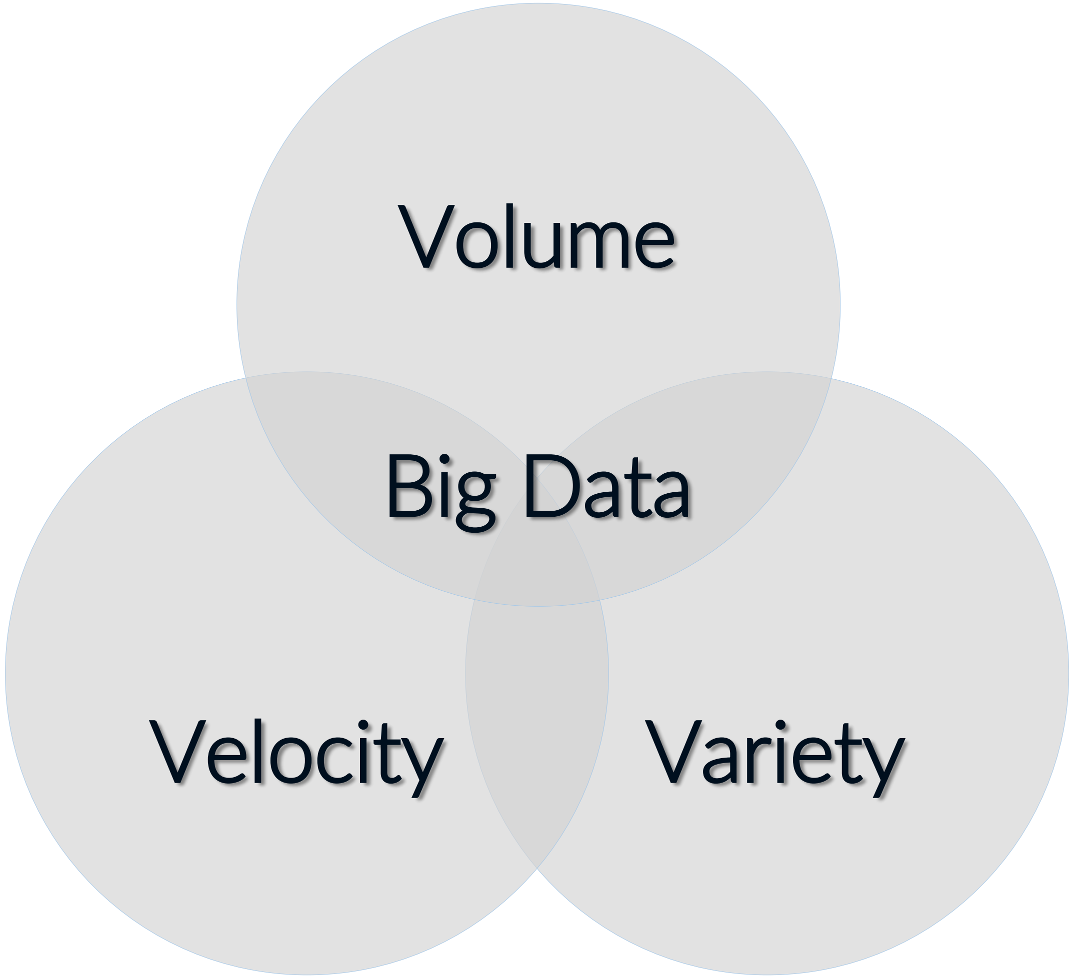
Big Data Storage
Challenges, Solutions & Trends

Oliver Bruski
Oleksandr Chumak

How do we define Big Data?



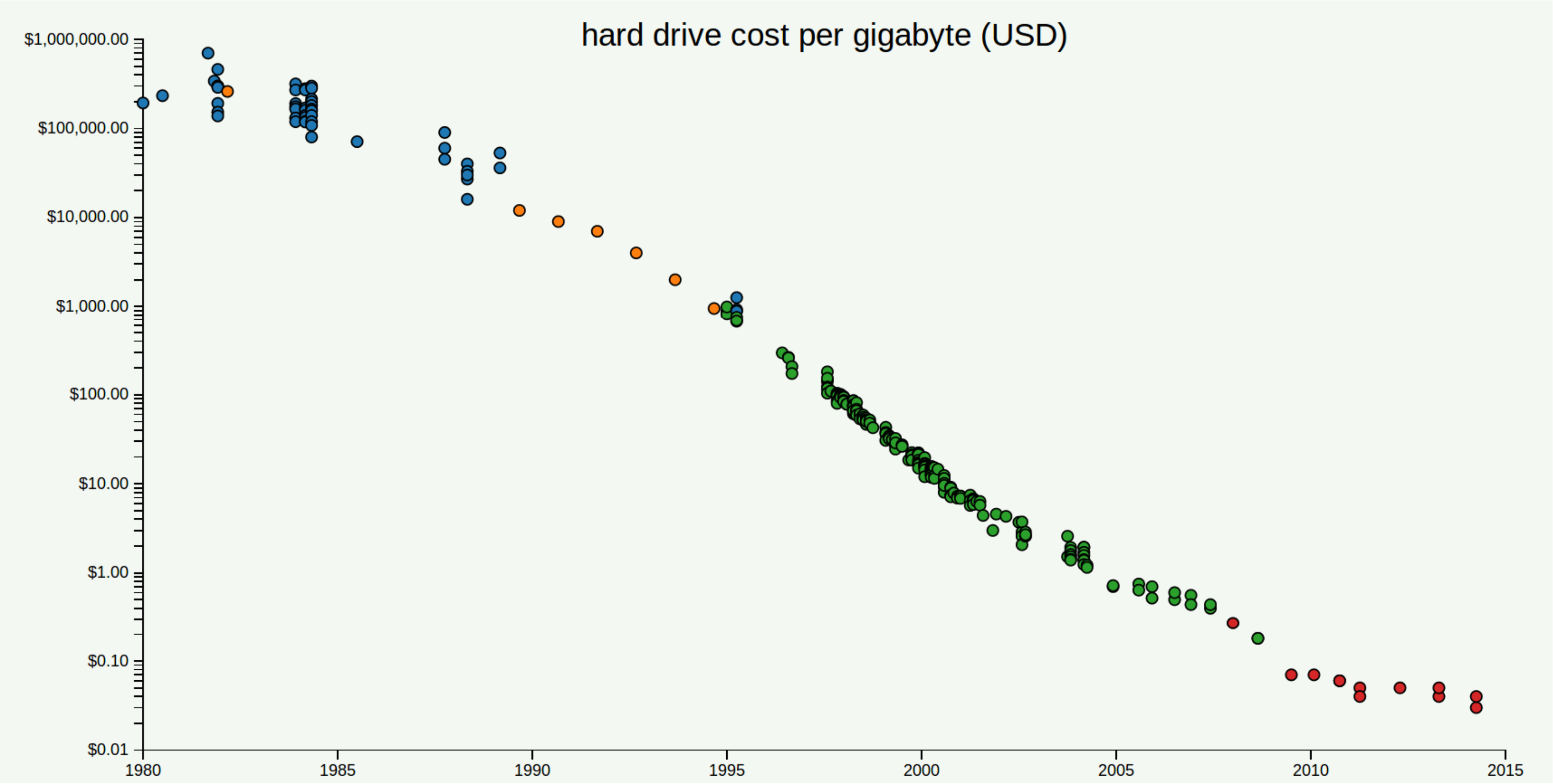
derived from Dominik Klein, Phuoc Tran-Gia, and Matthias Hartmann. "Big data." Informatik-Spektrum 36.3 (2013): 319-323.

Agenda

1. What Challenges come with Big Data?
2. Current Approaches to overcome the Challenges
    1. Distributed File Systems
    2. NoSQL Data Stores
    3. NewSQL Data Stores
    4. Cloud-Based Storage
3. Trends and Current Research

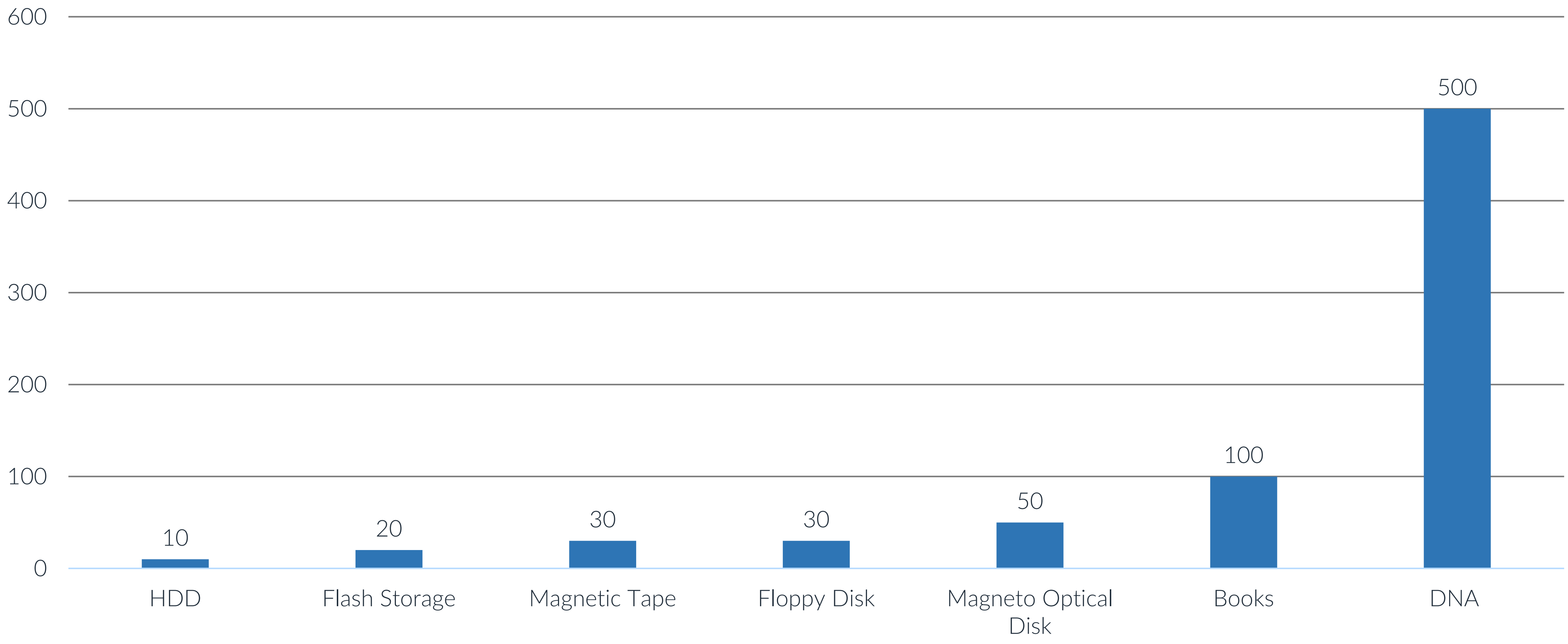
Big Data Challenges

The price to store a GB of data is decreasing continuously



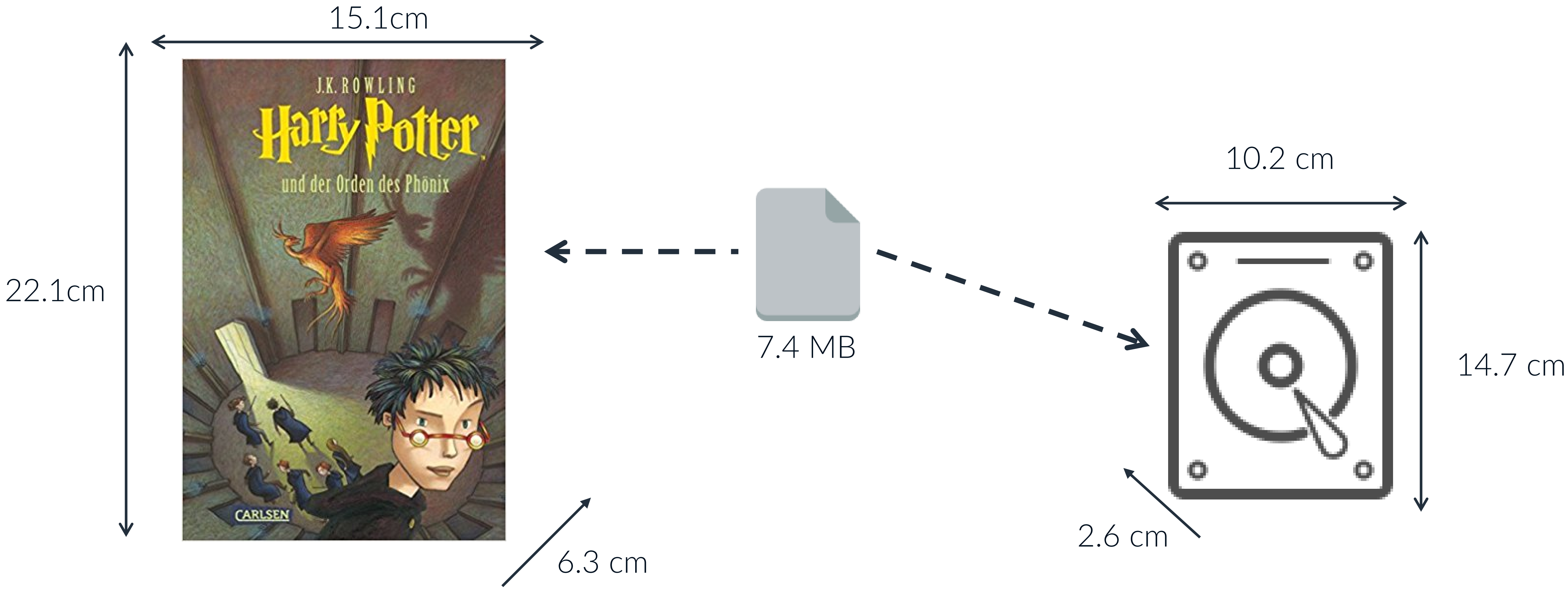
Source: http://www.mkomo.com/assets/cost-per-gigabyte-large.png

The expected Lifespan of storage systems



Source: Schwens, Ute, and Hans Liegmann. "Langzeitarchivierung digitaler ressourcen." *Grundlagen der praktischen Information und Dokumentation* 5 (2004): 567-570.

Storage systems from a different point of view – Data Density



Source: https://www.amazon.de/Harry-Potter-Ph%C3%B6nix-Joanne-Rowling/dp/3551555559/ref=tmm_hrd_swatch_0?_encoding=UTF8&qid=1495797735&sr=8-1; http://www.seagate.com/www-content/product-content/ironwolf/en-us/docs/ironwolf-hdd-ds-1904-8-1703us.pdf

The Volume of Data created is doubling every 2 years

- Study by IDC expects the size of the digital universe to be 44 ZB
- Not every created Bit needs to be stored
- Researchers at CERN only store data from certain experiments
  - 0.1% of data created is analyzed

Source: Turner, Vernon, et al. "The digital universe of opportunities: Rich data and the increasing value of the internet of things." IDC Analyze the Future (2014).
https://neic.nordforsk.org/2017/04/07/the-smartest-storage-for-the-worlds-largest-data-volumes.html

Source: https://www.domo.com/blog/data-never-sleeps-4-0/

What the Challenges Big Data creates

Volume

- Physical space to store data
- Keep data over time -> Replication
- Data locality

Variety

- New ways to store and query unstructered data
- Combination of heterogenous data stores

Velocity

- Store generated data and analyze in real-time
- Generate Value before data becomes outdated

Source: Pusala, Murali K., et al. "Massive Data Analysis: Tasks, Tools, Applications, and Challenges." Big Data Analytics. Springer India, 2016. 11-40.

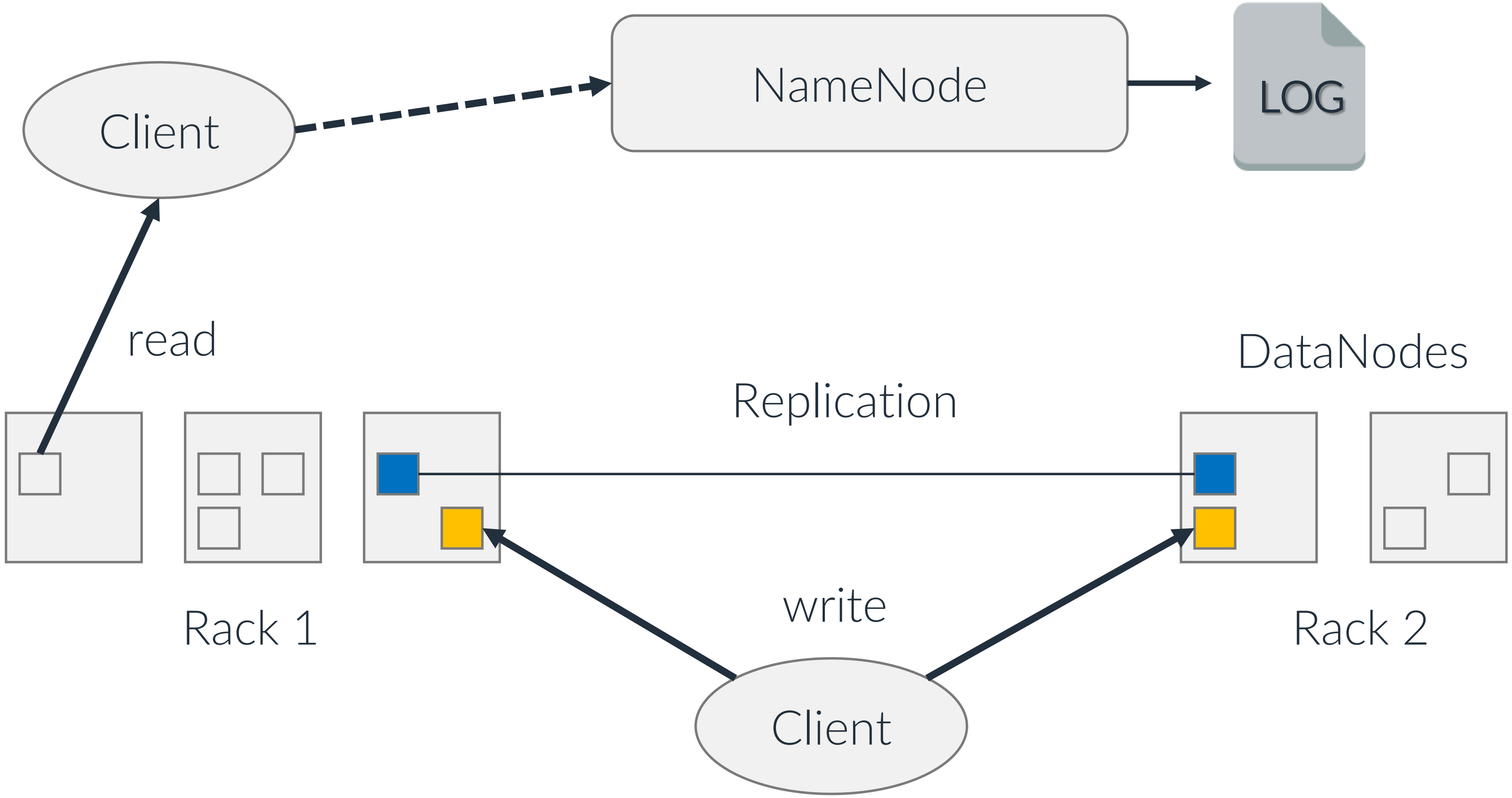

Solutions

What are the main concepts of a distributed file system

- Storage of data on several physical nodes rather than on local resources of a single machine
- Build on commodity hardware
  - Fault-tolerant + highly scalable file system
- Stores large data in blocks
- Master-slave architecture
- Typically write once and read multiple times

Source: Borthakur, Dhruba. "HDFS architecture guide." *Hadoop Apache Project* 53 (2008).
Strohbach, Martin, et al. "Big Data Storage." *New Horizons for a Data-Driven Economy*. Springer International Publishing, 2016. 119-141.

A generic architecture of HDFS



derived from https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

The features of HDFS

- Keeps track of changes using an in-memory EditLog
- DataNode stores files physical on its local file system
  - Data blocks are replicated among several DataNodes, typically 3
  - HDFS uses rack-aware replication (2 +1)
- DataNodes periodically send heartbeat to NameNode
- Writing a block is done in small chunks -> data piplining is possible

Source: Borthakur, Dhruba. "HDFS architecture guide." Hadoop Apache Project 53 (2008).
Strohbach, Martin, et al. "Big Data Storage." *New Horizons for a Data-Driven Economy*. Springer International Publishing, 2016. 119-141.

Alternative distributed file systems

GFS

- Proprietary distributed file system for mainly reading large files
- Master server and several Chunkservers

Ceph

- Interface for object, block and file storage
- No single-point-of-failure using cluster monitors + metadata servers

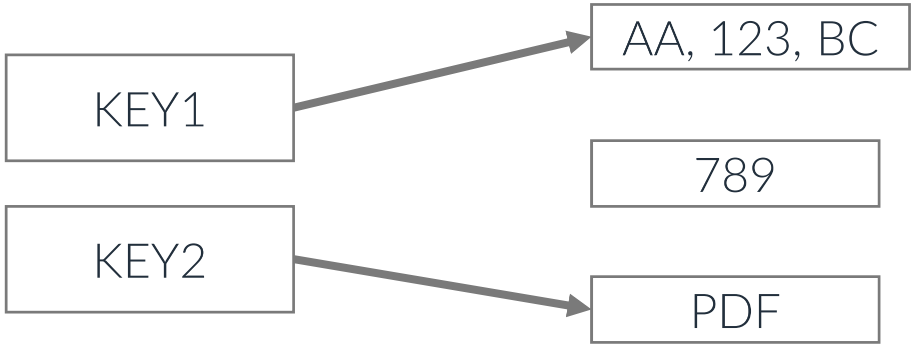
Source: Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. "The Google file system." ACM SIGOPS operating systems review. Vol. 37. No. 5. ACM, 2003.; Weil, Sage A., et al. "Ceph: A scalable, high-performance distributed file system." *Proceedings of the 7th symposium on Operating systems design and implementation*. USENIX Association, 2006.

A new form of databases has emerged NoSQL

- Focus is on fast read/write
- Highly scalable on commodity hardware
- Atomicity, Consistency, Isolation, Durability
- Trade-off ACID for better performance
- Basically Available, Soft State, Eventually Consistent (BASE)

Source: Sakr, Sherif. "Big Data 2.0 Processing Systems: A Survey." SpringerBriefs in computer science ((2016).
Han, Jing, et al. "Survey on NoSQL database." *Pervasive computing and applications (ICPCA), 2011 6th international conference on*. IEEE, 2011.

The different types of NoSQL databases



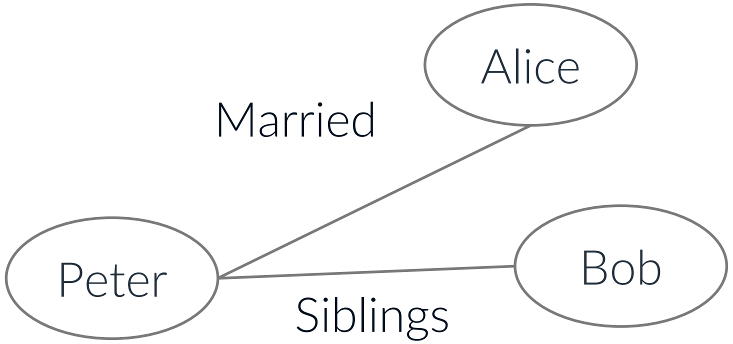
Key-Value Store

ID	Name	Age	CV
1	John	23	/CV/john.pdf
2	Jane	34	/CV/jane.pdf
3	John2	45	/CV/john2.pdf

Wide Column Store

```
{       _id: “123”,
        name: “john doe”,
        email: “john.doe@example.com”,
        address: {
                street: “Main Street 1”,
                city: “Berlin”,
}       }
```

Document Database



Graph Store

Key-Value Stores

- Store data as tuples of Key and Value (K, V)
- Key is usually an autogenerated ID
- Value can be of **any** type
- Data is stored in buckets
- Retrieve data fast and easy, no need for complex queries

Source: Strohbach, Martin, et al. "Big Data Storage." New Horizons for a Data-Driven Economy. Springer International Publishing, 2016. 119-141.

Wide Column Store

- Store a column in continuous blocks on disk
- Speeds up aggregations over a attribute
- Values in cells can be of any type
- Values have timestamps -> Versioning
- Columns can be added dynamically

Source: Strohbach, Martin, et al. "Big Data Storage." New Horizons for a Data-Driven Economy. Springer International Publishing, 2016. 119-141.

Document Database

- Also key-value but with deeper nesting format
- Value is stored as document e.g. JSON
- Easier to distribute and maintain data locality
- Need to load a lot of data to update a single value in a record

Source: Strohbach, Martin, et al. "Big Data Storage." New Horizons for a Data-Driven Economy. Springer International Publishing, 2016. 119-141.

Graph Database

- Stores data as a Graph
- Vertices represent entities
- Edges represent the relations between entities
- Edges can have timestamps
- Often used for social media analysis

Source: Strohbach, Martin, et al. "Big Data Storage." New Horizons for a Data-Driven Economy. Springer International Publishing, 2016. 119-141.

The next Buzzword – NewSQL

- Better scalability in comparison to NoSQL
- ACID support for transactions
- A non-locking concurrency control mechanism. MVCC (Multi Version Concurrency Control)
    - provides each transaction a snapshot of the data thus each transaction gets a consistent view of the database
    - Instead of overwriting existing documents, a completely new version of the document is generated
- A scale-out, shared-nothing architecture, capable of running on a large number of nodes without suffering bottlenecks
- About 50 times faster than traditional OLTP RDBMS
- VoltDB scales linearly in the case of single-partition queries

Source: Strohbach, M., Daubert, J., Ravkin, H., & Lischka, M. (2016). Big Data Storage. In New Horizons for a Data-Driven Economy (pp. 119-141). Springer International Publishing.

The State of the Art

- Main memory storage
  - identify which tuples are not being accessed anymore and then choose them for eviction
  - H-Store’s anti-caching
  - VoltDB retains the keys for evicted tuples in databases’ indexes
- Partitioning/Sharding
  - The database’s tables are horizontally divided into multiple fragments
  - Related fragments from multiple tables are combined together to form a partition that is managed by a single node
  - NuoDB: nodes divided into storage managers (SM, split DB into blocks) and transaction engines (TE)
- Concurrency control
  - Timestamp ordering, MVCC
- Secondary Indexes
  - Partitioned. Each node stores a portion of the index
  - If a transaction updates an index it will only have to modify one node

Source: Pavlo, Andrew, and Matthew Aslett. "What's Really New with NewSQL?." ACM SIGMOD Record 45.2 (2016): 45-55.

NewSQL Databases overview

		Year Released	Main Memory Storage	Partitioning	Concurrency Control	Replication	Summary
NEW ARCHITECTURES	**Clustrix** [6]	2006	No	Yes	MVCC+2PL	Strong+Passive	MySQL-compatible DBMS that supports shared-nothing, distributed execution.
	CockroachDB [7]	2014	No	Yes	MVCC	Strong+Passive	Built on top of distributed key/value store. Uses software hybrid clocks for WAN replication.
	Google Spanner [24]	2012	No	Yes	MVCC+2PL	Strong+Passive	WAN-replicated, shared-nothing DBMS that uses special hardware for timestamp generation.
	H-Store [8]	2007	Yes	Yes	TO	Strong+Active	Single-threaded execution engines per partition. Optimized for stored procedures.
	HyPer [9]	2010	Yes	Yes	MVCC	Strong+Passive	HTAP DBMS that uses query compilation and memory efficient indexes.
	MemSQL [11]	2012	Yes	Yes	MVCC	Strong+Passive	Distributed, shared-nothing DBMS using compiled queries. Supports MySQL wire protocol.
	NuoDB [14]	2013	Yes	Yes	MVCC	Strong+Passive	Split architecture with multiple in-memory executor nodes and a single shared storage node.
	SAP HANA [55]	2010	Yes	Yes	MVCC	Strong+Passive	Hybrid storage (rows + cols). Amalgamation of previous TREX, P*TIME, and MaxDB systems.
	VoltDB [17]	2008	Yes	Yes	TO	Strong+Active	Single-threaded execution engines per partition. Supports streaming operators.
MIDDLEWARE	**AgilData** [1]	2007	No	Yes	MVCC+2PL	Strong+Passive	Shared-nothing database sharding over single-node MySQL instances.
	MariaDB MaxScale [10]	2015	No	Yes	MVCC+2PL	Strong+Passive	Query router that supports custom SQL rewriting. Relies on MySQL Cluster for coordination.
	ScaleArc [15]	2009	No	Yes	Mixed	Strong+Passive	Rule-based query router for MySQL, SQL Server, and Oracle.
DBAAS	**Amazon Aurora** [3]	2014	No	No	MVCC	Strong+Passive	Custom log-structured MySQL engine for RDS.
	ClearDB [5]	2010	No	No	MVCC+2PL	Strong+Active	Centralized router that mirrors a single-node MySQL instance in multiple data centers.

Source: Pavlo, Andrew, and Matthew Aslett. "What's Really New with NewSQL?." ACM SIGMOD Record 45.2 (2016): 45-55.

	amazon web services	cloudera Ask Bigger Questions	hp
Analytical DBMS	Amazon Redshift service (based on ParAccel engine); Amazon Relational Database Service	HBase, and although not a DBMS, Cloudera Impala supports SQL querying on top of Hadoop	HP Vertica Analytics Platform Version 7
In-memory DBMS	None. Third-party options on AWS include Altibase, SAP Hana, and ScaleOut.	Although not a DBMS, Apache Spark supports in-memory analysis on top of Hadoop	Vertica is not an in-memory database, but with high RAM-to-disk ratios the company says it can ensure near-real-time query performance
Hadoop distributions	Amazon Elastic MapReduce. Third-party options include Cloudera and MapR.	CDH open-source distribution, Cloudera Standard, Cloudera Enterprise	None
Stream-processing technology	Amazon Kinesis	Open-source stream-processing options on Hadoop include Storm	None
Hardware/software systems	Not applicable	Partner appliances, preconfigured hardware, or both available from Cisco, Dell, HP, IBM, NetApp, and Oracle	HP ConvergedSystem 300 for Vertica, plus a choice of reference architectures for Cloudera, Hortonworks, and MapR Hadoop distributions

Source: http://www.informationweek.com/big-data/big-data-analytics/16-top-big-data-analytics-platforms/d/d-id/1113609

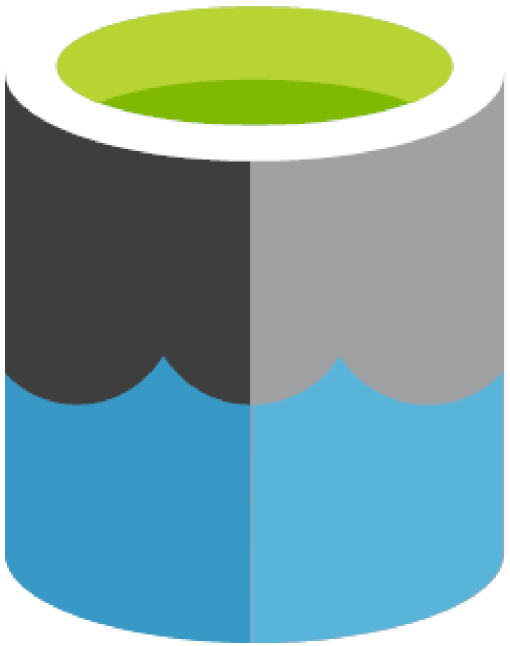
	ORACLE	TERADATA
Analytical DBMS	Oracle Database, Oracle MySQL, Oracle Essbase	Teradata, Teradata Aster
In-memory DBMS	Oracle TimesTen, Oracle Database 12c In-Memory Option	Not an in-memory DBMS, Teradata Intelligent Memory monitors queries and automatically moves the most-requested data to the fastest storage tiers available, options including RAM, flash, SSD, spinning discs of various speed
Hadoop distributions	Resells and supports Cloudera Enterprise	None
Stream-processing technology	Oracle Event Processing	Resells and supports the Hortonworks Data Platform
Hardware/software systems	Exadata, Exalytics, Oracle Big Data Appliance	Teradata and Teradata Aster are integrated software/hardware systems. Hadoop is supported with two Teradata appliance offerings as well as standardized Dell configurations

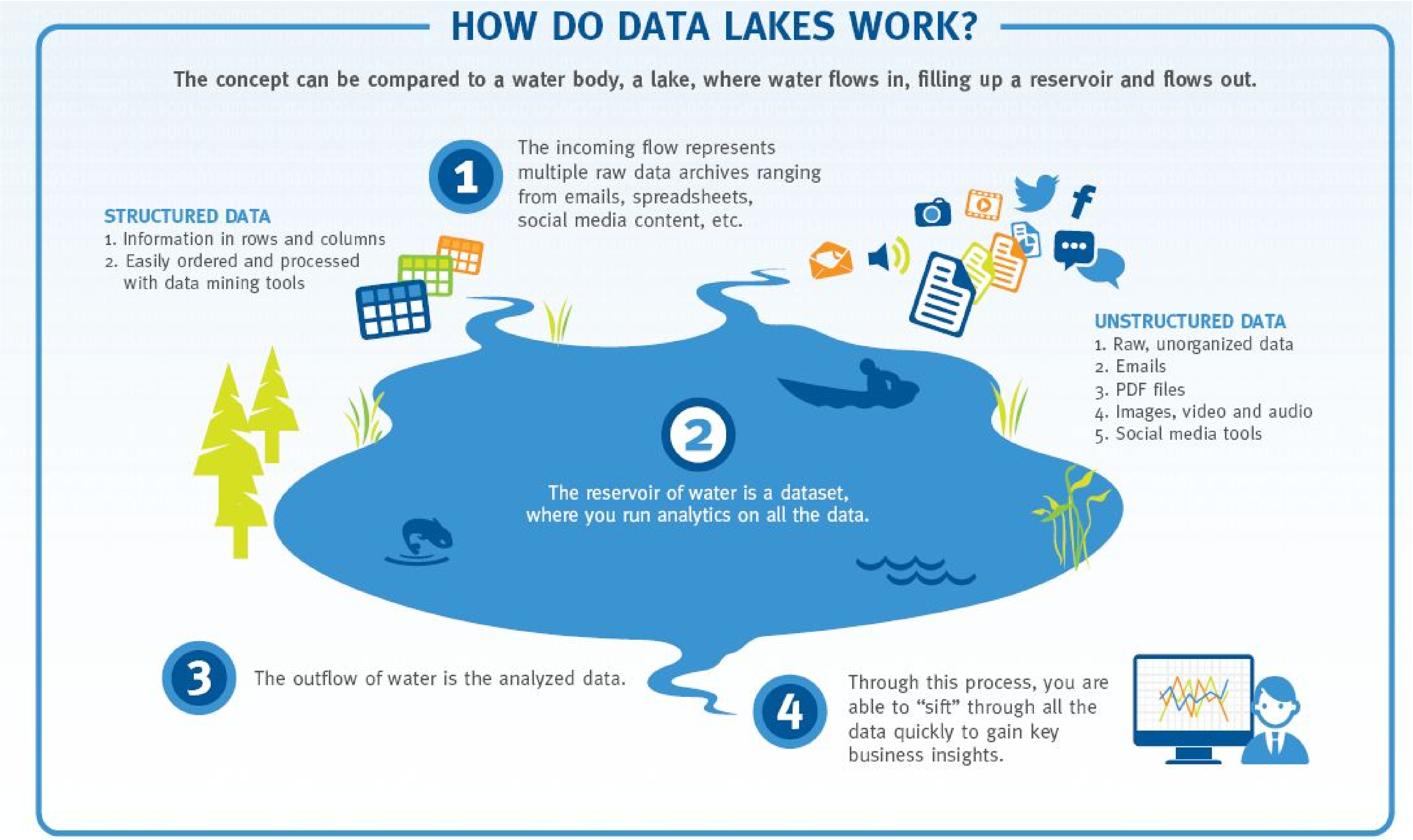
Source: http://www.informationweek.com/big-data/big-data-analytics/16-top-big-data-analytics-platforms/d/d-id/1113609

Upcoming Trends

A central repository for data – Data Lakes

- Primary repository of raw data
- The actual value of the data is unknown when it first arrives
- An infrastructure with low cost storage is needed
- It need not provide all the classic ACID properties of a database

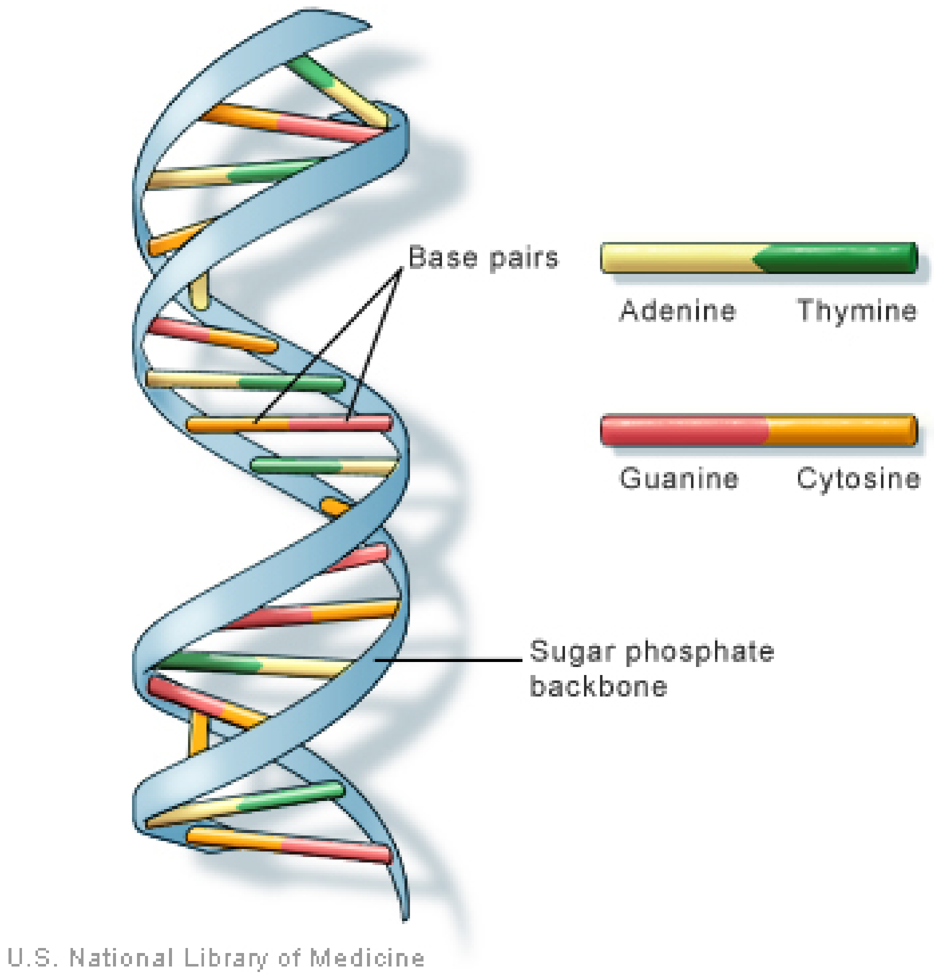




Source: https://goo.gl/images/45tctF

What are the main components of a DNA Strand

- A molecule that carries the genetic instructions used in the growth, development, functioning and reproduction.
- Two DNA strands (polynucleotides)
  - Composed of nucleotides
  - Each nucleotide is composed of one of four nitrogen-containing nucleobases — cytosine (C), guanine (G), adenine (A), or thymine (T)

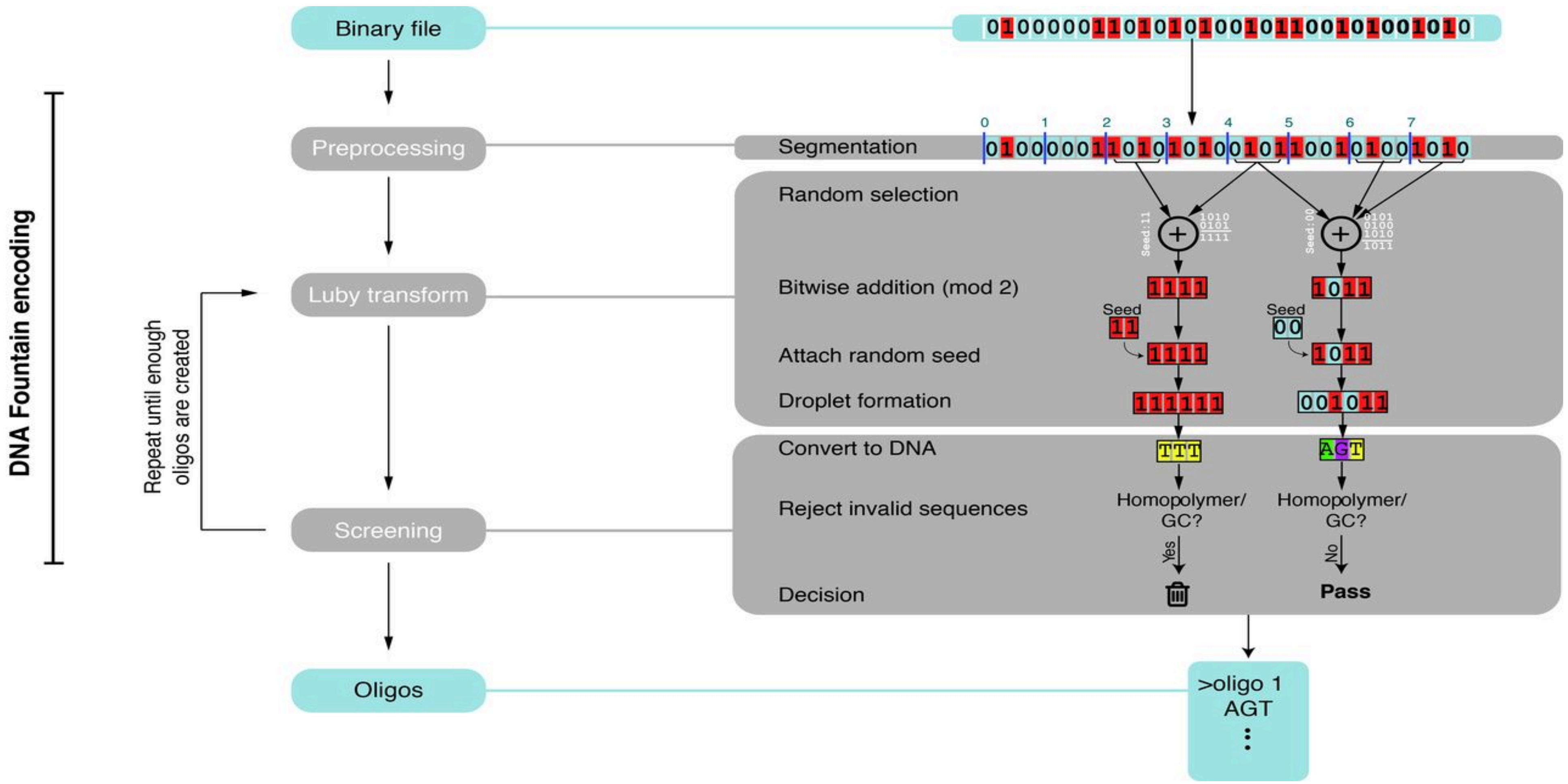


DNA as a storage solution

- 2.14×10^6 bytes of data were successfully stored and retrieved
- 2.18×10^{15} successful retrievals using the original DNA sample
- Limit of storage density – 215 petabytes per gram
- Shannon information capacity
  - capacity of each nucleotide can't 2 bits
  - not all DNA sequences are created equal – homopolymer runs (e.g., AAAAAA…) are undesirable
  - overall Shannon information capacity of a DNA storage device is ~1.83 bits per nucleotide
- Robustness against data corruption
- Overcome both oligo dropouts and the biochemical constraints of DNA storage
- {00,01,10,11} to {A,C,G,T}

Source: http://science.sciencemag.org/content/355/6328/950.full

How can we encode Binary Bits in a DNA Strand



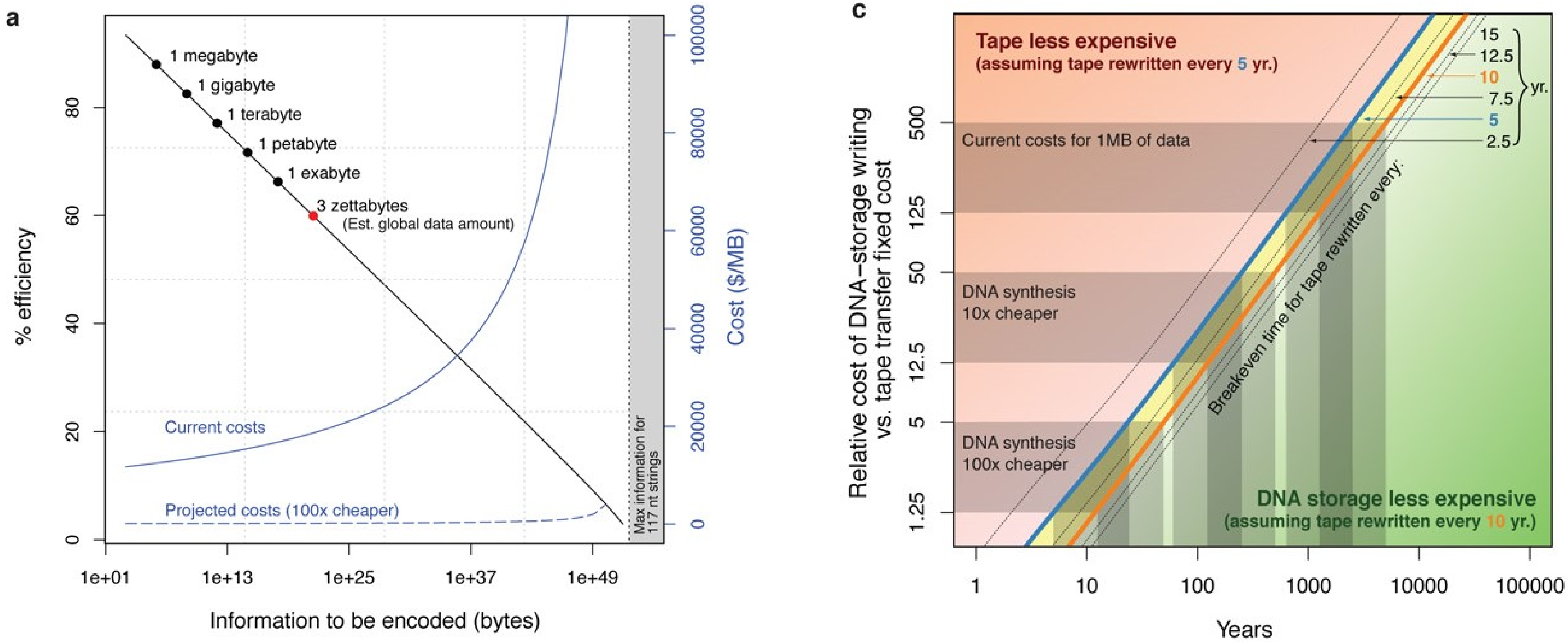
Source: http://science.sciencemag.org/content/355/6328/950.full

Study results of the DNA Fountain Encoding

- Encoding of a 2,146,816 byte in 2.5 minutes
- Density achieved – 1.57 bits/nt – 14% from the Shannon capacity of DNA storage and 60% more than previous studies with a similar scale of data
- Huge storage costs – write 3500$/Mbyte (read 1000$/Mbyte)
  - continuous improvements to the DNA synthesis chemistry
  - exploring quick-and-dirty oligo synthesis methods that consume less machine time and fewer reagents and, therefore, are more cost-effective
- Recovering with 0 errors
- Decoding took ~9 min with a Python script on a single CPU

Source: http://science.sciencemag.org/content/355/6328/950.full

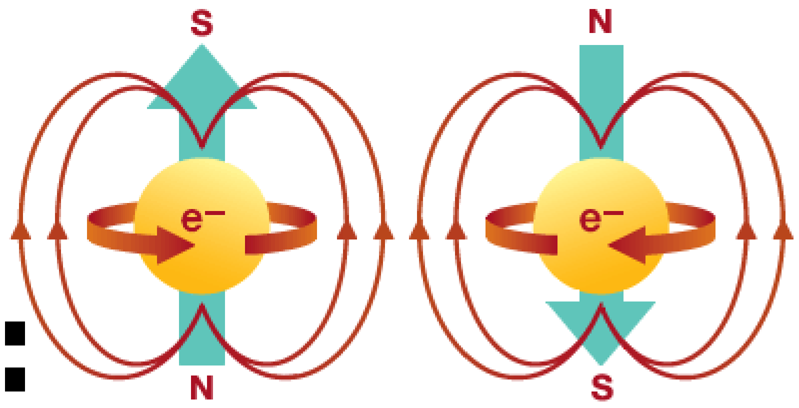
Progress in DNA sequencing makes DNA storage a viable option



Source: Goldman, N.; Bertone, P.; Chen, S.; Dessimoz, C.; Leproust, E. M.; Sipos, B.; Birney, E. (2013). "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA". Nature. 494 (7435): 77–80

A new way of representing a Bit – Spintronics

- Spin-based electronic
- Manipulating spin is faster and requires far less energy
- MRAM -- magnetoresistive random access memory
- Traditional representing of a bit – as charge in a capacitor or as the state of an interconnected set of transistors
- New approach: store data using spin of electrons in ferromagnetic substance
- Spin up means 0, and spin down means 1
- Compact, speedy, low-power, and nonvolatile
- (cache, RAM, HDD) -> working memory
- However
    - The density of bits is low, and the cost of chips is high



Source: http://spectrum.ieee.org/semiconductors/memory/spintronic-memories-to-revolutionize-data-storage

How a Spin Memory works

- Gallium manganese nitride
  - semiconductor whose magnetic properties we can manipulate electrically
- *n*egative-type semiconductor
- *p*ositive-type semiconductor