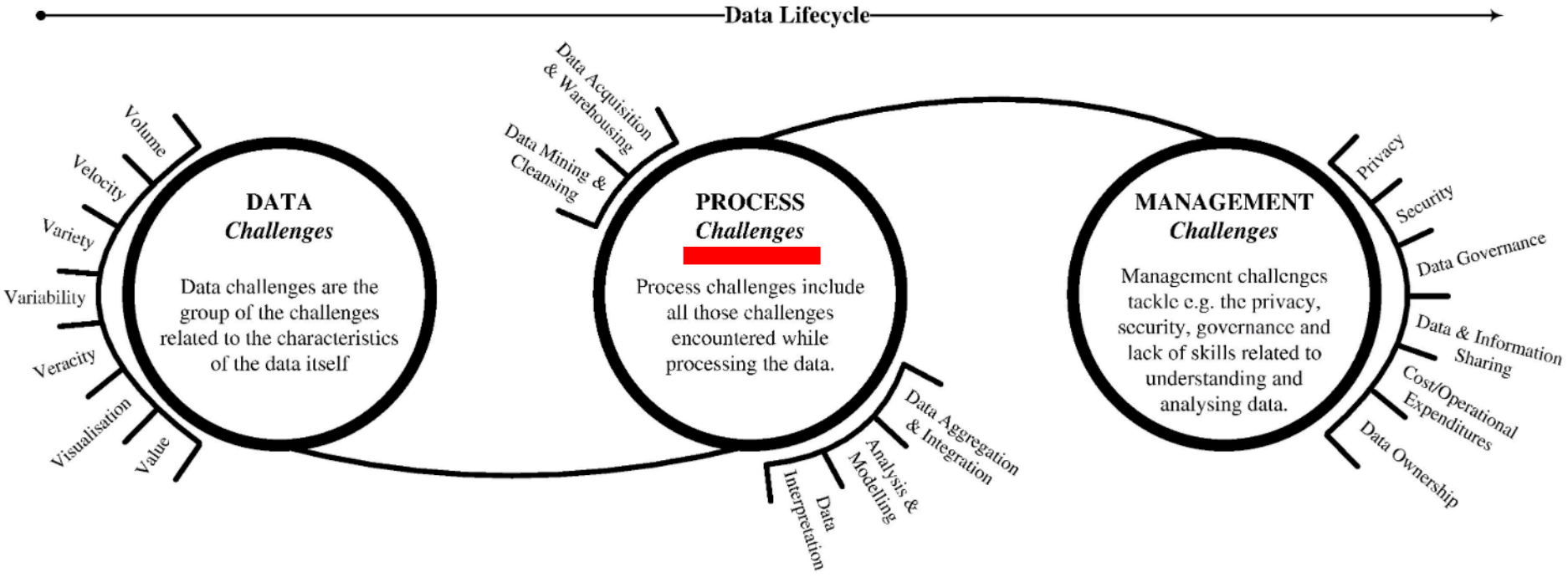
Big Data Processes and Management

COMP9313: Big Data Management

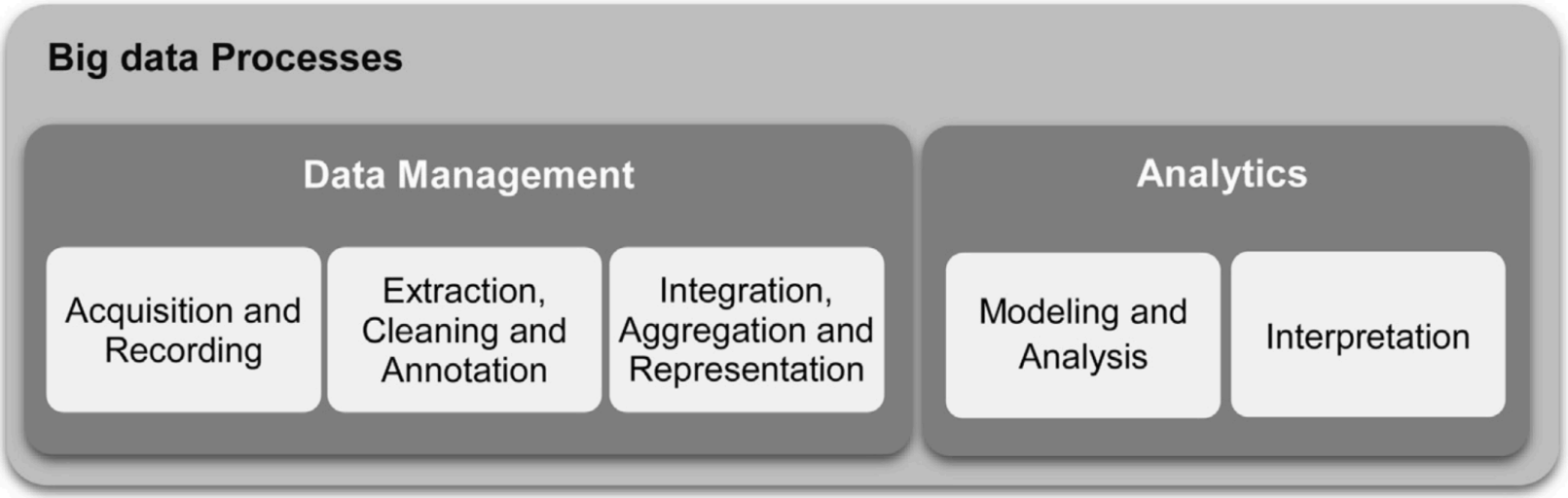
Big Data Lifecyle



https://www.sciencedirect.com/science/article/pii/S014829631630488X#bb9030

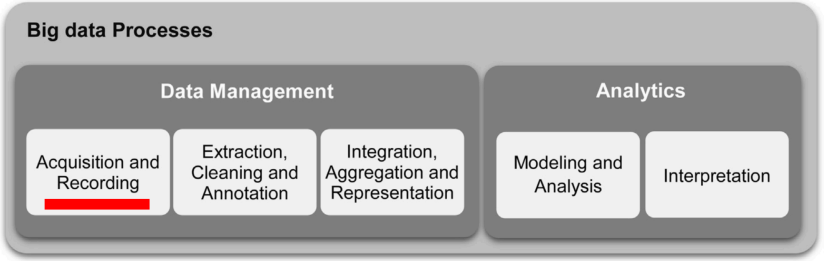
Big Data Processes

Big Data Processes



https://www.sciencedirect.com/science/article/pii/S0268401214001066

Data Management: Acquisition and Storage

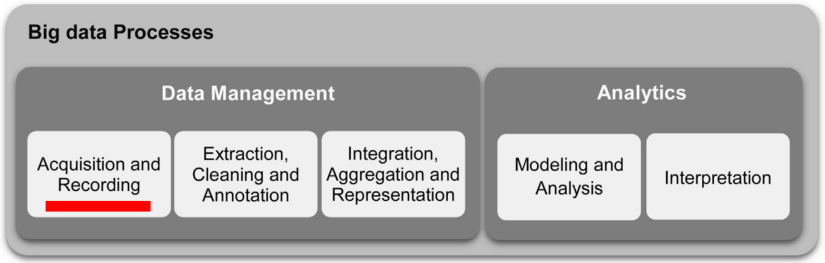


Data Acquisition:

- Samples of a physical phenomenon (from real-world sources)
- Conversion into digital representation (suitable for information systems)
- Sources: Smart phones, social media, IoT devices, online retail systems, etc.

source

Data Management: Acquisition and Storage

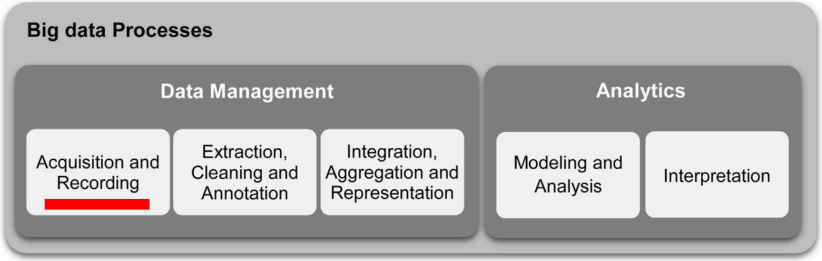


Sources of Big Data:

- User generated content
  - Applications with massive users
  - Unstructured data
  - Raw data -> Information extraction techniques needed
- Transactional data
  - Data generated by large scale systems
  - Web logs, bank transactions, sensor readings
  - Typically structured (pre-defined schema)
- Scientific Data
  - Data-intensive experiments (e.g. physics and genome data)
  - Structured, semi-structured and unstructured data
  - Provenance is important (e.g. for reproducibility)

https://link.springer.com/content/pdf/10.1007%2Fs11704-013-3903-7.pdf

Data Management: Acquisition and Storage

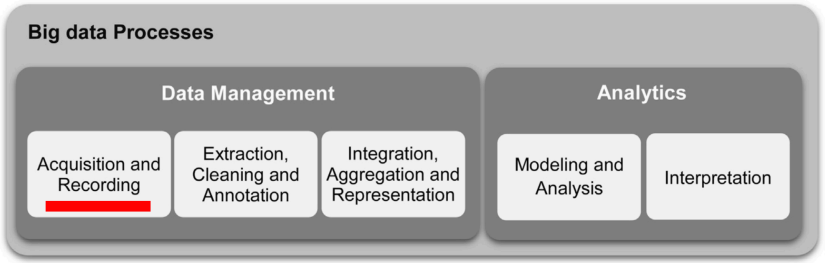


Sources of Big Data:

- Web data
  - Crawled data (e.g. for web search and indexing)
  - Unstructured and semi-structured in nature
- Graph data
  - Knowledge graphs (e.g. babelnet.org)
  - Large number of nodes and relationships
  - Ad-hoc topology -> harder to process

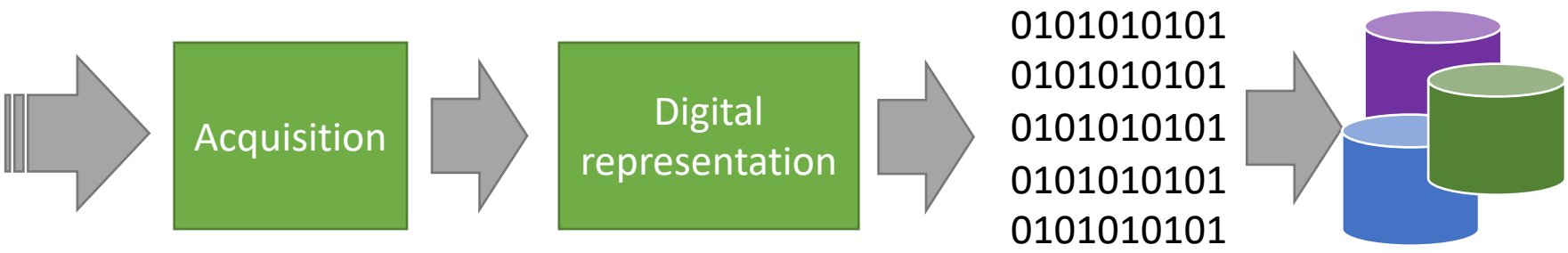
https://link.springer.com/content/pdf/10.1007%2Fs11704-013-3903-7.pdf

Data Management: Acquisition and Recording



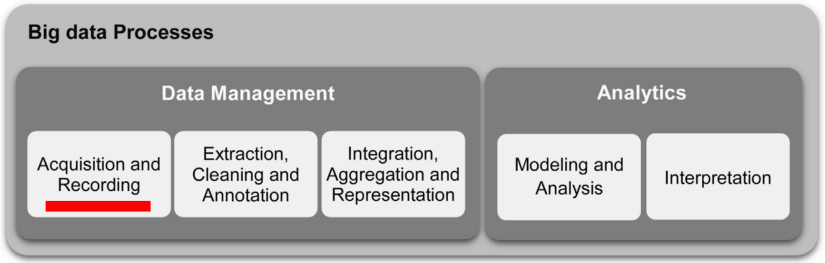
Storage:

- Different storage technologies
  - Traditional RDMS (e.g. Postgres and MySQL)
  - NoSQL (e.g. HBase and Hive)
  - Distributed file systems (e.g. HDFS)
  - Graph databases (e.g. Neo4J)
- Different data representations
  - Structured (e.g. relational data)
  - Semi-structured (e.g. JSON)
  - Unstructured data (e.g. text and images)



https://link.springer.com/content/pdf/10.1007%2Fs11704-013-3903-7.pdf

Data Management: Acquisition and Recording

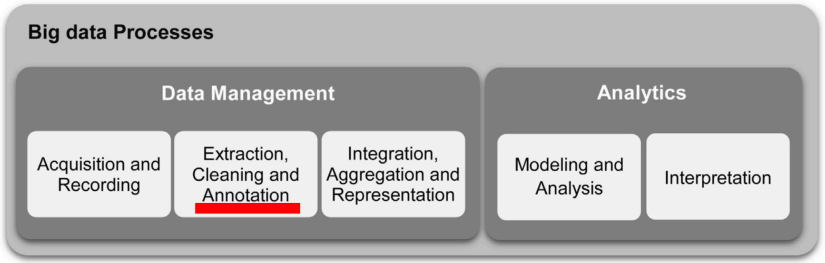


Storage:

- Three Big Data Vs to consider:
  - Volume: How large is (will be) the data we acquire?
  - Velocity: At what pace are we sampling the data?
  - Variety: How diverse is the sampled data?

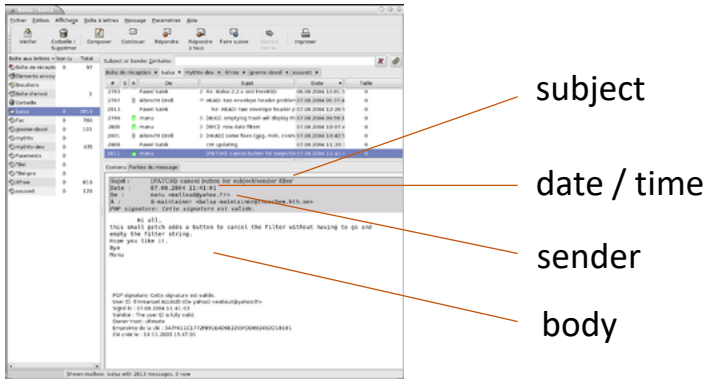
https://link.springer.com/content/pdf/10.1007%2Fs11704-013-3903-7.pdf

Data Management: Extraction, Cleaning and Annotation



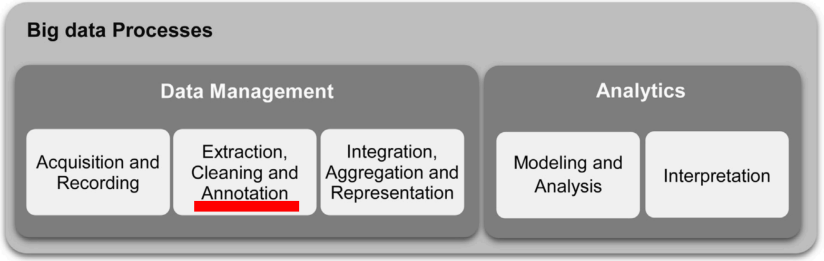
Extraction:

- Extract only relevant information (e.g. metadata in emails or named-entities from a corpus)
- Handing duplicate data (e.g. papers in arXiv.org vs. digital libraries)
- Use of complex techniques for data extraction (e.g. pattern matching and ML techniques)



source

Data Management: Extraction, Cleaning and Annotation

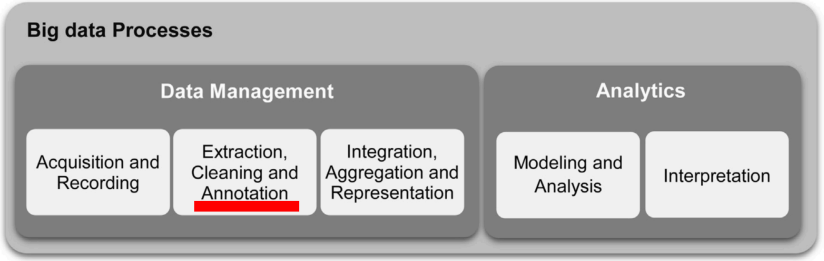


Example: Web Data Extraction

- Mostly HTML documents -> Semi-structured data
- Extraction techniques
  - Tree-based techniques
    - Leverages the tree-based structure of HTML documents
    - Typically uses XPath + Tree matching algorithms
  - Web wrappers
    - Wrappers -> extracts structured data from unstructured or semi-structured data
    - Based on regular expressions, logic (wrapping languages), machine learning (e.g. supervised learning)

https://www.sciencedirect.com/science/article/pii/S0950705114002640

Data Management: Extraction, Cleaning and Annotation

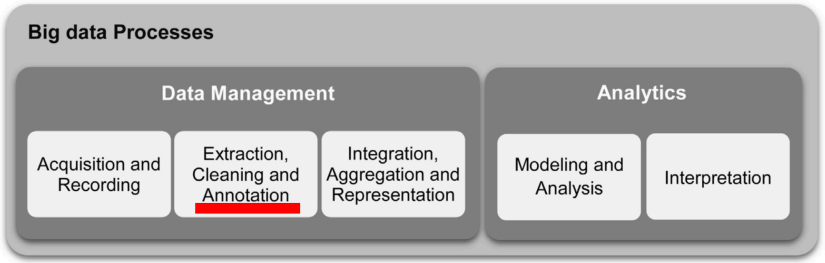


Cleaning:

- Any good big data project needs to satisfy certain quality criteria (garbage in -> garbage out)
- Examples of data quality dimensions:
  - Completeness
  - Timeliness
  - Free-of-error
  - Relevancy
  - Value-added
  - Believability
  - Interpretability
  - ... and more

https://dl.acm.org/citation.cfm?id=506010

Data Management: Extraction, Cleaning and Annotation

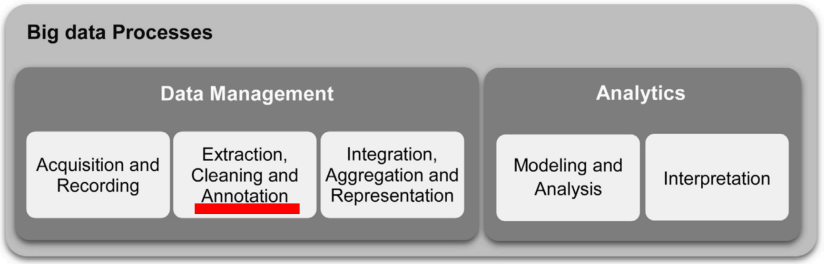


Cleaning Phases:

- Error detection:
  - Error type: Definition of what errors should be detected
  - Automation: To what extent the task is automated (are humans involved at all?)
  - Business Intelligence Layer: Where in the data processing pipeline should the error be detected?
- Error repairing:
  - Repair target: What errors to focus on? Interactions among errors?
  - Automation: Again, to what extent the task is automated (are humans involved at all?)
  - Repair model: In-situ repairs vs. Models for repairs

https://dl.acm.org/citation.cfm?id=2912574

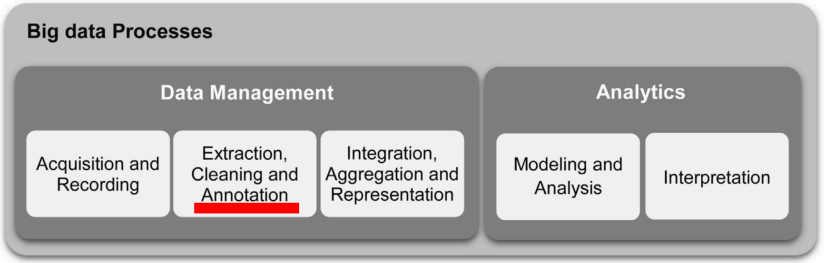
Data Management: Extraction, Cleaning and Annotation



Annotation:

- Sometimes annotation comes from the sources (e.g. tags in Twitter)
- Annotation helps with:
  - Adding semantics to data
  - Structuralizing raw, unstructured data
- Big data -> Automated techniques for annotation / tagging
  - Video tagging (e.g. lecture videos)
  - Text tagging (e.g. part-of-speech tagging)
  - Speech audio tagging (e.g. radio program tagging)

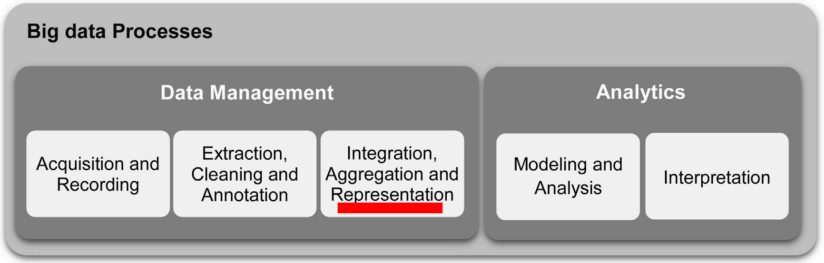
Data Management: Extraction, Cleaning and Annotation



Annotation techniques:

- Crowd-based annotation
  - Tags provided by the source (e.g. Twitter, Stack Overflow)
  - Tagging through Crowdsourcing platforms (e.g. Mturk)
- Automated annotation:
  - Using ad-hoc Machine Learning techniques
  - NLP-based tagging (e.g. POS tagging)
- Hybrid approach
  - Combine crowd-based annotation (labeling) and automated techniques (e.g. ML/Classification)

Data Management: Integration, Aggregation and Representation

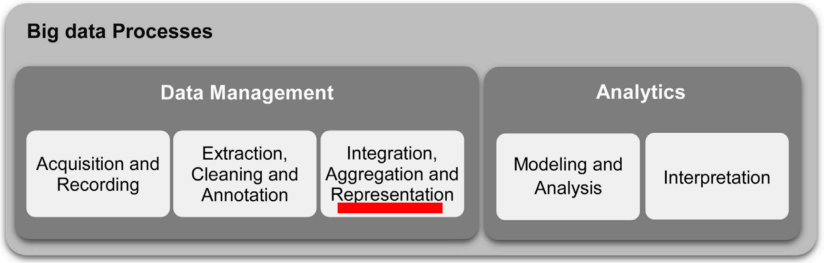


Integration:

- Involves combining data from multiple, complex, heterogenous resources
- Main goal -> Provide a uniform view of data
- Mature field in traditional databases (ETL, data federation)
- Existing traditional approaches -> Based on traditional models
- The Vs of Big Data pose challenges to traditional approaches

https://link.springer.com/article/10.1007/s11036-013-0489-0

Data Management: Integration, Aggregation and Representation

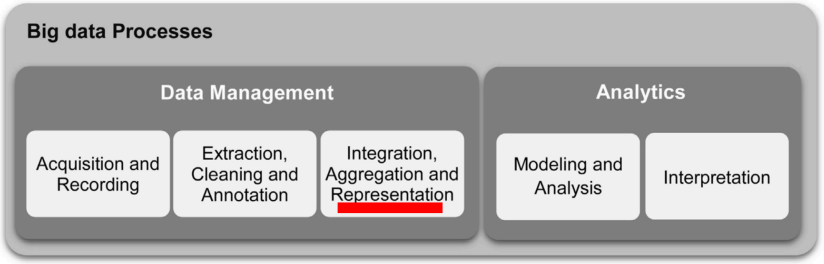


Integration:

- Schema mapping
  - Global schema creation
  - Mapping of global-to-local schema
- Record linkage
  - Same logical entities, different data sources
  - Traditional Record Linkage -> static/structured records, same schema
  - Record Linkage in Big Data -> heterogenous sources, dynamic and continuously evolving
- Data fusion
  - Resolving conflicts
  - Finding truth about real-world -> veracity of data

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6544914

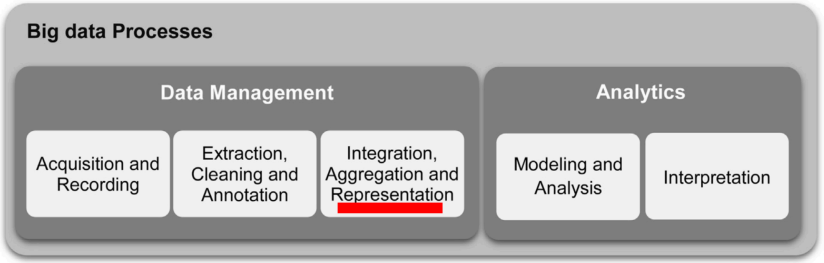
Data Management: Integration, Aggregation and Representation



Aggregation:

- Aggregation of heterogenous information sources
- Distributed, collaborative aggregation (e.g. using Hadoop / MapReduce)
- Privacy, security and compliance (e.g. EU/GDPR)

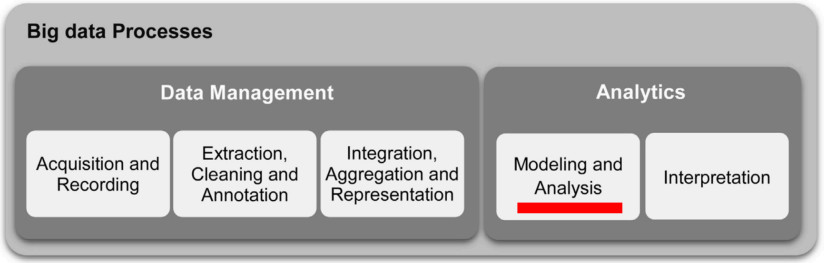
Data Management: Integration, Aggregation and Representation



Representation:

- Different representation models
  - Structured (e.g. relational data)
  - Unstructured (e.g. text and images)
  - Semi-structured (e.g. JSON)
- Mapping from one representation to another
  - Relational to JSON (e.g. MySQL tables to JSON)
  - Text to graph (e.g. Wikipedia corpus to knowledge graphs)
  - Actions/Interactions to graph (e.g. underlying networks in Twitter)

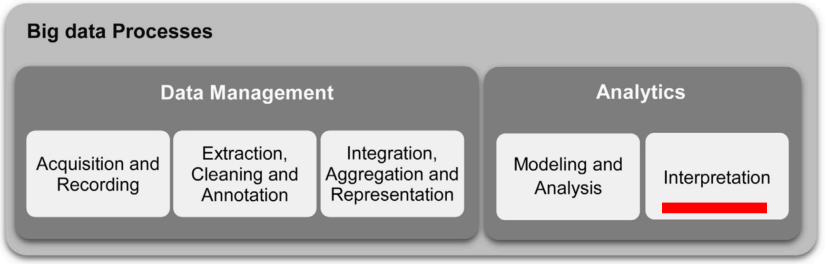
Analytics: Modeling and Analysis



Modeling and Analysis:

- Different modeling techniques based on the analytics goals
  - Descriptive statistics (e.g. histograms, measures of centrality)
  - ML (e.g. classification, regression, clustering)
  - Data mining (e.g. sequential patterns, frequent items)
  - Graph mining (e.g. sub-graph mining)
  - Outliers detection (e.g. clustering)
  - Recommendation (e.g. collaborative filtering)
  - Process mining (e.g. process discovery)

Analytics: Interpretation

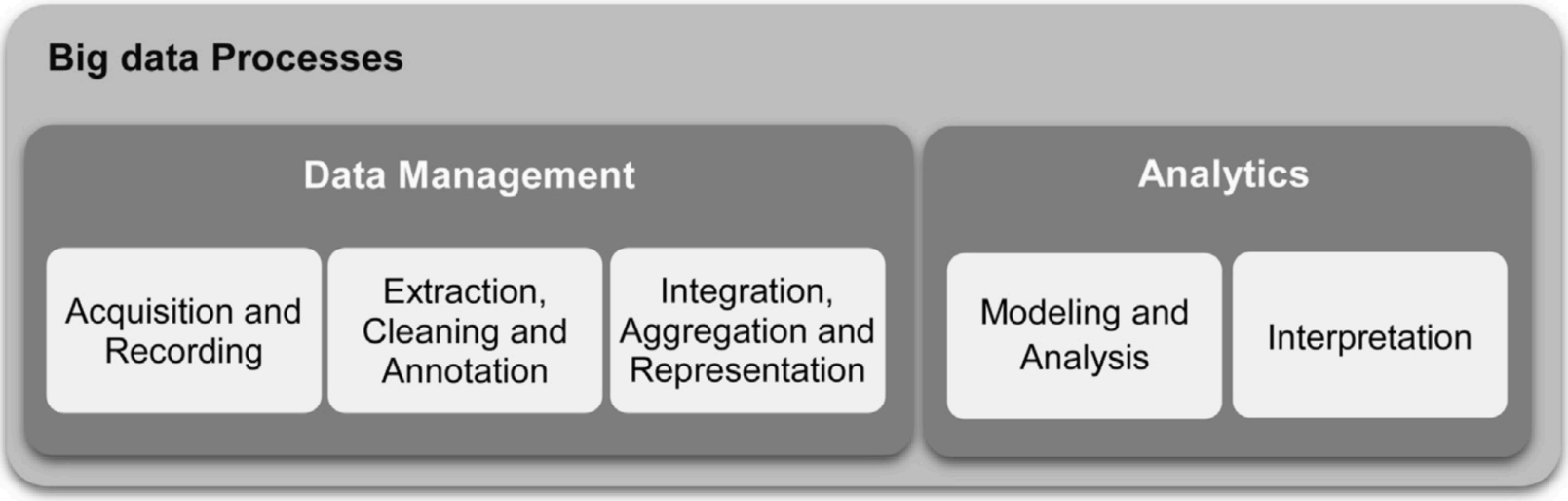


Interpretation:

- Domain experts involvement
- Data/results visualization
  - Spatial data (e.g. bacteria spread over tissue)
  - Geospatial data (e.g. heat map forecast for weather temperatures)
  - Time-oriented data (e.g. interest rates forecast)
  - Multivariate Data (e.g. individual income vs. literacy)
  - Trees, Graphs, Networks (e.g. social network visualization)
  - Text / Document visualizations (e.g. sentiment analysis)

Ward, M. et al. (2015). Interactive data visualization: foundations, techniques, and applications. AK Peters/CRC Press.

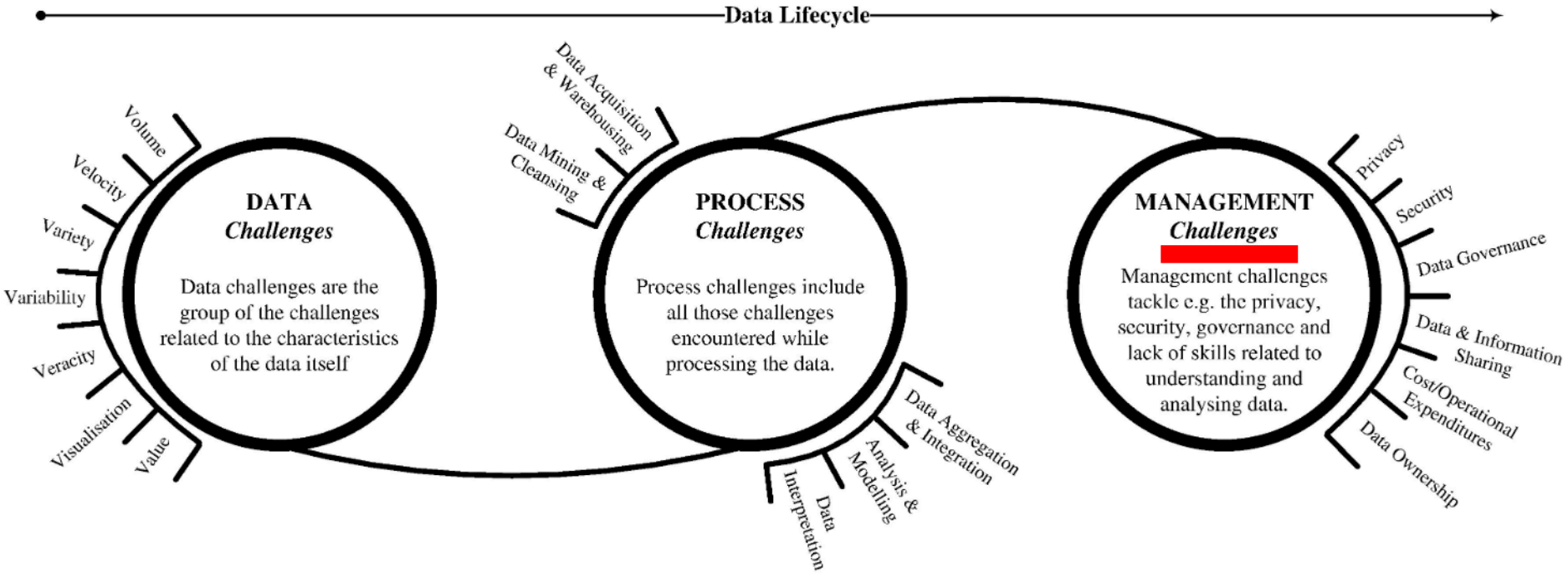
Recap: Big Data Processes



https://www.sciencedirect.com/science/article/pii/S0268401214001066

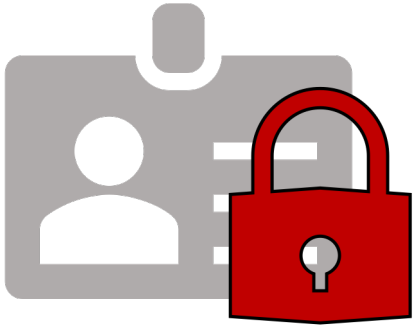
Big Data: Management Challenges

Data Lifecyle



https://www.sciencedirect.com/science/article/pii/S014829631630488X#bb9030

Big Data Privacy



Big Data from two perspectives:

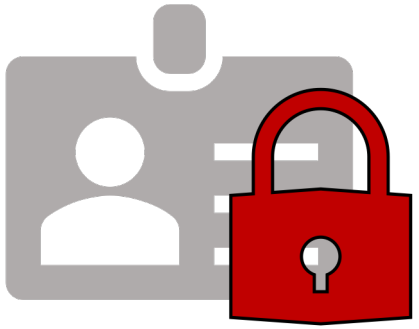
- As a technological challenge
  - Dealing with huge amounts of data
  - The 6 Vs: Volume, Velocity, Variety, Veracity, Visibility, Value
- As a sociological problem
  - Collection of personal data
  - Social media, business, healthcare, government, etc.

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6227909

Big Data Privacy

Uses of Big Data:

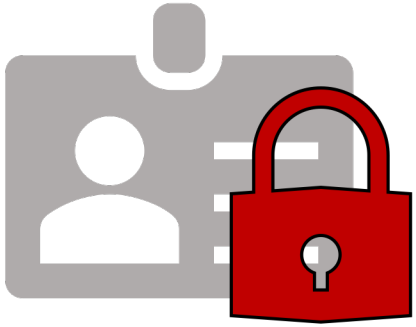
- Market research
- Targeted advertisement
- Workflow improvements
- National security
- Taxation
- and more...



How can we be sure that our data is not used for (intended / unintended) “questionable" purposes?

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6227909

Big Data Privacy



The case of social media

- People uploading their data voluntarily
  - Profiles
  - Pictures
  - Activities
  - and more...
- **Small** Data problem:
  - How can I control access to my own data?
- **Big Data** problem:
  - What can the controlling company do with our data?

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6227909

Big Data Privacy

Different user roles and their concerns

- Data Provider
  - Can I control the sensitivity of the data I provide?
  - Control what the collector can get access to
  - Compensation for privacy loss
- Data Collector
  - Prevent releasing sensitive information to data miner
  - Data must be modified / adapted
  - Modified data that is still useful

https://www.researchgate.net/publication/271967382_Information_Security_in_Big_Data_Privacy_and_Data_Mining

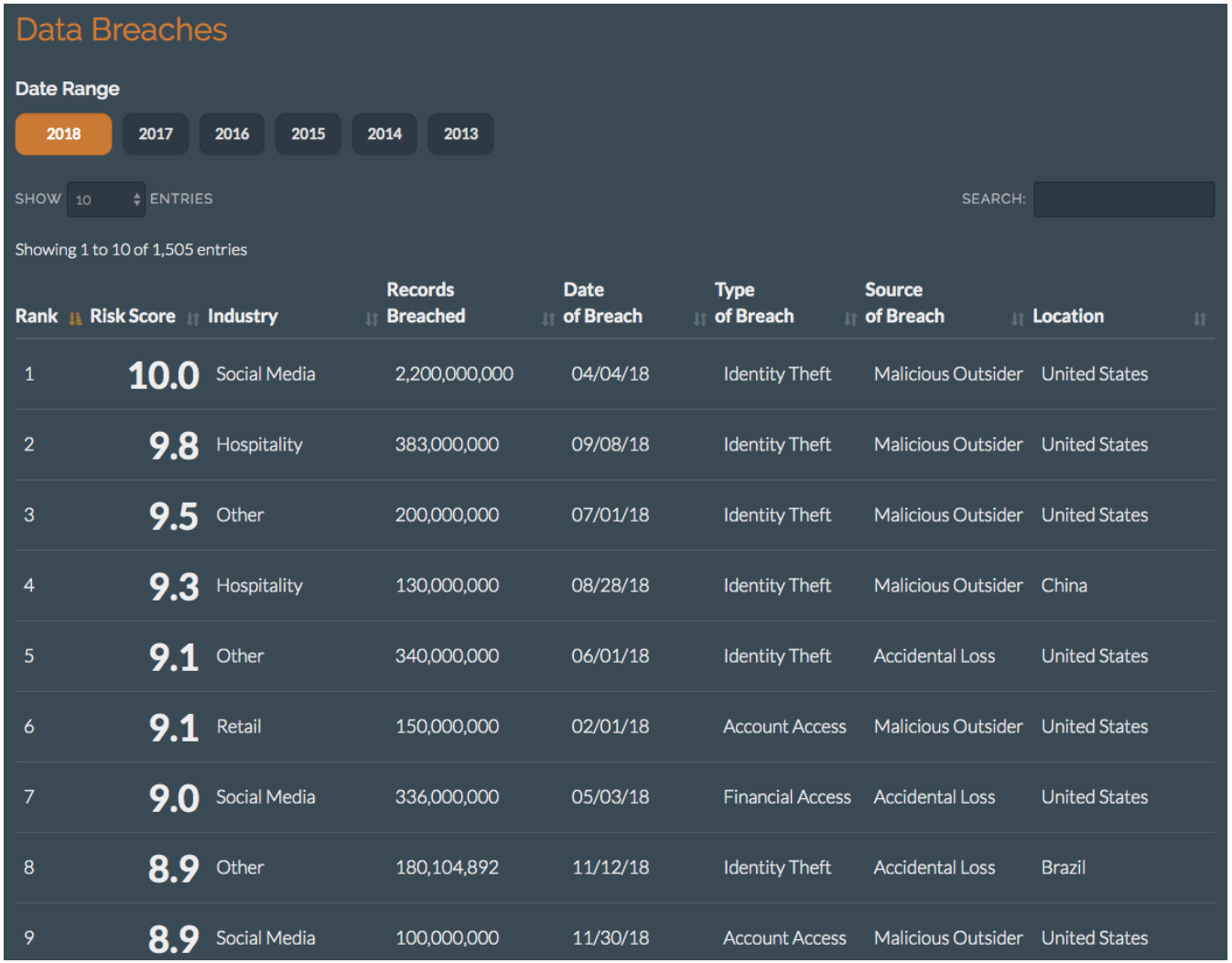
Big Data Privacy

Different user roles and their concerns

- Data Miner
  - Mines data provided by Data Collector
  - Protection of sensitive mining results
- Decision Maker
  - Are the mining results credible?

https://www.researchgate.net/publication/271967382_Information_Security_in_Big_Data_Privacy_and_Data_Mining

Big Data Privacy



Data Breaches

Date Range

2018 | 2017 | 2016 | 2015 | 2014 | 2013

SHOW 10 ENTRIES SEARCH:

Showing 1 to 10 of 1,505 entries

Rank	Risk Score	Industry	Records Breached	Date of Breach	Type of Breach	Source of Breach	Location
1	10.0	Social Media	2,200,000,000	04/04/18	Identity Theft	Malicious Outsider	United States
2	9.8	Hospitality	383,000,000	09/08/18	Identity Theft	Malicious Outsider	United States
3	9.5	Other	200,000,000	07/01/18	Identity Theft	Malicious Outsider	United States
4	9.3	Hospitality	130,000,000	08/28/18	Identity Theft	Malicious Outsider	China
5	9.1	Other	340,000,000	06/01/18	Identity Theft	Accidental Loss	United States
6	9.1	Retail	150,000,000	02/01/18	Account Access	Malicious Outsider	United States
7	9.0	Social Media	336,000,000	05/03/18	Financial Access	Accidental Loss	United States
8	8.9	Other	180,104,892	11/12/18	Identity Theft	Accidental Loss	Brazil
9	8.9	Social Media	100,000,000	11/30/18	Account Access	Malicious Outsider	United States

https://breachlevelindex.com/data-breach-database
(accessed: 16 May 2019)

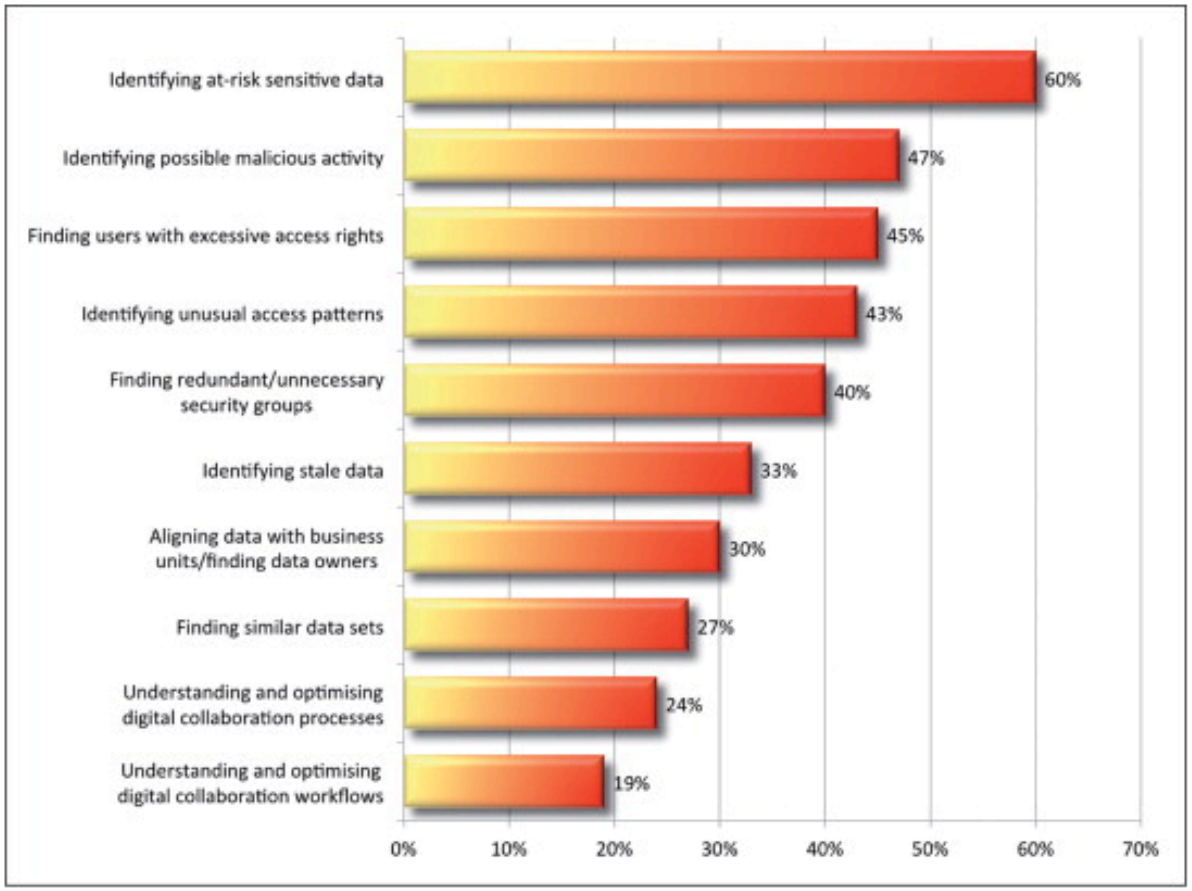
Big Data Security

- Big Data -> Big amount of sensitive information
  - Personal information
  - Intellectual property
  - Trade secrets
  - Financial information
- Huge data breaches
  - Security risks
  - Reputation loss
  - Financial loss

https://www.sciencedirect.com/science/article/pii/S1353485815700097

https://www.sciencedirect.com/science/article/pii/S1353485812700636

Big Data Security



Most significant challenges in managing big data (survey: 151 federal IT professionals, US Government IT community)

https://www.sciencedirect.com/science/article/pii/S1353485812700636

Big Data Security

Data Anonymization

- Removal of sensitive data from records
- Data utility vs. Data privacy
- Removing unique identifiers -> Privacy not guaranteed
- Cross-references and de-anonymization

https://www.sciencedirect.com/science/article/pii/S1353485815700097

Big Data Security

Data Encryption

- Encrypt stored data
- Attribute-based encryption -> fine-grained access control

https://www.sciencedirect.com/science/article/pii/S1353485815700097

Big Data Security

Access control and Monitoring

- Access control policy for privileged operations (e.g. encrypt / decrypt)
- Authentication mechanisms not always available in big data frameworks
- Real-time security monitoring -> Attacks detected in real-time and mitigated properly

https://www.sciencedirect.com/science/article/pii/S1353485815700097

Big Data Security

Policy approaches

- Identification of sensitive pieces of information
- Isolation of sensitive information
- Third-party involvement -> Compliance with regulations and standards
- Processes to manage and protect data effectively

https://www.sciencedirect.com/science/article/pii/S1353485815700097

Big Data Security

Governance and Frameworks

- Use of governance frameworks to protect data (e.g. ISO/IEC 27001 - Information Security)
- Big data is relatively new -> No major frameworks with policies and procedures specifically addressing the security big data challenges
- Main Challenge:
  - Variety -> Difficulty in categorizing, modeling and mapping big data

https://www.sciencedirect.com/science/article/pii/S1353485815700097

Big Data Security

Some good practices

- Cloud provider vetting (i.e. assessment)
- Adequate access control policies
- Data protection (e.g. using encryption)
- Communications protection (e.g. for confidentiality and integrity)
- Real-time security monitoring

https://www.sciencedirect.com/science/article/pii/S1353485815700097

Compliance

- **The Oxford dictionary[1]:**
  - "The state or fact of according with or meeting rules or standards"
- **Wikipedia[2]:**
  - "Conforming to a rule, such as a specification, policy, standard or law"

[1] https://www.oxforddictionaries.com/

[2] https://en.wikipedia.org/wiki/Regulatory_compliance

Compliance

Example regulations:

source

- Health Insurance Portability & Accountability Act
- Protection of Personal Identifiable Information from fraud and theft
- Mainly applicable to healthcare and healthcare insurance companies
- Consists of 5 ”Titles”
- Privacy is treated mostly under Title II - *Preventing Health Care Fraud and Abuse*

- General Data Protection Regulation
- EU regulation on data protection and privacy for citizens of the EU and EEA
- Export of data outside EU and EEA
- Gives controls to individuals over their personal data
- Data Controllers -> Must provide technical and organizational measures to implement GDPR
- Hefty fines for violations of GDPR

https://digital.library.unt.edu/explore/collections/CRSR/browse/?q=hipaa&t=fulltext

https://ec.europa.eu/info/law/law-topic/data-protection_en

Compliance

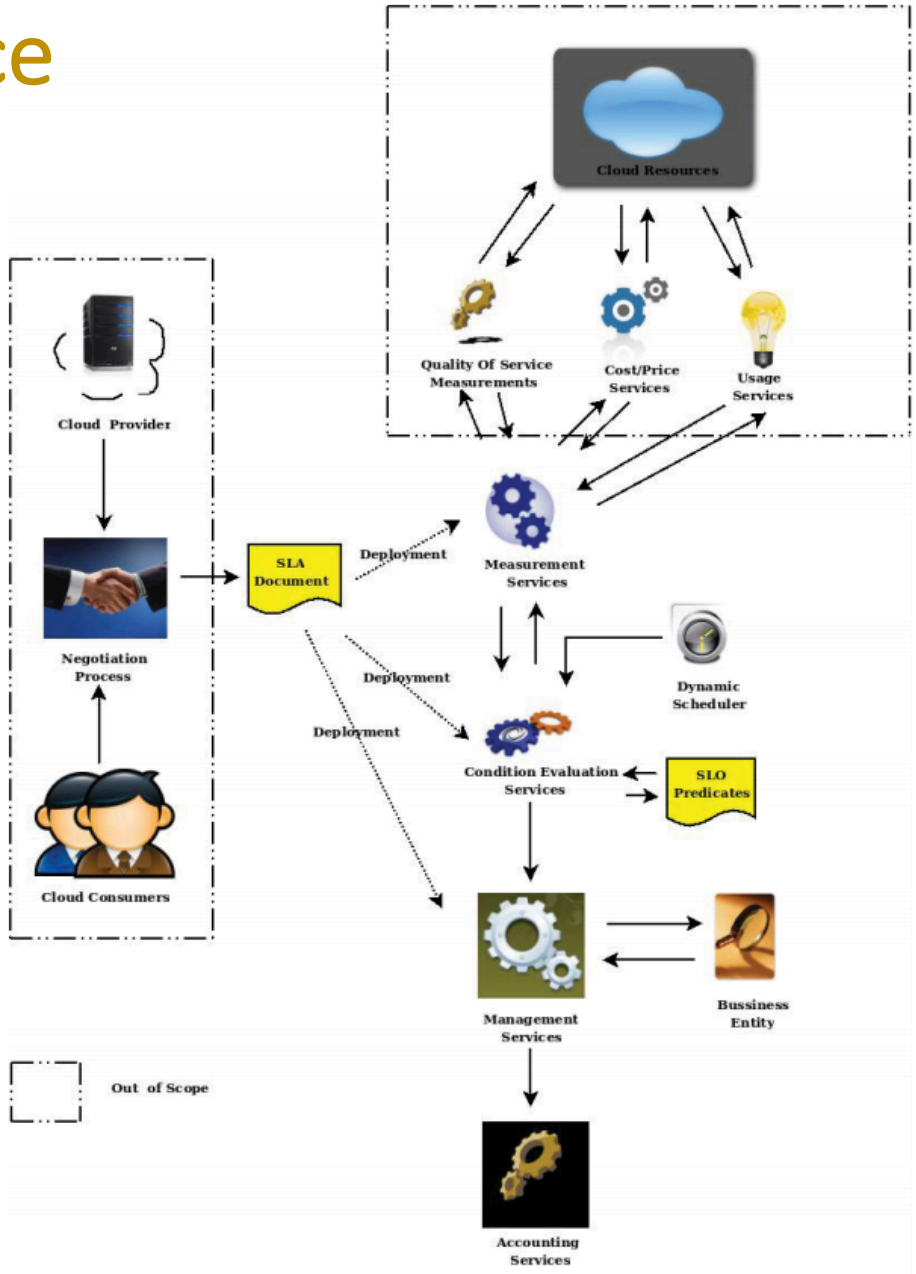
Big Data SLAs (Service-Level Agreements)

- Companies moving from experimentation to production (mission critical)
- Getting more serious about QoS (Quality of Service)
- Minimum performance and service levels required (e.g. response time and throughput)
- Service-Level Agreements:
  - Commitment between a provider and a user of service

https://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=1077&context=knoesis

https://www.techrepublic.com/article/how-to-start-using-big-data-slas/

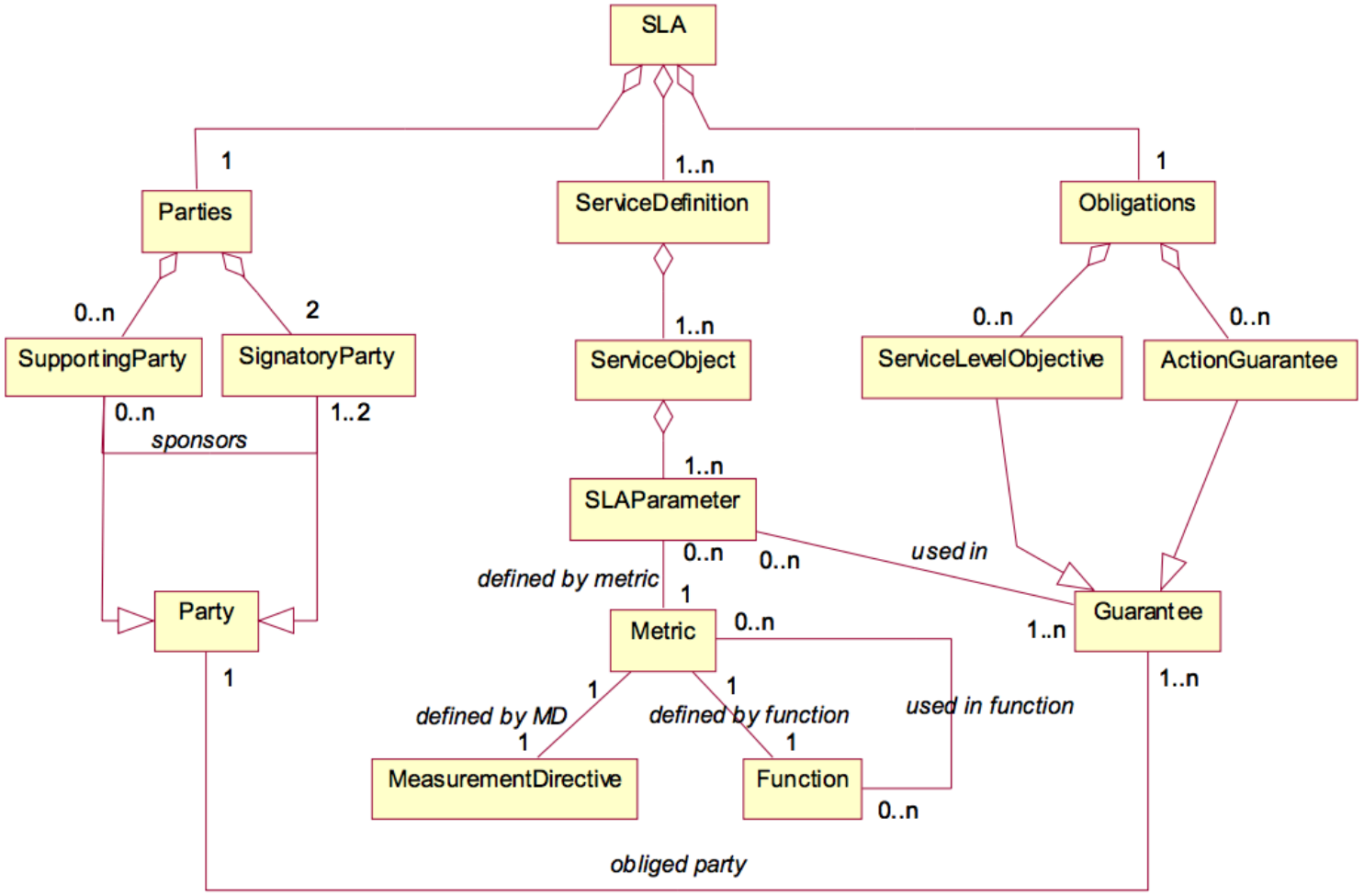
Compliance



https://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=1077&context=knoesis

Compliance

Web Service Level Agreement (WSLA)



http://cliplab.org/Projects/S-CUBE/papers/ludwig03:wsla-ibm.pdf

Compliance

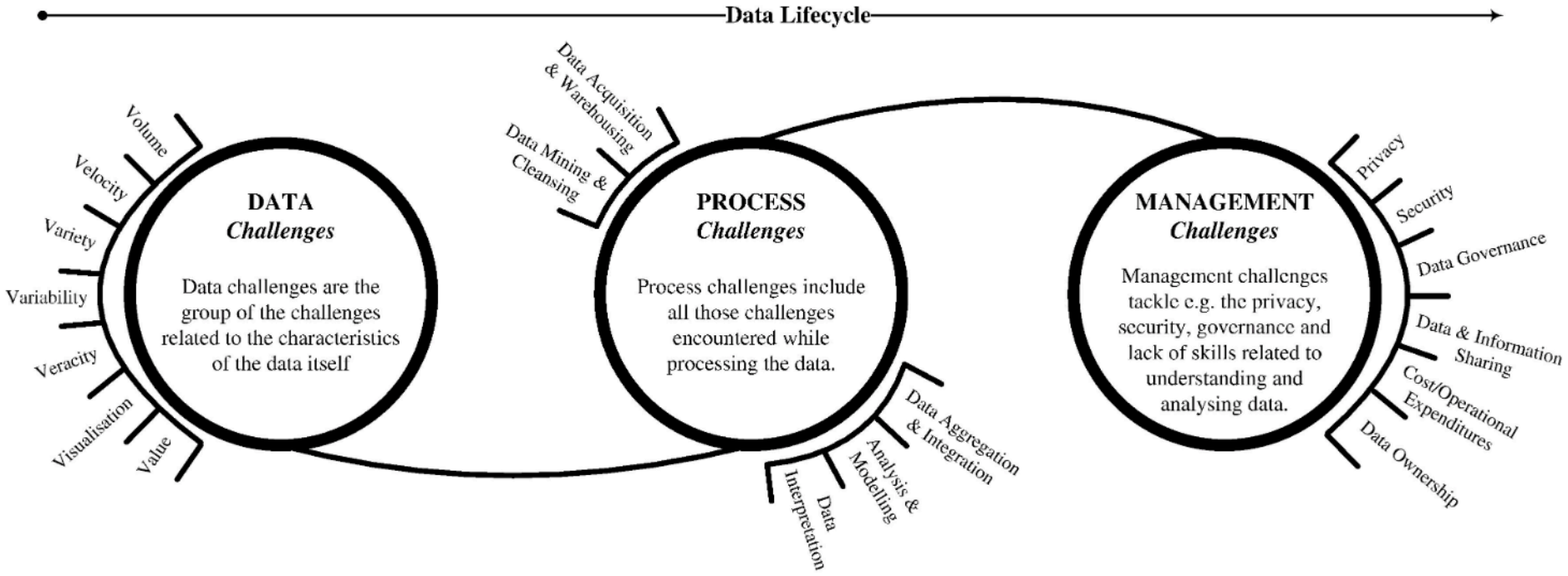
Web Service Level Agreement (WSLA)

```
<ServiceProvider
    name="ACMEProvider">
    <Contact>
        <Street>PO BOX 218</Street>
        <City>Yorktown, NY 10598, USA</City>
    </Contact>
    <Action name="notification"
            partyName="ACMEProvider"
            xsi:type="WSDLSOAPActionDescriptionType">
       <WSDLFile>notification.wsdl</WSDLFile>
       <SOAPBindingName>soapnotification</SOAPBindingName>
       <SOAPOperationName>notification</SOAPOperationName>
    </Action>
</ServiceProvider>
<ServiceConsumer
    name="XInc">
    <Contact>
        <Street>30 Saw Mill River RD</Street>
        <City>Hawthorne, NY 10532, USA</City>
    </Contact>
    <Action name="notification"
            partyName="XInc"
            xsi:type="WSDLSOAPActionDescriptionType">
       <WSDLFile>notification.wsdl</WSDLFile>
       <SOAPBindingName>soapnotification</SOAPBindingName>
       <SOAPOperationName>notification</SOAPOperationName>
    </Action>
</ServiceConsumer>
```

Example: Definition of service provider and consumer

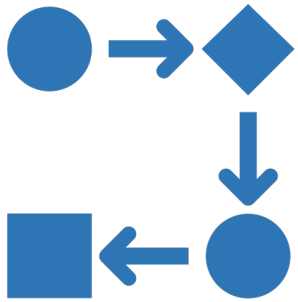
http://cliplab.org/Projects/S-CUBE/papers/ludwig03:wsla-ibm.pdf

Big Data Lifecyle



https://www.sciencedirect.com/science/article/pii/S014829631630488X#bb9030

Data Governance



Definition:

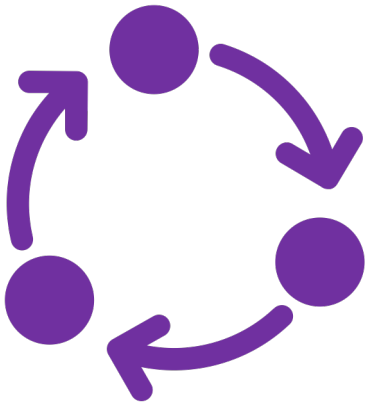
"Data governance is the overall management of the availability, usability, integrity and security of data used in an enterprise. A sound data governance program includes a governing body or council, a defined set of procedures and a plan to execute those procedures."

https://searchdatamanagement.techtarget.com/definition/data-governance

- Data Governance as an approach to:
  - Warranting data quality
  - Improving / leveraging information
  - Support for insights into business decisions and operations

https://www.sciencedirect.com/science/article/pii/S014829631630488X#bb9030

Data and Information Sharing



- Helps establish close connections with business partners
- Main challenges:
  - Sharing / integration of key information across different organizations
  - Disparate information / Data warehouses
  - Sensitive information / User privacy rights

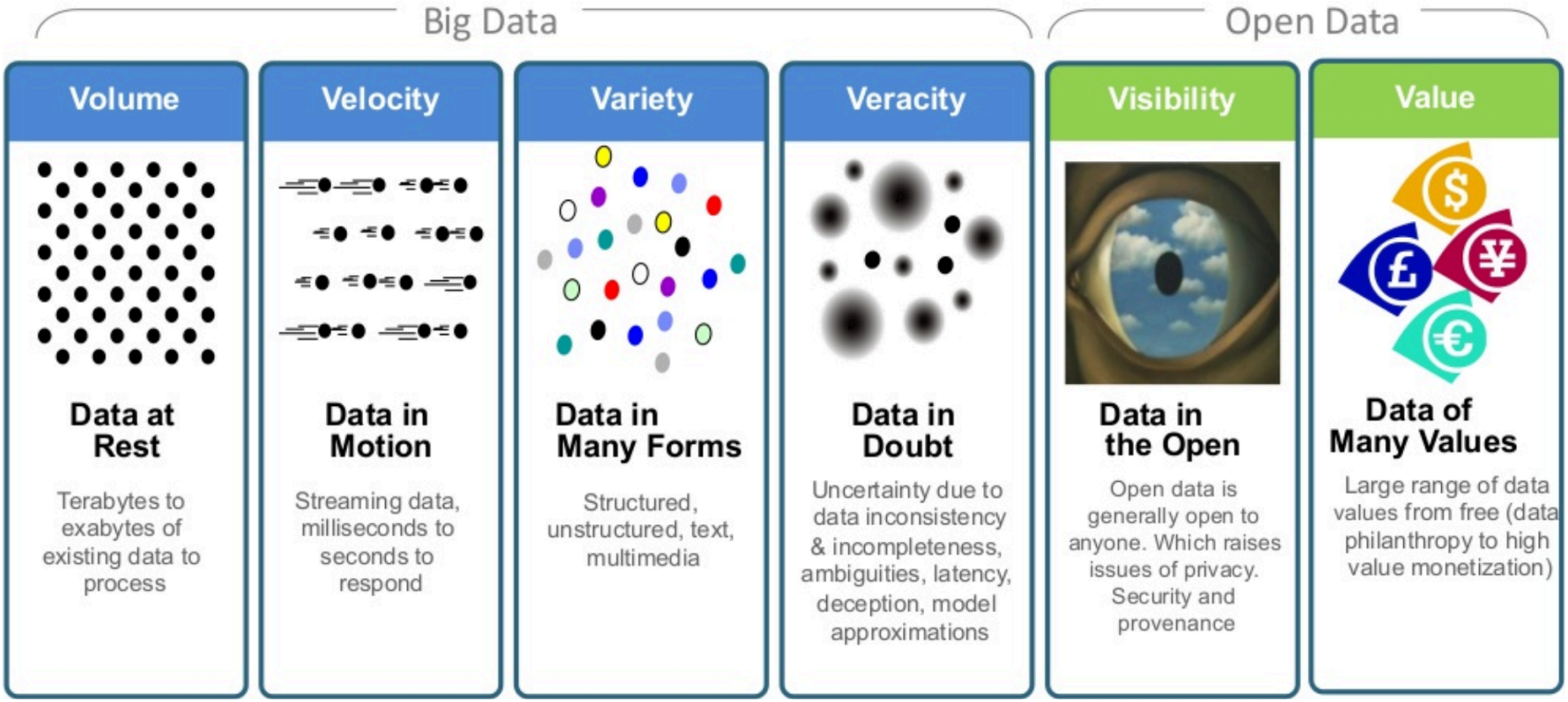
https://www.sciencedirect.com/science/article/pii/S014829631630488X#bb9030

Cost / Operational Expenditure

- Rising demand for Big Data processing
- Spread through geographical regions -> Build resilience and spread risk
- Big IT companies -> Data centers in different countries and continents
- Resources allocated to key operations:
  - Acquisition
  - Warehousing
  - Mining and cleaning
  - Aggregation and integration
  - Processing and interpretation

https://www.sciencedirect.com/science/article/pii/S014829631630488X#bb9030

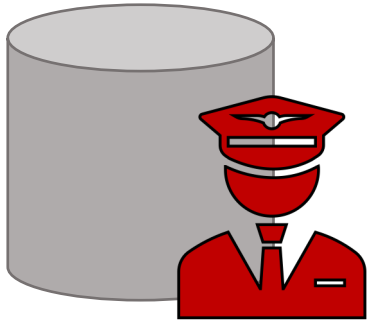
Cost / Operational Expenditure



Transforming Energy and Utilities through Big Data & Analytics
By Anders Quitzau, IBM, 2014

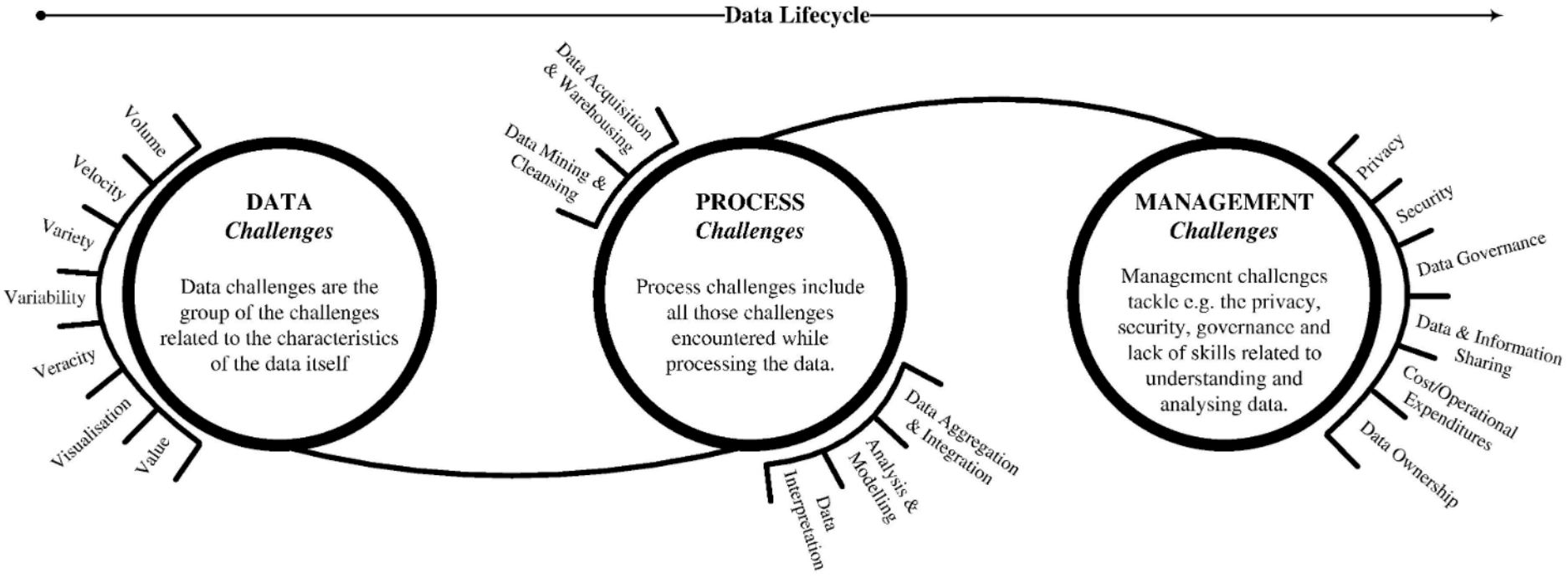
Data Ownership

- Complex issue, specially in the context of real time sharing of data
- Social media:
  - What's own by users who post content?
  - What's own by social media providers?
- Data ownership also involves:
  - Monitoring data accuracy
  - Ensuring data data accuracy



https://www.sciencedirect.com/science/article/pii/S014829631630488X#bb9030

Big Data Lifecyle



https://www.sciencedirect.com/science/article/pii/S014829631630488X#bb9030

Big Data Management Dos and Don’ts

Big Data Management Don'ts...

Assume very ambitious approaches will result in best returns

- Embracing complex, heterogenous, rich data and trying to do all at once
- Instead:
  - Focus on smaller, useful pilot applications
  - Build on top of initial small successes
  - Pick "the right" project to test your organization's capabilities

https://www.sas.com/en_us/insights/articles/big-data/big-data-dos-and-donts.html

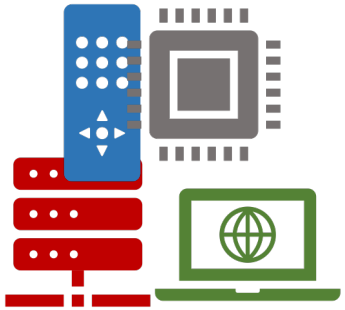
Big Data Management Don’ts…

Focusing all efforts on the needs of particular business units

- Think globally, act locally
- Safer to start by focusing on decreasing risks, than creating new opportunities
- The latter is harder, and takes longer to measure -> Requires experience

https://www.sas.com/en_us/insights/articles/big-data/big-data-dos-and-donts.html

Big Data Management Don’ts...



Assuming that technology alone will guarantee desired results

- Lack of technology is one of the top obstacles for business success
- Technology is already there...
- How to properly leverage on technology in the organization ecosystem?
- Focus should be on achieving good ROI

https://www.sas.com/en_us/insights/articles/big-data/big-data-dos-and-donts.html

Big Data Management Dos...

Building collaborative capabilities

- Processes and toolsets -> Instrumental to making big data more approachable
- Big data growing faster than capabilities
- Environment should facilitate ease of use
- Business and IT collaborations are key to harness opportunities of big data

https://www.sas.com/en_us/insights/articles/big-data/big-data-dos-and-donts.html

Big Data Management Dos...

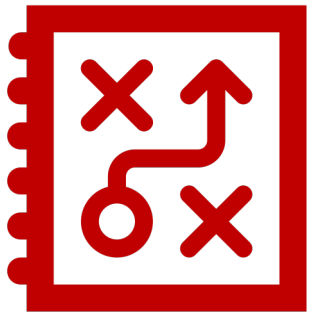
Use a stepwise approach

X Tackling unknow problem with unknown data

- More convenient:
  - Starting with a known problem
  - Solving it in an innovative way
- Next: Solve same problem, use new data
- Finally: Solve new problem, use new data
- Stepwise approach -> Higher likelihood of success

https://www.sas.com/en_us/insights/articles/big-data/big-data-dos-and-donts.html

Big Data Management Dos...

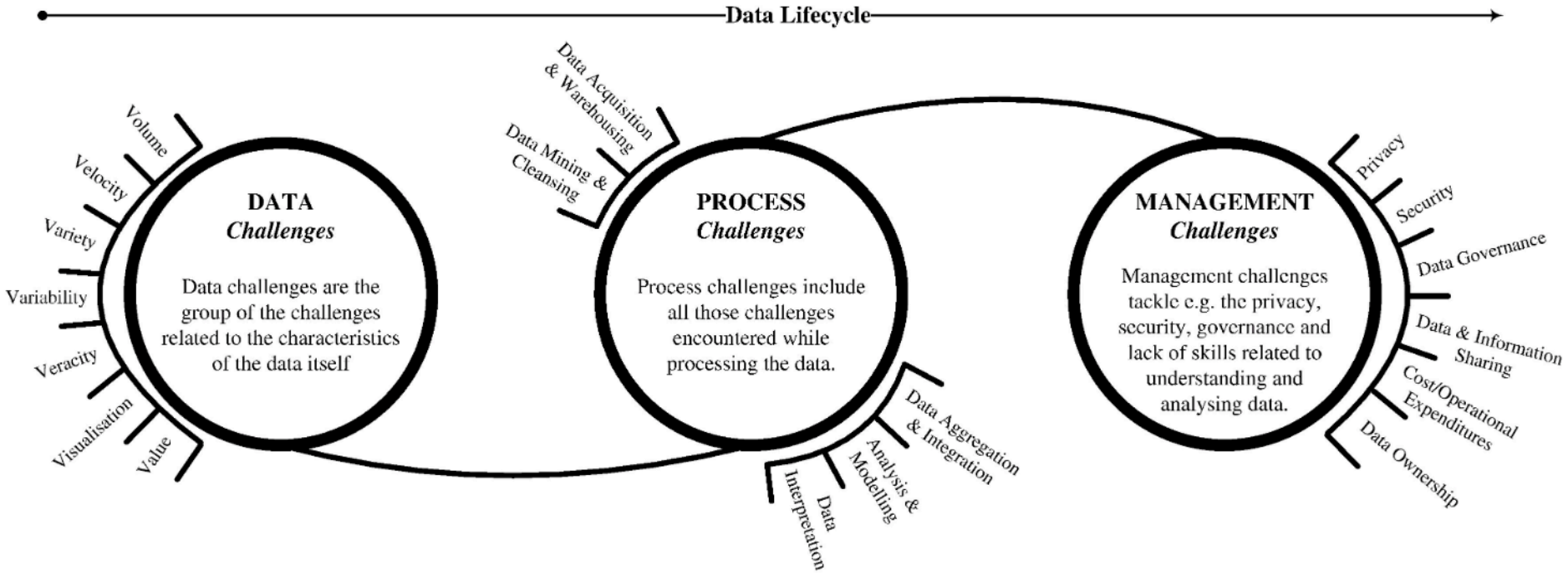


Think strategically, act tactically

- Organizations tend to rush and build platforms to solve very specific problems:
  - Program seen as experiment only
  - Hard to evolve and integrate as a business asset
- **Better:** Establish strategic goals -> Makes it easier to envision other applications of Big Data
- Continuous application of Big Data -> Better realization of its true value and potential

https://www.sas.com/en_us/insights/articles/big-data/big-data-dos-and-donts.html

Recap: Big Data Lifecyle



https://www.sciencedirect.com/science/article/pii/S014829631630488X#bb9030

Thanks