CIPHER

Patinformatics, LLC®
Patent Landscape Reports

Measuring the accuracy of AI for classifying patents – *what's the Gold Standard?*

Steve Harris, CTO, Cipher

Tony Trippe, Managing Director, Patinformatics LLC

Moderator: Nigel Swycher, CEO, Cipher

Patinformatics, LLC®
Patent Landscape Reports

CIPHER

Your speakers

Steve Harris

CTO, Cipher

Tony Trippe

Managing Director, Patinformatics LLC

Moderated by

Nigel Swycher, CEO, Cipher

> If the AI search tool worked well, we wouldn't need it. It would automatically classify every patent and converge on a single source of truth …
> I also wouldn't laugh when I got the results.

Matthew Wahlrab, CEO, rapid alpha

Construction and evaluation of Gold Standards for patent classification

Construction and evaluation of gold standards for patent classification

Steve Harris[1], Anthony Trippe[1], David Challis[1], Nigel Swycher[1]

[a]info@cipher.ai, *Aistemos Ltd, 39-41 Charing Cross Road, WC2H 0AR, London, UK*
[b]info@patinformatics.com, *Patinformatics LLC, 565 Metro Place S. Suite 3033, Dublin, OH 43017, USA*

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.

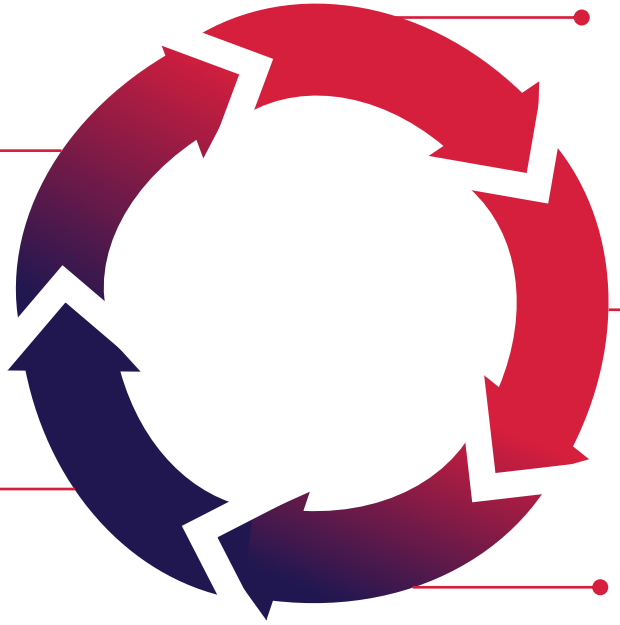
In the following text, we use the binary classification convention of denoting the data labelled as positive (examples of in-scope patents) with $T_{\oplus}$, and those labelled as negative (counter-examples) with $T_{\ominus}$, where T denotes the training set, G the gold standard as a whole, and so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

Advances in AI across the entire patent lifecycle



FTO/novelty | Drafting and optimising

Accelerating the time to reinvent | **Inventing**

Examination | Identifying prior art, beating the system

Prior art, both patents and NPL | **Invalidation**

Exploitation | Litigating and licensing incl. quality and valuation

References

AI-assisted patent prior art searching, UKIPO, April 2020

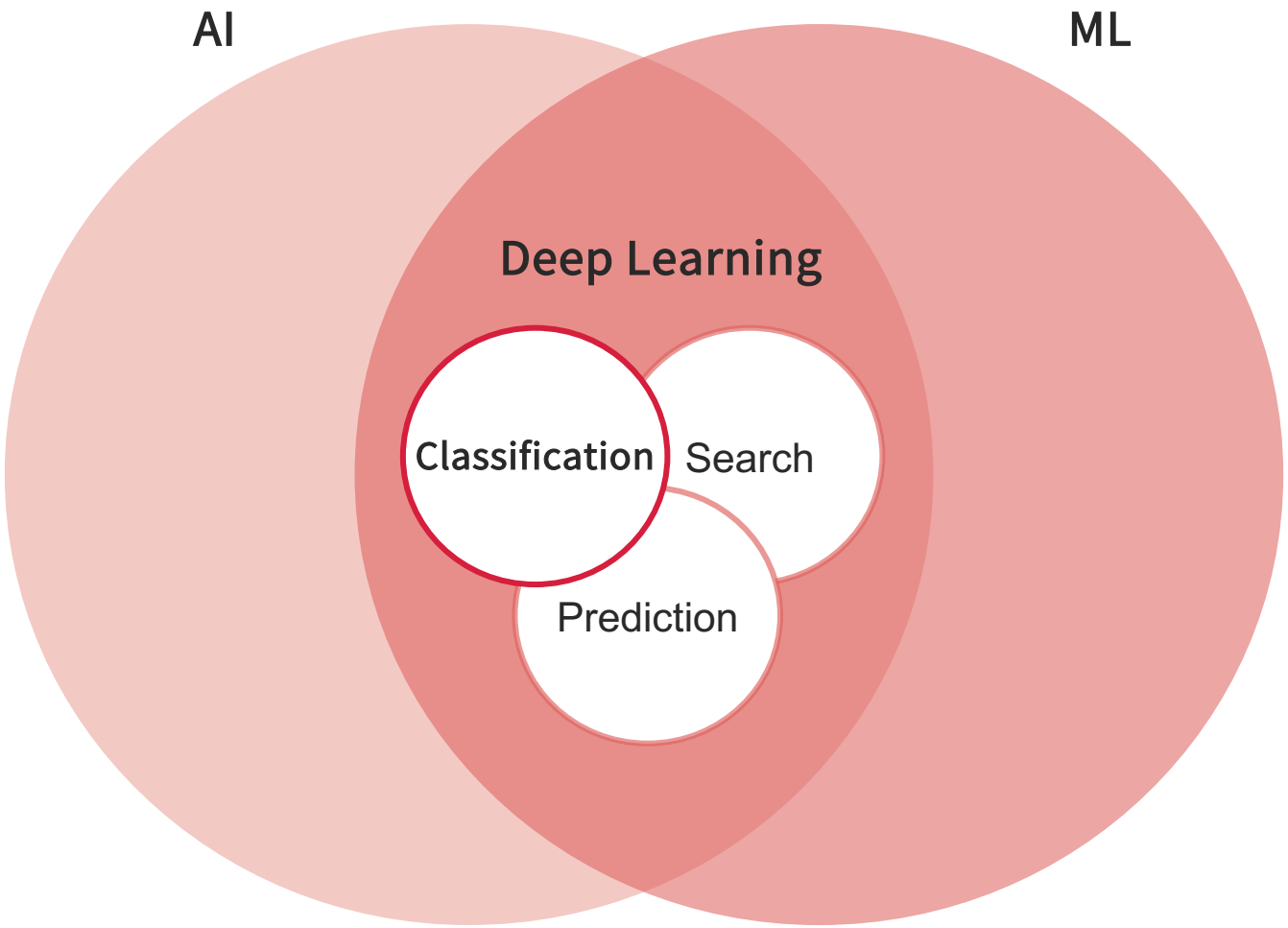
IP Automation - What's Here Today, Not Years Away, September 2019

Meeting of Intellectual Property Offices (IPOs) On ICT Strategies and Artificial Intelligence (AI) for IP Administration, Geneva, May 23 to 25, 2018

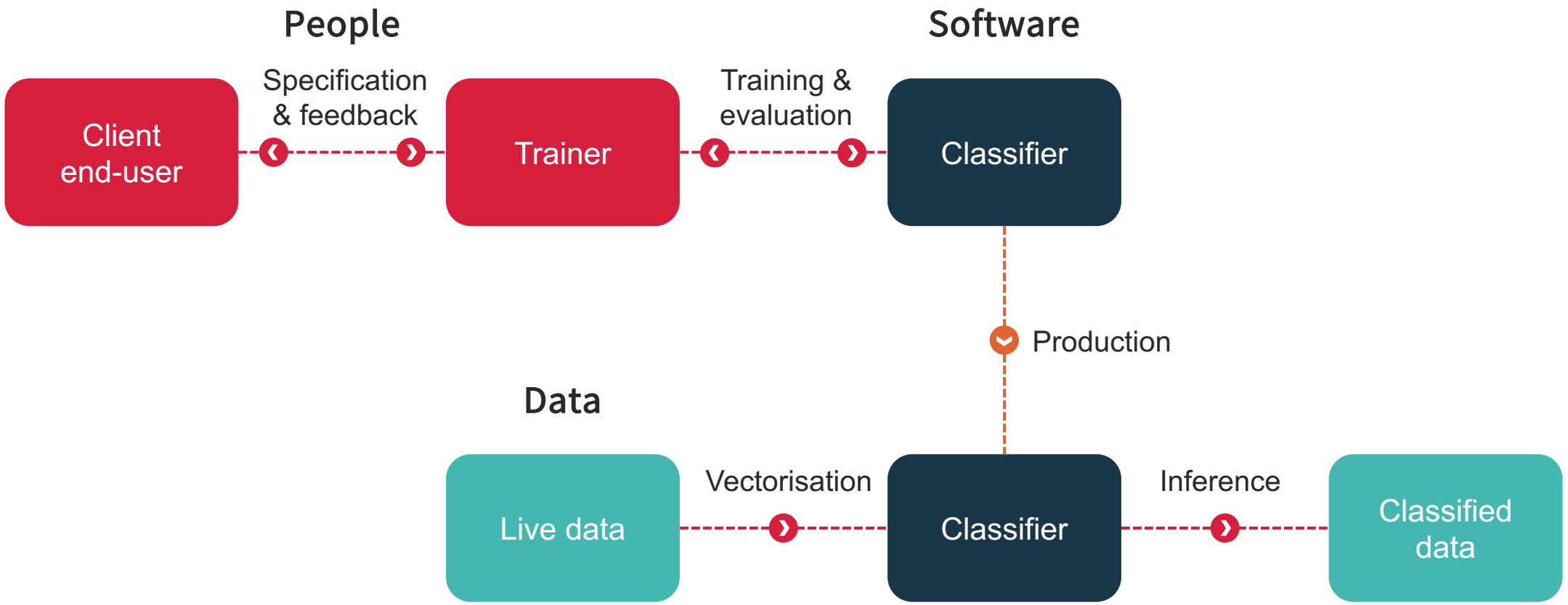
Why patents?


IP strategists lack good tools
Difficult and expensive to get data
Labour intensive
Lots of difficult, repetitive tasks in the industry
Technology cost
ML technology had reached a practical level

What are AI and ML?



Classifiers – investment and return



Classification's strengths and weaknesses

Upfront investment makes sense for tasks that are repetitive, e.g.

- Huge numbers of patents (thousands, millions)
- Tasks that need to be repeated often
- Large numbers of technologies at once
- Consistent results desired

Bad Uses

FTO search, prior art search

- Looking for small numbers of patents (or NPL)
- Novel topics, not repeated often
- Only of interest for a short period of time

Good Uses

Landscaping, gathering strategic data, asset tagging, analysing competitor portfolios

- Need for repeatable results
- Lots of patents need to be studied
- Many technologies at once

CIPHER

AI / ML Mythbusting

Not ‘self-taught’

Cannot “read”

Look at data differently

Very fast

What is a Gold Standard collection?

According to Wikipedia

- In medicine and statistics, **a gold standard test** is usually *the diagnostic test or benchmark that is the best available under reasonable conditions*. Other times, a gold standard is *the most accurate test possible without restrictions.*

- A hypothetical ideal "gold standard" test has a sensitivity of 100% with respect to the presence of the disease (it identifies all individuals with a well defined disease process; it does not have any false-negative results) and a specificity of 100% (it does not falsely identify someone with a condition that does not have the condition; it does not have any false-positive results).

What is a Gold Standard collection?

How does this compare to a Ground Truth?

According to Wikipedia

- In machine learning, the term "**ground truth**" *refers to the accuracy of the training set's classification for supervised learning techniques*. This is used in statistical models to prove or disprove research hypotheses. The term "ground truthing" refers to the process of gathering the proper objective (provable) data for this test.
  - Bayesian spam filtering is a common example of supervised learning. In this system, *the algorithm is manually taught the differences between spam and non-spam*. This depends on the ground truth of the messages used to train the algorithm – inaccuracies in the ground truth will correlate to inaccuracies in the resulting spam/non-spam verdicts.
- The term *ground truth* refers to the *underlying absolute state of information; the gold standard strives to represent the ground truth as closely as possible*. While *the gold standard is a best effort to obtain the truth*, ground truth is typically collected by direct observations. In machine learning and information retrieval, "ground truth" is the preferred term even when classifications may be imperfect; *the gold standard is assumed to be the ground truth.*

Why do we need Gold Standards/Ground Truths?

Mechanisms for achieving patent information retrieval

- Keywords and Boolean Logic
- Classification codes
- Citations
  - There is a visible and verifiable cause and effect with these methods

- Machine learning techniques
  - Most methods provide output without a visible path on a query-by-query basis on how the results were generated by the system and how it ranked them
  - Standards are required for these systems to ”*teach*” them and for practitioners to “*evaluate*” the corresponding output.

Desirable characteristics of a Patent Classification Gold Standard

Scope / Agreement

Scope

- Defining a scope which is both clear enough to offer a reasonable level of agreement between subject matter experts, and reflective of real-world use cases.

Agreement

- Ideally the gold standard covering each topic would be reviewed by multiple subject matter experts — allowing testing against the consensus, most generous, and most narrow definitions.

Diversity / Collection Size

Diversity of technology

- Different patented technology areas have quite differing characteristics in terms of variety of terminology, density of class codes, and quantity of patents, so it's reasonable to assume that different systems will perform with differing degrees of accuracy against each.

Size of dataset

- There is a tension between selecting technologies that are precise enough to be representative of real requirements, yet large enough that multiple experiments can be run without substantial overlap and withholding enough data for the evaluation to be robust and representative.

Challenging / Independent / Identification

Challenging

- Classifying against the gold standard should be sufficiently difficult that existing solutions cannot easily achieve 100% accuracy, which would render any comparison impossible.

Independent

- The gold standard should be created without reference to any existing system, independently, and as far as possible through manual research, to avoid systematic bias – such as the preponderance of a small number of class codes.

Identification

- One of the more trivial though persistent problems in patent data is the lack of standardization of patent serial number formatting. The gold standard should use whatever format is the most widely understood.

Practical guidelines for building Patent Classification Gold Standard Collections

Scope / Agreement

Scope

- These collections need to be relatively specific in scope
  - Qubit generation for quantum computing
    - Specific enough to be useful, but large enough to be practical
    - Can be identified or evaluated against other aspects of quantum computing
  - Cannabinoids edibles

Agreement

- The initial two collections were validated by a single individual after the generation by a first individual
  - Ideally at least three people would evaluate or independently generate collections

Specific scope for current Gold Standards

Qubit

- Qubit Generation for Quantum Computing refers to patents that discuss the various means of generating qubits for use in a quantum mechanics-based computing system. Types of qubits included superconducting loops, topological, quantum dot based and ion-trap methods as well as others. The excluded technologies are applications, algorithms and other auxiliary aspects of quantum computing that do not mention a hardware component, and hardware for other quantum phenomena outside of qubit generation

Cannabinoid edibles

- The positive collection discuss edible items, which can include lozenges, beverages, or powders containing a cannabinoid substance that can be used directly by oral absorption, or by formulating into a foodstuff for oral consumption. Cannabinoid substances include products from Cannabis sativa, ruderalis, or indica as well as products coming from the processing of hemp including hemp seeds, fibers, or oils.
- All the records in the negative collection mention an edible item of one sort or another and, specifically a foodstuff. The records labelled “easier" are publications with a substance like a cannabinoid included but not cannabinoids themselves. The “harder" collection discusses edible items with a plant extract of one sort or another included in the composition.

Diversity / Collection Size

Diversity of technology

- The two existing sets are intentionally very different from one another covering areas of technology that are in different parts of the major classification coding systems (G & H for qubit, A61/A23 for cannabinoid edibles).

Size of dataset

- 500 positives and at least 500 negatives were used in these collections.
  - In both current examples there are 1000 negatives divided into “easier” and “harder”.

How similar should positives & negatives be?

- **Apples and Astronauts** – way too easy
- **Apples and Fish** – still pretty easy
- **Apples and Oranges** – probably just right
- **Fuji and Red Delicious Apples** – likely too hard, especially for practical purposes

Challenging / Independent / Identification

Challenging

- Apples vs. oranges as opposed to apples vs. astronauts

Independent

- Use all available searching methods to create the queries including keywords/Boolean, classification codes, keywords and codes, citations
- Also take advantage of value-added indexing where available

Identification

- All INPADOC family members are included in the positive collections
- This removes family issues with training and during the evaluation of the results the results should be family reduced

Where can I find the existing Gold Standards?

- The data for the quantum computing and cannabinoid edibles gold standards can be found at:

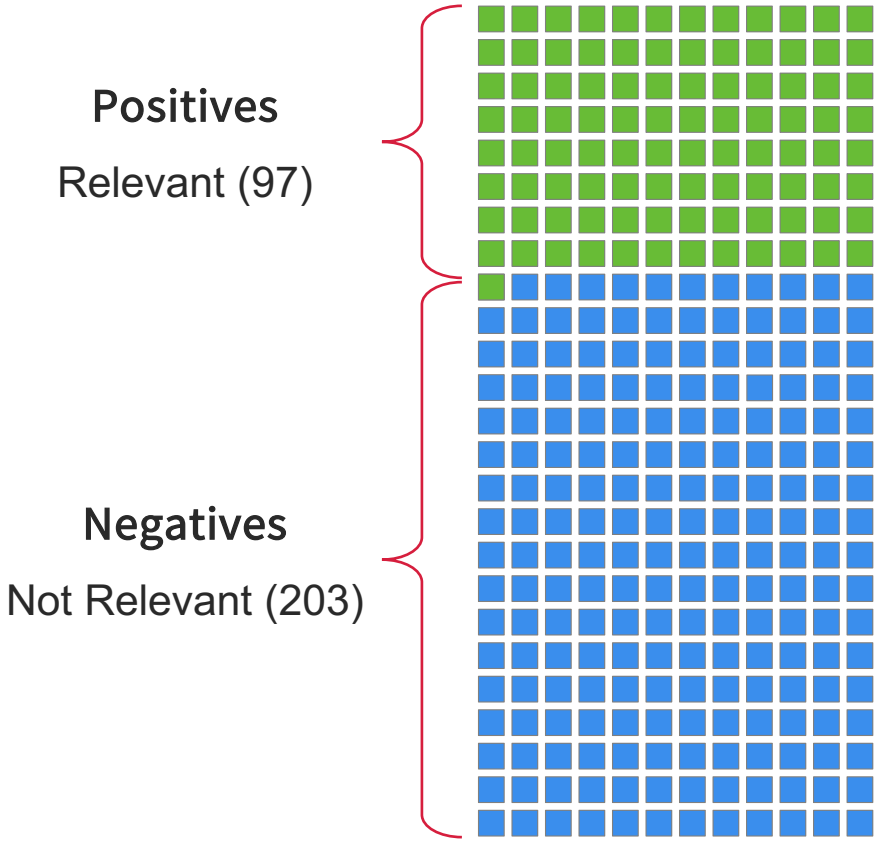
  https://github.com/swh/classification-gold-standard/tree/master/data

- It is made available under the BSD 3-Clause License, to allow reuse in other projects in a variety of ways. The site includes documentation for the file and data format the gold standard is represented in.

What's next - Community Support

- There are currently two data collections
- It would be ideal to have additional collections in each of the major topic areas based on the top levels of the patent classification systems
  - With 7-9 diverse collections we could cover most of technology at a high level
- When used for evaluation this would give a more comprehensive description of the strengths and weaknesses of each product
- Additional information professionals should be used to build the collections
  - Additional stakeholders should step forward to sponsor a study

Measuring accuracy – Gold Standard



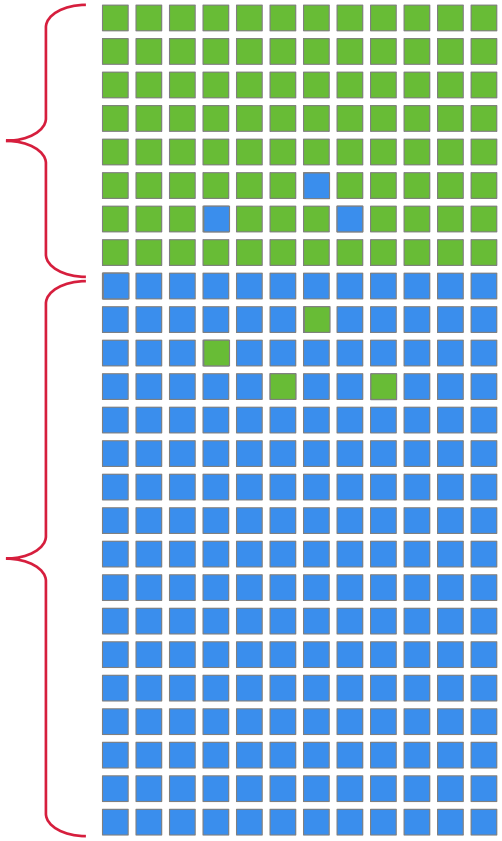
Measuring accuracy – returned results

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"

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

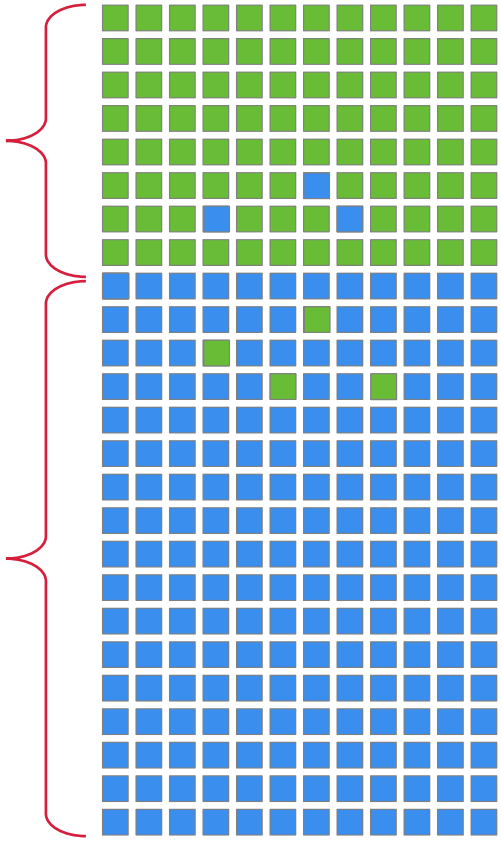
Measuring accuracy – returned results

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
$F_1 = 0.964$

Measuring accuracy – caveats

These numbers only relate to this test – there's no absolute precision and recall.

Unless your test is very, very, carefully constructed the results are misleading.

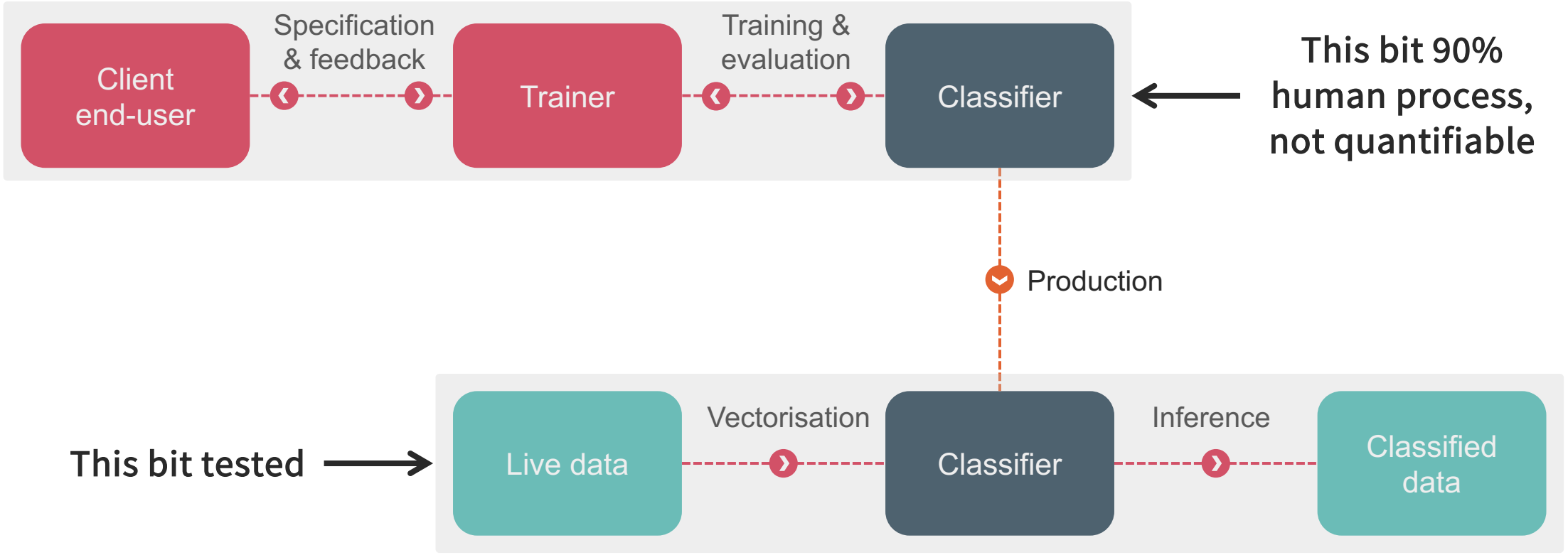
When doing scientific testing we average hundreds of runs.

Measuring accuracy – the real world

The best test is

> **“Are these results useful for what I want to achieve?**
>
> the only way to tell if a system is useful to *you*.

Results


Client end-user
Specification & feedback
Trainer
Training & evaluation
Classifier
This bit 90% human process, not quantifiable
Production
This bit tested
Live data
Vectorisation
Classifier
Inference
Classified data

Cipher's results

Average over 200 runs:

Test	Precision	Recall	F_1
Quantum	0.971	0.971	0.971
Cannabinoid	0.977	0.964	0.971

So what?

Computers can *"understand"* the topic of patents to a similar level as a human expert.

N.B. not the same as e.g. judging essentiality, or litigation worthy-ness.

There's a simple ROI for classification:
Is the 1-2 hours per topic to specify, and the system cost worth the speed and repeatability benefits?

Patent owners are leading the way

"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

Red Hat

"Data science and machine learning helps us better manage and shape our portfolio. The ML tools and models we've built have enabled us to operate more efficiently so that we can execute on our patent strategy."

Head of Patents

Google

"The main strategic benefits of Cipher for ARM are the accuracy of the classifiers, the ability to continue to run those classifiers over time."

Vice President, IP & Litigation

arm

"With improvements in AI technology in analytics platforms such as Cipher, we are able understand the numbers of patents that are relevant to certain technology areas at a push of a button."

Head of Patents

facebook


CIPHER