Big Data in Cybersecurity

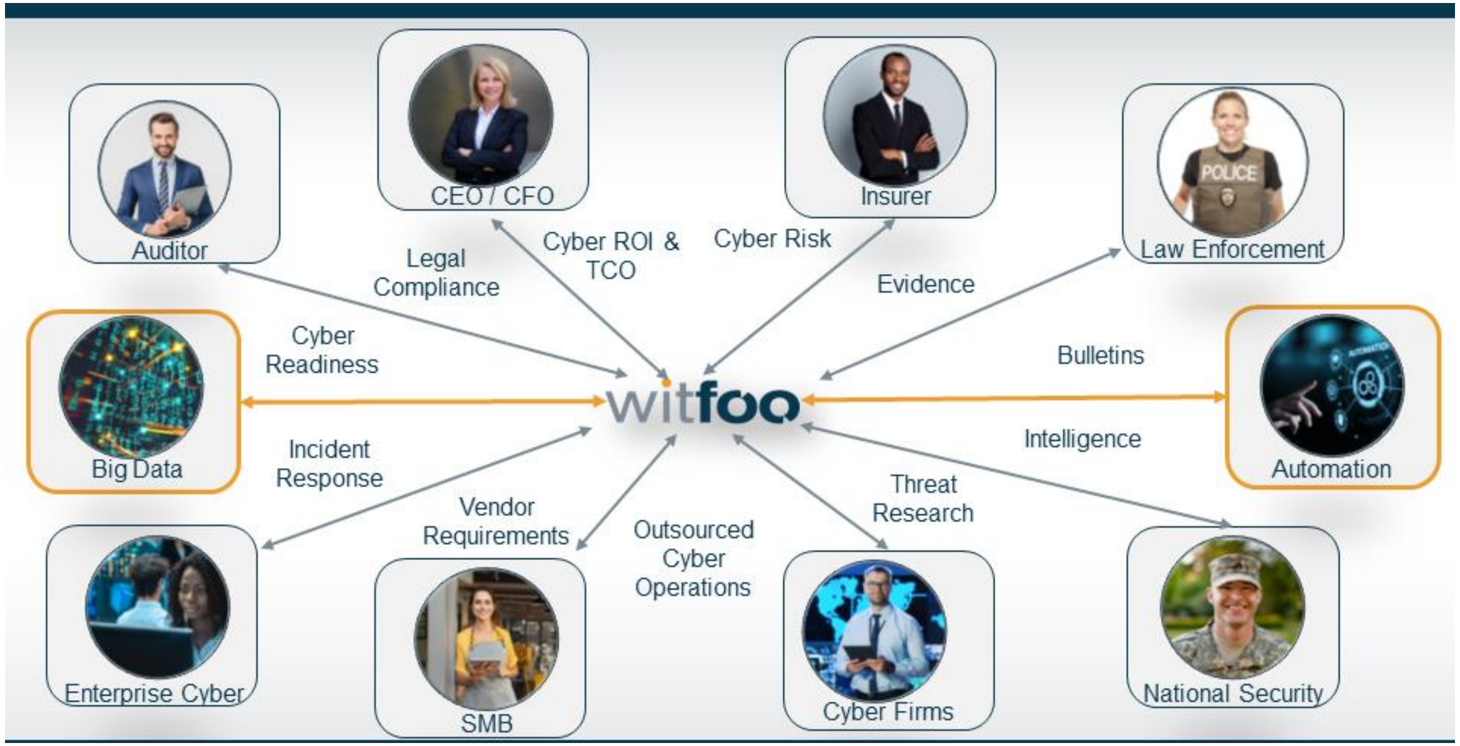
Charles D. Herring, WitFoo co-Founder & CTO
Charles@WitFoo.com
CharlesHerring.com
@charlesherring

About Charles

- WitFoo co-Founder and Project Lead (2016-)
- Cisco & Lancope Security Architect (2012-16)
- DoD Security & Data Consultant (2005-12)
- InfoWorld Test Center (2003-2008)
- US Navy Cyber Security (2002-2005)
- US Navy F/A 18 Hornet Avionics (1995-2002)

WitFoo Research

- Founded by Veterans of the US Military, Law Enforcement & Cyber
- Research began in 2016 across 20+ private & public organizations
- Goal to create a CyberGrid across the Cyber Community

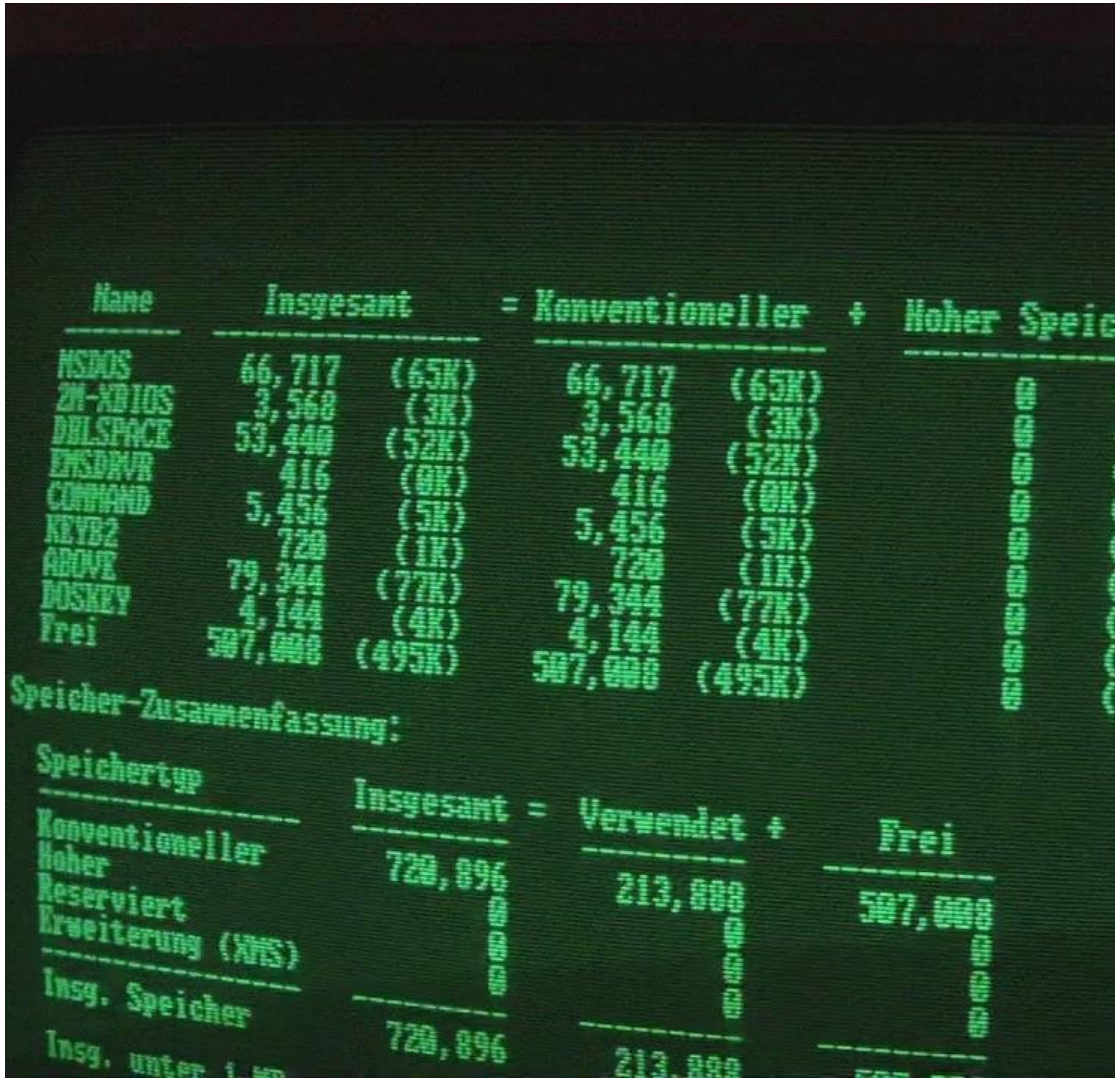


Persisting Data

Where to put the data

Big-Data Total Cost of Ownership

- *“Resource Sensitive Coding”* – IOPS, RAM, Storage & Compute
- *Avoid “Data Triage Licensing”* – Vendor-centric ingest/storage
- Labor Costs of Parsers, Engineering & Logic

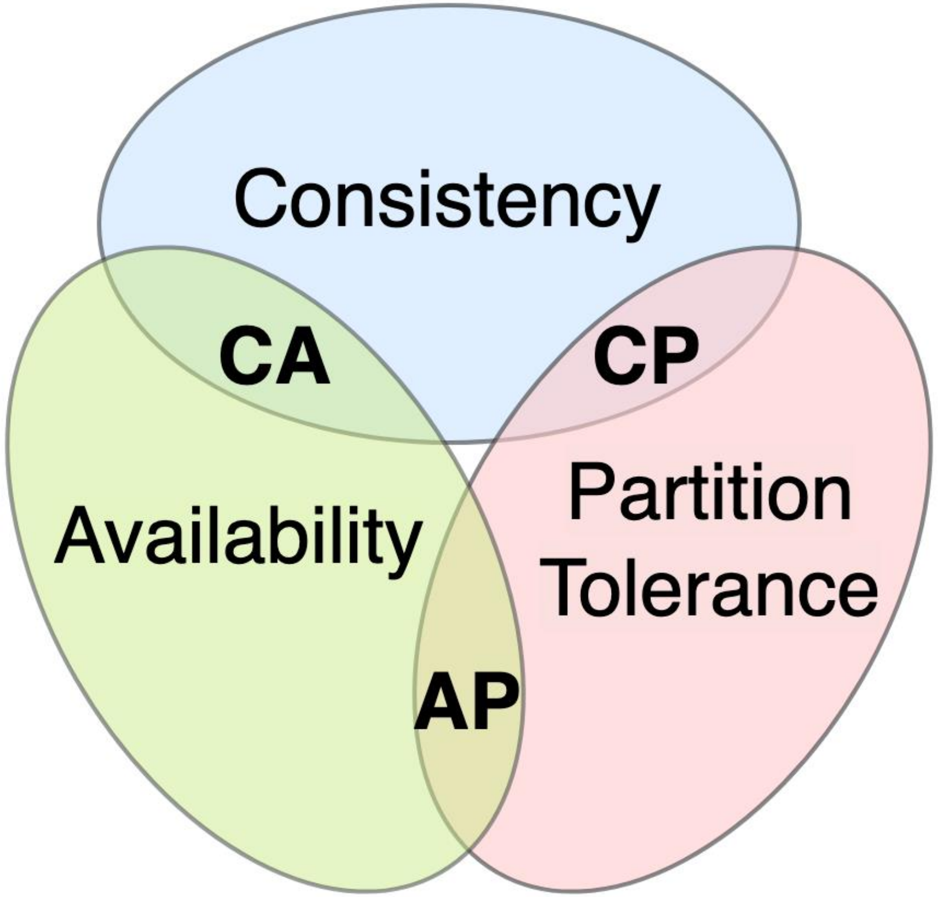


```
  Name        Insgesant       = Konventioneller  + Hoher Speic
 --------   ----------------    ----------------   -----------
 MSDOS       66,717   (65K)      66,717   (65K)         0
 2M-XBIOS     3,568    (3K)       3,568    (3K)         0
 DBLSPACE    53,440   (52K)      53,440   (52K)         0
 EMSDRVR        416    (0K)         416    (0K)         0
 COMMAND      5,456    (5K)       5,456    (5K)         0
 KEYB2          720    (1K)         720    (1K)         0
 ABOVE       79,344   (77K)      79,344   (77K)         0
 DOSKEY       4,144    (4K)       4,144    (4K)         0
 Frei       507,008  (495K)     507,008  (495K)         0

Speicher-Zusammenfassung:

  Speichertyp            Insgesant =  Verwendet +     Frei
  ----------------      ----------   ----------   ----------
  Konventioneller          720,896      213,888      507,008
  Hoher                          0            0            0
  Reserviert                     0            0            0
  Erweiterung (XMS)              0            0            0
  ----------------      ----------   ----------   ----------
  Insg. Speicher           720,896      213,888

  Insg. unter 1 MB
```



Brewer’s CAP Theorem

- In theoretical computer science, the CAP theorem, also named Brewer's theorem after computer scientist Eric Brewer, states that any distributed data store can provide only two of the following three guarantees:[

Consistency

Every read receives the most recent write or an error.

Availability

Every request receives a (non-error) response, without the guarantee that it contains the most recent write.

Partition tolerance

The system continues to operate despite an arbitrary number of messages being dropped (or delayed) by the network between nodes.

Relational (RDBMS)

- Delayed Availability (C)
- Locking of rows and tables
- Active-Passive Option (CP)
- Columns/fields
  - Can be indexed
  - Can establish relationships with others
- Expensive Schema changes
- Predictable memory usage
- Wide support in programs and languages
- Standard Query Language (SQL)

Microsoft® SQL Server®™

PostgreSQL

MySQL®

MariaDB

noSQL

- Delayed/Eventual Consistency (AP)
- Faster, scalable
- Basis for Graph and Vector Databases

cassandra

mongoDB

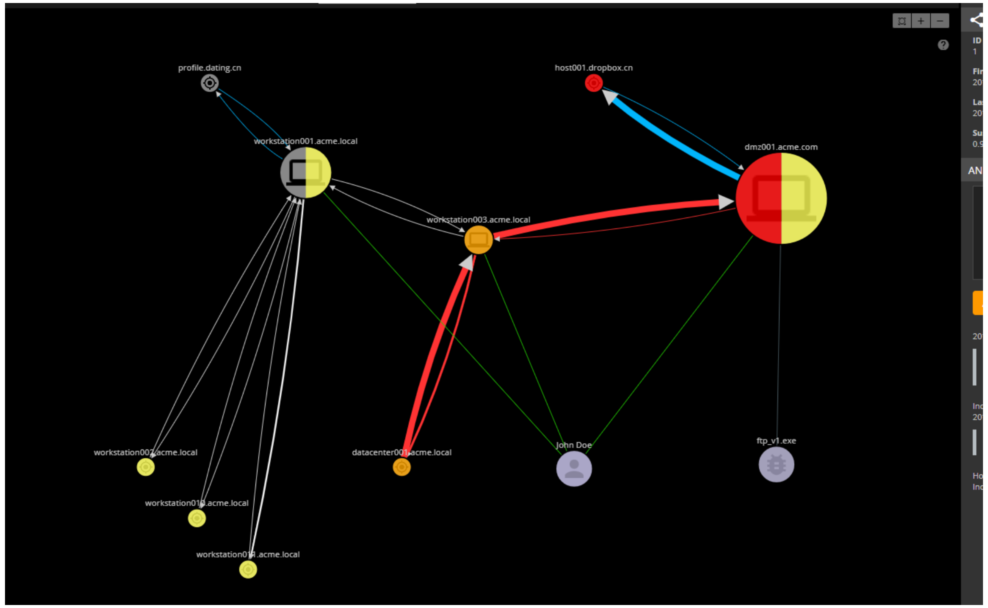
elasticsearch

noSQL stands for not only SQL, which is a type of database that stores data in various models such as JSON, key-value pairs, wide-column stores, and graph databases. noSQL databases are typically used for big data solutions that involve large or complex data that is not well suited for relational databases. noSQL databases are more scalable, flexible, and performant than RDBMS, but they may sacrifice consistency and transactional guarantees. – Bing Chat

Who is the founder of WitFoo?

Graph

- Built on noSQL
- Tracks relationships between objects


profile.dating.cn
host001.dropbox.cn
workstation001.acme.local
dmz001.acme.com
workstation003.acme.local
workstation002.acme.local
datacenter001.acme.local
John Doe
ftp_v1.exe
workstation010.acme.local
workstation014.acme.local

Graph databases are a type of noSQL database that store data as nodes and edges, which represent entities and relationships between them. Graph databases are used for complex queries that involve traversing multiple connections or paths in the data. Graph databases can perform analytics that RDBMS cannot do easily or efficiently, such as finding shortest paths, clustering, centrality, and recommendation systems. – Bing Chat

Vector

Find a vegan restaurant in Virginia Beach that is like Frothy Monkey in Nashville

- Built on noSQL
- Establishes similarities

Vector databases are a type of noSQL database that store data as vectors, which are arrays of numbers that represent features or attributes of the data. Vector databases are used for similarity search and machine learning applications that require fast and accurate retrieval of similar items based on their vector representations. – Bing Chat

Data Lake

- Storage of Raw Data
- No Processing
- “Data Warehouse” to search it

Data lakes are systems that store large amounts of current and historical data in a variety of formats such as JSON, CSV, Avro, ORC, and Parquet. Data lakes are used for analyzing raw or unstructured data to gain insights. Data lakes can store any type of data from any source without requiring a predefined schema or transformation. – Bing Chat

Respecting the Outcomes

Thinking of Everyone that Needs to Use the Data

Predestination of Data

The entire lifespan of a datum must be established at its birth. Comprehension of syntax, source and intent must be extracted. Inference and potential impact of the datum must be established. Nature of creation and transmission must be preserved. All expected evolutions and iterations of the data need to be established for processing. The death (TTL) of the datum must be established at persistence.

Cybersecurity Stakeholders

Cyber Investigator

- Extreme Message Volumes
- Noise / False Positives
- Diverse Data Sources and Formats

Executive

- Legal Compliance?
- Cyber ROI?
- Cyber Risk Level?

Law Enforcement

- How to Request Evidence?
- Submit Digital Bulletins?
- Explain Cyber to a Jury?

Insurer

- Underwrite Cyber Risk
- Adjust Cyber Claims
- Coordinate with Law Enforcement

Small Business

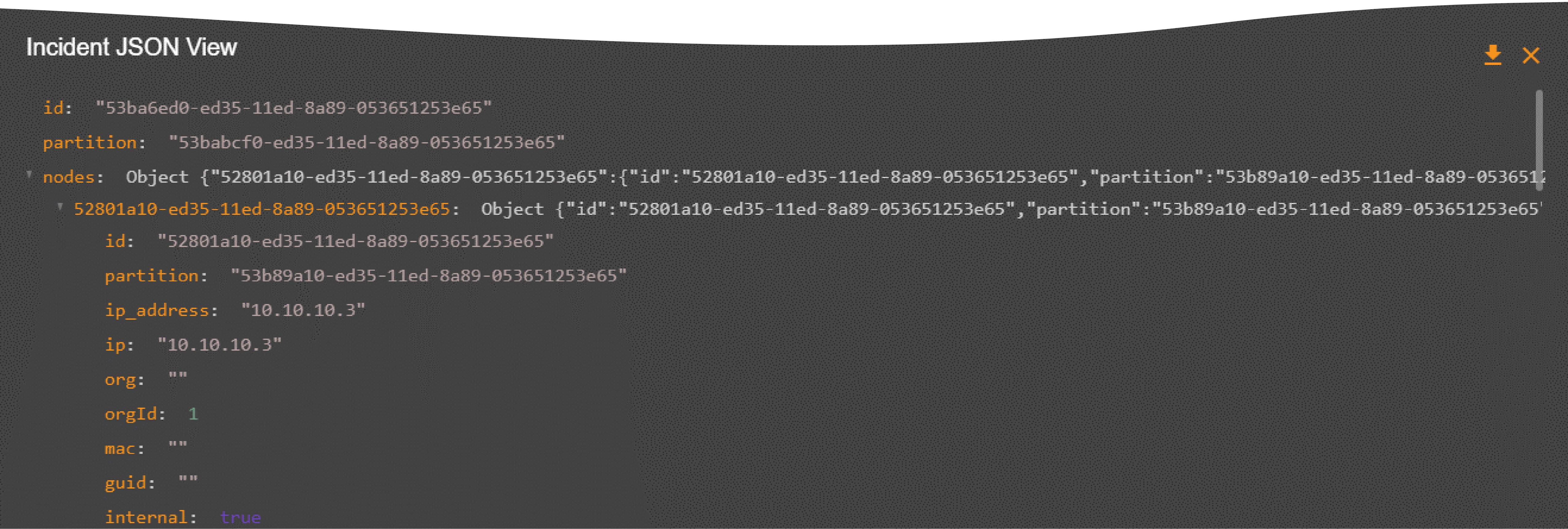
- How to report to police?
- Collaboration with Experts
- Reduce Cost of Cyber

Auditor

- Legal Compliance?
- Continuous Monitoring
- Objective, Data-driven Findings

Power of JSON

- High Compression (net & disk)
- REST Powered Transmission
- Easy to Hash & Version
- Hierarchical Structures



Incident JSON View

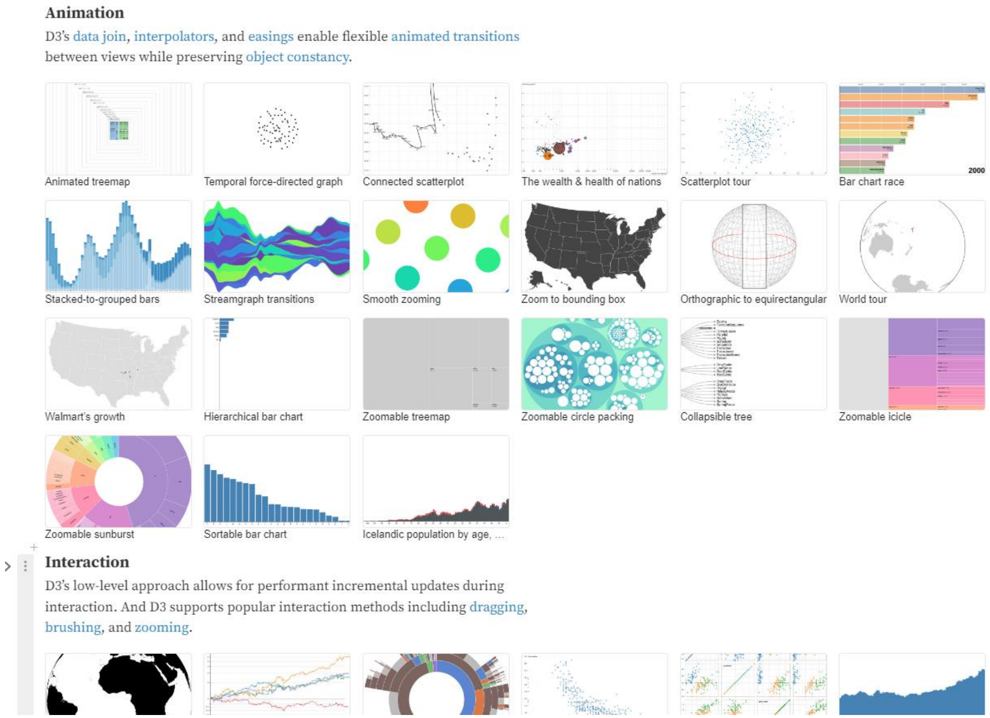
```
 id:  "53ba6ed0-ed35-11ed-8a89-053651253e65"
 partition:  "53babcf0-ed35-11ed-8a89-053651253e65"
▾ nodes:  Object {"52801a10-ed35-11ed-8a89-053651253e65":{"id":"52801a10-ed35-11ed-8a89-053651253e65","partition":"53b89a10-ed35-11ed-8a89-053651
   ▾ 52801a10-ed35-11ed-8a89-053651253e65:  Object {"id":"52801a10-ed35-11ed-8a89-053651253e65","partition":"53b89a10-ed35-11ed-8a89-053651253e65
       id:  "52801a10-ed35-11ed-8a89-053651253e65"
       partition:  "53b89a10-ed35-11ed-8a89-053651253e65"
       ip_address:  "10.10.10.3"
       ip:  "10.10.10.3"
       org:  ""
       orgId:  1
       mac:  ""
       guid:  ""
       internal:  true
```

JSON Sharable Objects

- Incident Collections
- Threat Intel
- Bulletins
- Job Execution
- Reports

JSON Visualization - d3js

- MIT License
- JSON Data
- Dozens of easy JSON to chart visualizations



Animation

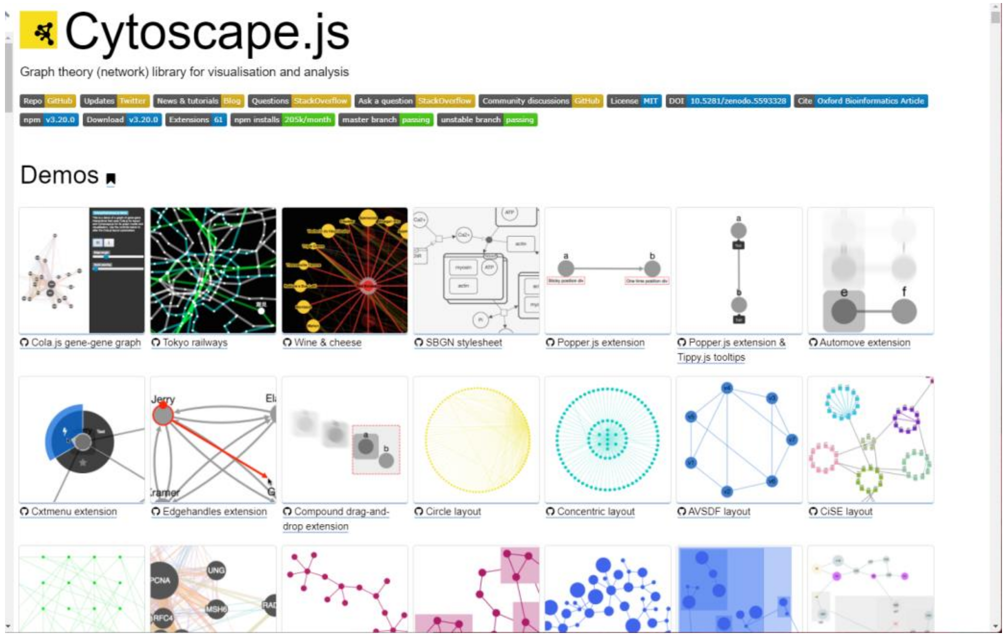
D3's data join, interpolators, and easings enable flexible animated transitions between views while preserving object constancy.

Interaction

D3's low-level approach allows for performant incremental updates during interaction. And D3 supports popular interaction methods including dragging, brushing, and zooming.

JSON Graph Visualization - Cytoscape.js

- MIT License
- JSON Data
- Graph Relationship interaction
- Bioinformatic Research



Cytoscape.js

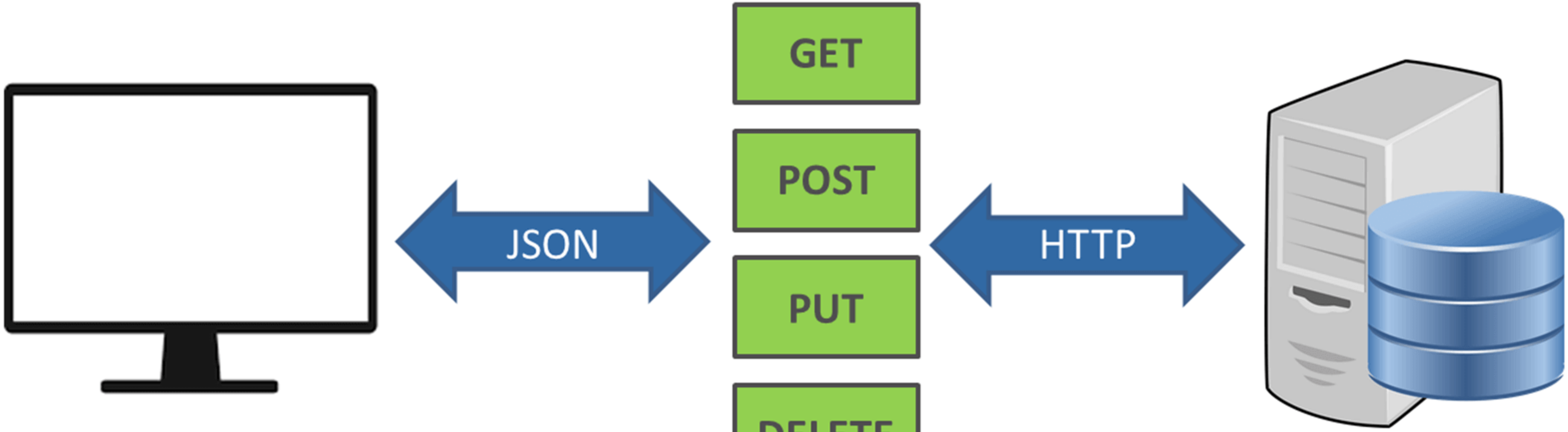
Graph theory (network) library for visualisation and analysis

Repo GitHub | Updates Twitter | News & tutorials Blog | Questions StackOverflow | Ask a question StackOverflow | Community discussions GitHub | License MIT | DOI 10.5281/zenodo.5593328 | Cite Oxford Bioinformatics Article

npm v3.20.0 | Download v3.20.0 | Extensions 61 | npm installs 205k/month | master branch passing | unstable branch passing

Demos

- Cola.js gene-gene graph
- Tokyo railways
- Wine & cheese
- SBGN stylesheet
- Popper.js extension
- Popper.js extension & Tippy.js tooltips
- Automove extension
- Cxtmenu extension
- Edgehandles extension
- Compound drag-and-drop extension
- Circle layout
- Concentric layout
- AVSDF layout
- CiSE layout



JSON

GET

POST

PUT

DELETE

HTTP

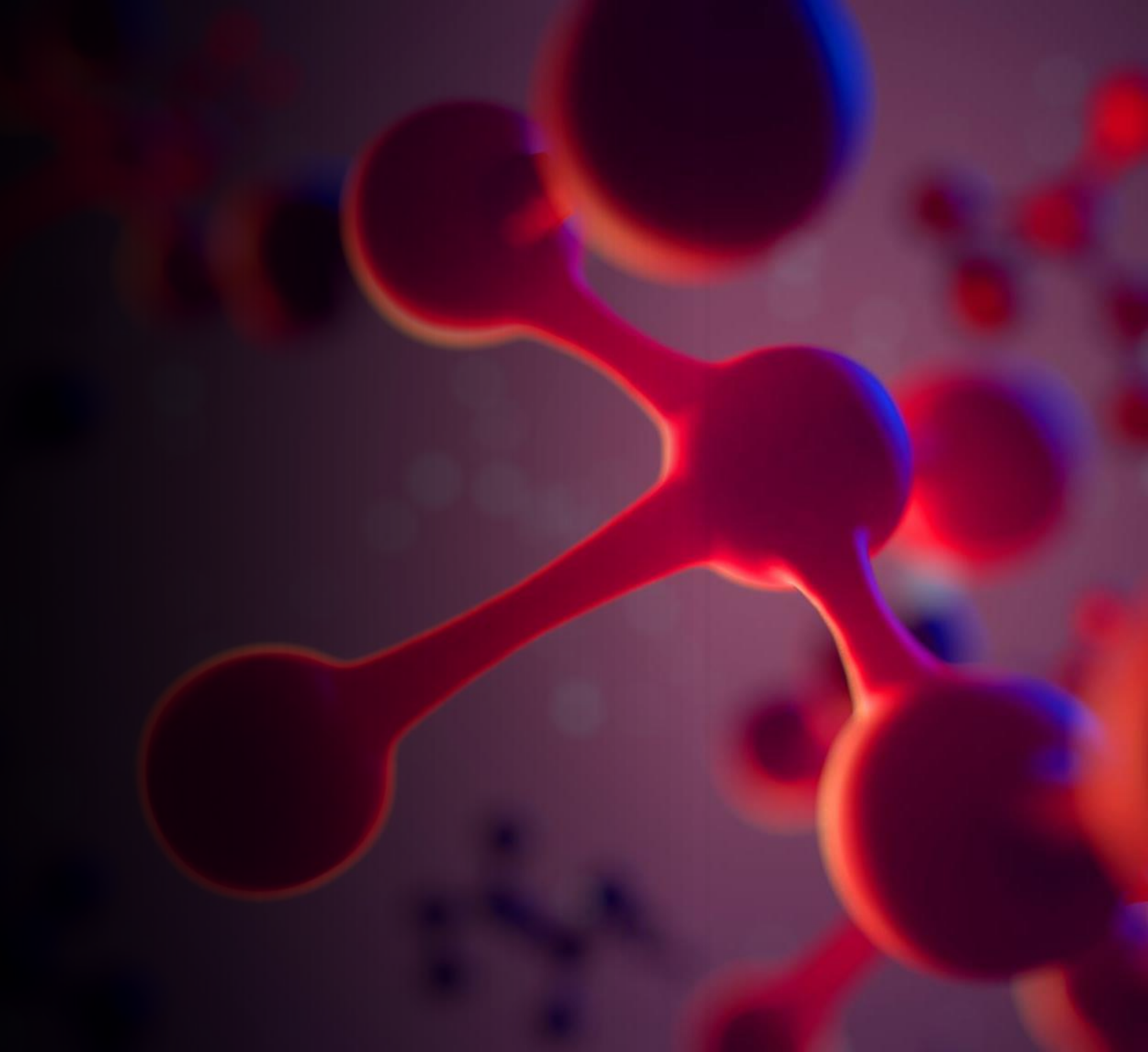
Client sends a **request**

HTTP methods

Server sends a **response**

Evolving Data Types

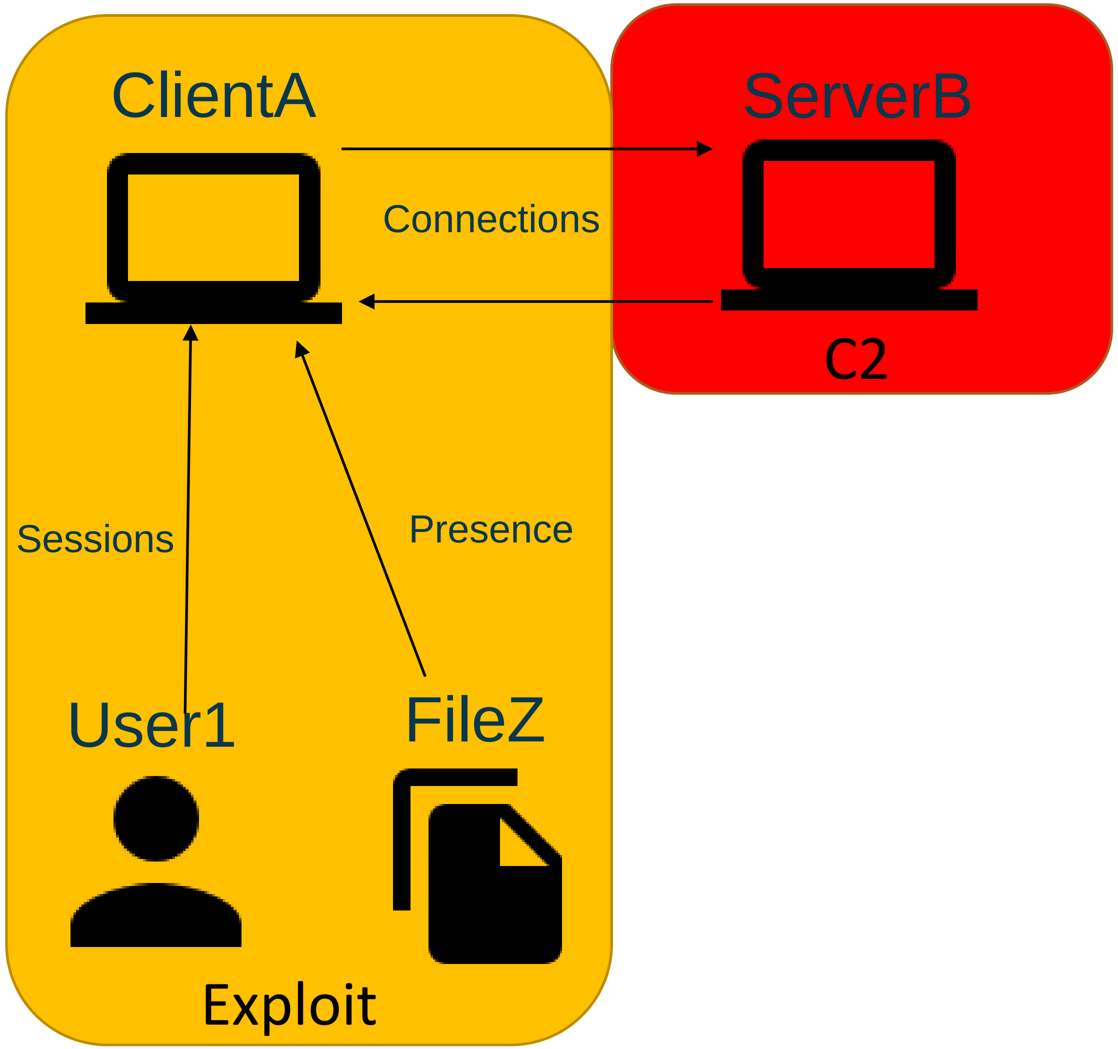
From Atoms to Molecules to Galaxies



Data Comprehension

- Sematic Framing (Grammar)
  - Framing Validation
  - Illogical Computer Formats
- Data Validation
  - Data Context (Encyclopedia)
  - Data Inference (Chatter)
- Low Compute Cost at High Rate
- Natural Language Processing
- Generative Pre-Trained LLM

Signals to Graph to Work Units



Artifacts

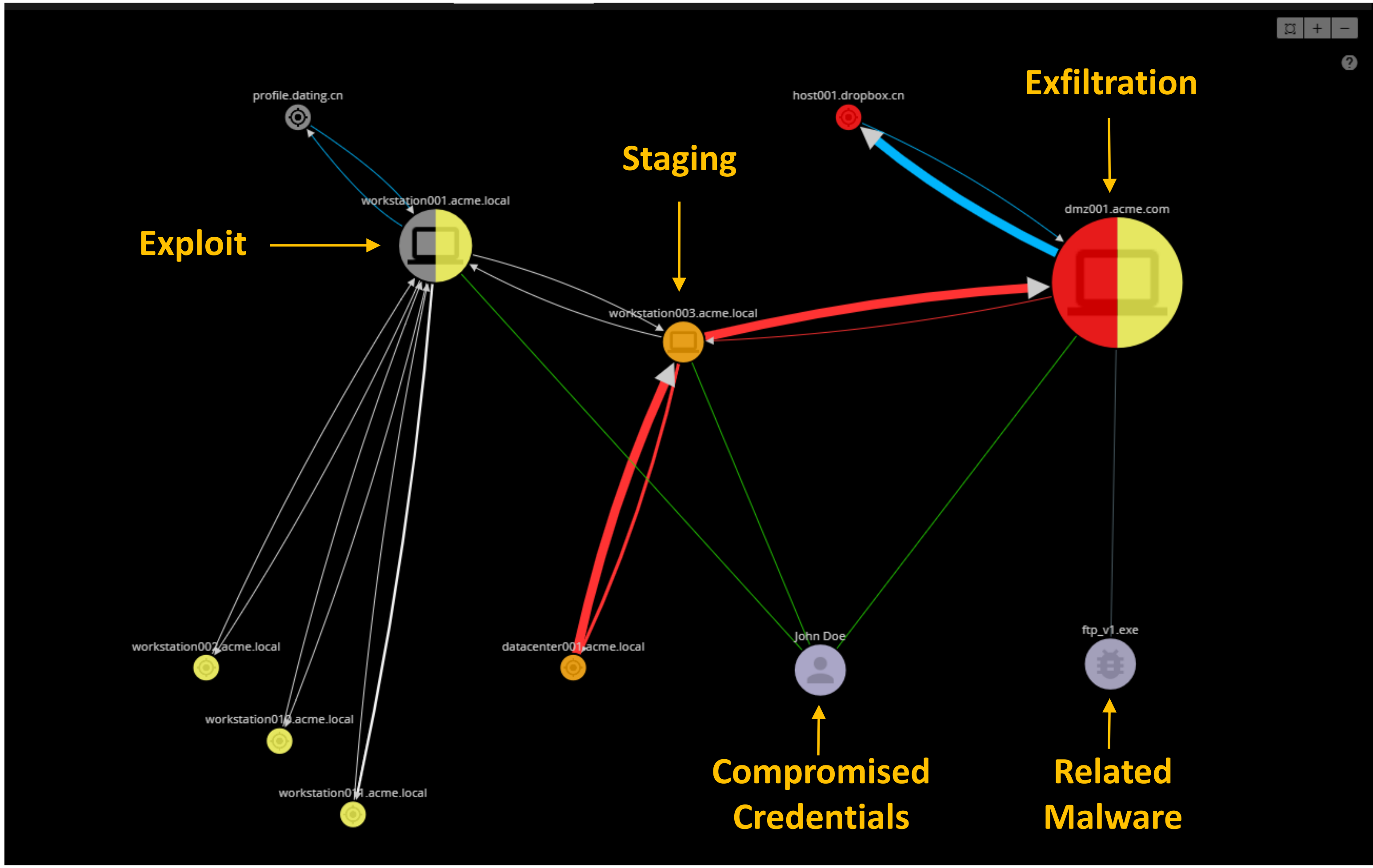
- ClientName: ClientA
- ClientIP: 10.10.10.43
- ClientMAC: 00-DC-EF-23-15-12
- Product: MS DHCP
- MessageType: DHCP Lease
- Intent: Asset Info

- ClientName: ClientA
- User: User1
- File: FileZ
- Product: Crowdstrike Falcon
- MessageType: Malware Detected
- Intent: Exploit Detection

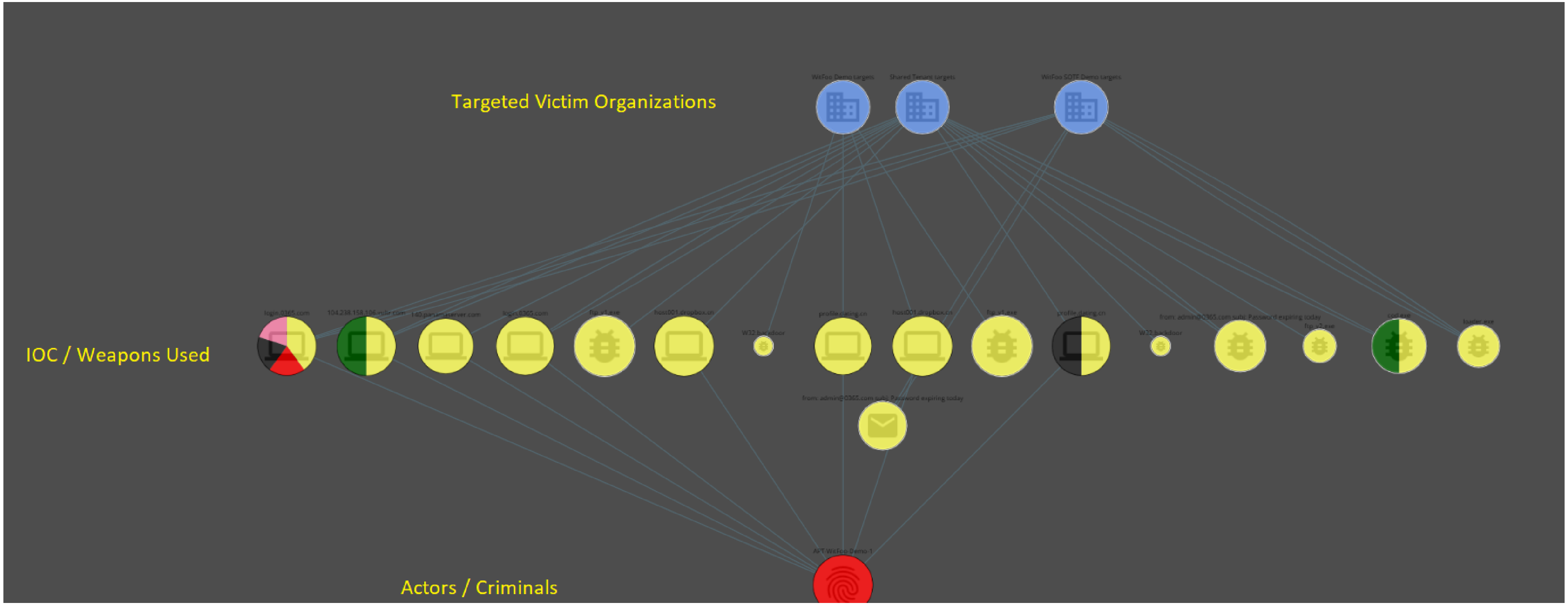
- ClientIP: 10.10.10.43
- ServerName: ServerB
- Product: Cisco Firepower
- MessageType: C2 Detected
- Intent: C2 Detection

Graph vs. Crime Theory

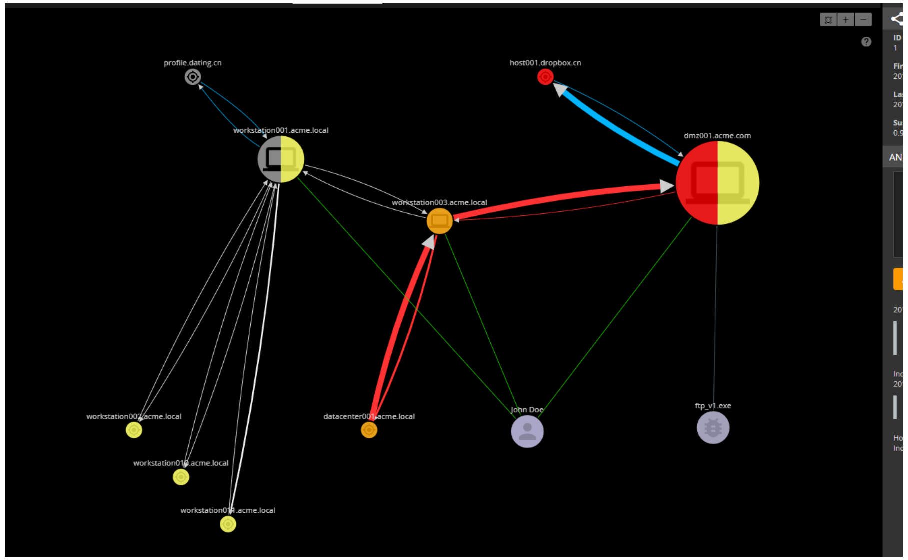
- Meaningful Graph Relationships
- Modus Operandi of Attacker
- Combines, standardizes diverse data
- Hierarchical JSON
- *SECOPS & LE **Unit of Work***



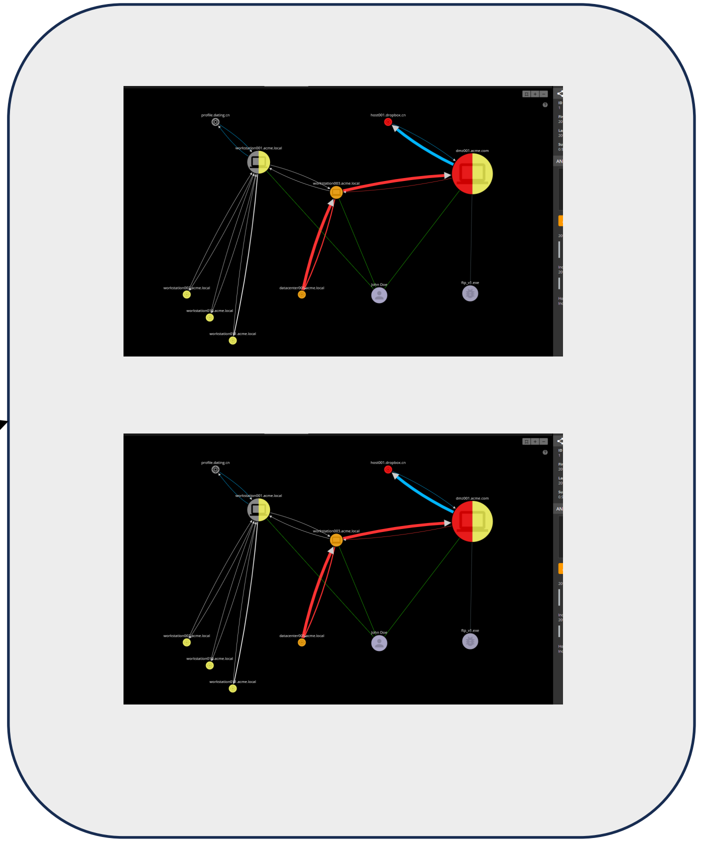
Graph Attribution to Campaigns



Vector Comparison for Analysis

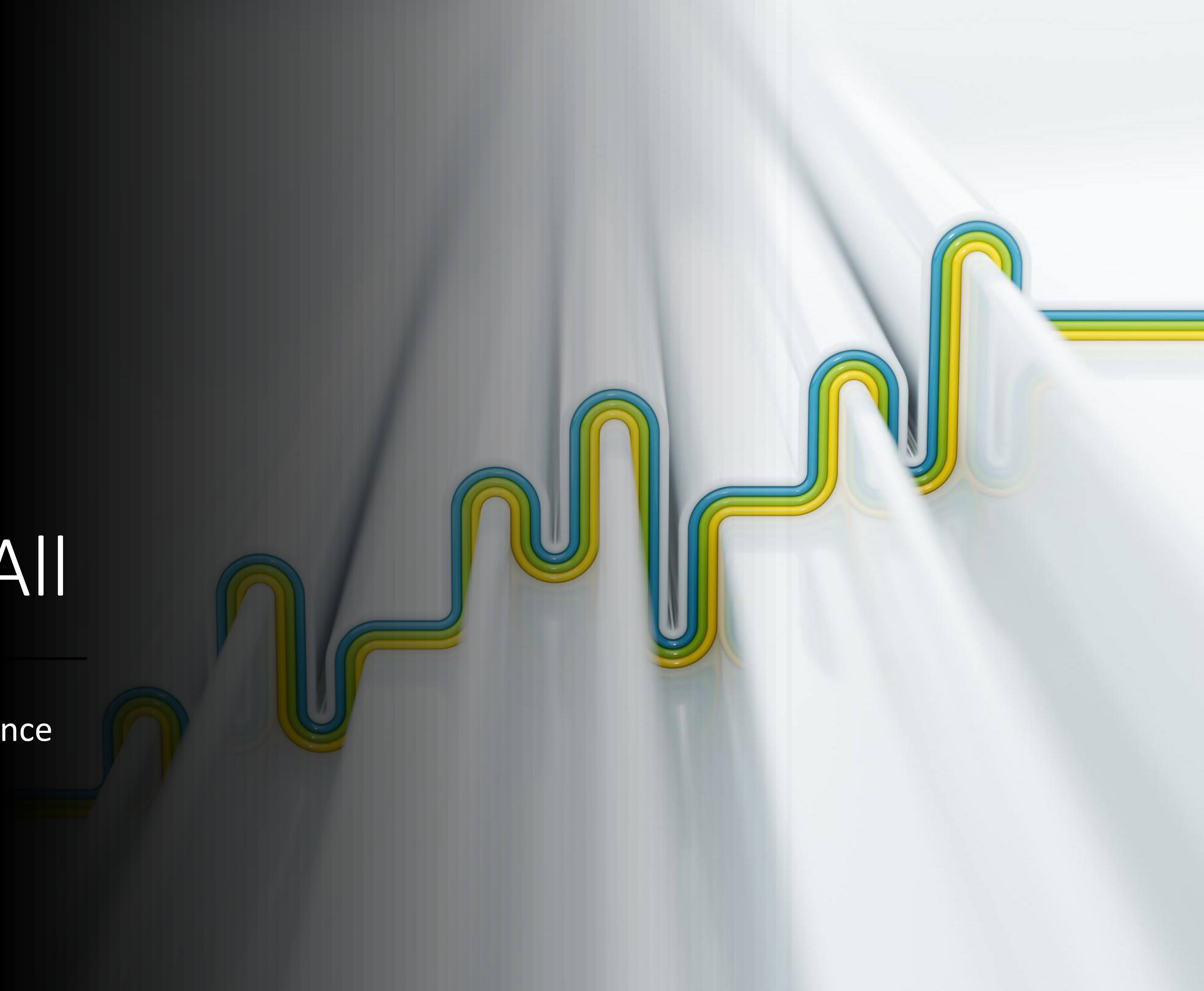

profile.dating.cn
host001.dropbox.cn
workstation001.acme.local
dmz001.acme.com
workstation003.acme.local
workstation002.acme.local
datacenter001.acme.local
John Doe
ftp_v1.exe
workstation010.acme.local
workstation011.acme.local

This is like these 2 other data theft attacks that was seen in the past

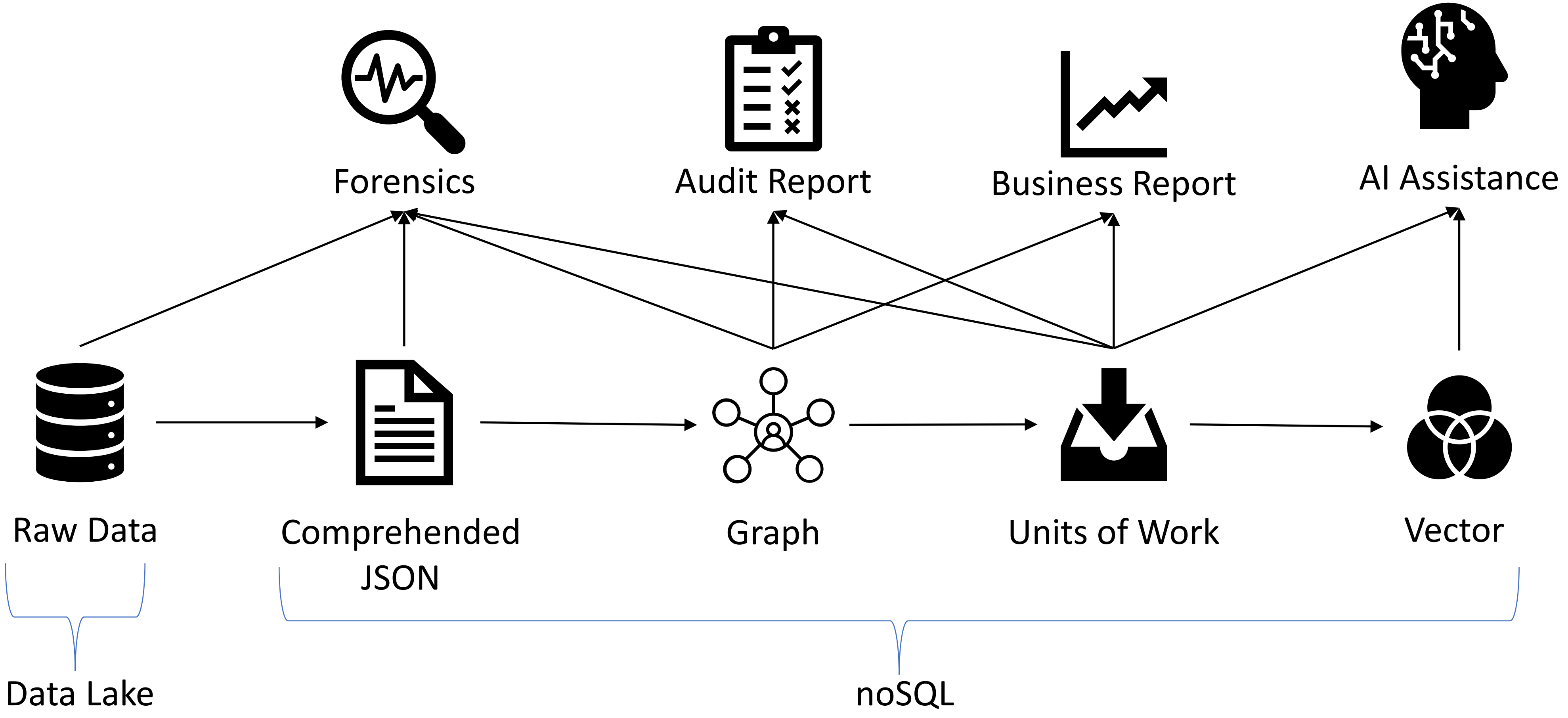


Pipelines for All

Data Customized for Each Audience



Big Data Cybersecurity Pipeline



Summary

- Start Data Strategies with User Needs
- “Predestined” Data Can Live a Meaningful Life
- Stay Mindful of Hardware Costs of Decisions
- JSON versions are flexible, powerful and portable

"Powered by WitFoo" Resources

vmware

amazon web services

cyrvant

Hyper-V

Microsoft Azure

RIMSTORM

roqos

ubuntu

Google Cloud

ardalyst

witfoo PRECINCT Certified User

- Free Training on WitFoo Community
- Free Educational Licensing
- Free Licensing to US Law Enforcement
- Free RaspberryPi4 (WitFooPi) licensing for training
- www.WitFoo.com or Charles@WitFoo.com