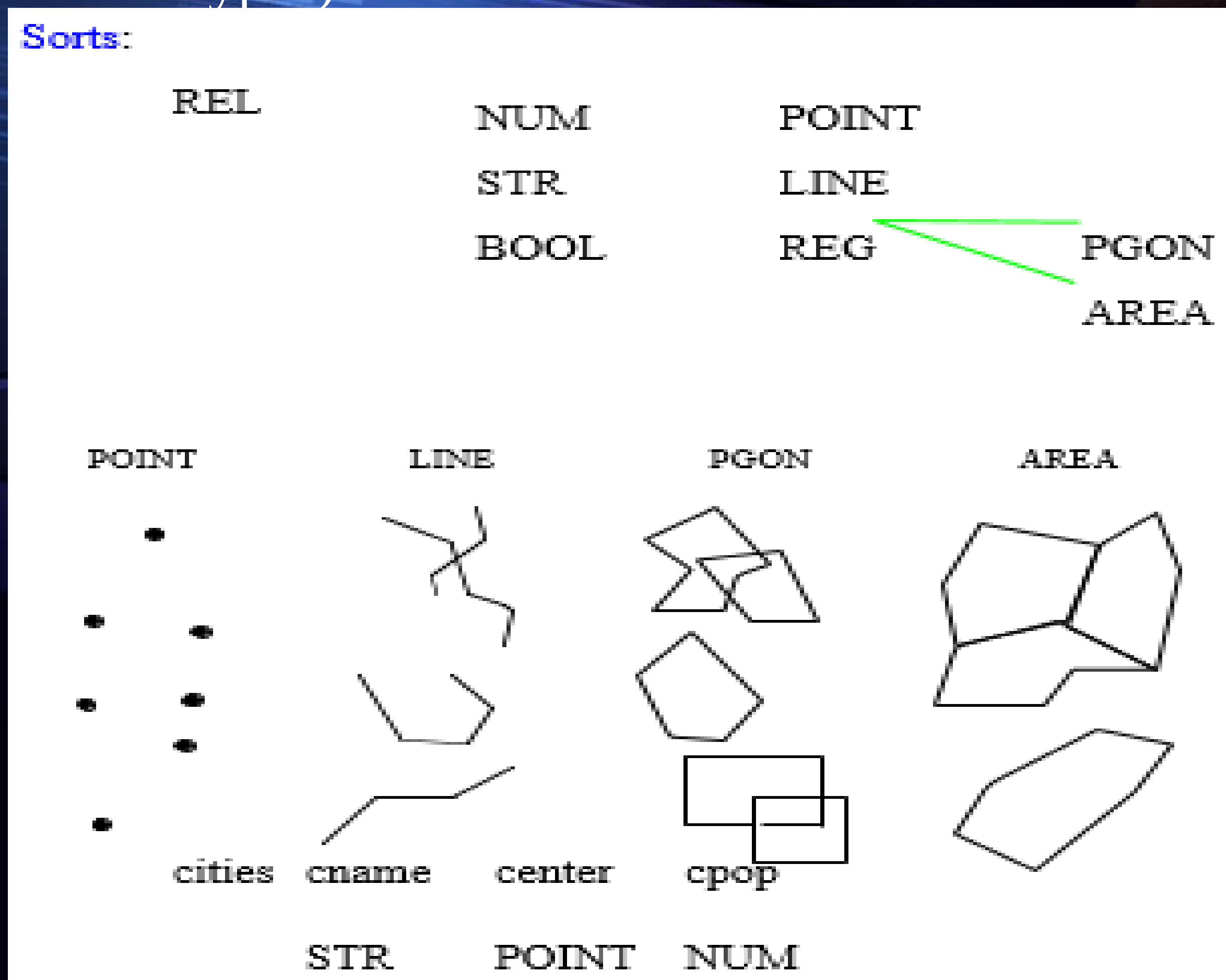
Most important operations of spatial algebras (predicates). E.g. find all objects in a given relationship to a query object.

- topological: inside, intersects, adjacent ... (invariant under translation, rotation, scaling)
- direction: above, below, north_of, ...
- metric: distance < 100



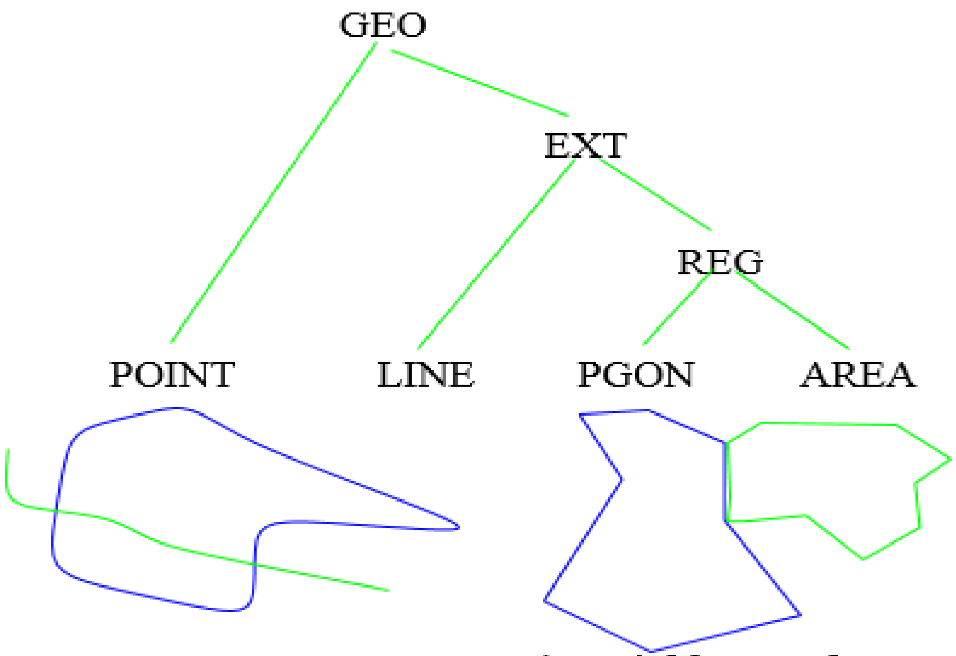
Geo-Relational Algebra

- Relational algebra viewed as a *many-sorted algebra* (relations
- + atomic data types)



Operations: Geometric Predicates

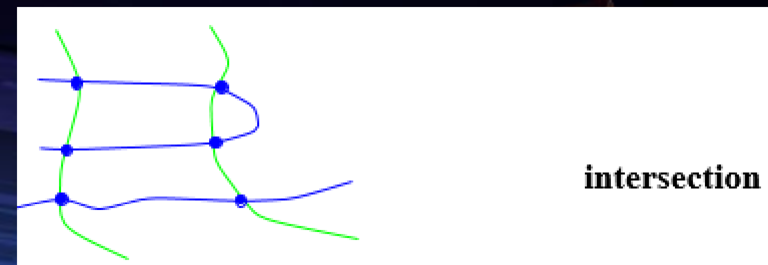
POINT × POINT	→ BOOL	=, ≠
LINE × LINE	→ BOOL	
REG × REG	→ BOOL	
GEO × REG	→ BOOL	**inside**
EXT × EXT	→ BOOL	**intersects**
AREA × AREA	→ BOOL	**is_neighbour_of**

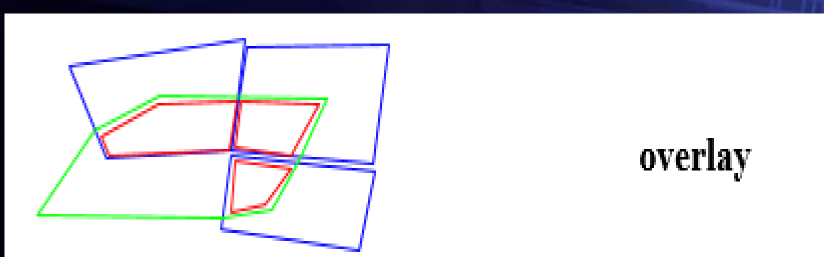


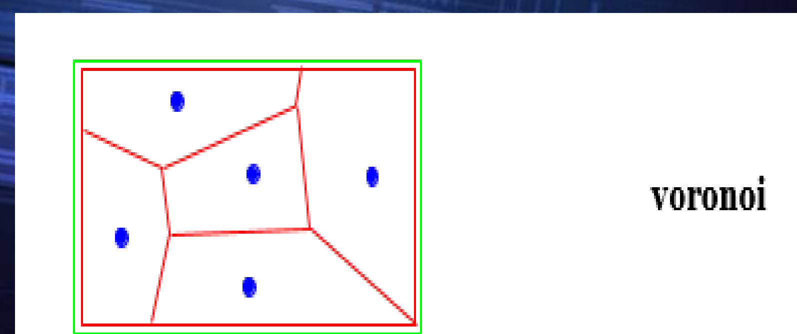
intersects **is_neighbour_of**

Geomteric Relations operations

LINE* × LINE*	→ POINT*	**intersection**
LINE* × REG*	→ LINE*	
PGON* × REG*	→ PGON*	
AREA* × AREA*	→ AREA*	**overlay**
EXT*	→ POINT*	**vertices**
POINT* × REG	→ AREA*	**voronoi**
POINT* × POINT	→ REL	**closest**





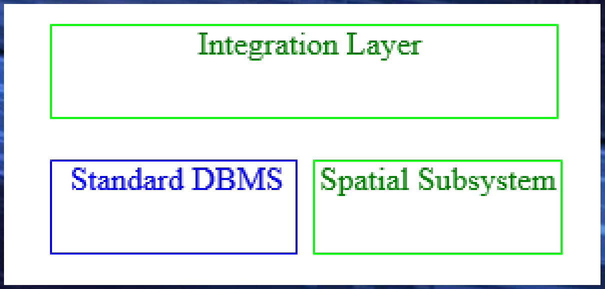


5- System Architecture

- Integrate the tools from Section 4 into the system architecture. Accommodate the following extensions:
  - representations for data types of a spatial algebra
  - procedures for atomic operations
  - spatial index structures
  - access operations for spatial indices
  - spatial join algorithms
  - cost functions for all these operations
  - statistics for estimating selectivity of spatial selection and spatial join
  - extensions of the optimizer to map queries into the spe- cialized query processing methods
  - spatial data types and operations within data definition and query language
  - user interface extensions for graphical I/O

GIS - Architectures — Using a Closed DBMS

- First generation: built on top of file system
  - → no high level data definition, no flexible querying,
    - no transaction management,
  - …
- Using a standard (mostly relational) DBMS:
  - layered architecture
  - dual architecture



- Layered architecture

Representation of SDT values:

- (1) Decompose SDT value into a set of tuples, one tuple per point or line segment
- DBMS handles geometries only as uninterpreted byte strings; any predicate or other operation on the exact geometry can only be evaluated in the top layer.
- Indexing: maintain sets of z-elements in special relations; index these with a B-tree.

References

1. B. Brown, M. Chuiu and J. Manyika, “Are you ready for the era of Big Data?” McKinsey Quarterly, Oct 2011, McKinsey Global Institute
2. C. Bizer, P. Bonez, M. L. Bordie and O. Erling, “The Meaningful Use of Big Data: Four Perspective – Four Challenges” SIGMOD Vol. 40, No. 4, December 2011
3. D. Boyd and K. Crawford, “Six Provation for Big Data” A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011, Oxford Internet Institute
4. D. Agrawal, S. Das and A. E. Abbadi, “Big Data and Cloud Computing: Current State and Future Opportunities” ETDB 2011, Uppsala, Sweden
5. D. Agrawal, S. Das and A. E. Abbadi, “Big Data and Cloud Computing: New Wine or Just New Bottles?” VLDB 2010, Vol. 3, No. 2
6. F. J. Alexander, A. Hoisie and A. Szalay, “Big Data” IEEE Computing in Science and Engineering journal 2011
7. O. Trelles, P Prins, M. Snir and R. C. Jansen, “Big Data, but are we ready?” Nature Reviews, Feb 2011
8. K. Bakhshi, “Considerations for Big data: Architecture and approach” Aerospace Conference, 2012 IEEE
8. S. Lohr, “The Age of Big Data” Thr New York times Publication, February 2012
10. M. Nielsen, “Aguide to the day of big data”, Nature, vol. 462, December 2009
11. Ralf Hartmut Güting Fernuniversität Hagen Praktische Informatik IV D-58084 Hagen Germany,

Thank You !

Mulhim Al-Doori