

# ML4PATENTS.COM WEBINAR SERIES

EPISODE 2

## YOU HAVE MACHINE LEARNING QUESTIONS, WE HAVE ANSWERS!

Wednesday, August 25 | 11:00 am CST

Featured speakers:



Nigel Swycher  
CEO of Cipher



Steve Harris  
CTO of Cipher



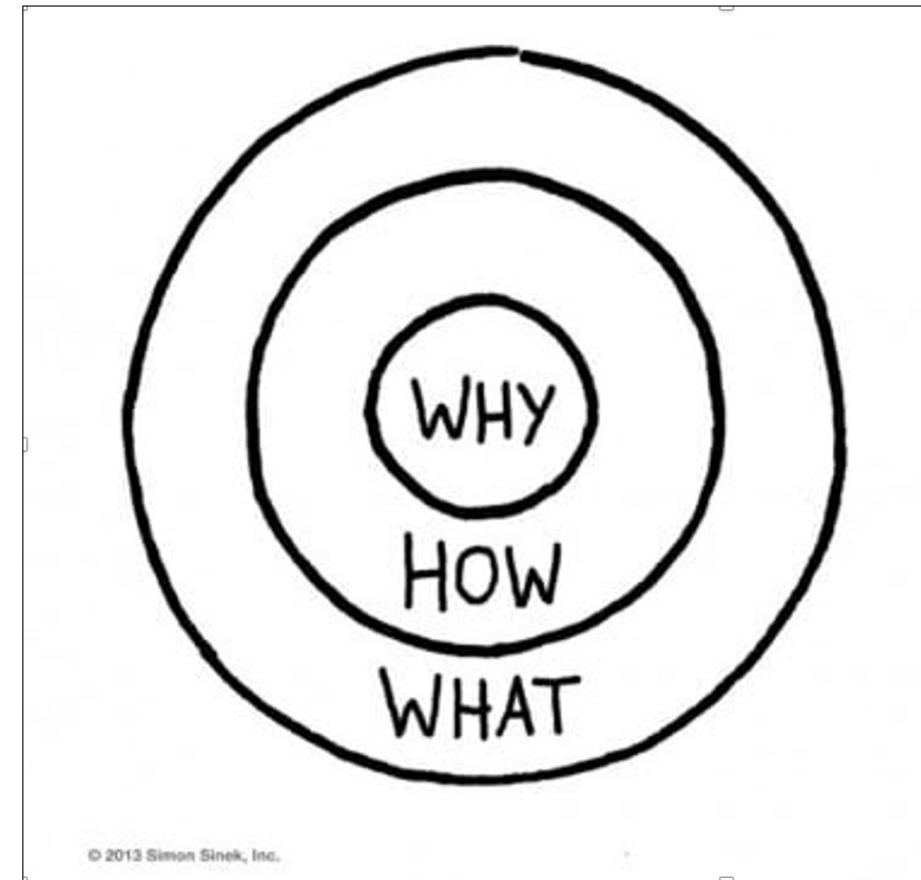
Tony Trippé  
Patinformatics

# Start with Why

## Linear Law of Patent Analysis

- Create a toolkit of analysis tools
- Understand the business need and the need behind the need
- The need drives the question
- The question drives the data
- The data drives the tool

<https://patinformatics.com/the-linear-law-of-patent-analysis-revisited/>





## Unleashing the strategic value of patents

*Cipher enables patent owners to make rational decisions by providing strategic patent intelligence, powered by machine learning*

*“Cipher provides the data that we need, almost magically, using a lens that aligns with our company’s view of the world.”*

Head of Patent Development



# Strategic Patent Intelligence is different



## Saves time & money

Conventional Boolean search requires time consuming manual cleaning. Cipher can read **61 million** in an hour



## Automated patent to technology mapping

Patent classification system (CPC codes) cannot be used to map patents to company taxonomies



## Objective, repeatable results

Automation provides analytics on demand, and achieves consistency



## Strategic insights

Cipher delivers evidence for strategic decision-making in response to increasing pressure to communicate patent strategy

# The importance of patent strategy

Corporate benchmarking is critical for patent strategy, competitive intelligence, budget management and pruning

“It’s critical to have benchmarking data because you don’t want to operate in a vacuum.”

Andreas Iwerback

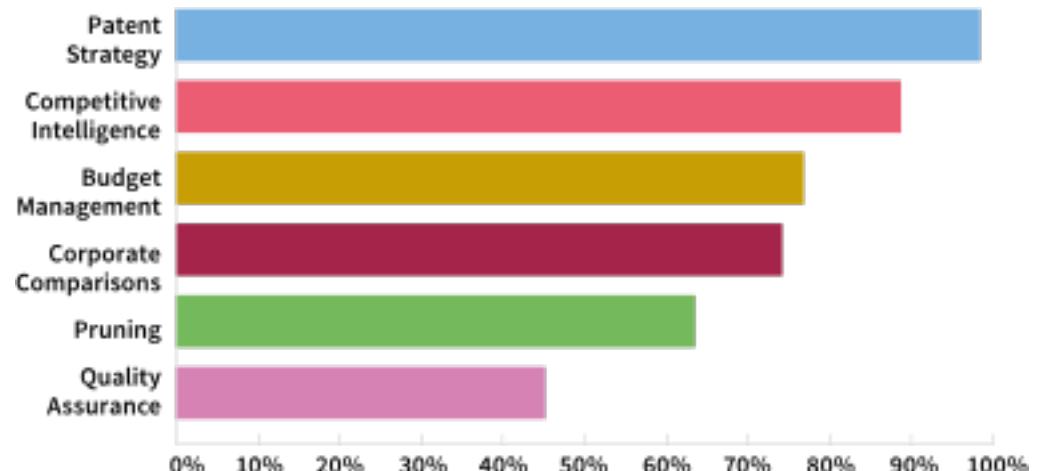
*Head of Technology & IP Intelligence, Husqvarna*

“Benchmarking supports the communication of our approach, our strategy and why we are building the portfolio in the way that we are.”

Gilbert Wong

*Associate General Counsel for Patents, Facebook*

Why do you benchmark your patent portfolio?



*Source: Cipher IAM Benchmarking Survey 2020*

Refer to Dispelling the Benchmarking Myth

*How Machine Learning increases efficiency and reduces cost, March 2021*

# Technology level analysis

There is a need for a granular and consistent approach

“Benchmarking at the technology level is the only way to compare ‘apples with apples’ and to spot meaningful trends.”

*IP expert at a major European Semiconductor company*

What granularity do you benchmark?

Technology grouping | 75%

Your whole portfolio | 67%

Business unit or division | 41%

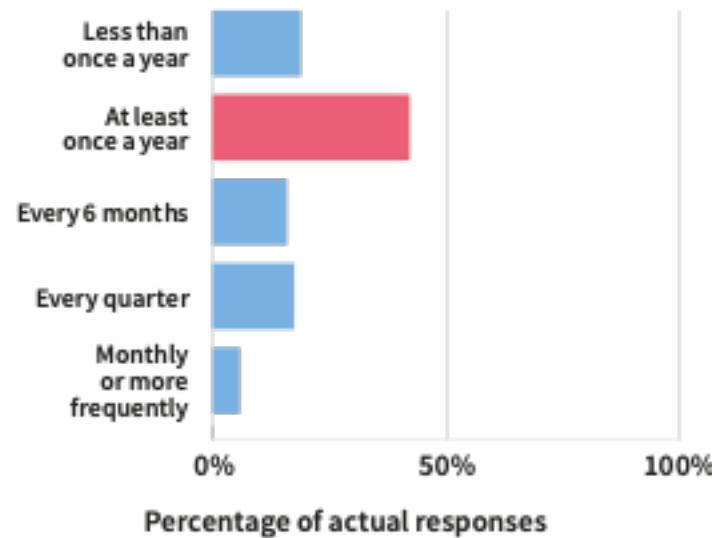
Individual patents | 22%

Source: Cipher IAM Benchmarking Survey 2020

# Current data challenges

The size of the data challenge is not to be underestimated

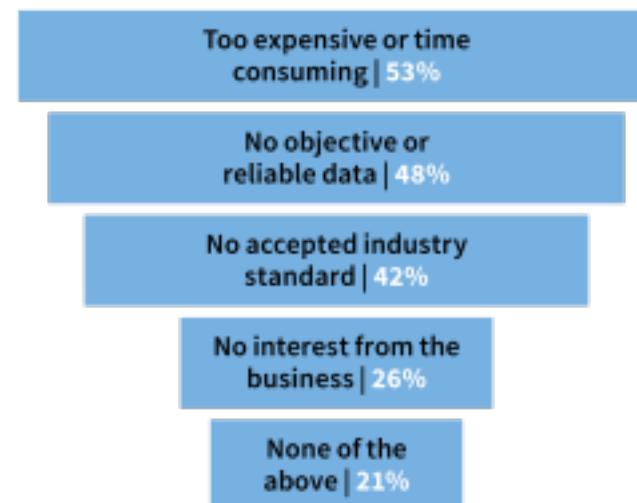
## How often do you benchmark?



Source: Cipher IAM Benchmarking Survey 2020

**88%** of companies are benchmarking at least once per year - structured data and process is needed.

## What are your challenges?



Source: Cipher IAM Benchmarking Survey 2020

**Time** and **money** are challenges - new approaches required.  
Data **reliability** and **consistency** are a significant challenge for many.

# The qubit generation test

Construction and evaluation of gold standards for patent classification

Steve Harris<sup>1</sup>, Anthony Trippe<sup>1</sup>, David Challis<sup>1</sup>, Nigel Swycher<sup>1</sup>

<sup>a</sup>[info@cipher.ai](mailto:info@cipher.ai), Aistemos Ltd, 39-41 Charing Cross Road, WC2H 0AR, London, UK

<sup>b</sup>[info@patinformatics.com](mailto:info@patinformatics.com), Patinformatics LLC, 565 Metro Place S. Suite 3033, Dublin, OH 43017, USA

# World Patent Information

THE JOURNAL FOR INTELLECTUAL  
PROPERTY INFORMATION AND ITS RETRIEVAL,  
DOCUMENTATION, CLASSIFICATION, SEARCH,  
ANALYSIS AND IP MANAGEMENT

## Abstract

This article discusses options for evaluation of patent and/or patent family classification algorithms by means of “gold standards”. It covers the creation criteria, and desirable attributes of evaluation mechanisms, then proposes an example gold standard, and discusses the results of applying the evaluation mechanism against the proposed gold standard and an existing commercial implementation.

**Keywords:** Patent Classification, Evaluation, Artificial Intelligence, Information Retrieval, Deep Learning, Gold Standard

## 1. Introduction

There are a number of problems in the strategic patent decision making and portfolio management domain where artificial intelligence techniques can be applied. One of the more common is that of mapping patent assets to technologies, for example to perform patent landscaping, or for reporting on the contents of your own, or competitor portfolios. This is also one of the hardest tasks to perform mechanically, and has been identified as a source of friction in strategic patent decision making[1].

Conventional “mandrolic”, or semi-automated solutions typically revolve around performing a boolean search over the assets to discover a superset of the assets to be identified, then manually reviewing returned results to determine if each individual asset falls into the desired class.

There are a number of compromises involved in this approach – predominantly related to the time taken to perform a thorough review of the technology domain, or the cost of outsourcing this work to external experts.

In addition there is also the issue of inconsistency of results

rithms in a neutral way is extremely difficult, even for experts in the field, which makes it very difficult to answer questions such as “which operations are viable to automate?”, and “how does the accuracy of AI algorithms compare to manual work?”.

This article proposes an approach for generating gold standards for machine classification of patents, and presents one such example. It then describes a methodology to test against that gold standard, and presents the results of evaluation of a commercially available system against it.

In the following text, we use the binary classification convention of denoting the data labelled as positive (examples of in-scope patents) with  $T_{\oplus}$ , and those labelled as negative (counter-examples) with  $T_{\ominus}$ , where  $T$  denotes the training set,  $G$  the gold standard as a whole, and so on.

We will also describe the processes in set notation for brevity, though restrict the use to just  $\cup$  (union),  $\cap$  (intersection),  $\setminus$  (set difference), and  $|X|$  (set cardinality).

## 2. Prior work

### 2.1. Existing gold standards

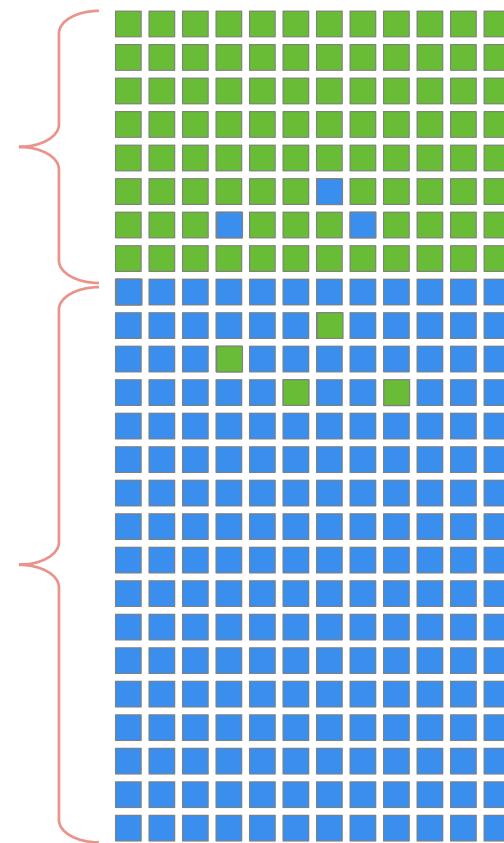
# Measuring accuracy – returned results

## Predicted Positives

Identified by classifier (96)

## Predicted Negatives

Not Identified by classifier (204)



93 true positives  
3 false positives

200 true negatives  
4 false negatives

## Precision

“Of the results found, what proportion are positive”

$$\frac{TP}{TP + FP} = 0.969$$

## Recall

“Of all the positives out there, what proportion were found”

$$\frac{TP}{TP + FN} = 0.959$$

**Can be combined to one number**  
 $F_1 = 0.964$

# Some common misunderstandings

## “Building test sets is easy”

You can't use any old data as a test set – it has to be very carefully checked, and representative.

Even tiny errors make the results meaningless.

## “Testing is hard”

If you want to know how well a classifier does for your use case, just test it how you would use it for real.

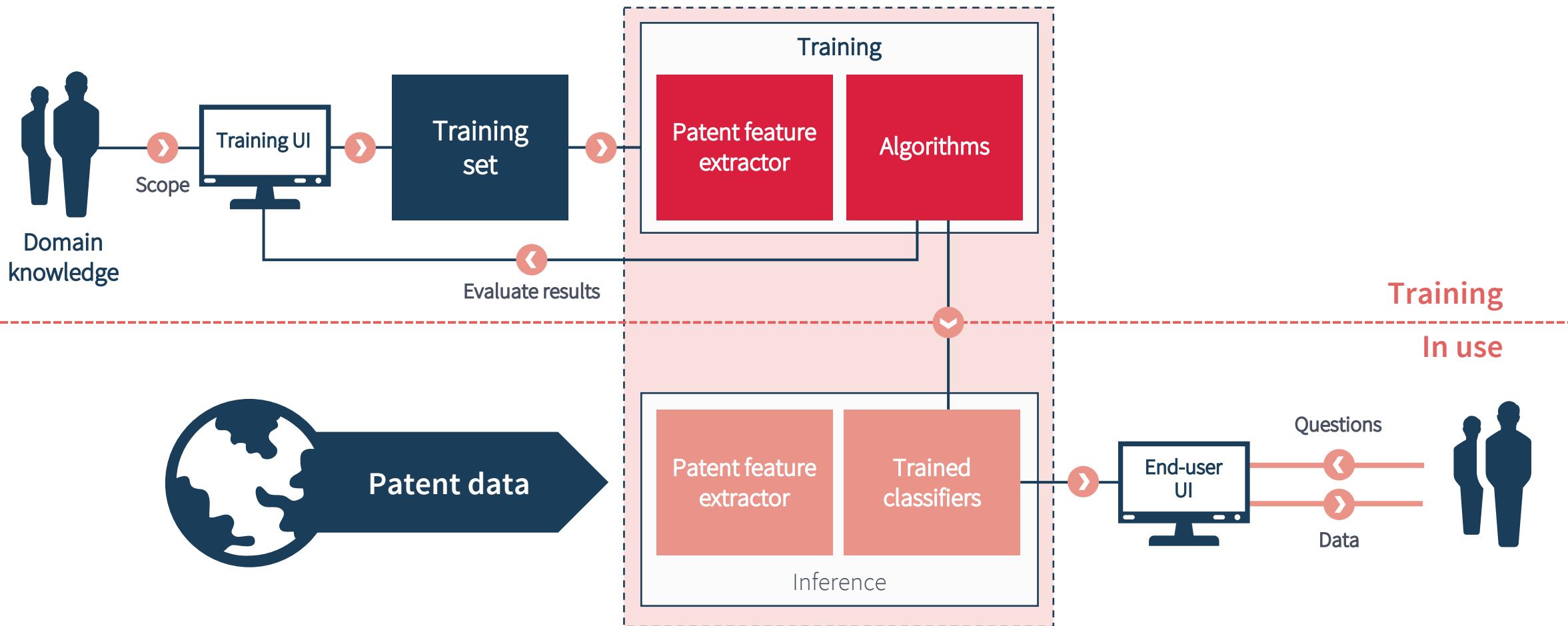
No synthetic test will ever be as useful.

## “This classifier has 0.9 precision”

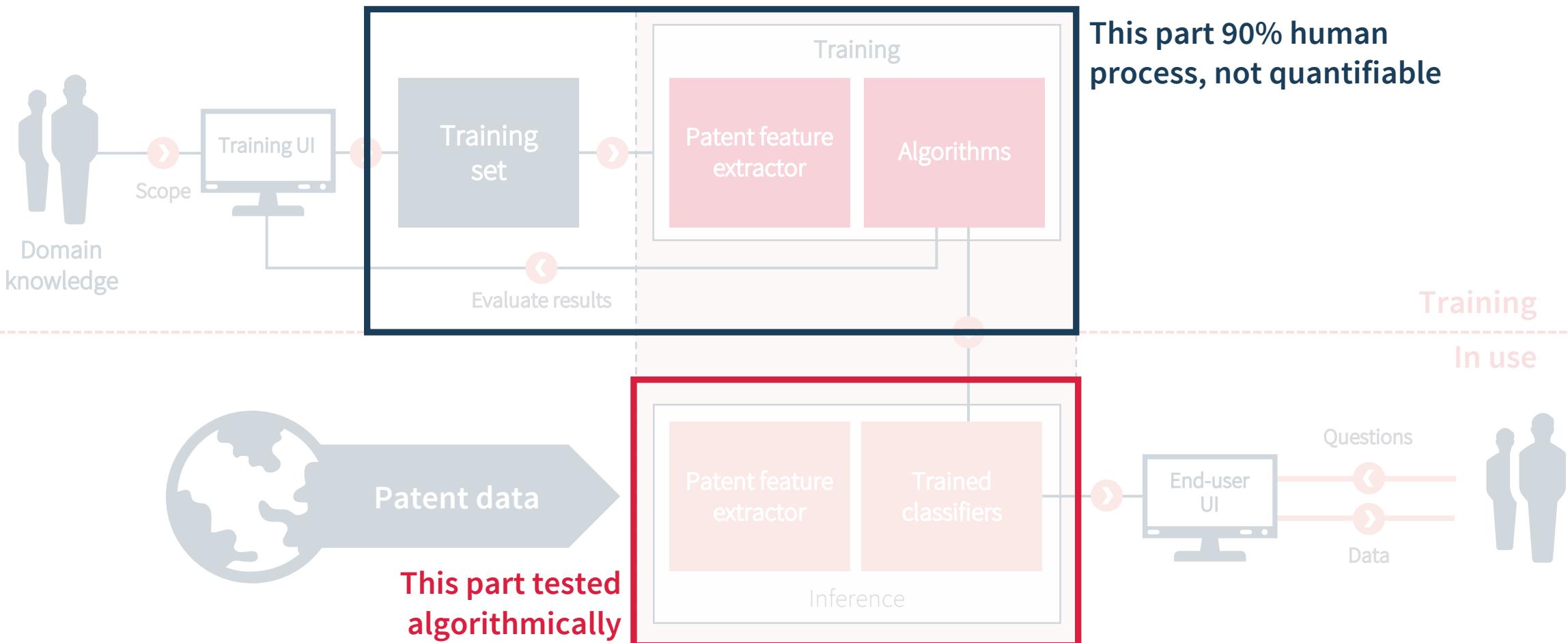
Classifiers do not “have” a precision and recall.

Precision and recall only mean anything with reference to a labelled test set.

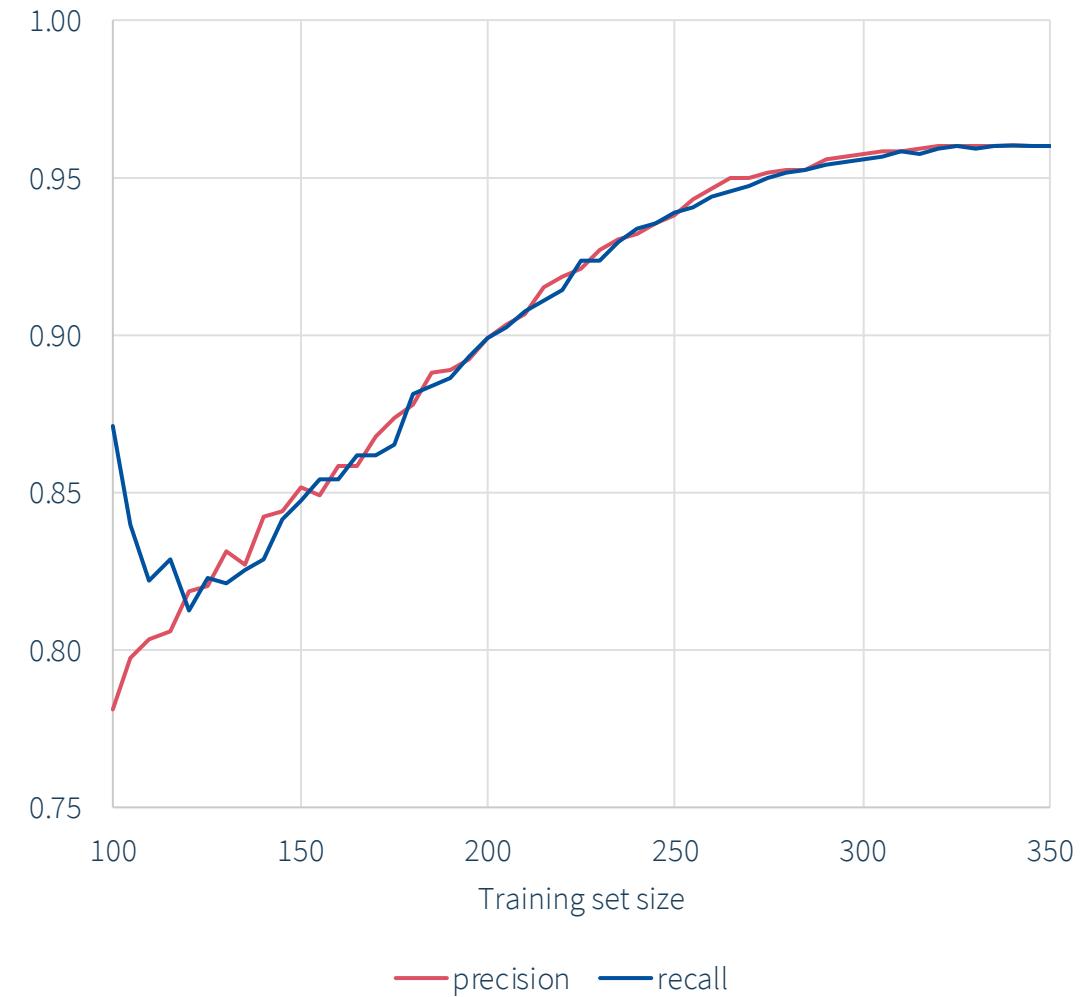
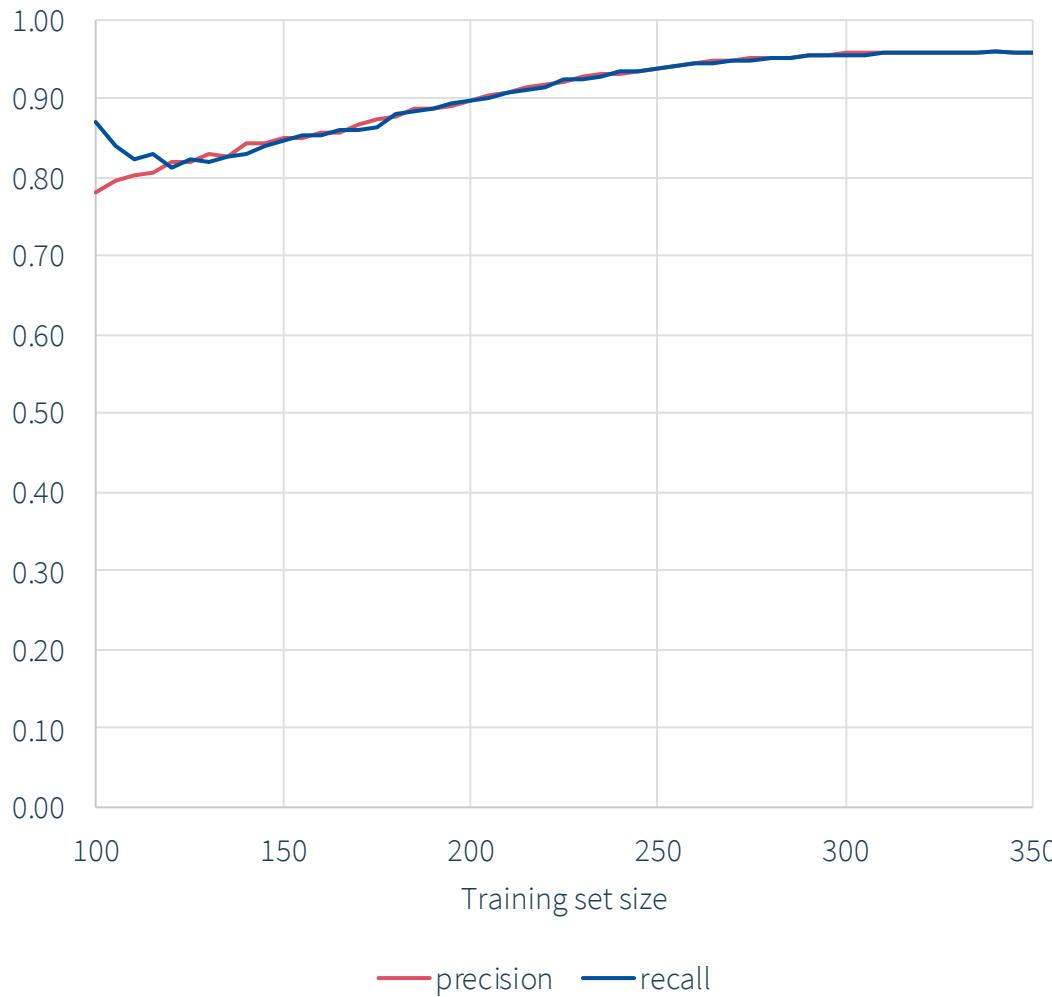
# Training and using a classifier



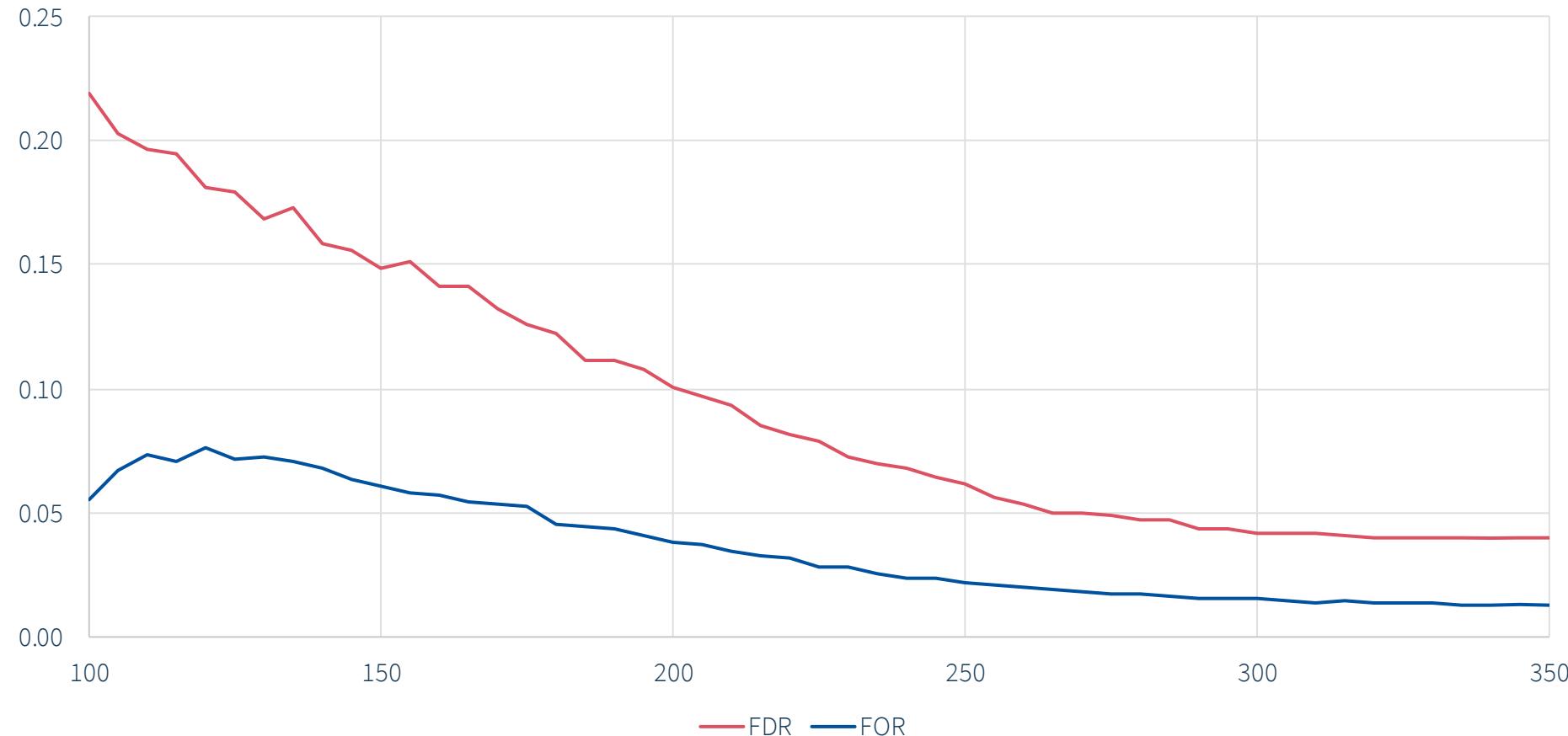
# What can you measure?



# Results – precision and recall

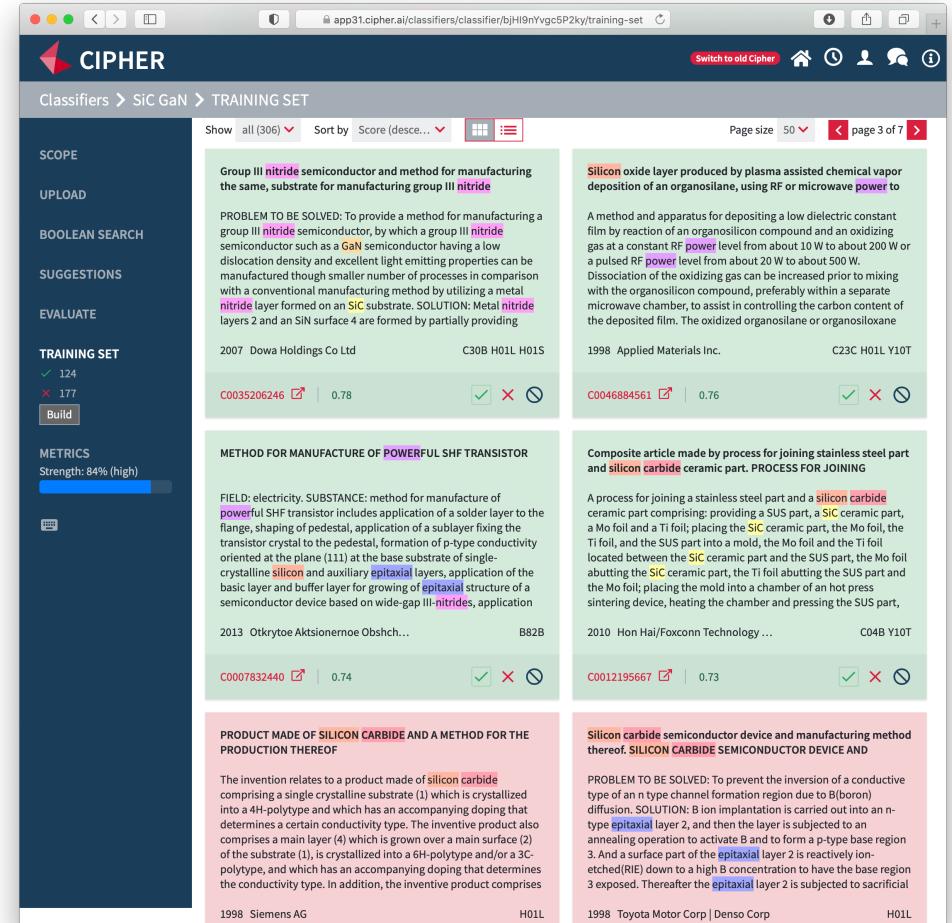


# False discovery and omission rates



# What does training involve?

- Custom user interface for quickly training and evaluating classifiers.
- Allows you to upload known examples.
- Run evaluations and correct errors.
- See related patents that would be helpful for training.
- Train and cross-validate classifiers.



The screenshot shows the CIPHER AI web interface for patent classification. The main title is "Classifiers > SIC GaN > TRAINING SET". On the left, there's a sidebar with sections for SCOPE, UPLOAD, BOOLEAN SEARCH, SUGGESTIONS, and EVALUATE. Under EVALUATE, there's a "TRAINING SET" section showing 124 successful uploads and 177 failed ones, with a "Build" button. Below that is a "METRICS" section with a strength of 84% (high). The main area displays a grid of patent cards. The first card is for a "Group III nitride semiconductor and method for manufacturing the same, substrate for manufacturing group III nitride" (2007 Dowa Holdings Co Ltd, C30B H01L H01S, score 0.78). The second card is for a "Silicon oxide layer produced by plasma assisted chemical vapor deposition of an organosilane, using RF or microwave power to" (1998 Applied Materials Inc., C23C H01L Y10T, score 0.76). The third card is for a "METHOD FOR MANUFACTURE OF POWERFUL SHF TRANSISTOR" (2013 Otkrytie Aktzionernoe Obsch..., B82B, score 0.74). The fourth card is for a "Composite article made by process for joining stainless steel part and silicon carbide ceramic part. PROCESS FOR JOINING" (2010 Hon Hai/Foxconn Technology ..., C04B Y10T, score 0.73). The fifth card is for a "PRODUCT MADE OF SILICON CARBIDE AND A METHOD FOR THE PRODUCTION THEREOF" (1998 Siemens AG, H01L, score 0.73). The sixth card is for a "Silicon carbide semiconductor device and manufacturing method thereof. SILICON CARBIDE SEMICONDUCTOR DEVICE AND" (1998 Toyota Motor Corp | Denso Corp, H01L, score 0.73).

# What data does the training process use?

## Data that's used

**Title and abstract** – general background information

**Claims** – obvious why, uses a different embedding to title and abstract

**Citations** – information about topics that related to the patent

**CPC class codes** – component technologies, but it's less useful than you might think (see paper)

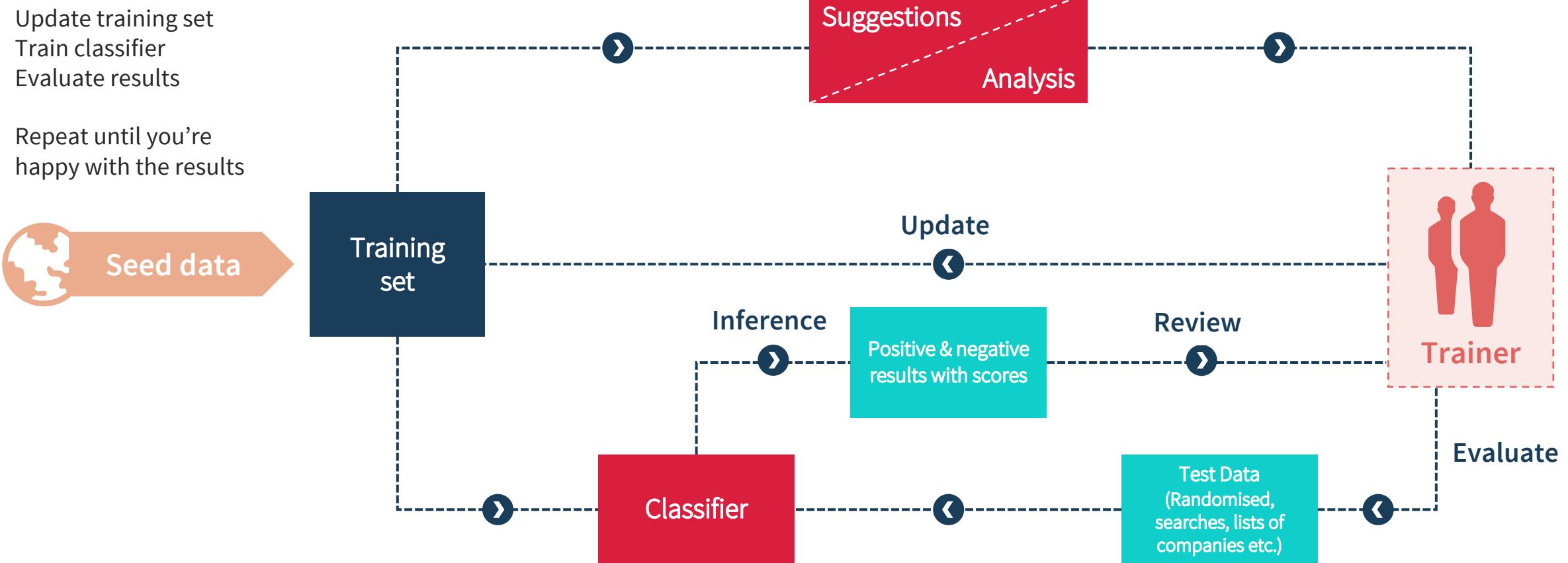
## Data that's explicitly not used

**Owner** – we want to avoid any biases for/against specific companies

**Inventor names** – inventors correlate to companies

**Non-English text** – we machine translate to English where necessary

# Cyclical training process



# Custom taxonomies and automated classification

**AUTOMOTIVE CLASSIFIERS**

[expand] [collapse]

- ✓  Automotive
  - >  OVERVIEW
  - >  Autonomous systems
  - >  Cabin systems
  - >  Chassis and body
  - >  Complete vehicle systems
- ✓  Drivetrain
  - >  Automatic transmission
    - Automatic gearbox (i)
    - Continuously variable transmissions (CVT) (i)
    - Dual-clutch transmissions (i)
    - E-clutch (i)
    - Semi-automatic gearbox (i)
    - Torque converters (i)
  - >  Hybrid
  - >  Manual transmission

✓  ADAS components

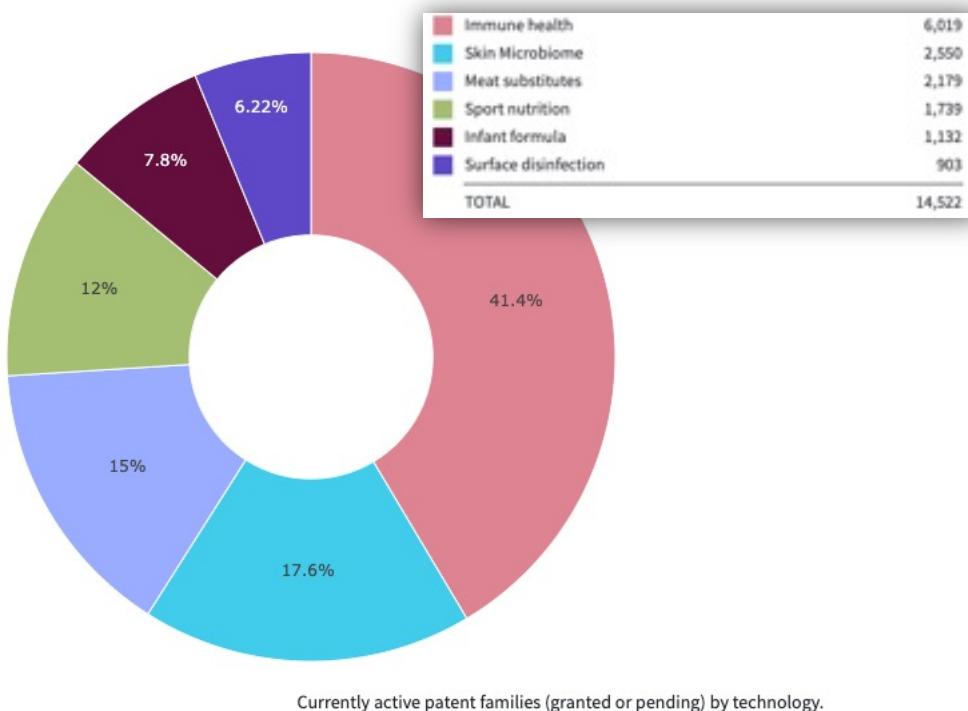
- Infrared sensors (i)
- Lidar sensors (i)
- Millimeter wave radar (MMW) (i)
- Night vision (i)
- Panoramic and overhead cameras (i)
- Radar sensors (i)
- Rear and parking cameras (i)
- Ultrasonic sensors (i)

Lidar/ladar (Light Detection And Ranging/Light Imaging, Detection, And Ranging) sensors are a type of electronic instrument that use light pulses as a medium to survey the environment. Lidar sensor systems contain transmitters, receivers, and image processing capabilities, for the detection, creation and visual illustration of objects, such as pedestrians or obstacles, in the exterior environment of a vehicle.

**Taxonomy design and scope are all activities exclusively within the human domain.**

# Demo | classification without limits

## 1. Global Landscapes



## 2. Selection of technologies against specific companies

Lidar sensors
Radar sensors
Panoramic and overhead cameras
Rear and parking cameras
Ultrasonic sensors
Millimeter wave radar (MMW)
Night vision
Infrared sensors

	Toyota	Volkswagen	GM	Hyundai	Ford	Next 47	TOTAL
164	73	188	109	167	505	1,206	
126	160	98	41	21	289	735	
82	58	32	58	19	441	690	
69	36	23	68	28	433	657	
52	36	1	37	13	122	261	
45	0	1	1	0	120	167	
12	13	2	15	9	81	132	
4	3	2	6	6	37	58	
<b>TOTAL</b>	<b>554</b>	<b>379</b>	<b>347</b>	<b>335</b>	<b>263</b>	<b>2,028</b>	<b>3,906</b>

Currently active patent families (granted or pending) by organisation and technology.

<https://app.cipher.ai/report/1870e7083c/g/U3oWO/d/cljAE>

<https://app.cipher.ai/report/81a8b3128b/g/selected/d/P11hY>



# CIPHER