Where we’ve been, where we’re going

- Two classes ago: classic model of learning concepts based on combinations of features.
  - [Is it *really* learning? Fodor’s challenge.]
  - Theories of when learning is possible:
    - Identifiability in the limit: Subset principle
    - PAC: The importance of choosing a good representation -- choosing good features (c.f., Goodman, sinusoidal interpolation, etc.)

Where we've been, where we're going

- Last class (TA):
  - Varieties of representations:
    - Features
    - Multidimensional vector spaces
    - Trees (hierarchically nested features/classes)
  - How does a learner construct good representations? A cognitive modeler?
  - Algorithms for *compression* on similarities or coarse features dimensions.
    - Multidimensional scaling / PCA
    - Hierarchical clustering
    - Additive clustering

Where we’ve been, where we’re going

- For this week:
  - “Structured” representations:
    - Grammars
    - First-order logic
  - What do structured representations offer?
    - How different from “unstructured” reps?
    - What is the evidence that human cognition uses structured representations?
  - How are structured representations constructed?
    - Compression algorithms for inducing grammars and first-order logical theories.
    - An answer to Fodor’s challenge?

Where we’ve been, where we’re going

- Next few weeks:
  - Learning and inference with probabilistic models.
  - Applications to categorization, unsupervised learning, semi-supervised learning.
- Then:
  - Causal models for categorization, learning and reasoning.
- After that:
  - Integrating probabilistic learning and inference with structured representations.
  - Applications to modeling intuitive theories of biology, physics, psychology.

Outline for today

- PAC learning analyses, and the importance of choosing good features.
- Grammars
- First-order logic
- Learning a theory and new concepts in first-order logic.

Computational analysis

- Can learning succeed under weaker conditions? PAC.
  - The true concept is not in the hypothesis space.
  - We are not willing to wait for infinite data, but we can live with a nonzero error rate.
- What can we say about more complex cases of learning?
  - Richer hypothesis spaces?
  - More powerful learning algorithms?

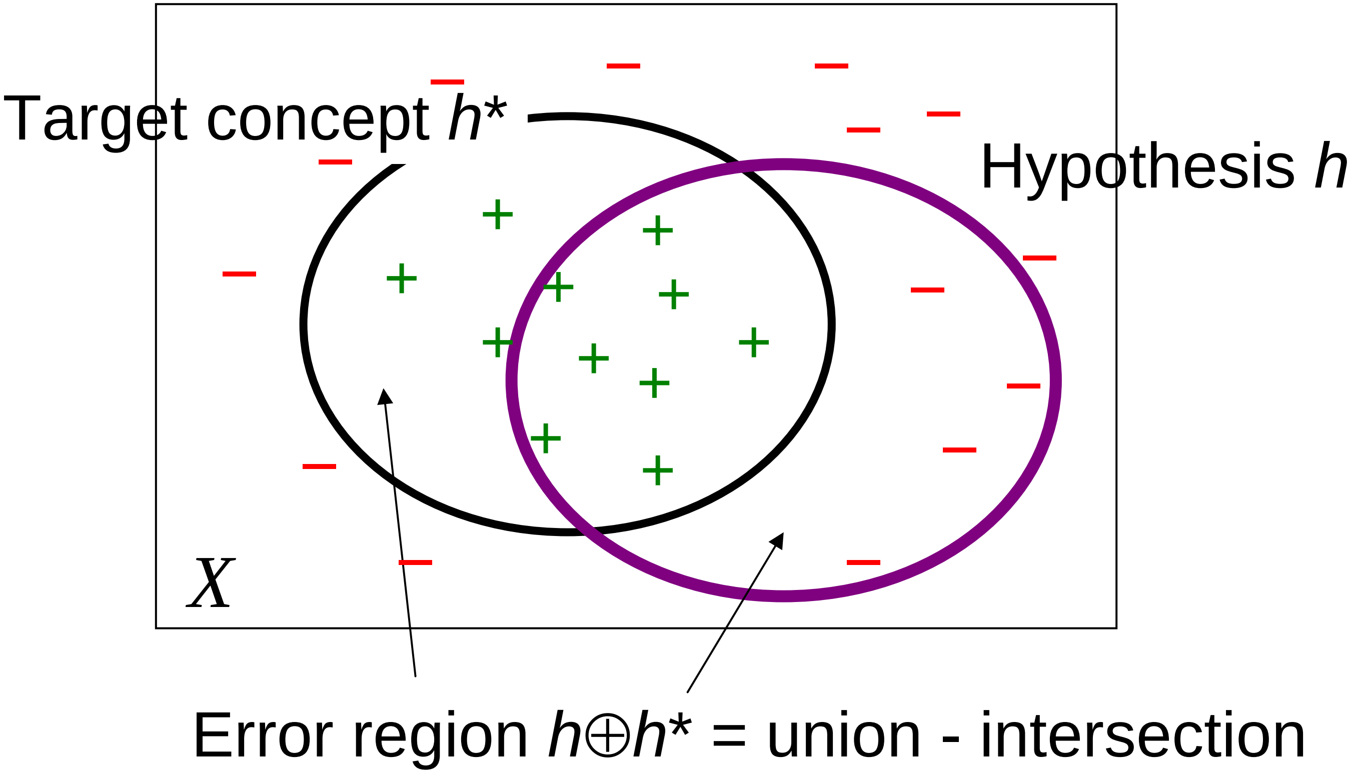
Probably Approximately Correct (PAC)

- The intuition: Want to be confident that a hypothesis which looks good on the training examples (i.e., appears to have zero error rate) in fact has a low true error rate (ε), and thus will generalize well on the test instances.

- Note we do not require that the true rule is in the hypothesis space, or that if it is, we must find it. We are willing to live with a low but nonzero error rate, as long as we can be pretty sure that it is low.

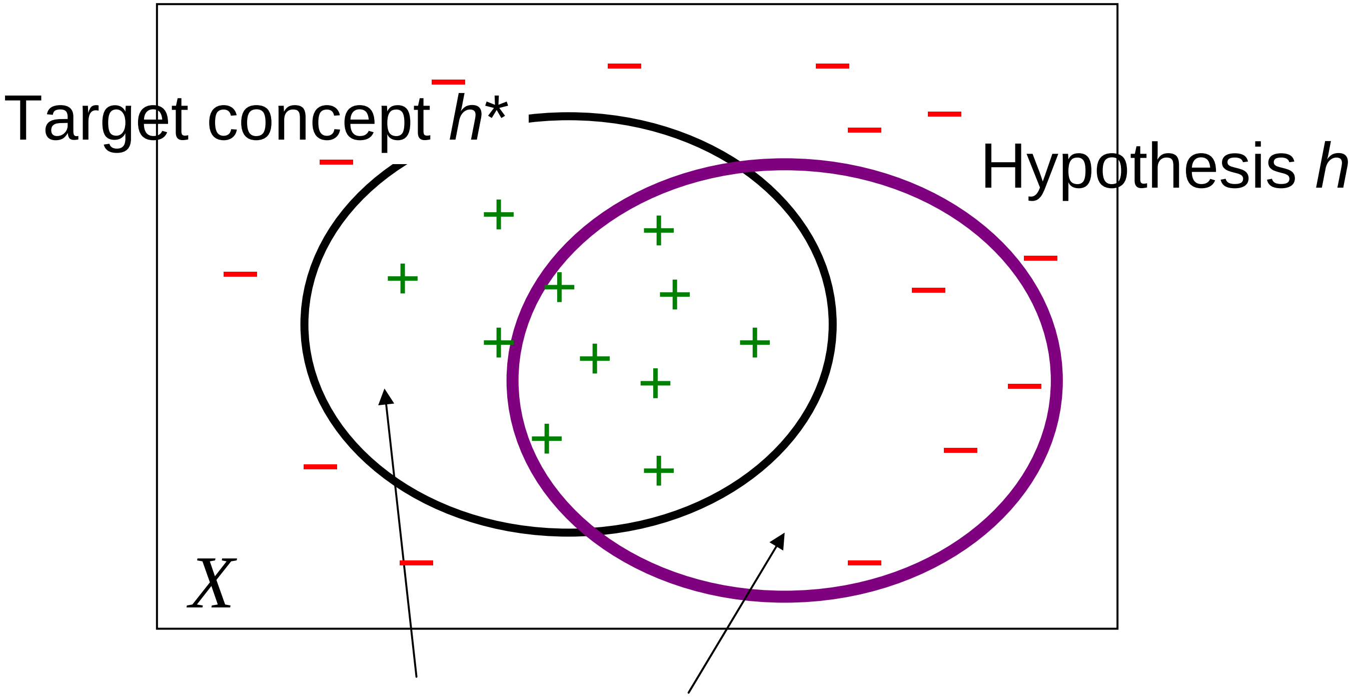
Probably Approximately Correct (PAC)

- Assumption of “uniformity of nature”:
  - Training and test instances drawn from some fixed probability distribution on the space *X*.



Probably Approximately Correct (PAC)

- Assumption of “uniformity of nature”:
  - Training and test instances drawn from some fixed probability distribution on the space *X*.



Error rate ε = probability of drawing from these regions

Probably Approximately Correct (PAC)

- The intuition: Want to be confident that a hypothesis which looks good on the training examples (i.e., appears to have zero error rate) in fact has a low true error rate (ε), and thus will generalize well on the test instances.

- PAC theorem: With probability 1-δ, a hypothesis consistent with *N* training examples will have true error rate at most ε whenever

$$N \geq \frac{1}{\varepsilon}\left(\log|H| + \frac{1}{\delta} \right).$$

Probably Approximately Correct (PAC)

- PAC theorem: With probability 1-δ, a hypothesis consistent with *N* training examples will have true error rate at most ε whenever

$$N \geq \frac{1}{\varepsilon}\left(\log\left|H\right| + \frac{1}{\delta}\right).$$

- How does *N*, the amount of data required for good generalization, change with problem parameters?
  - As allowable error (ε) decreases, *N* increases.
  - As desired confidence (1-δ) increases, *N* increases.
  - As the size of the hypothesis space (log |*H*|) increases, *N* increases.

Implications for what makes a good hypothesis space or inductive bias.

Probably Approximately Correct (PAC)

$$N \geq \frac{1}{\varepsilon}\left(\log|H| + \frac{1}{\delta}\right).$$

Why does N depend on number of hypotheses, $|H|$?

- Consider the set of "bad" hypotheses, H_{bad}: hypotheses with true error rate greater than or equal to ε.
- We want to be confident that a hypothesis which looks good on N examples is not actually in H_{bad} .
- Each example on average rules out at least a fraction ε of the hypotheses in H_{bad} . The bigger H_{bad} is, the more examples we need to see to be confident that all bad hypotheses have been eliminated.
- The learner doesn't know how big H_{bad} is, but can use $|H|$ as an upper bound on $|H_{\text{bad}}|$.

PAC analyses of other hypothesis spaces

- Single features
- Conjunctions
- Disjunctions
- Conjunctions plus *k* exceptions
- Disjunction of *k* conjunctive concepts
- All logically possible Boolean concepts
- Also:
  - Regions in Euclidean space: rectangles, polygons, …
  - Branches of trees….

- Single features:

h_1: f_2
h_2: f_5

- Conjunctions:

h_1: f_2 AND f_5
h_2: f_1 AND f_2 AND f_5

- Disjunctions:

h_1: f_2 OR f_5
h_2: f_1 OR f_2 OR f_5

- Conjunctions plus *k* exceptions:

h_1: (f_1 AND f_2) OR (0 1 0 1 1 0)
h_2: (f_1 AND f_2 AND f_5) OR (0 1 0 1 1 0) OR (1 1 0 0 0 0)

- Disjunction of *k* conjunctive concepts:

h_1: (f_1 AND f_2 AND f_5) OR (f_1 AND f_4)
h_2: (f_1 AND f_2) OR (f_1 AND f_4) OR (f_3)

- All logically possible Boolean concepts:

h_1: (1 1 1 0 0 0), (1 1 1 0 0 1), (1 1 1 0 1 0), ...
h_2: (0 1 0 1 1 0), (1 1 0 0 0 0), (1 0 0 1 1 1), ...

- Single features:

  $\log |H| = \log k \quad (k = \text{\# features})$

- Conjunctions:

  $\log |H| = k$

$$\boxed{N \geq \frac{1}{\varepsilon}\left(\log|H| + \frac{1}{\delta}\right).}$$

- Disjunctions:

  $\log |H| = k$

- Conjunctions plus *m* exceptions:

  $\log |H| \sim km$

- Disjunction of *m* conjunctive concepts:

  $\log |H| \sim km$

- All logically possible Boolean concepts:

  $\log |H| = 2$^k = number of objects in world.

The role of inductive bias

- Inductive bias = constraints on hypotheses.
- Learning with no bias (i.e., *H* = all possible Boolean concepts) is impossible.
  - PAC result
  - A simpler argument by induction.

The role of inductive bias

- Inductive bias = constraints on hypotheses.
- Relation to Ockham's razor: $N \geq \frac{1}{\varepsilon}\left(\log|H| + \frac{1}{\delta} \right).$
  - "Given two hypotheses that are both consistent with the data, choose the simpler one."
  - log |*H*| = number of bits needed to specify each hypothesis *h* in *H*. Simpler hypotheses have fewer alternatives, and shorter descriptions.
  - E.g. Avoid disjunctions unless necessary:

    "All emeralds are green and less than 1 ft. in diameter" vs. "All emeralds are green and less than 1 ft. in diameter, or made of cheese".

The importance of choosing good features

- Choose features such that all the concepts we need to learn will expressible as simple conjunctions.
- Failing that…. conjunctions with few exceptions, or disjunctions of few conjunctive terms.
- If we choose "any old features", we may express any possible concept as some Boolean function of those features, but learning will be impossible.

What this doesn’t tell us

- Why conjunctions easier than disjunctions?
  - C.f., Why is “all emeralds are green and less than 1 ft. in diameter” better than “all emeralds are green or less than 1 ft. in diameter”?
  - What are concepts useful for in the real world?
  - What is structure of natural categories?

What this doesn’t tell us

- Why conjunctions easier than disjunctions?
- How we choose the appropriate generality of a concept, given one or a few examples?
  - Subset principle says to choose a hypothesis that is as small as possible.
  - Ockham’s razor says to choose a hypothesis that is as simple as possible.
  - But these are often in conflict, e.g. with conjunctions. People usually choose something in between, particularly with just one example. Consider word learning ….

What this doesn’t tell us

- Why conjunctions easier than disjunctions?
- How we choose the appropriate generality of a concept, given one or a few examples?
- How we should (or do) handle uncertainty?
  - How confident that we have the correct concept?
  - When to stop learning?
  - What would the best example to look at next?
  - What about noise (so that we cannot just look for a consistent hypothesis)?
  - Should we maintain multiple hypotheses? How?

What this doesn’t tell us

Compare PAC bounds with typical performance in Bruner’s experiments or the real world.

- E.g., need > 200 examples to have 95% confidence that error is < 10%
- Bruner experiments: 5-7 examples
- Children learning words:

Images removed due to copyright considerations.

Other learning algorithms

- Current-best-hypothesis search

Images removed due to copyright considerations.

- Version spaces:

Images removed due to copyright considerations.

Summary: Inductive learning as search

- Rigorous analyses of learnability.
  - Explains when and why learning can work.
  - Shows clearly the need for inductive bias and gives a formal basis for Occam’s razor.
- Many open questions for computational models of human learning, and building more human-like machine learning systems.
  - Where do the hypothesis spaces come from? Why are some kinds of concepts more natural than others?
  - How do we handle uncertainty in learning?
  - How do we learn quickly and reliably with very flexible hypothesis spaces?

Outline for today

- PAC learning analyses, and the importance of choosing good features.
- Grammars
- First-order logic
- Learning a theory and new concepts in first-order logic.

High-level points

- The basic idea of a grammar: rules for generating structures.
- Some lessons for cognition more generally:
  - The importance of structured representations.
  - How structure and statistics interact.

What was Chomsky attacking?

- Models of language as a *sequential* object.
  - E.g., n-th order Markov models:

$$P(w_{i+n} \mid w_i, \ldots, w_{i+n-1})$$

  - Or, n-grams:

$$P(w_i, \ldots, w_{i+n})$$

$P(\text{model}, \text{of}, \text{language})$

Image removed due to copyright considerations.

$$P(\text{model}, \text{of}, \text{quickly})$$

Image removed due to copyright considerations.

The classic example where frequency fails

- “Colorless green ideas sleep furiously.”
- “Furiously sleep ideas green colorless.”

Image removed due to copyright considerations.

The new example where frequency fails

- “Stupendously beige ideas circumnavigate mercilessly.”

Image removed due to copyright considerations.

What is a grammar?

- A system for representing structures that “makes infinite use of finite means” (von Humboldt, ~1830’s).

A grammar is like a theory

"The grammar of a language can be viewed as a theory of the structure of this language. Any scientific theory is based on a certain finite set of observations and, by establishing general laws stated in terms of certain hypothetical constructs, it attempts to account for these observations, to show how they are interrelated, and to predict an indefinite number of new phenomena…. Similarly, a grammar is based on a finite number of observed sentences… and it 'projects' this set to an infinite set of grammatical sentences by establishing general 'laws'… [framed in terms of] phonemes, words, phrases, and so on…."

Chomsky (1956), "Three models for the description of language"

Finite-state grammar

- The “minimal linguistic theory”.

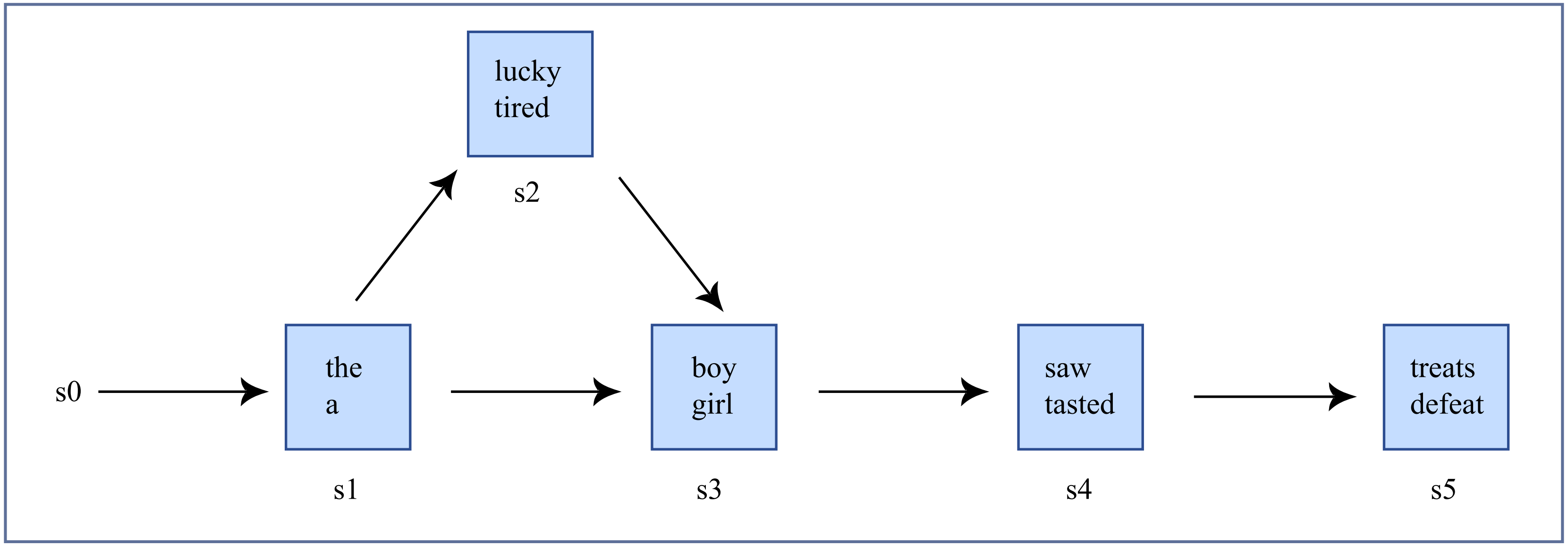


Figure by MIT OCW.

E.g., “The lucky boy tasted defeat.”

Generating infinite strings

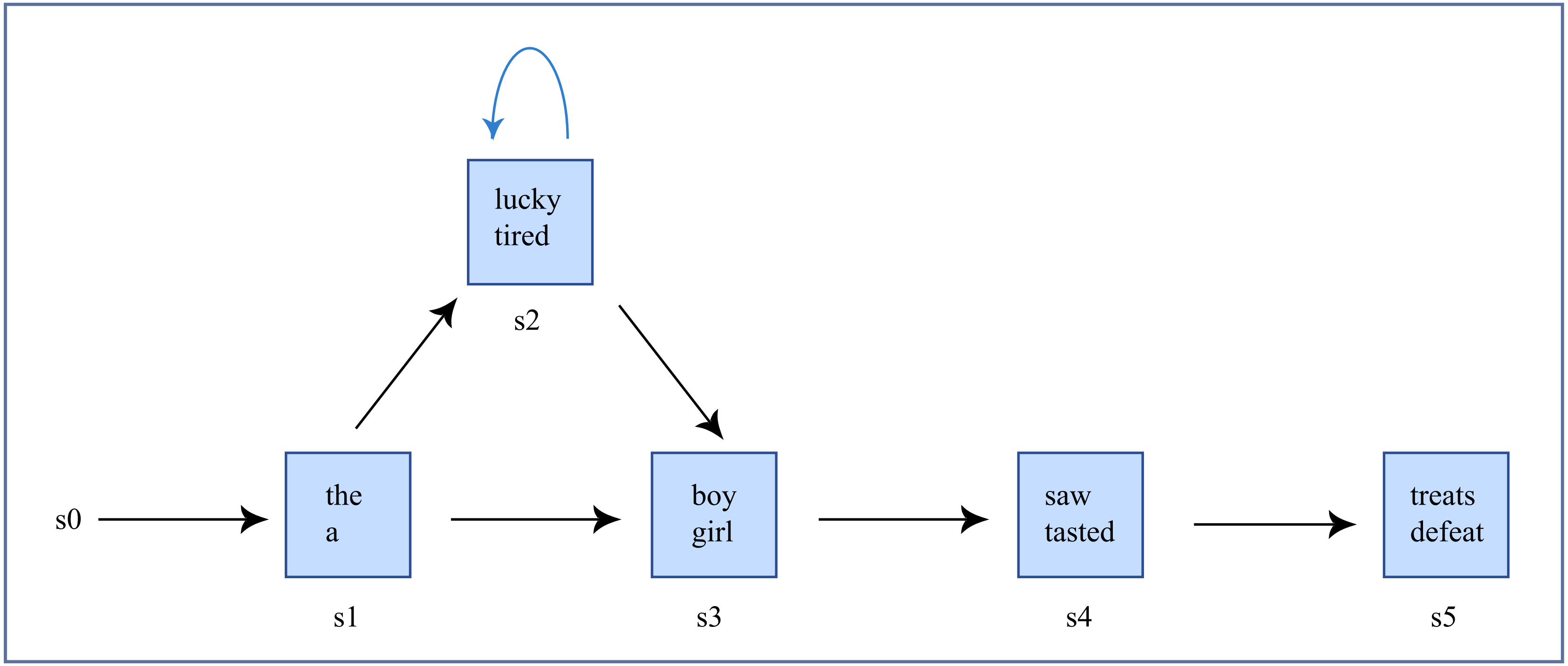


Figure by MIT OCW.

E.g., “The lucky tired tired … boy tasted defeat.”

Parses novel sequences

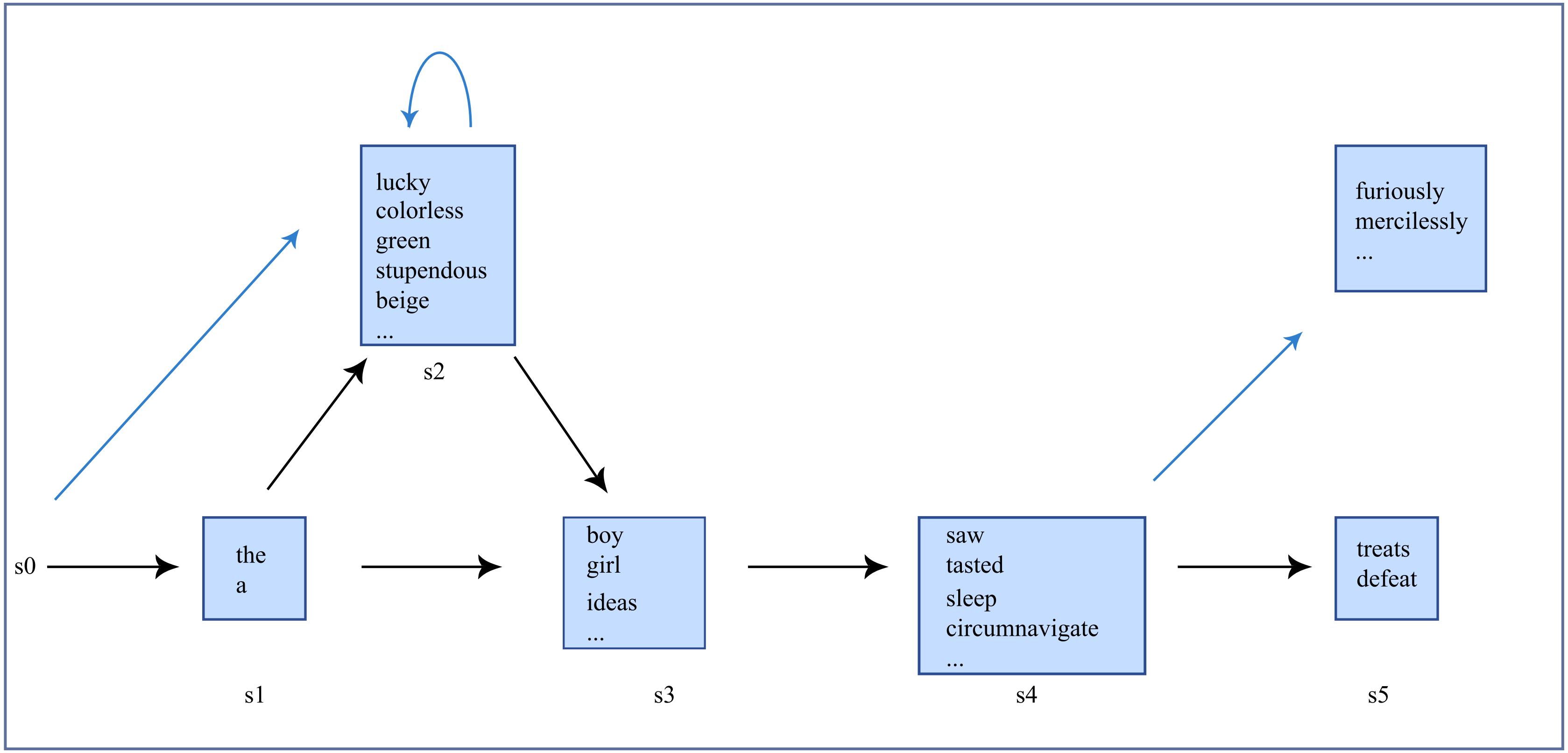


Figure by MIT OCW.

E.g., “Colorless green ideas sleep furiously”

What’s really wrong

- The problem with the “statistical” models isn’t that they are statistical.
- Nature of representation:
  - N-grams: Perceptual/motor/superficial/concrete
    - Utterance is a *sequence* of *words*.
  - Chomsky: Cognitive/conceptual/deep/abstract
    - Finite-state grammar: utterance is a *sequence* of *states*.
    - Phrase-structure grammar: utterance is a *hierarchical structure* of *phrases*.

Counterexamples to sequential models

- Center embedding
  - The man dies.
  - The man that the racoons bite dies.
  - The man that the racoons that the dog chases bite dies.
  - The man that the racoons that the dog that the horses kick chases bite dies.

Generating embedded clauses

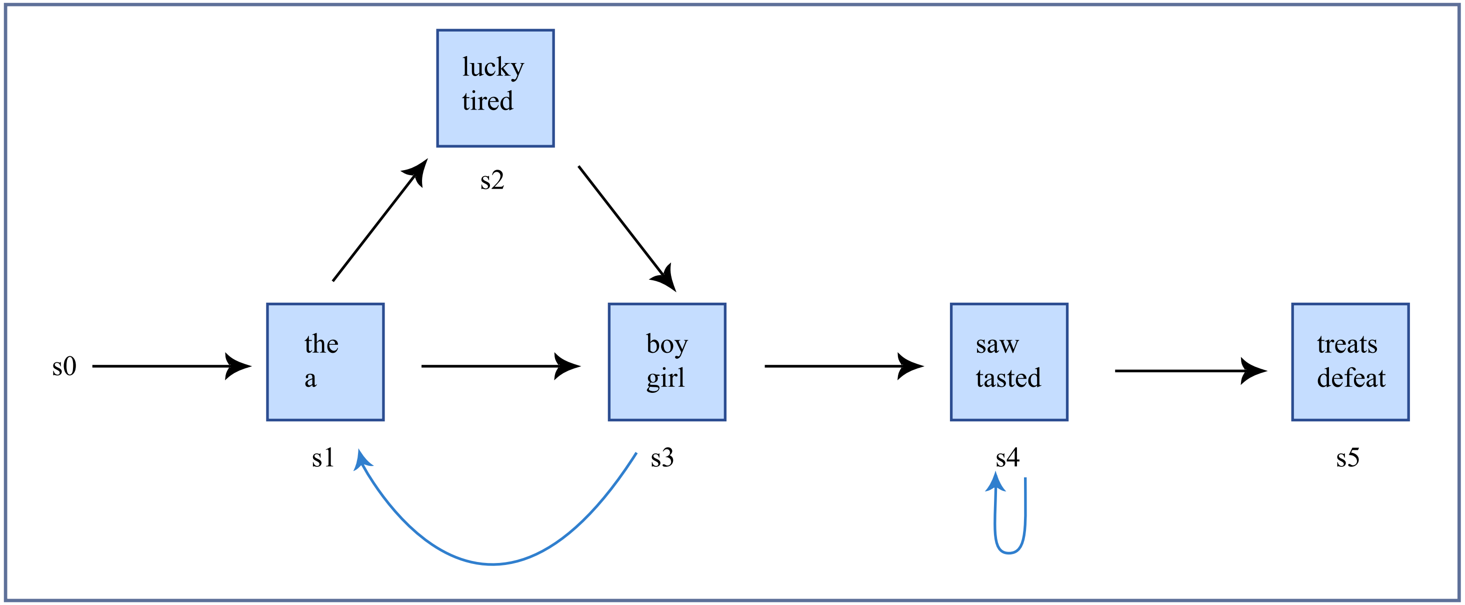


Figure by MIT OCW.

E.g., “A lucky boy the tired girl saw tasted defeat.”

No way to ensure subject-verb agreement.

Counterexamples to sequential models

- Tail-embedding
  - The horse that kicked the dog that chased the racoons that bit the man is alive.
  - The horses that kicked the dog that chased the racoons that bit the man are alive.

- Fundamental problem: Dependencies may be arbitrarily long-range in the sequence of words. Only local in the *phrase structure*.