

Machine Learning

What is machine learning?

- Creating computer programs that can “learn,” that is, they improve their performance by leveraging **data** to perform one or more **tasks**.
- This is in contrast to “regular programs” or programs that don’t learn: they must be programmed to do their tasks.

Syllabus things

- Website: pkirlin.github.io/ml-s23/
- Textbook: none, but there will be readings posted
- Prerequisites: math math math

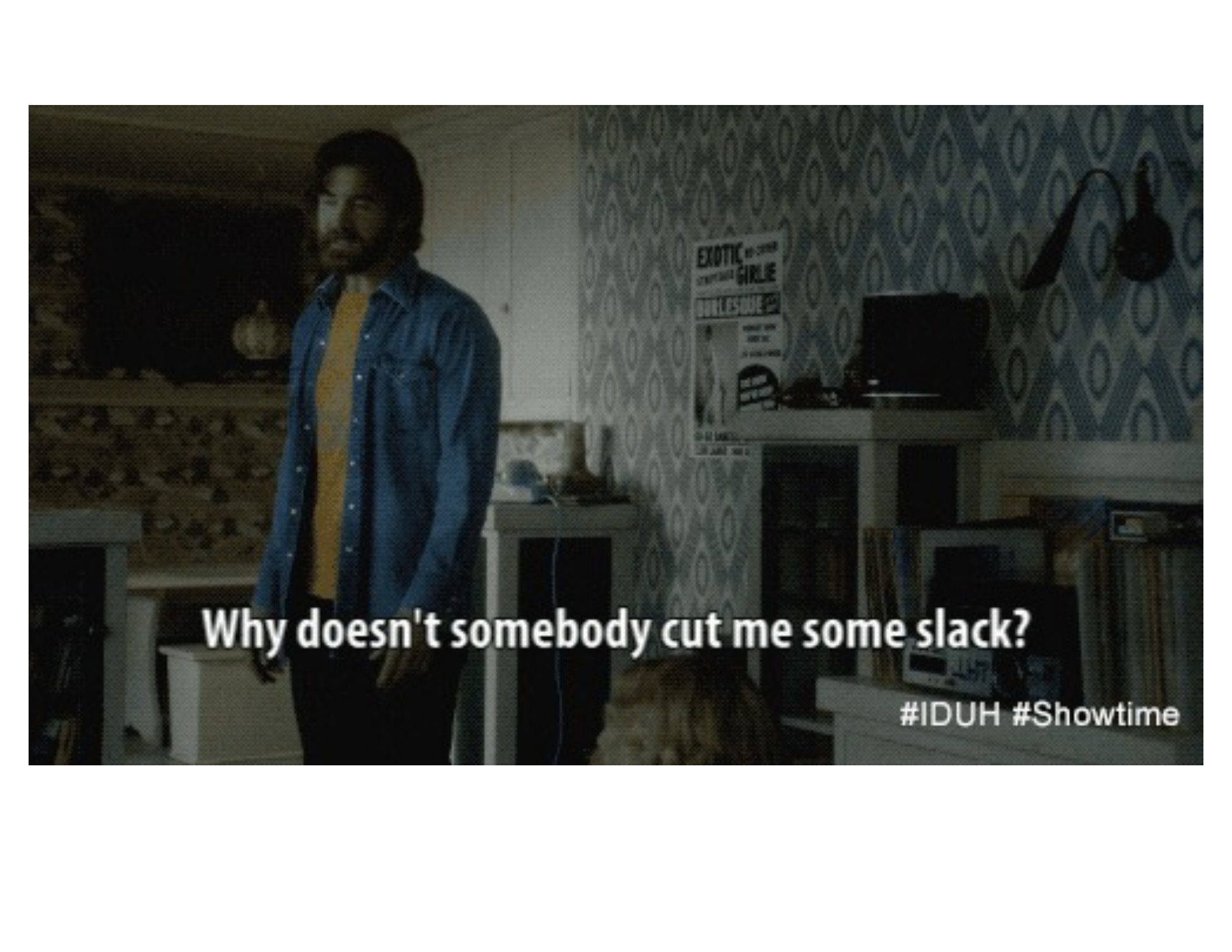
Coursework

	Tentative weight
Homework/Programming Projects	40%
Midterm 1	20%
Midterm 2	20%
Final Project	20%

Working independently

- Out-of-class assignments must be done independently, however, you may ask others for help.
- Rule 1: Do not look at anyone else's code for the same project or a similar project.
- Rule 2: Do not write code or pseudocode with anyone else. (This now includes ChatGPT!) 😊

Late Work and Makeups

A man with a beard and long hair, wearing a blue denim shirt over a yellow t-shirt, stands in a room. The room has patterned wallpaper and a shelf with a magazine titled "EXOTIC GIRLIE". The text "Why doesn't somebody cut me some slack?" is overlaid on the image.

Why doesn't somebody cut me some slack?

#IDUH #Showtime

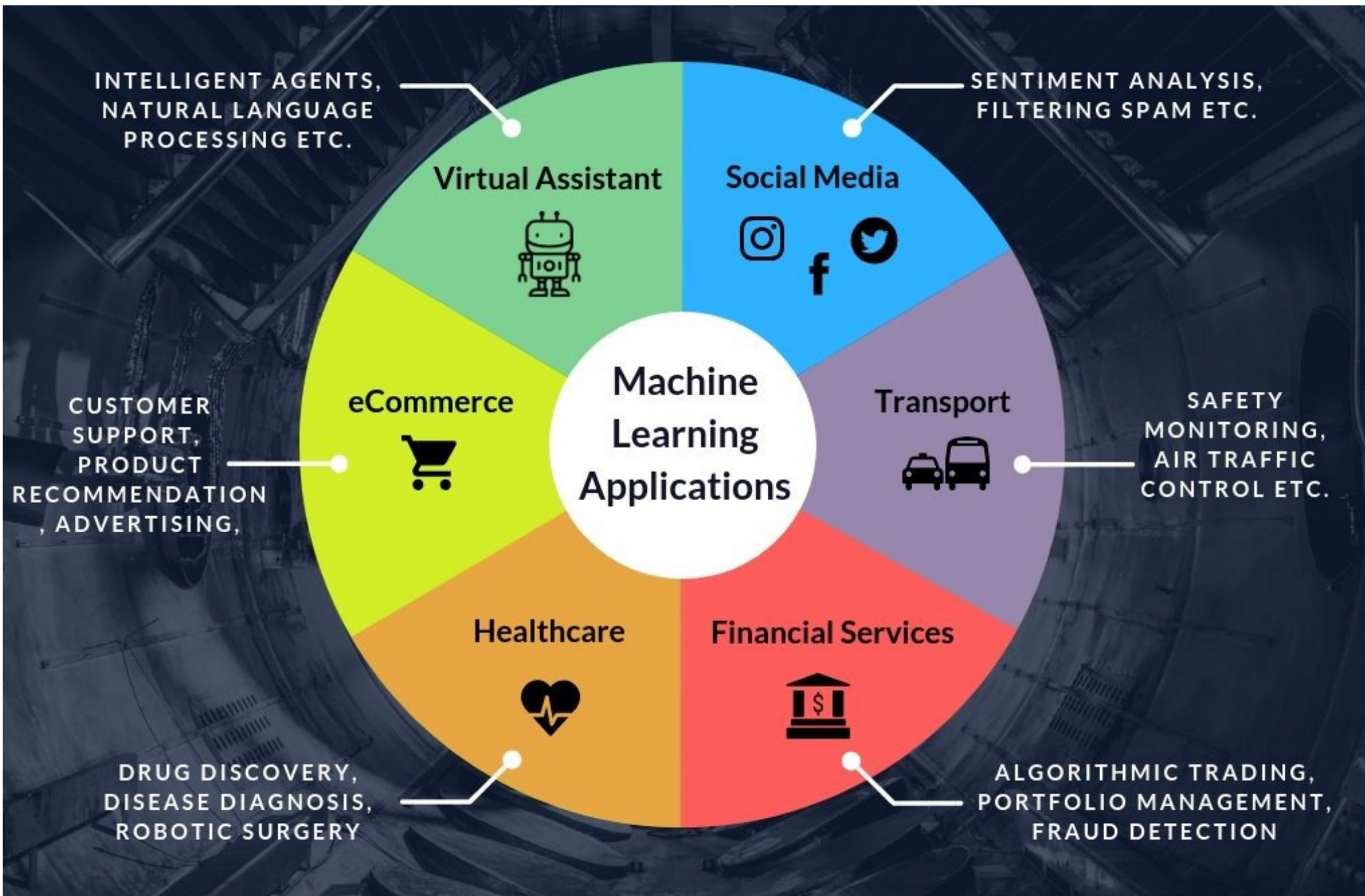
About Me

About You

- Name (what you want to be called)
- Pronouns (if you want)
- Where you're from (interpret however you want)
- Class year (first-year, sophomore, junior, senior)
- Something boring that you have a strong opinion about.

What is machine learning?

- What comes to mind when you hear the term?



AI in art



AI in art





- <https://deepai.org/machine-learning-model/text2img>
- <https://creator.nightcafe.studio/>

Machine learning in writing

Machine learning in face generation

- <https://thispersondoesnotexist.com/>

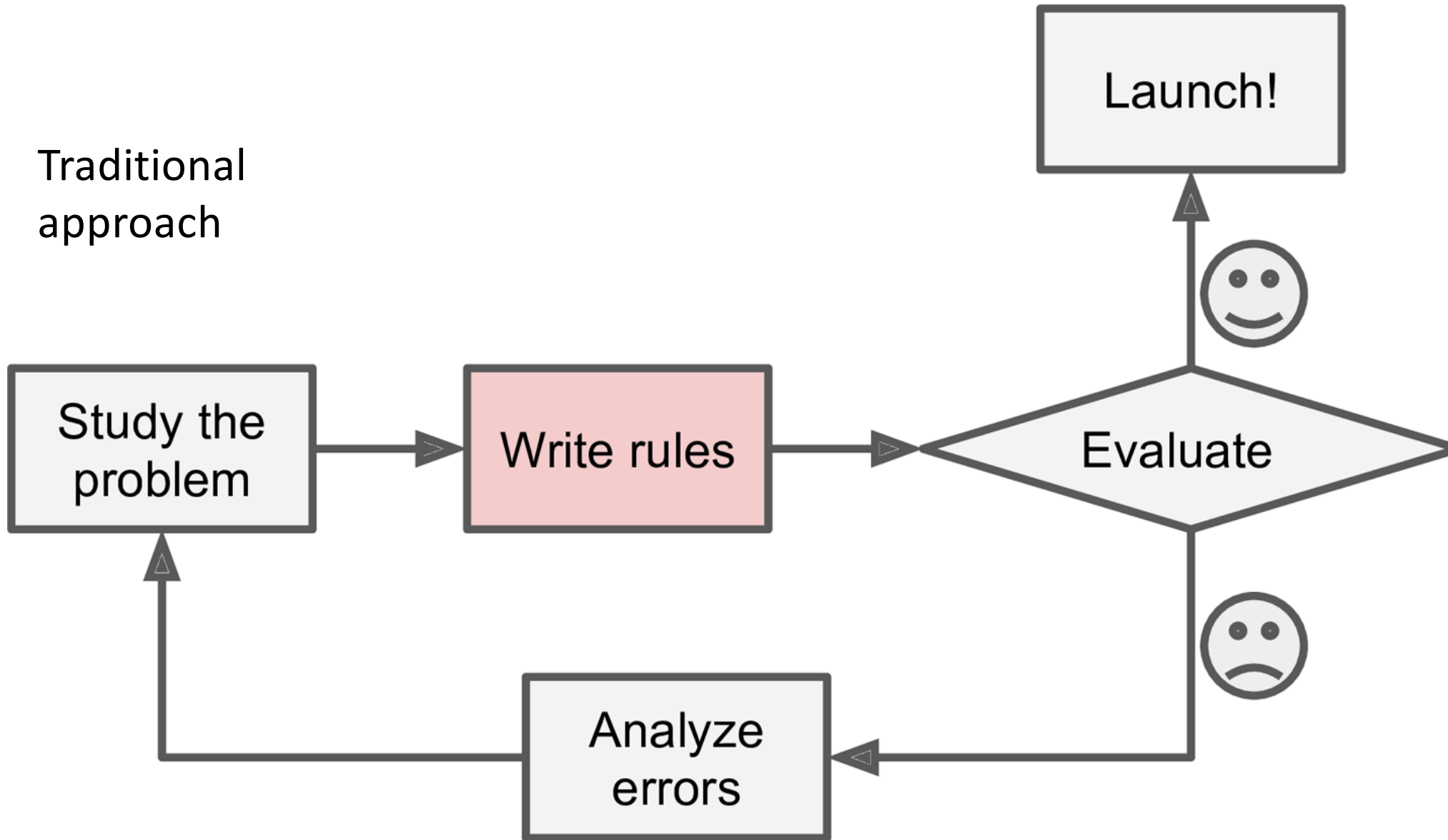
Machine learning in game playing

- https://www.youtube.com/watch?v=8tq1C8spV_g&t=1s

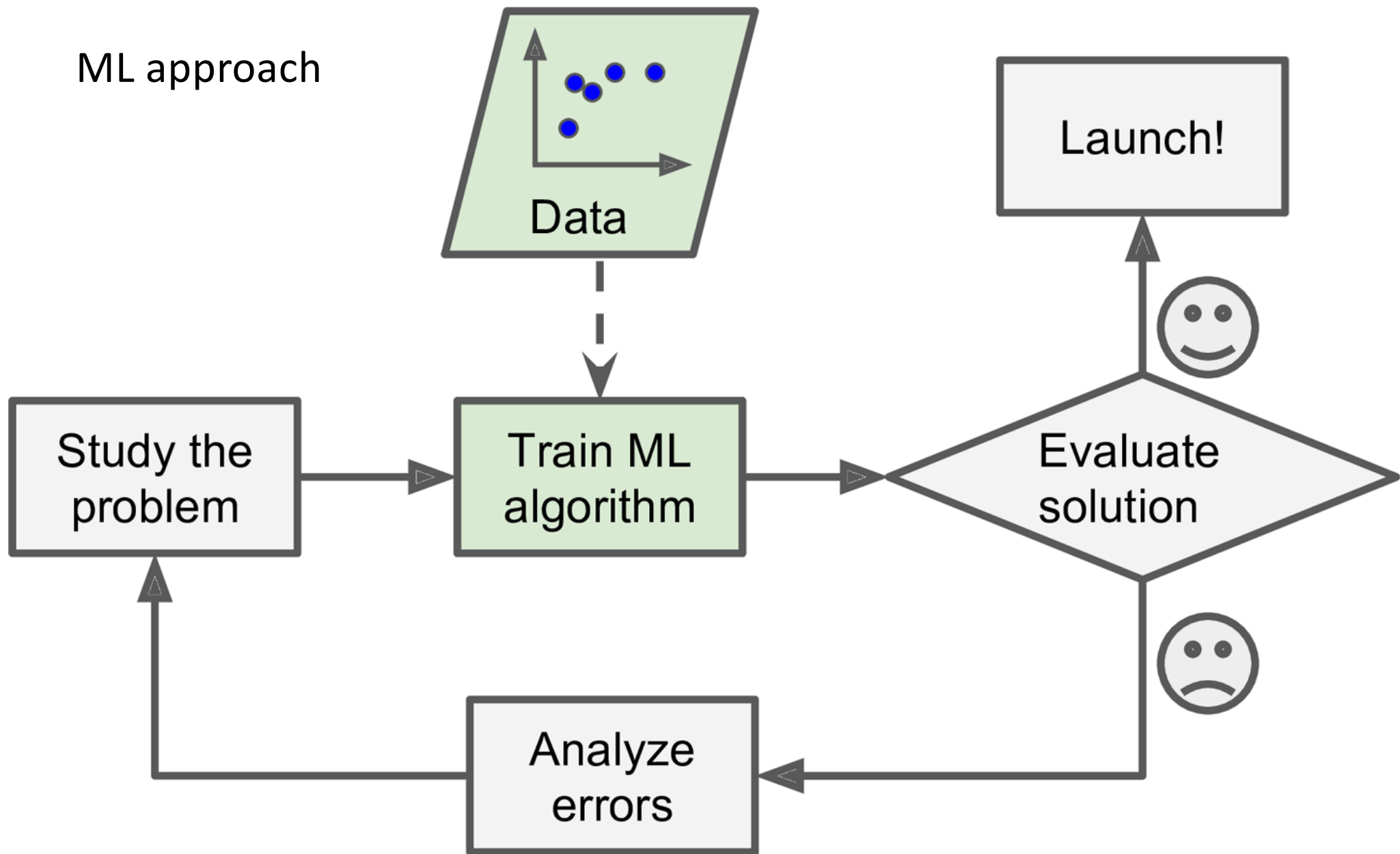
- Machine Learning is the science (and art) of programming computers so they can *learn from data*.
- Here is a slightly more general definition:
- *[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.*
 - Arthur Samuel, 1959
- And a more engineering-oriented one:
- *A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E .*
 - Tom Mitchell, 1997

Spam Filtering

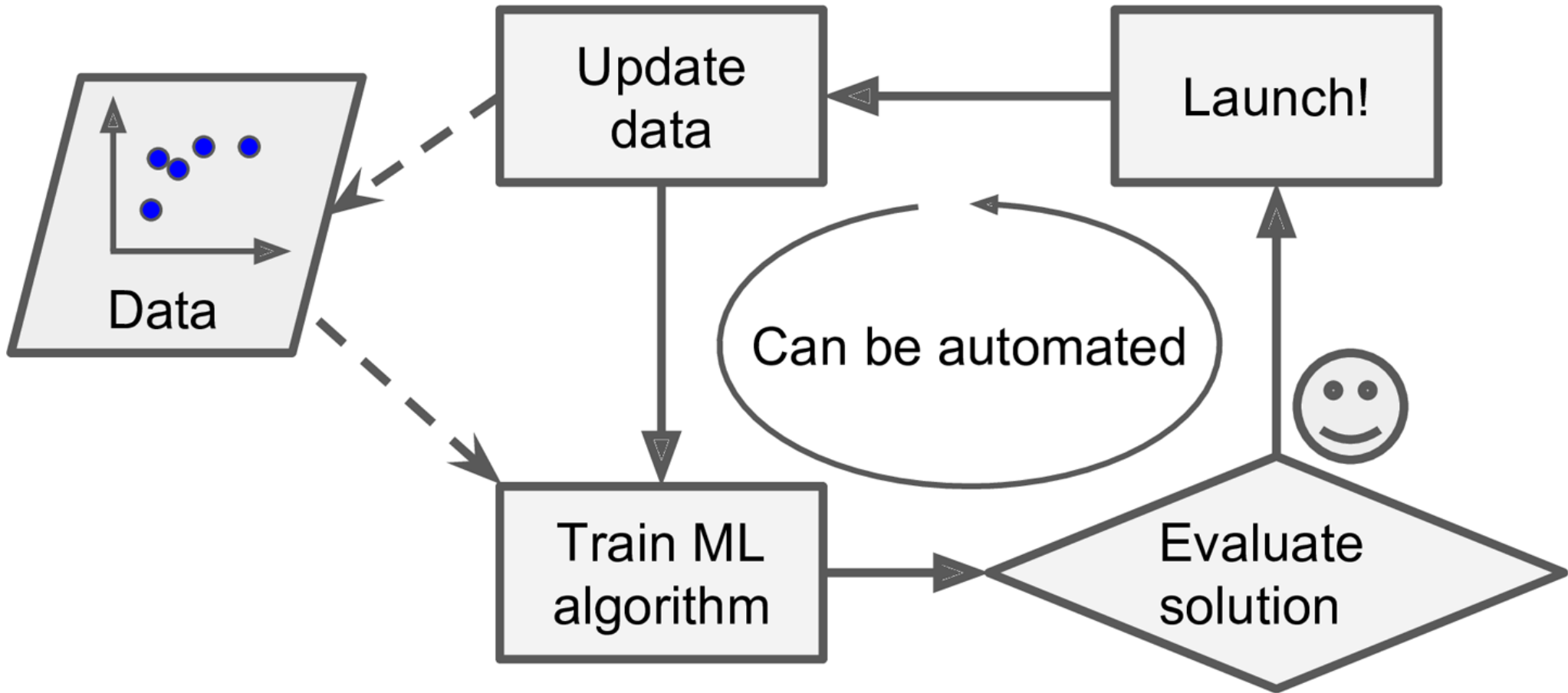
Traditional approach



ML approach



ML approach



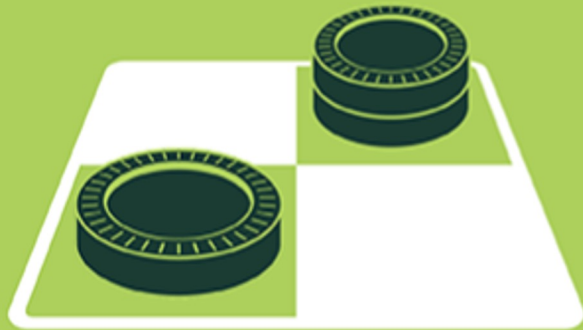
ML is good for:

- Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one machine learning algorithm can often simplify code and perform better.
- Complex problems for which there is no good solution at all using a traditional approach: the best machine learning techniques can find a solution.
- Fluctuating environments: a machine learning system can adapt to new data.
- Getting insights about complex problems and large amounts of data.

ML timeline

ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



MACHINE LEARNING

Machine learning begins to flourish.



DEEP LEARNING

Deep learning breakthroughs drive AI boom.



1950's

1960's

1970's

1980's

1990's

2000's

2010's

“A computer program is said to **learn** from *experience E* with respect to some *task T* and some *performance P*, if its performance on *T*, as measured by *P*, improves with *E*.”

-- **Tom Mitchell, Computer Scientist, 1997**

T = The thing we are trying to learn how to do.

P = The way we measure how well we're doing.

E = The data the algorithm will learn from.

Experience E: the data

- Most learning algorithms are allowed to experience a **dataset**, a collection of many **examples** or **data points** or **instances**.
- Each example or data point is a collection of one or more **features**: each **feature** is usually a number, or a string representing a category (like blood type).
- More complex data points are usually broken down. (e.g., speech, language, music, etc).

Performance P

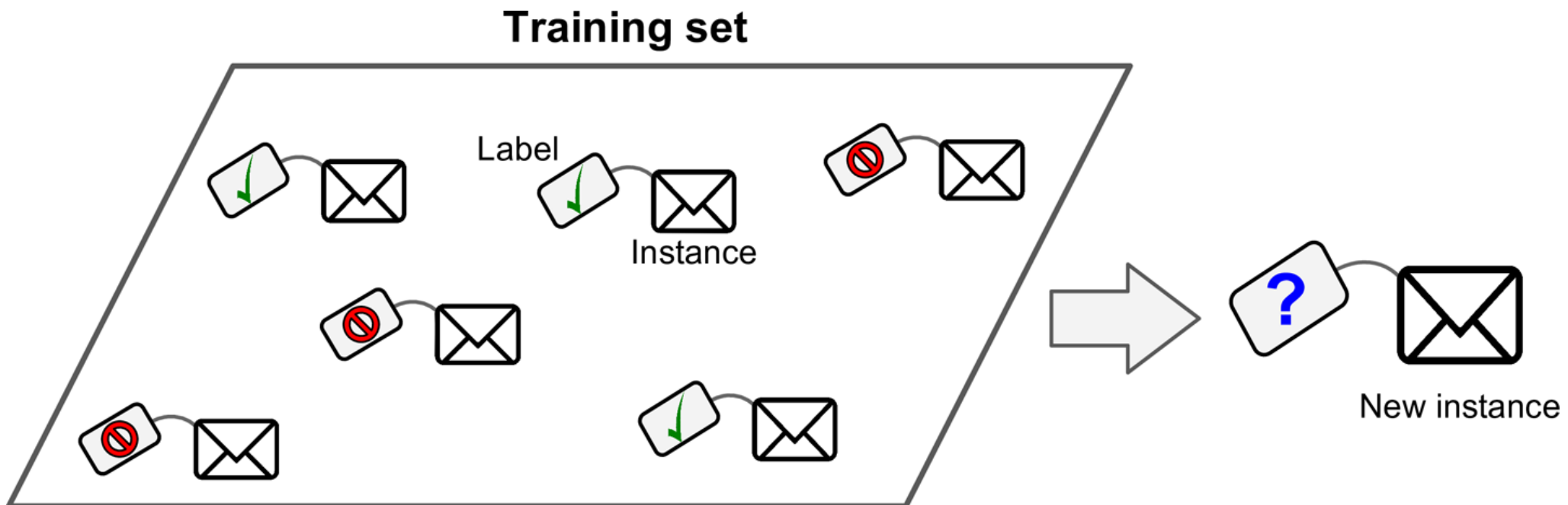
- We need some quantitative way to measure how well an ML algorithm is doing.
- Ideas:
 - Accuracy
 - Error
- Training set: set of examples used to train the ML algorithm.
- Testing set: set of examples used to evaluate (test) the ML algorithm.

Types of ML systems

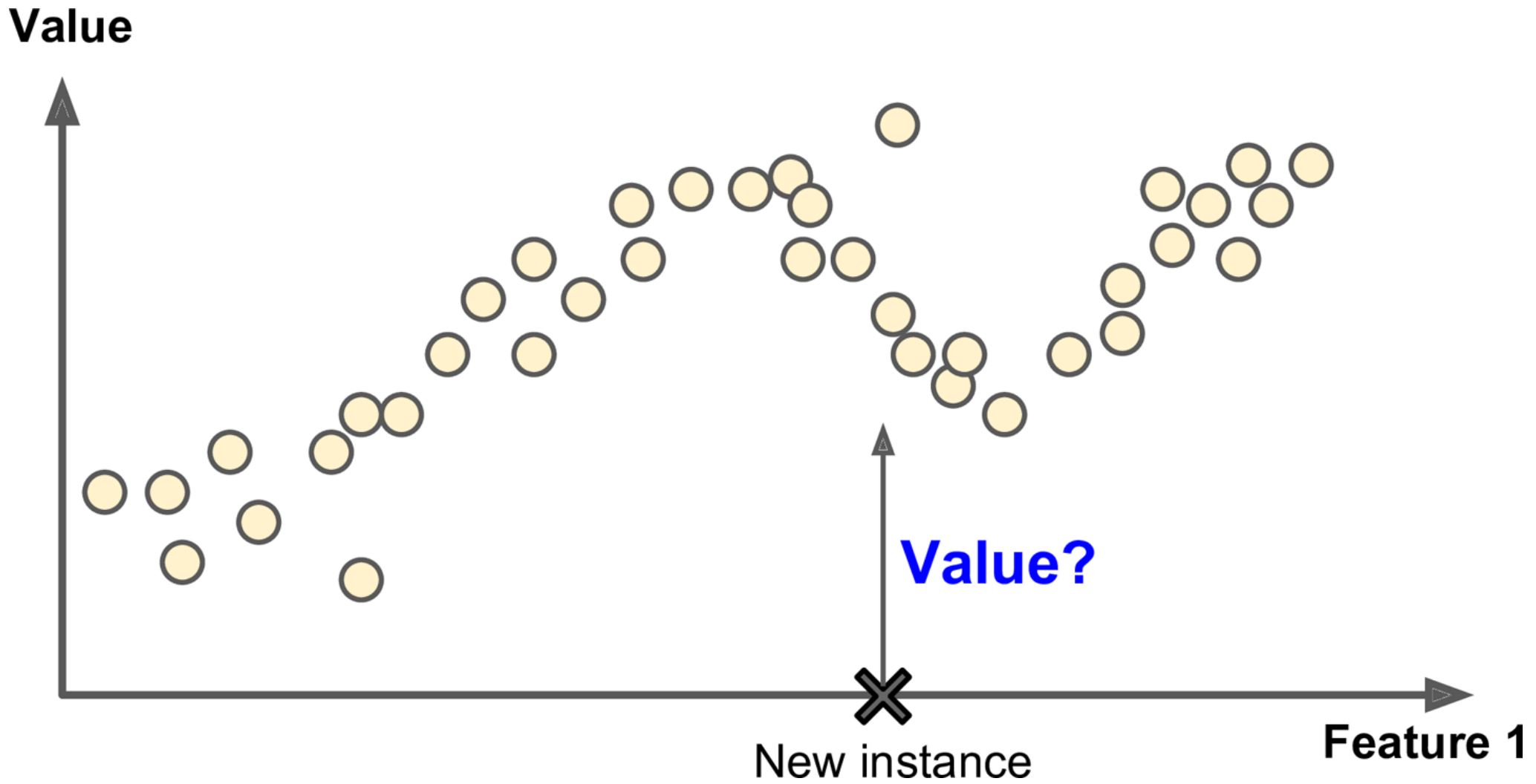
- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

Supervised learning

- Supervised learning algorithms receive data that is “labeled.”
- The labels are the desired solutions (outputs).
- Produces a system which can predict a new label for a previously-unseen input.



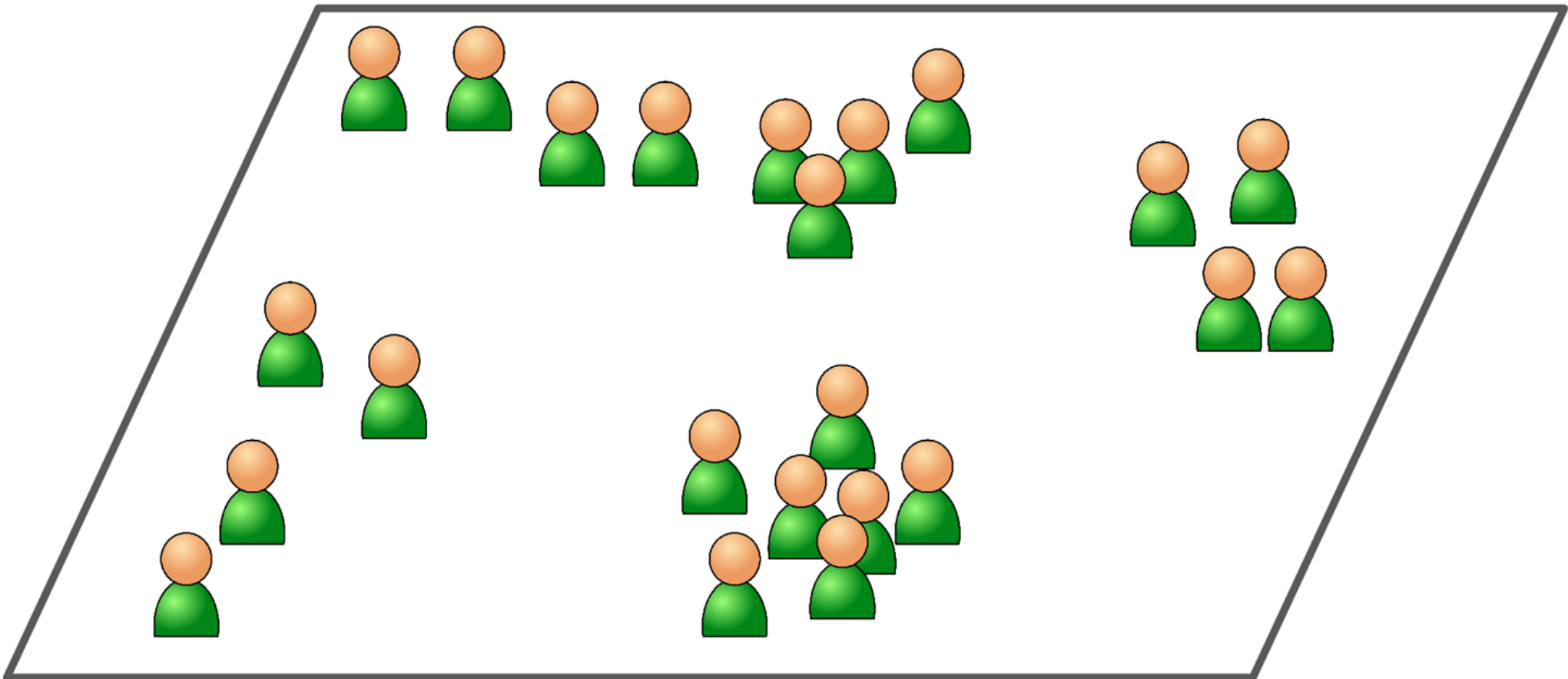
Regression



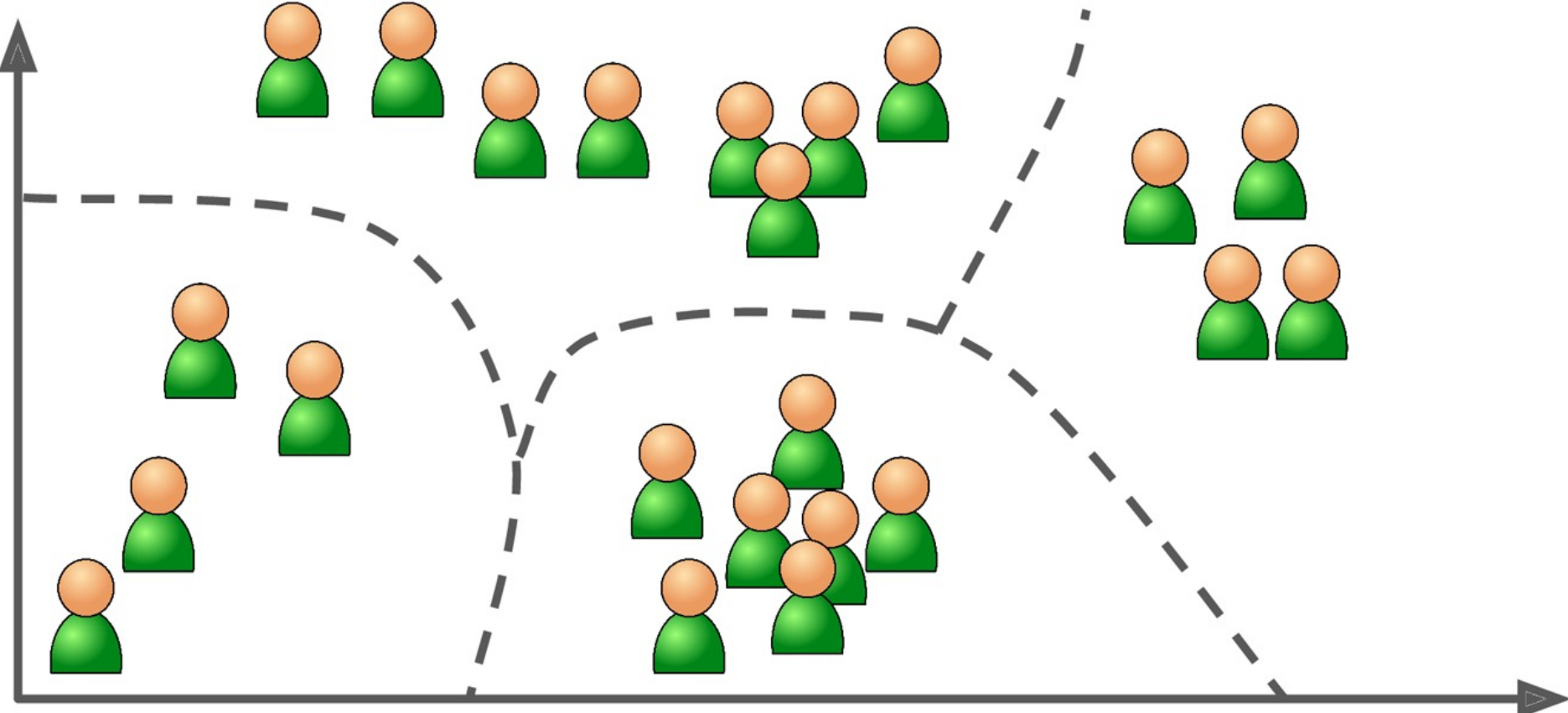
Unsupervised learning

- Unsupervised ML algorithms receive unlabeled data.

Training set

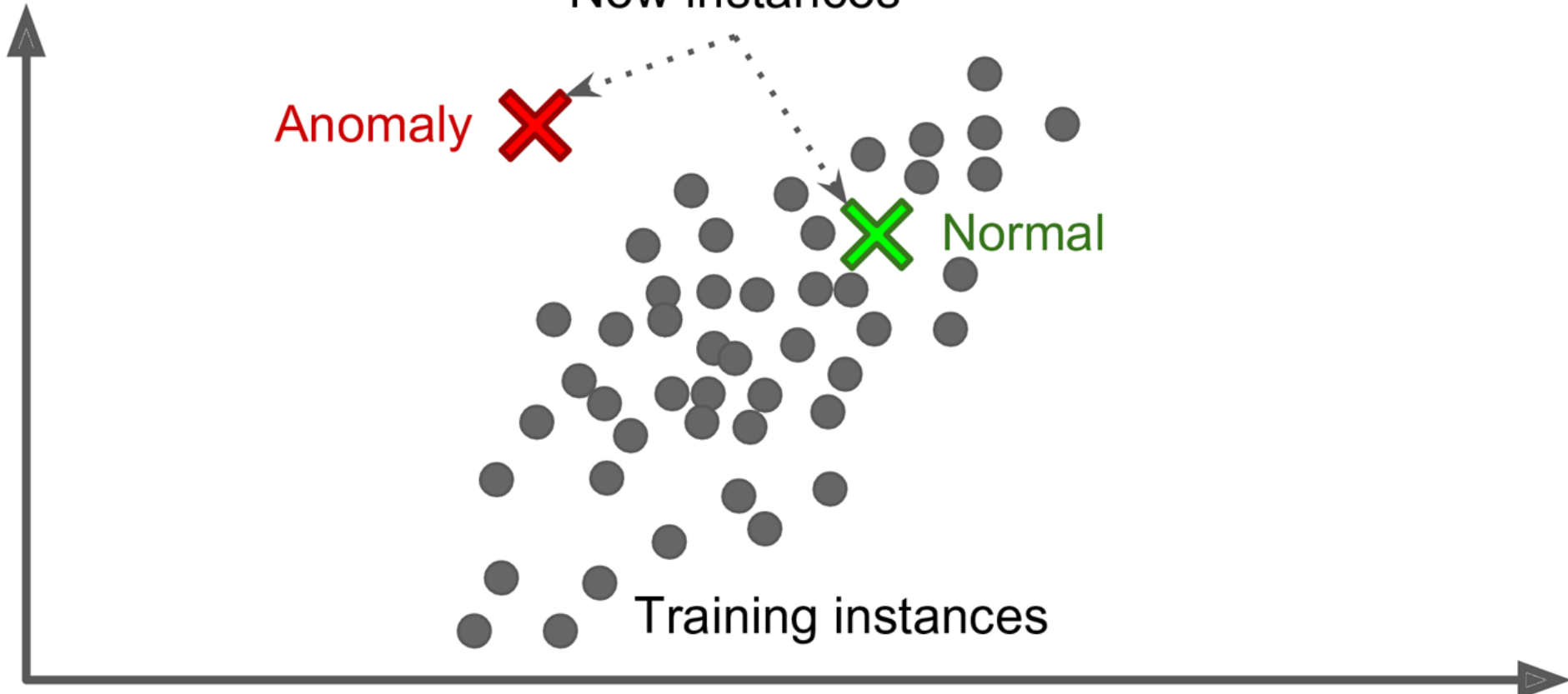


Feature 2



Feature 1

Feature 2



New instances

Anomaly



Normal

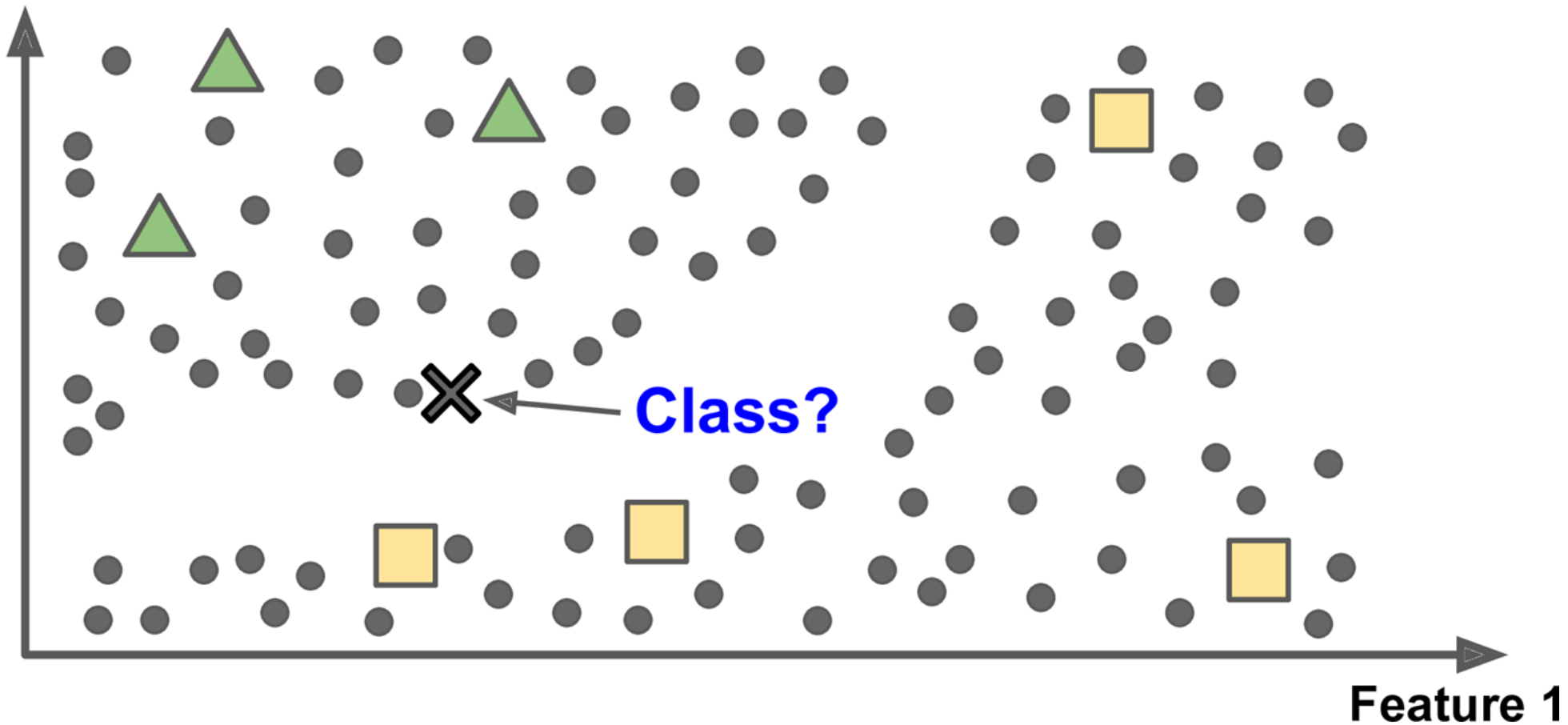
Training instances

Feature 1

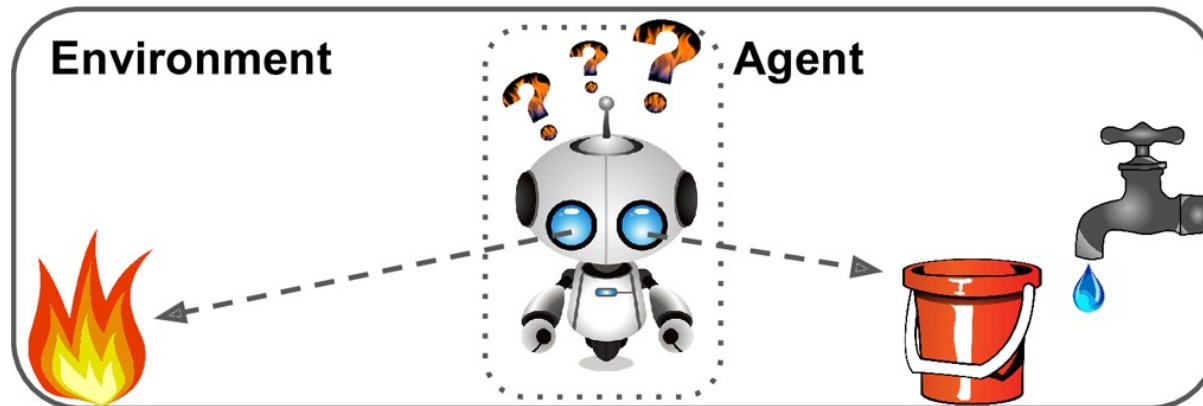
Semisupervised learning

- Receive partially-labeled data.

Feature 2

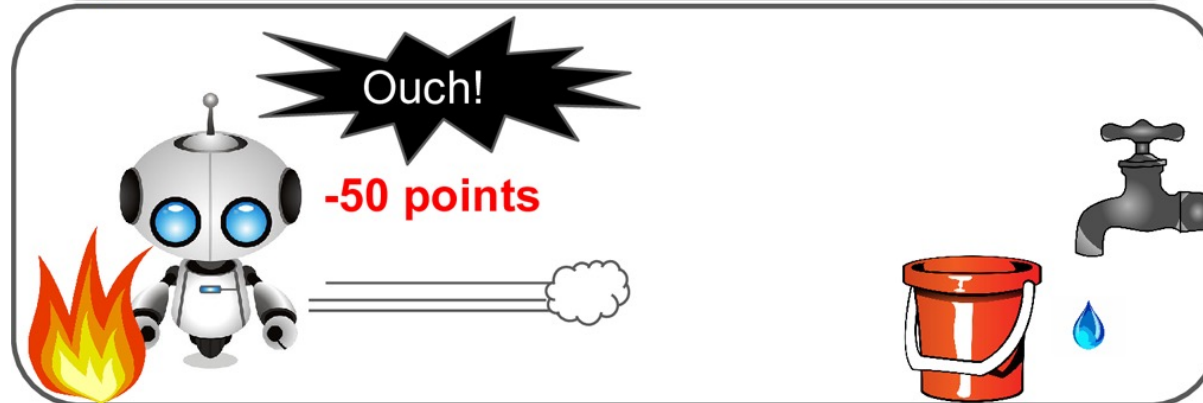


Reinforcement learning



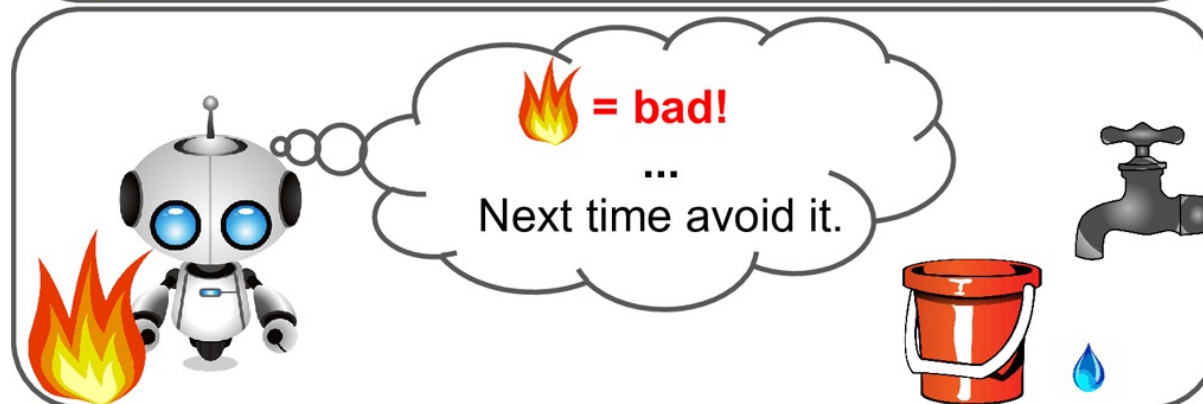
1 Observe

2 Select action using policy



3 Action!

4 Get reward or penalty



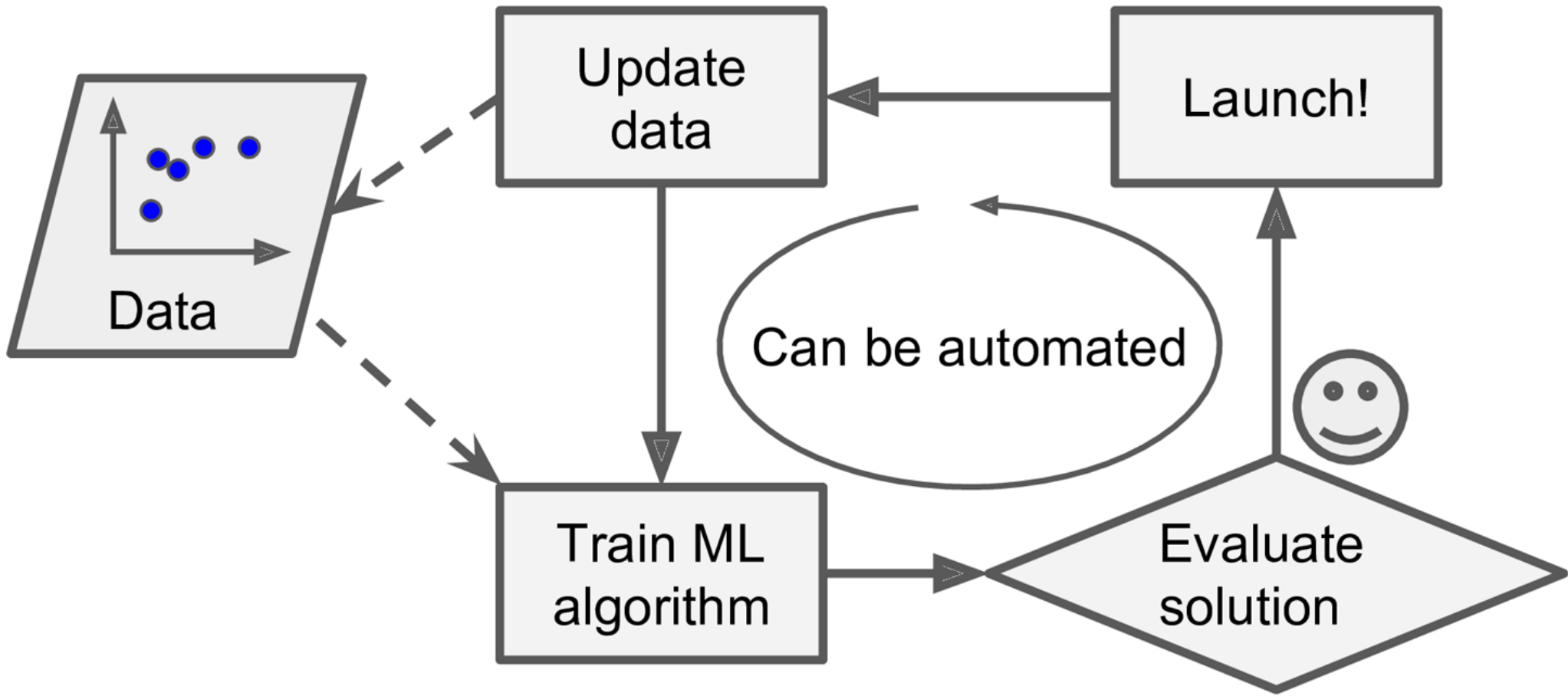
5 Update policy (learning step)

6 Iterate until an optimal policy is found

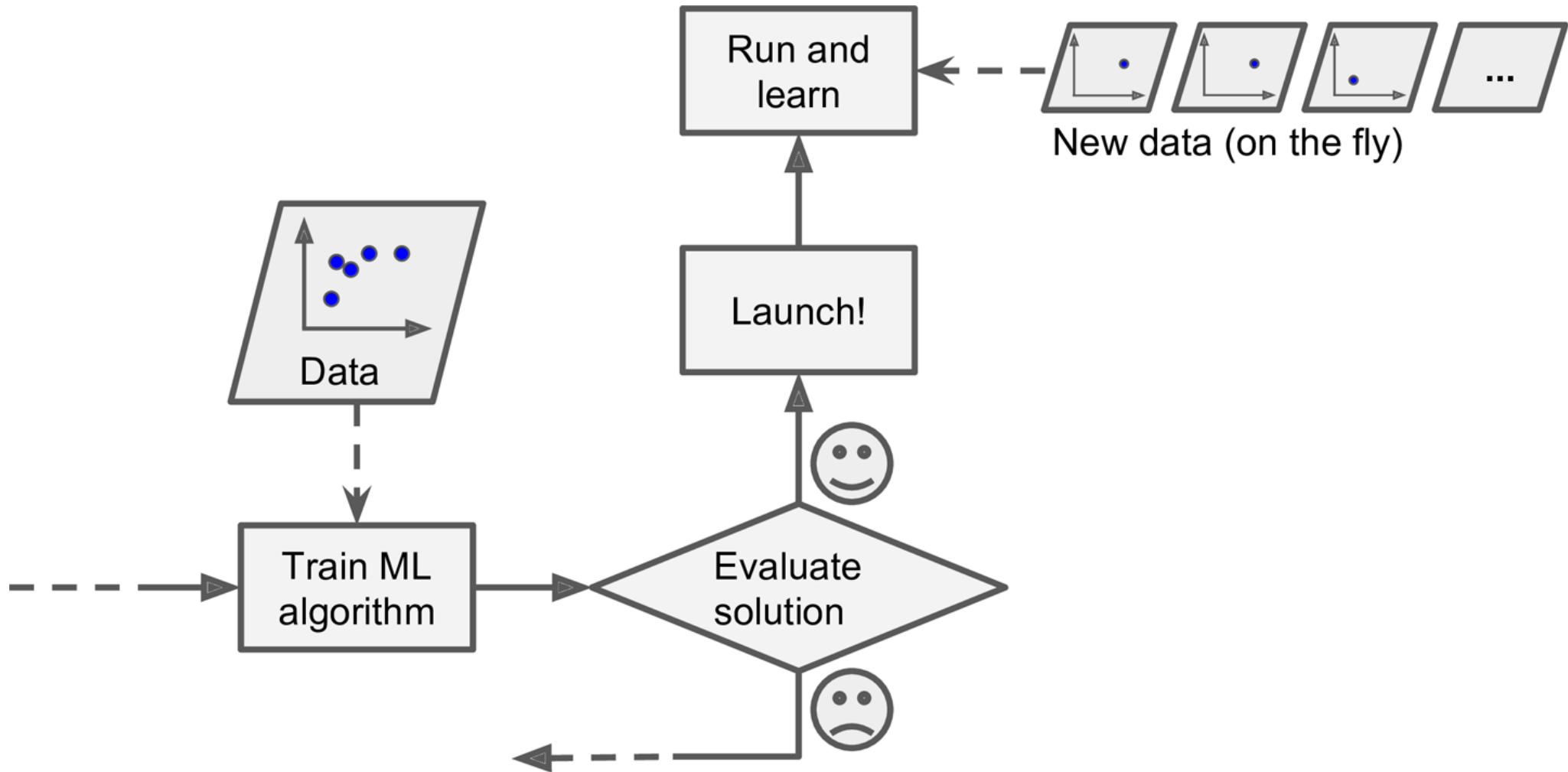
Batch vs online learning

- Batch learning:
 - Algorithm trains on all the data at once, and after this is done, new data cannot be incorporated in.
 - If new data becomes available, the algorithm must be trained from scratch.
- Online learning:
 - Algorithm can be trained incrementally, so if new data arrives, it can be incorporated without starting from scratch.

Batch learning



Online learning

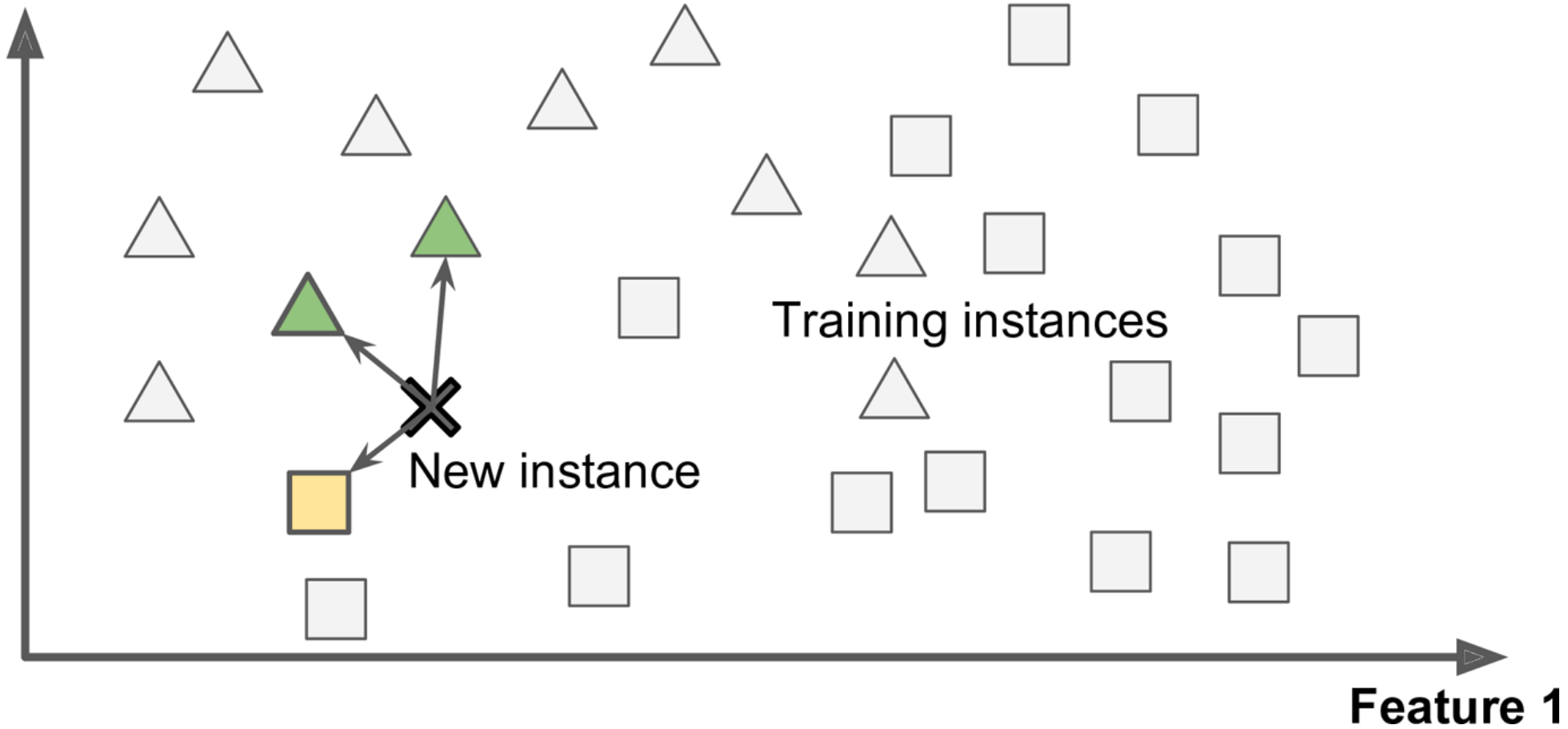


Instance-based vs model-based

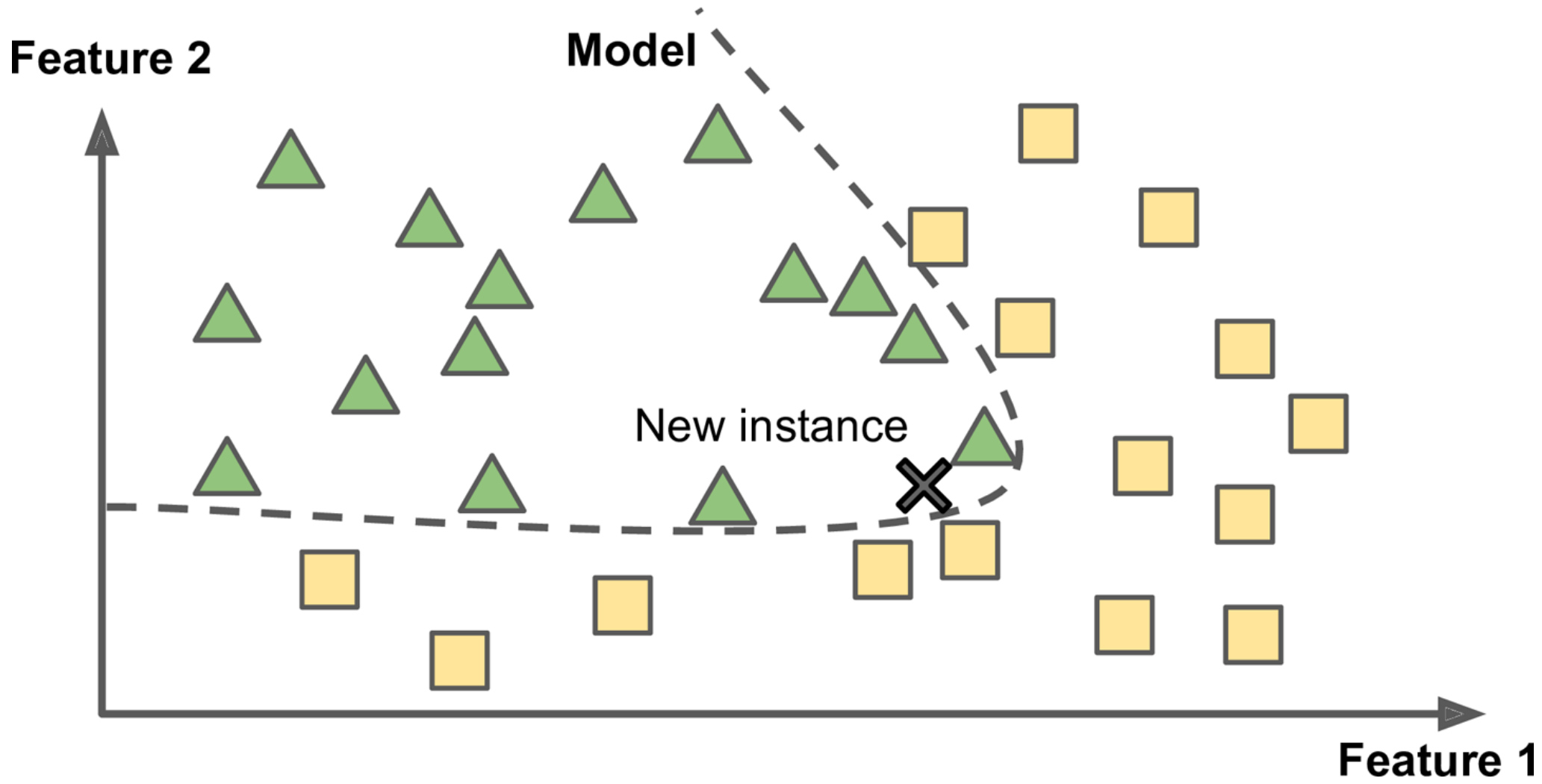
- Instance-based:
 - An algorithm that makes new predictions by analyzing the training data (and comparing the data that we want a prediction about against the training data).
 - The training data must be kept in memory to make predictions.
- Model-based:
 - An algorithm that builds a model of the training data and uses that model to make predictions.
 - The training data can be discarded after the model is created.

Instance-based

Feature 2

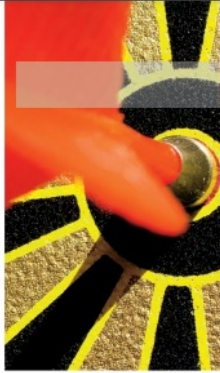


Model-based



Main challenges of ML

- Insufficient quantity of training data
- Non-representative training data
- Poor-quality data
- Irrelevant features
- Overfitting the training data
- Underfitting the training data



EXPERT OPINION

Contact Editor: **Brian Brannon**, bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, Google

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"¹ examines why so much of physics can be neatly explained with simple mathematical formulas

such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that involve human beings rather than elementary particles have proven more resistant to elegant mathematics. Economists suffer from physics envy over their inability to neatly model human behavior. An informal, incomplete grammar of the English language runs over 1,700 pages.² Perhaps when it comes to natural language processing and related fields, we're doomed to complex theories that will never have the elegance of physics equations. But if that's so, we should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data.

One of us, as an undergraduate at Brown University, remembers the excitement of having access to the Brown Corpus, containing one million English words.³ Since then, our field has seen several notable corpora that are about 100 times larger, and in 2006, Google released a trillion-word corpus with frequency counts for all sequences up to five words long.⁴ In some ways this corpus is a step backwards from the Brown Corpus: it's taken from unfiltered Web pages and thus contains incomplete sentences, spelling errors, grammatical errors, and all sorts of other errors. It's not annotated with carefully hand-corrected part-of-speech tags. But the fact that it's a million times larger than the Brown Corpus outweighs these drawbacks. A trillion-word corpus—along with other Web-derived corpora of millions, billions, or trillions of links, videos, images, tables, and user interactions—captures even very rare aspects of human

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The reason for these successes is not that these tasks are easier than other tasks; they are in fact much harder than tasks such as document classification that extract just a few bits of information from each document. The reason is that translation is a natural task routinely done every day for a real human need (think of the operations of the European Union or of news agencies). The same is true of speech transcription (think of closed-caption broadcasts). In other words, a large training set of the input-output behavior that we seek to automate is available to us *in the wild*. In contrast, traditional natural language processing problems such as document classification, part-of-speech tagging, named-entity recognition, or parsing are not routine tasks, so they have no large corpus available in the wild. Instead, a corpus for these tasks requires skilled human annotation. Such annotation is not only slow and expensive to acquire but also difficult for experts to agree on, being bedeviled by many of the difficulties we discuss later in relation to the Semantic Web. The first lesson of Web-scale learning is to use available large-scale data rather than hoping for annotated data that isn't available. For instance, we find that useful semantic relationships can be automatically learned from the statistics of search queries and the corresponding results⁵ or from the accumulated evidence of Web-based text patterns and formatted tables,⁶ in both cases without needing any manually annotated data.

Main challenges of ML

- Insufficient quantity of training data
- Non-representative training data
- Poor-quality data
- Irrelevant features
- Overfitting the training data
- Underfitting the training data

1936 Presidential Election

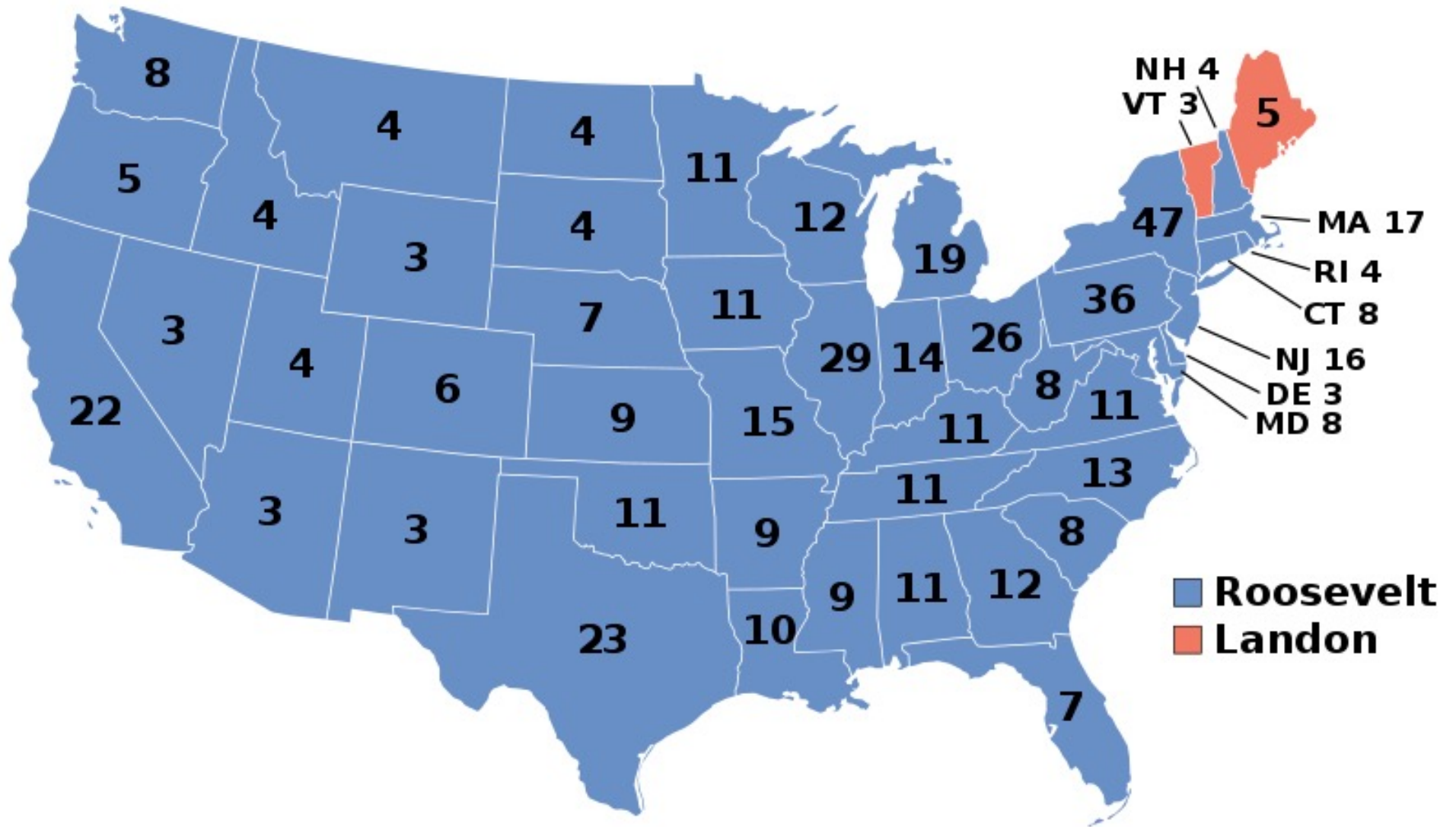


Franklin D. Roosevelt



Alf Landon

This election is notable for [*The Literary Digest*](#) poll, which was based on ten million questionnaires mailed to readers and potential readers; 2.27 million were returned. The *Literary Digest* had correctly predicted the winner of the last five elections, and announced in its October 31 issue that Landon would be the winner with 57.1% of the vote (v Roosevelt) and 370 electoral votes.



Main challenges of ML

- Insufficient quantity of training data
- Non-representative training data
- Poor-quality data
- Irrelevant features
- Overfitting the training data
- Underfitting the training data

Testing and Validation

- Training set and testing set
- Validation set
- Cross-validation