

Crossing the Chasm Between AI Services and AI-Powered Workflows

Fabio Casati

Credits and sources

The Science of Rejection: A Research Area for Human Computation

Burcu Sayin¹, Jie Yang², Andrea Passerini¹, Fabio Casati³

¹University of Trento

Via Calepina, 14, 38122 Trento TN, Italy

²Delft University of Technology
Mekelweg 5, 2628 CD Delft, Netherlands

³Servicenow
Santa Clara, CA, USA

On the Value of ML Models

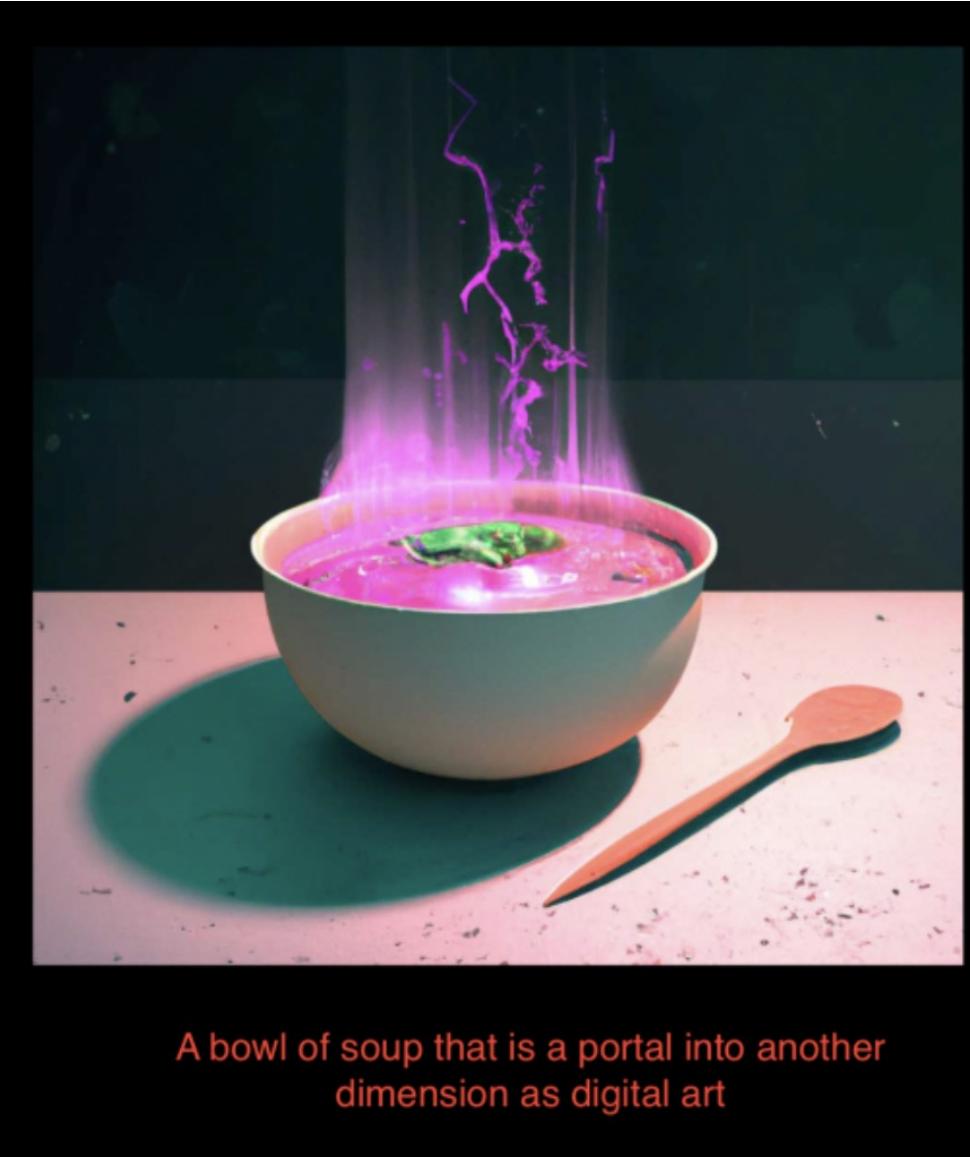
Fabio Casati and Pierre-André Noël,
Element AI, a ServiceNow company

Jie Yang
TU Delft

Stop Using Accuracy to Assess your ML Models.

And why the implications of a correct ML model assessment will redefine what we understand with “learning” in ML.

"Consumer" AI today: simple, effective, impressive

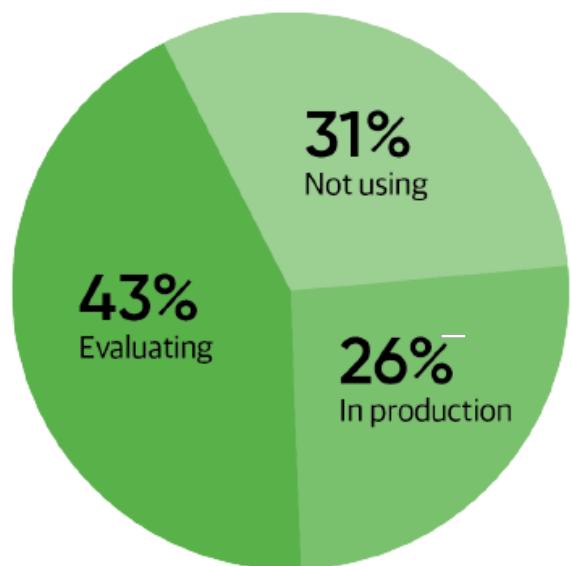


Scharfsinn/Shutterstock

A bowl of soup that is a portal into another dimension as digital art

But what about AI in the Enterprise?

*... the results were surprisingly similar to 2021.
Furthermore, if you go back another year, the 2021
results were themselves surprisingly similar to
2020. [...]*
*The same percentage of respondents said that their
organizations had AI projects in production (26%).*



O'Reilly. AI Adoption in the Enterprise 2022

*Findings from the 2021 survey indicate that
AI adoption is continuing its steady rise: 56
percent of all respondents report AI adoption
in at least one function, up from 50 percent in
2020.*

McKinsey Global Survey - State of AI

Why? Commonly cited reasons / blockers

[...] the biggest bottlenecks were lack of skilled people and lack of data or data quality issues (both at 20%), followed by finding appropriate use cases (16%).

[O'Reilly. AI Adoption in the Enterprise 2022](#)

The most frequently cited barriers to AI adoption are a lack of a clear strategy, a lack of talent, and functional silos.

[McKinsey - AI adoption advances, but foundational barriers remain](#)

My own experience (from in depth engagements):

- 1. Risk**
- 2. Value**
- 3. Adoption and Maintenance Journey**

AI-Powered “Workflows” and AI Services in the Enterprise

Workflows vs guided actions vs journeys

Focus is on a **case**



Approval flow for requested items

TRIGGER



Service Catalog

Waits for an item to be requested.

ACTIONS

1 Get Catalog Variables from Standard Laptop
Select the catalog item and variables to use in your flow

2 ▾ If Price is over \$1000.00

3 then Ask For Approval
Ask the requester's Manager for approval

4 ▾ Else

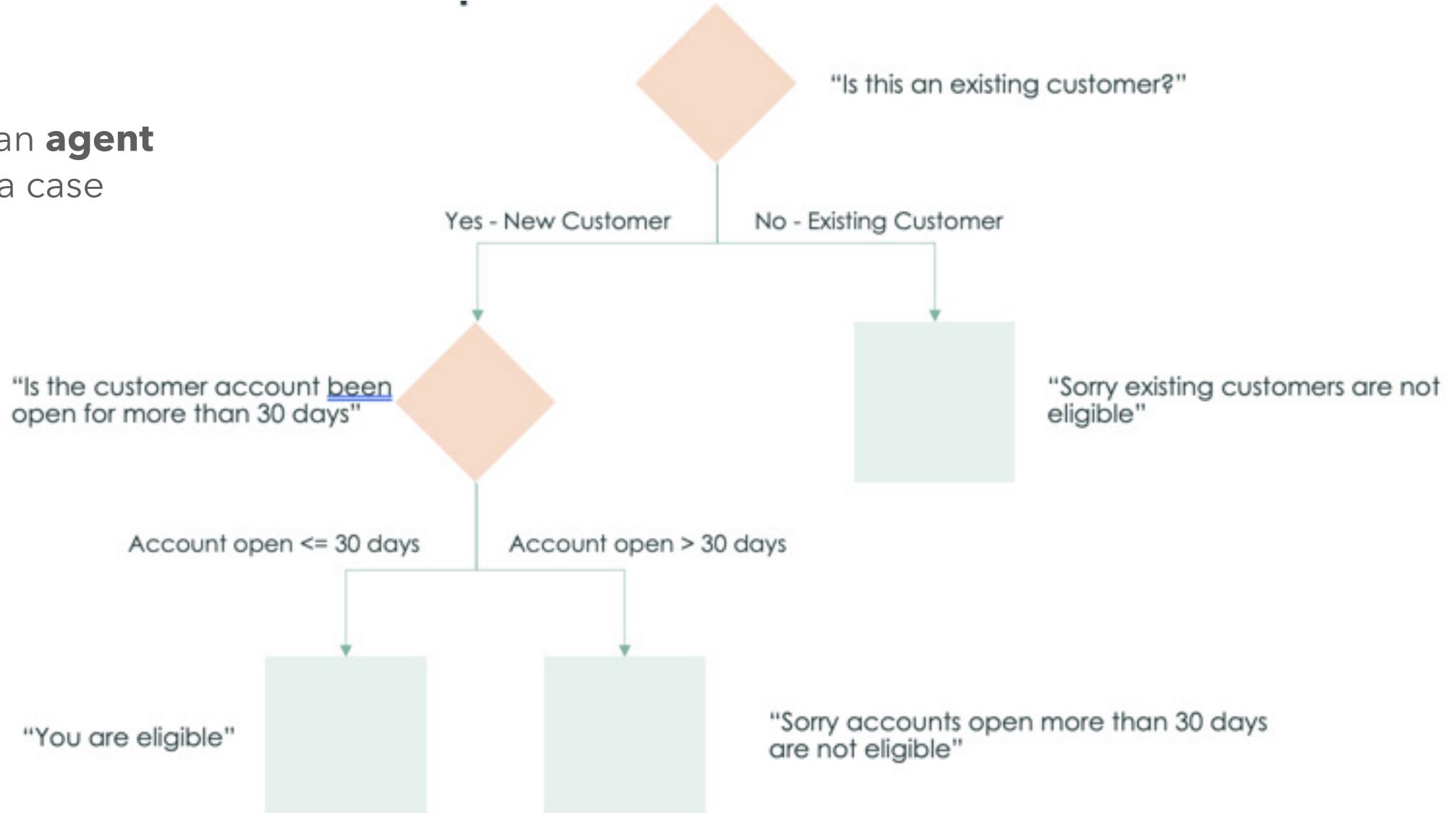
5 Update Requested Item Record



Add an Action, Flow Logic, or Subflow

Workflows vs guided actions vs journeys

"consumer" is an **agent**
supporting a case



Workflows vs guided actions vs journeys

see recommended readings

search

chat with virtual agent

submit a ticket



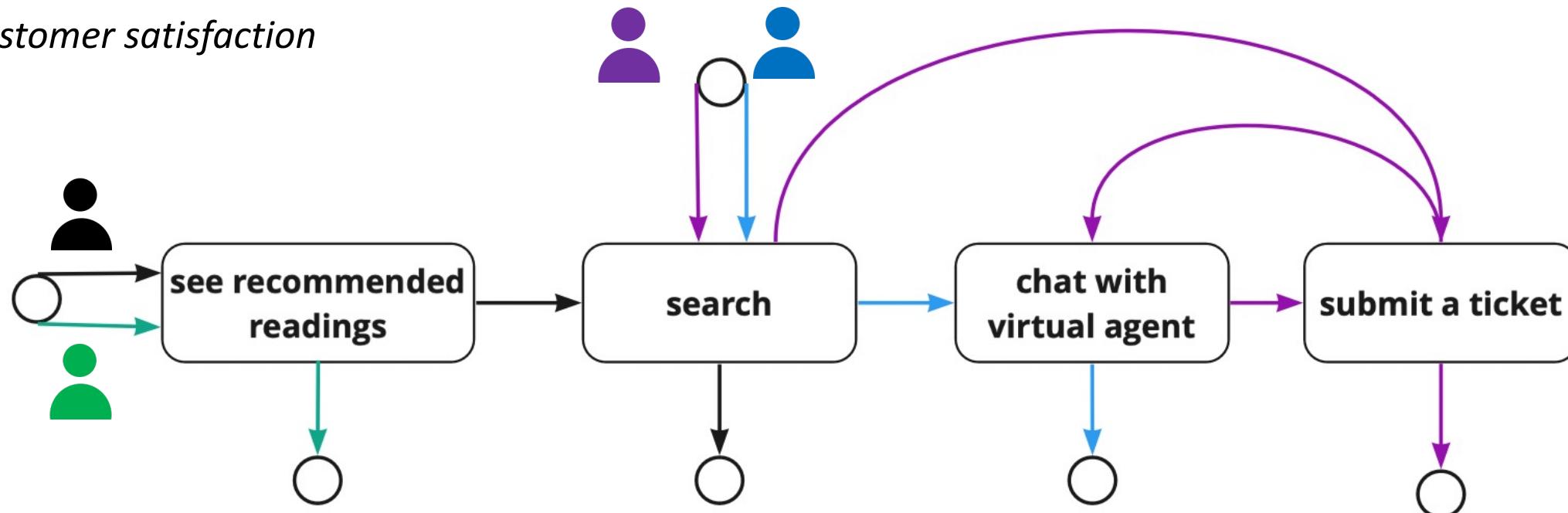
Workflows vs guided actions vs journeys

Cost

User experience

Customer satisfaction

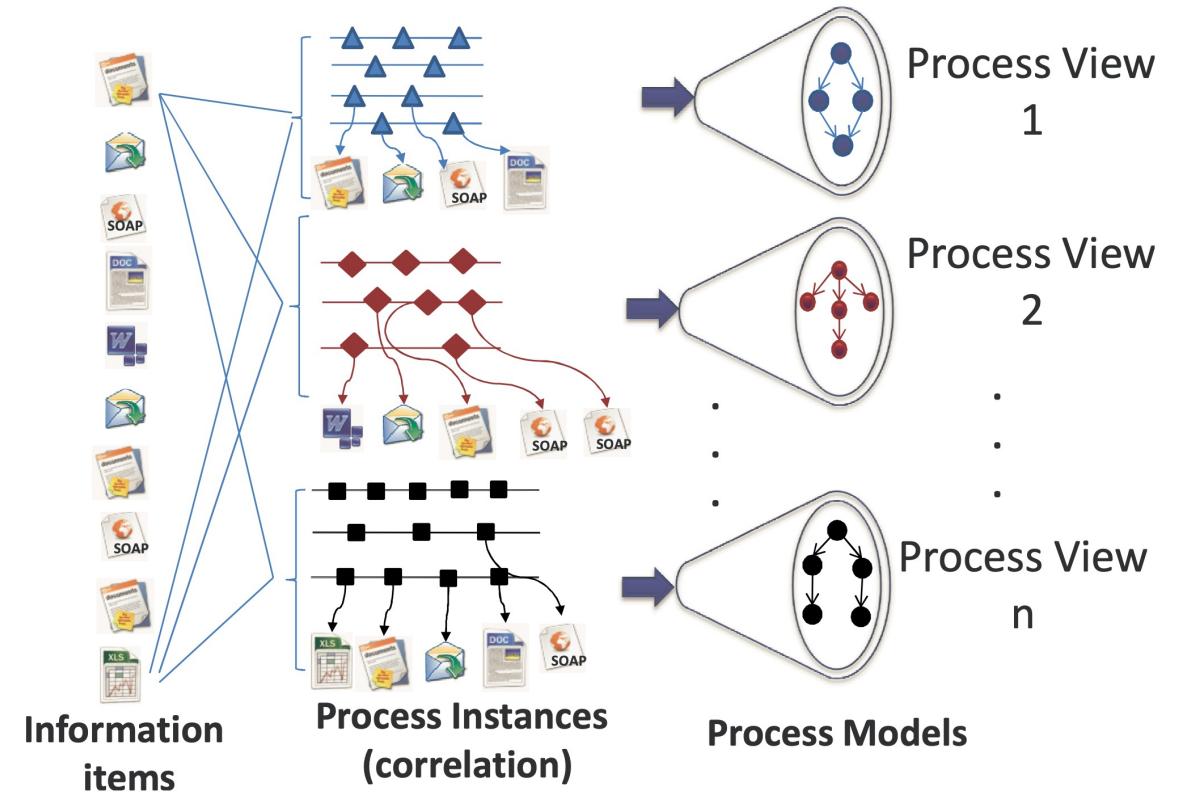
...



Focus is on the **user** (customer with a need) or an **agent** or a **case**

We need tools to monitor journeys

- Logging abstractions
- Context and correlation



Process spaceship: Discovering and exploring process views from event logs in data spaces. VLDB 2008

AI Services supporting journeys and workflows

see recommended readings

search

chat with virtual agent

submit a ticket



AI powered search

conversational AI

ticket routing

detect major incident

genius results

next best action

ticket categorization

detect root cause

recommender system

root cause of SLA violations

Smart work assignments

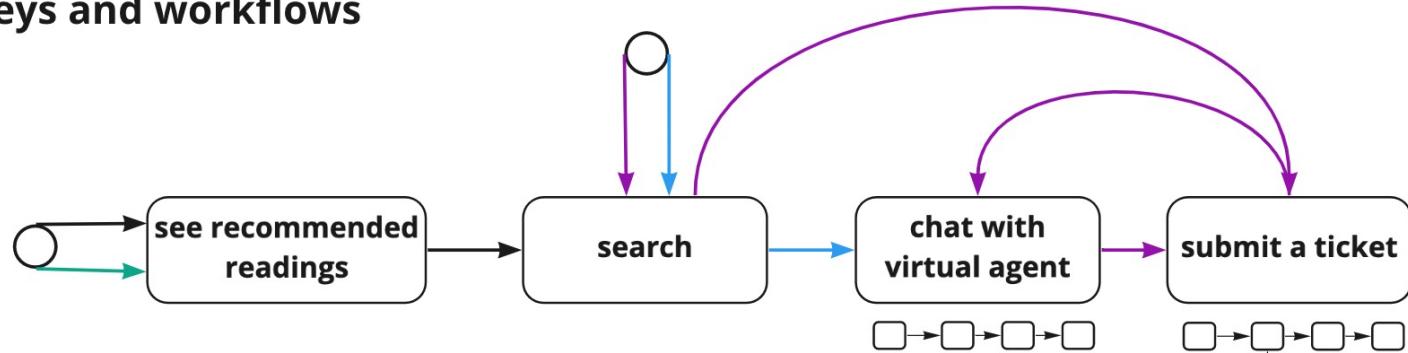
Find similar tickets

...

Here is where we measure value.
This is what we care about



Journeys and workflows



Enterprise AI Service Customer

AI Services Integration workflows

AI Services

Smart work assignments

ticket routing

...

AI Service Provider

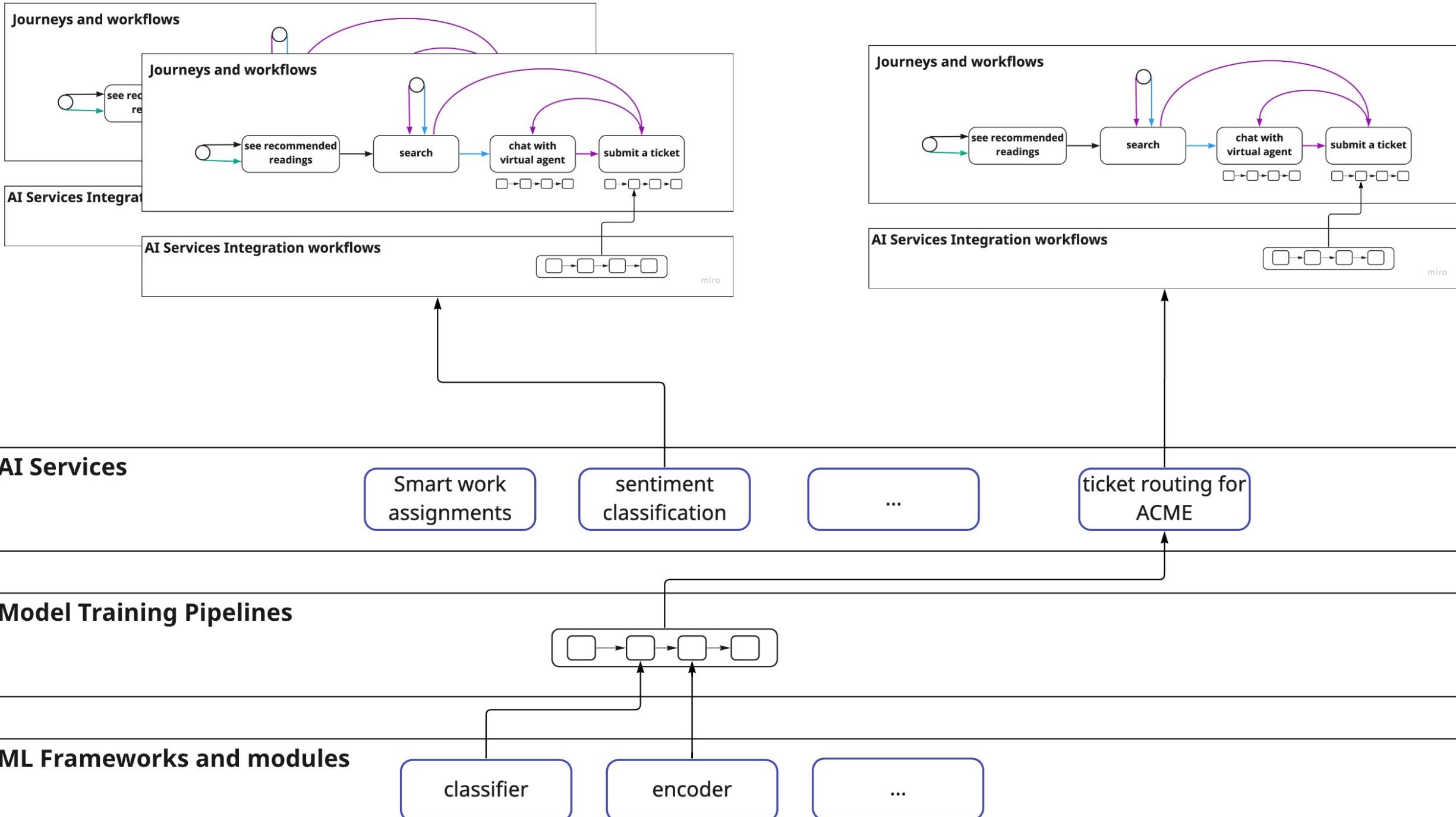
Model Training Pipelines

ML Frameworks and modules

classifier

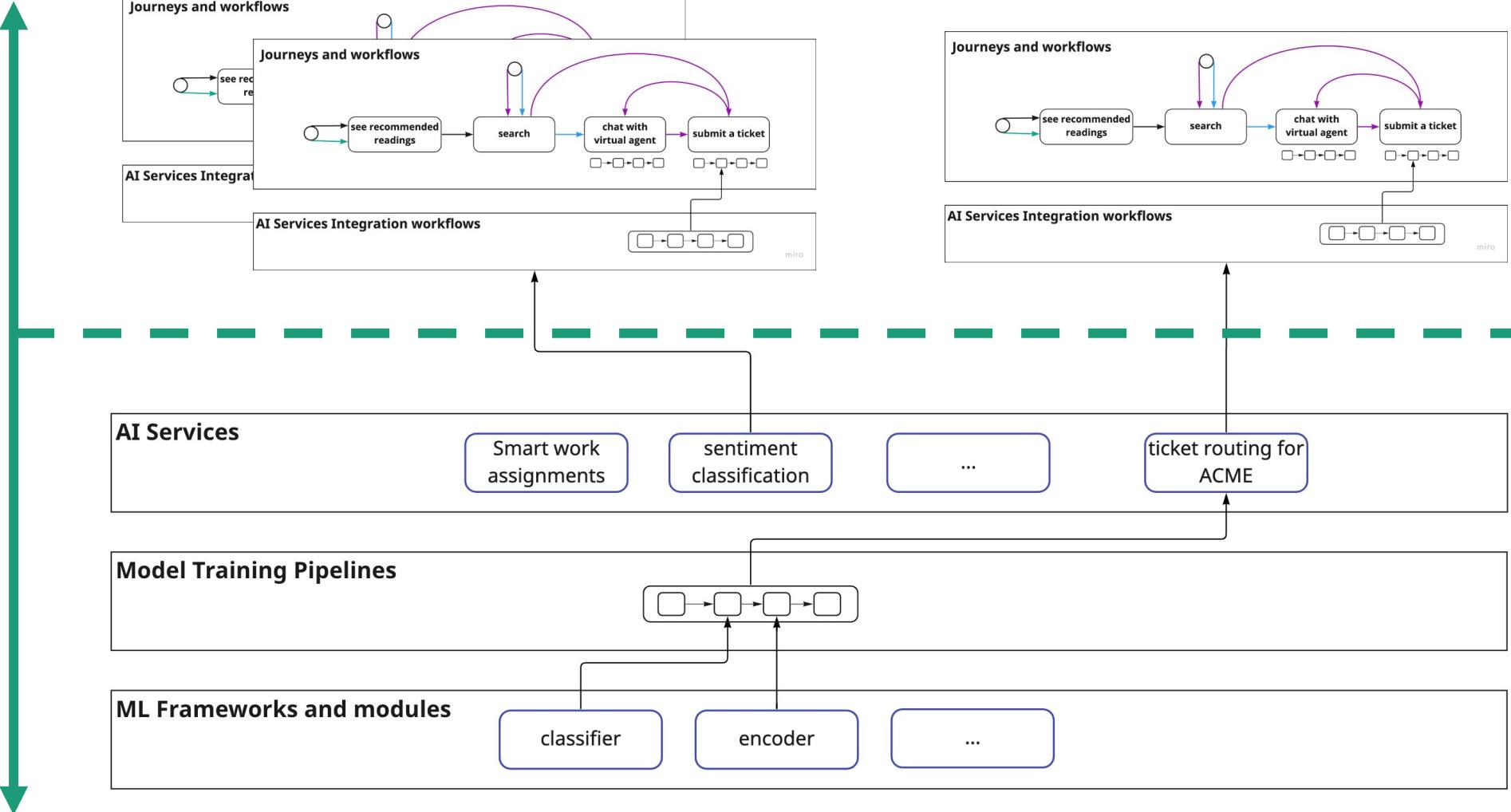
encoder

...

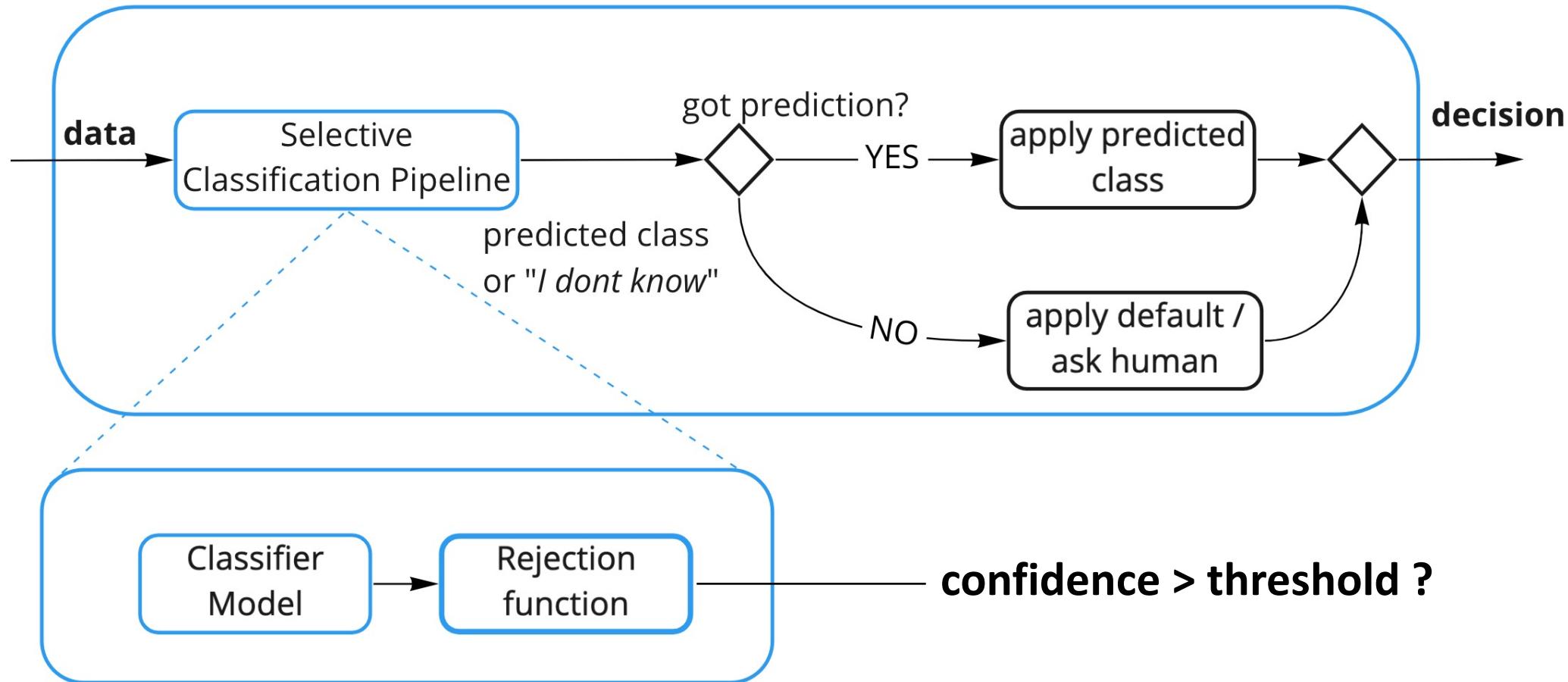


Forward and backward prop

Enterprise AI Service Customer



AI powered steps are not steps, and they are not just AI service invocations



This is not an exception: there are (almost) always defaults or humans-in-the-loop

Even search

Google X |

All Maps Images News Videos More Tools

About 3,290 results (0.41 seconds)

Did you mean: **mispelled** some **search** query

It looks like there aren't many great matches for your search

Tip: Try using words that might appear on the page you're looking for. For example, "cake recipes" instead of "how to make a cake."

Need help? Check out other [tips](#) for searching on Google.

<http://kmeleonbrowser.org> › forum › read ...
Update Search Engines for KM 1.6 - K-Meleon
Aug 24, 2010 · 24 posts · 8 authors
I worked in his update **some** time ago but after a hard month working I lost the final file after ...
<http://yellowpages.aol.com/search?query=>

<https://scienceon.kisti.re.kr> › srch › s... · [Translate this page](#) ...
Generating synonyms based on query log data - ScienceON
An approach is described for generating synonyms to supplement at least one information item, such as, in one case, a set of related items.
<https://scienceon.kisti.re.kr> › srch › selectPORsRchPatent ...
Data input and processing apparatus having spelling-check ...
A data input and processing apparatus as used in a typewriter, having a function of checking the spelling of entered words which are separated by word- ...

Key points so far

1. “Value” is measured on journeys and workflows
2. The value and impact of AI services depends on the application use case
3. AI services are almost invariably applied as a selective inferences – we’ll see why this matters

AI Service Quality

“Obvious things that need to be said”

Accuracy and friends

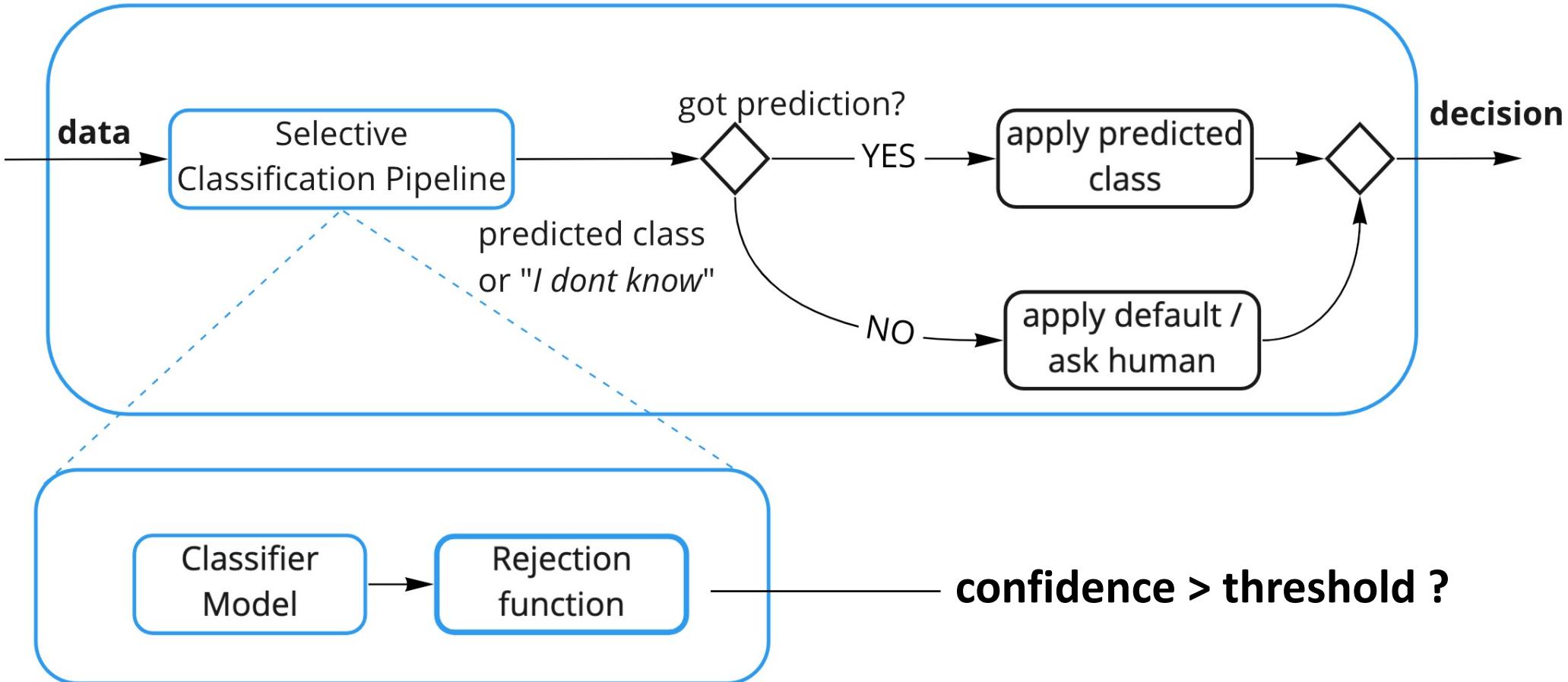
How do we measure AI services today?

Overwhelmingly, we use **accuracy**.

Leaderboards do it too.

Customers tend to do it too, at first. Why?

Accuracy is the metric you use if you don't care about accuracy



Value of an integrated AI service in a use case (impact on business KPI)

The first question we need to answer is: what is the impact of a correct, wrong, skipped prediction

$$V = \begin{cases} V_r \text{ if correct} \\ V_w \text{ if wrong} \\ V_s \text{ if skipped} \end{cases} \quad (\text{can be a matrix})$$

These are parameters we can (**should!**) measure, or estimate - if we have the proper infra in place

Computing Value for $g(f(x))$

$$V(g, D) = \rho V_r + (1 - \rho)(\alpha V_c + \sum_{ij} [\Omega \odot V_W]_{ij})$$

$$V(g, D) = \rho V_r + (1 - \rho)(\alpha V_c + (1 - \alpha)V_w)$$

$$V(g, D) = (1 - \rho)(\alpha - k(1 - \alpha))$$

$V_r = \phi$ BASELINE

***model quality is use-case specific!
or, at least, it depends on the cost ratio k***

“Obvious things that need to be said”

1. Measuring accuracy implies assuming $k=0$
2. Because the value of an AI service depends on the use case, I cannot say that S_1 is better than S_2 “in general”
3. Every well-calibrated/ “properly rejected”* AI service is valuable, no matter how bad it's accuracy is

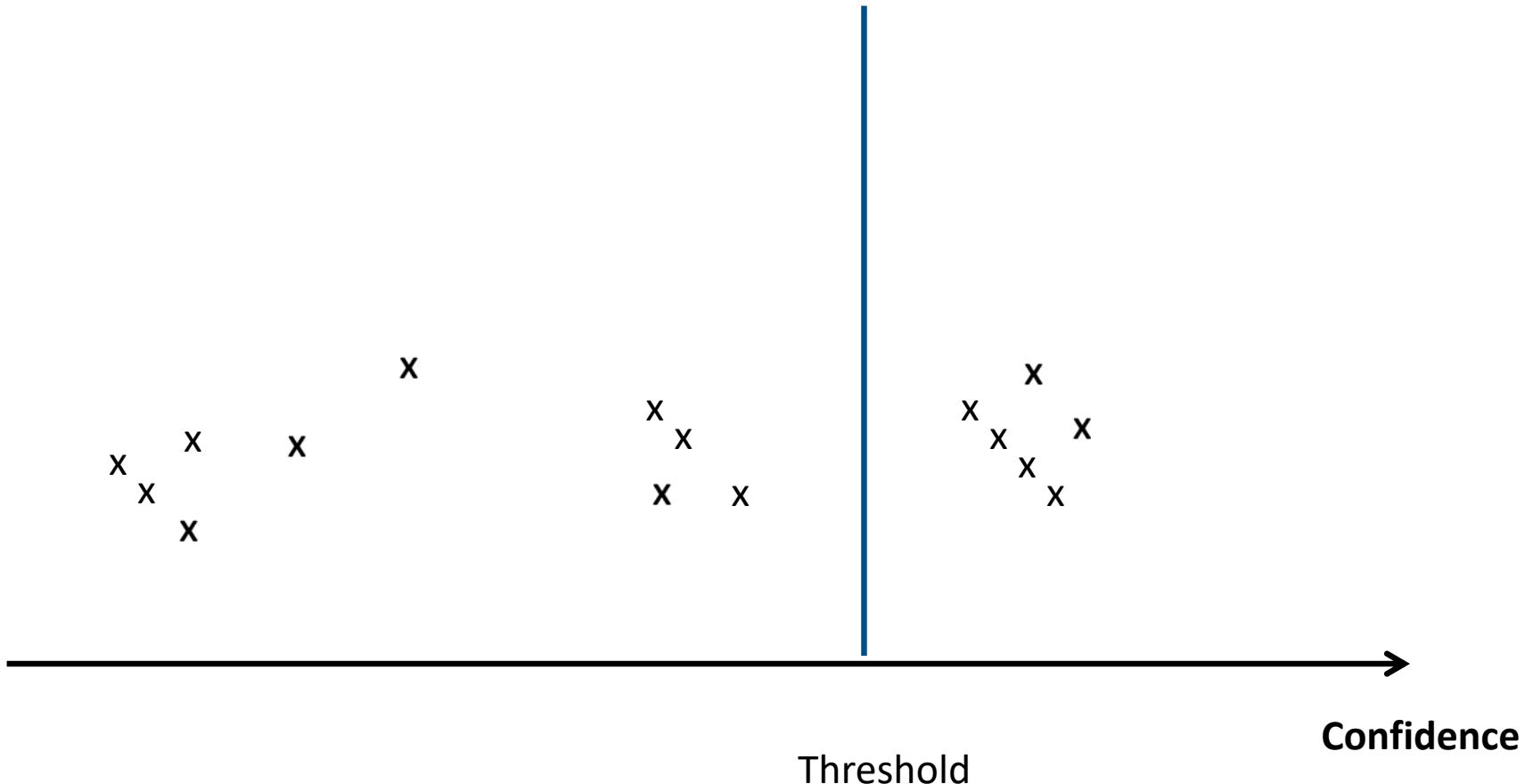
*more on calibration, another misunderstood concept, later in the presentation (if time permits)

Is accuracy a good proxy for business value?

- Sometimes value is negative
- Without confidence (or without adapting thresholds to costs), values decrease linearly
- Even large models perform badly (and sometimes way worse than models from the last century) as costs grow

$$V(g, D) = (1 - \rho)(\alpha - k(1 - \alpha))$$

Implication on active learning



In summary

On the Workflow Side

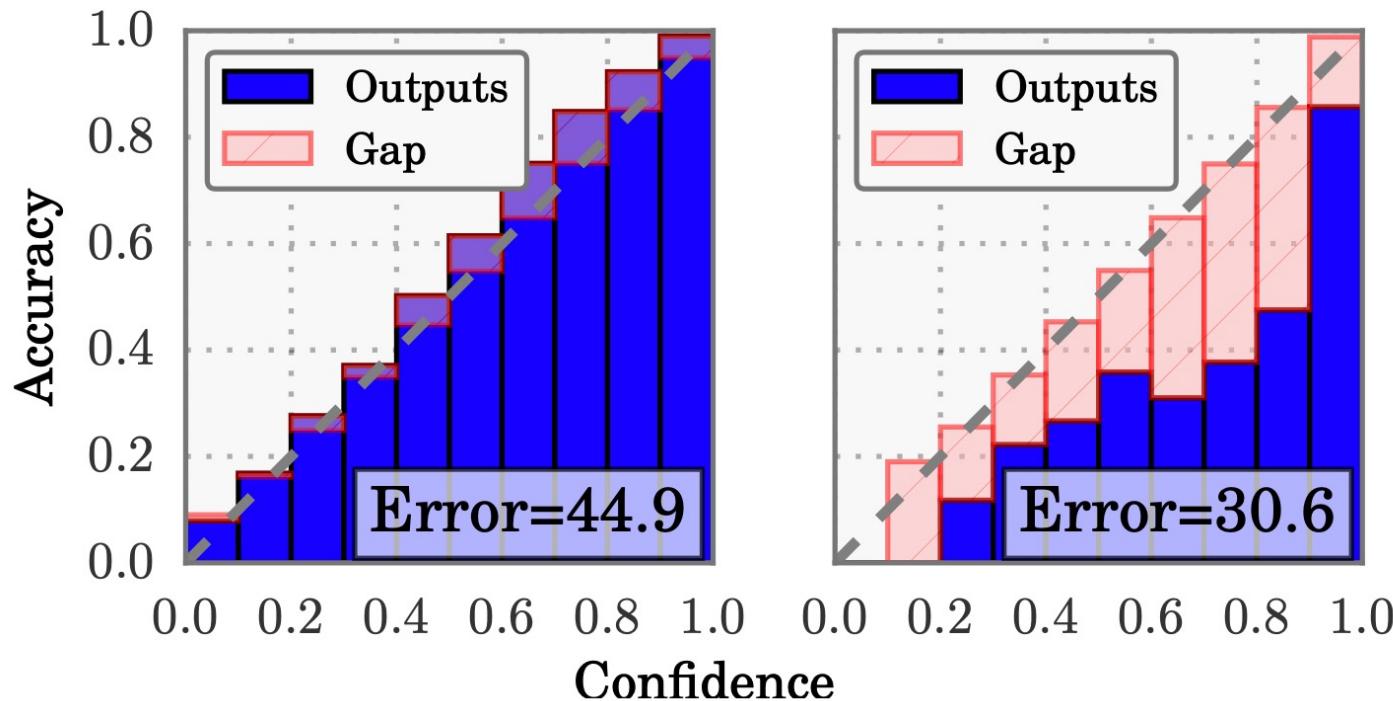
1. Link services up to value
2. KPI Impact of skipped, rejected, correct, wrong
3. Pick the right service, set the right thresholds. This can make a massive difference
4. Tune the threshold (rather than calibrate, rather than trusting service confidence)

On the AI service side

1. Pick the right metric, plot value curve, give guidance
2. The quality of a model depends on how well it knows if it knows, not just in getting the right inferences
3. Also in active learning context, we prioritize value, not accuracy

Calibration

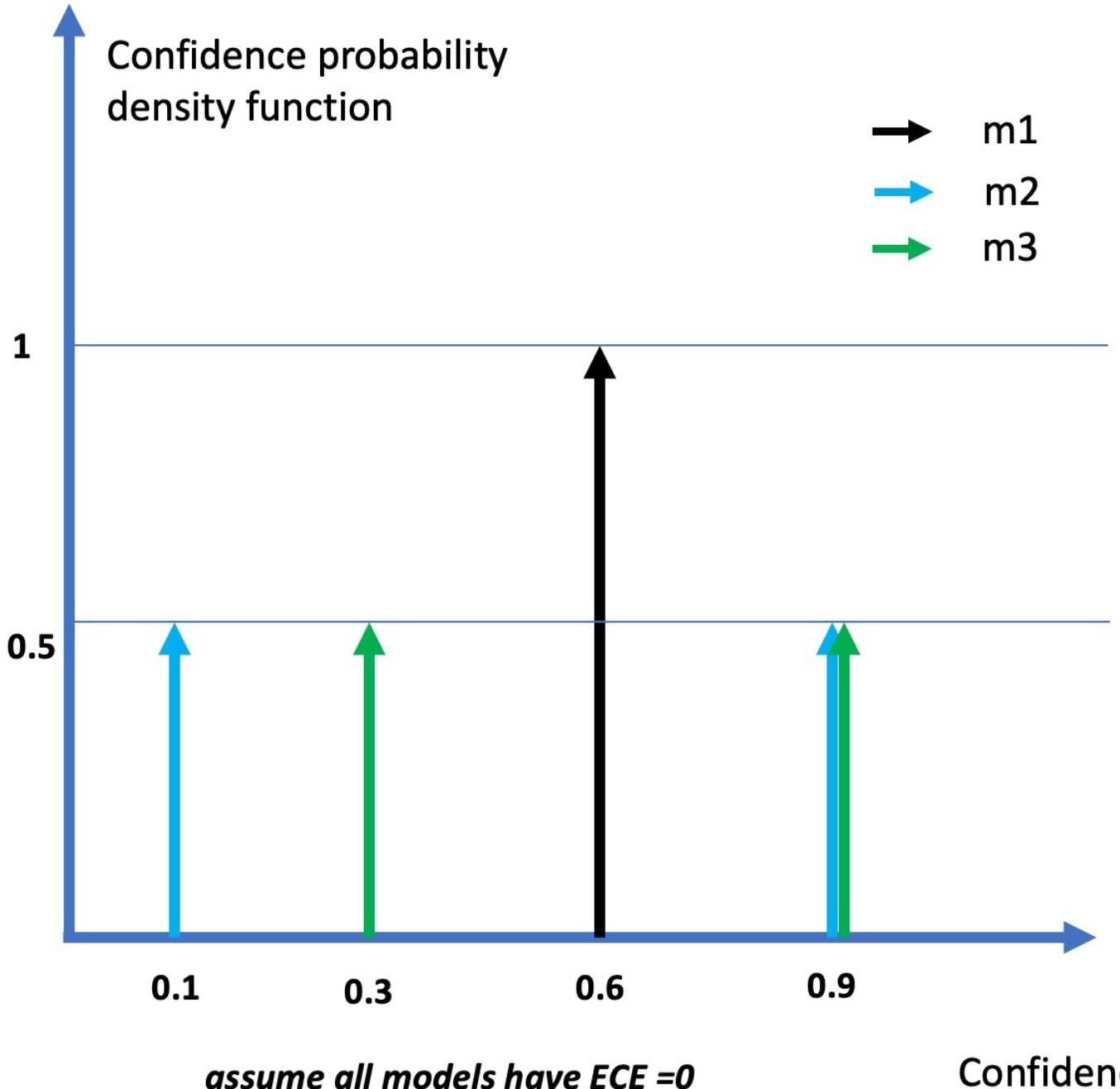
The importance of the most misunderstood and undervalued concept in ML: confidence scores



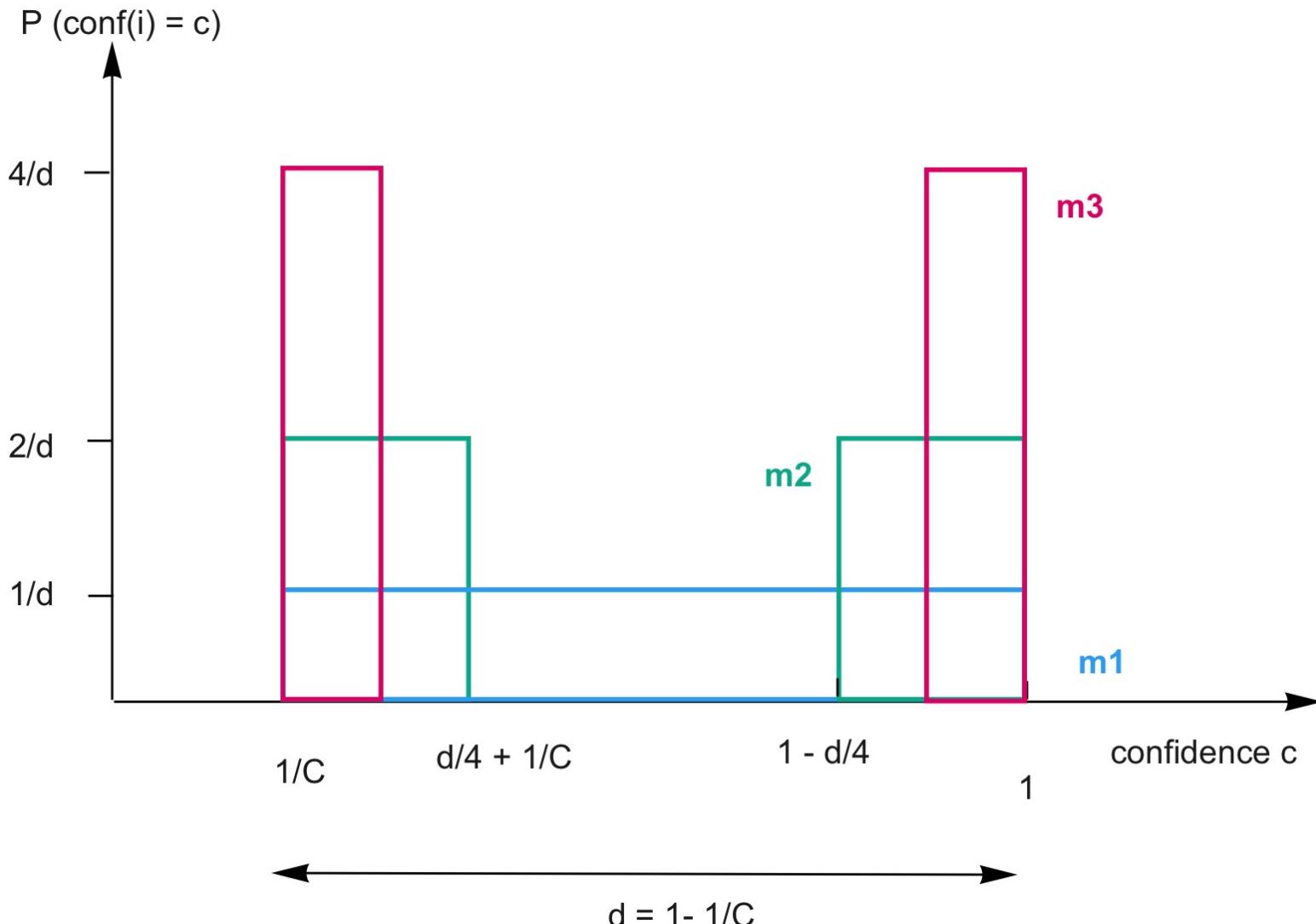
*Guo et al.
On Calibration of Modern Neural
Networks*

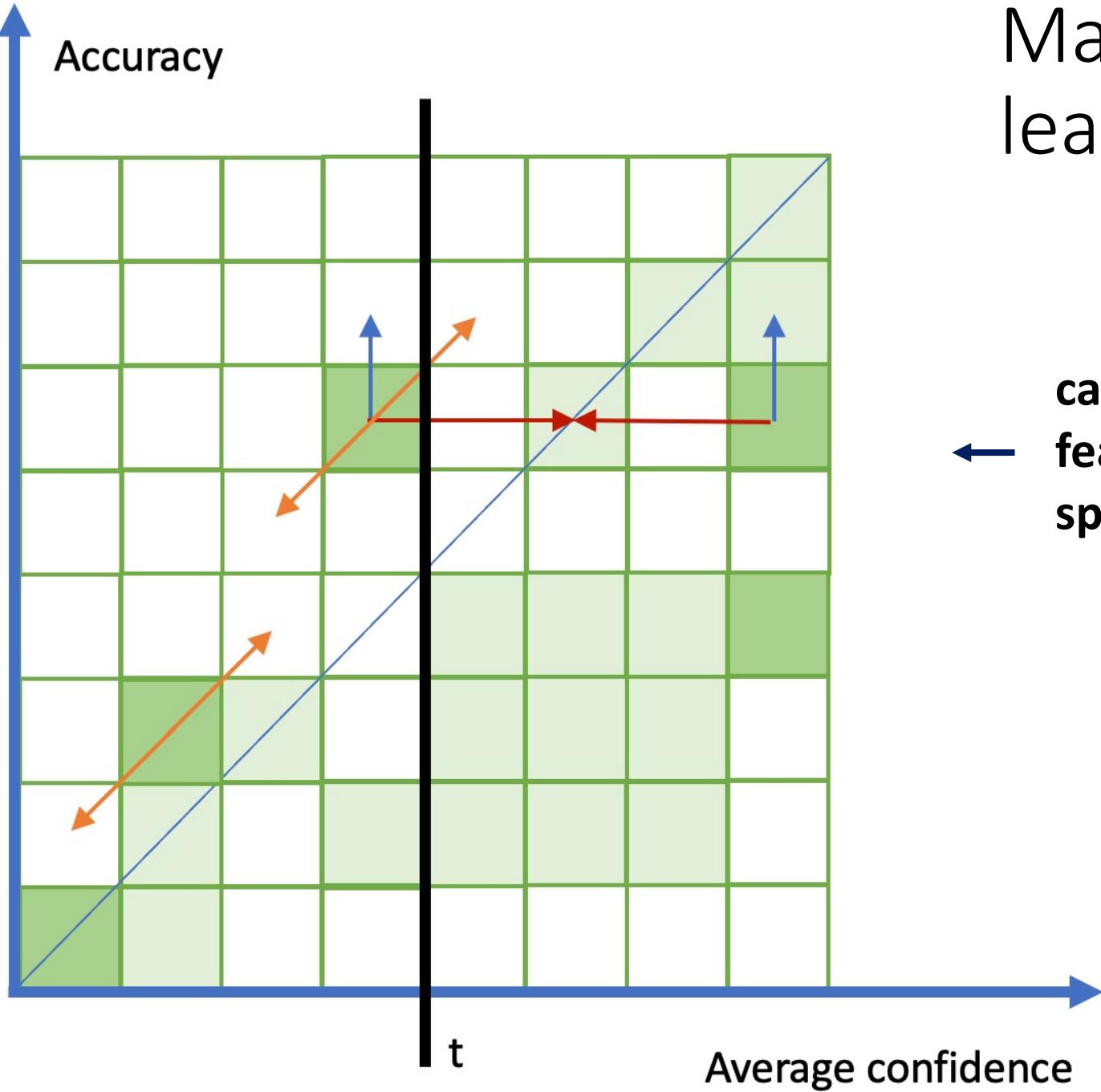
What is, really, “calibration”?

Think about your
doctor



Consider these perfectly calibrated models wits same accuracy: which one is better? has m₃ “learned”?





Many types of learning

← can be a
feature
space

“

So I left him, saying to myself, as I went away: Well, although I do not suppose that either of us knows anything really beautiful and good, I am better off than he is - for he knows nothing, and thinks that he knows. I neither know nor think that I know. In this latter particular, then, I seem to have slightly the advantage of him[...] I found that the men most in repute were all but the most foolish; and that some inferior men were really wiser and better.

— The Apology of Socrates

Thanks

fabio.casati@servicenow.com

<https://medium.com/@sphoebs>