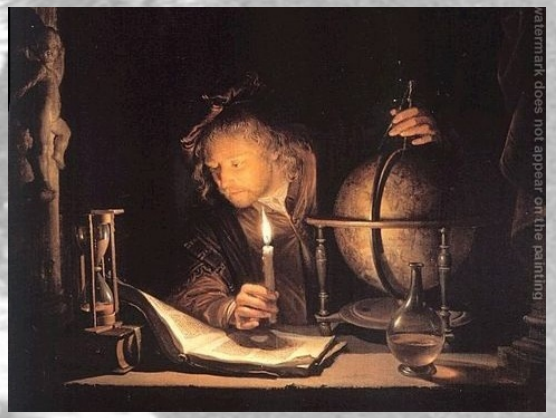
The Art of Data Science

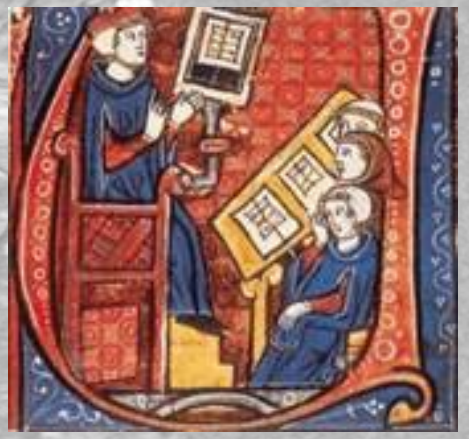
Matthew J. Graham (Caltech, VAO)

The dawn of a new era?

- "We're facing a data flood/tsunami/ explosion/firehose/glut/challenge"
- By the end of the decade, petascale data sets will be a regular feature of daily life
- "To solve it, we need…"
  - The fourth paradigm
  - *x*-informatics
  - E-/Data science
- "We've been doing this for *x* years" where $x \in [1, \infty]$

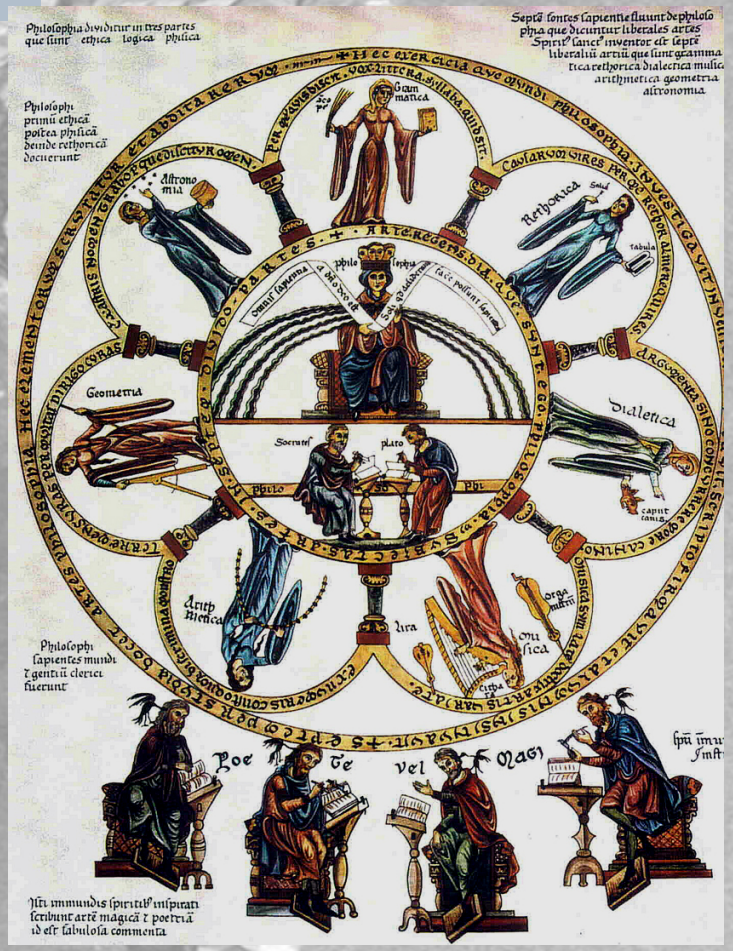


Teaching the new science

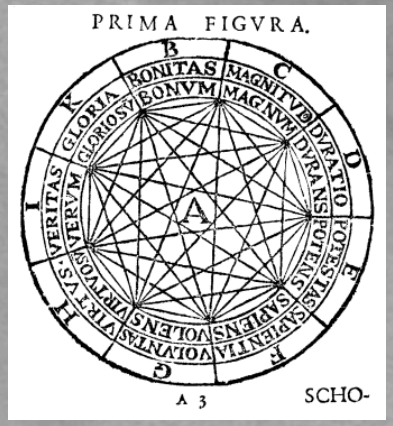


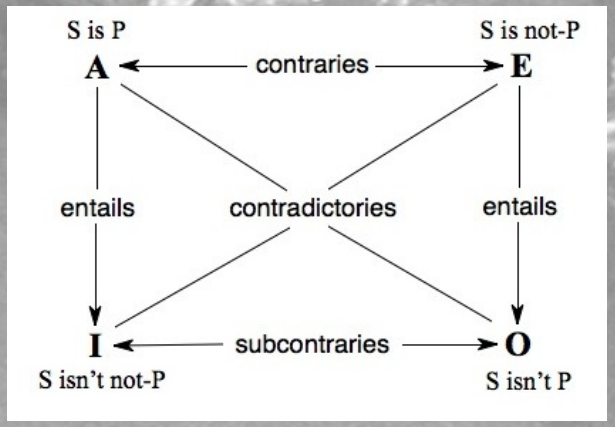
The seven liberal arts

- Trivium:
  - Logic
  - Grammar
  - Rhetoric
- Quadrivium:
  - Arithmetic
  - Geometry
  - Music
  - Astronomy
- A systemization of knowledge:
  - What rules does it obey
  - How is symbolized
  - How is communicated
  - What is its relationship to physical space and time



the nature of knowledge

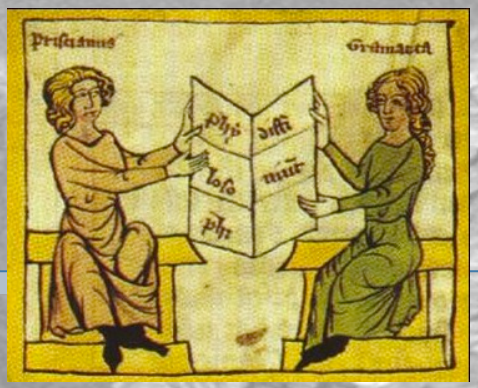




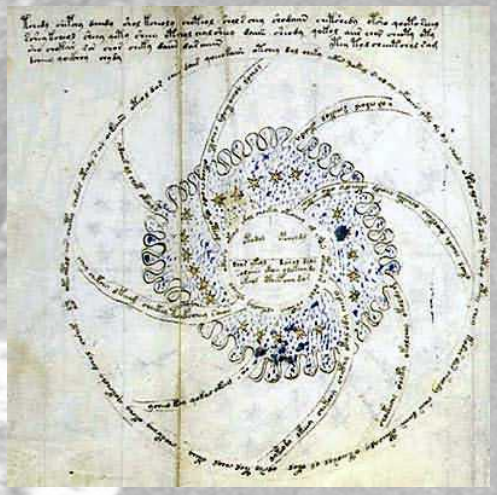

Sicut dicit philosoph⁹ sexto metaphysice tres sūt partes ꝑncipales scie speculatiue. s. nālis: mathematica. ⁊ diuina. Et rō hui⁹ ꝓpositiōis ē. q: scie distīguūt secundū distinctionē suox obiectorū de quibus sūt. sicut patet ꝙ ꝑbm tertio de aīa. ſz res a nobis speculabiles sūt tres naturales mathematice. ⁊ diuine. ergo sunt tres scientie speculatiue principales. s. naturalis: mathematica: ⁊ diuina.

The logic of data

- Anything beyond raw data values is metadata and inferior
- Description of associated knowledge required:
  - Bayesian posterior probability model
  - Semantic constructs
- Ask questions of data via inferencing:
  - Statistical:
    - hypothesis matching
  - Logical:
    - determine inconsistencies
- Ultimate descriptions

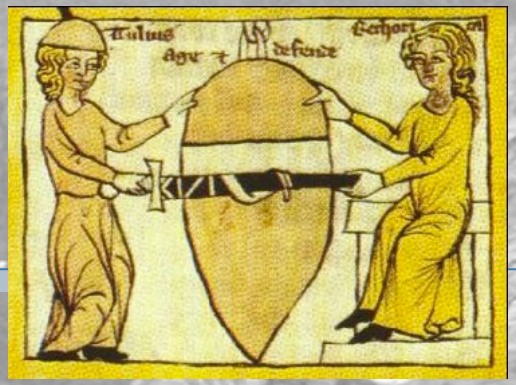


the symbolism of knowledge

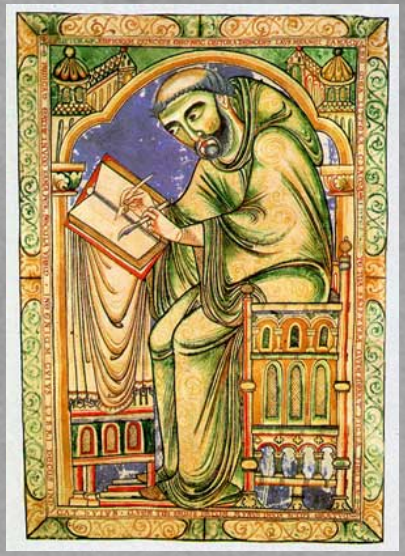


The grammar of data

- Data is just numbers or symbols stored in a digital representation
- Structure and its description makes it manipulable
- Raw binary:
  - Blob: Streaming multimedia
  - Separable textual description:
    - FITS, HDF5
  - Common data model with API:
    - CDF, netCDF
- Textual:
  - Separable schema:
    - XML, JSON
  - Interface description language:
    - Protocol Buffers, Avro

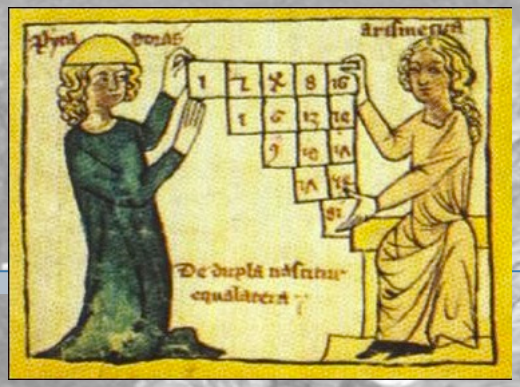


the communication of knowledge

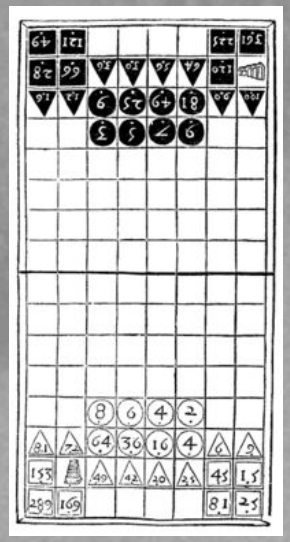


The rhetoric of data

- Data needs to be communicated well and correctly
- Physical transport ("sneakernet"):
  - efficient and reliable
  - sacrifices latency for throughput
  - bespoke
- The future looks bright for advanced capabilities – Internet2
- Maximizing conventional means:
  - Multiple streams: GridFTP
  - Fine-tuning/modifying TCP
  - UDP packets: UDT
  - Newer protocols: SCTP
  - Compression: *zip; FITS tile

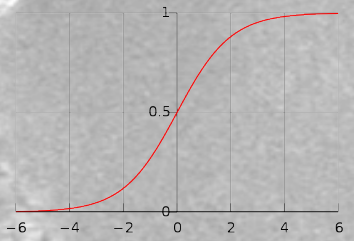


the properties and relationships of pure number

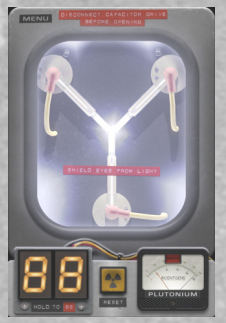


The arithmetic of data

- The utility of data lies in its ability to convey information
- The relative utility follows a logistic trend:
  - Initial data is approximately exponential
  - Progressively more data creates saturation
  - At maturity, zero utility



- Megatrend is succession of or multiply logistic
- Unprecedented progress along these trends:
  - Szalay's law: the utility of N comparable data sets is N^2
  - Exponential growth rates (Moore's law)
- Power considerations:
  - Exascale requires ~100 MW
  - GPUs, trans-silicon technologies

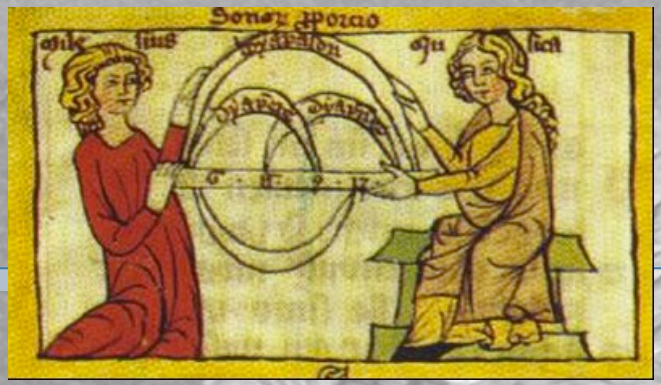


the patterns of number in nature

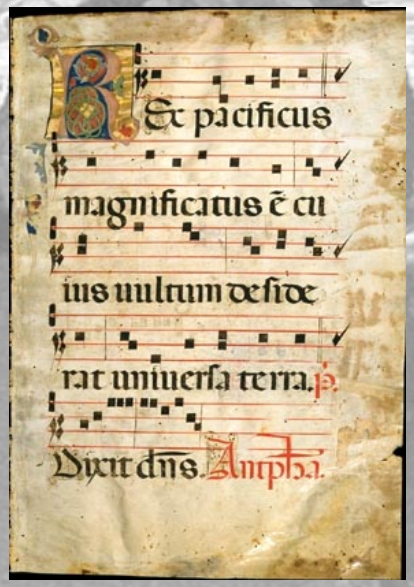
The geometry of data

- The architectural order of collections of data facilitates the study of the universe
- Cost of petascale storage: $40k – 3M/PB
- Standard approach for layering high throughput data (GFS/ HDFS):
  - Divide into fixed size chunks (~64MB)
  - Distribute multiple copies across disk cluster
  - Central/master node maintains list of chunk locations, metadata, and data operations
  - Sequence files for large numbers of small files

- Low latency random access (high availability), large numbers of varying sized files, multiple concurrent writes:
  - Distributed multi-dimensional sorted maps (BigTable/Hbase)
  - Swift, iRODs
- RDBMSs do not function well beyond 100TB (Gray & Hey):
  - NoSQL
  - SciDB



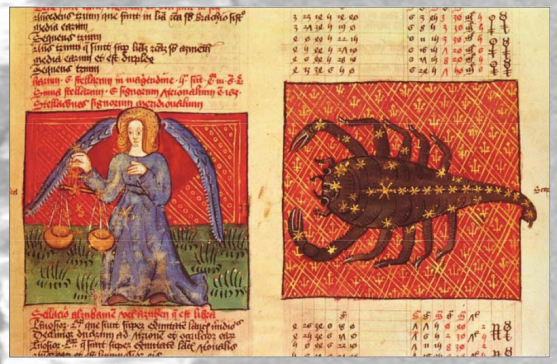
the progression of number through time



The music of data

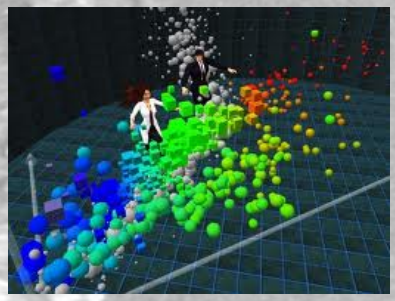
- Data computation follows identifiable patterns
- The embarrassingly parallel task:
  - Generic jobs on general resources from local dedicated clusters to scavenged cycles: Conder, BOINC
  - MapReduce (Hadoop) (summation form):
    - Primary choice for fault-tolerant and massively parallel data crunching
    - Mapper transforms input data to intermediate set of (key, value) pairs
    - Gathered, sorted by key and send to reducer
    - Reducer output collected
  - GPUs make brute force tractable
- The streaming (single pass) task:
  - Incremental formulation of algorithm
  - Further optimizations with learning types, e.g., stochastic gradient descent

the patterns of the universe



The astrology of data

- Data contains meaningful patterns
- Data mining has two primary goals:
  - Predicting the future behaviour of certain entities based on the existing behaviour of other entities in the data
  - Finding human-interpretable patterns describing the data
- Data mining techniques can be categorized as:
  - Classification, regression, clustering, summarization, dependency modelling, outlier detection
- Data mining has a process:
  - Collection and preparation/preprocessing
  - Assumption and limitations understood
  - Validation
  - Interpretation
- Incorporation of appropriate prior knowledge
  - Model-based (statistical inferencing)
  - As part of DM algorithm (semantic approach)



The systemization of data

- We are *still* concerned with:
  - Structural representations of our knowledge
  - Communicating well and correctly
  - Meaningful architectures
  - What are we studying
  - What rules apply
  - Meaningful patterns
- International and interdisciplinary effort
  - More universal than the IVOA
- Data must be a first-class entity in our worldview
- The Dark Ages of data end today