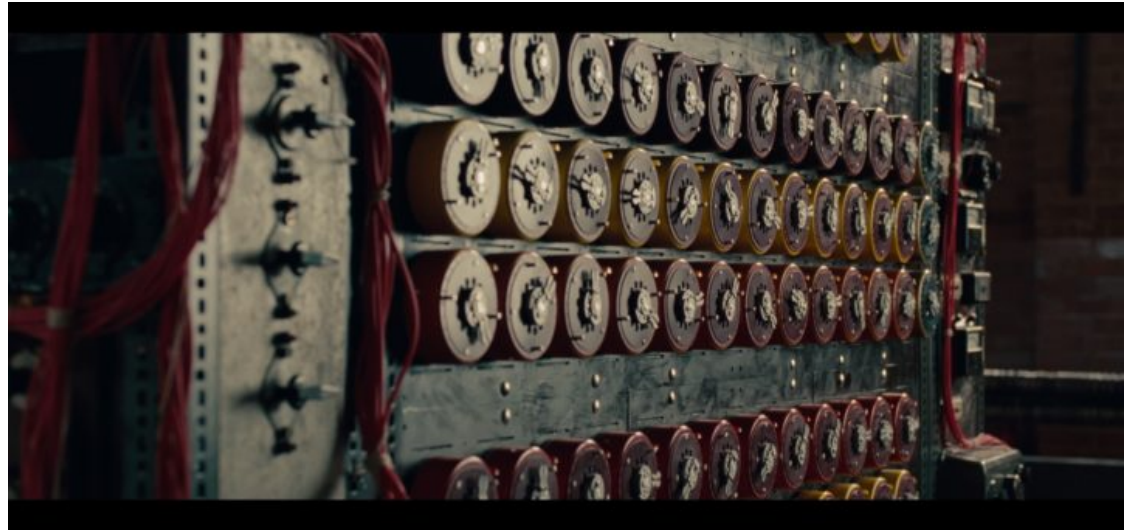
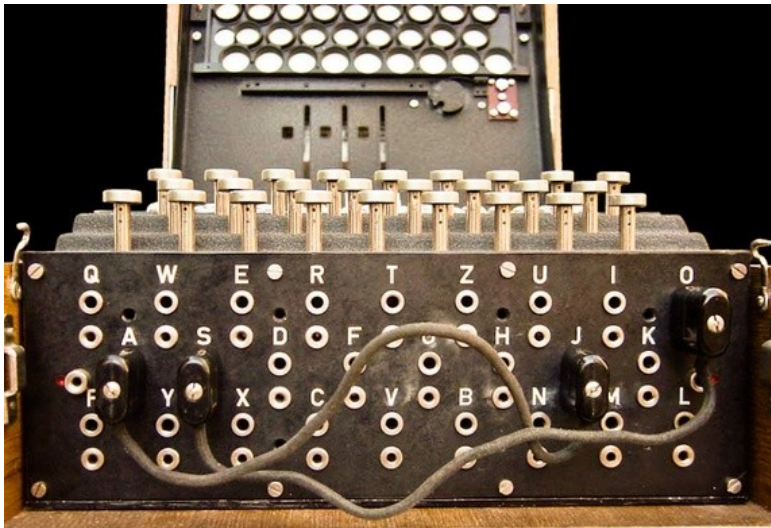


The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer.

Alan Turing

THE IMITATION GAME



THE IMITATION GAME

Input data



Combinatorial problem

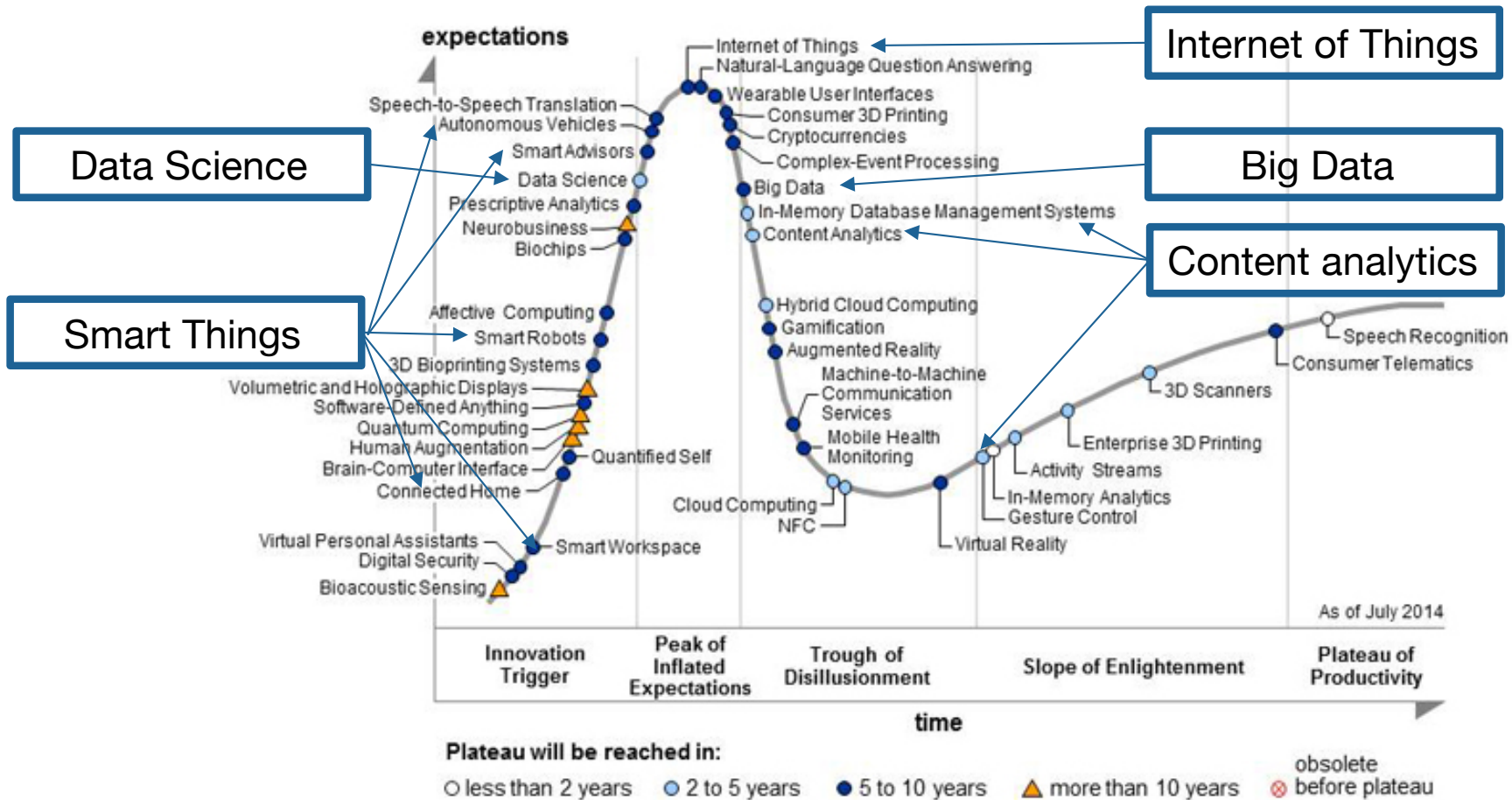


Critical variable



... CONTEMPORARY CHALLENGES





<http://www.gartner.com/newsroom/id/2819918>

Artificial Intelligence and Big Data management the dynamic duo for moving forward data centric sciences

Genoveva Vargas-Solar

Senior Scientist, French Council of Scientific Research, LIG-LAFMIA

genoveva.vargas@imag.fr

<http://vargas-solar.com>



WHAT ARE DATA CENTRIC SCIENCES ?
THE STUDY OF COMPLEX SYSTEMS



What makes Bach sound like Bach?

The composer Johann Sebastian Bach left behind an incomplete fugue upon his death, either as an unfinished work or perhaps as a puzzle for future composers to solve



The Art of Fugue is based on a single subject employed in some variation in each canon and fugue

- **Simple fugues** (Contrapunctus I-IV, 4 voices)
- **Counter fugues** subject used simultaneously in regular, inverted, augmented, and diminished forms (Contrapunctus V- VII)
- **Double and triple fugues**, employing two and three subjects respectively (Contrapunctus VIII – XI)
- **Mirror fugues**, a piece is notated once and then with voices and counterpoint completely inverted, without violating contrapuntal rules or musicality (Contrapunctus XII – XIII)
- **Canons**, labelled by interval and technique (Augmentationem in Contrario Motu, alla Ottava, Decima in Contrapunto alla Terza, Duodecima in Contrapunto alla Quinta)

... UNFINISHED FUGUE

Fuga a 3 Soggetti (Contrapunctus XIV):

- 4-voice triple fugue
- the third subject of which is based on the

B A C H motif



« At the point where the composer introduces the name BACH in the countersubject to this fugue, the composer died. »

CHALLENGING PUZZLES

- Identify the *notes* performed at specific times in a recording
- Classify the *instruments* that perform in a recording
- Classify the *composer* of a recording
- Identify precise *onset* times of the notes in a recording
- Predict the *next note* in a recording, conditioned on history

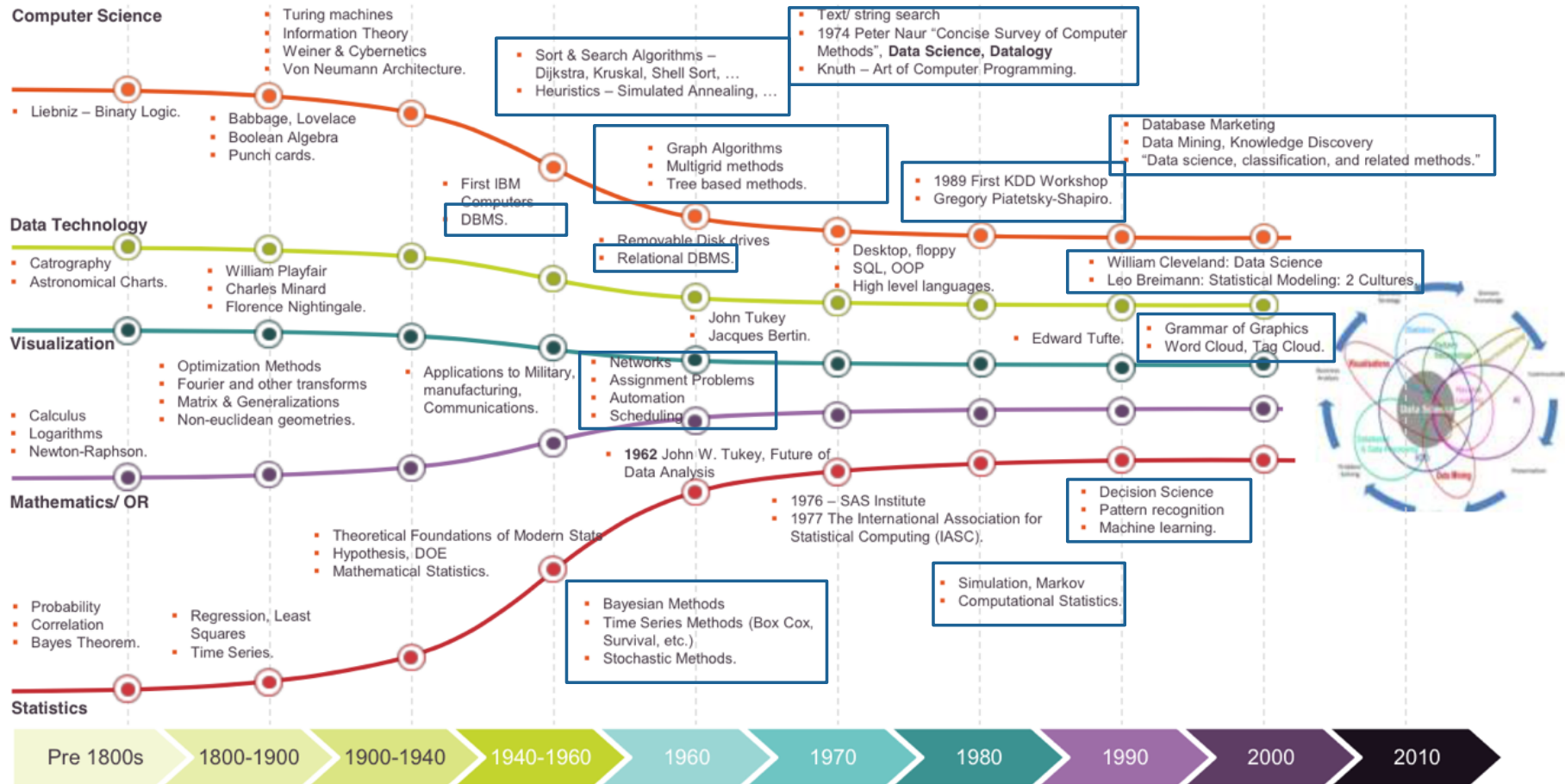
Music information retrieval

- Automatic music transcription
- Inferring a musical score from a recording

Generative models fabricating performances under various constraints

- Can we learn to synthesize a performance given a score?
- Can we generate a fugue in the style of Bach using a melody by Brahms?

DATA AS BACKBONE



Social Data Science



Data Science



Network Science



Computational Science



Digital humanities



Computation
(Algorithm: mathematical model)

Experiment
(Architecture: computing environment)

Velocity

Volume

DATA

Variety

Value

Veracity

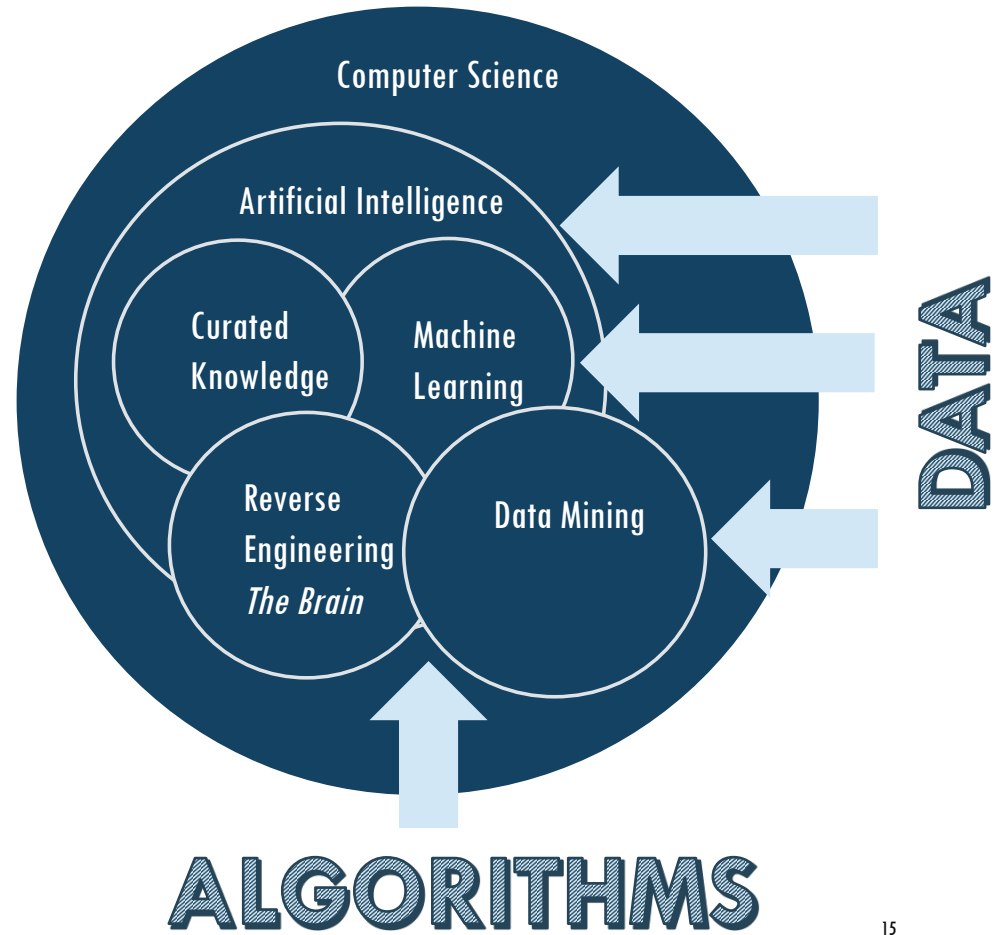
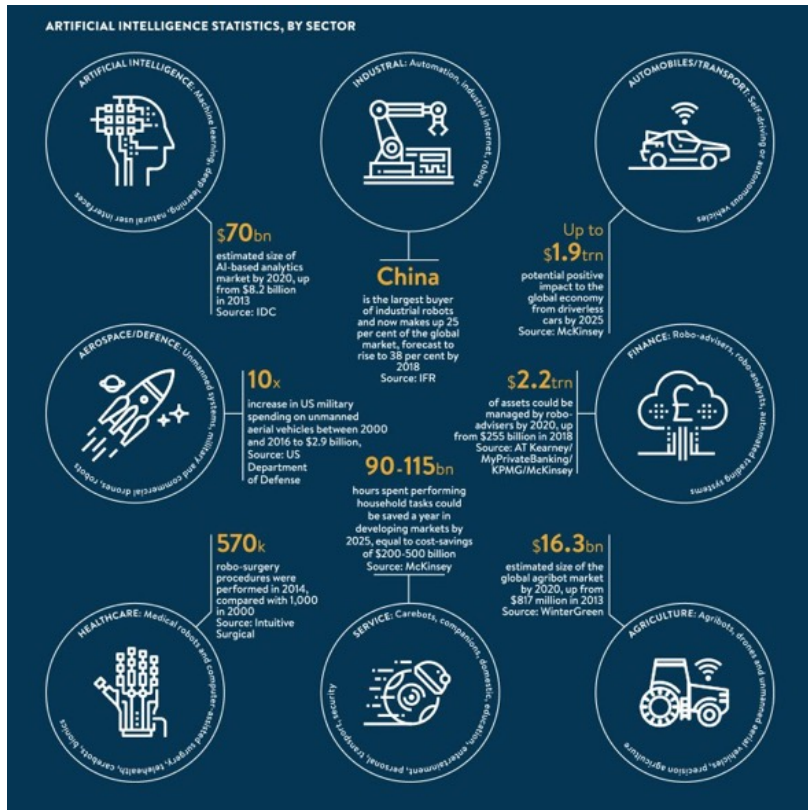
1000 Yottabytes	1 Brontobyte
1000 Brontobytes	1 Geopbyte

ARTIFICIAL INTELLIGENCE UNDERSTANDING & SIMULATING COMPLEX SYSTEMS

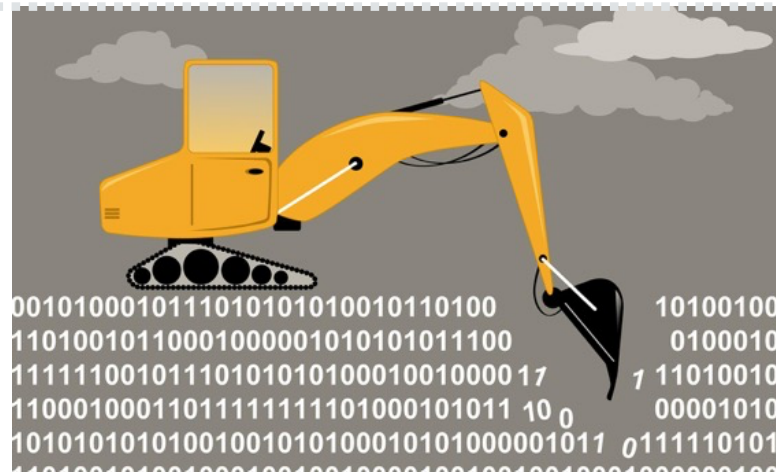


<https://ai100.stanford.edu/2016-report>





WHAT ABOUT DATA ?





5v: Value

Which is the real value of data?



VOLUME
DATA SIZE



VELOCITY
SPEED OF CHANGE



VARIETY
DIFFERENT FORMS
OF DATA SOURCES



VERACITY
UNCERTAINTY OF
DATA



Consumed data: quality, conditions in which data is retrieved; explicit cultural, contextual, background properties; uncertainty, ambiguity degree
Conditions of consumption: reproducibility, transparency degree (avoid “software artefacts”)

VOLUME
DATA SIZE

VELOCITY
SPEED OF CHANGE

VARIETY
DIFFERENT FORMS
OF DATA SOURCES

VERACITY
UNCERTAINTY OF
DATA



Data Science



DATA COLLECTIONS

Different sizes, evolution in structure, completeness, production conditions & content, access policies modification ...



NOT MANAGEABLE NEITHER EXPLOITABLE AS SUCH

RAW DATA:

heterogeneous (*variety*), huge (*volume*), incomplete, unprecise, missing, contradictory (*veracity*), continuous releases produced at different rates (*velocity*), proprietary, critical, private (*value*)



Computing resources



Applications & Data consumers

Data cleaning, processing and storage requires a lot of
DECISION MAKING

Data scientist requires knowledge about data collections content




Data collections releases

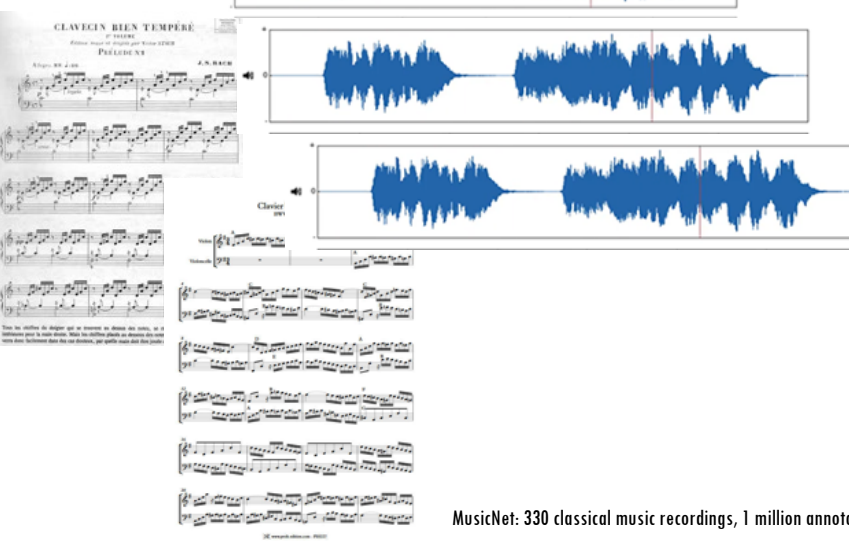


DATA SHARDING

Tocatta and Fugue in D Minor



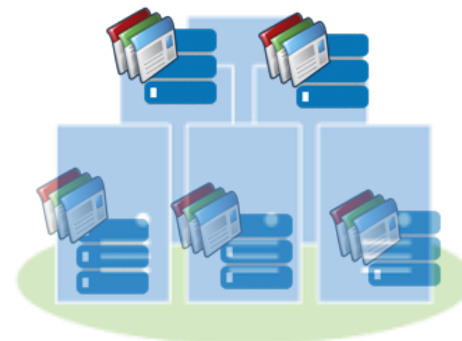
Multimedia multiform data



Clavier

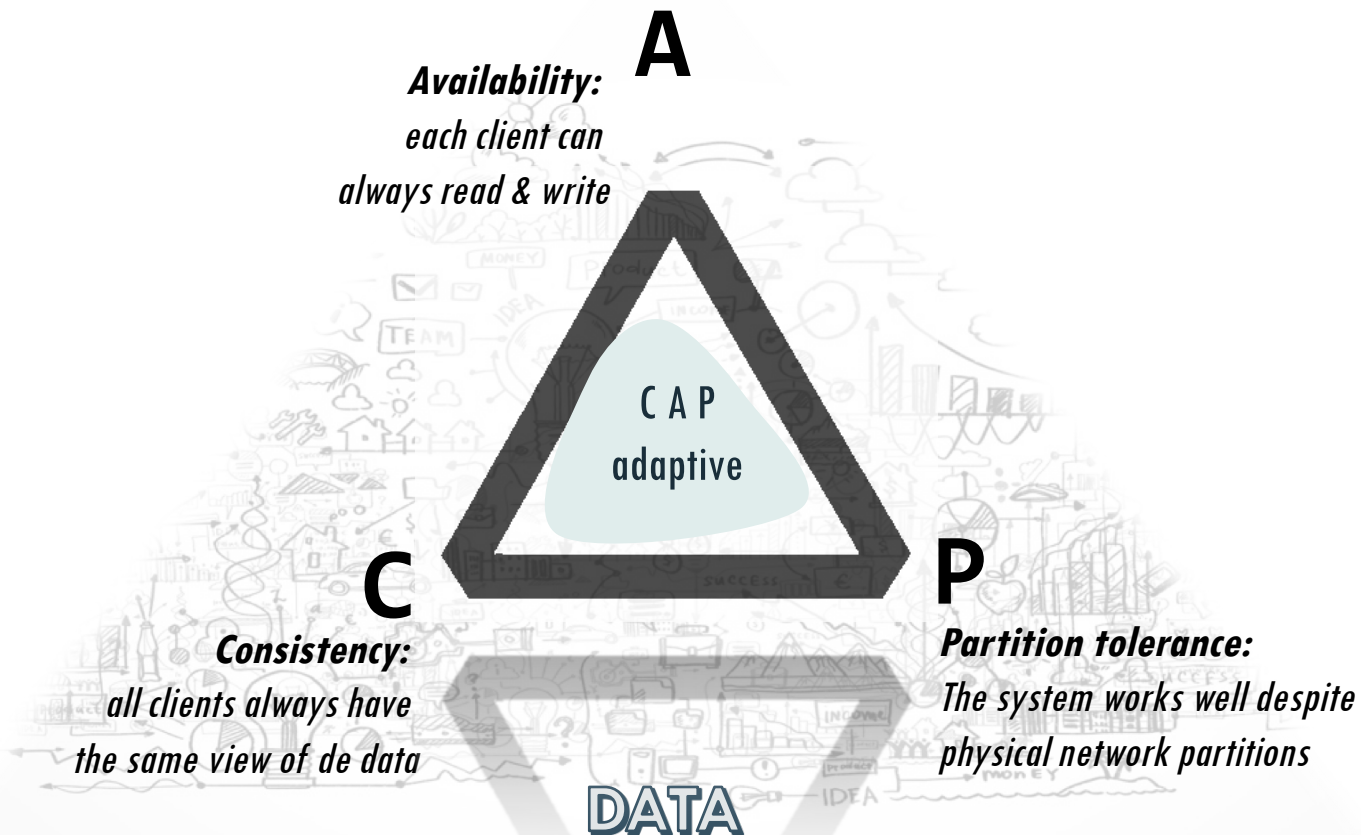
This image illustrates multimedia multiform data. It shows two musical scores: 'Tocatta and Fugue in D Minor' and 'Clavier Bien Tempere'. Below each score are three audio waveforms, representing the audio data corresponding to the musical notation. The waveforms are blue and show the amplitude of the sound over time.

Distributed File System



Sharded & colocated Input data

DATA HARVESTING & STORAGE



Requirements

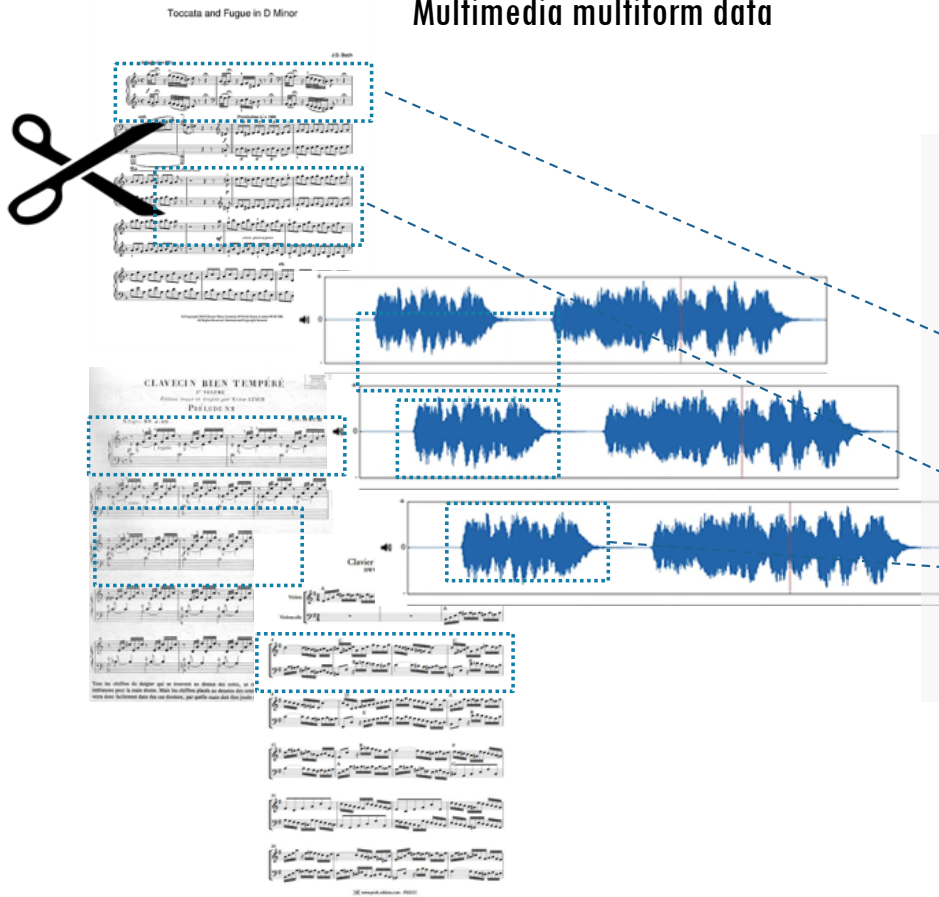
Operations

Hardware

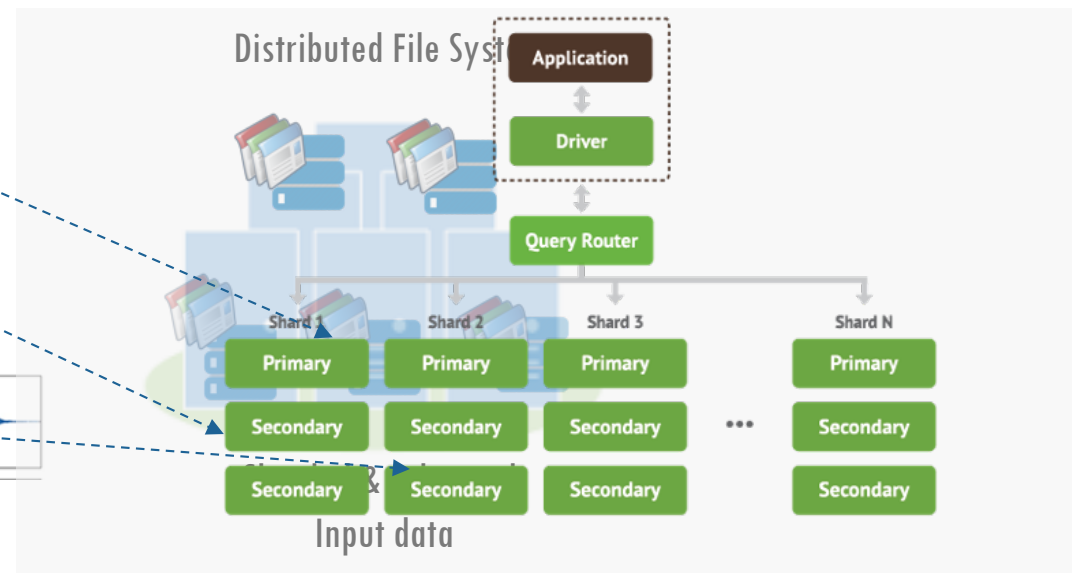


SHARDING ACROSS DIFFERENT STORES

Multimedia multiform data



Sharded data architecture



Factors:

- RAM
- CPU
- Disk
- Network

Tocatta and Fugue in D Minor

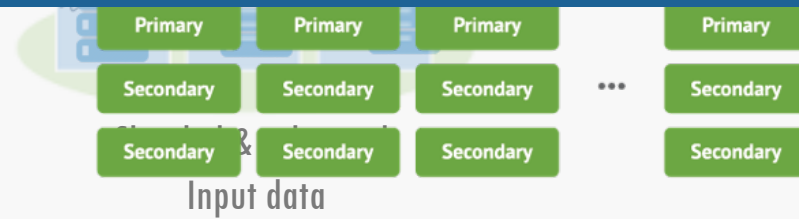
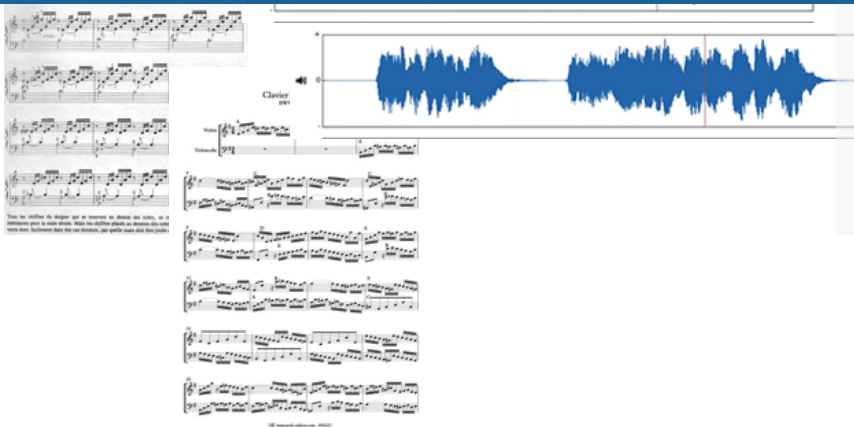
Multimedia multiform data



Sharded data architecture

Distributed File System

- Which attribute can be used to shard the collection?
- Is there critical data with particular availability requirements?
- Should some fragments be collocated?



Factors:

- RAM
- CPU
- Disk
- Network

Tocatta and Fugue in D Minor

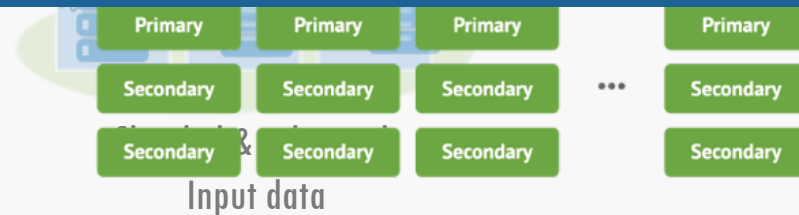
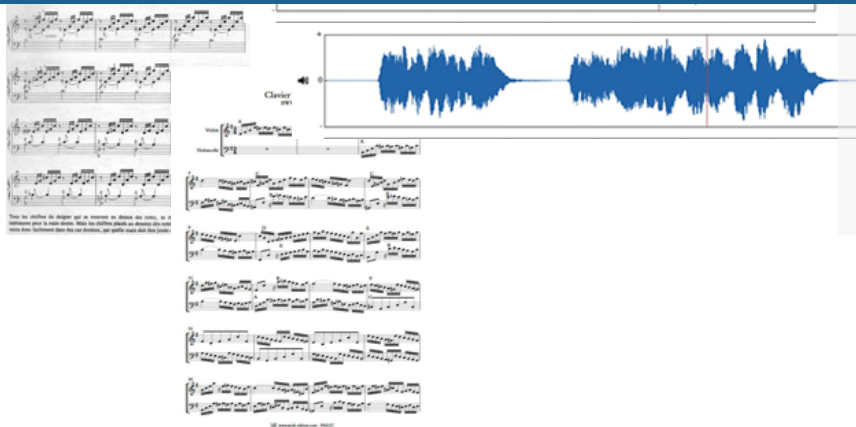
Multimedia multiform data



Sharded data architecture

Distributed File System

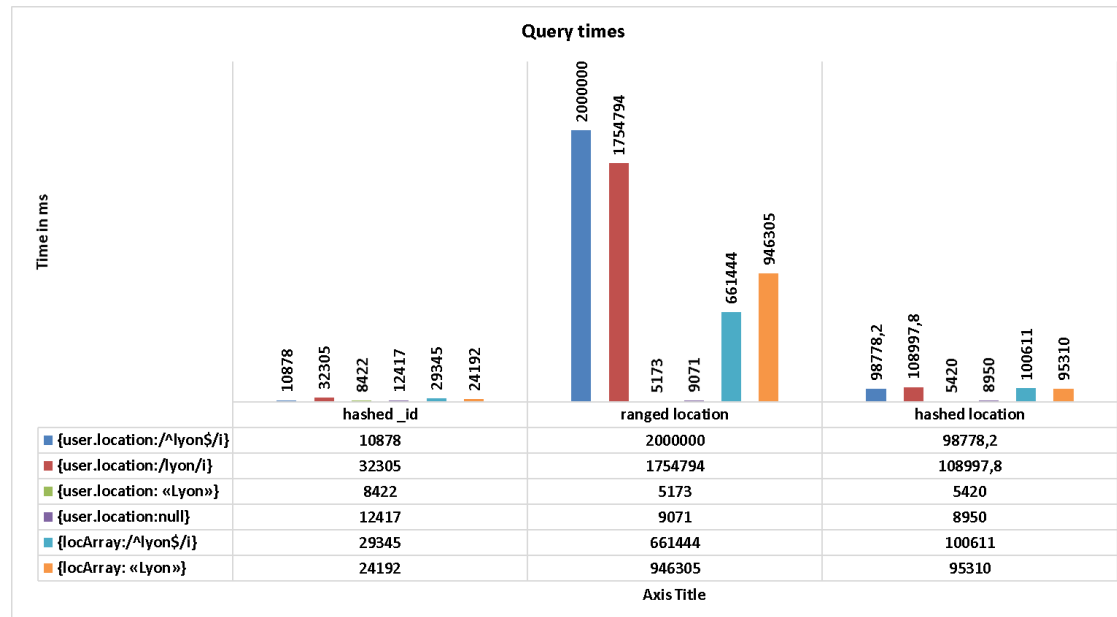
Decision making depends on: item data structure, distribution of values of each attribute of the data structure, dependency among attributes ...
This information must be discovered, computed and/or extracted



Factors:

- RAM
- CPU
- Disk
- Network

EXPERIMENTAL RESULTS: QUERIES



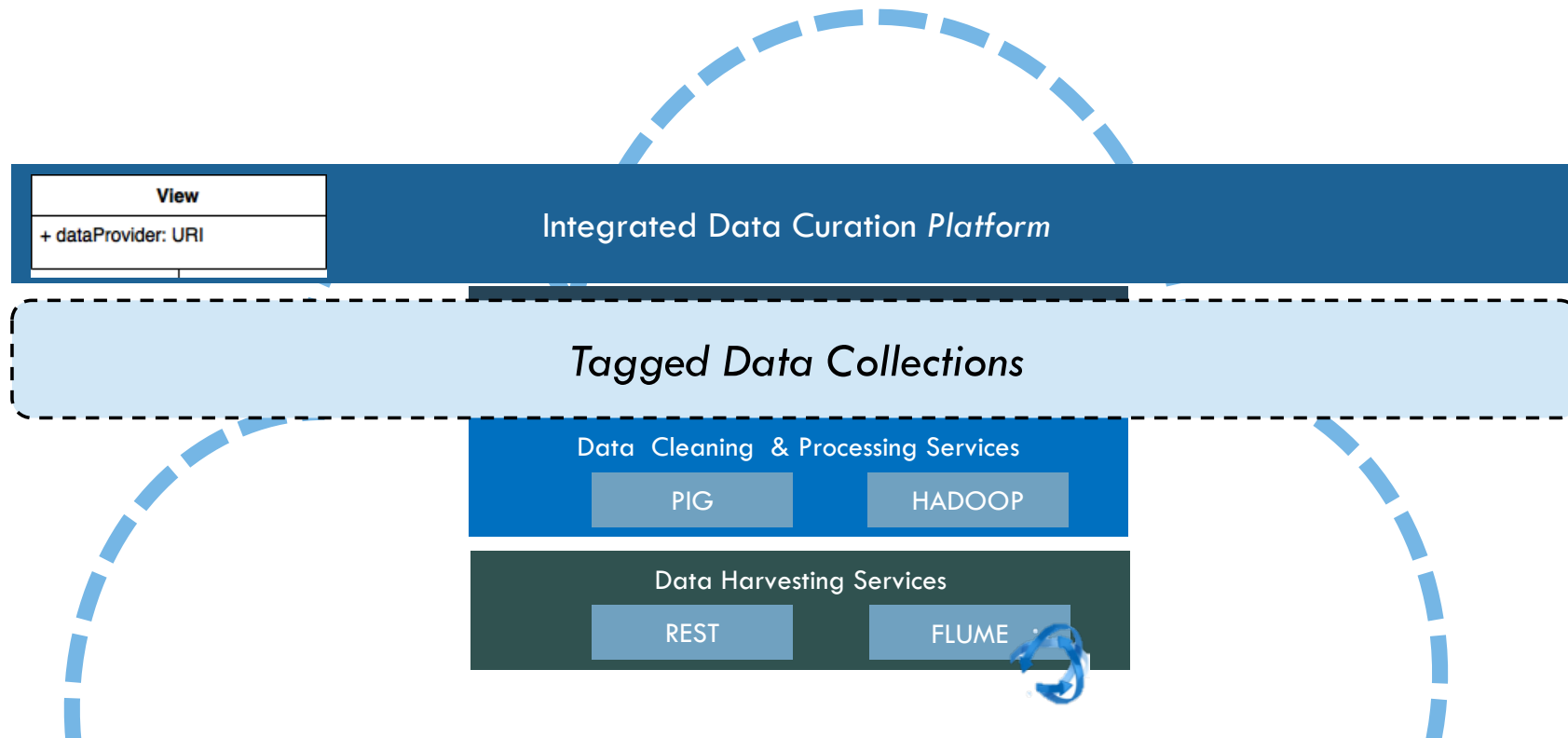
Ranged sharding gives both the best and the worst performances

Hashed _id gives the most consistent and generally low time

Hashed location gives some of the best times

- The most complex queries remain relatively quick

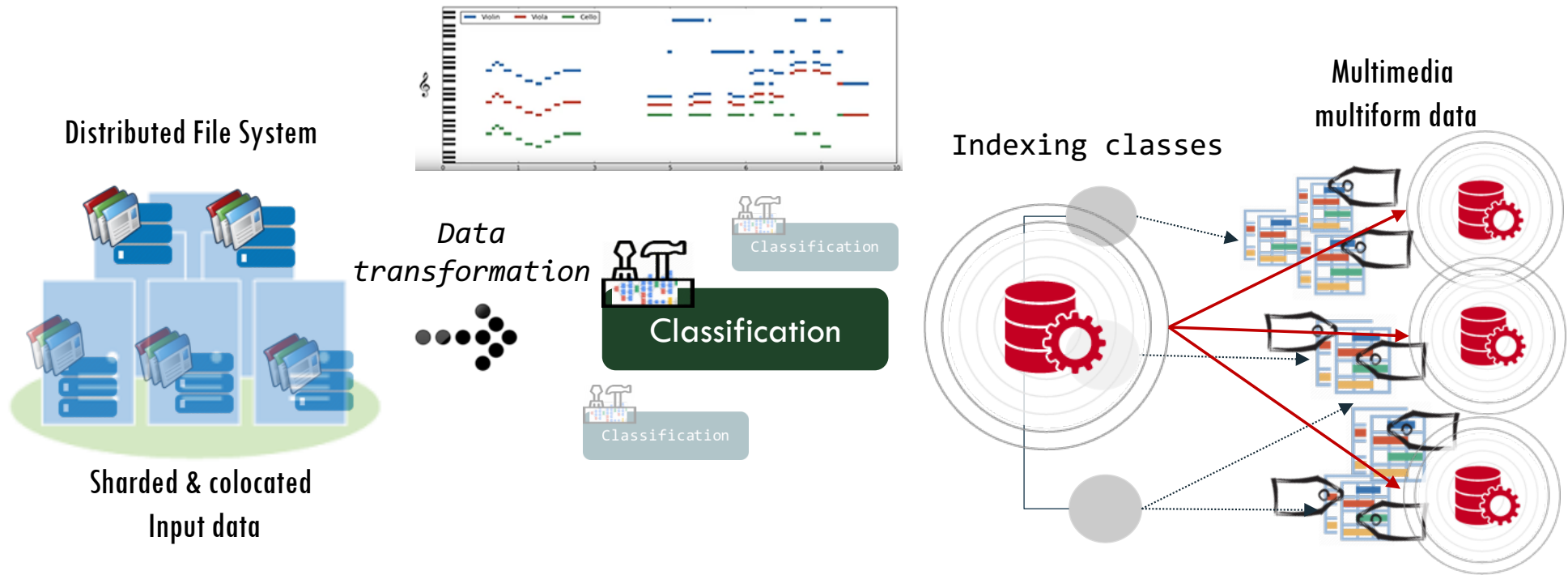
DATA CURATION ENVIRONMENT



Towards Cloud big data services for intelligent transport systems; Gavin Kemp, Genoveva Vargas-Solar, Catarina Ferreira da Silva, Parisa Ghodous, Christine Collet, Pedropablo Lopez. concurrent engineering 2015, Jul 2015, Delft, Netherlands

Service Oriented Big Data Management for Transport; G. Kemp, G. Vargas-Solar, C. Ferreira Da Silva, P. Ghodous, C. Collet; *Smart Cities, Green Technologies, and Intelligent Transport Systems / series Communications in Computer and Information Science*, Springer, 579, pp. 267-281, 2016

INDEXING & STORING



- the precise time of each note every recording,
- the instrument that plays each note,
- the note's position in the metrical structure of the composition

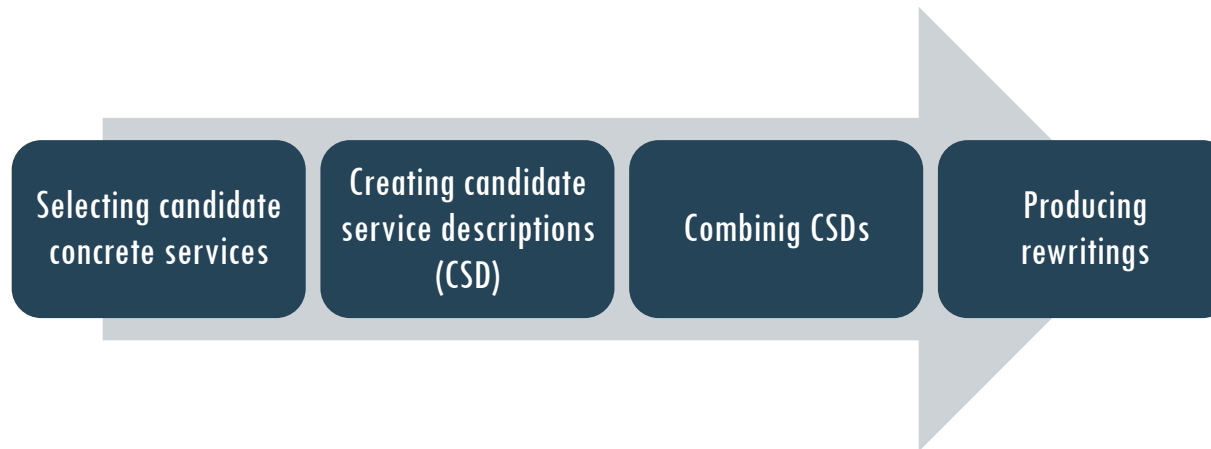
"SUR MESURE" DYNAMIC DATA INTEGRATION

A combinatorial problem where a query result is a data collection integrated by

- composing different data providers
- data processing (cloud) services

that fulfill quality constraints and SLAs specified by a data consumer

"SUR MESURE" DYNAMIC DATA INTEGRATION

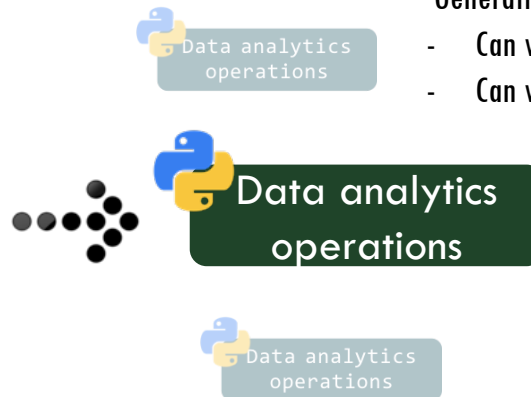
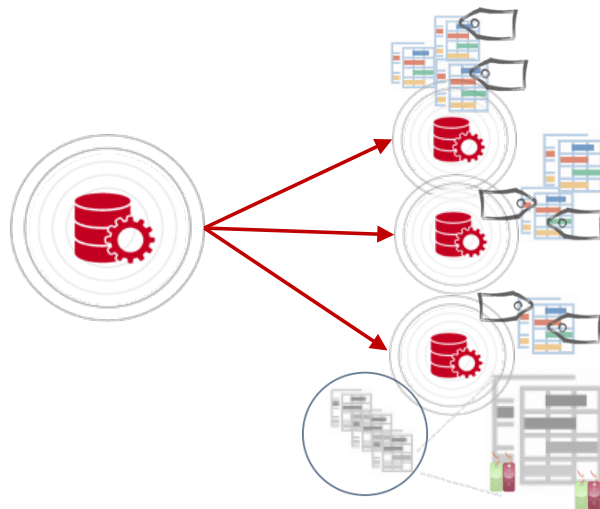


- **A rewriting algorithm customizing**

- data providers (services) **look up**
- data integration considering different data consumers requirements and expectations
- requirements & expectations depend on the context in which they consume data (e.g., mobile devices with few physical capacities, critical decision making)

¹ D. A. S. Carvalho, P. A. S. Neto, C. Ghedira, G. Vargas-Solar, N. Bennani. **Rhone: a quality-based query rewriting algorithm for data integration**. East-European Conference on Advances in Databases and Information Systems, Aug 2016, Prague, France. ADBIS East-European Conference on Advances in Databases and Information Systems, 2016.

LOADING FOR ANALYTICS



Music information retrieval

- Automatic music transcription
- Inferring a musical score from a recording

Generative models that can fabricate performances under various constraints

- Can we learn to synthesize a performance given a score?
- Can we generate a fugue in the style of Bach using a melody by Brahms?

- Identify the *notes* performed at specific times in a recording
- Classify the *instruments* that perform in a recording
- Classify the *composer* of a recording
- Identify precise *onset* times of the notes in a recording
- Predict the *next note* in a recording, conditioned on history

what can go wrong?

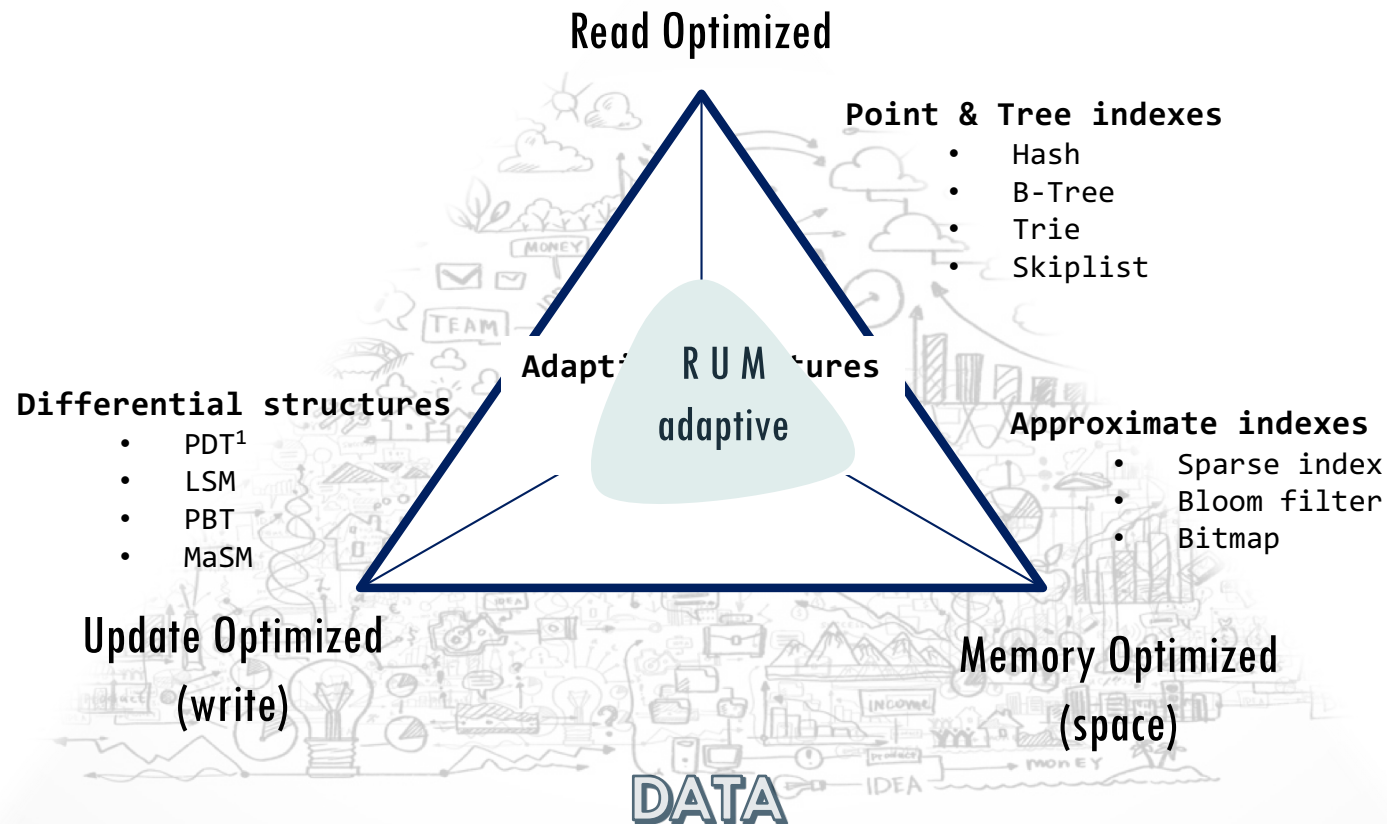
not enough space to index all data

not enough idle time to finish proper tuning

by the time we finish tuning, the workload changes

not enough money - energy - resources

ACCESS METHODS



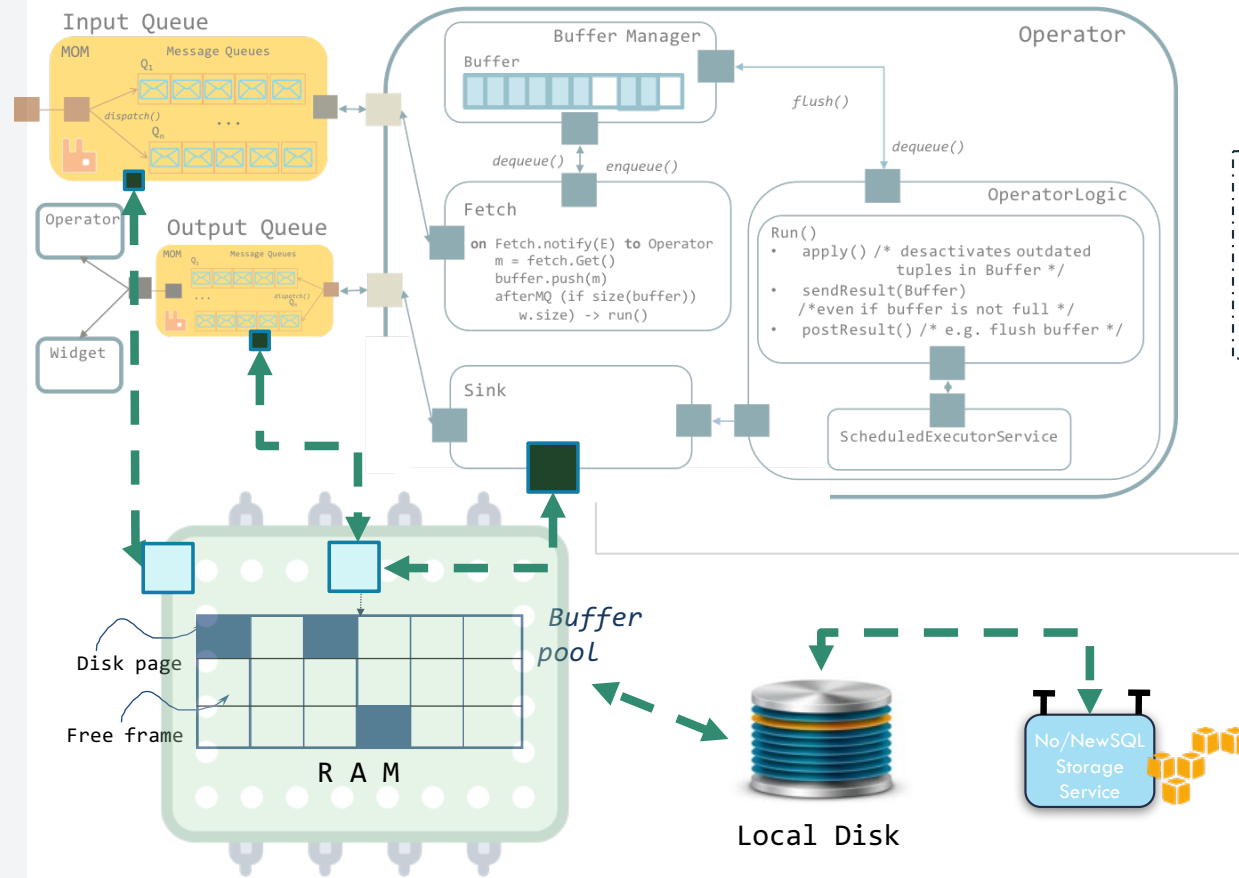
Requirements

Operations

Hardware

DEALING WITH VOLUME

IoT world
(objects farms)



Persistence management strategies for reading/writing/updating data from RAM, Cache and Disk for dealing with volume



FINAL COMMENTS

Infrastructure	Analytics	Applications
<p>NoSQL Databases</p>	<p>Analytics Platforms</p>	<p>Ad Optimization</p>
<p>NewSQL Databases</p>	<p>BI Platforms</p>	<p>Marketing</p>
<p>Cluster Service</p>	<p>Unstructured Data</p>	<p>Finance</p>
<p>Machine Learning</p>	<p>Data Visualization</p>	<p>Human Capital</p>

Avoid getting lost in the dense complexity of technological chaotic forest

<p>Cross Infrastructure /</p>	<p>Health</p>	<p>Industries</p>
<p>Open Source</p>	<p>Real Time</p>	<p>Stat Tools</p>
<p>Data Sources</p>	<p>Sensor Data</p>	<p>Incubators & School</p>

Addressing **data centric sciences** problems is a matter of designing complex systems according to a **multidisciplinary vision**

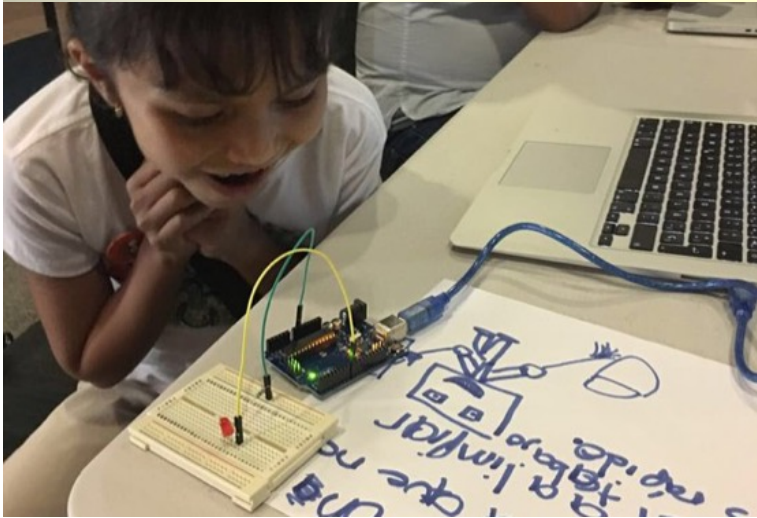


Move from **design based on intuition & experience** to a more **formal and systematic way** to design systems



Let's move forward data centric sciences







Genoveva Vargas-Solar

CR1, CNRS, LIG-LAFMIA

Genoveva.Vargas@imag.fr

<http://vargas-solar.com>

VISUAL GUIDE TO NOSQL SYSTEMS

