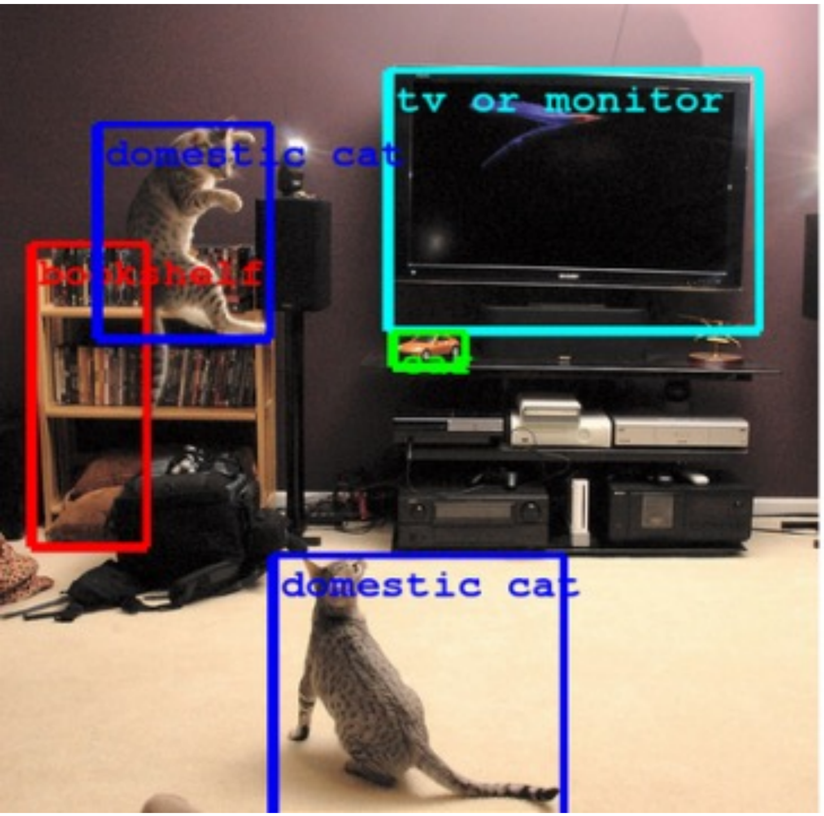
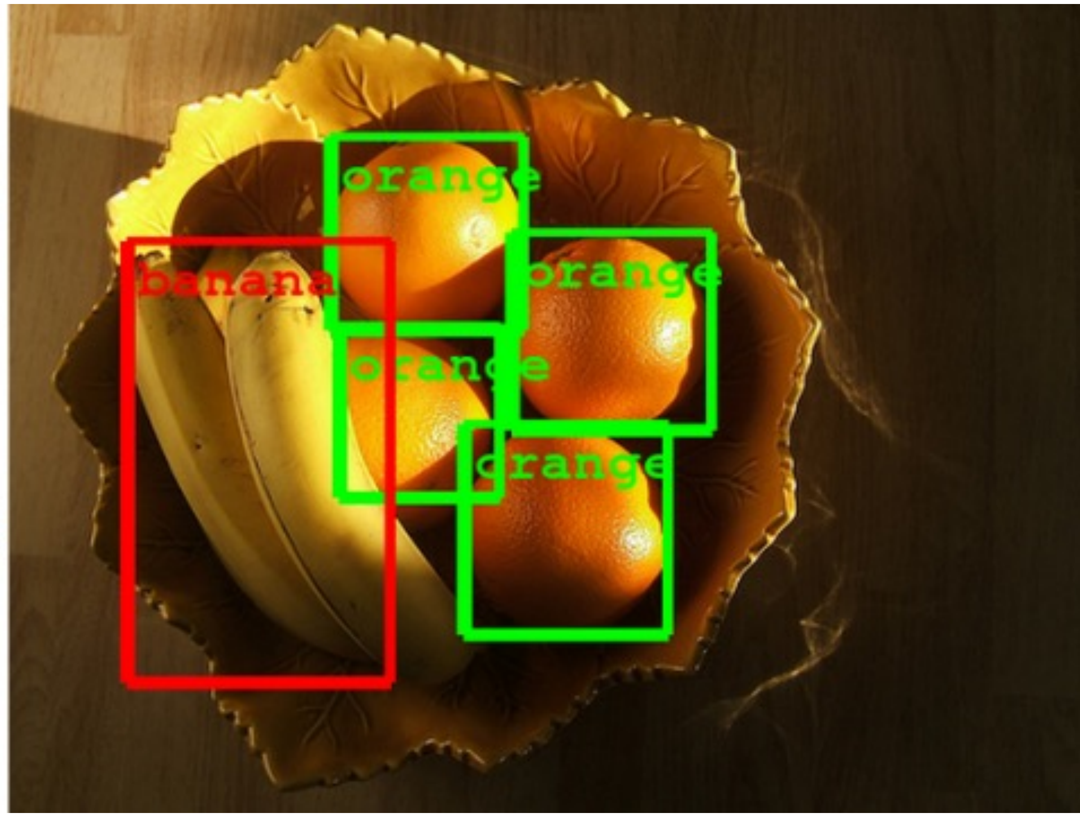
Hidden Markov Models

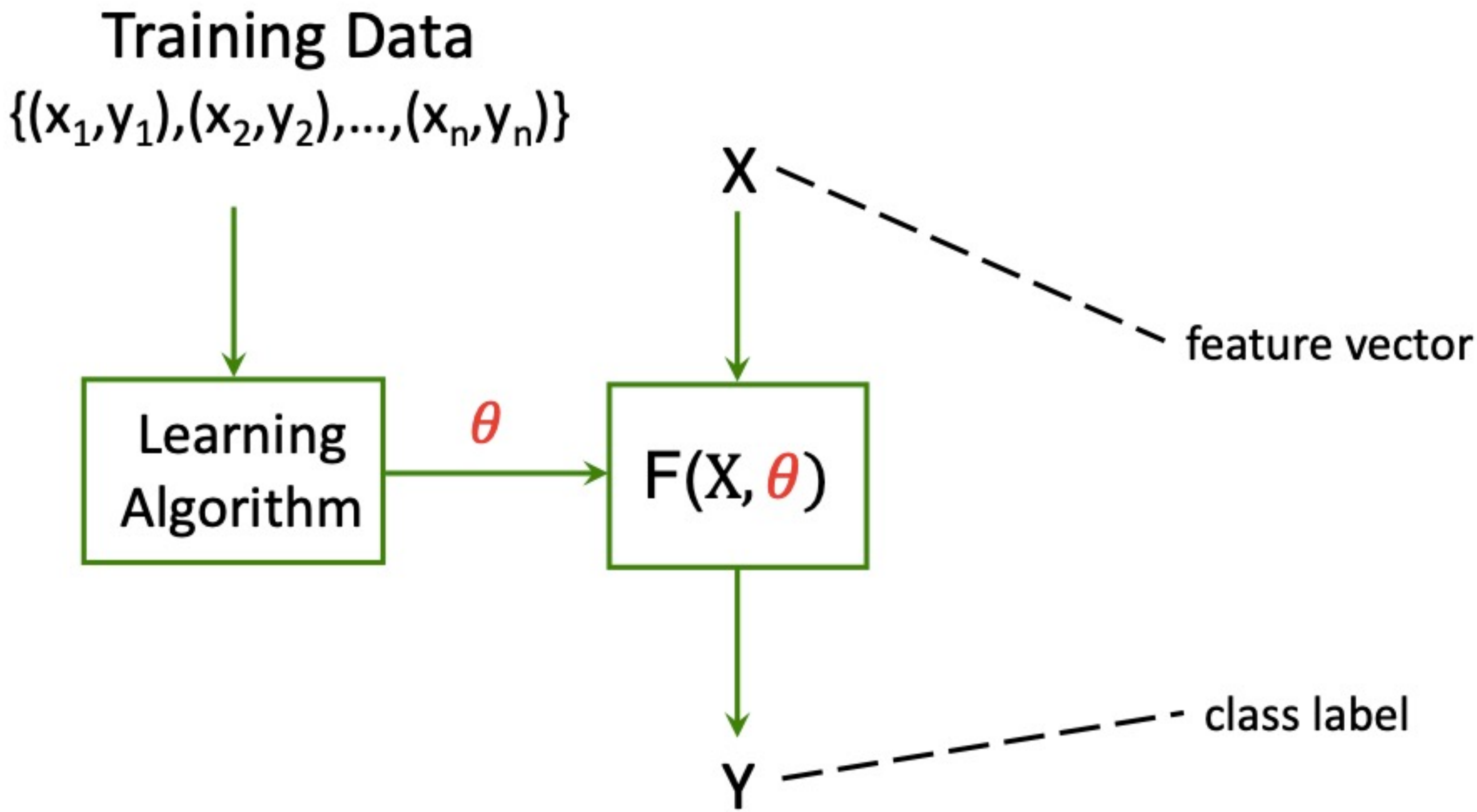




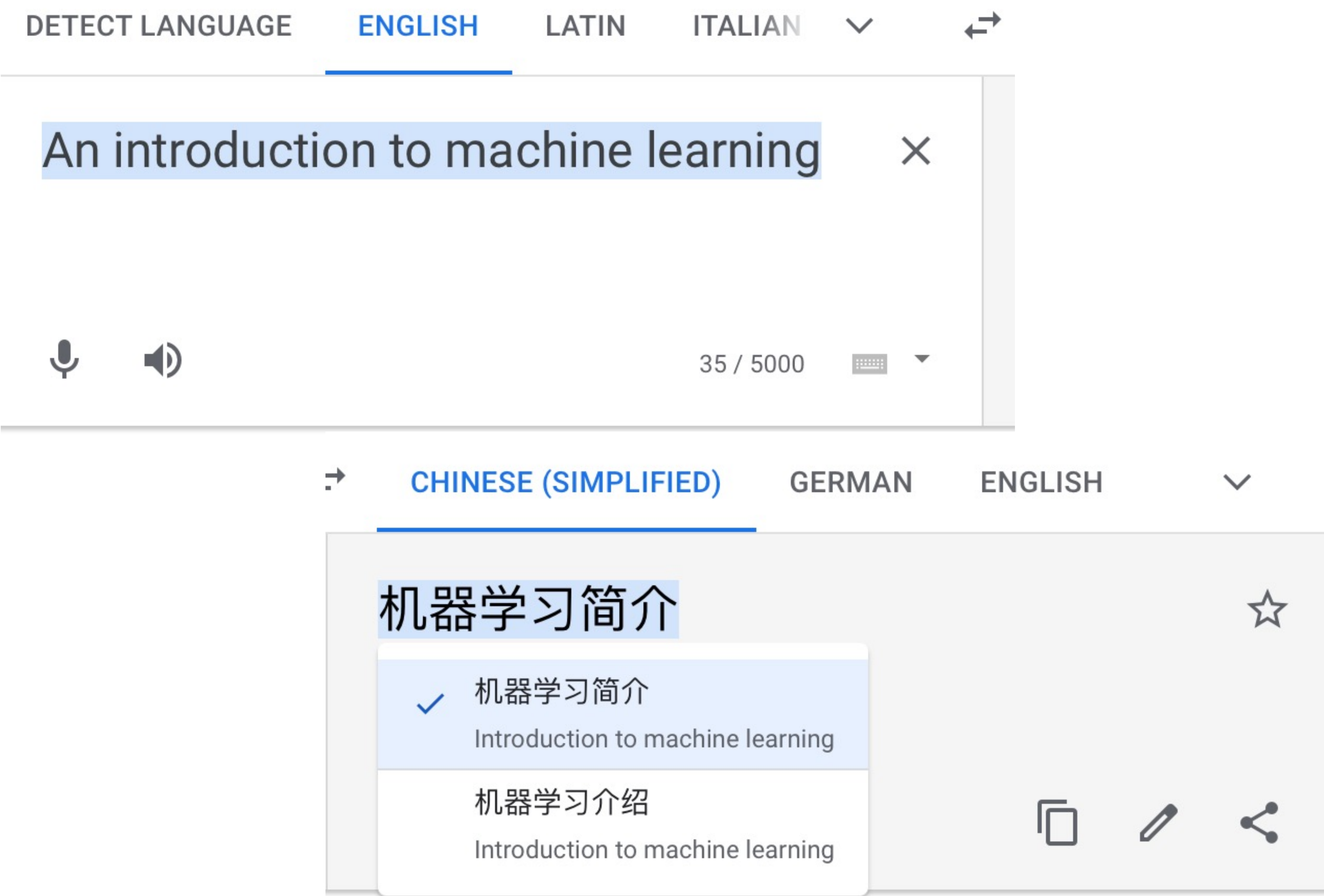
Zach Wood-Doughty and Bryan Pardo

Some slides borrowed from Mark Dredze

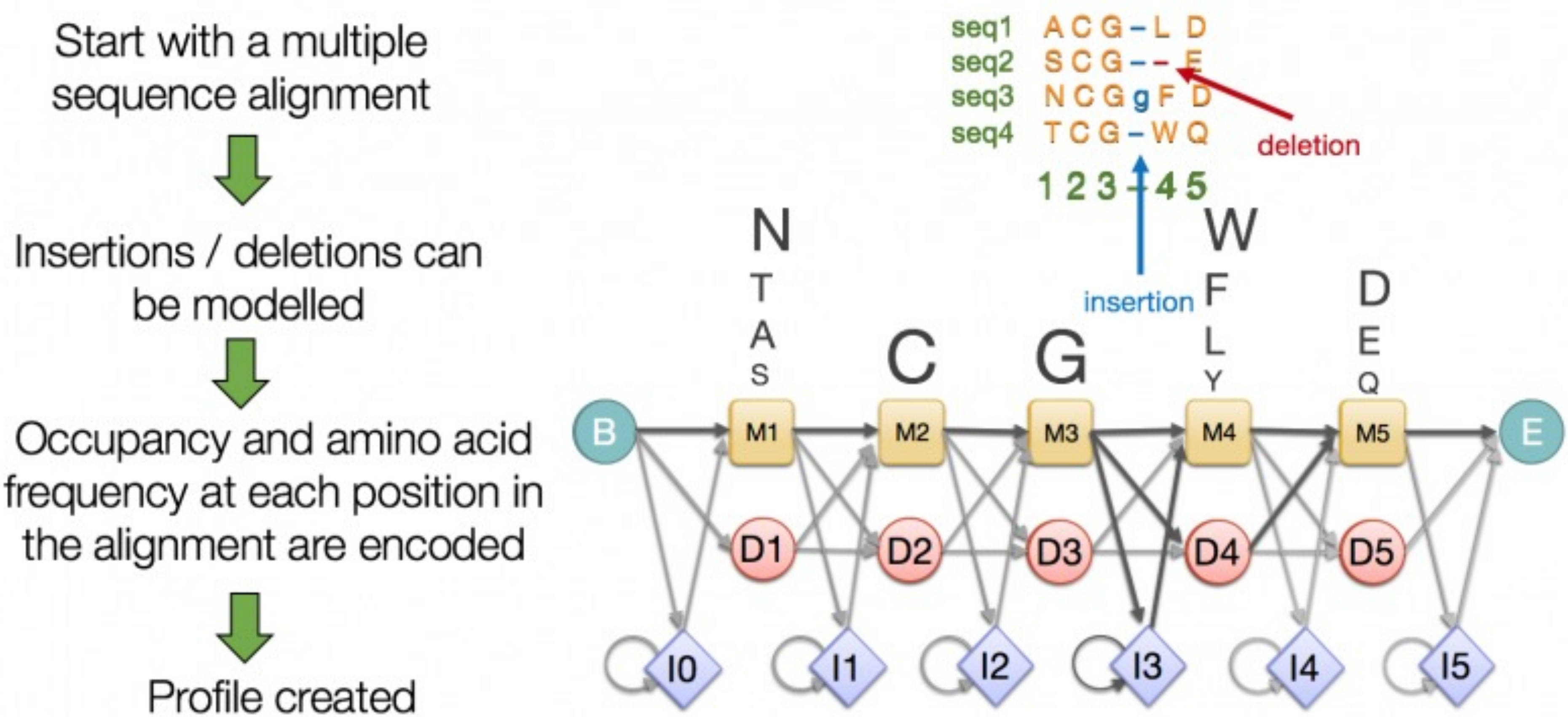
Classification & Regression



What about the more complex tasks we know that ML can be used for?

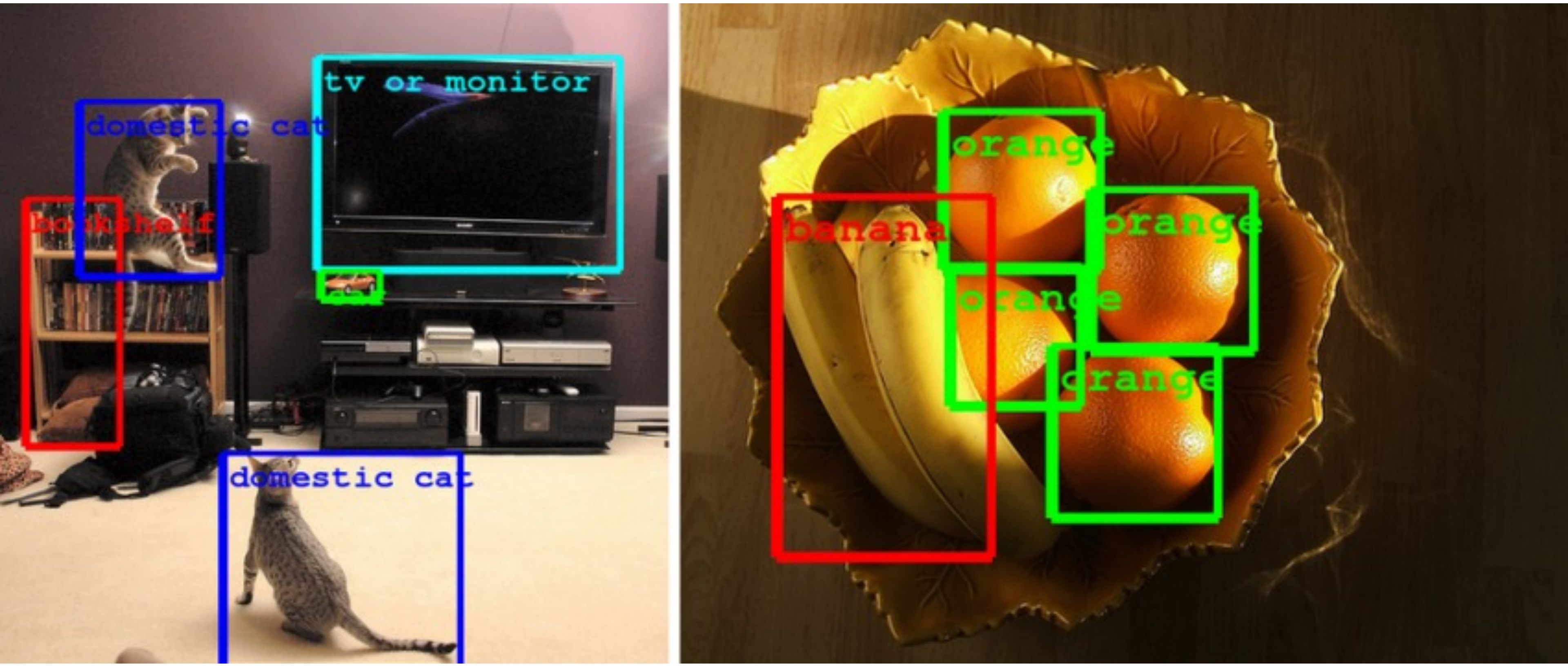

DETECT LANGUAGE
ENGLISH
LATIN
ITALIAN
An introduction to machine learning
35 / 5000
CHINESE (SIMPLIFIED)
GERMAN
ENGLISH
机器学习简介
机器学习简介
Introduction to machine learning
机器学习介绍
Introduction to machine learning

Biological Sequence Alignment



Search and Ranking

Object Recognition

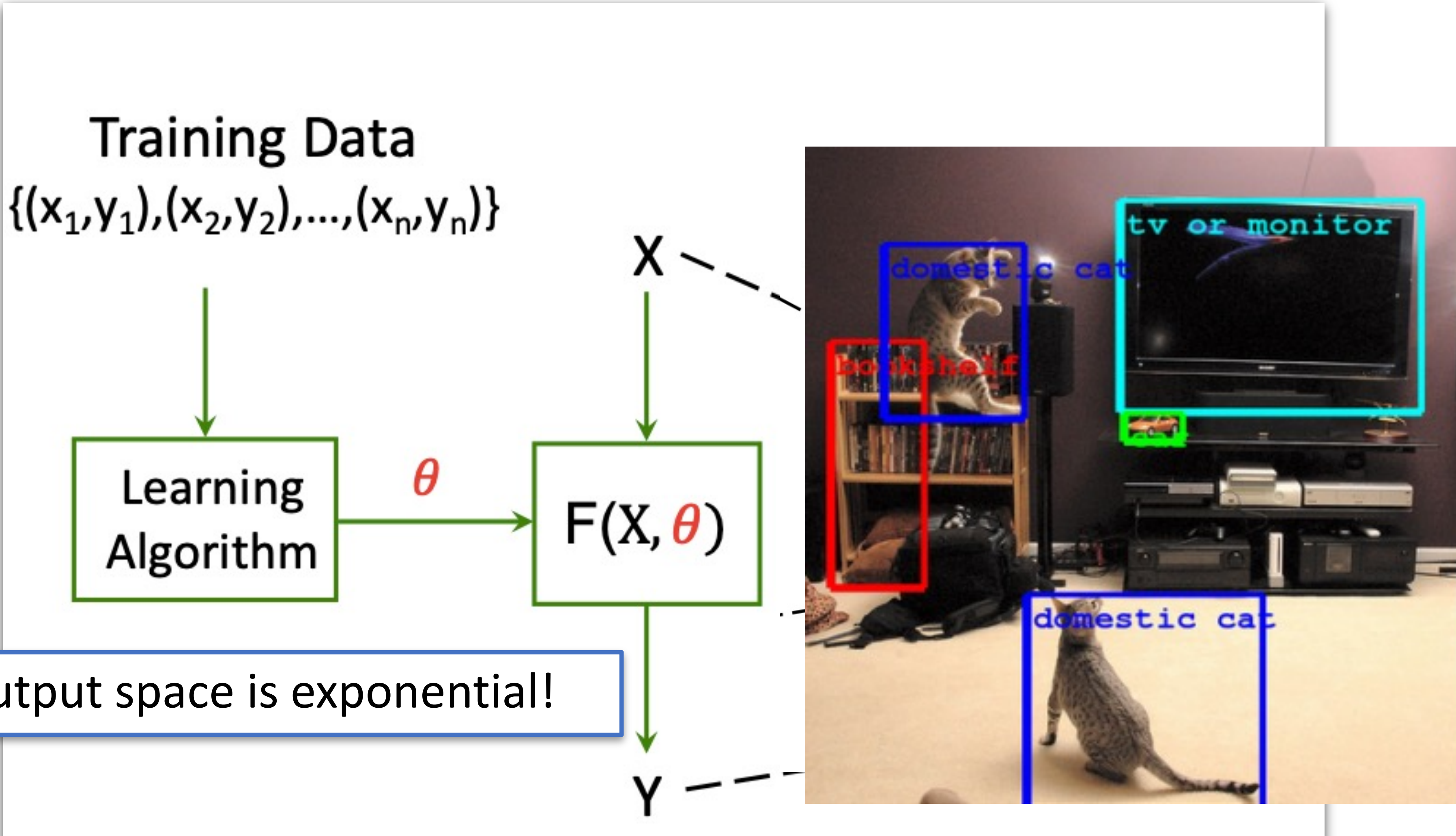


https://3c1703fe8d.site.internapcdn.net/newman/gfx/news/hires/2014/googleteamri.png

What is Structured Prediction?

- Input: x
  - Typically, a structured input (string, tree, image)
  - Maintain structure of input in x
    - Don’t just flatten into a real-valued vector
- Output: y
  - y is now from a large set of possible outputs
  - Outputs connected to input, often exponential in size of input

Structured Prediction



Output space is exponential!

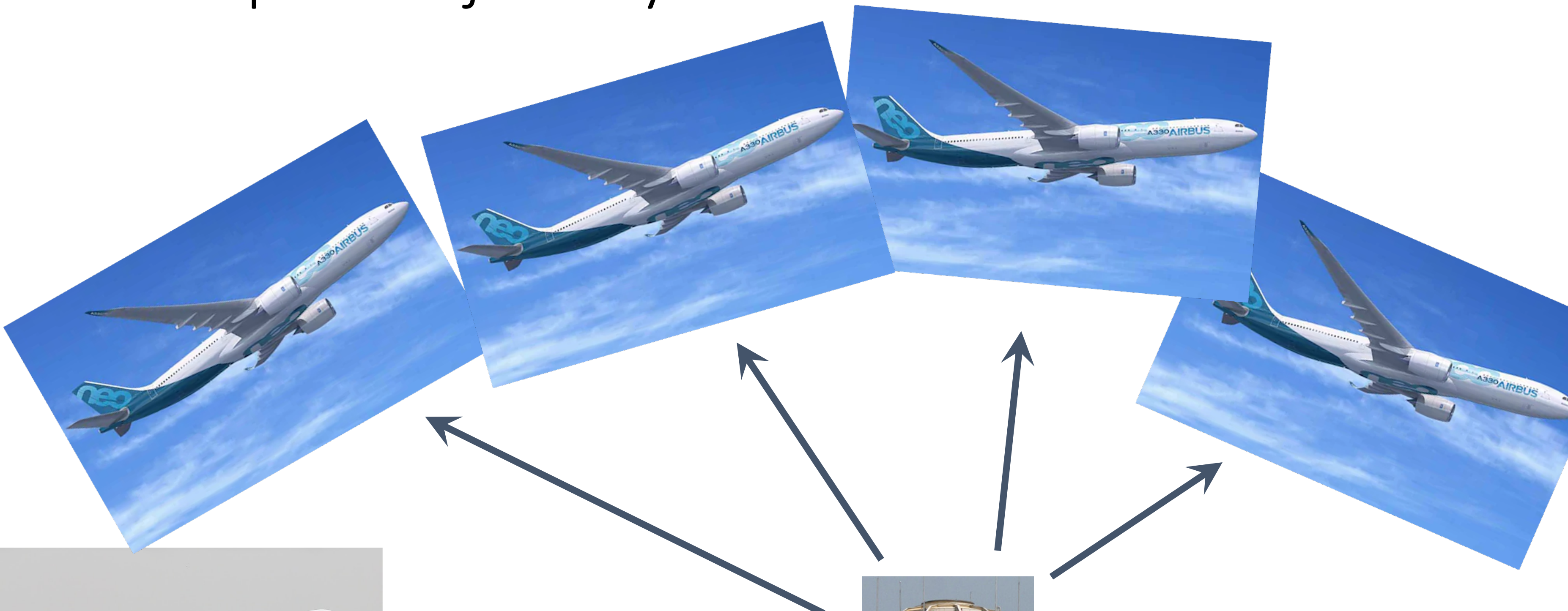
Previous Approaches

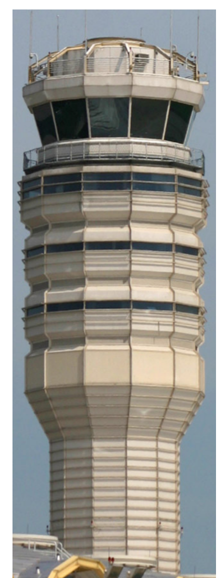
- Multi-class classifiers:
  - e.g. neural networks or decision trees
  - Train one classifier per label
- These won’t work when:
  - Exponential number of possible outputs
  - Outputs defined based on input
  - Output subcomponents highly correlated with each other

Sequential Events

- Many events happen in sequence
  - Weather on consecutive days
  - Words in a written sentence
  - Speech or music
  - Movements in the stock market
  - DNA base pairs
- Let's model sequences with a graphical model
  - Input $p(X_1, X_2, \dots X_N)$
  - Output $p(X_{N+1}, X_{N+2} \mid X_1, X_2, \dots X_N)$

Example: Trajectory Prediction





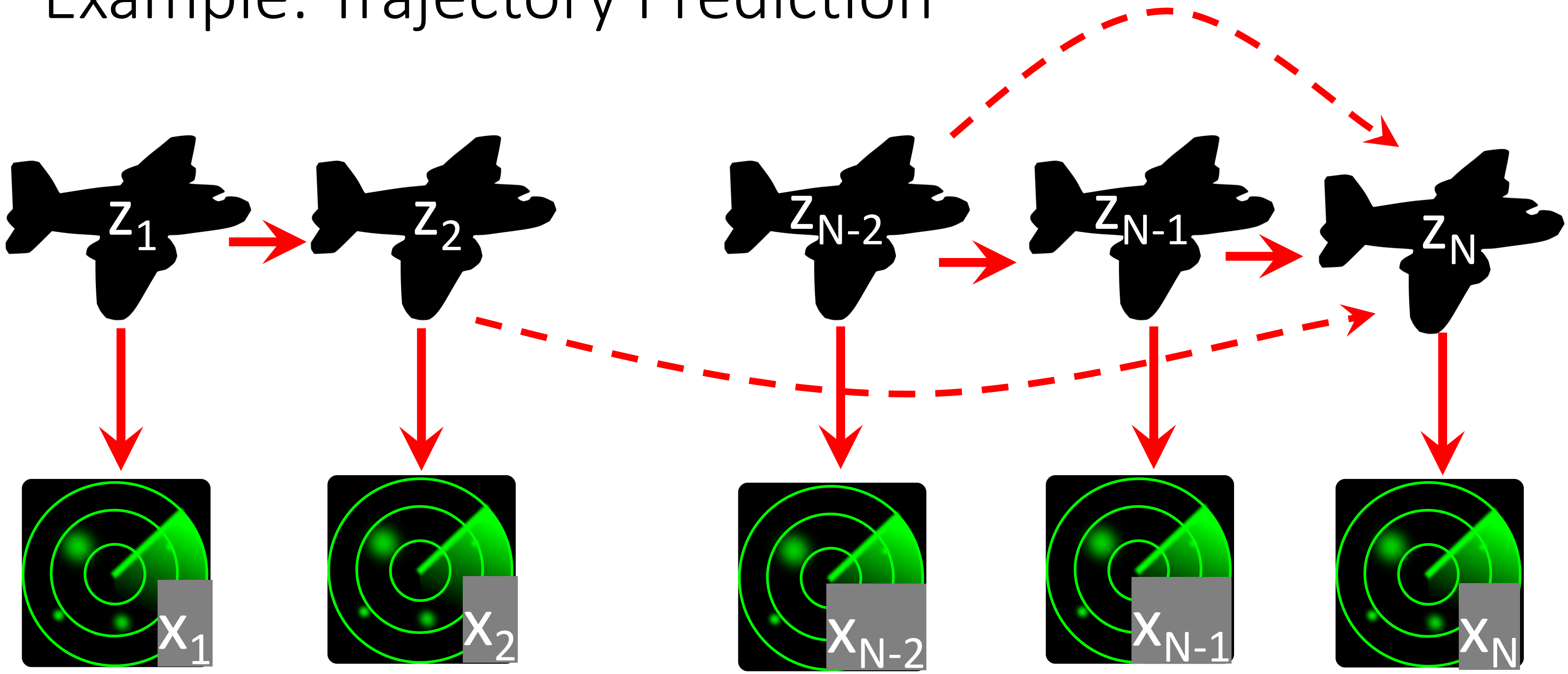
Where is the aircraft now?
Where will it be in 10 minutes?

Credit: David Sontag

Example: Trajectory Prediction

- There are two kinds of variables in our problem.
  At every time t = 1, 2, 3, …
    - We see the measured radar location of the airplane
    - We don’t see the true location of the airplane
- How would we graphically represent this?

Example: Trajectory Prediction



Sequential Models

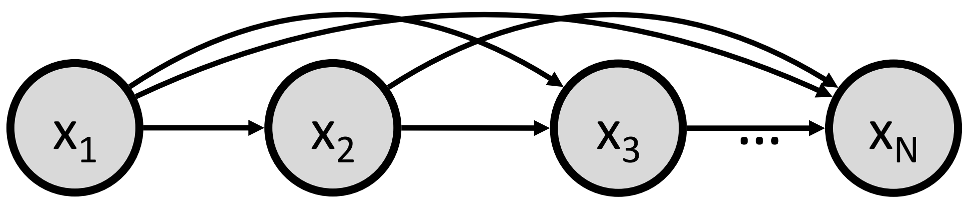
- Simplest approach
  - Each event is independent

$$p(x_1, x_2 \ldots x_N) = \prod_{n=1}^{N} p(x_n)$$

  - Simple, but not very helpful

Sequential Models

- Complex approach
  - Each event is dependent on previous events



$$p(x_1, x_2 \ldots x_N) = \prod_{n=1}^{N} p(x_n \mid x_1 \ldots x_{n-1})$$

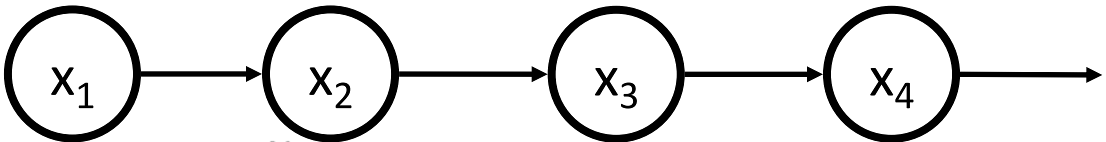
  - Captures dependencies, but way too complex

Markov Assumption

- The current state depends on a fixed number of previous states
  - The weather today depends on the past three days, but NOT two weeks ago
  - The next word in the sentence depends on the past three words, but nothing before
- Pro: makes for simpler models
- Con: doesn’t capture full history

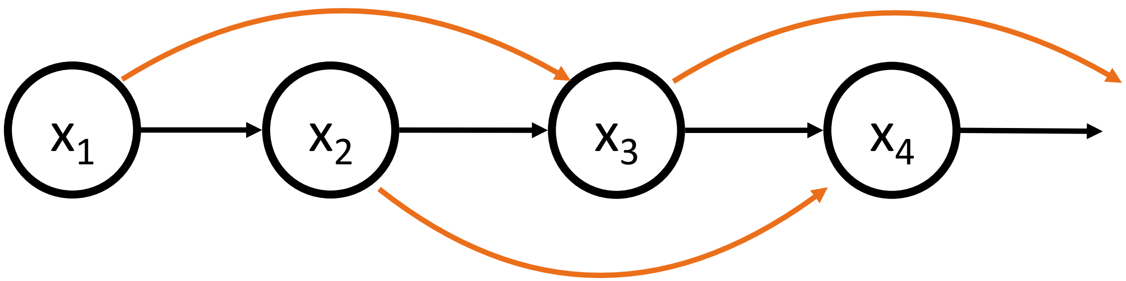
Markov Chains

- First order Markov chain



$$p(x_1, x_2 \ldots x_N) = \prod_{n=1}^{N} p(x_n \mid x_{n-1})$$

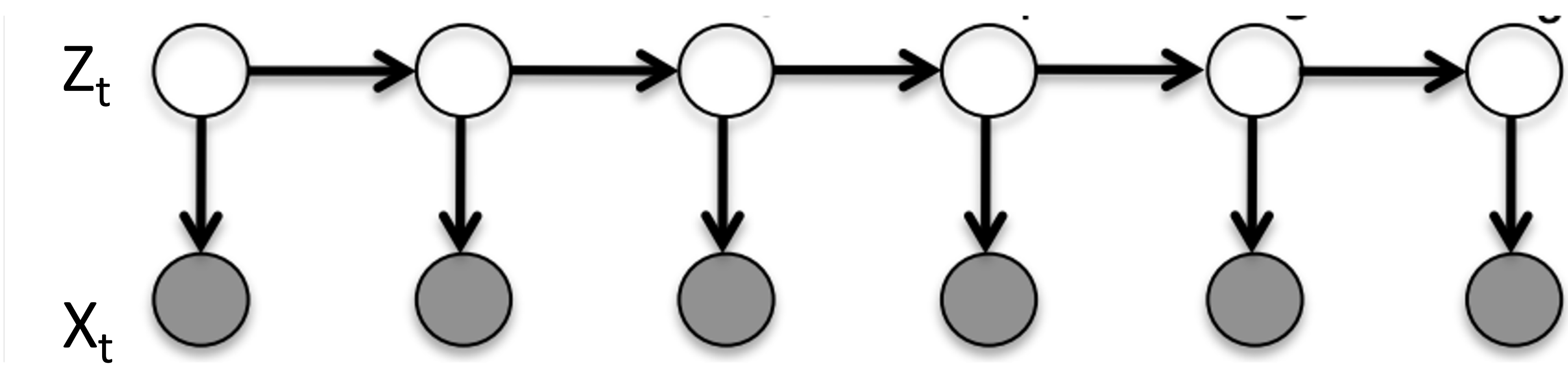
- Second order Markov chain

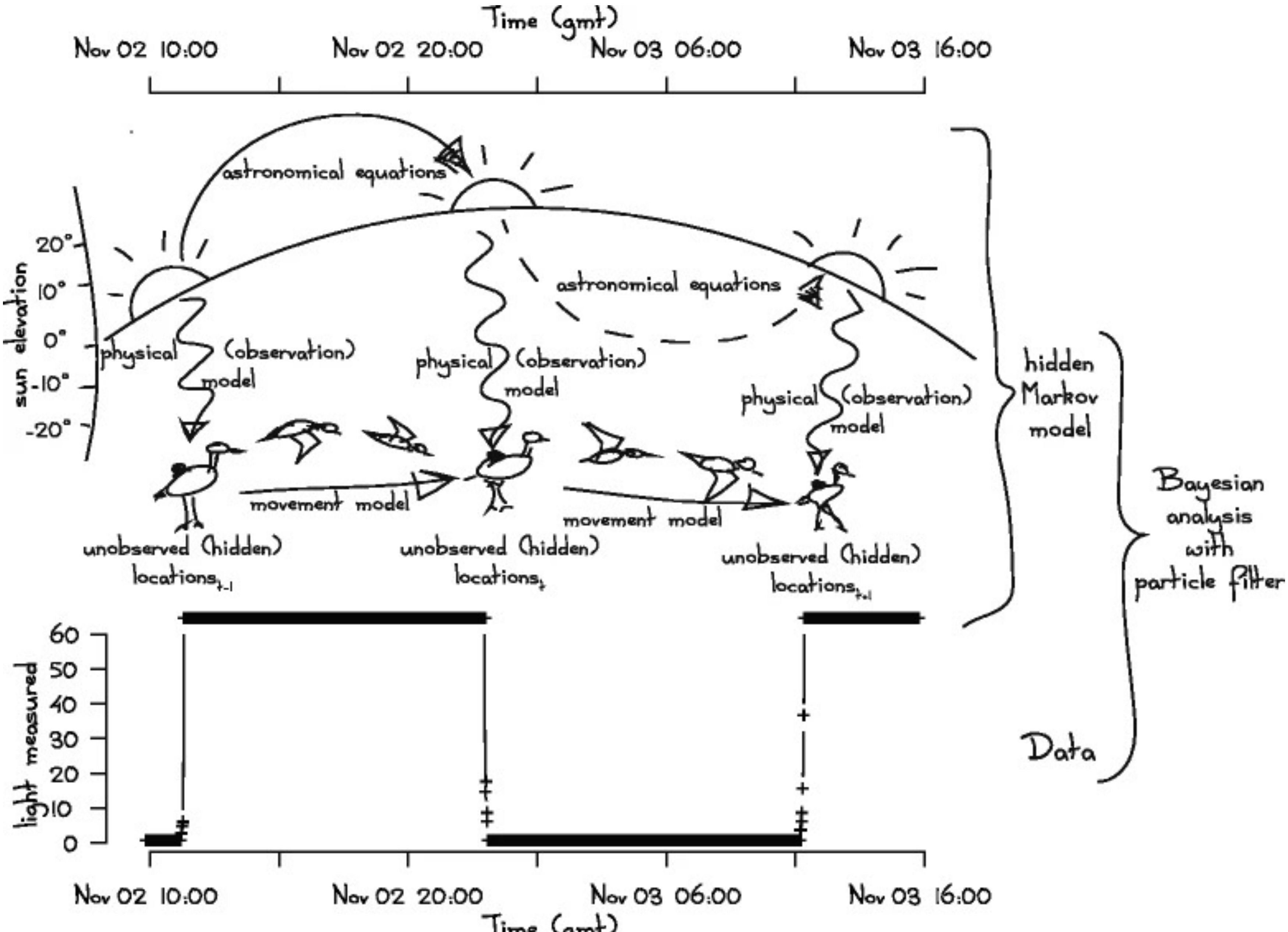


$$p(x_1, x_2 \ldots x_N) = \prod_{n=1}^{N} p(x_n \mid x_{n-1}, x_{n-2})$$

Example: Trajectory Prediction

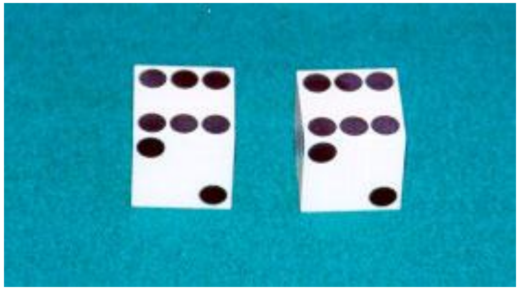
- There are two kinds of variables in our problem.
  At every time t = 1, 2, 3, …
  - We see the measured radar location of the airplane
  - We don't see the true location of the airplane
- How would we graphically represent this?





Rakhimberdiev, *et al.* HMM for reconstructing animal paths from solar geolocation loggers using templates for light intensity. *Mov Ecol* **3,** 25 (2015).

Casino Example

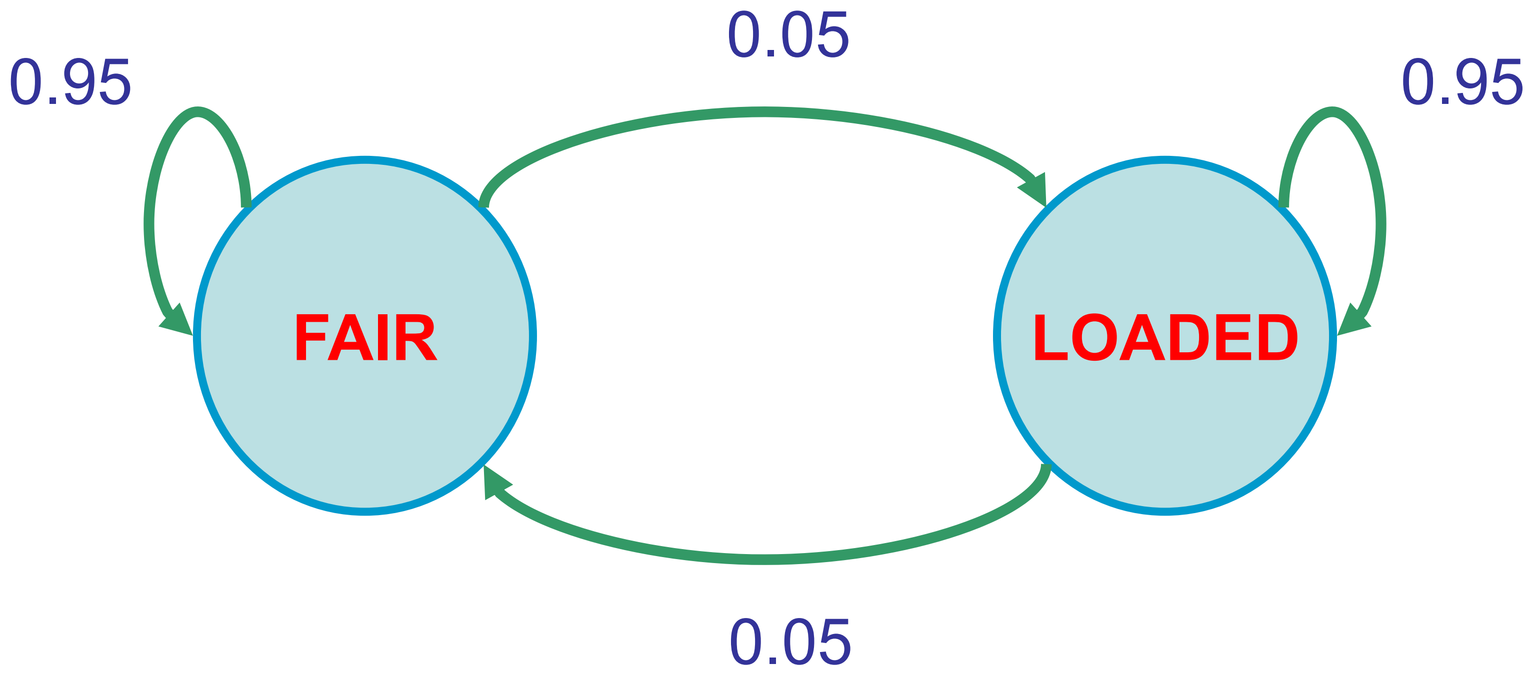


A casino has two dice:

- Fair die
  $P(1) = P(2) = P(3) = P(5) = P(6) = 1/6$
- Loaded die
  $P(1) = P(2) = P(3) = P(5) = 1/10$ $P(6) = ½$

I think the casino switches back and forth between fair and loaded die once every 20 turns, on average

Casino Example

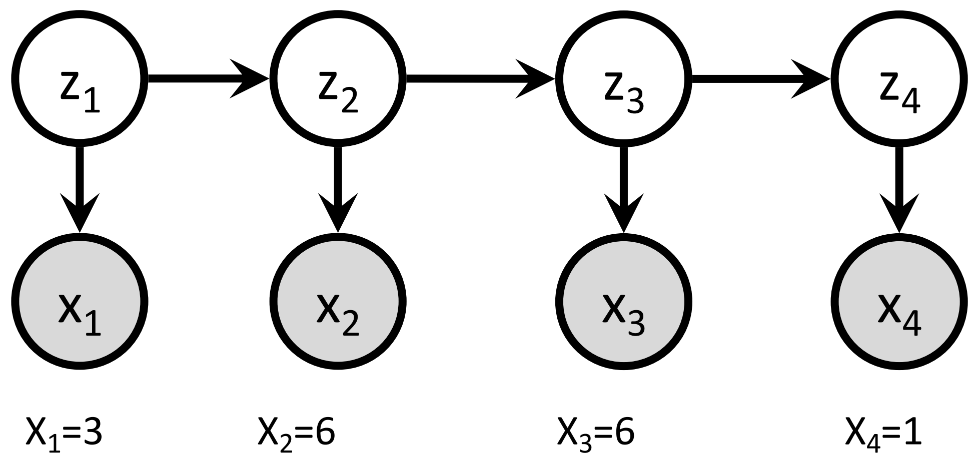


Casino Example

- Given a sequence of dice rolls, guess whether the fair or loaded die was used

Hidden states Z
(which die used)

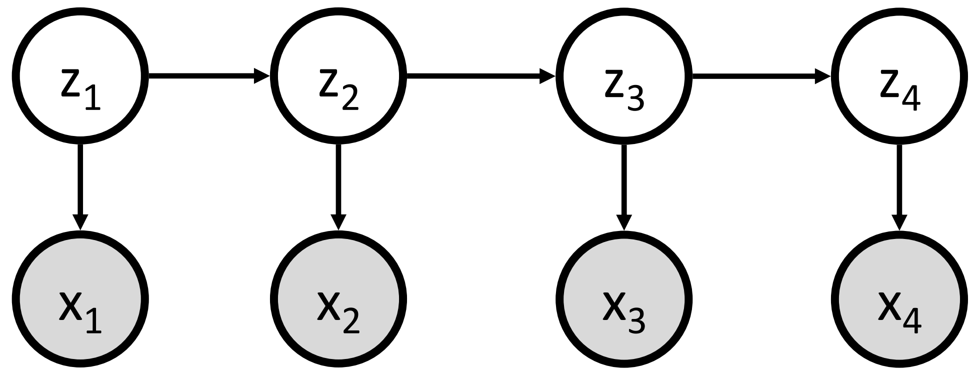
Observed emissions X
(The rolled value)



- This is called a Hidden Markov Model (HMM)

Conditional Probability Tables

	$P(z_n=1 \mid z_{n-1})$
$z_{n-1}=1$	
$z_{n-1}=0$	



	$P(x_n=1 \mid z_n)$
$z_n=1$	
$z_n=0$	

Joint Probability of HMM

- The joint probability of an HMM

$$p(\mathbf{X},\mathbf{Z} \mid \theta) = p(z_1 \mid \pi)\left[\prod_{n=2}^{N} p(z_n \mid z_{n-1},\mathbf{A})\right]\prod_{m=1}^{N} p(x_m \mid z_m,\phi)$$

- A- transition probabilities (matrix)
  - A_{ij} is the probability of moving from state i to j
- π- vector with starting probabilities
- ϕ – emission probabilities (matrix)
  - ϕ_{ij} is the probability of state i and emitting observation j

Unsupervised Training

- How do we train a probabilistic model?
    - As with GMMs and linear regression: maximum likelihood!

      $$\max_\theta \ p(\mathbf{X},\mathbf{Z} \mid \theta)$$

    - Problem: we only observe X, not Z
- Solution: EM
    - First, need to write the complete data likelihood

$$p(\mathbf{X} \mid \theta) = \sum_{\mathbf{Z}} p(z_1 \mid \pi) \left[\prod_{n=2}^{N} p(z_n \mid z_{n-1}, \mathbf{A}) \right] \prod_{m=1}^{N} p(x_m \mid z_m, \phi)$$

EM for HMMs

- E-Step
  - Find the expected values for the hidden variables Z given the model parameters
    - The most likely Z given X and current model parameters
- M-Step
  - Pretend to observe the values for Z
  - Update model parameters A, π, φ to maximize complete data likelihood

Why is the E-step hard?

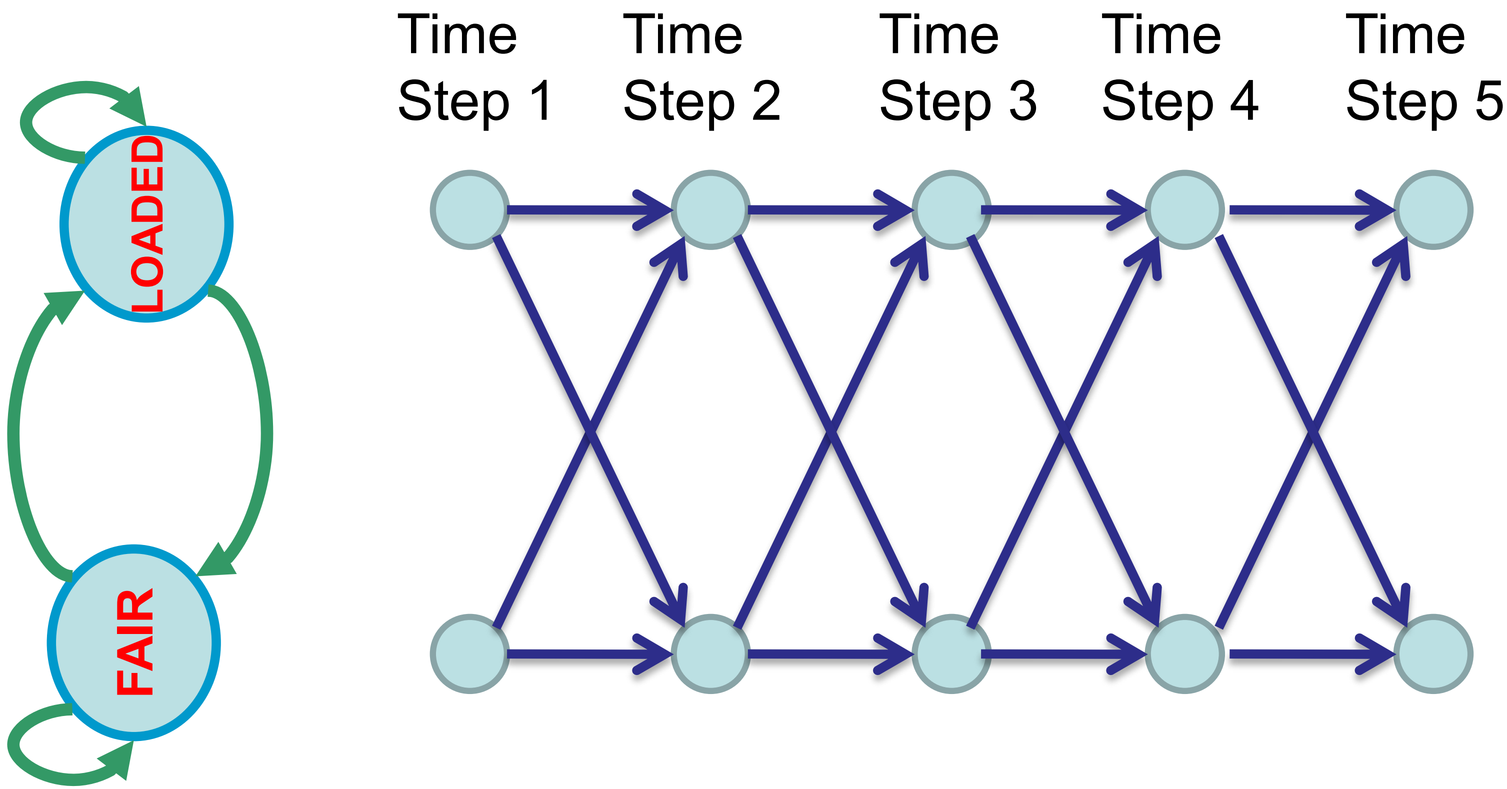
Naïve Bayes E-Step:

$$\delta(y|i) = p(y|\underline{x}^{(i)}; \underline{\theta}^{t-1}) = \frac{q^{t-1}(y) \prod_{j=1}^{d} q_j^{t-1}(x_j^{(i)}|y)}{\sum_{y=1}^{k} q^{t-1}(y) \prod_{j=1}^{d} q_j^{t-1}(x_j^{(i)}|y)}$$

GMM Bayes E-Step:

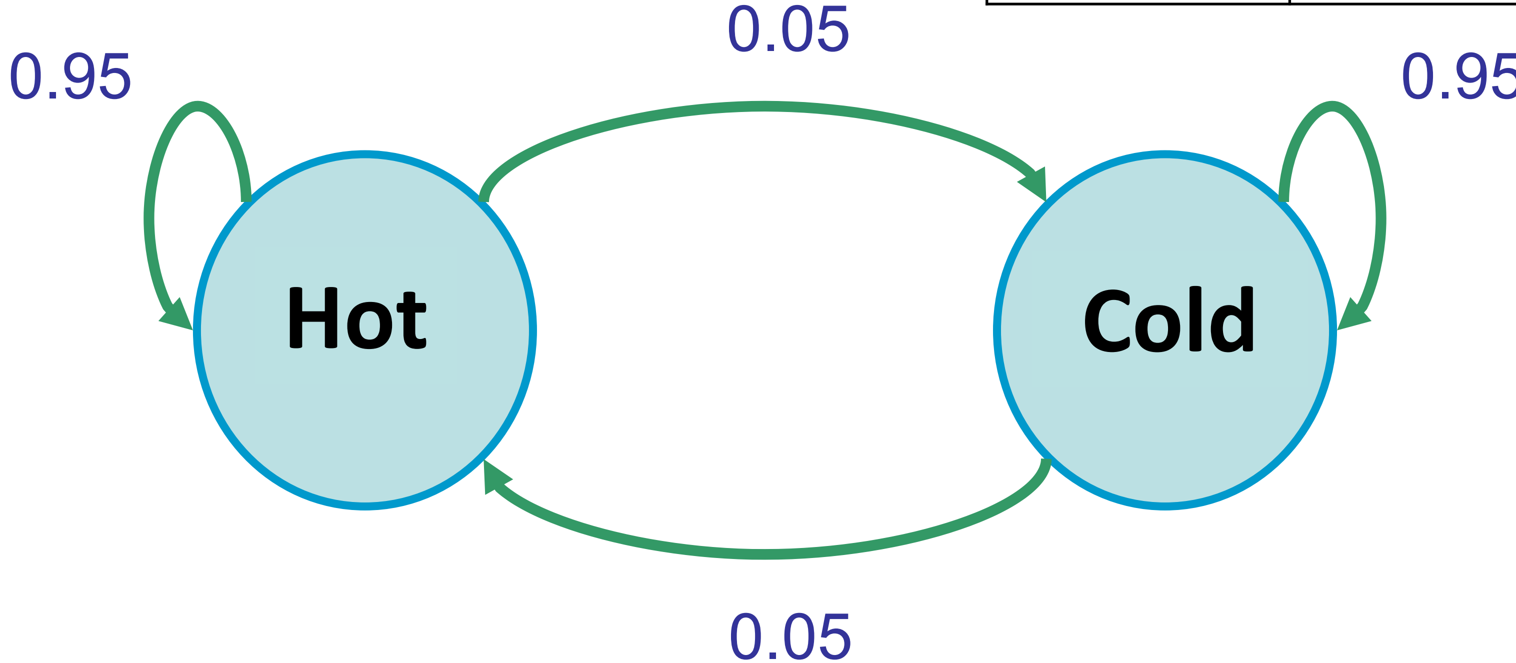
$$\gamma(z_{n,k}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n \mid \mu_j, \Sigma_j)}$$

How many possible assignments of Z?



Example: Ice Cream!

	$P(z_n = k \mid z_{n-1})$
$z_{n-1}=k$	0.8
$z_{n-1} \,!= k$	0.2



	$P(x_n=1 \mid z_n)$	$P(x_n=2 \mid z_n)$	$P(x_n=3 \mid z_n)$
z_{n-1}=Hot	0.1	0.2	0.7
z_{n-1}=Cold	0.7	0.2	0.1