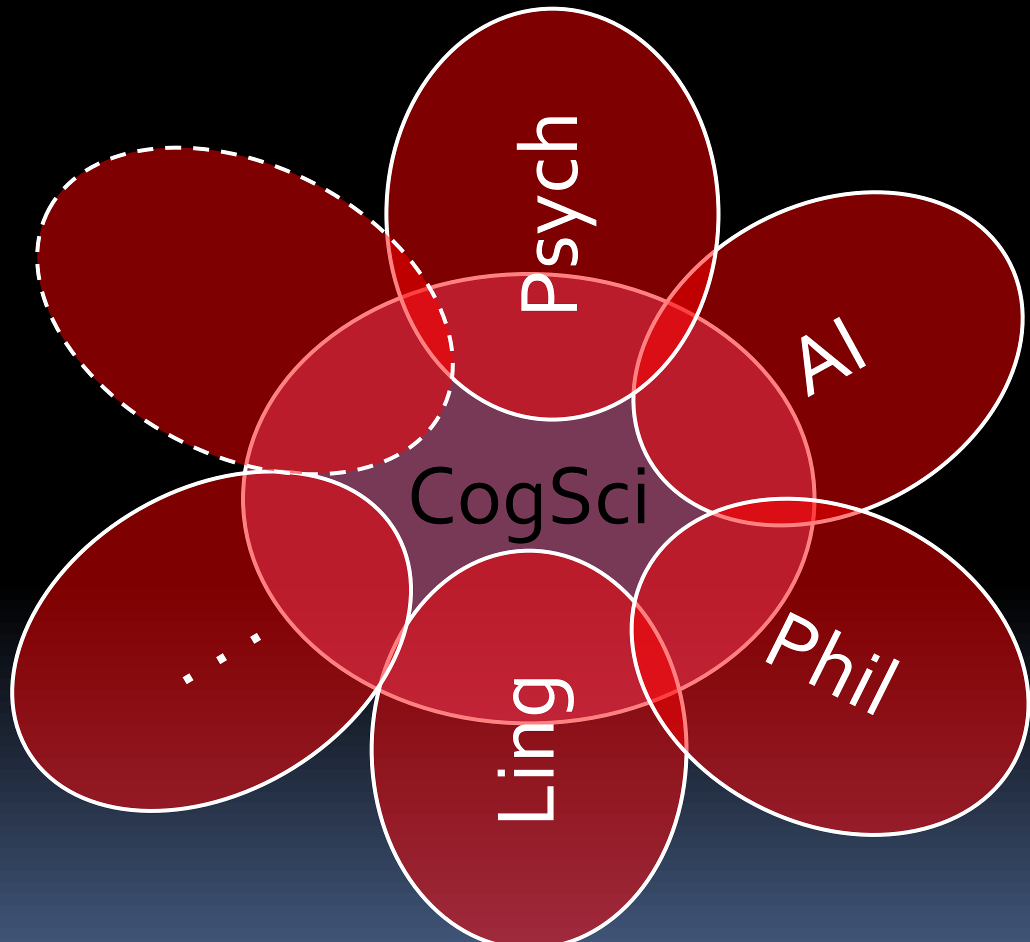
ARTIFICIAL INTELLIGENCE, COGNITIVE SCIENCE, AND EXPLANATION-BASED LEARNING

Gerald DeJong
University of Illinois

Thanks to
Shiau Hong Lim, Geoff Levine, Li Lun Wang,
Arkady Epshteyn, Qiang Sun, Adam Laud…
NSF, ONR, DARPA,…

Cognitive Science Petal Model



What is AI?

- Many diverse approaches
- Central focus: Implementability
  - Cognitive guidance / intuition / heuristic
- Model-Based Reasoning
  - What is a model?

Illustrations primarily from image processing
Message concerns
knowledge representation, expressiveness, & inference
Applies broadly to AI generally
Largely a personal interpretation

Old AI

- Deep, componential logic-like symbolic models
- Expressive knowledge representations
- Deductive inference (satisfiability)
- Impressive examples
- Laughable real-world behavior

Impressive NLP examples from Stone Age AI

Winograd, 1972; Schank et al, 1975

Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.

Shrdlu: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

Person: OTHELLO STRANGLED DESDEMONA.

Person: WHY DID DESDEMONA DIE?

Margie: DESDEMONA DIED BECAUSE SHE COULD NOT BREATHE AND SHE COULD NOT BREATHE BECAUSE OTHELLO CHOKED HER.

Person: JOHN DROVE HIS CAR TO BUY GROCERIES.

Shrdlu / Margie / … : READ UNHAPPY – MAKNAM
WARNING: INTERNED OBJECT NOT FOUND ON OBLIST

- WYSIWYG artificial intelligence systems
- “Brittleness is due to the lack of humans’ encyclopedic knowledge…”
- But there was a deeper flaw
- Generally applies to strong prior-model AI:
  planning, expert systems, machine learning, robotics, natural language processing, vision, …

Deeper Flaw:
No Formal Connection to the Modeled World

- Semantics of Logic representations
  - Δ is the Model: Constraints on Possible Worlds
  - Satisfiability / Set Membership: $\Delta \models \Phi$
  - Cars move; …
- The Qualification Problem

  "Most universally quantified expressions intended to apply to the real world require an unbounded number of qualifying expressions"
  - Birds Fly: $\forall x\ Bird(x) \Rightarrow Flies(x)$
- Formal properties all w.r.t. Δ (micro-world)
- Connection to the modeled world is beyond the scope of the formal theory

Computer Vision: classify by intrinsic features

Beach, but what features?
Define visually
    waves?   palm trees?
Tedious
Fragile behavior on new images
No significant improvement

Computer Vision left the AI fold early

WYSIWYG Computer Vision is not easily excused.

Beach / City / Kitchen; Face / Airplane / Motorcycle…

Beach

City

Kitchen

In search of Robust Behavior

- Robotics also soon broke away from AI
- NLP followed
  - Statistical approach
  - Robustness via Machine Learning
    (finding patterns in data)
  - Weaker prior model (e.g., bag of words)
  - fit to observations (training data)
- Gradual but persistent shift to statistical ML in AI
- Rely more on observations; less on prior commitments
- The weaker the better (the more robust)

FORMAL CONNECTION TO THE "REAL" (MODELED) WORLD

Input representation (raw features) → **CLASSIFIER** → Classification

Palm Trees? – No, too hard to find reliably

Good Features:

- Need not be semantically interpretable
- Easy to find
- Expose *some* information
- Learn to appreciate information

Examples:

Bag of words	Dot Prod. Kernels
Codewords	Clusters of patches

Transformed classification problem

- Much harder for people
- Much easier for computers

Weak Models

- Specialize w/ training examples
  - Training examples = labeled world observations
  - Formal connection to the modeled world
  - Optimization / parameter estimation (not satisfiability)
  - Statistical inference (weaker guarantees than logical entailment)

- NLP categorize newsgroup articles
  - Counts of reasonable words
  - Article -> Bag of words -> Newsgroup

- Support Vector Machines
  - RKHS – dot product kernels
  - Features + large margin classifiers together

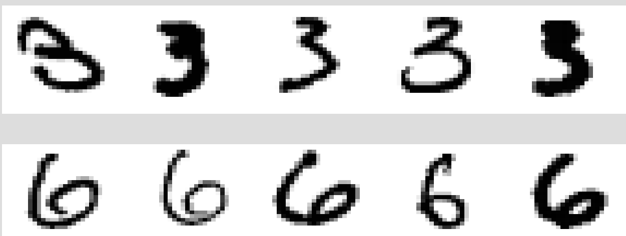
- Image classification
  - Clusters of patches / Codewords

Computer Strengths & Weakness

- Doing it “Like a person” may be a mistake

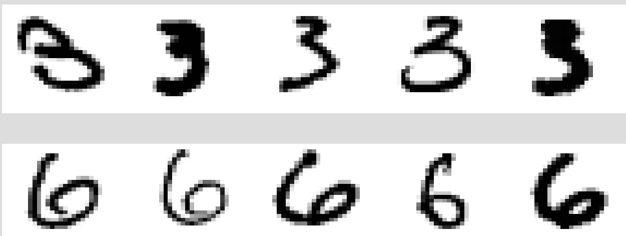
- Support Vector Machine
- Very high (infinite?) dimensional feature space
- Classifier
  - Linear in high dimensional space
  - Non-linear in original space
- Kernel function
  - Dot product of example to be classified
  - Applied to a few training examples (support vectors)

Two Related Classification Problems



	No. examples	error
Humans	< 100 ?	negligible
SVMs	60000	1.2%

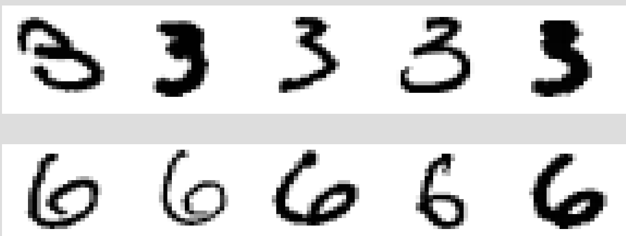
Two Related Classification Problems



a fixed permutation over pixels

	No. examples	error
Humans	< 100 ?	negligible
SVMs	60000	1.2%

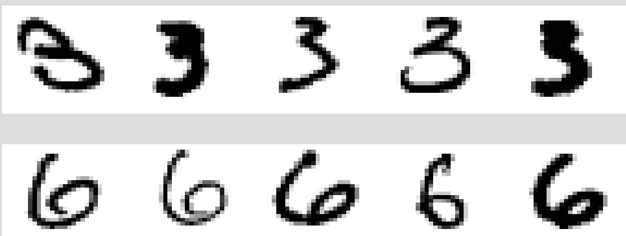
Two Related Classification Problems



a fixed permutation over pixels

	No. examples	error	No. examples	error
Humans	< 100 ?	negligible		
SVMs	60000	1.2%		

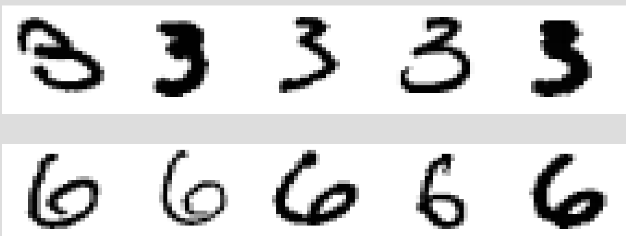
Two Related Classification Problems



a fixed
permutation
over pixels

	No. examples	error	No. examples	error
Humans	< 100 ?	negligible	NA	50%
SVMs	60000	1.2%		

Two Related Classification Problems



a fixed permutation over pixels

	No. examples	error	No. examples	error
Humans	< 100 ?	negligible	NA	50%
SVMs	60000	1.2%	60000	1.2%

To an SVM these are the *same* problem

Vapnik, LeCun, others

A More Difficult Task

- Offline handwritten Chinese characters

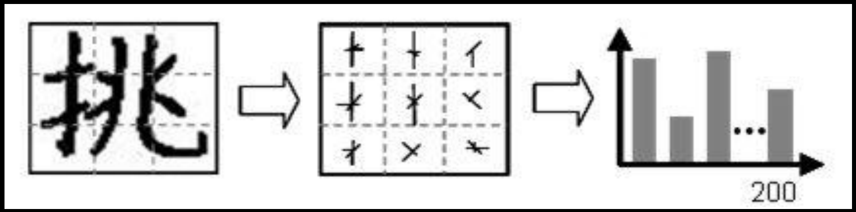
- Thousands of complex patterns
- Hundreds of training examples
- Large within-class differences

- Conventional SVM on pixels performs poorly
- Better features
  - Reduce dimensionality
  - improve signal / noise ratio

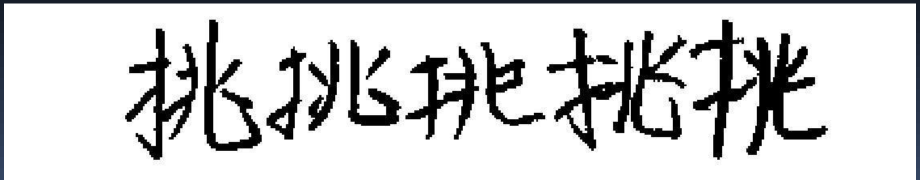
WDH Features (internal representation)

Kimura et al, 1997, Liu & Ding, 2005

- Register image
- Tile into windows (5 x 5 thru 11 x 11)
- Stretch / shrink to normalize intensities
- Histogram of oriented edges for each tile



- Features not easily semantically interpretable
- Form a distribution for each character class



- Classify by most likely distribution membership

Performance
(pairs of isolated Chinese characters)

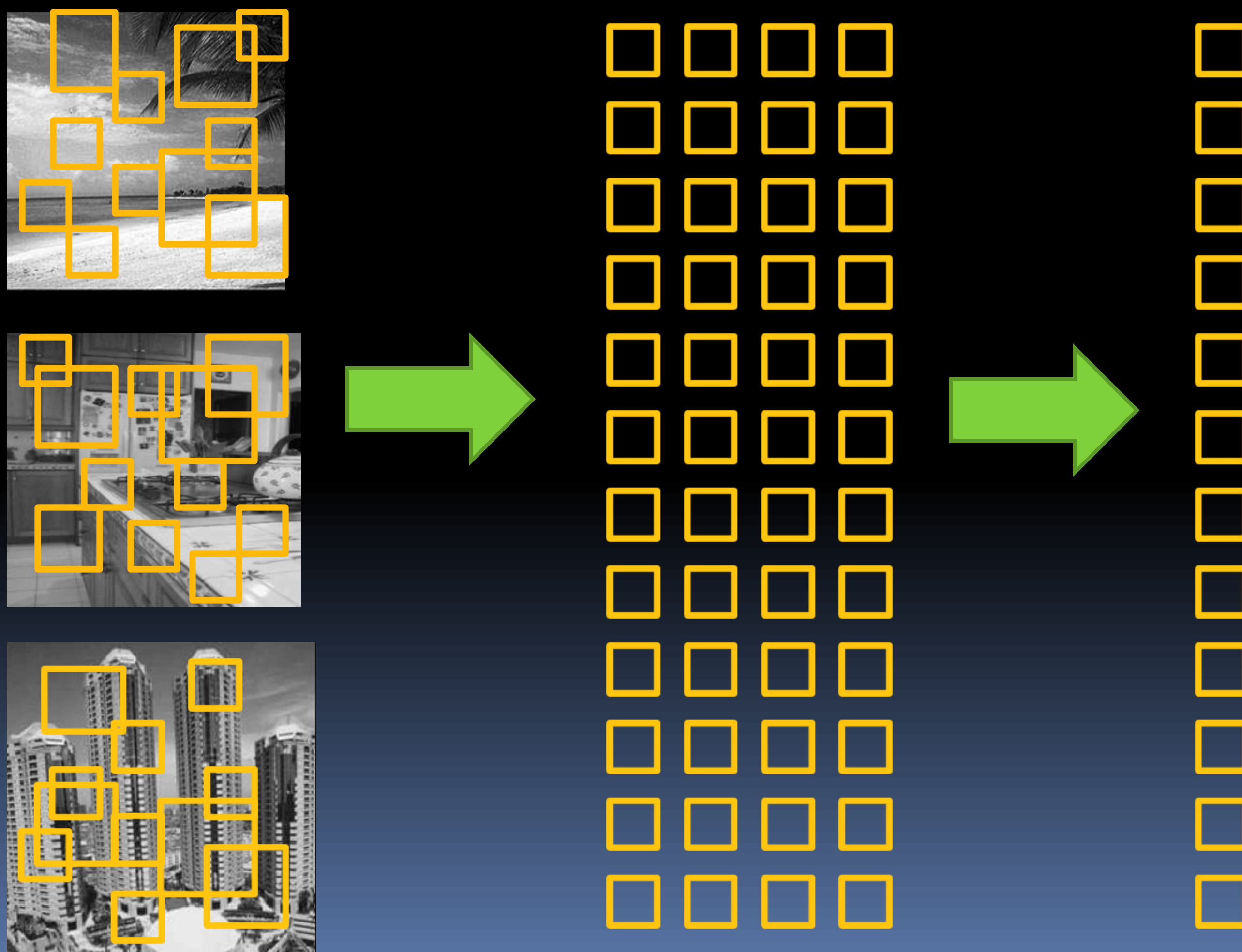
- Extremely good
  - >98% on easier datasets
  - ~90% on harder datasets
  - For most characters
- Bothersome: Quite poor on some character pairs
- Worse: No obvious extension (more later)

Weak Model Approach to Image Features

Collect Patch Samples

Normalize

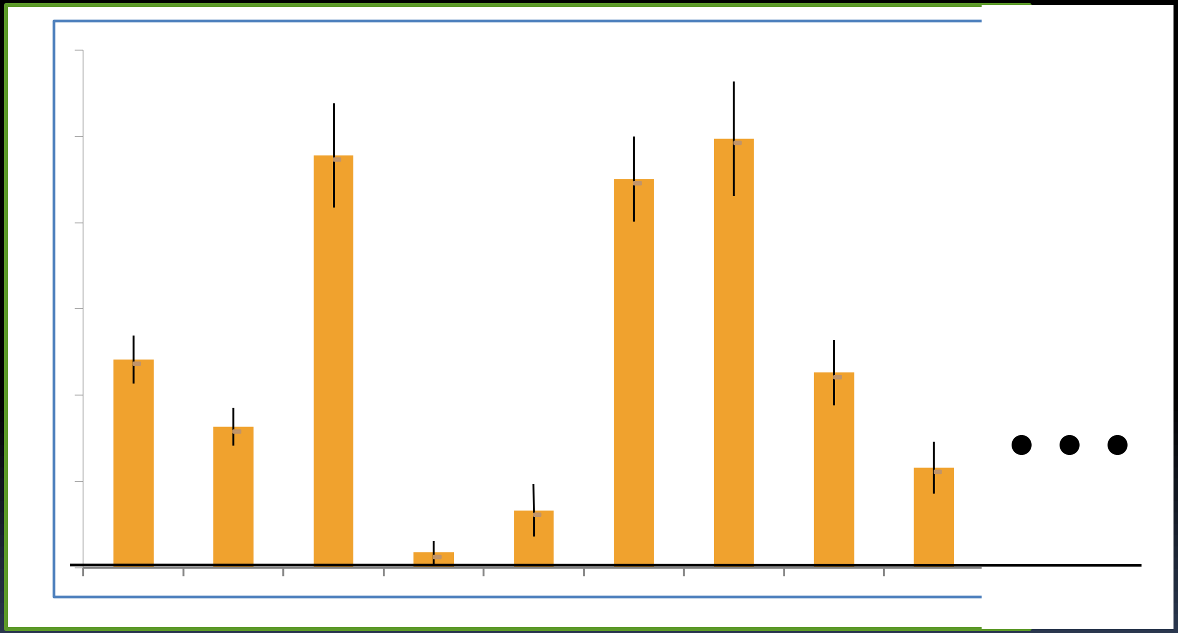
Form Clusters

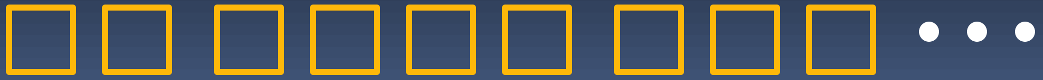


Class Conditional Distributions

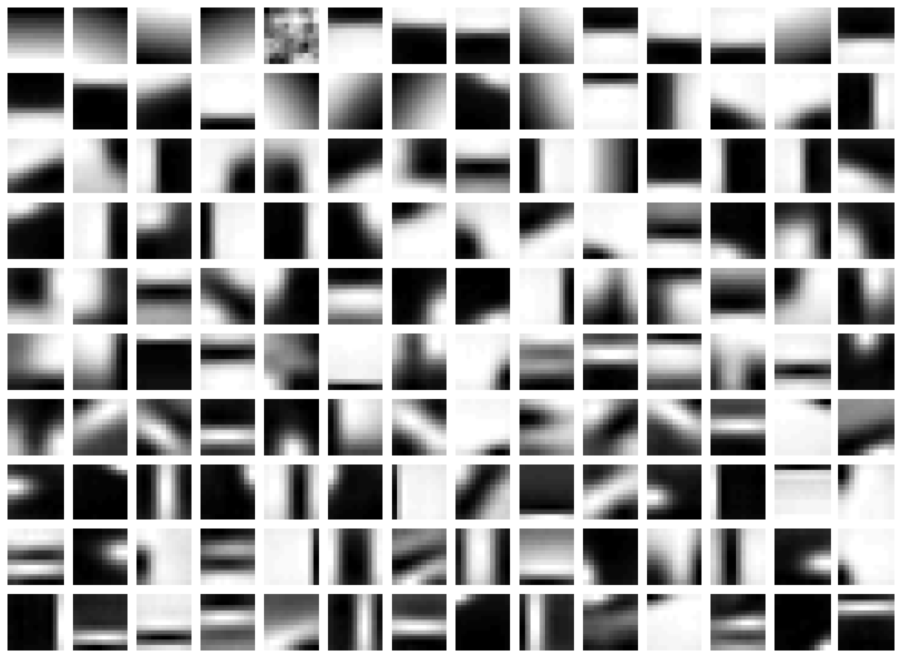
Beaches

Distribution of clusters

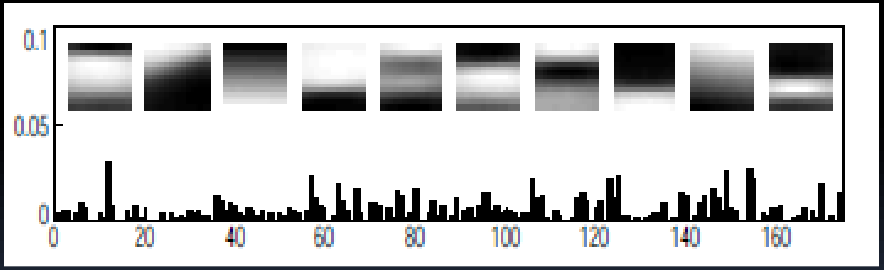




Codewords
Cluster Averages



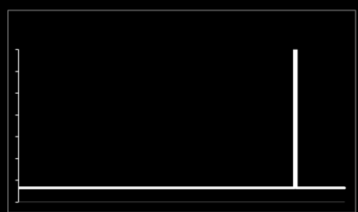
New Image



Best fits “beach” distribution of codewords

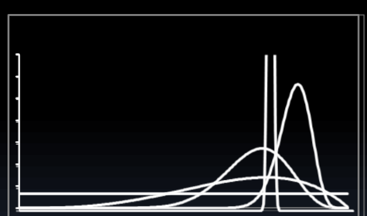
Old / New AI: a Useful Metaphor

Old Strong Model AI



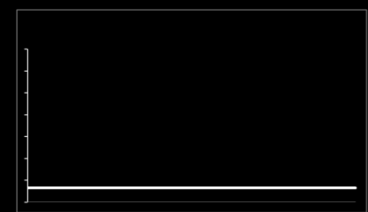
Fully formed complete model *a priori*

“Data? We don’t need no stinkin’ data”

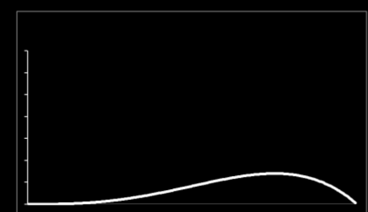


But little chance of obtaining the fully trained weak model

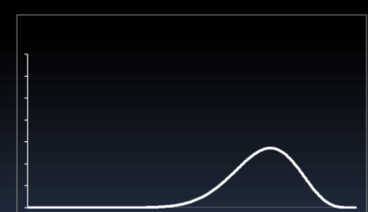
New Weak Model AI



Flexible model

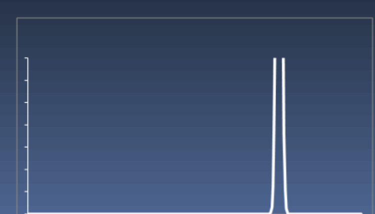


Initially no strong opinions on the world



With training stronger opinions emerge

Eventually much stronger opinions



Primarily formed by training data

Training Images

Fei Fei & Perona, 2005

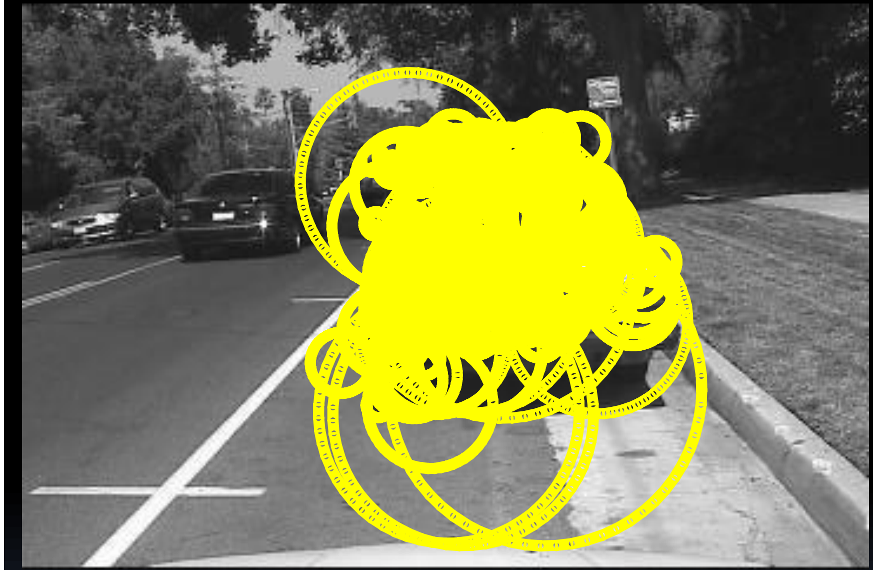
Faces

Motorbikes

Airplanes

Cars

What is this?



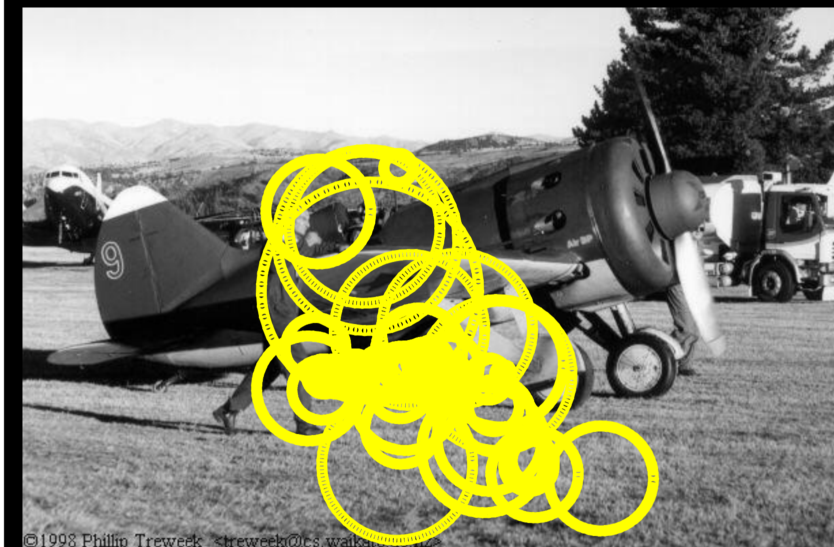
"It's a car"

What are these?

"It's a face"

"It's a motorbike"

What is this?



“It’s a face”

What did this state-of-the-art award-winning system find?

Felzenszwalb et al, 2008

Performance / Consequences

- Accuracy is quite high
- Robust on new inputs
- Capitalizes on computers strengths

BUT

- These are simple tasks

  (more in a moment)
- Non-human mistakes
- Fragile decisions even when correct
- Human behavior is quite different

Prior / Posterior Models

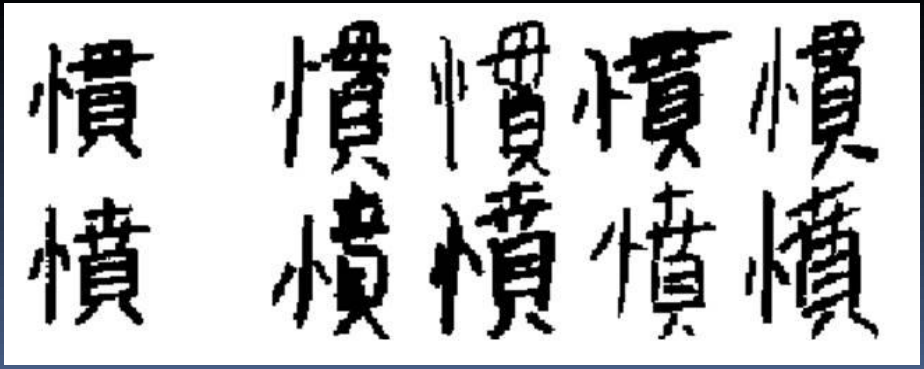
- Prior inductive bias
  - Necessary / Unavoidable
  - *Tabula rasa* learning is mathematically inconsistent
- Trained weak models work because
  - Careful prior choices (inductive bias)
  - Simple tasks
    - Classes are very different
    - Features provide low-grade class information
    - But are plentiful, almost everywhere you look

Codewords = Hubel & Wiesel –like features?

- No,
- not exactly,
- well perhaps,
- kind of, yes…
- What are we hoping for?
- A hint of trouble

A More Difficult Task

- These are difficult even with WDH features
- Small class differences
  Large individual variation



Another Difficult Task

Which is the Ketch,
Which is the Schooner?

- Patches / clusters drawn from almost identical populations
- Low signal to noise ratio
- The rule is…

Tuning Weak Models for Difficult Problems

- Relies on
  - Pre-defined features
  - Training observations
- Need (much) better features
  - Uncomfortably close to old AI
- Or (dramatically) more examples
  - Uncomfortably close to *tabula rasa* learning
  - Learn brain surgery by example?

$$\min_{\mathcal{D}_{H*}} KL(\mathcal{D}_{H*}, \mathcal{D}_{H^o}) \leq \mathbb{I}(Z)$$

EBL++ Explanation Based Learning w/ Statistics

- Learning brain surgery $\neq$ Inventing brain surgery
- Use available prior knowledge
  - Componential sub-models
  - Combinatorial inference
  - Can be First Order Logic sentences (not necessary)
  - Weaker semantics
- Combine "deductively" only to *explain* observations
  - Combine = well-formed assemblies (e.g., unification)
  - $\forall x\ Bird(x) \Rightarrow Flies(x)$
  - Have Bird, may conjecture it Flies if leads to an observed / assigned feature

Importance of Prior Domain Knowledge

"Depends if fore mast or aft mast is higher"

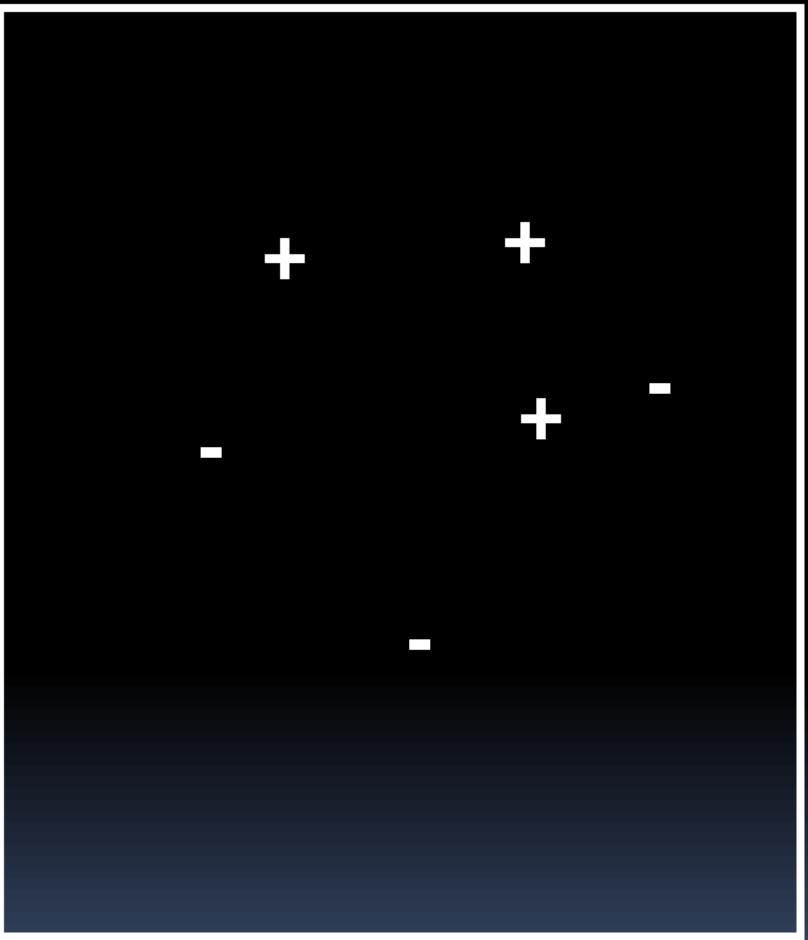
- Does not specify the answer
- Does not change the training observations
- But ketch/schooner becomes an easy problem
- Prior knowledge can magnify information of training examples

"Schooners are faster;
Ketches are more maneuverable"

$$\min_{\mathcal{D}_{H*}} KL(\mathcal{D}_{H*}, \mathcal{D}_{Ho}) \leq \mathbb{I}(\Delta \cup Z)$$

Potentially $\mathbb{I}(\Delta \cup Z) >> \mathbb{I}(\Delta) + \mathbb{I}(Z)$

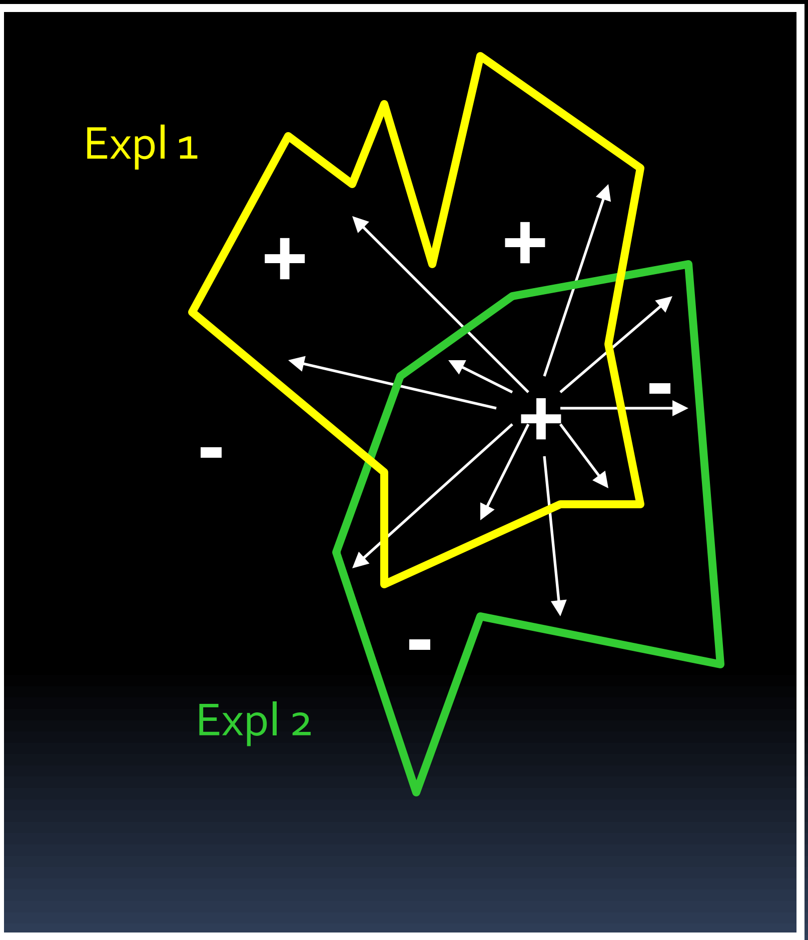
Conventional Machine Learning



Example Space X

- Training examples in a high-dimensional space
- Boundaries emerge
- Confidence increases with more examples

EBL++



Example Space X

- Build plausible explanations
- Each commits to additional labelings
- Rate each by posterior
- Learner should behave as if all explained data is seen, weighted by posterior

$$\mathbf{E}=\{E_i: \subseteq \Delta \cup X(d_k) \vdash_{\Delta} L(d_k)\}$$

$$C = \underset{h_j \in H}{\text{argmin}} \left\| \sum_i \Pr(E_i \mid D, \Delta) \,.\, E_i - h_j \right\|$$

Explanation Based Features

- Semantic features
- Automatically purpose-designed
- Learn
  - How to estimate / evaluate
  - How to combine
  - NOT their existence
- Find easy-to-find features, even if no class info
- Conceptually tied to hard-to-find class info features

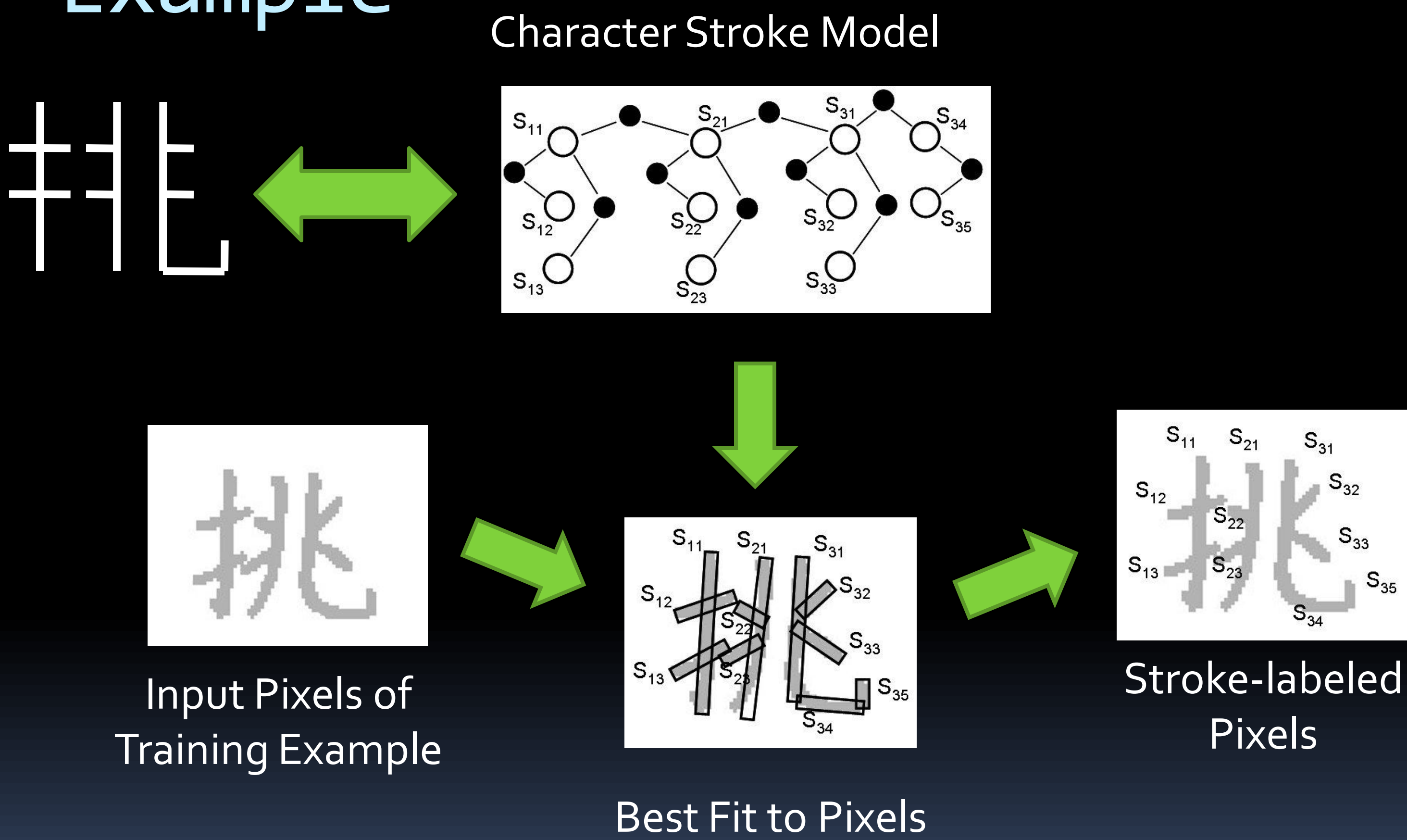
Example I: EBL Features in Chinese Characters

- Domain theory
  - Characters made of strokes
  - Pixels arise from strokes
    Strokes interact
  - Some strokes are easier to find than others
- Don’t need to invent notion of strokes from data

Prior Knowledge

- Ideal stroke model of each character
- Essentially a vector font
- Some strokes are easy to find (but not informative of class)
- Others are class-informative (but hard to find)
- Use explanation as bridge
- EBF: find what you don't care about but will help to find what you do

Explanation of a Labeled Example



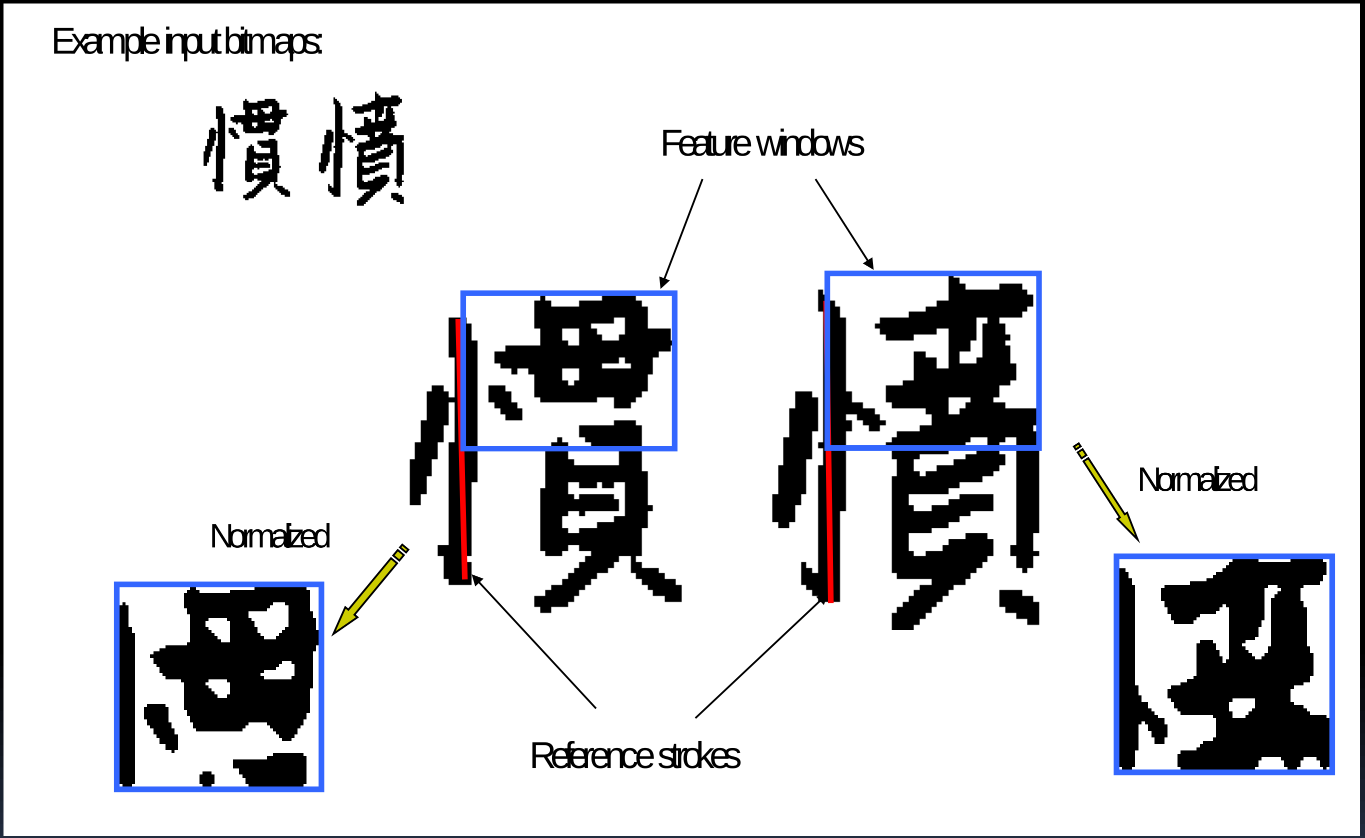
Finding Strokes from Pixels

- Much easier if stroke necessarily exists
- Easier if the stroke is longer
- Easier if near an edge
- We use a Hough transform
- Fitting pixels exposes stroke parameters
  - Angle, Length, Location, Width

Explanation Based Features

- Inferentially connected
  - perceptual features – class neutral, easy to find
  - concpetual features – class informative, hard to find
  - Use perceptual features to find conceptual ones
- Difficult = low signal / noise;
- Must have significant shared structure
- Find reference strokes:
  - Shared, Long, Near image edge
- Estimate high-information regions from reference

Explanation Based Feature Creation



Features are complex but automatically tailored to classification problem

Illustration on Ten Char 2 Test Images

SVM with WDH features

Char 1

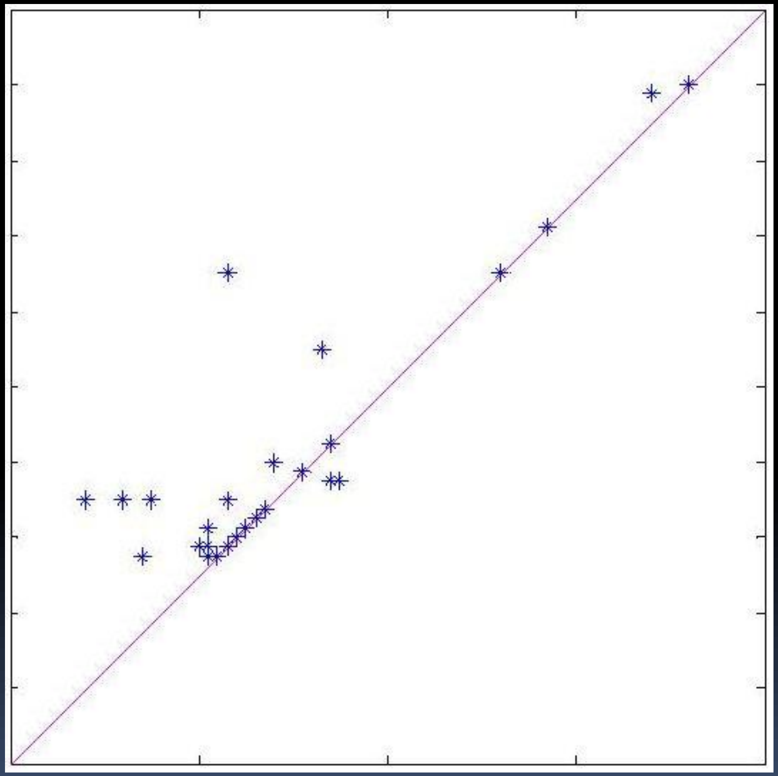
Char 2

Original	Window	Normalized	Full	EBF
			OK	OK
			ERROR	OK
			OK	OK
			OK	OK
			OK	OK
			ERROR	OK
			OK	OK
			OK	OK
			ERROR	OK
			ERROR	OK

Result on 30 hardest pairs

Lim et al, 2007

ETL9B database



Error Rate – WDH with Full Image

Error Rate – WDH with EBL Feature

Many Possible Improvements

- Disambiguate reference strokes
- Use multiple reference strokes
- Use multiple focus windows
- Entertain small strokes within focus window
- …

Example II: Ketch / Schooner

- Relative fore/aft mast height is important
- Mast height related to sail height
- Unfurled sails are like reference strokes
- Estimate Aft / Fore by geometry
- Boat = Hull + Sails
- Chooses
  - relative horizontal position
  - Highest point on sail
  - Sail center of geometry

Ketch / Schooner

Levine et al, 2008

Ketch

Schooner

Cognitive Implications

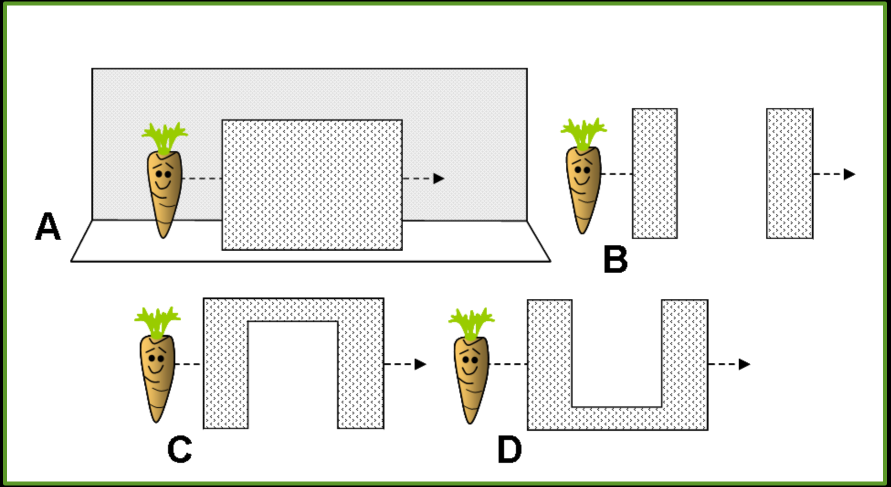
- Developmental evidence?
- We examine infant behavior
  - Best chance for statistically driven weak model
  - Prior theory for EBL?
  - Core knowledge, moderate nativist accounts
  - Exploit wife's expertise
- Examine / interpret existing results

EBL++ vs. Conventional ML
Differential Predictions

- Conventional statistical ML
  - Initially a weak flexible model
  - Improvement from accretion of observations
  - Uniform accuracy convergence in expectation
  - More examples
- EBL++
  - Explanations conjecture medium-strength models
  - Calibrated explanations may be eclipsed or rejected
  - Features come and go with explanations
  - Punctuated non-uniform accuracy improvement
  - Conceptual features

Occlusion Task 2.5-3.5 months

Baillargeon, 2004



Habituate to A

At 2.5 mo. surprised

- at non-appearance in B
- at appearance in C

At 3.0 mo. surprised

- at non-appearance in C
- at appearance in D

At 3.5 mo. mature behavior

Two weeks! (at home)

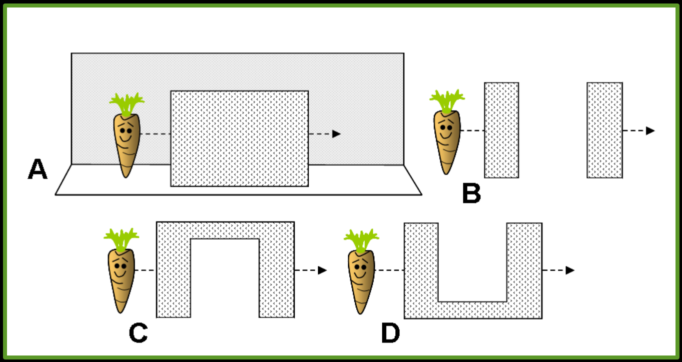
Very few examples

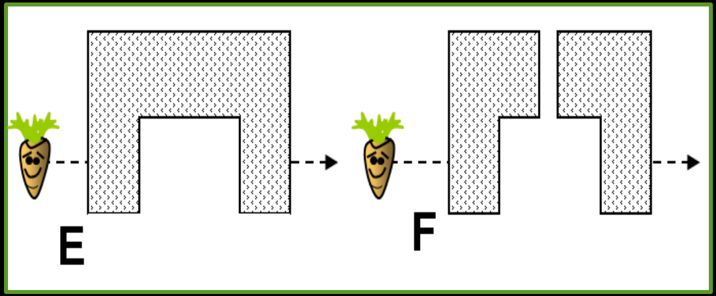
None with these objects

No explicit training

And yet, active opinions incompatible with any observations

At 2.5 Months





- Consider screens E and F w/ smaller object
- E is like C and babies (incorrectly) believe they both behave like A and should occlude
- F is E with a small (irrelevant) gap at the top
- But babies believe it should be like B and (correctly) expect it not to occlude
- Distinctions seem conceptual and not solely due to data accretion

Conclusions

- AI should drift back towards Cognitive Science
- Perhaps EBL++ can help
- Again a role for prior knowledge
- Combines deductive and inductive reasoning
- Promise of greater example efficiency
- Ongoing developmental investigations

Thank You!