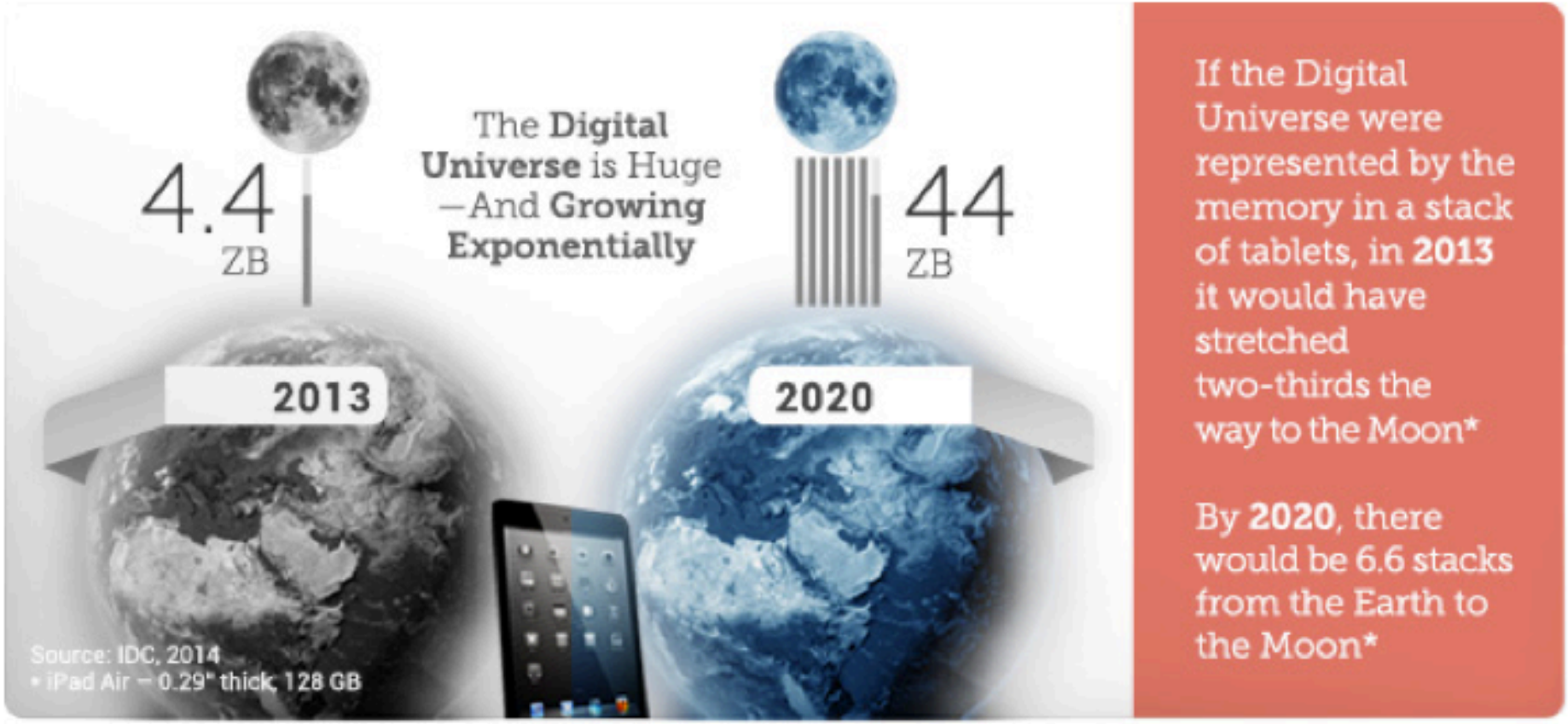
Module MCS 016
Machine Learning for Cyber-Security & Artififial Intelligence

Prof. Hermes Senger

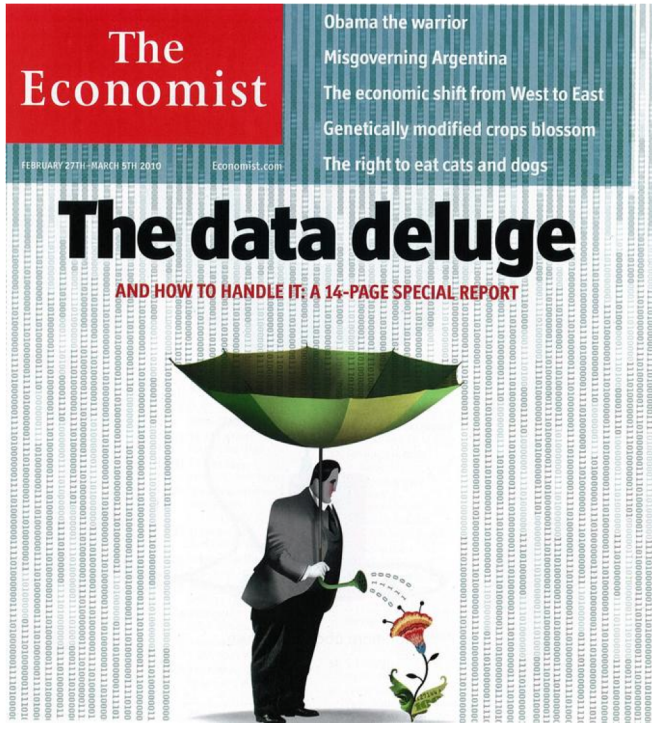
- Before starting our module, let’s see some facts and tendencies …

The data growth



A few (old) examples

- 2008: Google processes 20 PB a day
- 2009: Facebook has 2.5 PB user data + 1.5 TB a day
- 2009: eBay has 6.5 PB user data + 50 TB a day
- 2011: Yahoo! has 180-200 PB of data
- 2012: Facebook
  - +2.5 billion pieces of content and 500+ Terabytes of data each day
  - +2.7 billion Like actions
  - 300 million photos
  - scans ~ 105 Terabytes of data

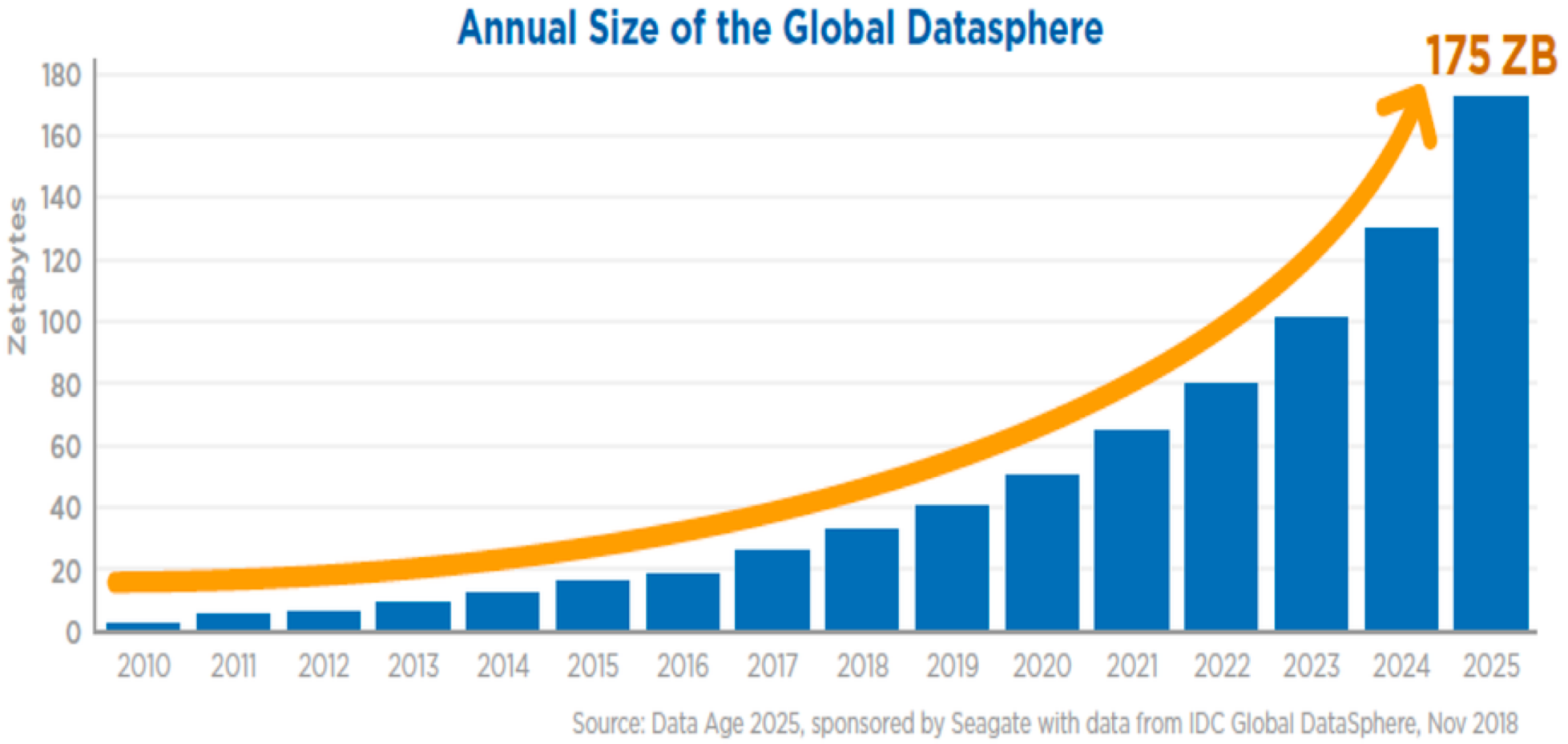


EMC - # of Storage Shipped

- Amount of data storage shipped
  - In 2005 the total amount of data storage shipped by the company over its history achieved **one Exabyte** (i.e., 10^{18} bytes).
  - In 2010, ~ one Exabyte of storage within a **single year**
  - In 2013 (June) ~ Exabyte in a **single month**



Data Growth in the World

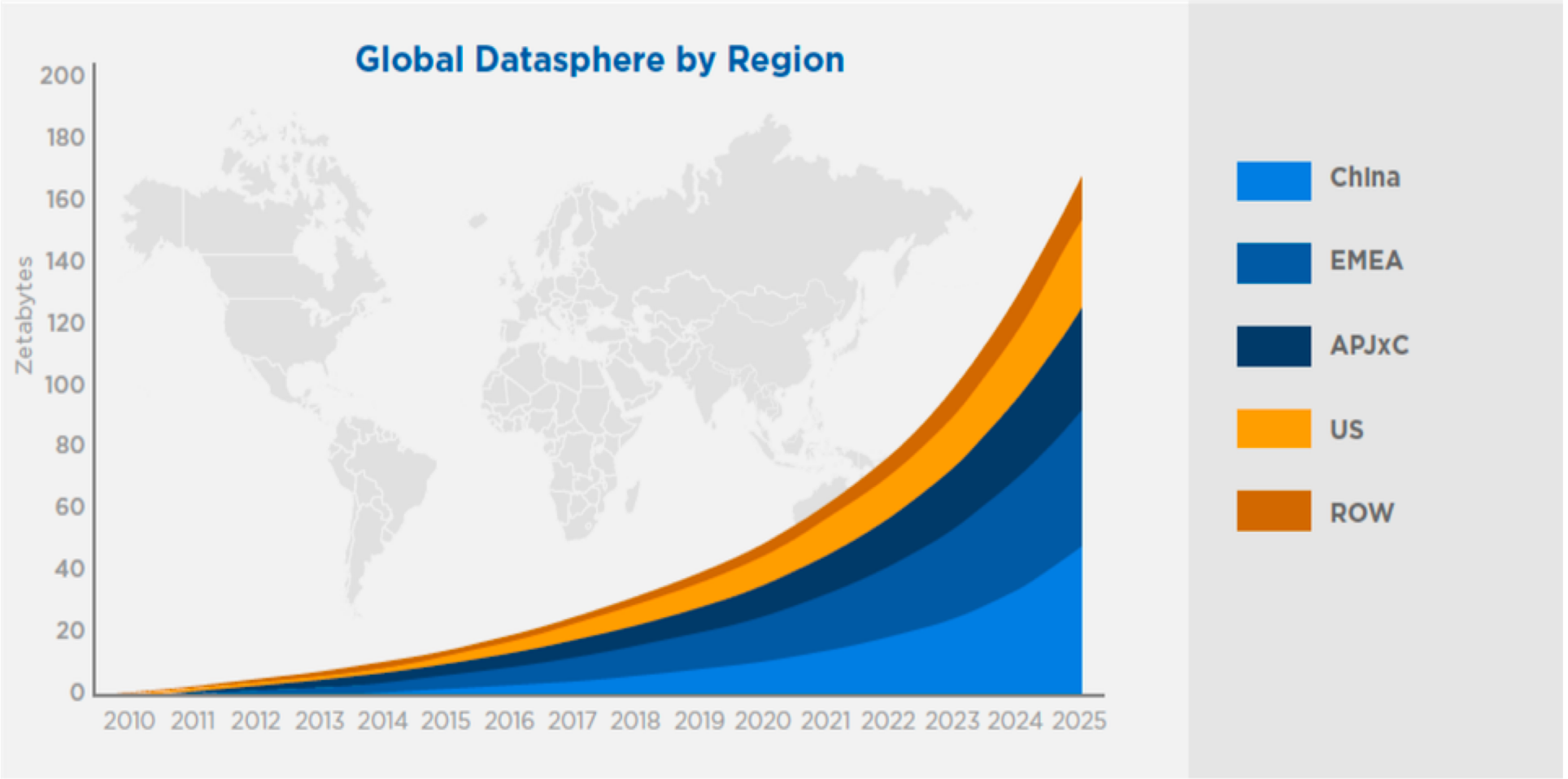


https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/#44e9324c5459

What is Yotabyte?

terabyte	2^{40} bytes	10^{12} bytes
petabyte	2^{50} bytes	10^{15} bytes
exabyte	2^{60} bytes	10^{18} bytes
zettabyte	2^{70} bytes	10^{21} bytes
yotabyte	2^{80} bytes	10^{24} bytes

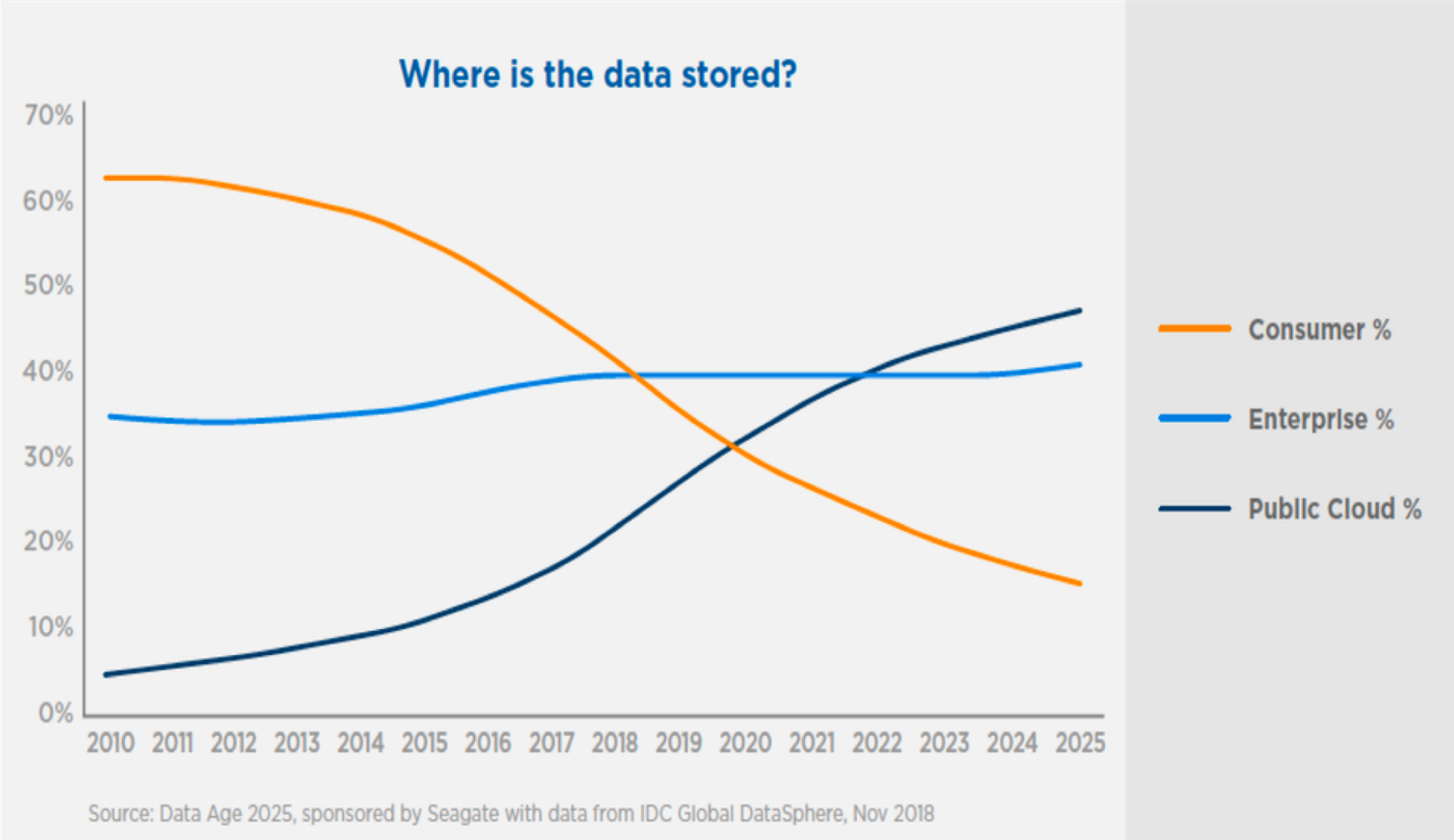
Data growth per region



Source: IDC's Data Age 2025 study, sponsored by Seagate

https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/#44e9324c5459

Where is the data stored?



A few statistics and estimates

- **Worldwide data is expected to hit 175 zettabytes by 2025, representing a 61% CAGR**
  - **51% of the data will be in data centers and 49% will be in the public cloud**
  - **90 ZB of this data will be from IoT devices in 2025**
- 80% of data will be **unstructured** by 2025
  - On top of business **documents**, **video** and **audio** are added new content such as **social media**, **IoT**, **streaming** and **geo data**
- There will be 4.8 billion internet users by 2022, up from 3.4 billion in 2017
- Worldwide public cloud revenue is expected to grow 17.5% in 2019 to $214.3B
  - The largest segment is Cloud Application Services (SaaS), expected to grow to $94.8B in 2019
- **200 billion** devices generating data in the **IoT** by 2020
- 90% of all data in existence today was created in the past two years

https://www.aparavi.com/data-growth-statistics-blow-your-mind/

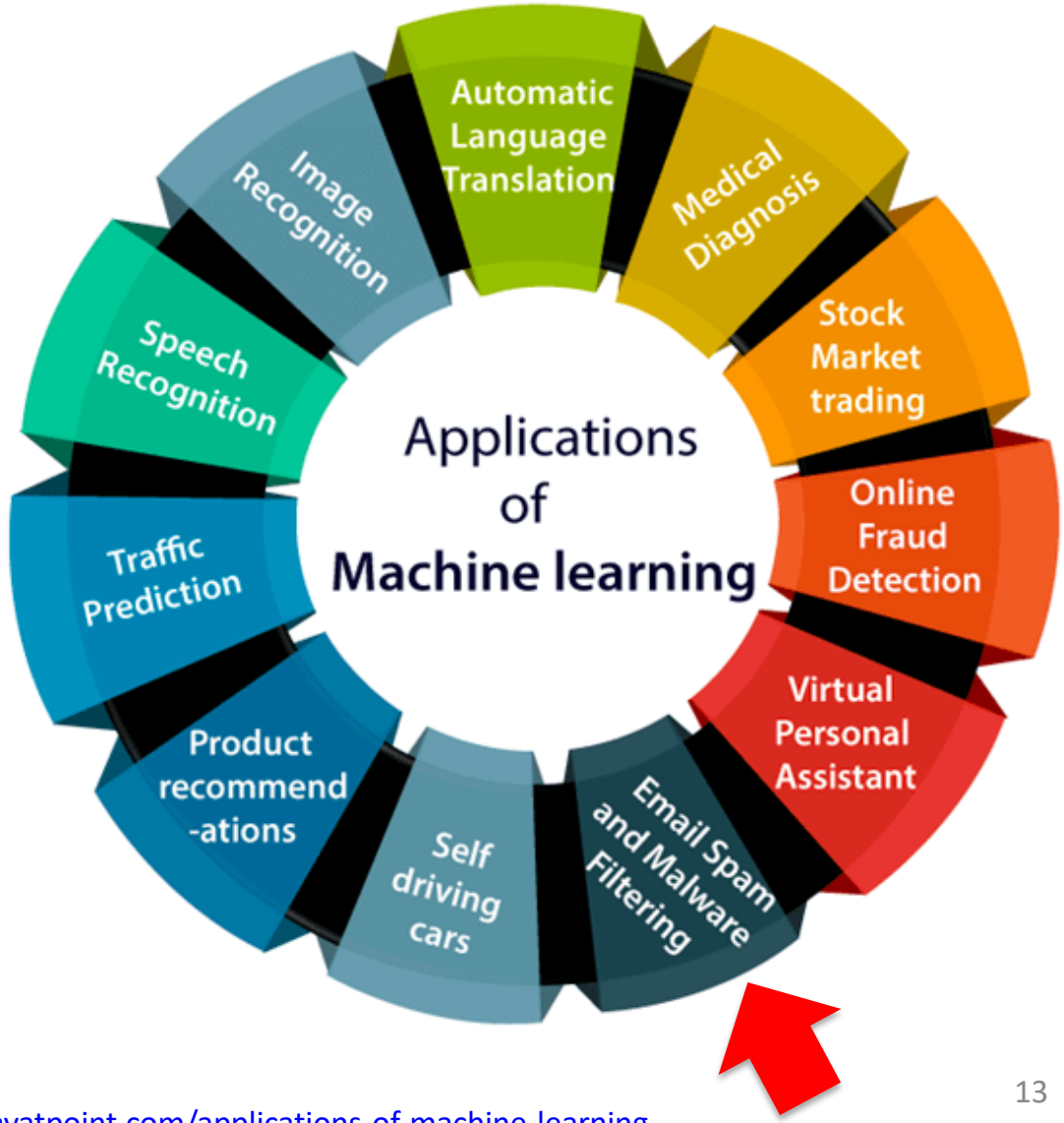
Data Growth Challenges

- The data challenges ahead
  - How to transmit?
  - How to store?
  - How to process?
  - How to learn from?
- Data is the new gold!
  - We can learn from data

Approaches

- Data Science
- Machine learning
- Data analytics
- Data Mining
- BigData
- Business Intelligence
- …

Application areas



https://www.javatpoint.com/applications-of-machine-learning

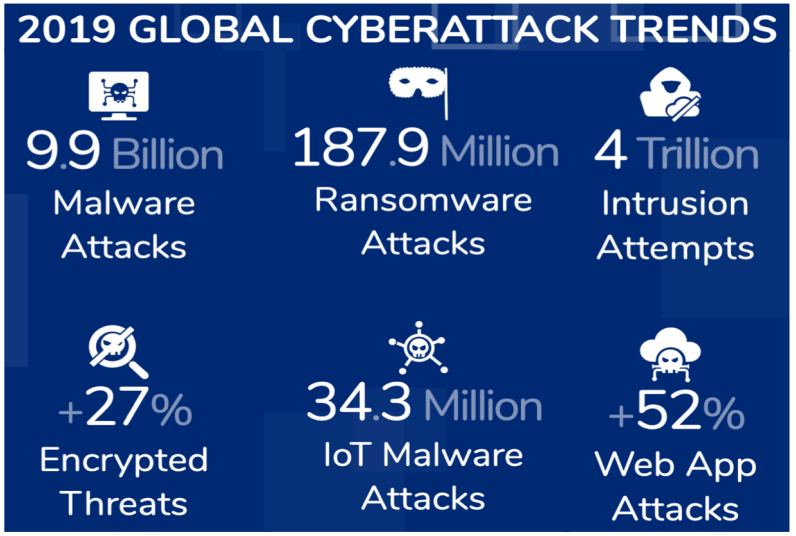
Logistics and supply chain

- AI in Logistics for Demand Prediction
- Supply Chain Planning using Machine Learning
- Warehouse Management
- Track and Warehouse Analysis
- Logistics Route Optimization
- Predicting Peak Hours using AI in Logistics Centers

https://addepto.com/use-cases-ai-machine-learning-logistics-supply-chain/

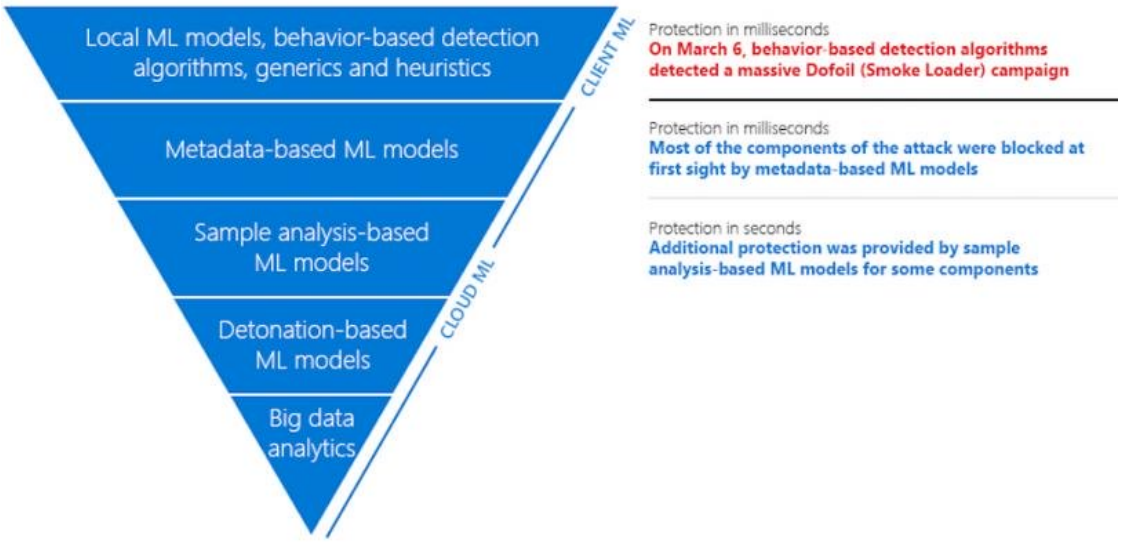
ML in Cybersecurity

- How can ML be used for cybersecurity?
  - It is virtual impossible for humand to analyze every and each attack tentative
  - How to detect “malicious behavior” ?
  - How to detect “zero day” attacks?



Windows Defender ATP

- Windows defender ATP employs multiple ML layers to Identify & stop malicious behavior
  - Ex: In 2018, blocked 10 .5 billion attacks
  - The crypto-miners were shut down almost as soon as they started digging.



MICROSOFT

Location: Redmond, Washington

How it's using machine learning: Microsoft uses its own cybersecurity platform, Windows Defender Advanced Threat Protection (ATP), for preventative protection, breach detection, automated investigation and response. Windows Defender ATP IS built into Windows 10 devices, automatically updates and employs cloud AI and multiple levels of machine learning algorithms to spot threats.

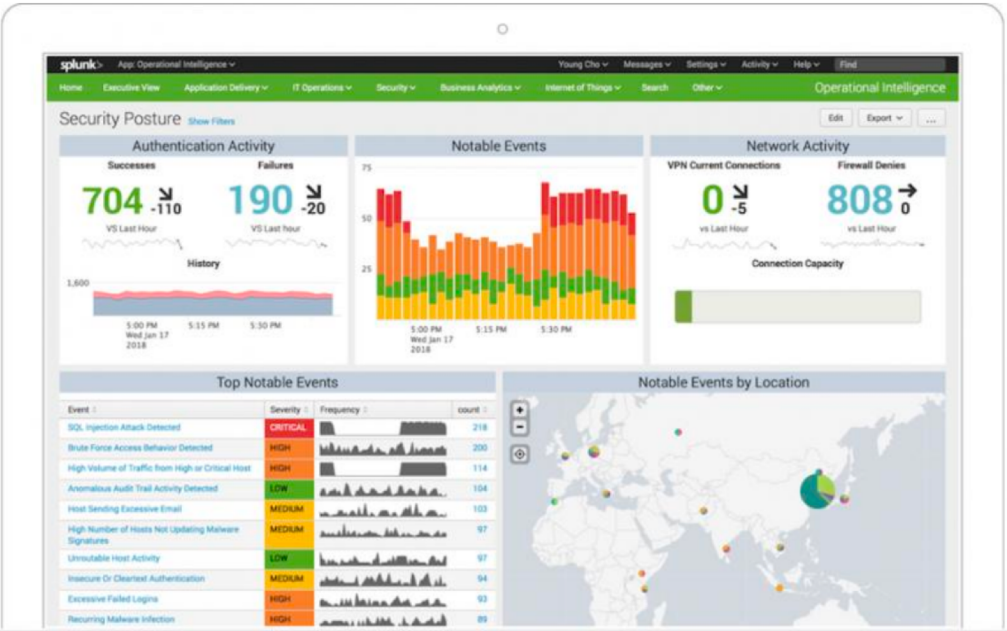
https://builtin.com/artificial-intelligence/machine-learning-cybersecurity

CHRONICLE

CHRONICLE

Location: Mountain View, California

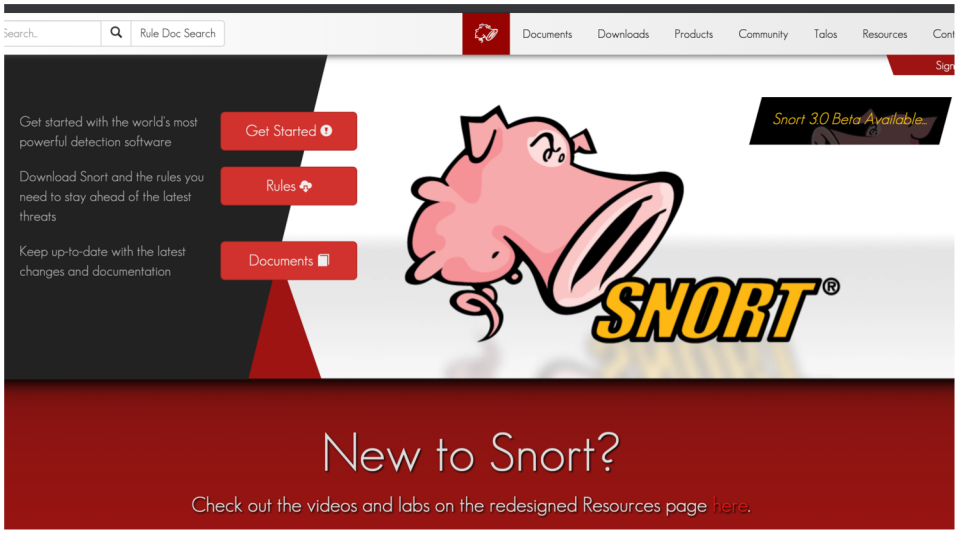
How it's using machine learning: Chronicle is a cybersecurity company that sprang from Google's parent company Alphabet. Its first product, Backstory, has been described as "designed for a world where companies generate massive amounts of security telemetry and struggle to hire enough trained analysts to make sense of it." Backstory analyzes large amounts of security data (such as internal network activity, known bad domains and suspected malware) and uses machine learning to condense it into more easily digestible insights.



https://builtin.com/artificial-intelligence/machine-learning-cybersecurity

SNORT and The Talos Group

- Snort: rule based intrusion detection system (IDS)
- It doesn't use ML itself ...




```
alert tcp $EXT_NET any -> $HOME_NET 53
(msg:"DNS named version attempt";
flow:to_server,established;
content:"|07|version"; offset:12; nocase;)
```

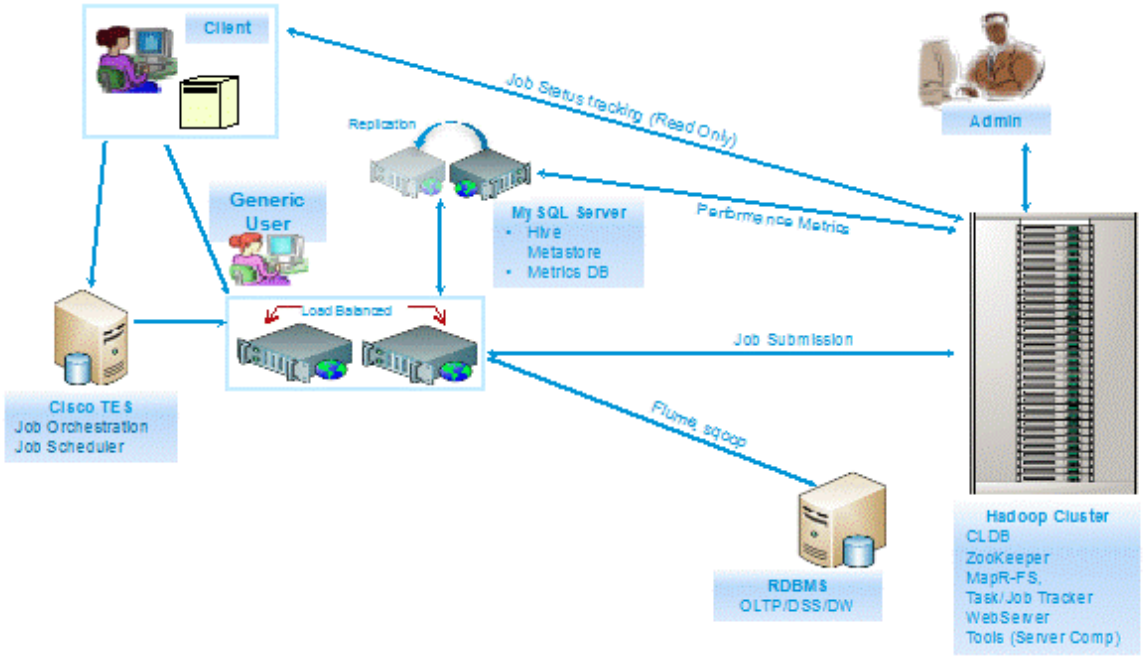
A SNORT rule

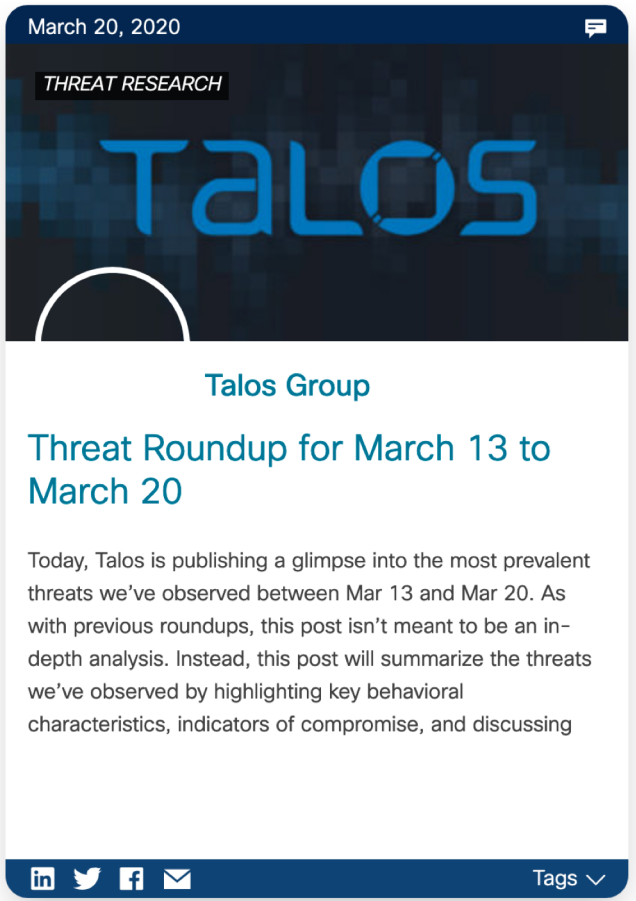
Talos (formerly the VRT) is a group of leading-edge network security experts working around the clock to proactively discover, assess, and respond to the latest trends in hacking activities, intrusion attempts, malware and vulnerabilities. Some of the most renowned security professionals in the industry, including the ClamAV Team and authors of several standard security reference books, are members of Talos. This team is supported by the vast resources of the Snort, ClamAV, and Spamcop.net communities, making it the largest group dedicated to advances in the network security industry.

https://www.snort.org/

The Talos Group (CISCO)

- Collect & analyze data from user's devices
- Combine ML & specialists



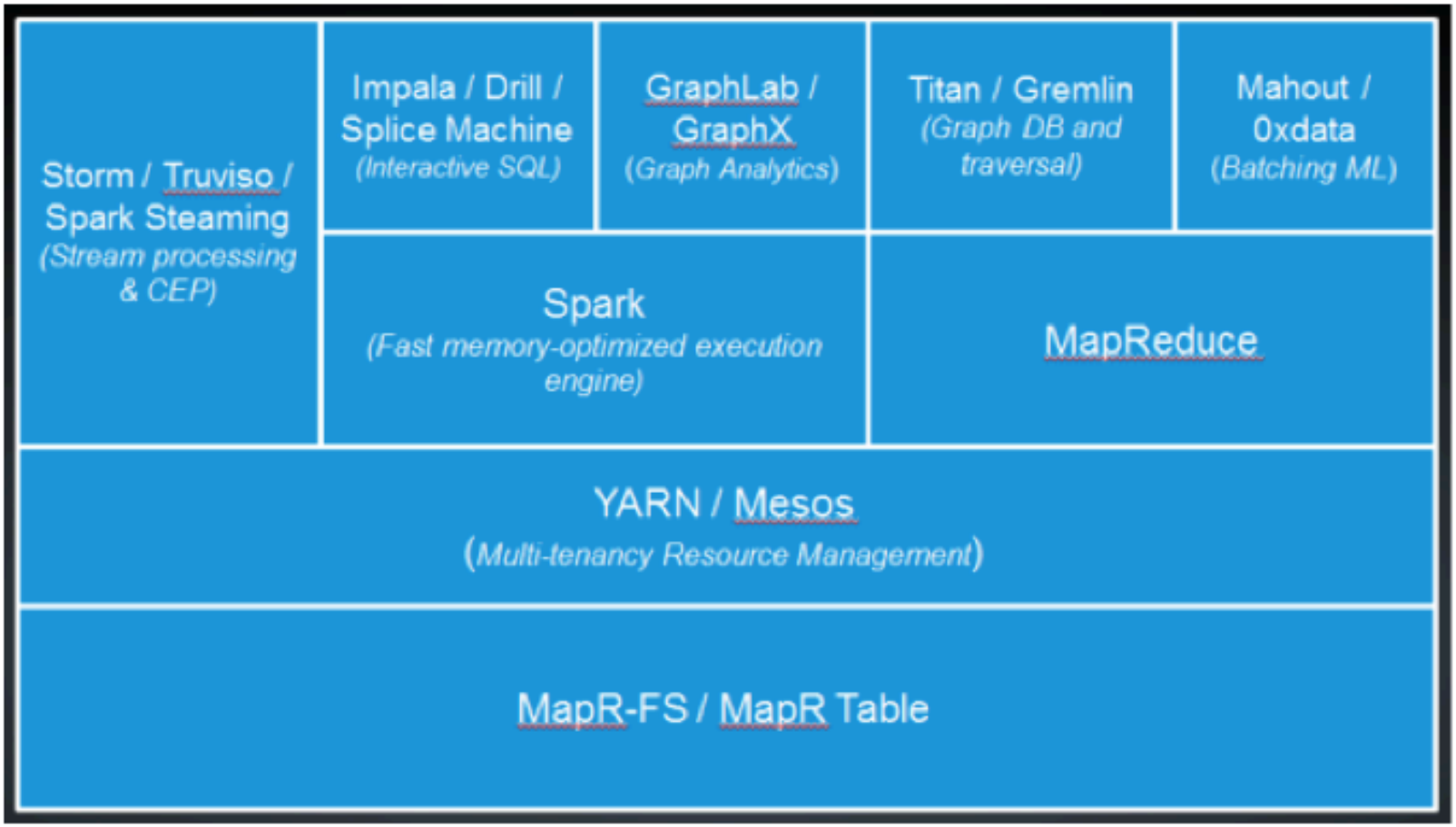


https://blogs.cisco.com/author/talos

Applications - Cisco

- Cisco's Global Security Intelligence Operations (SIO)
  - A 60-node, 1,000-core Hadoop cluster
  - Every day, receives ~ 20 TB of raw log data from local SIO's and data centers around the world
    - telemetry data collected from Cisco's IPS, firewall, email, and Web application logs;
    - freely sourced data from the Internet, e.g. data from Whois, GeoIP, and botnet/darknet data;
    - and malware sandboxing, fire repudiation, and end-user logs from SourceFire FireAMP currently hosted on Amazon Web Services.

Hadoop stack at Cisco's central SIO



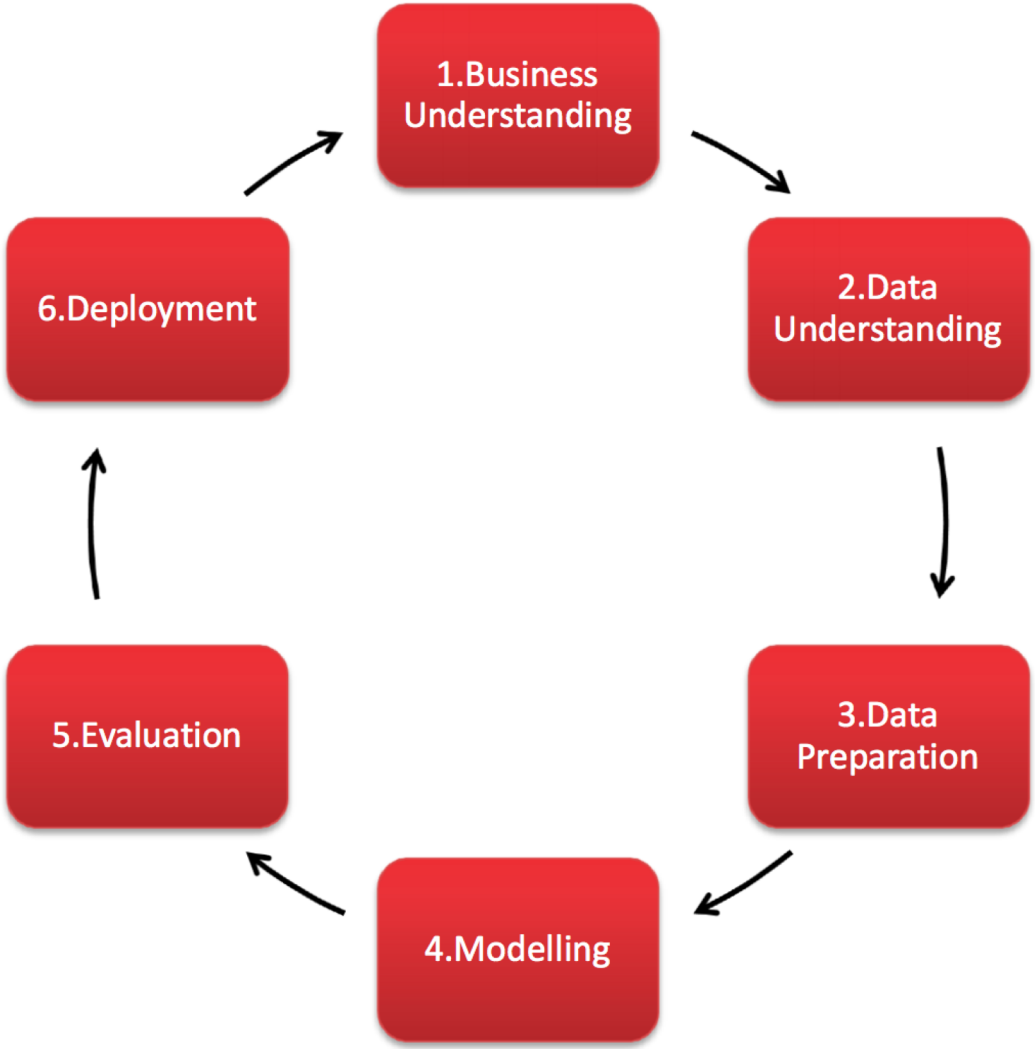
The Hadoop stack at Cisco Global SIO

ML in Cybersecurity

- It is impossible for human specialists to analyze massive amounts of data in real time
- ML can be very helpfull in this case
- It is not a panacea!

https://www.sonicwall.com/2020-cyber-threat-report/

The Process: CRISP-DM



https://www.sv-europe.com/crisp-dm-methodology/

THANKS !