
Sikt

Sikt

Cyber Defense and Big Data

Emil Henry Flakk, systems engineer
Geir Solskinsbakk, data scientist (PhD)
SIKT @ NTNU, 2022-02-03

Sikt

How modern defensive security works and applies big data techniques to achieve its goals.

Sikt

Om **Sikt** – Kunnskapssektorens tjenesteleverandør

- ~80 tjenester
- 370 ansatte
- 700 kunder
- Hovedkontor i Trondheim
- Mellom 1,3 og 1,5 millioner aktive brukere
- Kontorsted i Oslo, Bergen og på hjemmekontorene

Sikt

Samfunnsoppdrag

- **Sikt skal bidra til at virksomheter** (universiteter, forskningsinstitutter, grunnopplæringen osv.) **og brukere** (studenter, forelesere, elever, forskere etc) **i kunnskapssektoren når sine mål**
- **Sikt skal utvikle, forvalte og drifte digitale fellestjenester og infrastruktur**. Blant annet Feide, vitnemålsportalen, samordna opptak, Zoom og forskningsnettet.
- **Sikt skal bidra til sikker digitalisering i kunnskapssektoren** blant annet med cybersikkerhetssenteret og personverntjenester, **og legge til rette for innovasjon.**

Sikt

Noen **få** av tjenestene våre

- Feide
- Eduroam
- **Cybersikkerhetssenter**
- Personvern (NSD)
- Zoom
- Studentweb

Sikt

Cybersikkerhetssenter for forskning og utdanning

- Sektorvist responsmiljø (SRM)
- Varsler som netteier, informasjon fra samarbeidspartnere
- Sikkerhetstjenester

Virksomheten

Eget team for å håndtere hendelser (Incident Response Team, IRT).

Stort sett ansvar for egen drift og sikkerhet.

Sikt

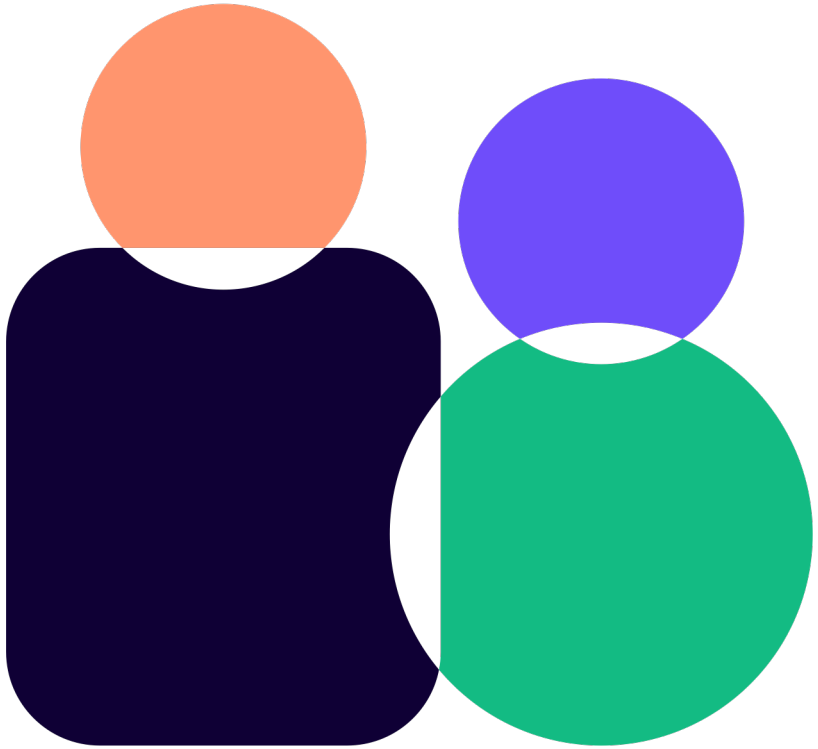
Basic terminology

Incident response (IR). Prioritizing, controlling and resolving (mitigating) incidents.

Forensics. Investigating and determining when, how and why an incident occured.

Threat intelligence. Semi-structured information about threats. Might contain **indicators of compromise** (IoCs), e.g. IPs.

Sensor. Anything that lets us observe something, e.g. application logs, traffic measurements.



Sikt

Business expectations

Protect the mission

Lessen the impact or likelihood of attacks affecting goals

Enable decisions

Provide feedback and suggestions for solutions to security concerns

Build confidence

"Guard rails" prevent FUD when making agile solutions

Sikt

It's kinda hard to win

> 400 Gbps optic fibers

> **1.5M** daily users of our services

> 500,000 *daily* security alerts, ~100 fields/details

~52M unique IP addresses (14d count)

More SaaS services. More data in general.

Encryption makes networks opaque

Attacker needs only succeed once ☺

Customers (e.g. universities) run the endpoints,

Sikt

Other engineering considerations

Threat detection: Soft realtime

Forensics: Long-term retention. When and how data was processed (provenance)

At-least-once delivery. High availability.

Ad-hoc queries to aid hypothesis-based investigations and innovation.

Needs big fat pipes, usually some kind of data bus (e.g. Kafka) and dataflow mgmt

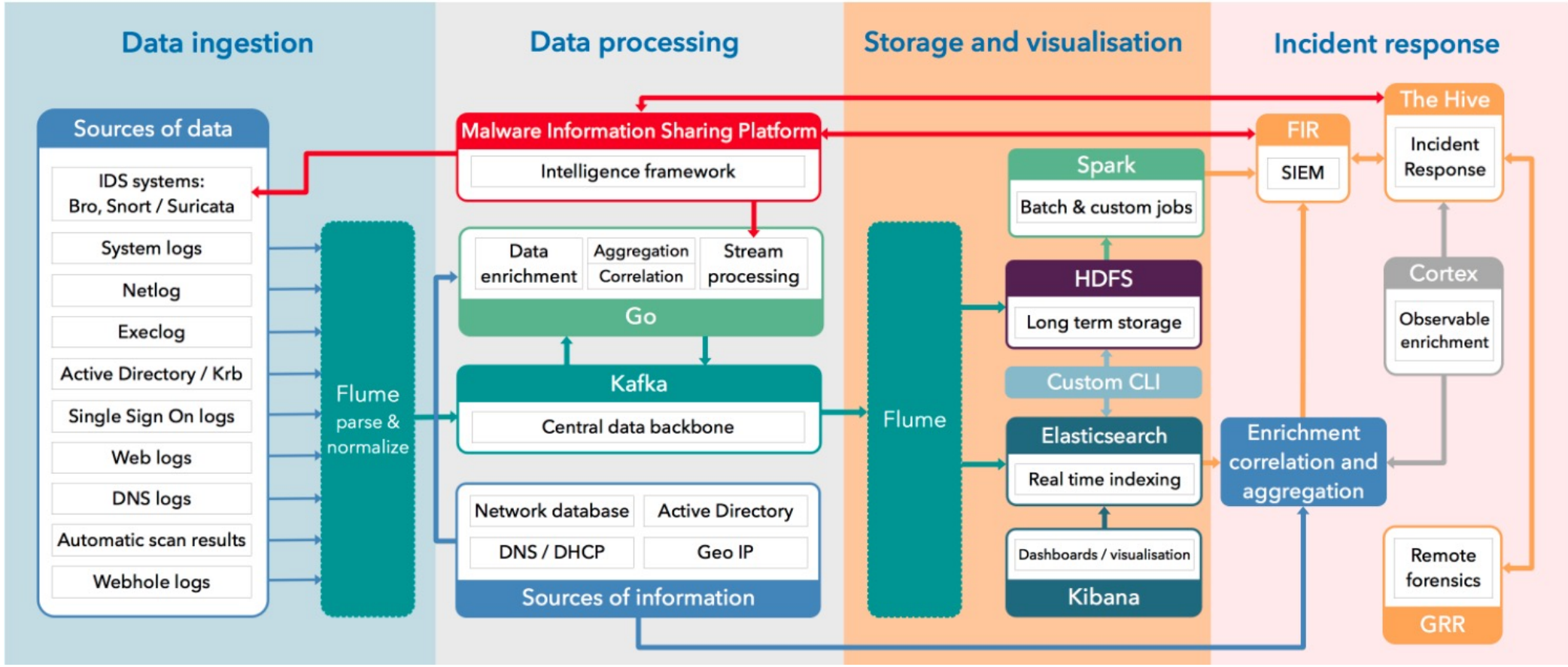
Sikt

Security analysis is big data

- 3V: Velocity, volume, variety.
- #1: How to store and process the data, quickly?
  - Stream analysis – mostly for detection
  - OLAP/data warehouses for investigations and producing more intelligence (insights)
- #2: Not enough *human* time to analyse this much data.
  - Bayesian base rate fallacy (too many FPs)
  - The future is machine learning

Sikt

Example: Modern security infrastructures (CERN)



Sikt

Example of interesting problems: Logs

Find keyword in 1 PB of Apache logs in < 60 seconds

Full-text search is costly and slow. Consider mixed approach:

- #1: Logs are time series. Index the timestamp: "Show me data from this week" (inverted, sorted index)
- #2: CPU is cheap and fast, storage is expensive and slow. Compress the data in order to read and store fewer bytes.
- #3: Bruteforce search on modern hyperscalar processors is actually fast. Use bloom filters to skip chunks of data.
- #4: Store columns separately. Avoids reading the whole row when need a few columns. Improves compresses too.

```
8.211.223.229 - - [03/Feb/2022:14:09:57 +0100] "GET /wp-content HTTP/1.0" 200 5
033 "http://www.murphy.com/wp-content/main.htm" "Mozilla/5.0 (Macintosh; PPC Ma
c OS X 10_10_4; rv:1.9.3.20) Gecko/2016-07-15 00:50:53 Firefox/3.8"
137.153.146.14 - - [03/Feb/2022:14:12:15 +0100] "DELETE /apps/cart.jsp?appID=69
54 HTTP/1.0" 200 4980 "http://shields-hancock.net/faq.html" "Mozilla/5.0 (Macin
tosh; Intel Mac OS X 10_10_9; rv:1.9.2.20) Gecko/2014-06-11 22:34:38 Firefox/3.
8"
119.240.226.82 - - [03/Feb/2022:14:12:45 +0100] "GET /list HTTP/1.0" 404 4892 "
https://powell-stanton.org/explore/login/" "Mozilla/5.0 (iPhone; CPU iPhone OS
14_2_1 like Mac OS X) AppleWebKit/531.2 (KHTML, like Gecko) CriOS/16.0.868.0 Mo
bile/99M735 Safari/531.2"
212.146.162.104 - - [03/Feb/2022:14:15:49 +0100] "DELETE /app/main/posts HTTP/1
.0" 200 5010 "http://jacobs.com/search/" "Mozilla/5.0 (X11; Linux i686) AppleWe
bKit/535.1 (KHTML, like Gecko) Chrome/21.0.838.0 Safari/535.1"
34.246.109.60 - - [03/Feb/2022:14:18:19 +0100] "POST /wp-content HTTP/1.0" 200
5007 "https://www.king-george.com/tags/app/explore/register.php" "Mozilla/5.0 (
X11; Linux x86_64) AppleWebKit/533.0 (KHTML, like Gecko) Chrome/23.0.880.0 Safa
ri/533.0"
91.218.12.226 - - [03/Feb/2022:14:21:22 +0100] "GET /posts/posts/explore HTTP/1
.0" 200 4967 "https://www.hall.net/categories/categories/search/home.html" "Moz
illa/5.0 (iPad; CPU iPad OS 3_1_3 like Mac OS X) AppleWebKit/531.0 (KHTML, like
 Gecko) FxiOS/15.0o1623.0 Mobile/47H311 Safari/531.0"
```

Sikt

Example of search optimalization

- Initially: **1 PB** of raw logs. Looking for **'contoso.com'**
- Compress them (10:1). **100 TB**.
- Only grab this week's logs (10:1). **10 TB**.
- Bloom filter: only chunks with 'contoso.com' (10:1). **1 TB**.
- Only read the columns we need to (10:1). **100 GB**.
- NVMe PCI 4.0 speed (on a good day): ~5 GB/s.
- 100 GB / 5 GB/s => **20 s**. Can go faster with multiple disks, preferably multiple storage buses

Sikt

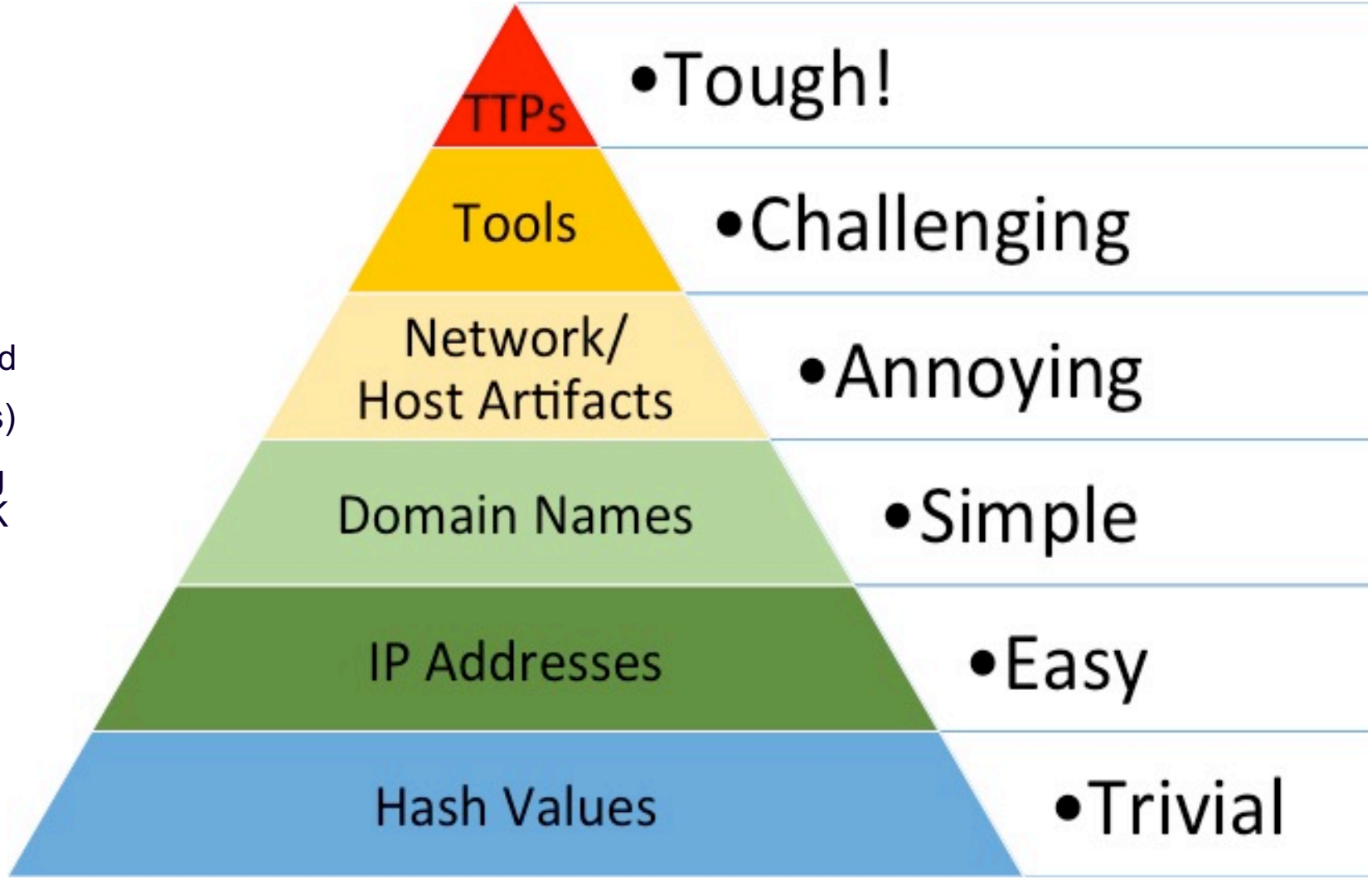
Overcoming adversity

Instead of chasing alerts, remove *opportunities* for attackers to succeed

Tools, Tactics and Procedures (TTPs)

Threat intelligence tracks this using structured data, e.g. MITRE ATT&CK

Build tools to observe and prevent TTPs relevant to our environment



Sikt

Fake login pages

Very convincing clones of service pages across multiple customers

Valid TLS certificates

Hundreds of domains at their disposal. Switches to new domains once detected

Goal: Steal credentials.

Tactic

Common patterns in the naming convention used for certificates

All LetsEncrypt certs are registered in a public log known as **Certificate Transparency (CT)**

Outcome

New streaming detections:

Alert on patterns in newly issued certificates via CT

Alert on DNS lookups to detect compromised users before they're abused.

Could reliably detect and block their attempts 30 minutes in advance

Sikt

Economic fraud

Very easy to start new fraud campaigns. Just change the template and pay Amazon some $$$ to send emails.

Valid TLS certificates

Change domain if detected. Hundreds of them available

Goal: Steal credentials.

Tactic

You need some kind of payment infrastructure, often reliant on *real* humans

Common theme: Financial fraud victims fooled into receiving and sending $$$

Examples of data: SWIFT/IBAN numbers, bitcoin wallet IDs

Outcome

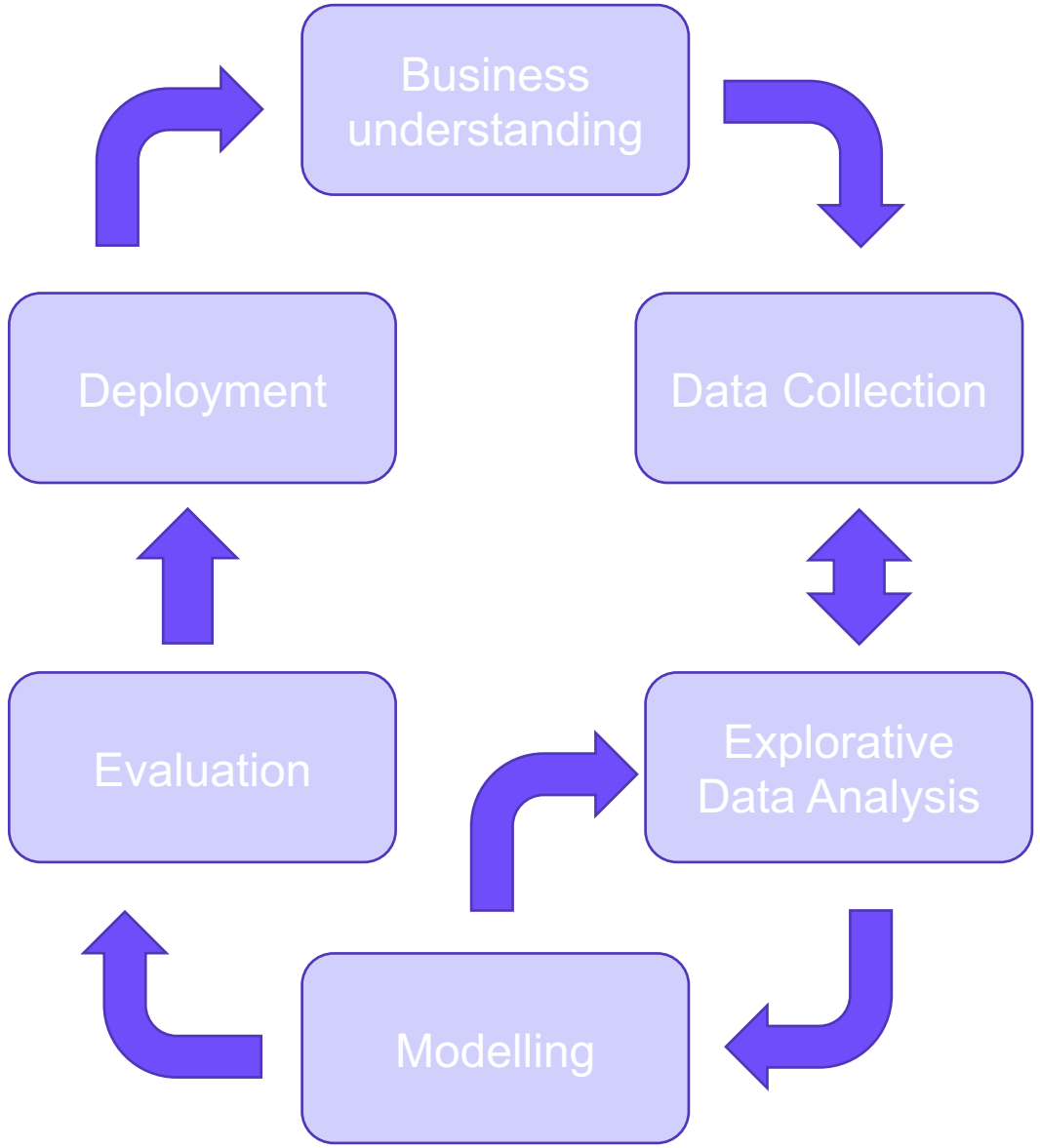
Track and report accounts. Share them with other organizations that might be targeted as well. (Streaming, OLAP analysis, even graph-based techniques)

Attackers run out of viable options to move money and switch to new careers

Sikt

Data Science – what and why?

- Extract knowledge from data
- Increasingly complex threats
  - Manual analysis increasingly complex
  - Machine learning can interpret signals that we as humans can't
- Automate analysis – humans have limited capacity
  - Continuous monitoring of multiple scenarios



Sikt

Goals for analytics

- Finding the needle in the haystack
  - A diverse set of logs – may need to combine many sources to get the big picture
- Extract insights / alerts / events that are interesting for further analysis
  - Answering one question opens up another
- Start with manual analysis
- Automation of analysis
  - Important factors: scale and response time
- False positives vs. False negatives

Sikt

OLAP example: Passive DNS intelligence

- DNS queries are usually not encrypted, so let's log them. Drop client IP for privacy reasons.
- Aggregate on (RRname, RRtype, answer), e.g. (vg.no, *A*, 195.88.54.16) + first/last seen.
- "Wow, this domain stores malware instructions in its TXT record"
- "When did someone first visit totallynotavirus[.]com?"
- "Has anyone queried Russian domains containing the name of our company?"

Sikt

Adding context to existing data

- Fact tables: Knowing what the network is used for makes triaging easier (e.g. GeoIP, ASN)
  - VPN traffic in the server DMZ doesn't look good
- What is the entropy of a domain?
  - English has skewed letter distribution and hence will be less random (less entropy)
  - More random domains are likely sampled from uniform distribution and hence machine generated, for use by malware etc

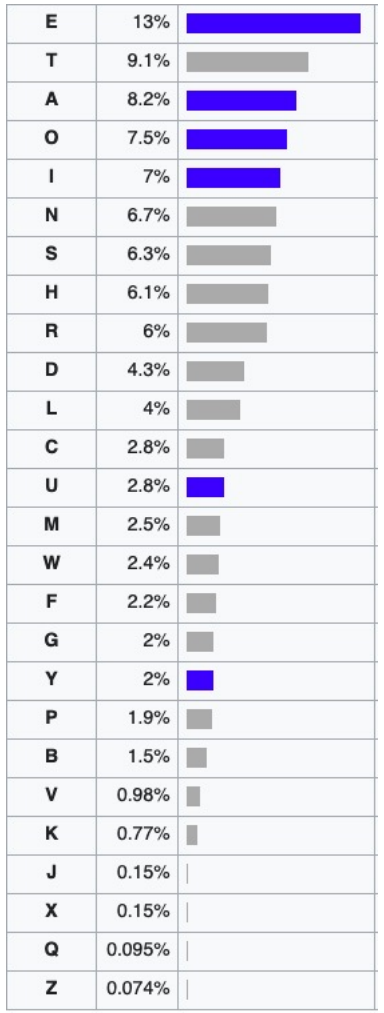


Table of letter frequencies in English language from: https://en.wikipedia.org/wiki/Letter_frequency

Sikt

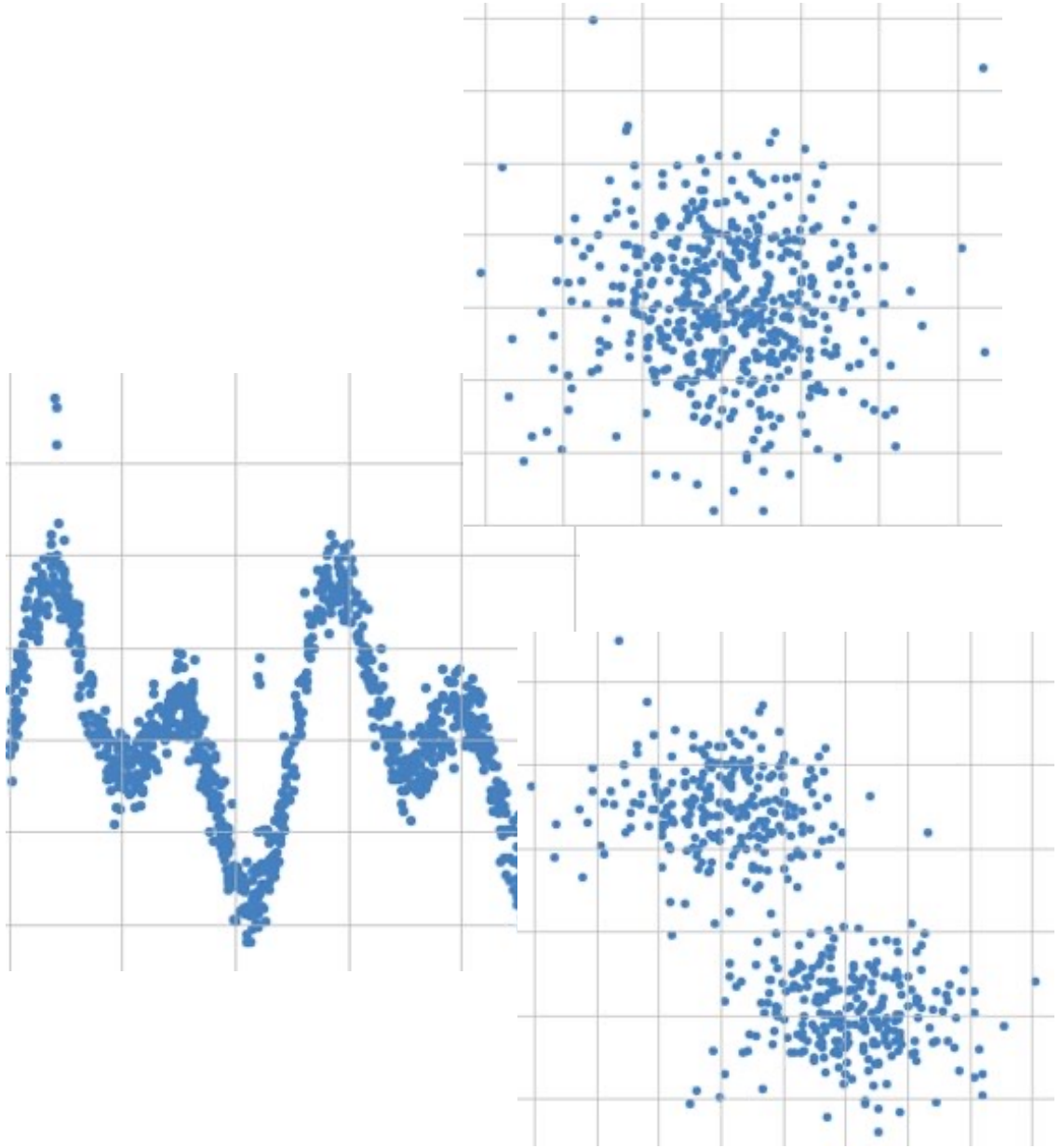
Machine Learning

- Supervised
  - Labeled data sets
    - Good labeled data sets are hard to get / costly to make
    - Outdated?
      - A dog will not stop being a dog
      - New attack vectors emerge
  - Classification / regression / ++
- Unsupervised
  - Unlabeled data
  - Discover structures in data
  - Clustering / anomaly detection / ++

Sikt

Anomaly detection

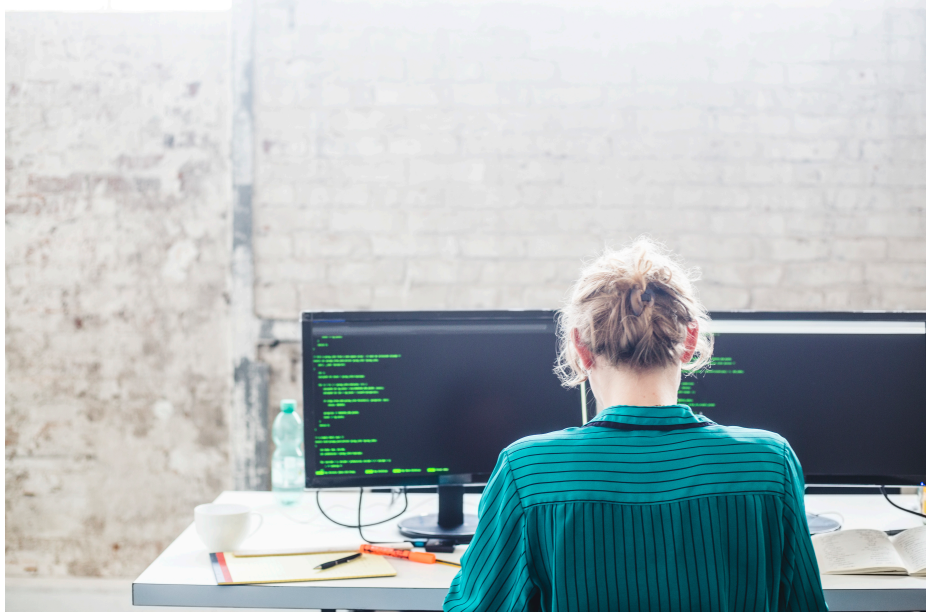
- What is an anomaly?
  - One or more data points that come from a distribution different from the regular data / data generated from a different process
- Measured vs. existing (baseline) data
  - Do we know for certain that the existing data is clean?
- Anomlies can point at interesting parts of the data
  - Pivot points
  - Combine with other data to make more sense



Sikt

Anomaly detection – malicious logins

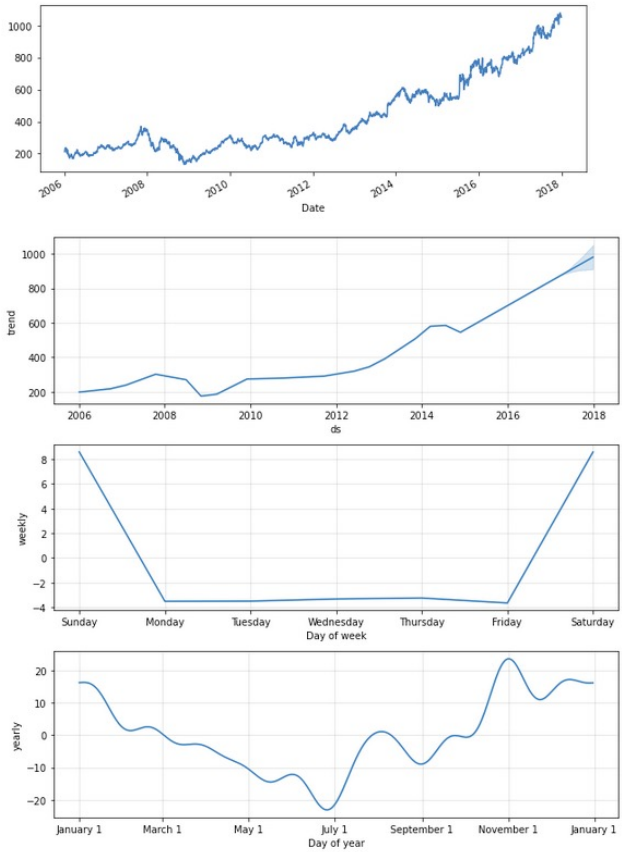
- Goal: Find logins that seem suspicious
  - E.g., password leaked / user hacked / ++
- "login heuristics"
  - Working hours?
  - Location?
  - Behaviour after login?
  - Tor / VPN?
- A login from a usual location in the middle of the night?
- Login from a different continent at an unusual time?



Sikt

Anomaly detection – time series

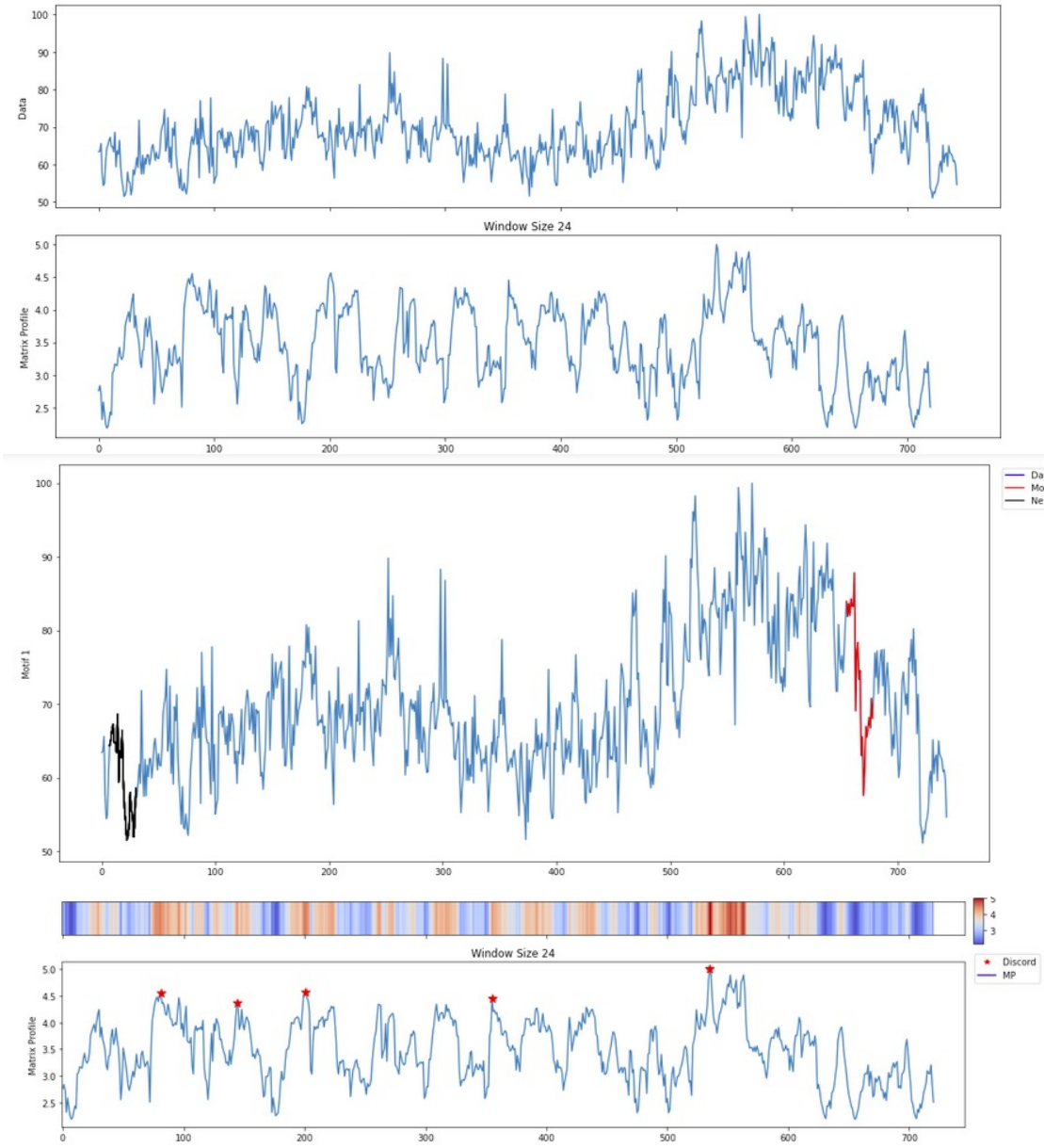
- Forecasting time series
  - Confidence intervals
  - Use this to find outliers / anomalies
    - Trends + uncertainty
  - E.g. prophet from Facebook
  - Alert on spikes in failed logins vs. Expected number of failed logins



Sikt

Anomaly detection – time series II

- Matrix profile
  - Time windows
  - For each time window store the similarity and location of the most similar time window
    - Z-normalized window
- Recognize repeating patterns / anomalous patterns
  - E.g. search for traffic into/out from a server that fits malicious profile
  - Find patterns that don't fit with the "baseline"



Sikt

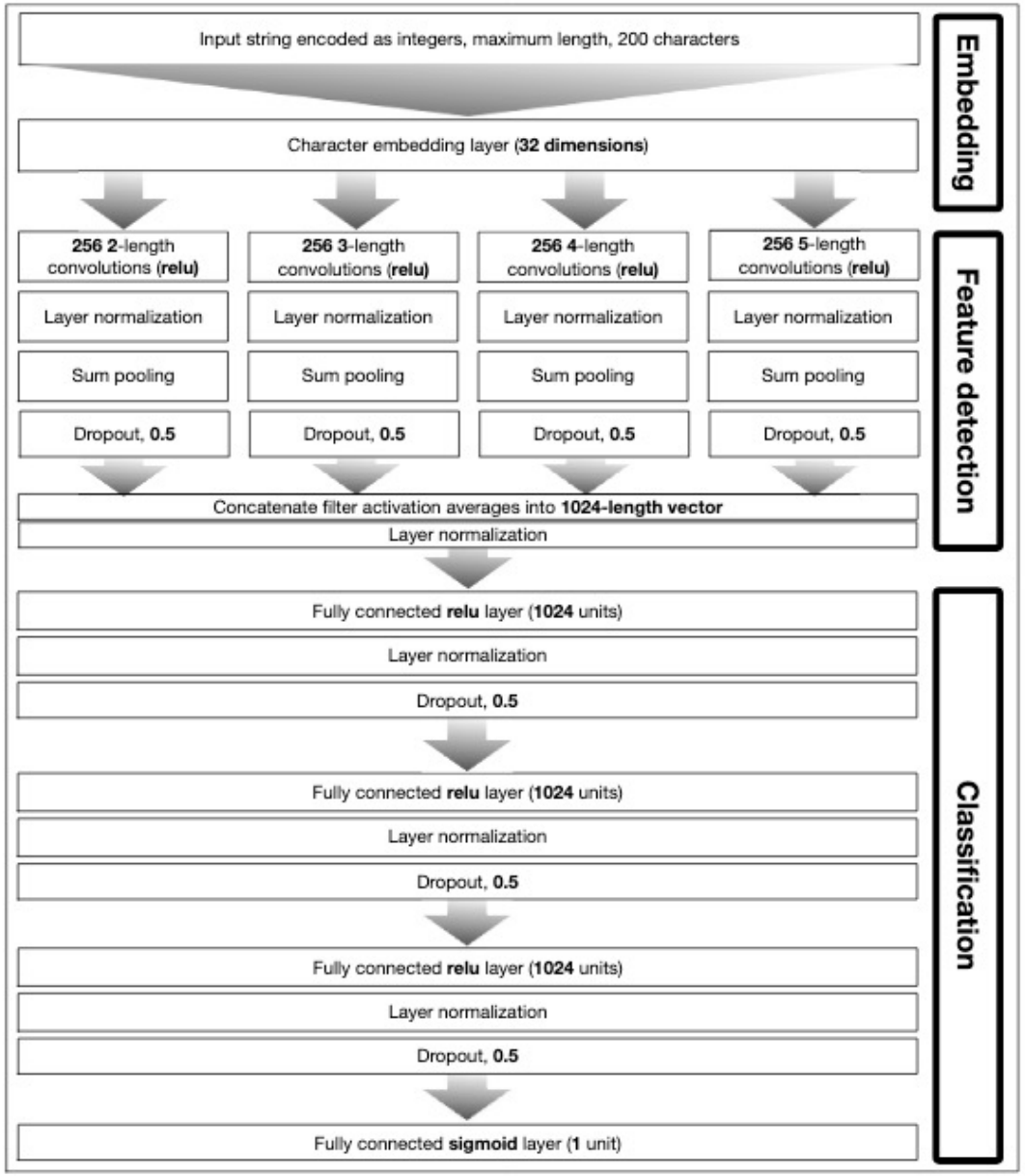
Traffic analysis

- Packet capture analysis
  - Indicators: ips, payloads, signatures ++
- With more encrypted traffic these indicators are to a larger extent hidden from view
- Encrypted traffic analysis
  - Is the conversation benign or malignant?
  - Can't inspect the packages
  - Inspect for (anomolous) patterns in the traffic
    - Timing
    - Sizes
    - Contact points
    - Ports
    - ++

Sikt

Classification – malicious URLs (supervised)

- Goal: recognize malicious URLs
- Dataset
  - Large number of benign URLs
  - Large number of malicious URLs
- Deep learning
  - Learns features – don't need to do a lot of feature engineering



Techinque and figure: Saxe & Berlin: **eXpose: A Character-Level ConvolutionalNeural Network with Embeddings For DetectingMalicious URLs, File Paths and Registry Keys** (https://arxiv.org/pdf/1702.08568.pdf)

Sikt

Analytics - Summary

- No silver bullets
  - Large number of scenarios that require different techniques (perhaps in combination)
    - Adding up evidence
  - Big variation in data sources
  - Understanding domain and exploration of the data space is key
- Need to keep the false positives at an acceptable level
  - Alert fatigue
- Scaling from manual analysis to automated production environment

Sikt

DIY

Want to play around with security analytics?

Our recommendation:

- Free data from https://opendata.rapid7.com/
- Shodan.io has a generous free tier for students ☺ Just email their support.

Common software for analytics

- ClickHouse (probably our favorite)
- Spark
- Jupyter notebooks / pandas / scikit-learn / ++
- Feel free to email us for further suggestions ☺

Sikt

Questions?