The problem with statements like "No party candidate has won the election without [state]." or "No president has been reelected under [circumstances]."

1804... No incumbent has beaten a challenger.
1808... No congressman has ever become president.
1812... No one can win without New York.
1816... No candidate who doesn't wear a wig can get elected.
1820... No one who wears pants instead of breeches can be reelected.
1824... No one has ever won without a popular majority.

... until Jefferson.
... until Madison.
... but Madison did.
... until Monroe was.
... but Monroe was.
... J.Q. Adams did.

1972... Quakers can't win twice.
1976... No one who lost New Mexico has won.
1980... No one has been elected president after a divorce.
1984... No left-handed president has been reelected.
1988... No one with two middle names has become president.
1992... No democrat has won without a majority of the Catholic vote.

... until Nixon did.
... but Carter did.
... until Reagan was.
... until Reagan was.
... until "Herbert Walker."
... until Clinton did.

1996... No dem. incumbent without combat experience has beaten someone whose first name is worth more in Scrabble.
2000... No republican has won without Vermont.
2004... No republican without combat experience has beaten someone two inches taller.
2008... No democrat can win without Missouri.
2012... Democratic incumbents never beat taller challengers.

No nominee whose first name contains a 'K' has lost.

Which streak will break?
Questions in Statistical Learning Theory:

– How good is the learned rule after n examples?
– How many examples do I need before the learned rule is accurate?
– What can be learned and what cannot?
– Is there a universally best learning algorithm?

In particular, we will address:

What is the true error of $h$ if we only know the training error of $h$?

– Finite hypothesis spaces and zero training error
– Finite hypothesis spaces and non-zero training error
– Infinite hypothesis spaces and VC dimension
Can you Convince me of your Psychic Abilities?

• Game
  – I think of n bits
  – If somebody in the class guesses my bit sequence, that person clearly has telepathic abilities – right?

• Question:
  – If at least one of $|H|$ players guesses the bit sequence correctly, is there any significant evidence that he/she has telepathic abilities?
  – How large would n and $|H|$ have to be?
Discriminative Learning and Prediction Reminder

• Goal: Find $h$ with small prediction error $Err_p(h)$ over $P(X,Y)$.

• Discriminative Learning: Given $H$, find $h$ with small error $Err_{S_{\text{train}}}(h)$ on training sample $S_{\text{train}}$.

• Training Error: Error $Err_{S_{\text{train}}}(h)$ on training sample.

• Test Error: Error $Err_{S_{\text{test}}}(h)$ on test sample is an estimate of $Err_p(h)$
Definition: A particular instance of a learning problem is described by a probability distribution $P(X,Y)$.

Definition: A sample $S = ((\tilde{x}_1, y_1), ..., (\tilde{x}_n, y_n))$ is independently identically distributed (i.i.d.) according to $P(X,Y)$.

Definition: The error on sample $S$ $Err_S(h)$ of a hypothesis $h$ is $Err_S(h) = \frac{1}{n} \sum_{i=1}^{n} \Delta(h(\tilde{x}_i), y_i)$.

Definition: The prediction/generalization/true error $Err_P(h)$ of a hypothesis $h$ for a learning task $P(X,Y)$ is

$$Err_P(h) = \sum_{\tilde{x} \in X, y \in Y} \Delta(h(\tilde{x}), y)P(X = \tilde{x}, Y = y).$$

Definition: The hypothesis space $H$ is the set of all possible classification rules available to the learner.
Generalization Error Bound: Finite H, Zero Error

- **Setting**
  - Sample of $n$ labeled instances $S_{\text{train}}$
  - Learning Algorithm $L$ with a finite hypothesis space $H$
  - At least one $h \in H$ has zero prediction error $Err_P(h) = 0$ ($\Rightarrow Err_{S_{\text{train}}}(h) = 0$)
  - Learning Algorithm $L$ returns zero training error hypothesis $\hat{h}$

- What is the probability that the prediction error of $\hat{h}$ is larger than $\varepsilon$?

$$P(Err_P(\hat{h}) \geq \varepsilon) \leq \frac{H}{n} e^{-\varepsilon n}$$

**Diagram**

- **Training Sample** $S_{\text{train}}$  
  $(x_1, y_1), \ldots, (x_n, y_n)$

- **Learner**

- **Test Sample** $S_{\text{test}}$  
  $(x_{n+1}, y_{n+1}), \ldots$
Useful Formulas

- **Binomial Distribution**: The probability of observing $x$ heads in a sample of $n$ independent coin tosses, where in each toss the probability of heads is $p$, is

$$P(X = x|p,n) = \frac{n!}{r!(n-r)!} p^x (1-p)^{n-x}$$

- **Union Bound**: 

$$P(X_1 = x_1 \lor X_2 = x_2 \lor \cdots \lor X_n = x_n) \leq \sum_{i=1}^{n} P(X_i = x_i)$$

- **Unnamed**: 

$$(1 - \epsilon) \leq e^{-\epsilon}$$
Sample Complexity: Finite H, Zero Error

- **Setting**
  - Sample of $n$ labeled instances $S_{\text{train}}$
  - Learning Algorithm $L$ with a finite hypothesis space $H$
  - At least one $h \in H$ has zero prediction error ($\Rightarrow Err_{S_{\text{train}}}(h)=0$)
  - Learning Algorithm $L$ returns zero training error hypothesis $\hat{h}$

- How many training examples does $L$ need so that with probability at least $(1-\delta)$ it learns an $\hat{h}$ with prediction error less than $\varepsilon$?

\[ n > \frac{1}{\varepsilon} \left( \log(|H|) - \log(\delta) \right) \]

**Diagram:**
- **Training Sample** $S_{\text{train}}$: $(x_1, y_1), \ldots, (x_n, y_n)$
- **Learner**
- **Test Sample** $S_{\text{test}}$: $(x_{n+1}, y_{n+1}), \ldots$
**Definition:**  
*C is PAC-learnable* by learning algorithm $\mathcal{L}$ using $H$ and a sample $S$ of $n$ examples drawn i.i.d. from some fixed distribution $P(X)$ and labeled by a concept $c \in C$, if for sufficiently large $n$

$$P(\text{Err}_P(h_{\mathcal{L}(S)}) \leq \varepsilon) \geq (1 - \delta)$$

for all $c \in C$, $\varepsilon > 0$, $\delta > 0$, and $P(X)$. $\mathcal{L}$ is required to run in polynomial time dependent on $1/\varepsilon$, $1/\delta$, $n$, the size of the training examples, and the size of $c$. 