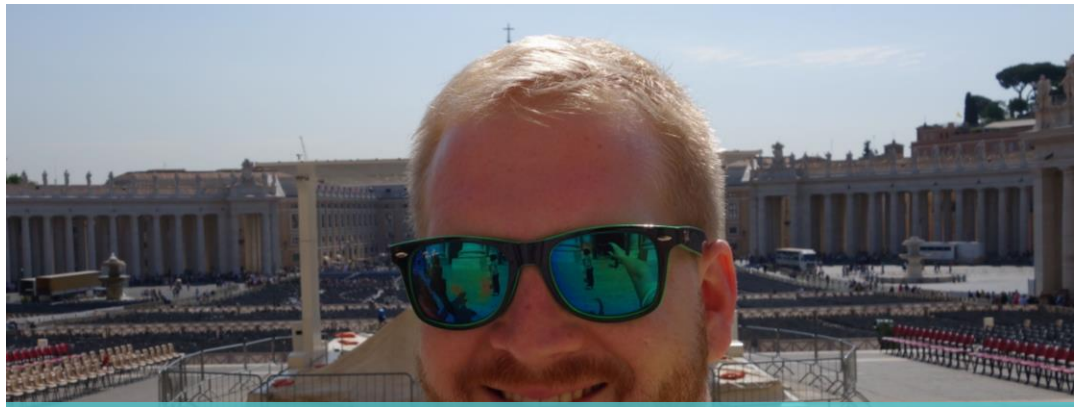
Holmes Processing

Cyber Threat Intelligence at Scale

Who we are



George Webster

Technical University of Munich

PhD Candidate

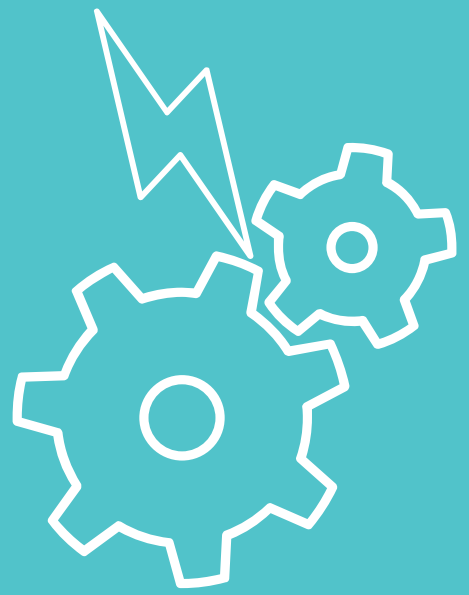
- Scalable Methods for Cyber Analytics
- Static and Dynamic Analysis
- Distributed Systems
- Over a Decade in Industry and Academia

Christian von Pentz

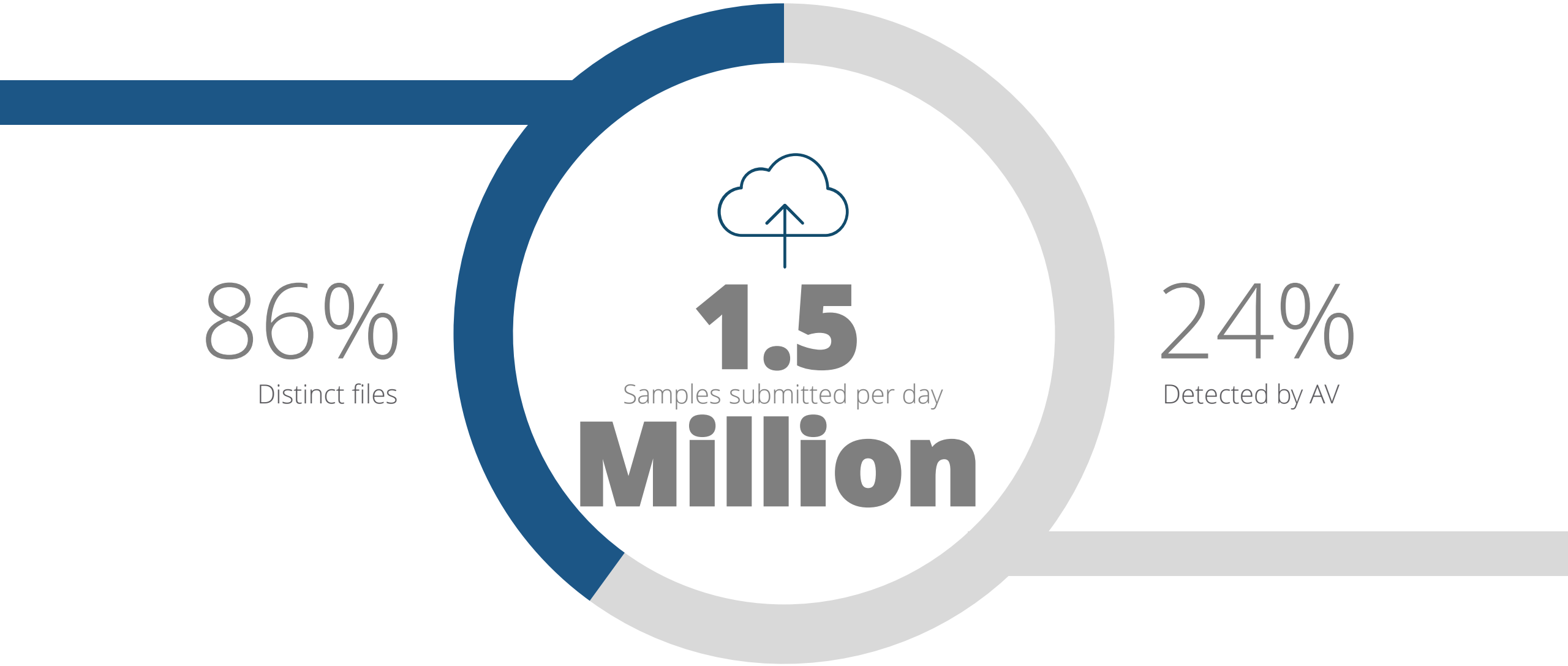
Technical University of Munich

Research Assistant

- Web Application Security
- Scalable Architecture Design
- Process Automation
- CTF player



Problem


86%
Distinct files
1.5
Samples submitted per day
Million
24%
Detected by AV

Current Solutions

Great for what they were designed for but...

cuckoo

yara

DO THEY SOLVE THE BIG PROBLEM?

CRITS

VIPER

01

Designed to solve one problem

Most solutions are designed to get the job done for a single purpose

02

...but they are disjointed

Our tools are not designed to interact with each other

03

Do not easily scale

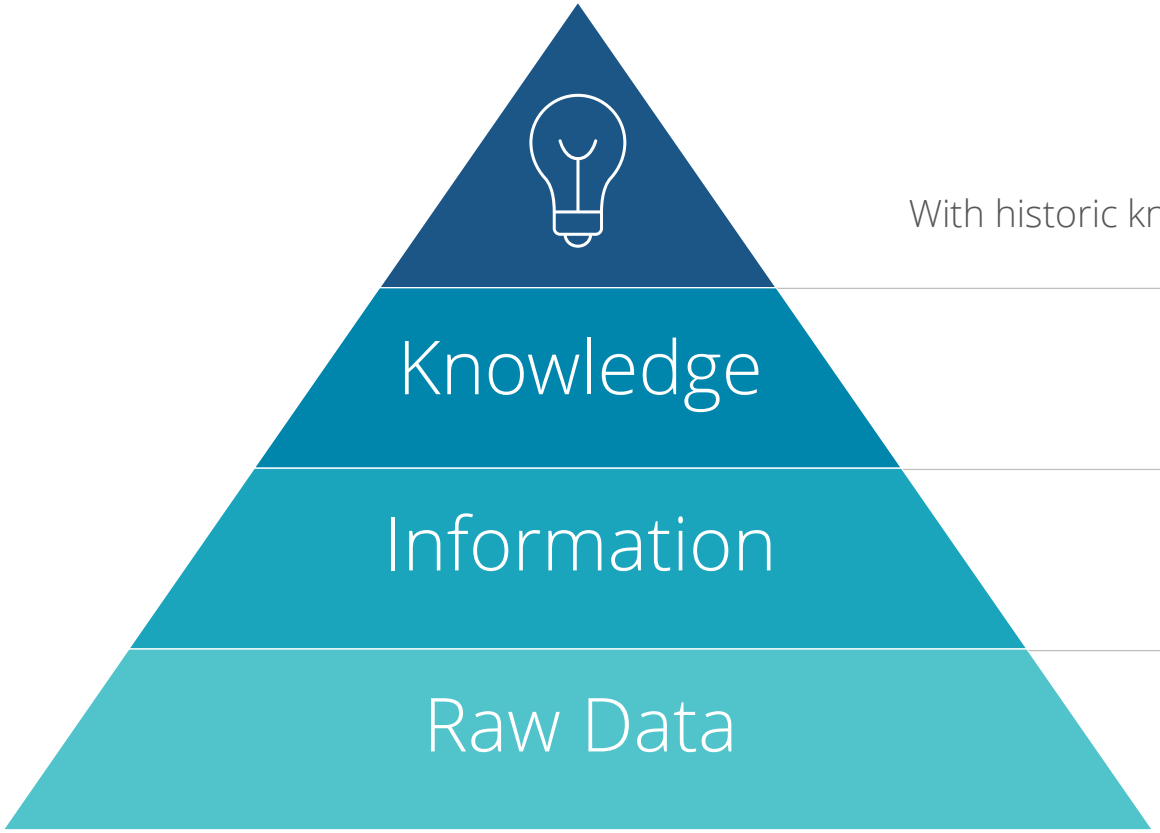
Often monolithic in design and focused on scaling with more powerful workstation

04

Do not support collaboration

How do we work together as teams with tools that are disjointed and do not scale? Let alone, not design to support teams

How to Achieve Our Goals

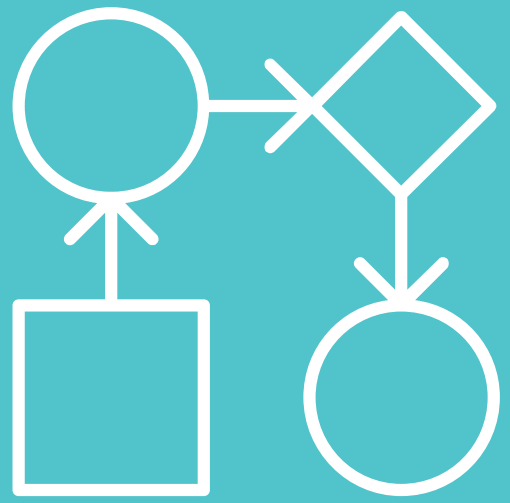


With historic knowledge and human power can be make decisions?

How do we then make sense of these features?

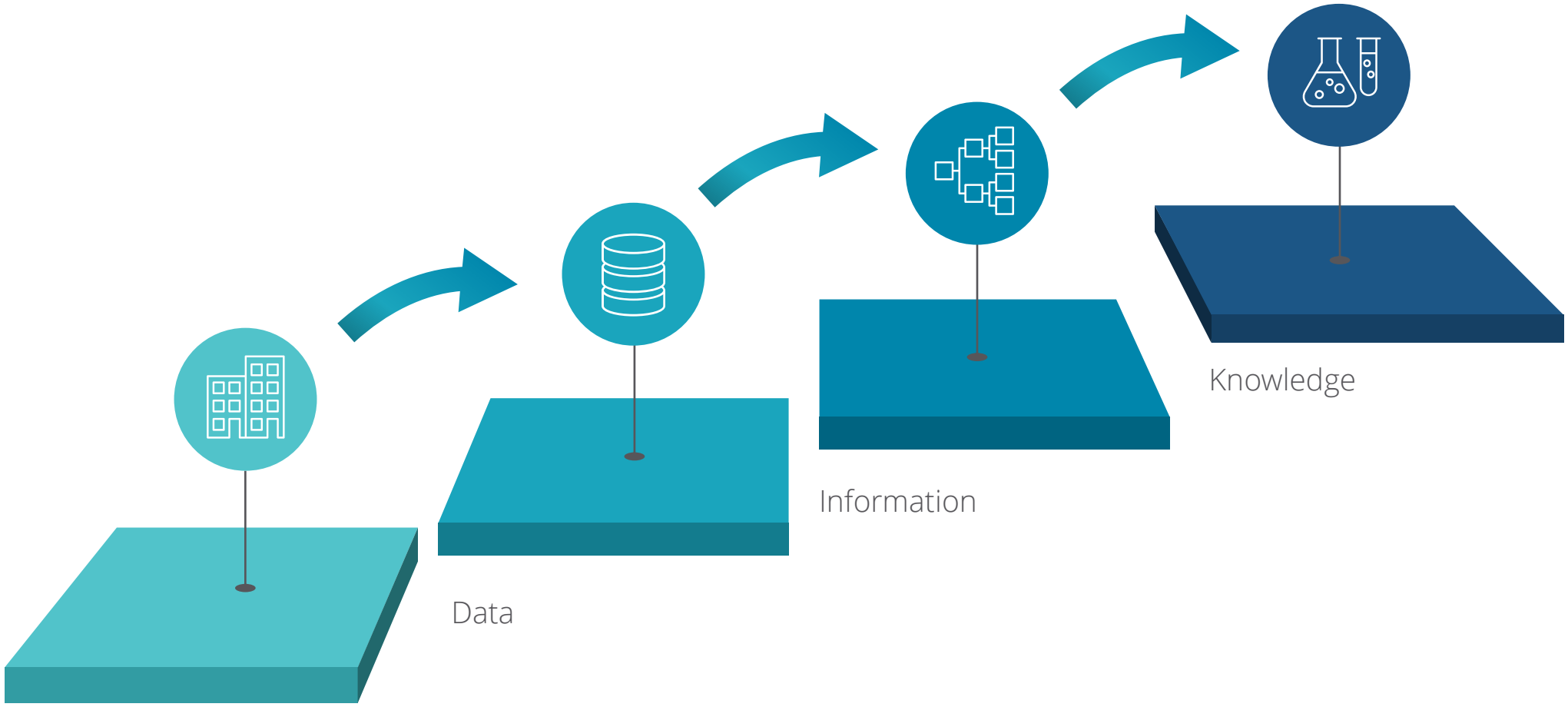
How do we extract useful features?

How do we manage the volume?

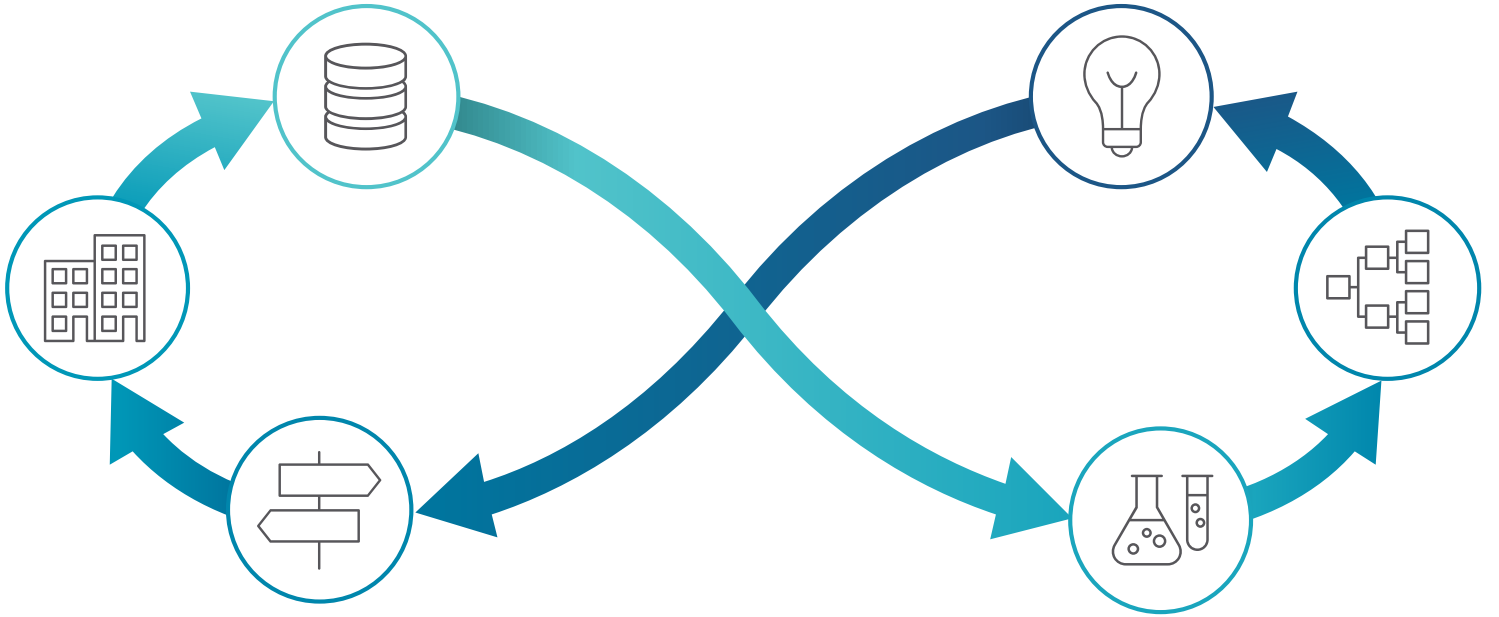


The Analytic Lifecycle

Analytic Lifecycle



Analytic Lifecycle



Raw Data
PE32, PDF, ELF, Domain Name,
IP Address, Zip, JS, ...

Knowledge
Make sense of the intelligence

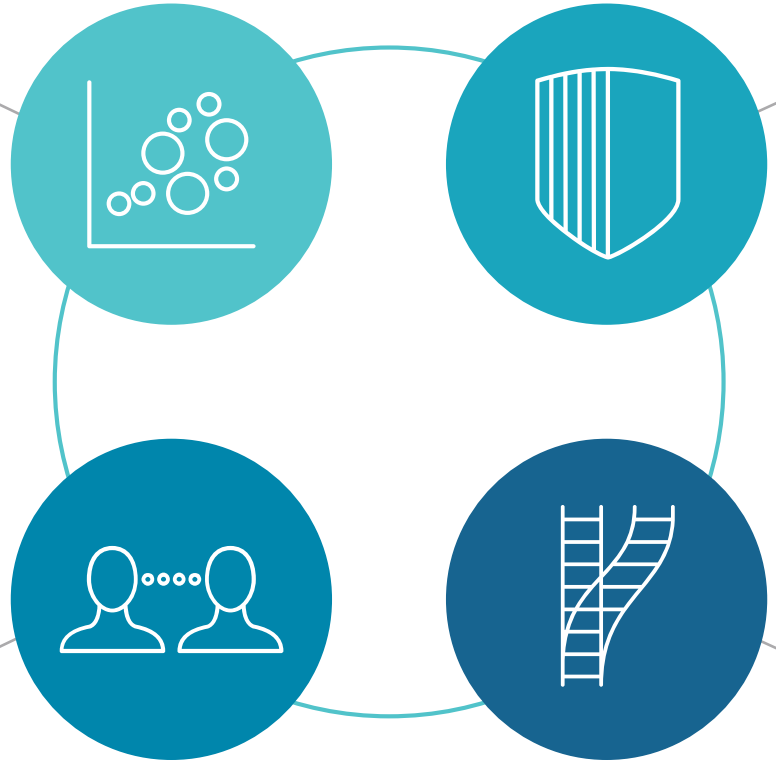
Operational Environment

Intelligence
Organize features in useful ways

Make Decisions
Do I need more data?
Do I need to adjust defenses?
Do I need to take major action?

Information
Extract features that can be useful
for analysis

Support for the Analytic Lifecycle



Scalable

Cope with the volume of data

Resilient

Handle errors gracefully, stable environment

Collaborative

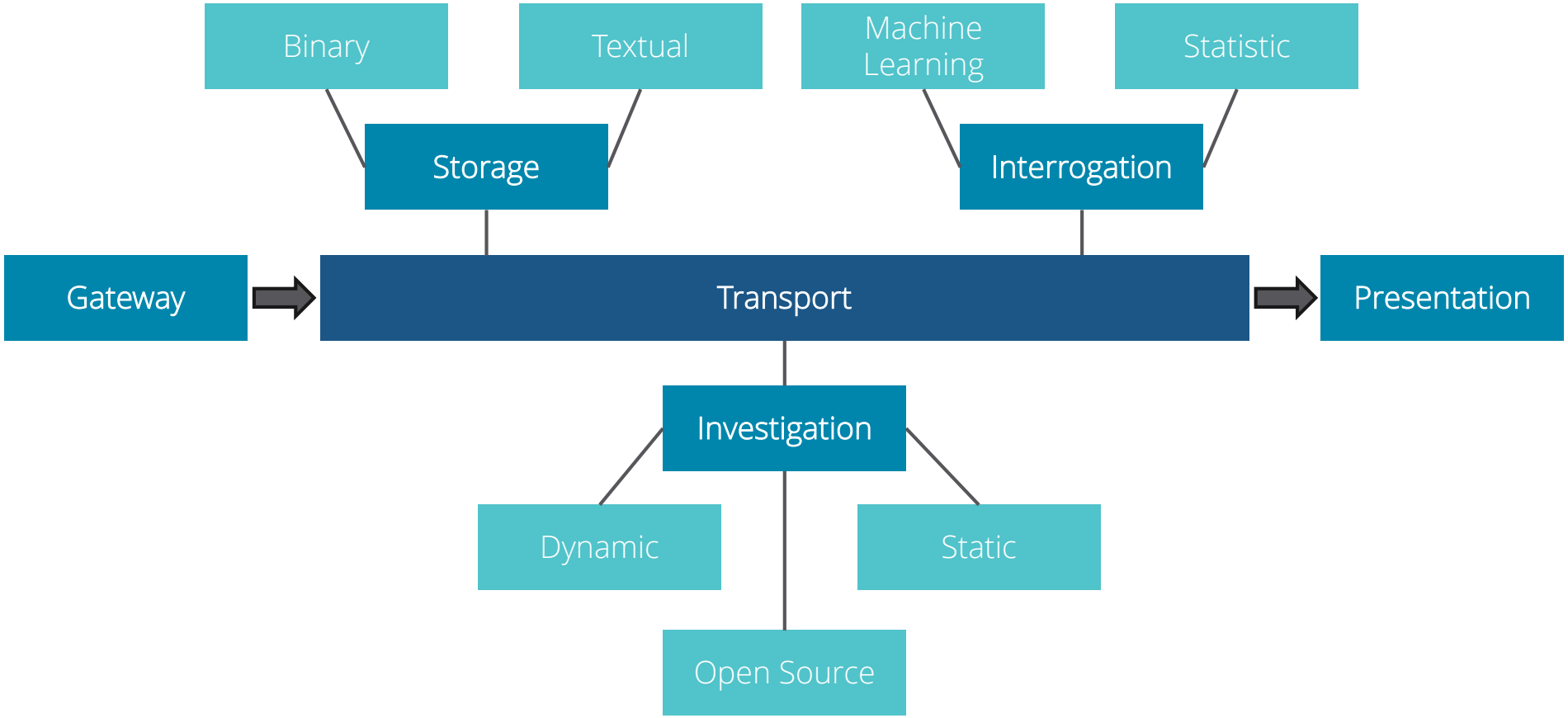
Enable teams and peers to work together

Flexible

Incorporate future trends and methods

SKALD &
Holmes Processing

SKALD

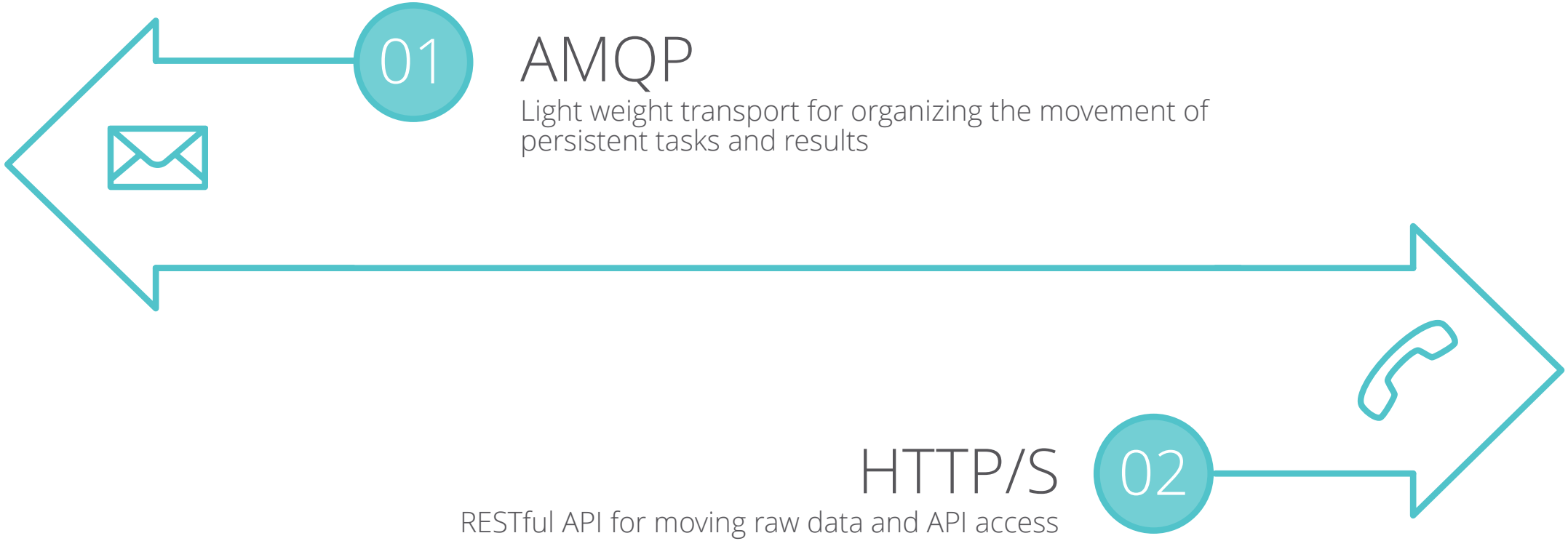


Transport – Moves data between planners

Planner – Orchestrates execution of taskings

Service – Executes work

Transport



01

AMQP

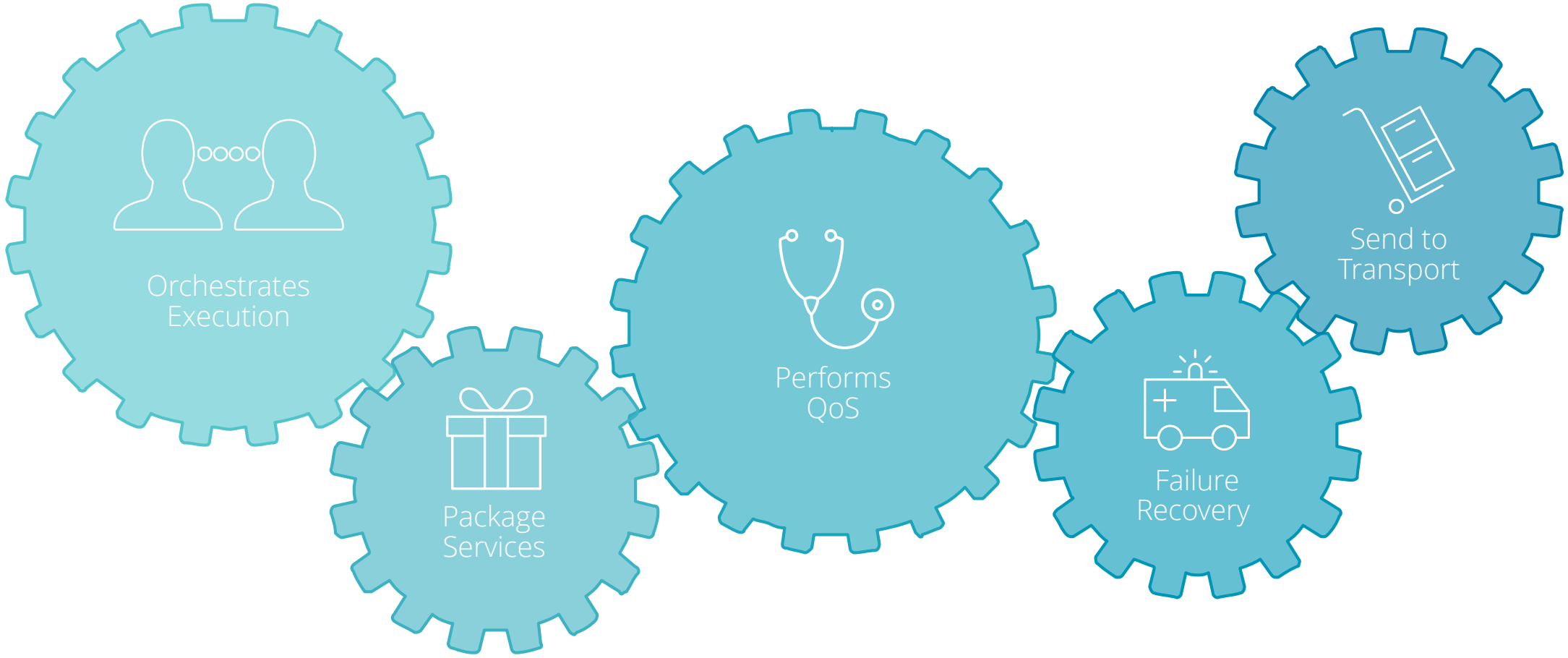
Light weight transport for organizing the movement of persistent tasks and results

HTTP/S

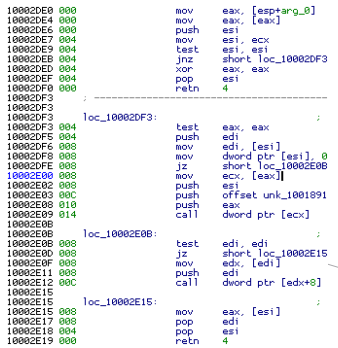
02

RESTful API for moving raw data and API access

Planner


Orchestrates Execution
Package Services
Performs QoS
Failure Recovery
Send to Transport

Service - Information



Static Analysis

yara

Third-Party Information

Dynamic Analysis

cuckoo

RESTful

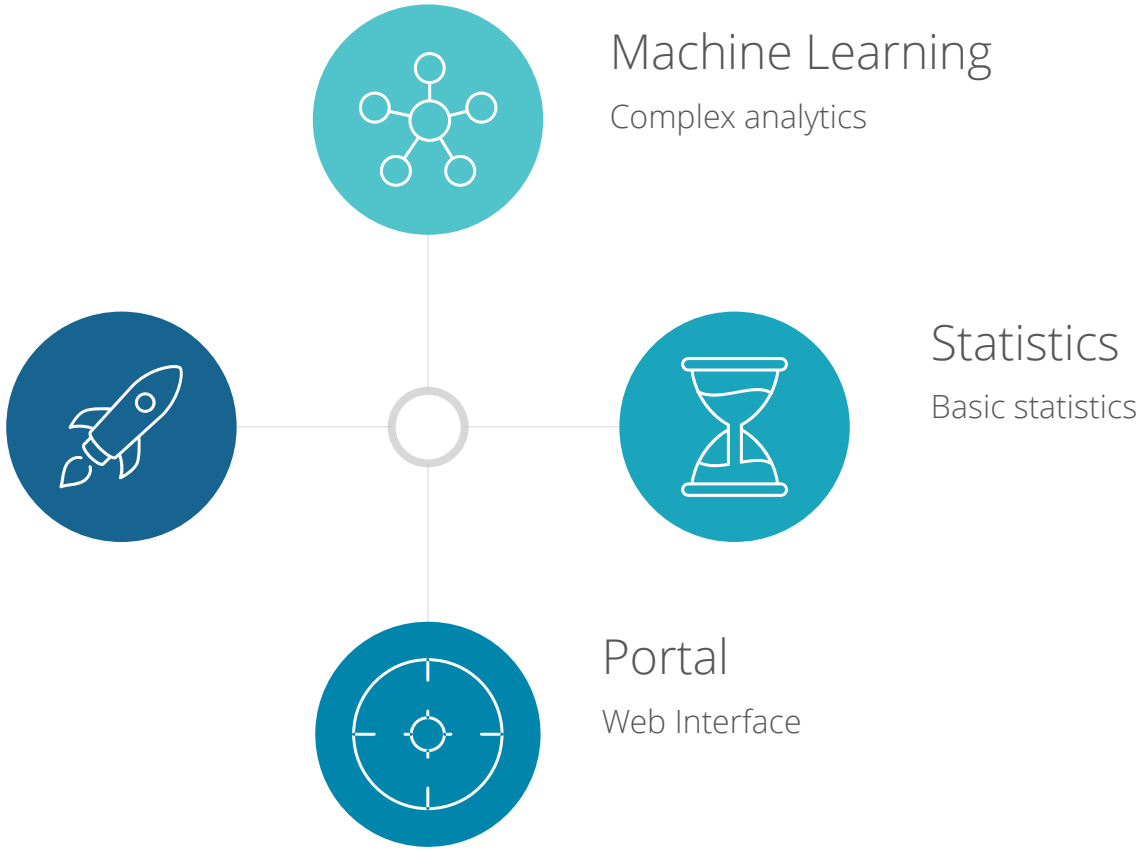
virustotal

Focused Work – Optimized to perform one job

Loosely Coupled – If a failure occures it does not propigate

RESTful – Easy to understand interaction methods

Service - Knowledge



Machine Learning

Complex analytics

Graphical Representation

Maltigo, Palentir, etc

Statistics

Basic statistics

Portal

Web Interface

What is our System?

Holmes Processing

Gateway

Receiving tasking and objects

Totem

High performance scheduler

Totem-Dynamic

Long running scheduler

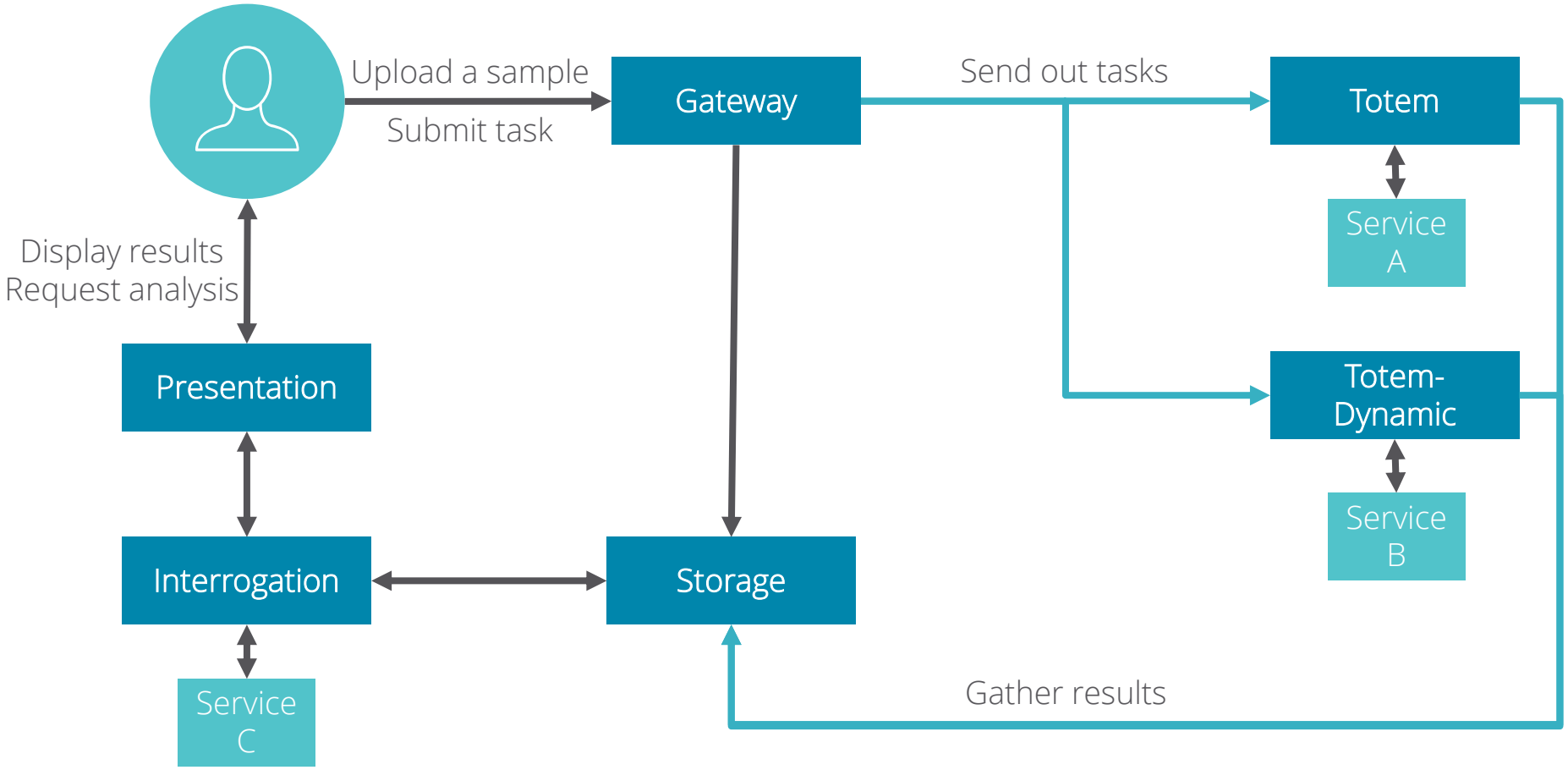
Storage

Lightweight data access and management

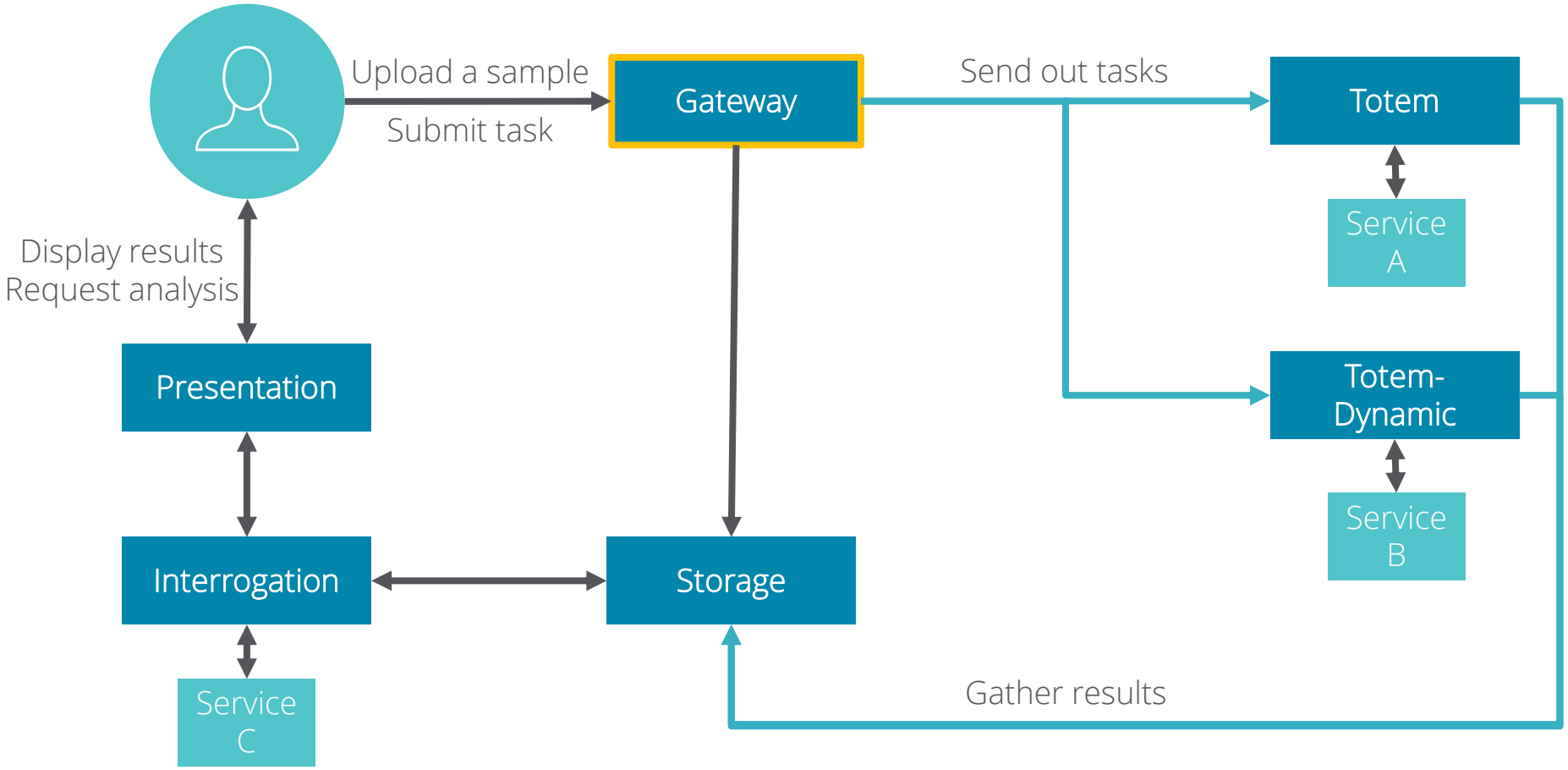
Interrog-ation

Data analysis and presentation

How it All Fits Together



What is the Gateway?



Holmes Gateway

A sophisticated router orchestrating the **submission of new samples** as well as **sending tasks** to Totem and Totem-Dynamic.

01

Go

One statically compiled binary for all platforms

02

HTTP API

Easy to access and integrate into your tools

03

UAC & ACL

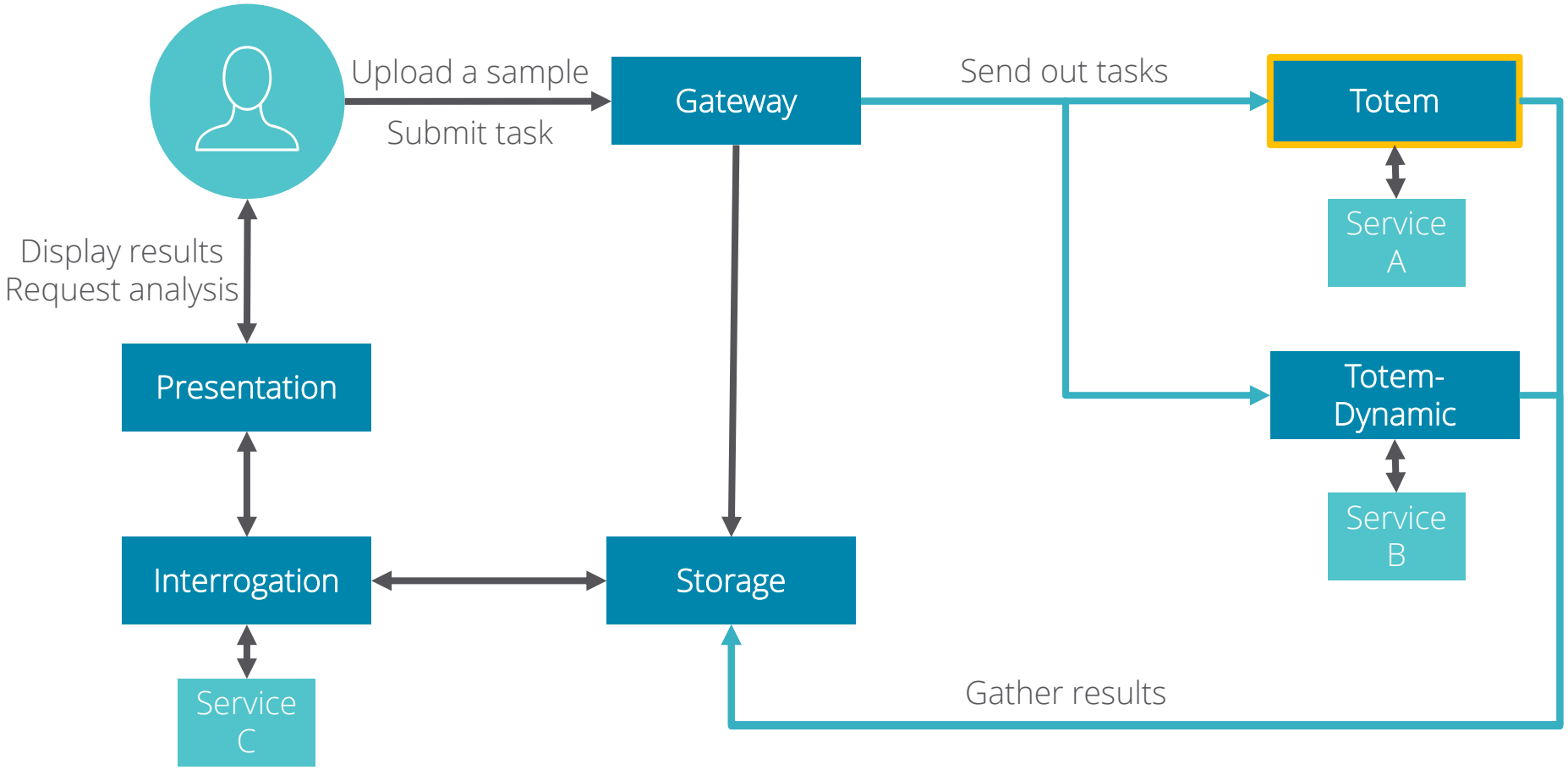
Enforce strict access policies onto all users

04

Sharing

Allow analysis without access to raw data

What is Totem?



Holmes Totem

High performance scheduler performing feature extraction against submitted objects. Optimized for fast Services. i.e. static analysis

01

JVM

Easy to build and deploy on commodity HW

02

Akka

Highly concurrent and memory efficient

03

AMQP

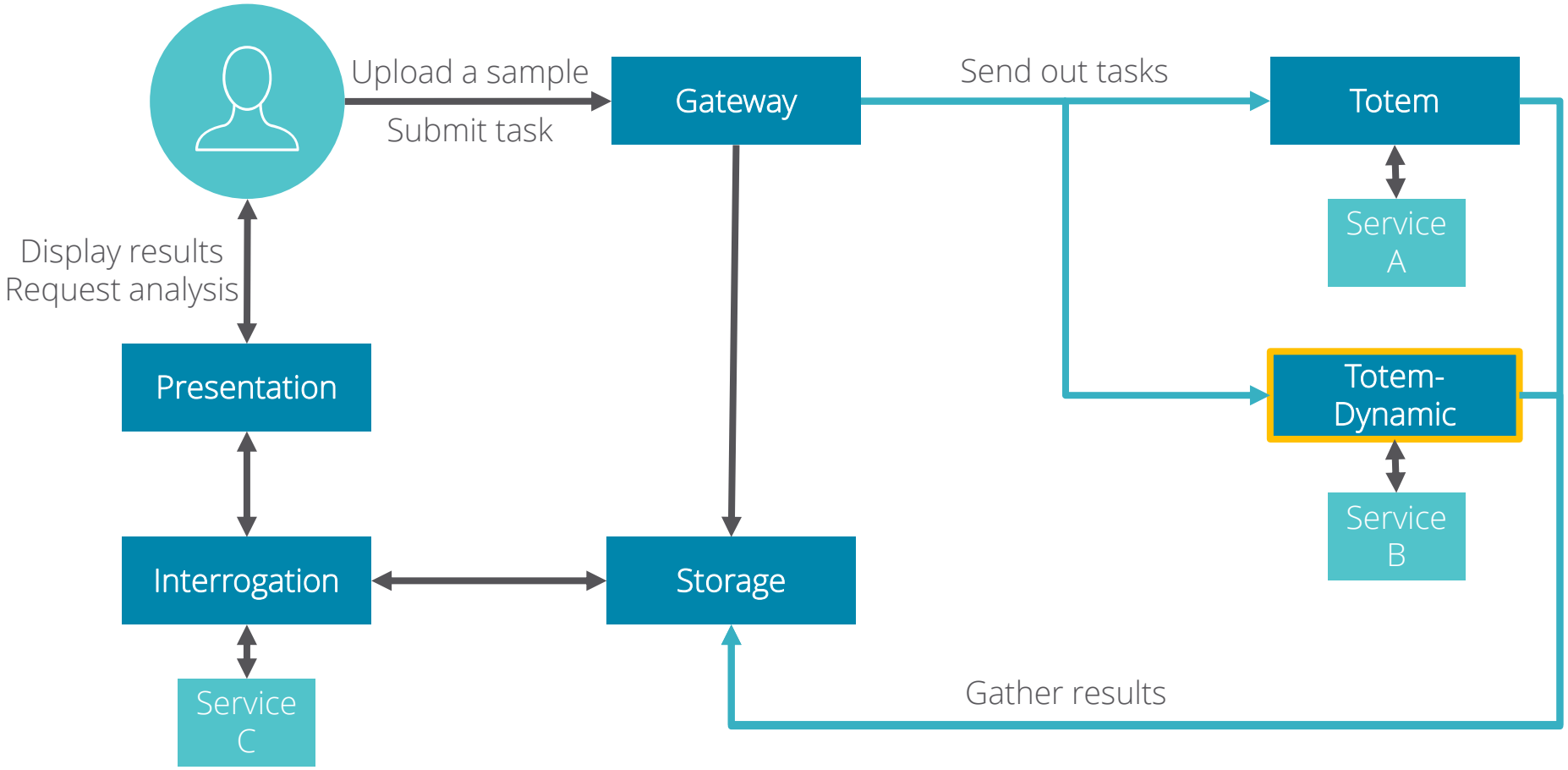
Small footprint sending and receiving

04

Docker

Containerized, easy to setup services

What is Totem-Dynamic?



Holmes Totem-Dynamic

A planner specifically designed for long running analysis and unpredictable third party Services. i.e. Dynamic analysis

01

Go

One statically compiled binary for all platforms

02

FCS

Feed, Check, Submit - a resilient concept

03

AMQP

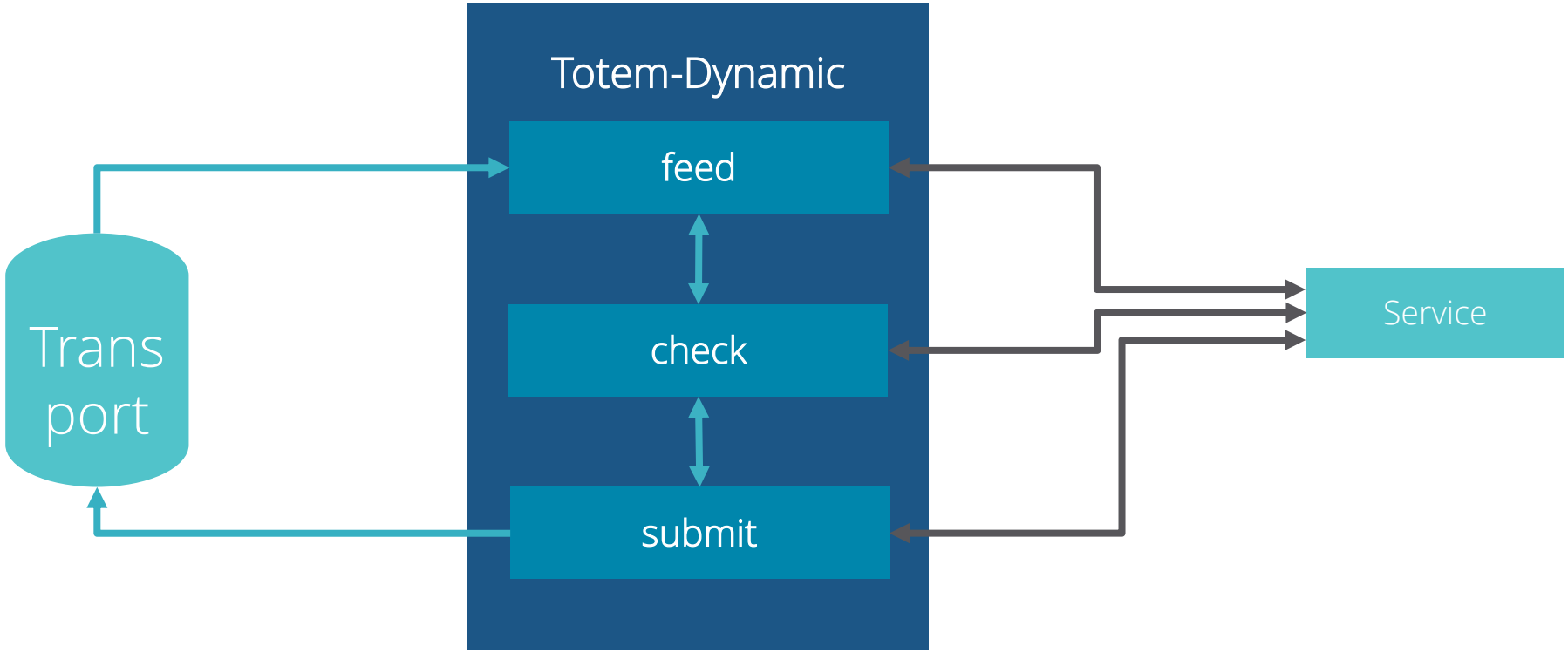
Small footprint sending and receiving

04

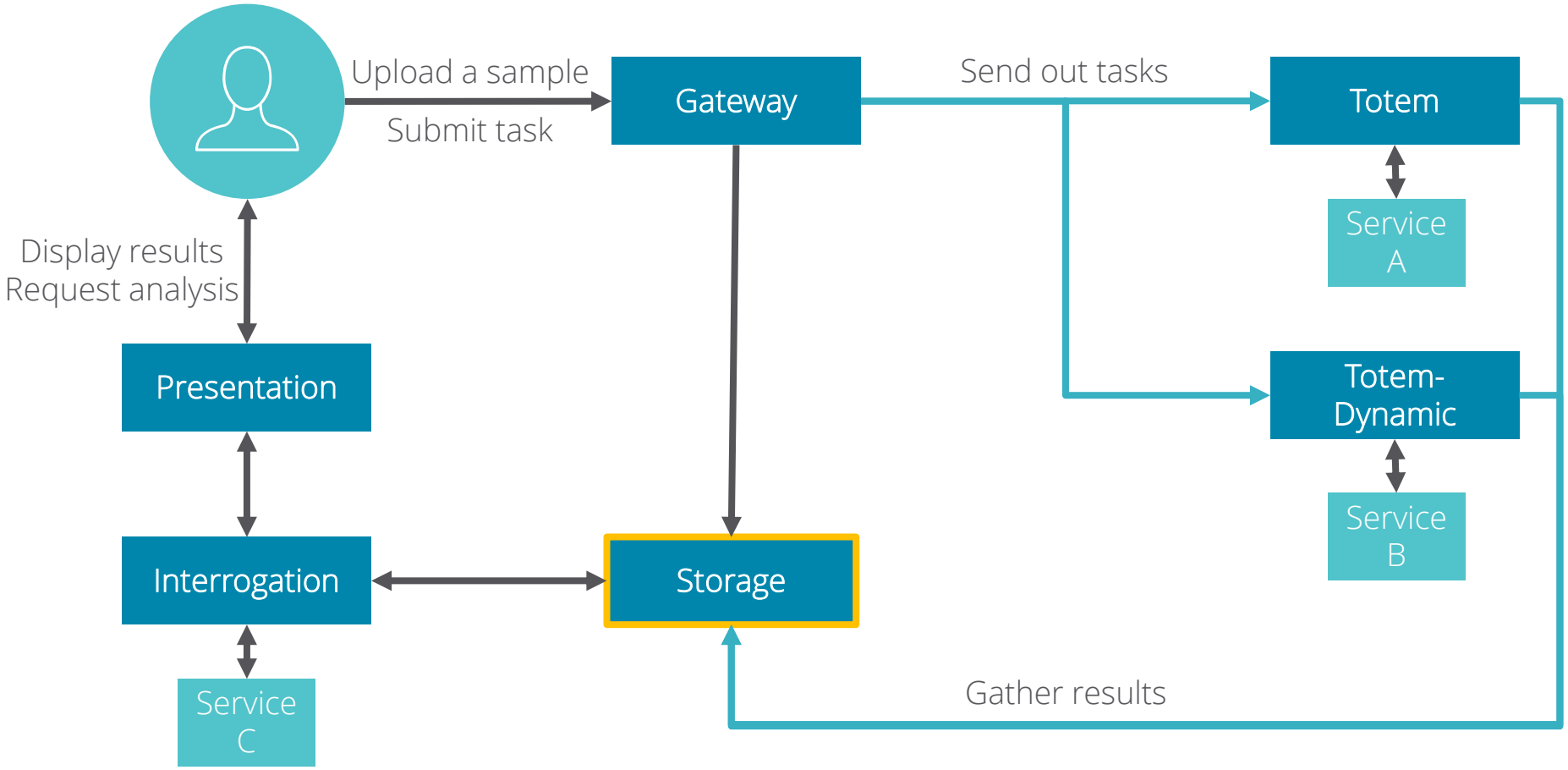
Docker

Containerized, easy to setup services

Totem-Dynamic Inner Workings



What is Storage?



Holmes Storage

A planner designed to manage data and sample storage as well as offering an API for other planners and services to easy and secure interact with the data.

01

Go

One statically compiled binary for all platforms

02

HTTP API

Easy to use and load balance, stateless

03

AMQP

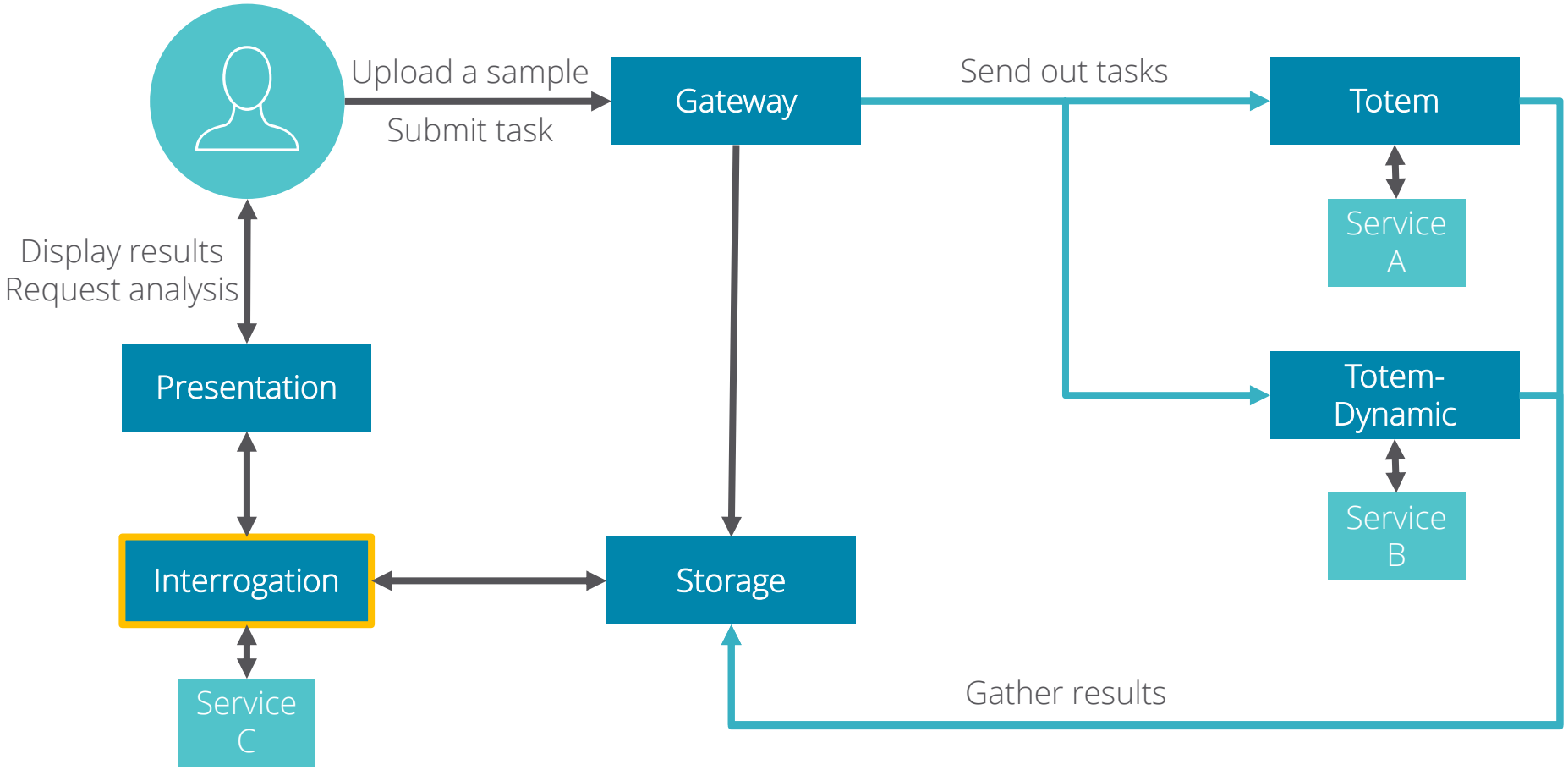
Small footprint receiving of analysis results

04

Versatile

MongoDB, Cassandra, S3, ...

What is Interrogation?



Holmes Interrogation

The Interrogation planner is defined by the SKALD architecture and serves as the focal point for performing analysis and render the data.

01

Go

One statically compiled binary for all platforms

02

Interface

Web Interface for easy of use and visualization

03

ML

Machine Learning Support

04

ALPHA

Very much under development!

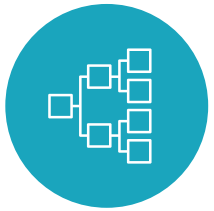
Case Study

Work cycle for the Rich Header

Data 01

Gather and import raw samples into the system [Gateway, Storage]

Information 02



Perform feature extraction and other needed analysis [Totem, Totem-Dynamic]

Knowledge 03

Execute algorithms for statistics and similarity matching [Interrogation (prototype)]

Wisdom 04

Interpret our results, refine our methods, write paper

Conclusion

Key Takeaways

APACHE 2 LICENSE

We developed this system in partnership with our industry partners to support our research and their operations. We have had great success and we hope it helps you as well.

01 **Supports the analytic life cycle**

02 **Scalable, resilient, flexible**

03 **Supports collaboration**

04 **Use the parts you want**

Hold Tight and Pretend There is a Plan

01

Release Interrogation Planner

Interrogation should simplify the execution and scheduling of machine learning and analytic code. We plan to release the code within the next few months

02

Implement advanced sharing model

Version 2 of the sharing model will simplify the exchange of information and sharing of infrastructure

03

Centralize configuration management

We plan to allow the ability to automatically configure Planners and Services through Storage

04

Monitoring of Planners and Services

We want to extend the monitoring capabilities of the system and output the information through an API and the website

Love Your Help

01

Write Services

We love receiving new extraction methods for Totem and Totem-dynamic. As we build out Interrogation, new ML methods will become critical

02

Add Elastic

We supply MV and 2nd indexing for Cassandra. Extending Storage to support Elastic across the raw information would be much appreciated

03

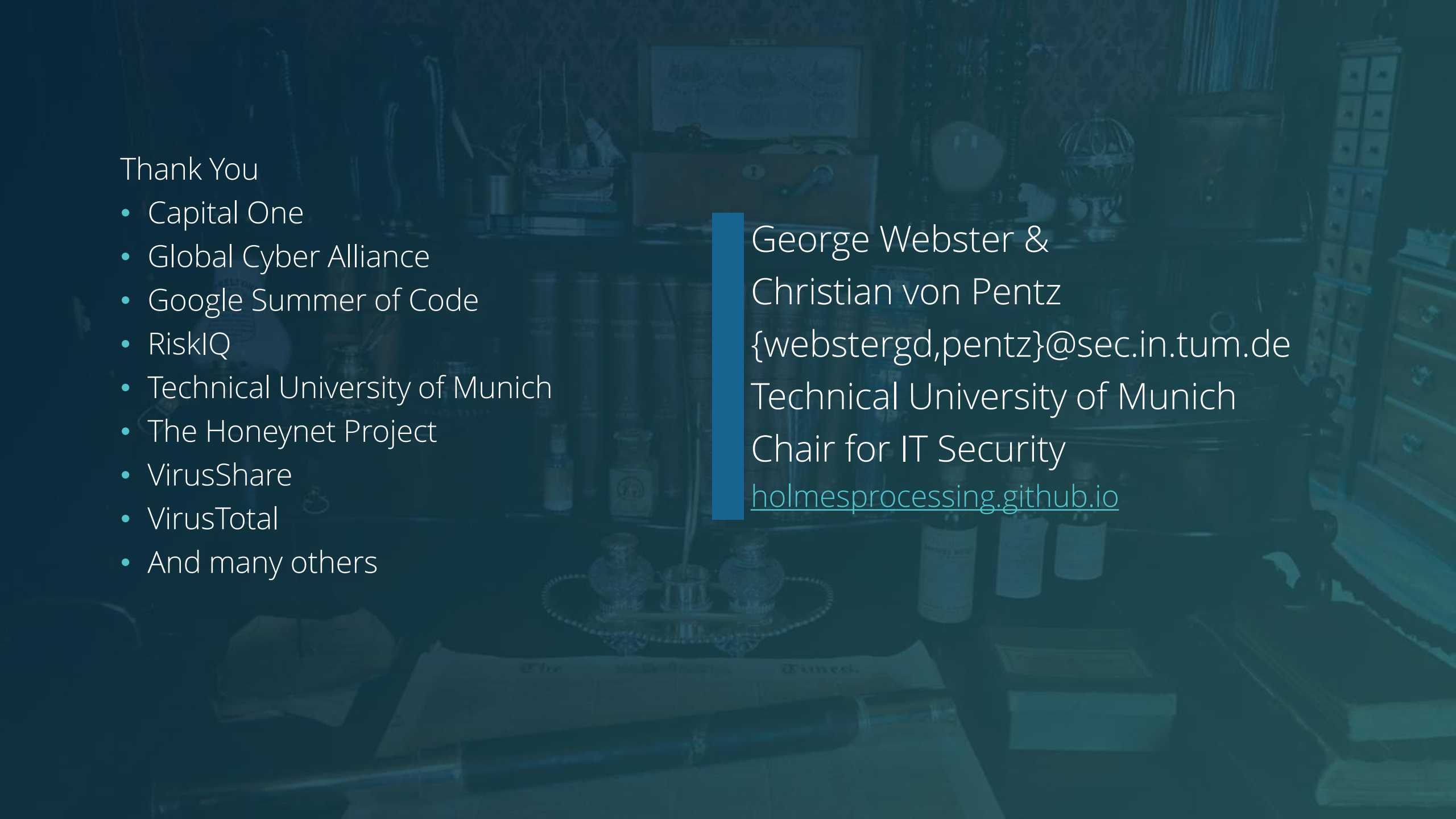
Improvements

We do this work on the side. Anything from bug fixes, to general improvements, to documentation, to pretty artwork would be wonderful

04

Samples

We need samples and love anything that has labels. It is a huge benefit to our research and on many different fronts



Thank You

- Capital One
- Global Cyber Alliance
- Google Summer of Code
- RiskIQ
- Technical University of Munich
- The Honeynet Project
- VirusShare
- VirusTotal
- And many others

George Webster &
Christian von Pentz
{webstergd,pentz}@sec.in.tum.de
Technical University of Munich
Chair for IT Security
holmesprocessing.github.io