Alchemy and AI

Michael Wollowski

Summary of the paper with the same title written by Hubert Dreyfus

Paper source: https://www.rand.org/content/dam/rand/pubs/papers/2006/P3244.pdf

Four Areas of Stagnation of Classical AI

- Dreyfus notes four areas of stagnation of classical AI:
  - Game playing,
  - Problem solving,
  - Language translation, and
  - Pattern recognition.
- We note that NN-based systems outperform humans in all but one area, namely that of problem solving.
- Pattern recognition is somewhat of an all-encompassing activity, and in many ways underlies the three other areas.
- For now, we point out that AlphaGoZero uses pattern recognition in the evaluation of board positions, and large-language models look at the patterns in sentences.

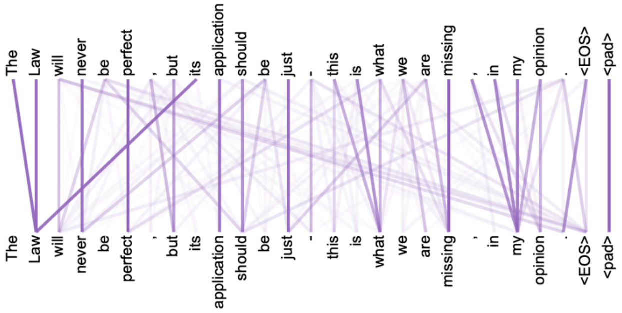
Features of Pattern Recognition

- Dreyfus: “The pattern may be skewed, incomplete, deformed and embedded in noise.”
- Within reason, this is something that CNNs as well as Transformers can process.
- In the case of Transformers, we are thinking of misspelled words, phrases, as well as sentences that do not follow the rules of grammar.

Features of Pattern Recognition

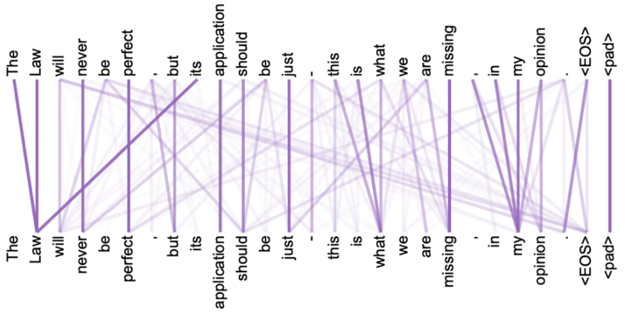
- Dreyfus: “The traits required for recognition may be ‘so fine and so numerous’ that, even if they could be formalized, a search through a branching list of such traits would soon become unmanageable as new patterns for discrimination were added.”

- NNs do not use lists of traits, and they do not use search.
- A Transformer processes an input based on training over many sentences and creates relationships between the words in a sentence.


The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>

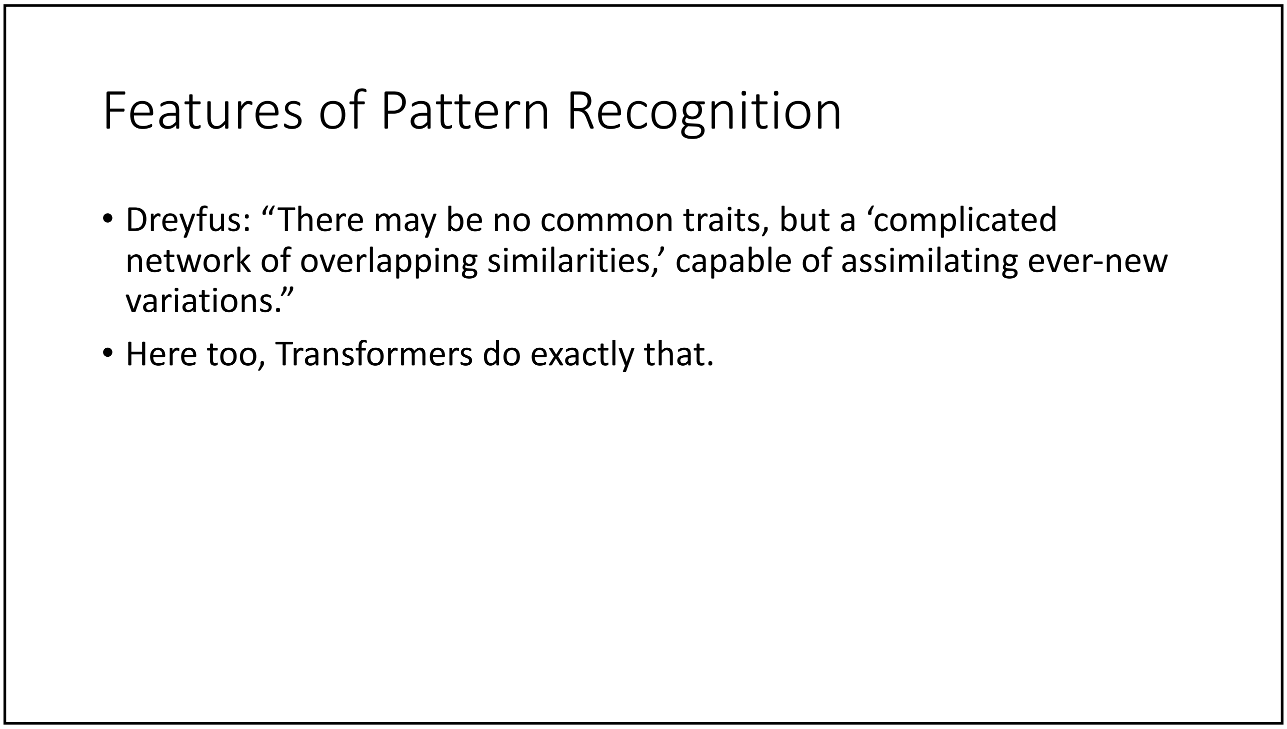
Features of Pattern Recognition

- The thickness of the lines indicate the strengths of the connections.
- We see that Transformers attend to the different token in a sentence with varying degrees.
- We also point out that below is a pattern for just one head, Transformers have many heads.



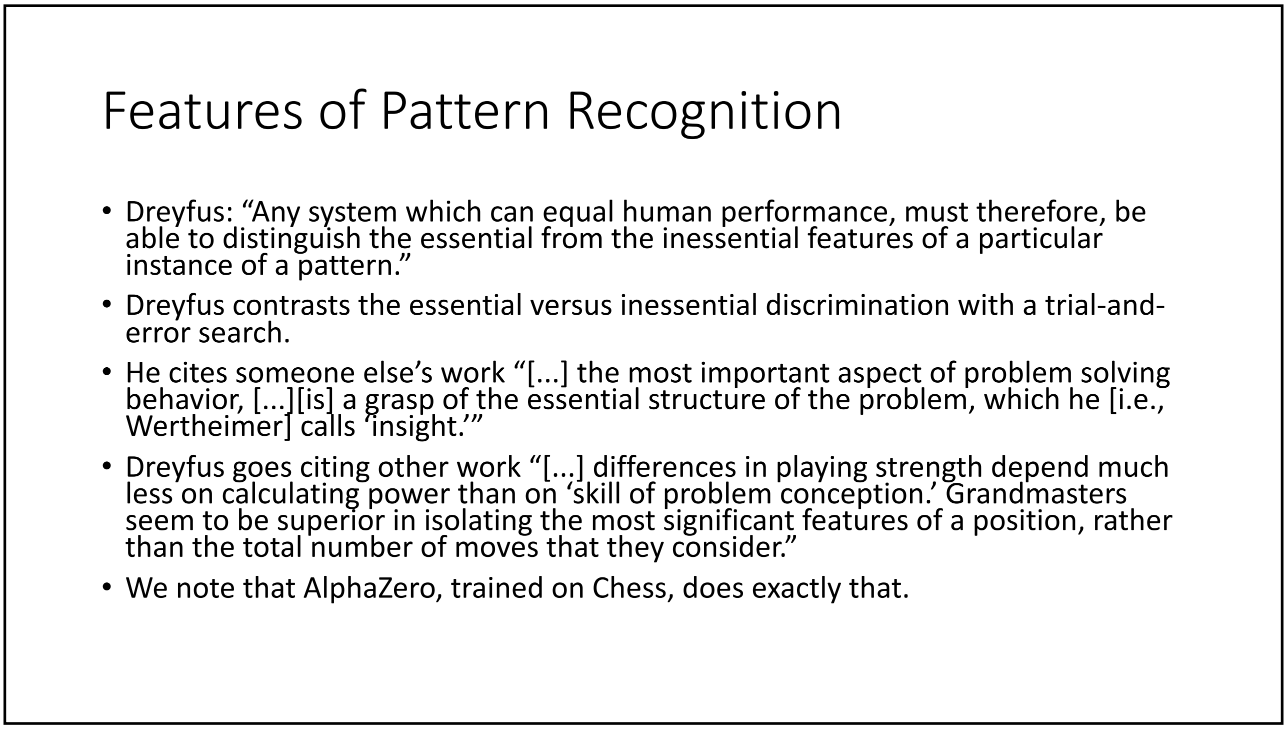
Features of Pattern Recognition

- Dreyfus: “The traits may depend upon internal and external context and are thus not isolable into lists.”
- Again, NNs do not use lists.
- The multiple heads of a Transformer are very capable of picking up multiple contexts within a sentence.



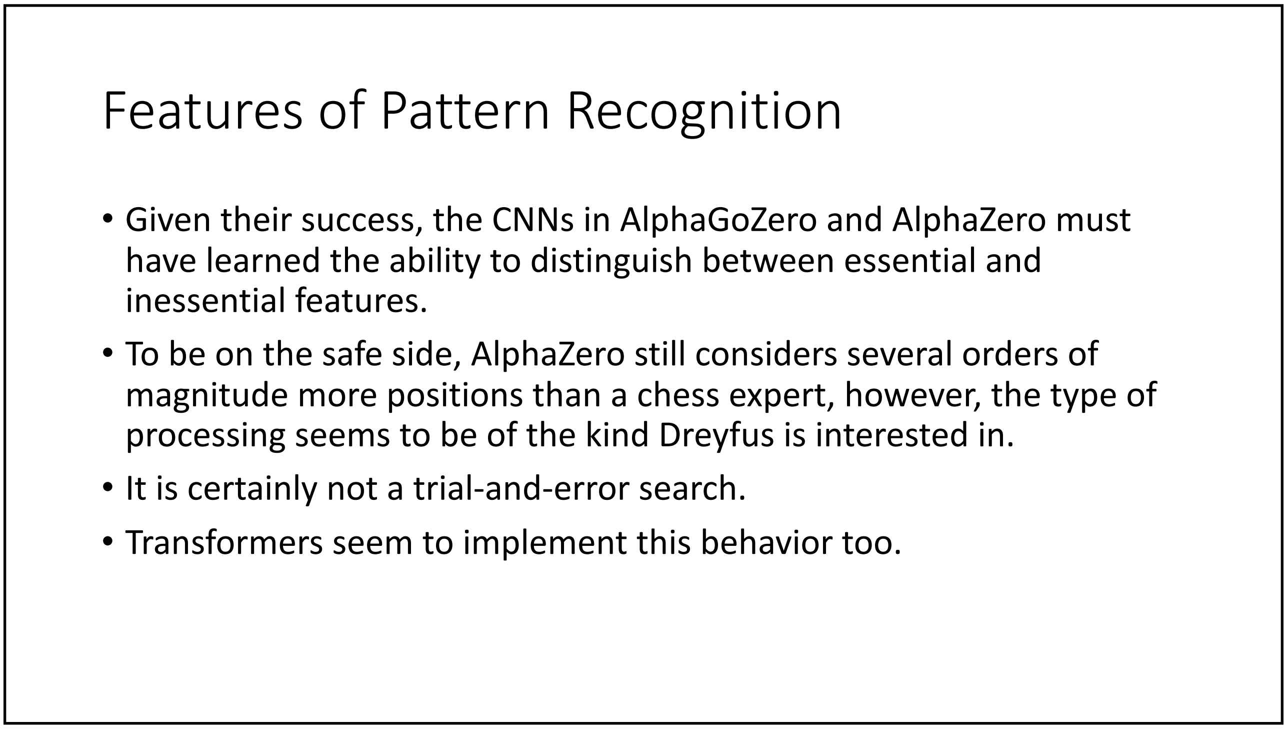
Features of Pattern Recognition

- Dreyfus: “There may be no common traits, but a ‘complicated network of overlapping similarities,’ capable of assimilating ever-new variations.”
- Here too, Transformers do exactly that.



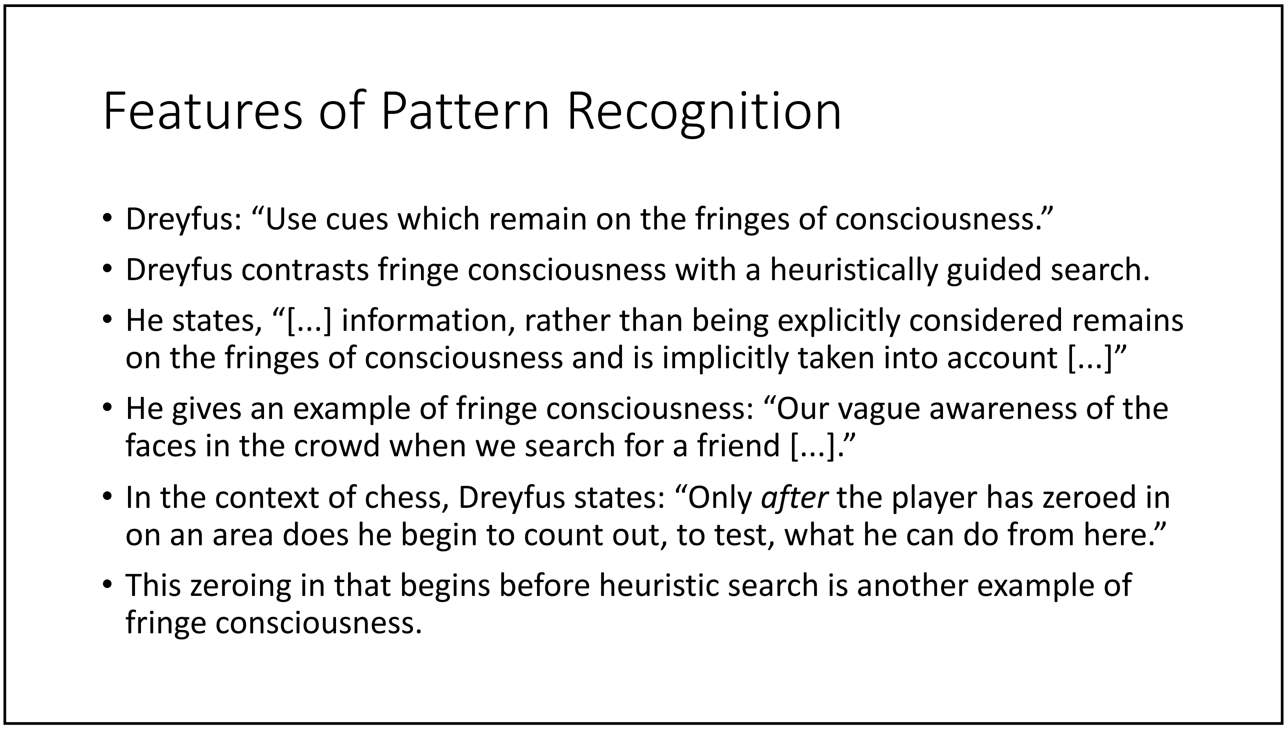
Features of Pattern Recognition

- Dreyfus: “Any system which can equal human performance, must therefore, be able to distinguish the essential from the inessential features of a particular instance of a pattern.”
- Dreyfus contrasts the essential versus inessential discrimination with a trial-and-error search.
- He cites someone else’s work “[...] the most important aspect of problem solving behavior, [...][is] a grasp of the essential structure of the problem, which he [i.e., Wertheimer] calls ‘insight.’”
- Dreyfus goes citing other work “[...] differences in playing strength depend much less on calculating power than on ‘skill of problem conception.’ Grandmasters seem to be superior in isolating the most significant features of a position, rather than the total number of moves that they consider.”
- We note that AlphaZero, trained on Chess, does exactly that.



Features of Pattern Recognition

- Given their success, the CNNs in AlphaGoZero and AlphaZero must have learned the ability to distinguish between essential and inessential features.
- To be on the safe side, AlphaZero still considers several orders of magnitude more positions than a chess expert, however, the type of processing seems to be of the kind Dreyfus is interested in.
- It is certainly not a trial-and-error search.
- Transformers seem to implement this behavior too.



Features of Pattern Recognition

- Dreyfus: “Use cues which remain on the fringes of consciousness.”
- Dreyfus contrasts fringe consciousness with a heuristically guided search.
- He states, “[...] information, rather than being explicitly considered remains on the fringes of consciousness and is implicitly taken into account [...]”
- He gives an example of fringe consciousness: “Our vague awareness of the faces in the crowd when we search for a friend [...].”
- In the context of chess, Dreyfus states: “Only *after* the player has zeroed in on an area does he begin to count out, to test, what he can do from here.”
- This zeroing in that begins before heuristic search is another example of fringe consciousness.

Features of Pattern Recognition

- He states: “[...] we can conclude that our subject’s familiarity with the overall chess pattern and with the past moves of this particular game enabled him to recognize the lines of force, the loci of strength and weakness, as well as specific positions.”
- Dreyfus quotes someone else about the sort of pattern recognition developed by an experienced chess player: “Because of the large number of prior associations which an experienced player has acquired, he does not visualize a chess position as a conglomerate of scattered squares and wooden pieces, but has an organized pattern [...]”.
- Fringe consciousness, while perhaps not the best term in this context, is what we see in the patterns that are learned in the CNN of AlphaGoZero as well as AlphaZero; those systems seem to be able to pay attention not just to a few key pieces but the entire board.
- Yet, they do not do this by searching through lists, but by returning a list of potentially successful moves to consider.

Features of Pattern Recognition

- Dreyfus: “Take account of the context.”
- Dreyfus contrasts ambiguity tolerance with context-free precision.
- He states: “Fringe consciousness takes account of the cues in the context, and probably some possible parsings and meanings, all of which would have to be made explicit in the output of a machine. Our sense of the situation then allows us to exclude most of these possibilities without explicit consideration. We shall call the ability to narrow down the spectrum of possible meanings as much as the situation requires ‘ambiguity tolerance’.”
- He goes on to require: “[...] deal with situations which are ambiguous without having to transform them by substituting a precise description.”

Features of Pattern Recognition

- The processing of an input by a Transformer is informed by a lengthy and extensive training period.
- The processing of the input by several heads causes the system to “take account of the cues in the context,” as Dreyfus writes.
- Certainly, the cues are not explicitly written down in lists but are captured in the weights of a very large multi-dimensional vector which are brought to bear when processing the input.

Features of Pattern Recognition

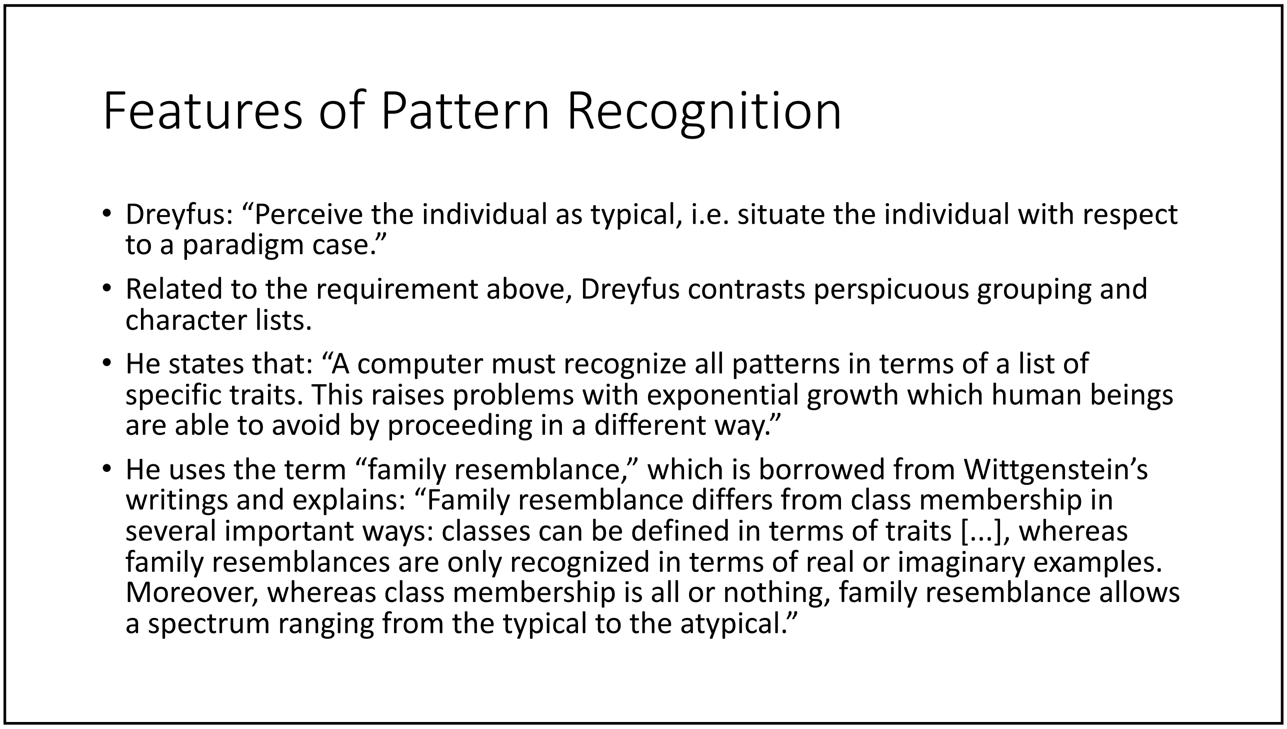
- Dreyfus considers the challenges of natural language translation: “[...] has realized that in order to translate a natural language, more is needed than a mechanical dictionary – no matter how complete – and the laws of grammar - no matter how sophisticated. The order of the words in a sentence does not provide enough information to enable a machine to determine which of several possible parsings is the appropriate one, nor do the surrounding words – the written context – always indicate which of several meanings is the one the author had in mind.”
- It is not clear what Dreyfus has in mind to ensure the grasp of a particular meaning that an author had in mind.
- Perhaps, he means common sense, or a reflection back to the use of a word in a prior portion of the writings.

Features of Pattern Recognition

- Transformers do not use dictionaries or grammar rules, instead they build internal representations by learning from an incredibly large training set.
- Just like AlphaGoZero learns the patterns that are essential to successful play, LLMs learn patterns of language, taking into account several aspects of a sentence and its context.
- They do this for language that is ambiguous and without transforming ambiguity into a precise description.
- Oh yet, recall the NYTimes article on Google’s Translate.

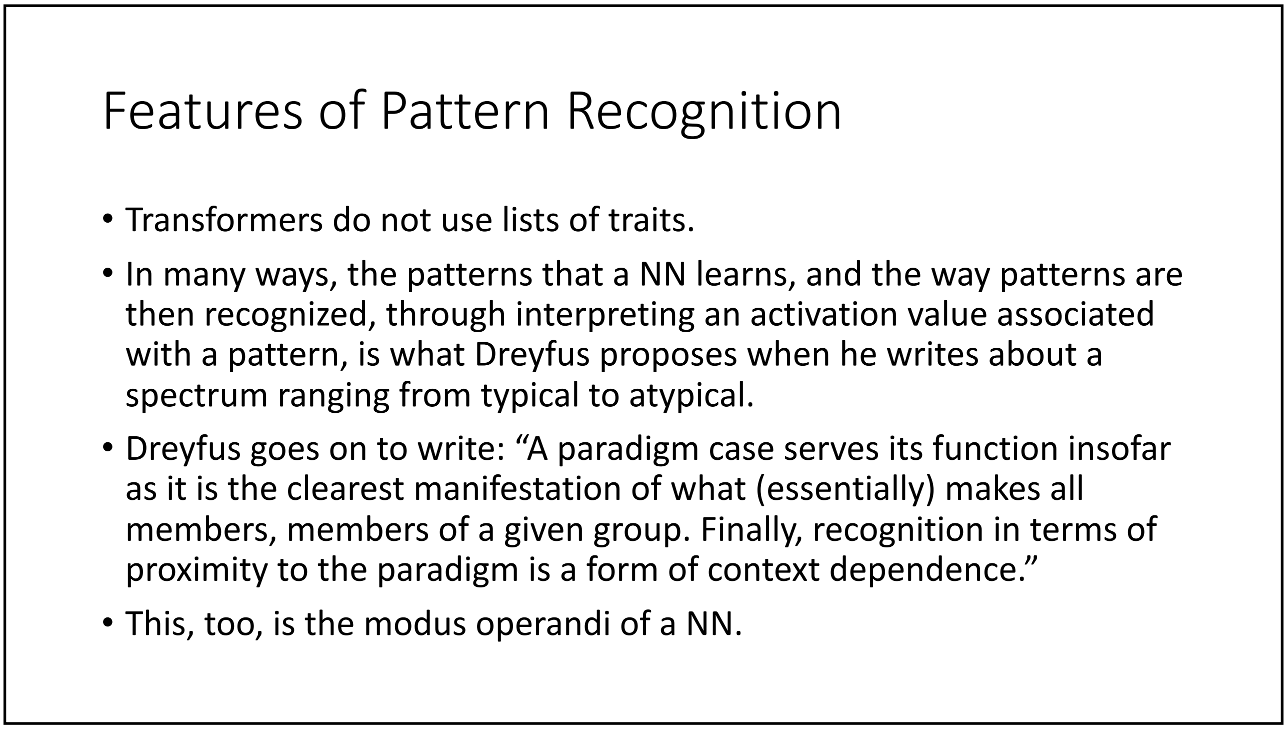
Features of Pattern Recognition

- Dreyfus states: “Since a human being using and understanding a sentence in a natural language requires an implicit knowledge of the sentence’s context-dependent use, the only way to make a computer that could understand and translate a natural language may well be […] to program it to learn about the world.”
- Learning about the world is what the creators of ChatGPT had it do, by ingesting publicly available writings and a lot of it.
- ChatGPT seems to have learned quite a bit about the way the world works.
- We should point out that embeddings, too, contain a fair amount of knowledge about the world.



Features of Pattern Recognition

- Dreyfus: “Perceive the individual as typical, i.e. situate the individual with respect to a paradigm case.”
- Related to the requirement above, Dreyfus contrasts perspicuous grouping and character lists.
- He states that: “A computer must recognize all patterns in terms of a list of specific traits. This raises problems with exponential growth which human beings are able to avoid by proceeding in a different way.”
- He uses the term “family resemblance,” which is borrowed from Wittgenstein’s writings and explains: “Family resemblance differs from class membership in several important ways: classes can be defined in terms of traits [...], whereas family resemblances are only recognized in terms of real or imaginary examples. Moreover, whereas class membership is all or nothing, family resemblance allows a spectrum ranging from the typical to the atypical.”



Features of Pattern Recognition

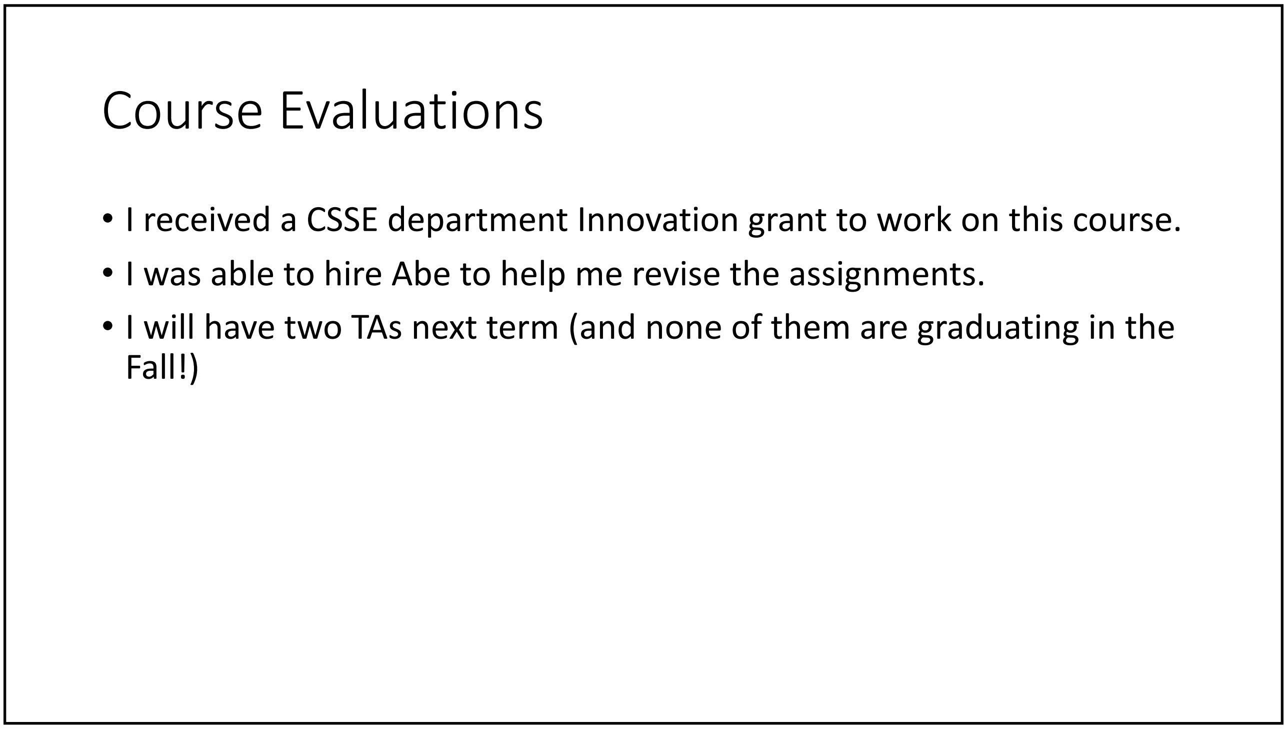
- Transformers do not use lists of traits.
- In many ways, the patterns that a NN learns, and the way patterns are then recognized, through interpreting an activation value associated with a pattern, is what Dreyfus proposes when he writes about a spectrum ranging from typical to atypical.
- Dreyfus goes on to write: “A paradigm case serves its function insofar as it is the clearest manifestation of what (essentially) makes all members, members of a given group. Finally, recognition in terms of proximity to the paradigm is a form of context dependence.”
- This, too, is the modus operandi of a NN.

Course Evaluations

- What I plan to do:
- Beef up the NN assignment:
  - Make instructions more explicit.
  - Implement FF from scratch
  - Revise experiments
  - Get rid of parity but portion
- Beef up CNN assignment:
  - Implement LeNet 5 from scratcj
- Get rid of ChatGPT experiments
- Possibly implement a small Transformer architecture

Course Evaluations

- Swap out “What kind of Mind does ChatGPT have?” article.
- Add data curation, preparation, selection to some of the assignments.
- Have learning materials on data prepartion.
- Add learning materials for:
  - NN
  - CNNs
  - Transformers
- Discuss more assignment details in class
- Add learning materials specific to the assignments
- Add three quizzes, align them with the prior three items.



Course Evaluations

- I received a CSSE department Innovation grant to work on this course.
- I was able to hire Abe to help me revise the assignments.
- I will have two TAs next term (and none of them are graduating in the Fall!)