NLP & Linguistics

Natural Language Processing
CS 4120/6120—Spring 2017
Northeastern University

David Smith
some slides from
Jason Eisner, Chris Manning & Roger Levy

Engineering vs. Science?

- One story
  - NLP took formal language theory and generative linguistics (same source?),
  - Built small AI systems for a while,
  - Then added statistics/machine learning (from speech recognition).
- What now?
  - Shouldn't AI tell us about natural intelligence?
  - Are all NLP models lousy linguistics?

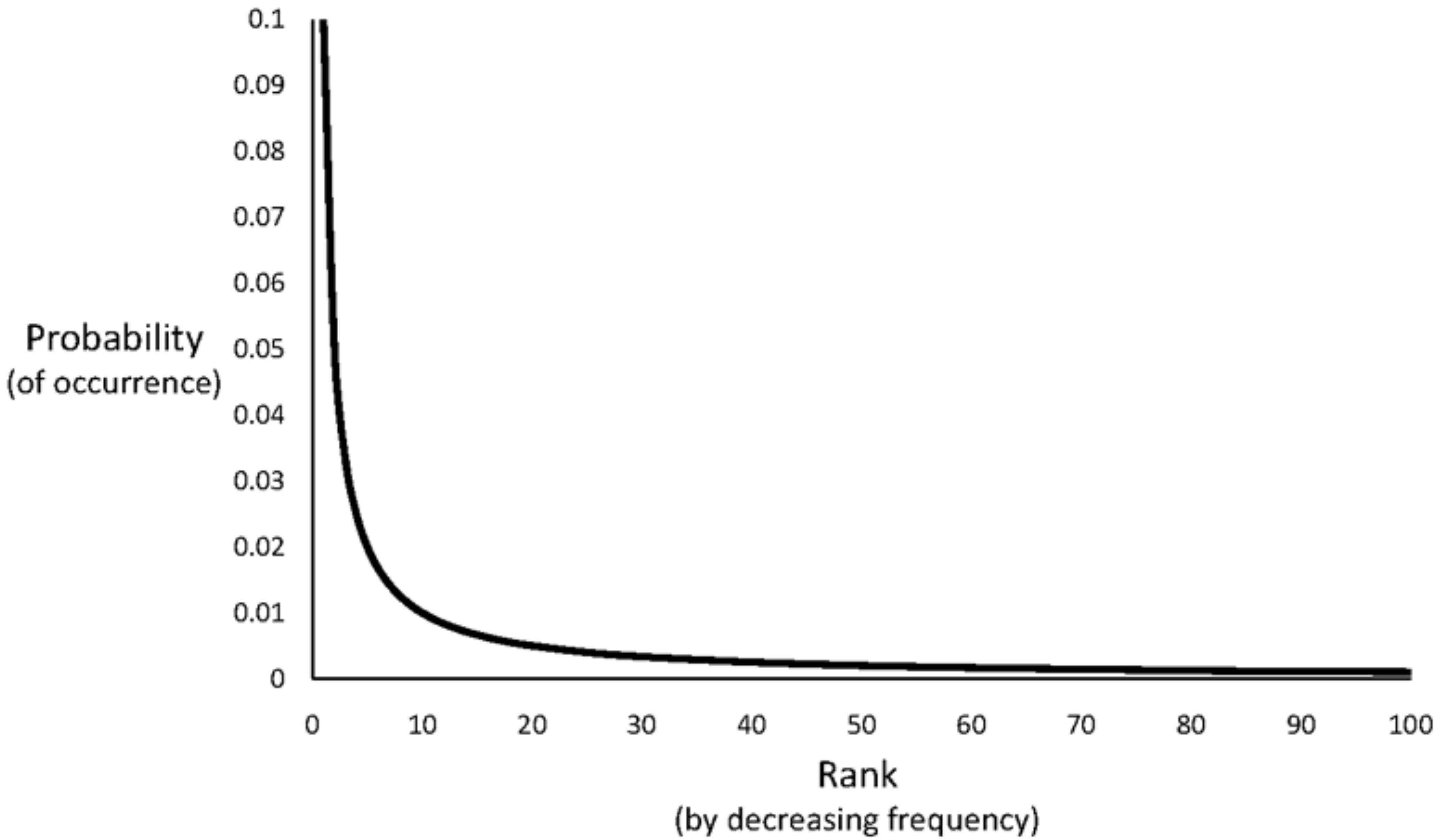
Zipf’s Law

The Roots of Quantitative Linguistics

Zipf’s Law

- Distribution of word frequencies is very *skewed*
  - a few words occur very often, many words hardly ever occur
  - e.g., two most common words (“the”, “of”) make up about 10% of all word occurrences in text documents
- Zipf’s “law” (more generally, a “power law”):
  - observation that rank (r) of a word times its frequency (f) is approximately a constant (k)
    - assuming words are ranked in order of decreasing frequency
  - i.e., $r.f \approx$ k or $r.P_r \approx c$, where P_r is probability of word occurrence and $c \approx 0.1$ for English

Zipf's Law


0.1
0.09
0.08
0.07
0.06
Probability
(of occurrence)
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30
40
50
60
70
80
90
100
Rank
(by decreasing frequency)

News Collection (AP89) Statistics

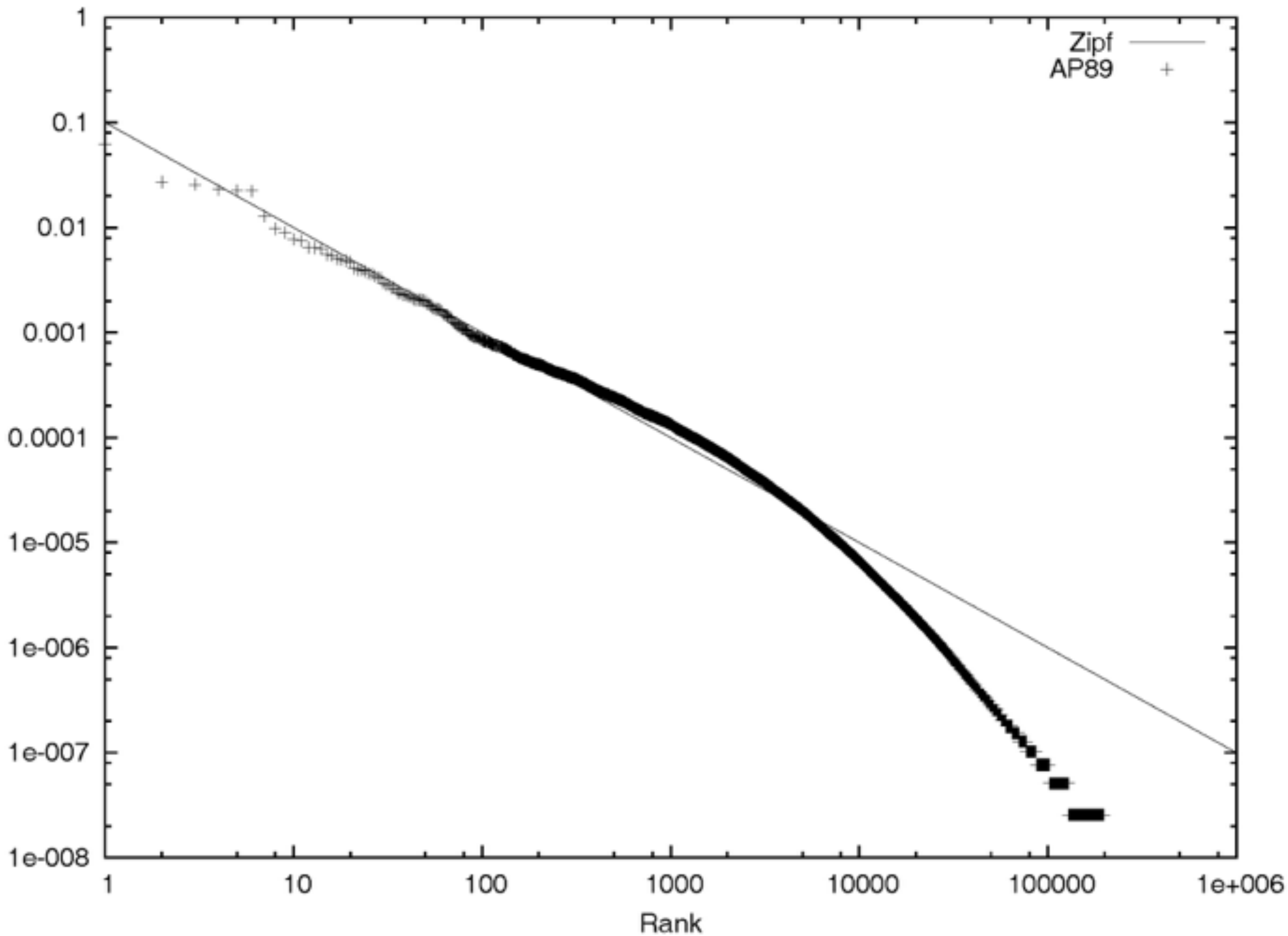
Total documents	84,678
Total word occurrences	39,749,179
Vocabulary size	198,763
Words occurring > 1000 times	4,169
Words occurring once	70,064

Word	*Freq.*	*r*	*Pr(%)*	*r.Pr*
assistant	5,095	1,021	.013	0.13
sewers	100	17,110	$2.56 \times 10-4$	0.04
toothbrush	10	51,555	$2.56 \times 10-5$	0.01
hazmat	1	166,945	$2.56 \times 10-6$	0.04

Top 50 Words from AP89

Word	*Freq.*	*r*	$P_r(\%)$	$r.P_r$	*Word*	*Freq*	*r*	$P_r(\%)$	$r.P_r$
the	2,420,778	1	6.49	0.065	has	136,007	26	0.37	0.095
of	1,045,733	2	2.80	0.056	are	130,322	27	0.35	0.094
to	968,882	3	2.60	0.078	not	127,493	28	0.34	0.096
a	892,429	4	2.39	0.096	who	116,364	29	0.31	0.090
and	865,644	5	2.32	0.120	they	111,024	30	0.30	0.089
in	847,825	6	2.27	0.140	its	111,021	31	0.30	0.092
said	504,593	7	1.35	0.095	had	103,943	32	0.28	0.089
for	363,865	8	0.98	0.078	will	102,949	33	0.28	0.091
that	347,072	9	0.93	0.084	would	99,503	34	0.27	0.091
was	293,027	10	0.79	0.079	about	92,983	35	0.25	0.087
on	291,947	11	0.78	0.086	i	92,005	36	0.25	0.089
he	250,919	12	0.67	0.081	been	88,786	37	0.24	0.088
is	245,843	13	0.65	0.086	this	87,286	38	0.23	0.089
with	223,846	14	0.60	0.084	their	84,638	39	0.23	0.089
at	210,064	15	0.56	0.085	new	83,449	40	0.22	0.090
by	209,586	16	0.56	0.090	or	81,796	41	0.22	0.090
it	195,621	17	0.52	0.089	which	80,385	42	0.22	0.091
from	189,451	18	0.51	0.091	we	80,245	43	0.22	0.093
as	181,714	19	0.49	0.093	more	76,388	44	0.21	0.090
be	157,300	20	0.42	0.084	after	75,165	45	0.20	0.091
were	153,913	21	0.41	0.087	us	72,045	46	0.19	0.089
an	152,576	22	0.41	0.090	percent	71,956	47	0.19	0.091
have	149,749	23	0.40	0.092	up	71,082	48	0.19	0.092
his	142,285	24	0.38	0.092	one	70,266	49	0.19	0.092
but	140,880	25	0.38	0.094	people	68,988	50	0.19	0.093

Zipf's Law for AP89



- Log-log plot: Note problems at high and low frequencies

Zipf's Law

- What is the proportion of words with a given frequency?
  - Word that occurs n times has rank $r_n = k/n$
  - Number of words with frequency n is
    - $r_n - r_{n+1} = k/n - k/(n + 1) = k/n(n + 1)$
  - Proportion found by dividing by total number of words = highest rank = k
  - So, proportion with frequency n is $1/n(n+1)$

Zipf's Law

- Example word frequency ranking

Rank	*Word*	*Frequency*
1000	concern	5,100
1001	spoke	5,100
1002	summit	5,100
1003	bring	5,099
1004	star	5,099
1005	immediate	5,099
1006	chemical	5,099
1007	african	5,098

- To compute number of words with frequency 5,099
  – rank of "chemical" minus the rank of "summit"
  – 1006 − 1002 = 4

Example

Number of Occurrences (n)	*Predicted Proportion (1/n(n+1))*	*Actual Proportion*	*Actual Number of Words*
1	.500	.402	204,357
2	.167	.132	67,082
3	.083	.069	35,083
4	.050	.046	23,271
5	.033	.032	16,332
6	.024	.024	12,421
7	.018	.019	9,766
8	.014	.016	8,200
9	.011	.014	6,907
10	.009	.012	5,893

- Proportions of words occurring *n* times in 336,310 TREC documents
- Vocabulary size is 508,209

Vocabulary Growth

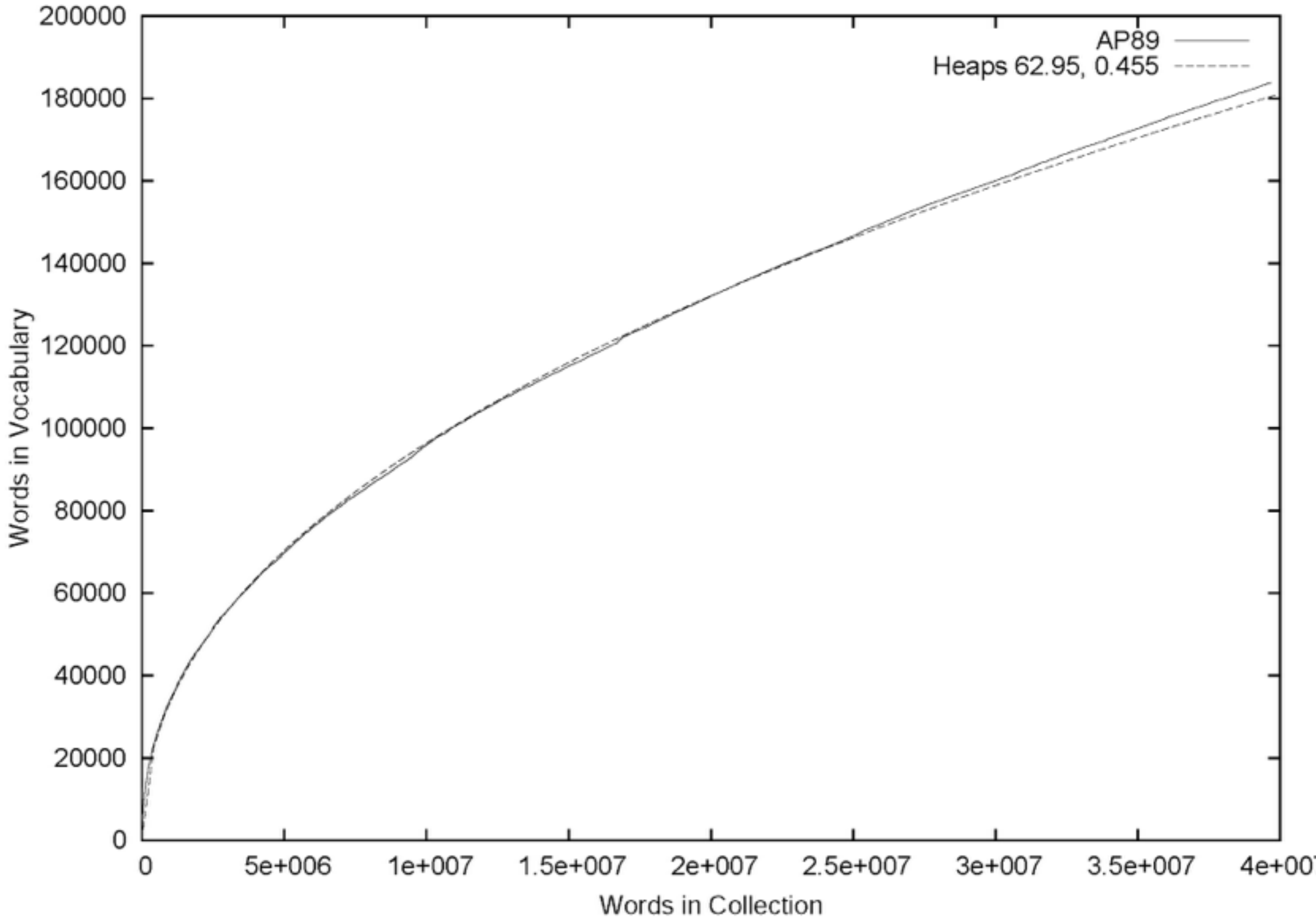
- As corpus grows, so does vocabulary size
  - Fewer new words when corpus is already large
- Observed relationship (*Heaps' Law*):

$$v = k.n^{\beta}$$

where v is vocabulary size (number of unique words), n is the number of words in corpus,

k, β are parameters that vary for each corpus (typical values given are $10 \le k \le 100$ and $\beta \approx 0.5$)

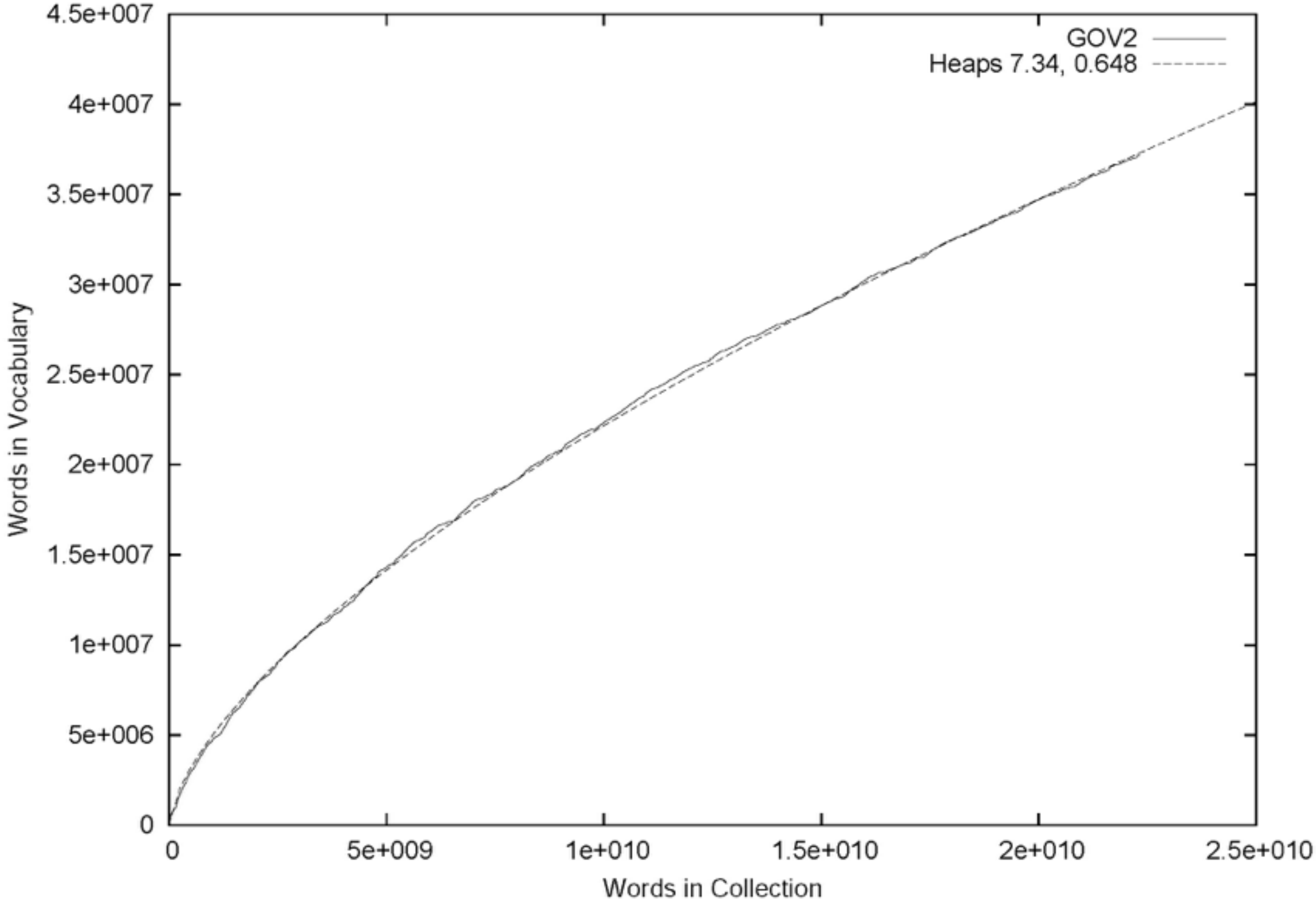
AP89 Example



Heaps' Law Predictions

- Predictions for TREC collections are accurate for large numbers of words
  - e.g., first 10,879,522 words of the AP89 collection scanned
  - prediction is 100,151 unique words
  - actual number is 100,024
- Predictions for small numbers of words (i.e. < 1000) are much worse

GOV2 (Web) Example



Ever Upwards

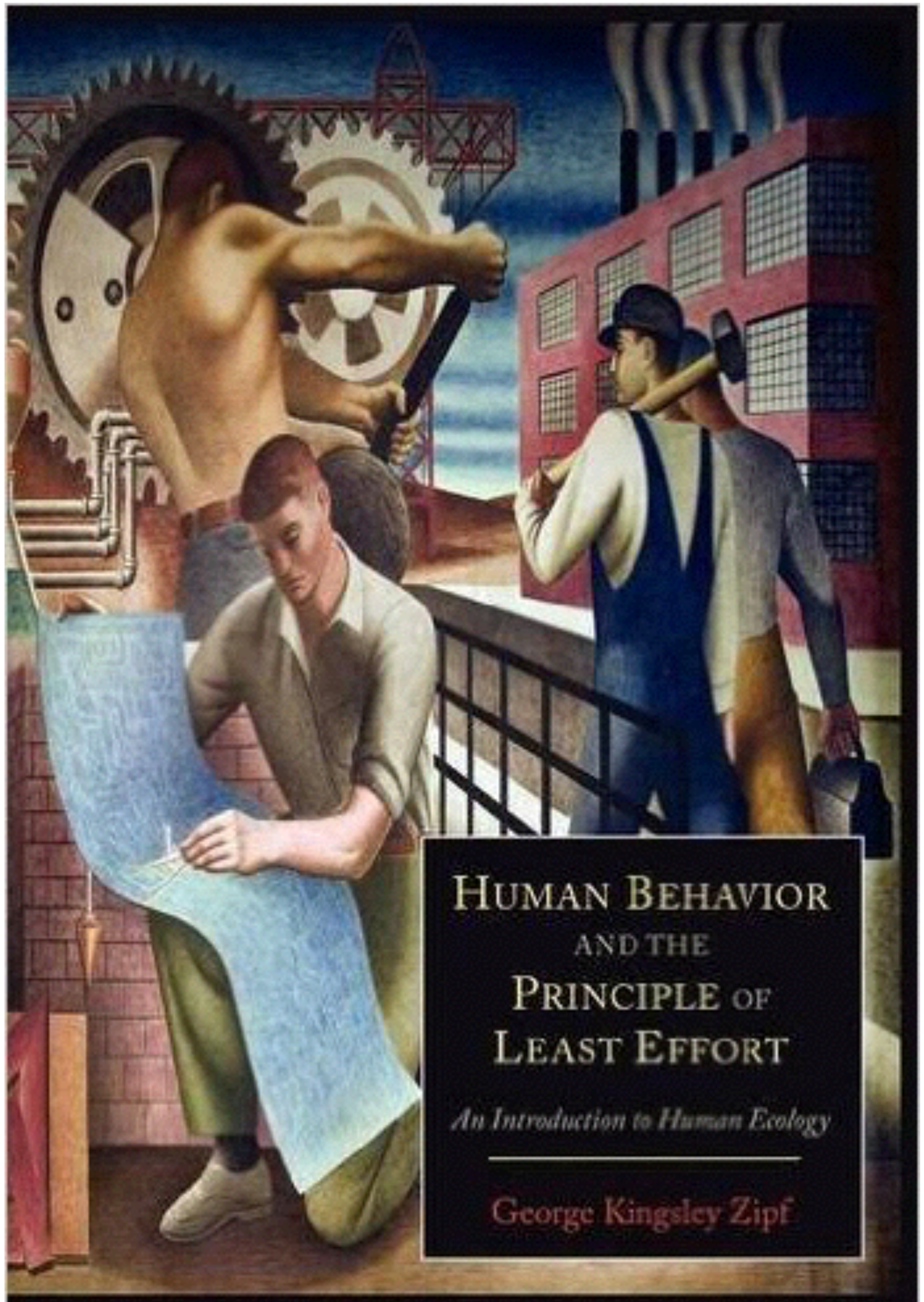
- Heaps' Law works with very large corpora
  - new words occurring even after seeing 30 million!
  - parameter values different than typical TREC values
- New words come from a variety of sources
  - spelling errors, invented words (e.g. product, company names), code, other languages, email addresses, etc.
- Language models (and other NLP and IR systems) need to handle open, growing vocabulary

Power-Law Distributions

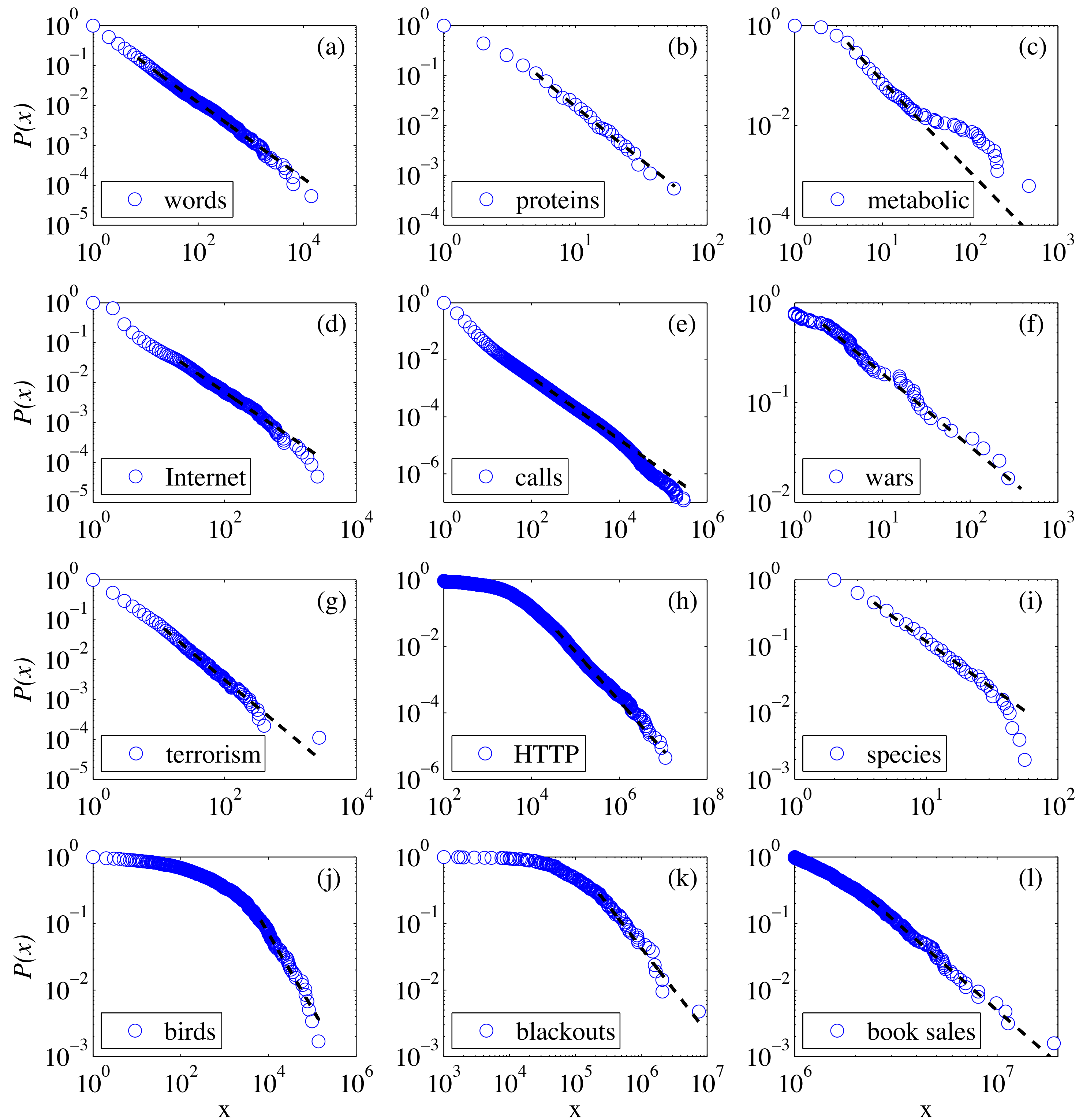
- For discrete data *(Clauset et al., 2009)*:

$$p(x) = \Pr(X = x) = Cx^{-\alpha}$$

- which diverges at 0, thus requiring a lower bound $x_{\min} > 0$
- which normalizes to $p(x) = \dfrac{x^{-\alpha}}{\zeta(\alpha, x_{\min})}$
- with Hurwitz zeta $\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha}$

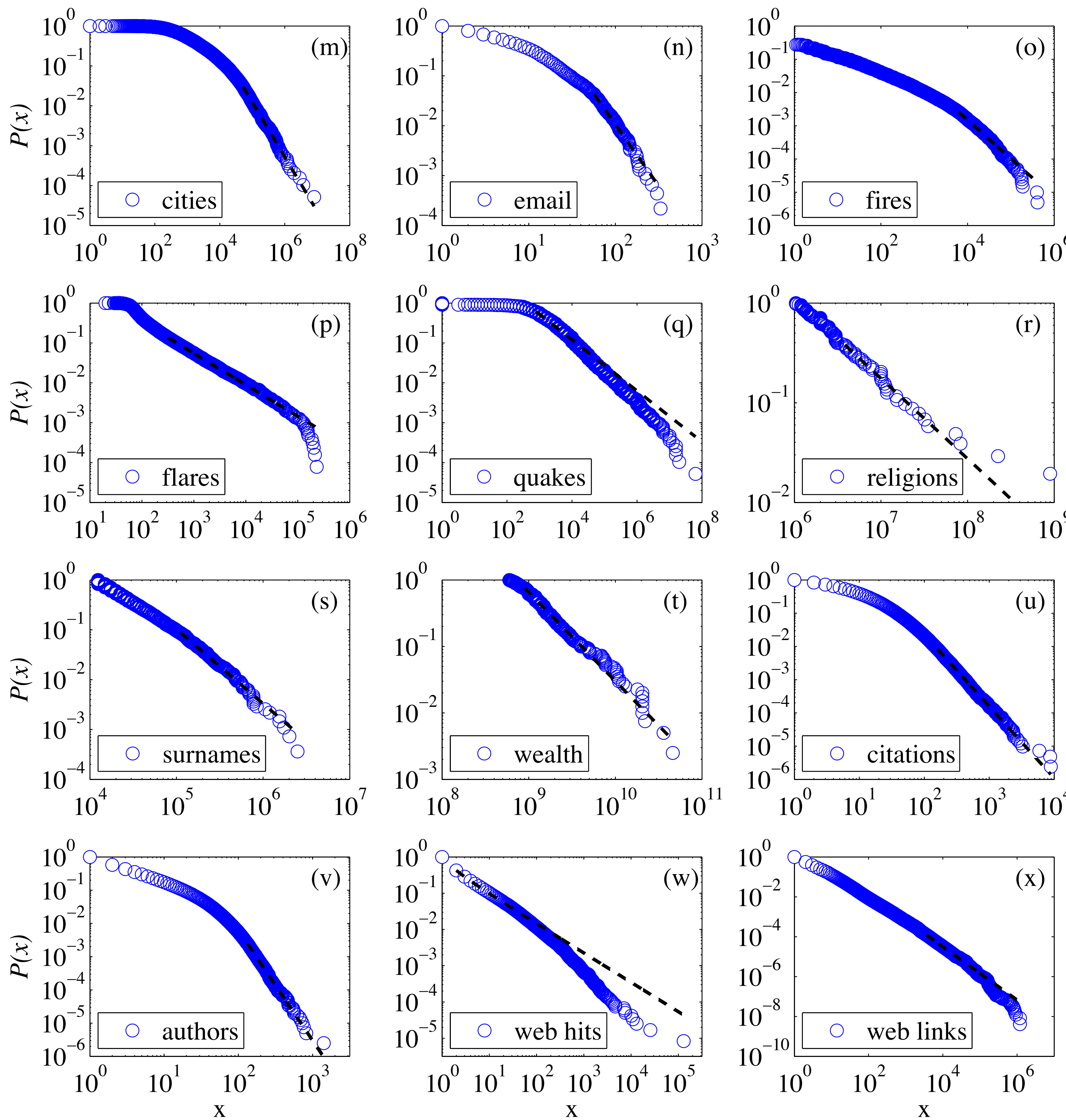

HUMAN BEHAVIOR
AND THE
PRINCIPLE OF
LEAST EFFORT
An Introduction to Human Ecology
George Kingsley Zipf

Power Laws Everywhere!



(Clauset et al., 2009)

Power Laws Everywhere!



(Clauset et al., 2009)

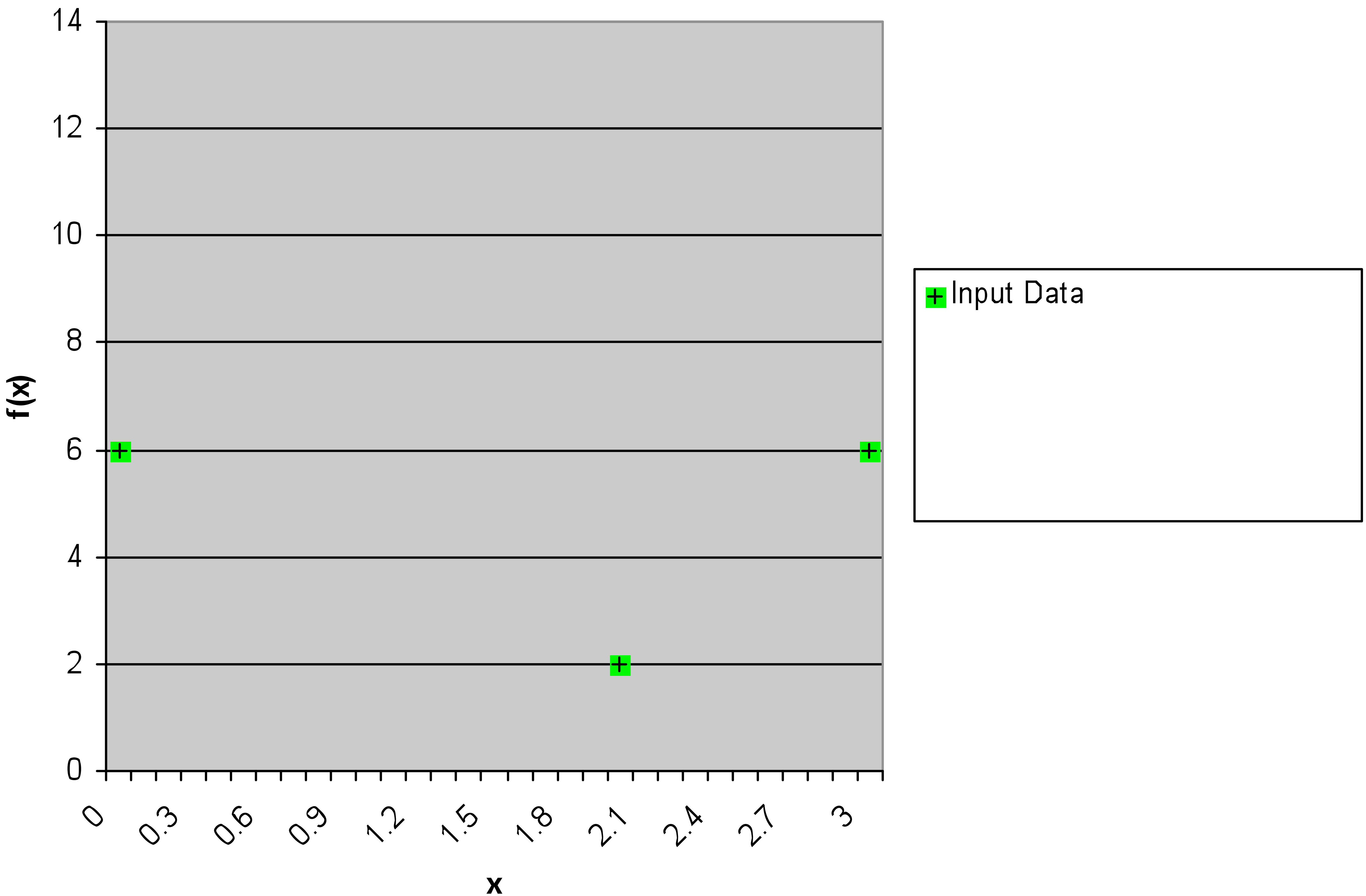
Power Laws Everywhere?

		Poisson		log-normal		exponential		stretched exp.		power law + cut-off		support for
data set	p	LR	p	LR	p	LR	p	LR	p	LR	p	power law
Internet	**0.29**	5.31	**0.00**	−0.807	0.42	6.49	**0.00**	0.493	0.62	−1.97	**0.05**	with cut-off
calls	**0.63**	17.9	**0.00**	−2.03	**0.04**	35.0	**0.00**	14.3	**0.00**	−30.2	**0.00**	with cut-off
citations	**0.20**	6.54	**0.00**	−0.141	0.89	5.91	**0.00**	1.72	**0.09**	−0.007	0.91	moderate
email	**0.16**	4.65	**0.00**	−1.10	0.27	0.639	0.52	−1.13	0.26	−1.89	**0.05**	with cut-off
metabolic	0.00	3.53	**0.00**	−1.05	0.29	5.59	**0.00**	3.66	**0.00**	0.000	1.00	none
papers	**0.90**	5.71	**0.00**	−0.091	0.93	3.08	**0.00**	0.709	0.48	−0.016	0.86	moderate
proteins	**0.31**	3.05	**0.00**	−0.456	0.65	2.21	**0.03**	0.055	0.96	−0.414	0.36	moderate
species	**0.10**	5.04	**0.00**	−1.63	0.10	2.39	**0.02**	−1.59	0.11	−3.80	**0.01**	with cut-off
terrorism	**0.68**	1.81	**0.07**	−0.278	0.78	2.457	**0.01**	0.772	0.44	−0.077	0.70	moderate
words	**0.49**	4.43	**0.00**	0.395	0.69	9.09	**0.00**	4.13	**0.00**	−0.899	0.18	good

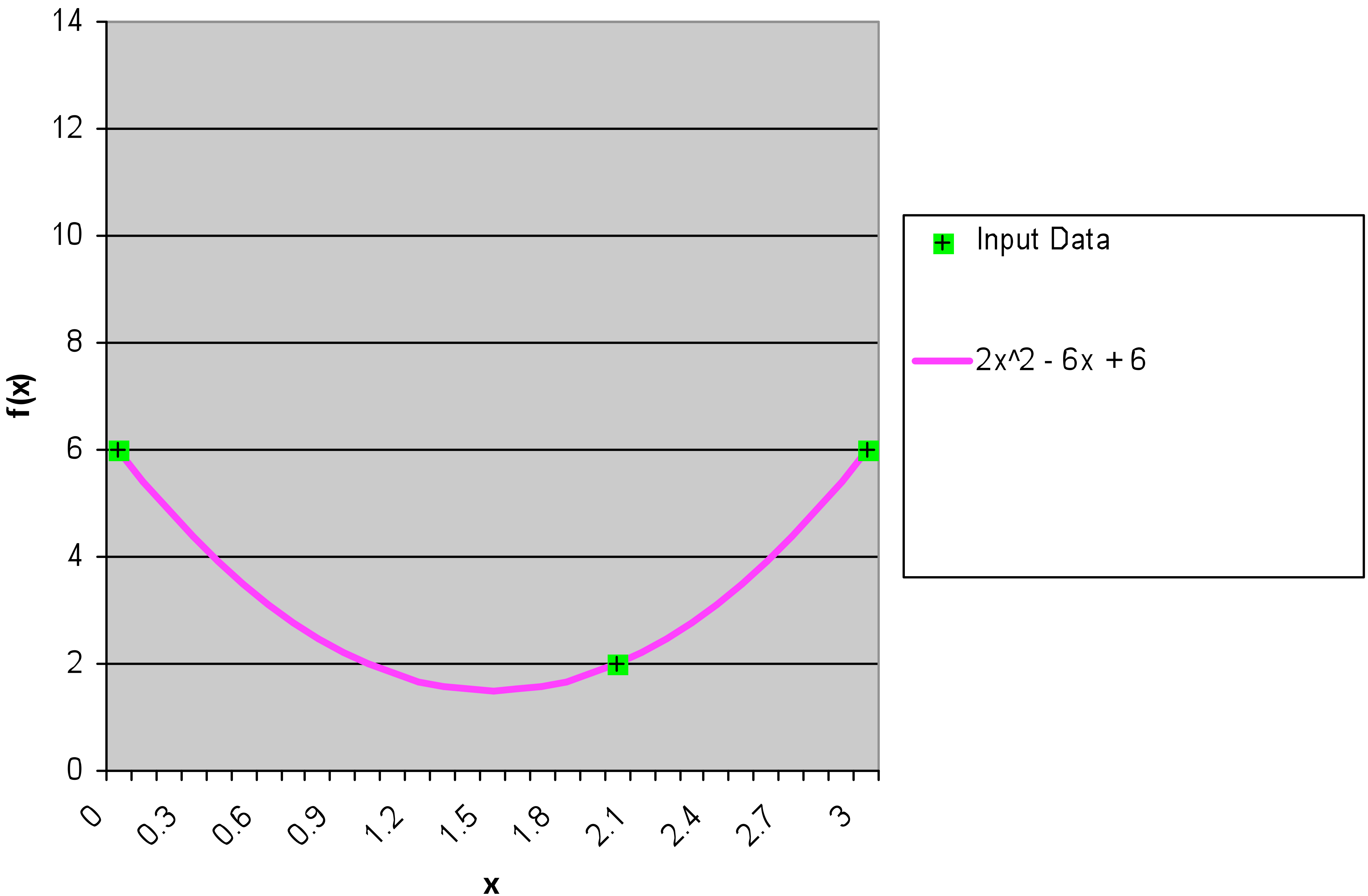
(Clauset et al., 2009)

Learning in the Limit
Gold's Theorem

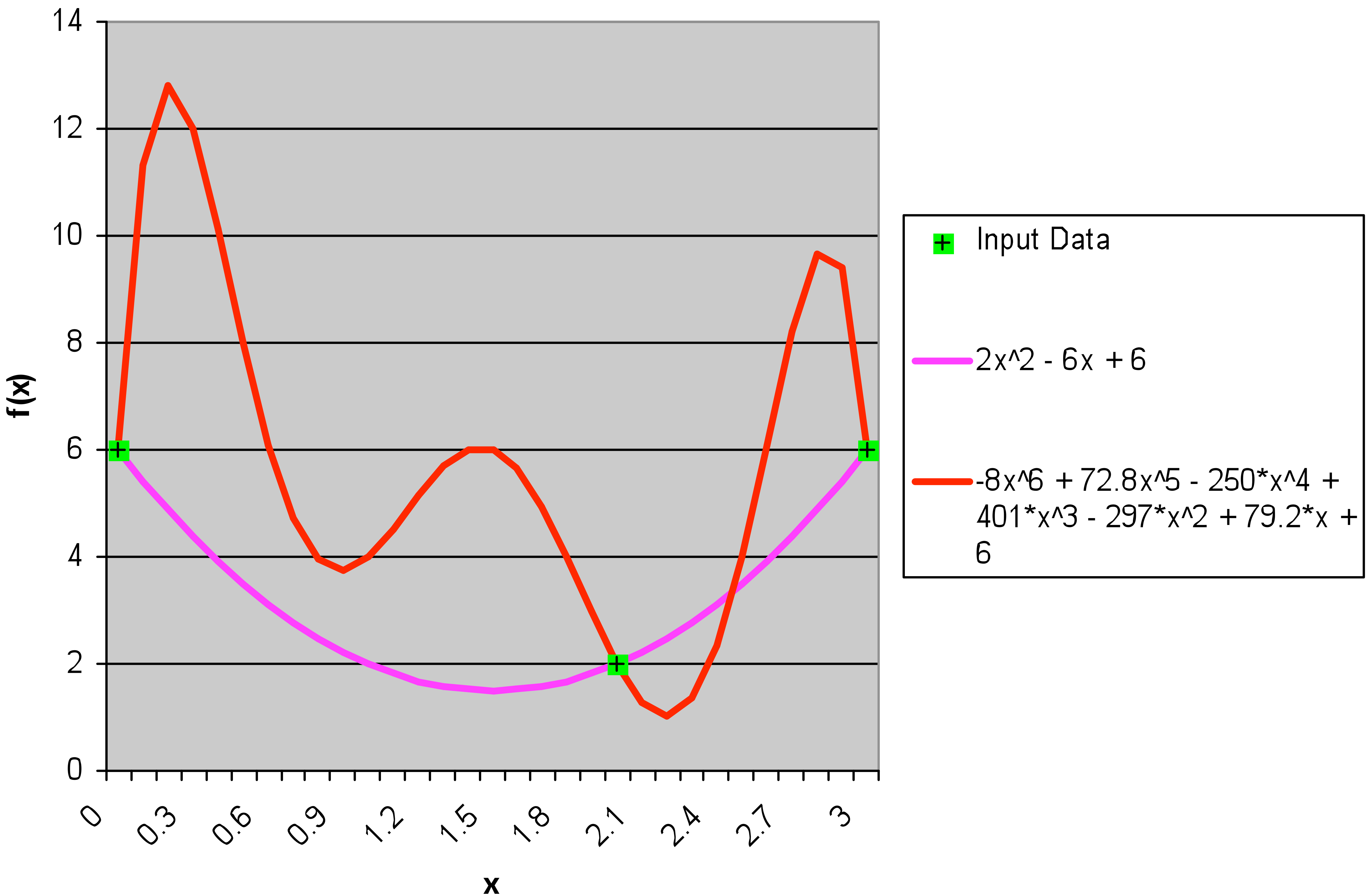
Observe some values of a function



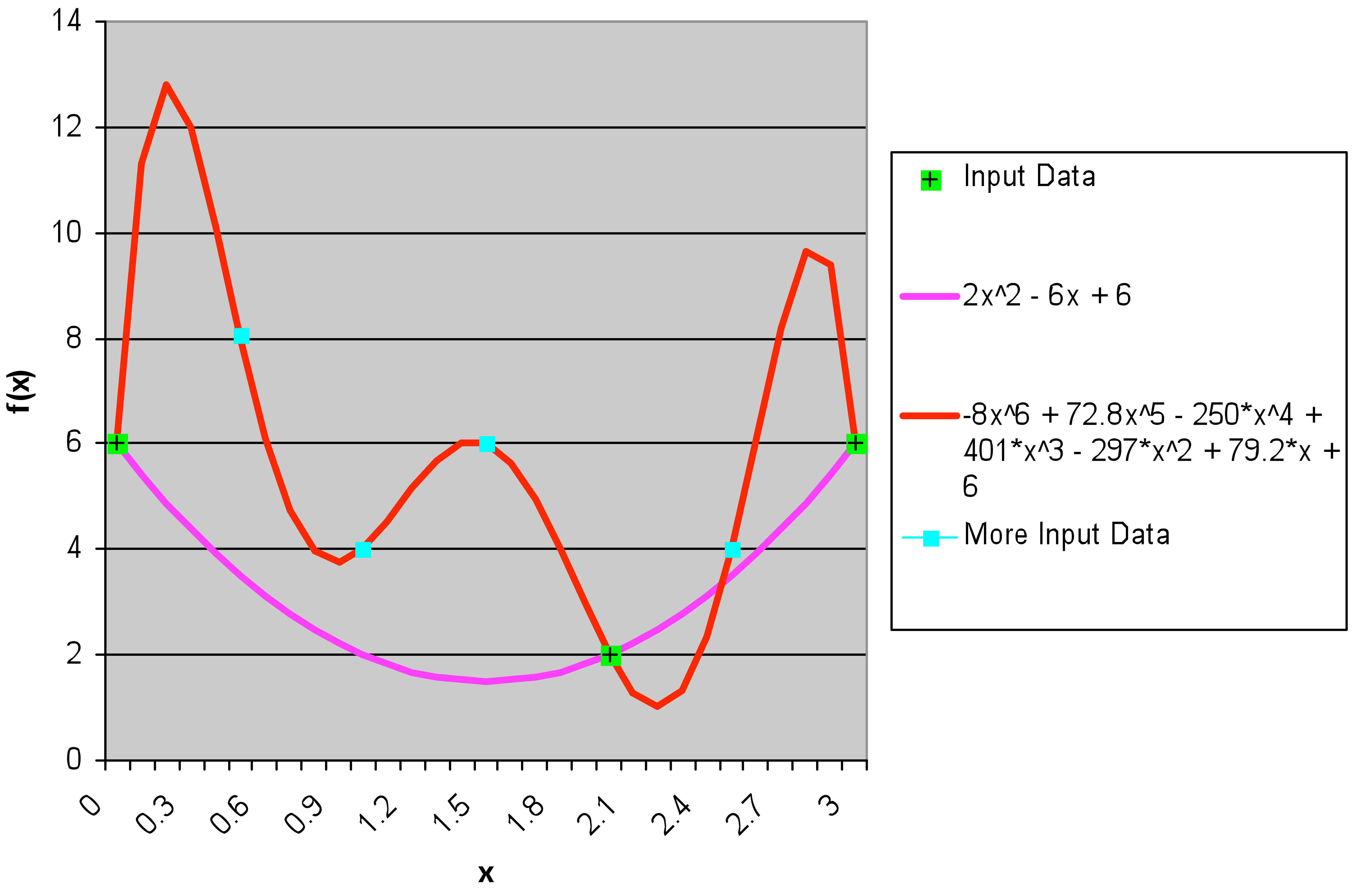
Guess the whole function



Another guess: Just as good?



More data needed to decide



Poverty of the Stimulus

Poverty of the Stimulus

- Never enough input data to completely determine the polynomial …
  - Always have infinitely many possibilities

- … unless you know the order of the polynomial ahead of time.
  - 2 points determine a line
  - 3 points determine a quadratic
  - etc.
- In language learning, is it enough to know that the target language is generated by a CFG?
  - without knowing the size of the CFG?

Language learning:

<u>Language</u> learning:

- Children listen to language [unsupervised]

<u>Language</u> learning:

- Children listen to language [unsupervised]
- Children are corrected?? [supervised]

<u>Language</u> learning:

- Children listen to language [unsupervised]
- Children are corrected?? [supervised]
- Children observe language in context

Language learning:

- Children listen to language [unsupervised]
- Children are corrected?? [supervised]
- Children observe language in context
- Children observe frequencies of language

<u>Language</u> learning:

- Children listen to language [unsupervised]
- Children are corrected?? [supervised]
- Children observe language in context
- Children observe frequencies of language

<u>Language</u> learning:

- Children listen to language [unsupervised]
- Children are corrected?? [supervised]
- Children observe language in context
- Children observe frequencies of language

Remember: Language = set of strings

Poverty of the Stimulus (1957)

- Children listen to language
- Children are corrected??
- Children observe language in context
- Children observe frequencies of language

Poverty of the Stimulus (1957)

Chomsky: Just like polynomials: never enough data unless you know something in advance. So kids must be born knowing what to expect in language.

- Children listen to language
- Children are corrected??
- Children observe language in context
- Children observe frequencies of language

Gold’s Theorem (1967)

a simple negative result along these lines:
kids (or computers) can’t learn much
without supervision, inborn knowledge, or statistics

- Children listen to language
- Children are corrected??
- Children observe language in context
- Children observe frequencies of language

The Idealized Situation

The Idealized Situation

- Mom talks

The Idealized Situation

- Mom talks
- Baby listens

The Idealized Situation

- Mom talks
- Baby listens

The Idealized Situation

- Mom talks
- Baby listens

- 1. Mom outputs a sentence

The Idealized Situation

- Mom talks
- Baby listens

- 1. Mom outputs a sentence
- 2. Baby hypothesizes what the language is
  (given all sentences so far)

The Idealized Situation

- Mom talks
- Baby listens

- 1. Mom outputs a sentence
- 2. Baby hypothesizes what the language is
  (given all sentences so far)
- 3. Goto step 1

The Idealized Situation

- Mom talks
- Baby listens

- 1. Mom outputs a sentence
- 2. Baby hypothesizes what the language is
  (given all sentences so far)
- 3. Goto step 1

The Idealized Situation

- Mom talks
- Baby listens

- 1. Mom outputs a sentence
- 2. Baby hypothesizes what the language is
  (given all sentences so far)
- 3. Goto step 1

- Guarantee: Mom's language *is* in the set of hypotheses that Baby is choosing among

The Idealized Situation

- Mom talks
- Baby listens

- 1. Mom outputs a sentence
- 2. Baby hypothesizes what the language is
  (given all sentences so far)
- 3. Goto step 1

- Guarantee: Mom's language *is* in the set of hypotheses that Baby is choosing among
- Guarantee: Any sentence of Mom's language is *eventually* uttered by Mom (even if infinitely many)

The Idealized Situation

- Mom talks
- Baby listens

- 1. Mom outputs a sentence
- 2. Baby hypothesizes what the language is
  (given all sentences so far)
- 3. Goto step 1

- Guarantee: Mom's language *is* in the set of hypotheses that Baby is choosing among
- Guarantee: Any sentence of Mom's language is *eventually* uttered by Mom (even if infinitely many)
- Assumption: Vocabulary (or alphabet) is finite.

Can Baby learn under these conditions?

Can Baby learn under these conditions?

- Learning in the limit:
  - There is some point at which Baby's hypothesis is correct and never changes again. Baby has converged!
  - Baby doesn't have to **know** that it's reached this point – it can keep an open mind about new evidence – but if its hypothesis is right, no such new evidence will ever come along.

Can Baby learn under these conditions?

- Learning in the limit:
  - There is some point at which Baby's hypothesis is correct and never changes again. Baby has converged!
  - Baby doesn't have to **know** that it's reached this point – it can keep an open mind about new evidence – but if its hypothesis is right, no such new evidence will ever come along.

- A class C of languages is **learnable in the limit** if one could construct a perfect C-Baby that can learn any language $L \in C$ in the limit from a Mom who speaks L.

Can Baby learn under these conditions?

- Learning in the limit:
  - There is some point at which Baby's hypothesis is correct and never changes again. Baby has converged!
  - Baby doesn't have to **know** that it's reached this point – it can keep an open mind about new evidence – but if its hypothesis is right, no such new evidence will ever come along.

- A class C of languages is **learnable in the limit** if one could construct a perfect C-Baby that can learn any language $L \in C$ in the limit from a Mom who speaks L.

Can Baby learn under these conditions?

- Learning in the limit:
  - There is some point at which Baby's hypothesis is correct and never changes again. Baby has converged!
  - Baby doesn't have to **know** that it's reached this point – it can keep an open mind about new evidence – but if its hypothesis is right, no such new evidence will ever come along.

- A class C of languages is **learnable in the limit** if one could construct a perfect C-Baby that can learn any language L $\in$ C in the limit from a Mom who speaks L.

- Baby knows the class C of possibilities, but not L.

Can Baby learn under these conditions?

- Learning in the limit:
  - There is some point at which Baby's hypothesis is correct and never changes again. Baby has converged!
  - Baby doesn't have to **know** that it's reached this point – it can keep an open mind about new evidence – but if its hypothesis is right, no such new evidence will ever come along.

- A class C of languages is **learnable in the limit** if one could construct a perfect C-Baby that can learn any language L $\in$ C in the limit from a Mom who speaks L.

- Baby knows the class C of possibilities, but not L.
- Is there a perfect finite-state Baby?

Can Baby learn under these conditions?

- Learning in the limit:
  - There is some point at which Baby's hypothesis is correct and never changes again. Baby has converged!
  - Baby doesn't have to **know** that it's reached this point – it can keep an open mind about new evidence – but if its hypothesis is right, no such new evidence will ever come along.

- A class C of languages is **learnable in the limit** if one could construct a perfect C-Baby that can learn any language L $\in$ C in the limit from a Mom who speaks L.

- Baby knows the class C of possibilities, but not L.
- Is there a perfect finite-state Baby?
- Is there a perfect context-free Baby?

Languages vs. Grammars

- Does Baby have to get the right grammar?
- (E.g., does VP have to be called VP?)

- Assumption: Finite vocabulary.

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom
Baby

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa
Baby	

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa
Baby	L3

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa	ab
Baby	L3	

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa	ab
Baby	L3	L1

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa	ab	ac
Baby	L3	L1	

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa	ab	ac
Baby	L3	L1	L1

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa	ab	ac	ab
Baby	L3	L1	L1	

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa	ab	ac	ab
Baby	L3	L1	L1	L1

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa	ab	ac	ab	aa
Baby	L3	L1	L1	L1	

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa	ab	ac	ab	aa
Baby	L3	L1	L1	L1	L1

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

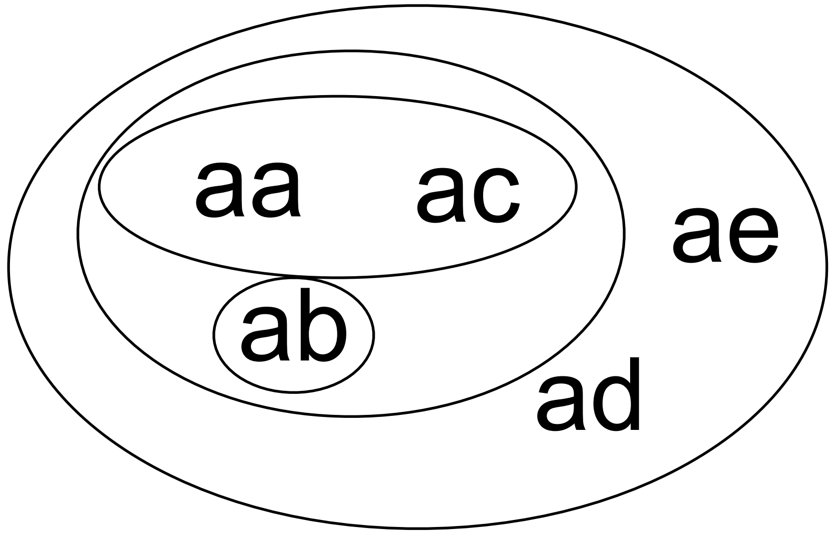
- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

Mom	aa	ab	ac	ab	aa	…
Baby	L3	L1	L1	L1	L1	

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

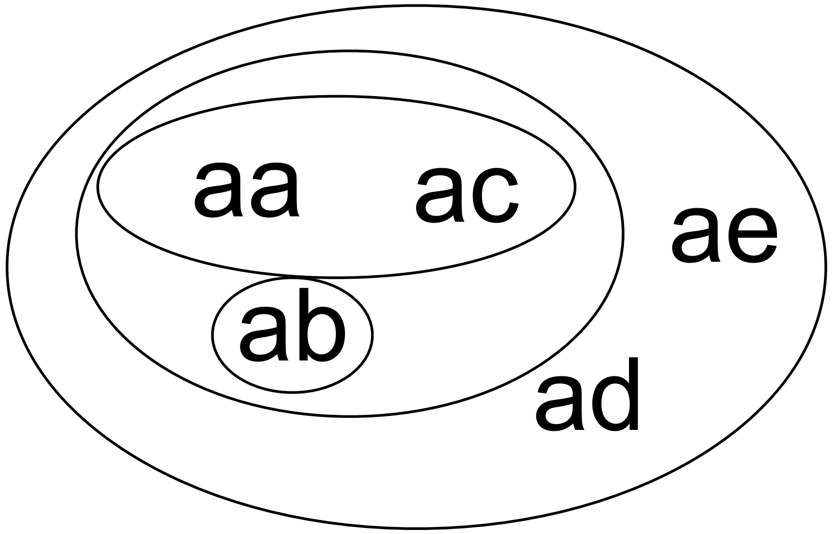


Mom
Baby

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

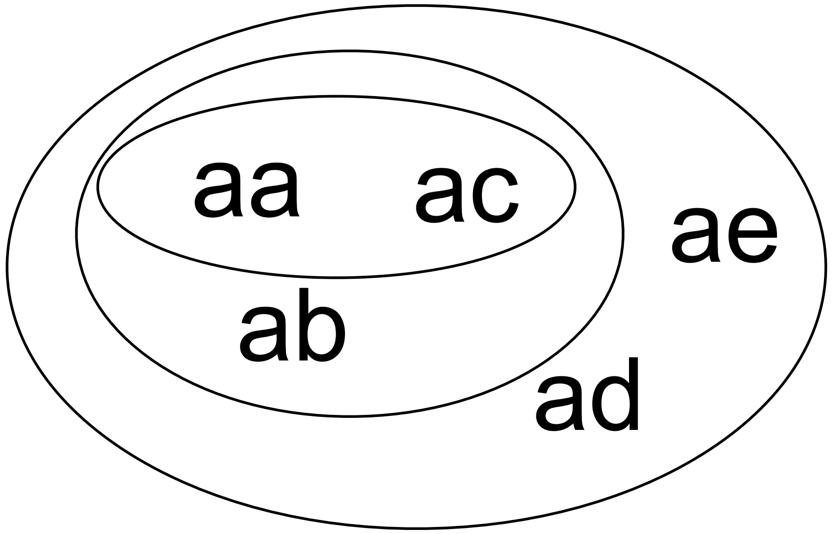


Mom aa
Baby

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

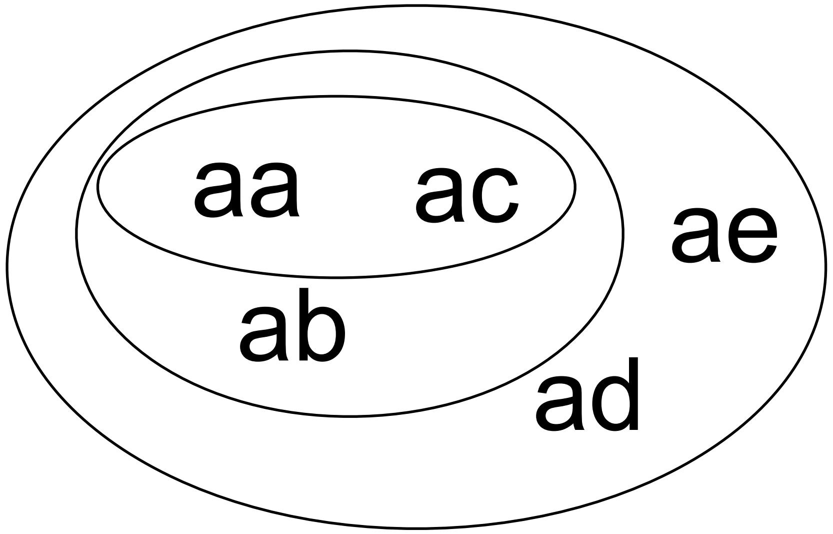


Mom aa
Baby L3

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

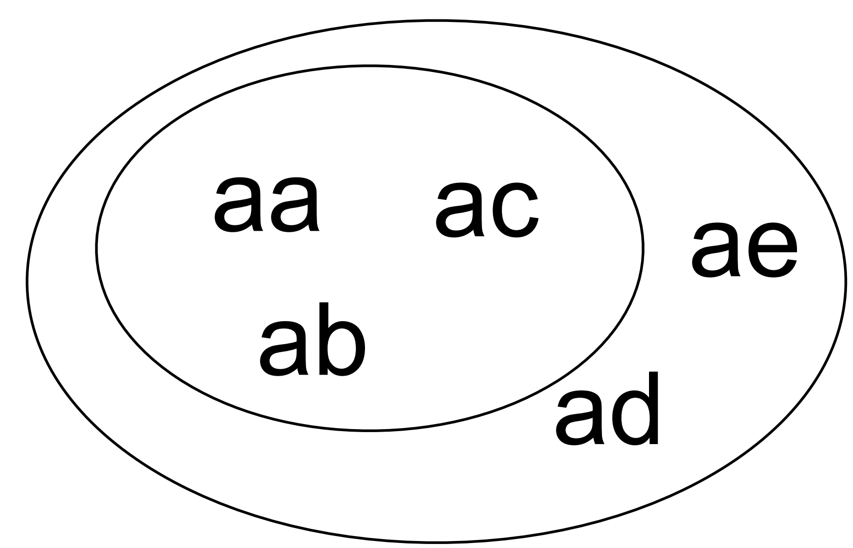


Mom	aa	ab
Baby	L3	

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

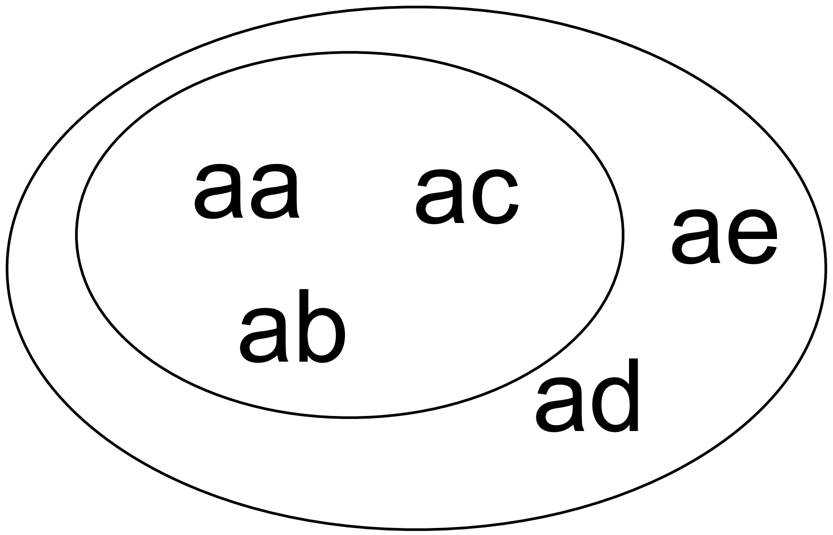


Mom	aa	ab
Baby	L3	L1

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

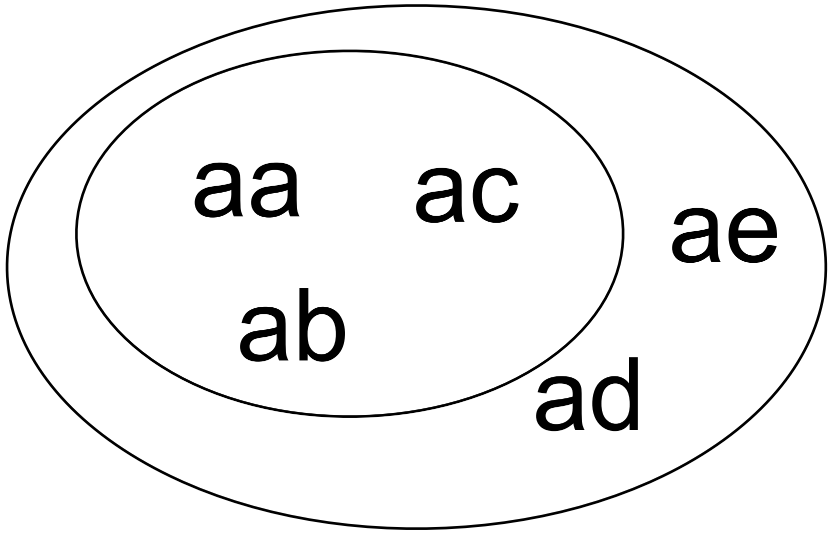


Mom	aa	ab	ac
Baby	L3	L1	

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

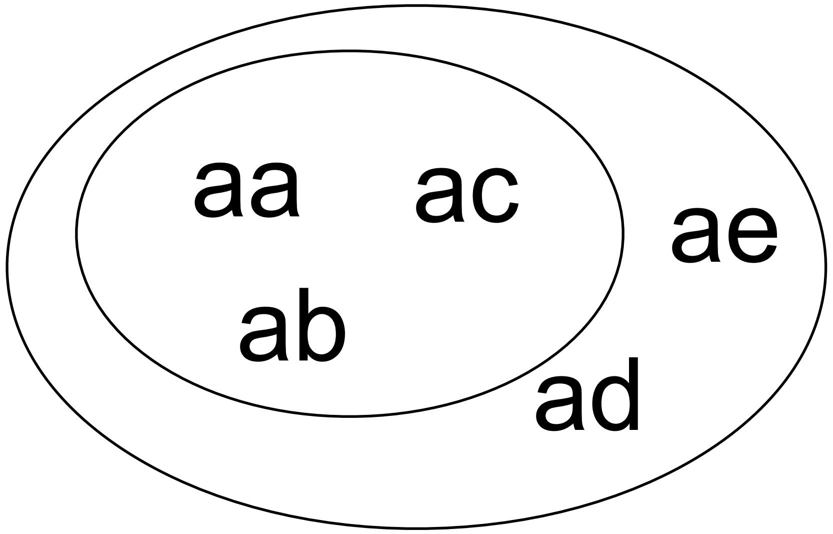


Mom	aa	ab	ac
Baby	L3	L1	L1

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

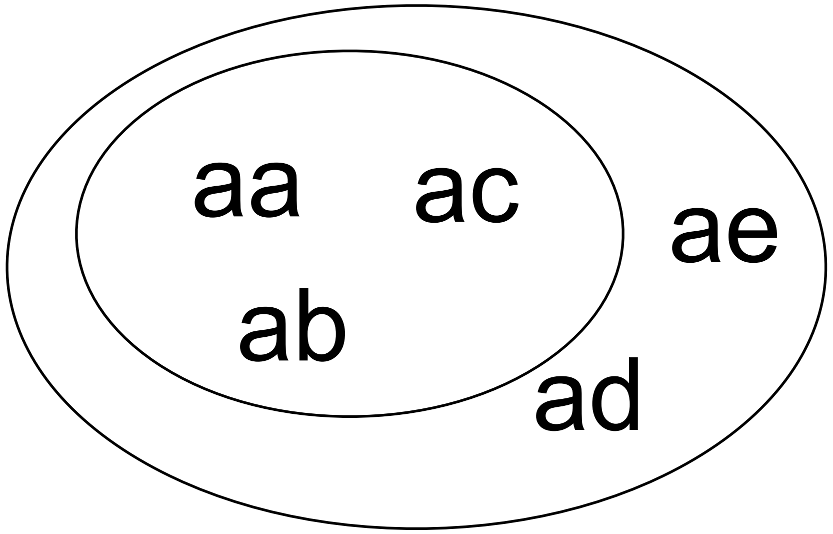


Mom	aa	ab	ac	ab
Baby	L3	L1	L1	

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

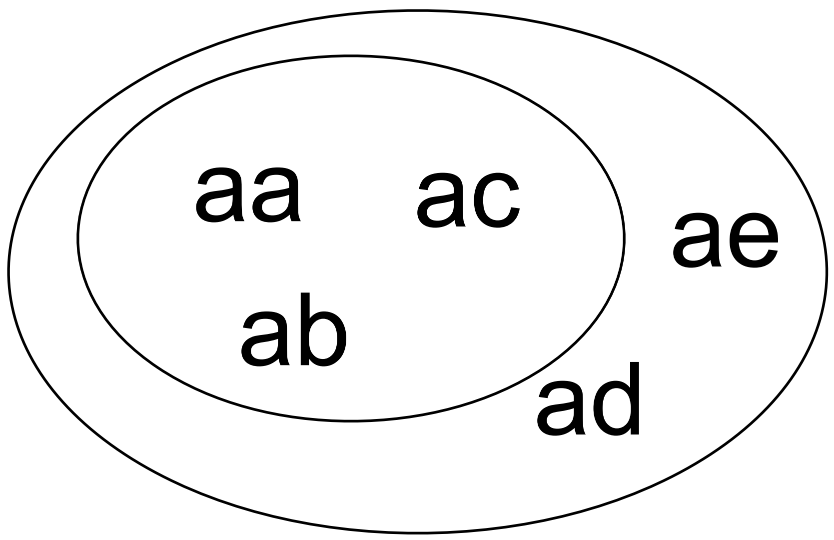


Mom	aa	ab	ac	ab
Baby	L3	L1	L1	L1

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

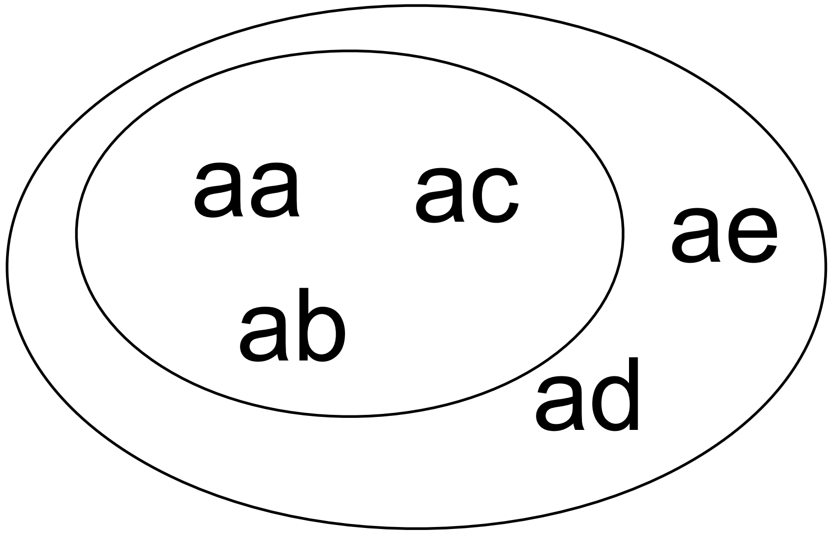


Mom	aa	ab	ac	ab	aa
Baby	L3	L1	L1	L1	

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}

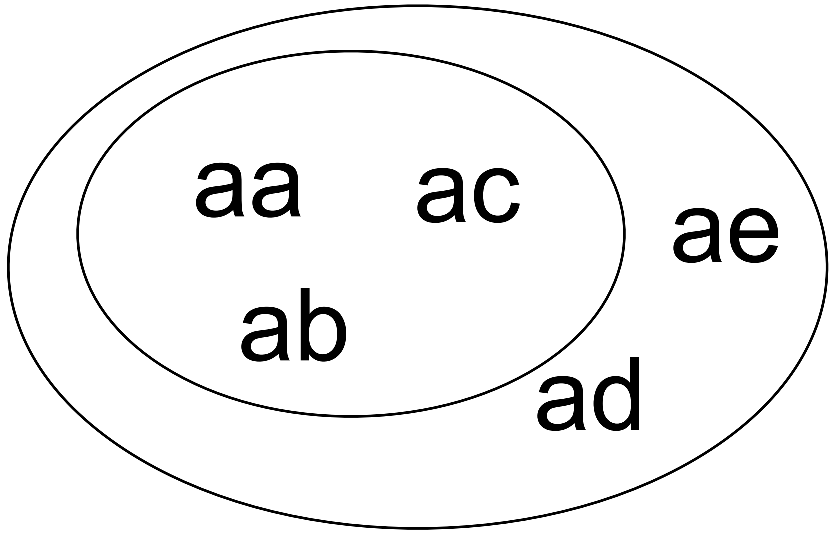


Mom	aa	ab	ac	ab	aa
Baby	L3	L1	L1	L1	L1

Conservative Strategy

- Baby's hypothesis should always be smallest language consistent with the data

- Works for finite languages? Let's try it …
  - Language 1: {aa,ab,ac}
  - Language 2: {aa,ab,ac,ad,ae}
  - Language 3: {aa,ac}
  - Language 4: {ab}



Mom	aa	ab	ac	ab	aa	…
Baby	L3	L1	L1	L1	L1	

Evil Mom

- To find out whether Baby is perfect, we have to see whether it gets 100% even in the most adversarial conditions
- Assume Mom is trying to fool Baby
  - although she must speak only sentences from L
  - and she must eventually speak each such sentence
- Does Baby's strategy work?

An Unlearnable Class

- Class of languages:
  - Let L_n = set of all strings of length < n
  - What is L_0?
  - What is L_1?
  - What is L_∞?
    - If the true language is L_∞, can Mom really follow rules?
    - Must eventually speak every sentence of L_∞. Possible?
    - Yes: ε; a, b; aa, ab, ba, bb; aaa, aab, aba, abb, baa, …
  - Our class is $C = \{L_0, L_1, \ldots L_\infty\}$

An Unlearnable Class

An Unlearnable Class

- Let L_n = set of all strings of length < n
  - What is L_0?
  - What is L_1?
  - What is L_∞?

An Unlearnable Class

- Let L_n = set of all strings of length < n
  - What is L_0?
  - What is L_1?
  - What is L_∞?
- Our class is $C = \{L_0, L_1, \ldots L_\infty\}$

An Unlearnable Class

- Let L_n = set of all strings of length < n
  - What is L_0?
  - What is L_1?
  - What is L_∞?
- Our class is $C = \{L_0, L_1, \ldots L_\infty\}$
- A perfect C-baby will distinguish among all of these depending on the input.

An Unlearnable Class

- Let L_n = set of all strings of length < n
  - What is L_0?
  - What is L_1?
  - What is L_∞?
- Our class is C = $\{L_0, L_1, \ldots L_\infty\}$
- A perfect C-baby will distinguish among all of these depending on the input.
- But there is no perfect C-baby …

An Unlearnable Class

An Unlearnable Class

- Our class is $C = \{L_0, L_1, \ldots L_\infty\}$

An Unlearnable Class

- Our class is $C = \{L_0, L_1, \ldots L_\infty\}$
- Suppose Baby adopts conservative strategy, always picking smallest possible language in C.

An Unlearnable Class

- Our class is $C = \{L_0, L_1, \ldots L_\infty\}$
- Suppose Baby adopts conservative strategy, always picking smallest possible language in C.
- So if Mom's longest sentence so far has 75 words, baby's hypothesis is L_{76}.

An Unlearnable Class

- Our class is $C = \{L_0, L_1, \ldots L_\infty\}$
- Suppose Baby adopts conservative strategy, always picking smallest possible language in C.
- So if Mom's longest sentence so far has 75 words, baby's hypothesis is L_{76}.
- This won't always work: What language can't a conservative Baby learn?

An Unlearnable Class

An Unlearnable Class

- Our class is $C = \{L_0, L_1, \ldots L_\infty\}$

An Unlearnable Class

- Our class is $C = \{L_0, L_1, \ldots L_\infty\}$
- Could a non-conservative baby be a perfect C-Baby, and eventually converge to any of these?

An Unlearnable Class

- Our class is C = $\{L_0, L_1, \ldots L_\infty\}$
- Could a non-conservative baby be a perfect C-Baby, and eventually converge to any of these?
- **Claim:** *Any* perfect C-Baby must be "quasi-conservative":

An Unlearnable Class

- Our class is C = $\{L_0, L_1, \ldots L_\infty\}$
- Could a non-conservative baby be a perfect C-Baby, and eventually converge to any of these?
- **Claim:** *Any* perfect C-Baby must be "quasi-conservative":
  - If true language is L_{76}, and baby posits something else, baby must still eventually come back and guess L_{76} (since it's perfect).

An Unlearnable Class

- Our class is C = {L_0, L_1, … L_∞}
- Could a non-conservative baby be a perfect C-Baby, and eventually converge to any of these?
- **Claim:** *Any* perfect C-Baby must be "quasi-conservative":
  - If true language is L_{76}, and baby posits something else, baby must still eventually come back and guess L_{76} (since it's perfect).
  - So if longest sentence so far is 75 words, and Mom keeps talking from L_{76}, then eventually baby must actually return to the conservative guess L_{76}.

An Unlearnable Class

- Our class is $C = \{L_0, L_1, \ldots L_\infty\}$
- Could a non-conservative baby be a perfect C-Baby, and eventually converge to any of these?
- **Claim:** *Any* perfect C-Baby must be "quasi-conservative":
  - If true language is L_{76}, and baby posits something else, baby must still eventually come back and guess L_{76} (since it's perfect).
  - So if longest sentence so far is 75 words, and Mom keeps talking from L_{76}, then eventually baby must actually return to the conservative guess L_{76}.
  - Agreed?

Mom's Revenge

If longest sentence so far is 75 words, and Mom keeps talking from L_{76}, then eventually a perfect C-baby must actually return to the conservative guess L_{76}.

- Suppose true language is L_∞.
- Evil Mom can prevent our supposedly perfect C-Baby from converging to it.
- If Baby ever guesses L_∞, say when the longest sentence is 75 words:
  - Then Evil Mom keeps talking from L_{76} until Baby capitulates and revises her guess to L_{76} – as any perfect C-Baby must.
  - So Baby has *not* stayed at L_∞ as required.
- Then Mom can go ahead with longer sentences. If Baby ever guesses L_∞ again, she plays the same trick again.

Mom's Revenge

If longest sentence so far is 75 words, and Mom keeps talking from L_{76}, then eventually a perfect C-baby must actually return to the conservative guess L_{76}.

- Suppose true language is L_∞.
- Evil Mom can prevent our supposedly perfect C-Baby from converging to it.
- If Baby ever guesses L_∞, say when the longest sentence is 75 words:
  - Then Evil Mom keeps talking from L_{76} until Baby capitulates and revises her guess to L_{76} – as any perfect C-Baby must.
  - So Baby has *not* stayed at L_∞ as required.
- Conclusion: There's no perfect Baby that is guaranteed to converge to L_0, L_1, … or L_∞ as appropriate. If it always succeeds on finite languages, Evil Mom can trick it on infinite language.

Implications

Implications

- We found that $C = \{L_0, L_1, \ldots L_\infty\}$ isn't learnable in the limit.

Implications

- We found that C = $\{L_0, L_1, \ldots L_\infty\}$ isn't learnable in the limit.

Implications

- We found that $C = \{L_0, L_1, \ldots L_\infty\}$ isn't learnable in the limit.

- How about class of finite-state languages?

Implications

- We found that C = $\{L_0, L_1, \ldots L_\infty\}$ isn't learnable in the limit.

- How about class of finite-state languages?
  - Not unless you limit it further (e.g., # of states)

Implications

- We found that C = $\{L_0, L_1, \dots L_\infty\}$ isn't learnable in the limit.

- How about class of finite-state languages?
  - Not unless you limit it further (e.g., # of states)
  - After all, it includes all languages in C, and more, so learner has harder choice

Implications

- We found that $C = \{L_0, L_1, \ldots L_\infty\}$ isn't learnable in the limit.

- How about class of finite-state languages?
  - Not unless you limit it further (e.g., # of states)
  - After all, it includes all languages in C, and more, so learner has harder choice

Implications

- We found that $C = \{L_0, L_1, \ldots L_\infty\}$ isn't learnable in the limit.

- How about class of finite-state languages?
  - Not unless you limit it further (e.g., # of states)
  - After all, it includes all languages in C, and more, so learner has harder choice

- How about class of context-free languages?

Implications

- We found that $C = \{L_0, L_1, \ldots L_\infty\}$ isn't learnable in the limit.

- How about class of finite-state languages?
  - Not unless you limit it further (e.g., # of states)
  - After all, it includes all languages in C, and more, so learner has harder choice

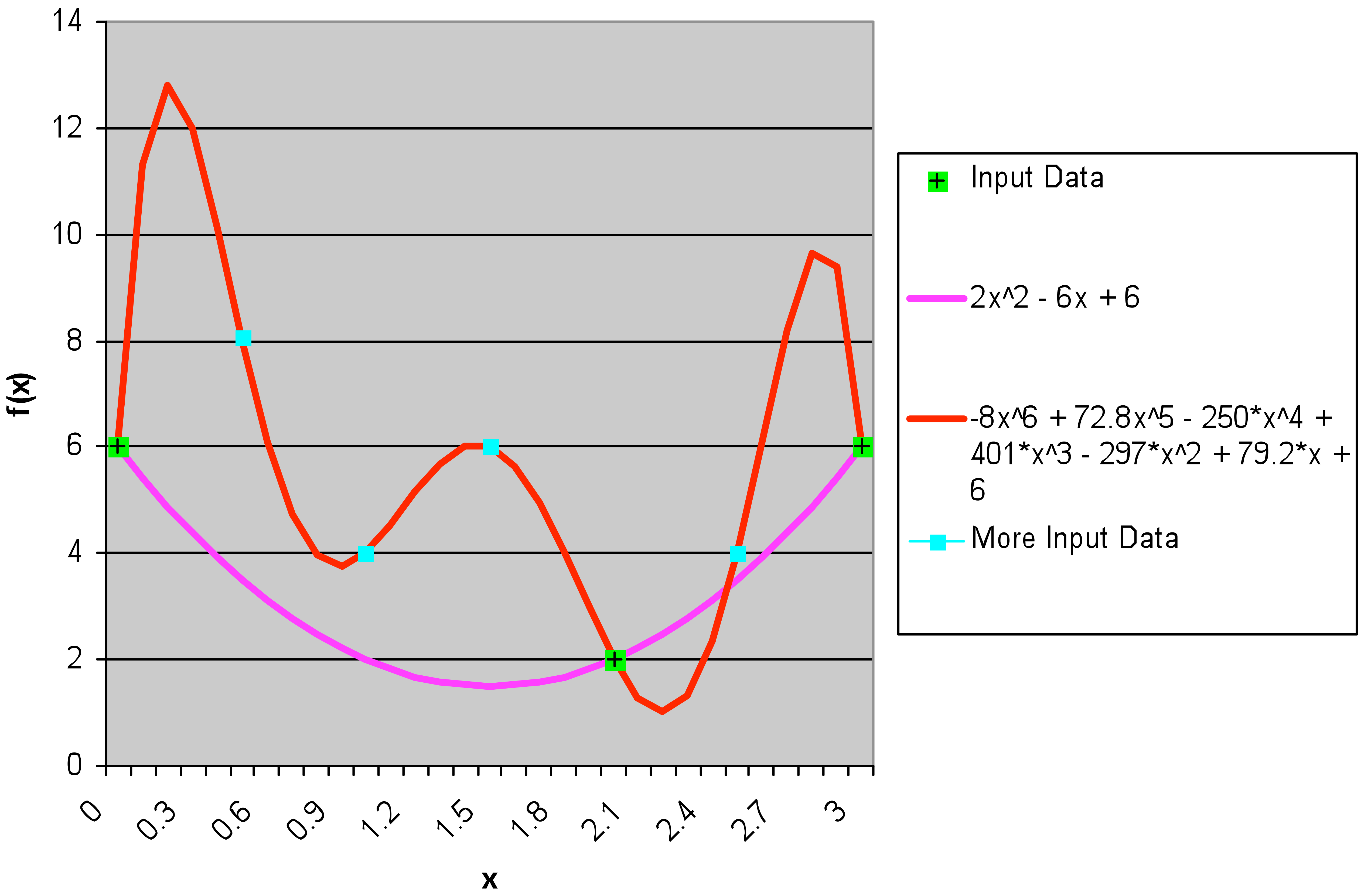
- How about class of context-free languages?
  - Not unless you limit it further (e.g., # of rules)

Punchline

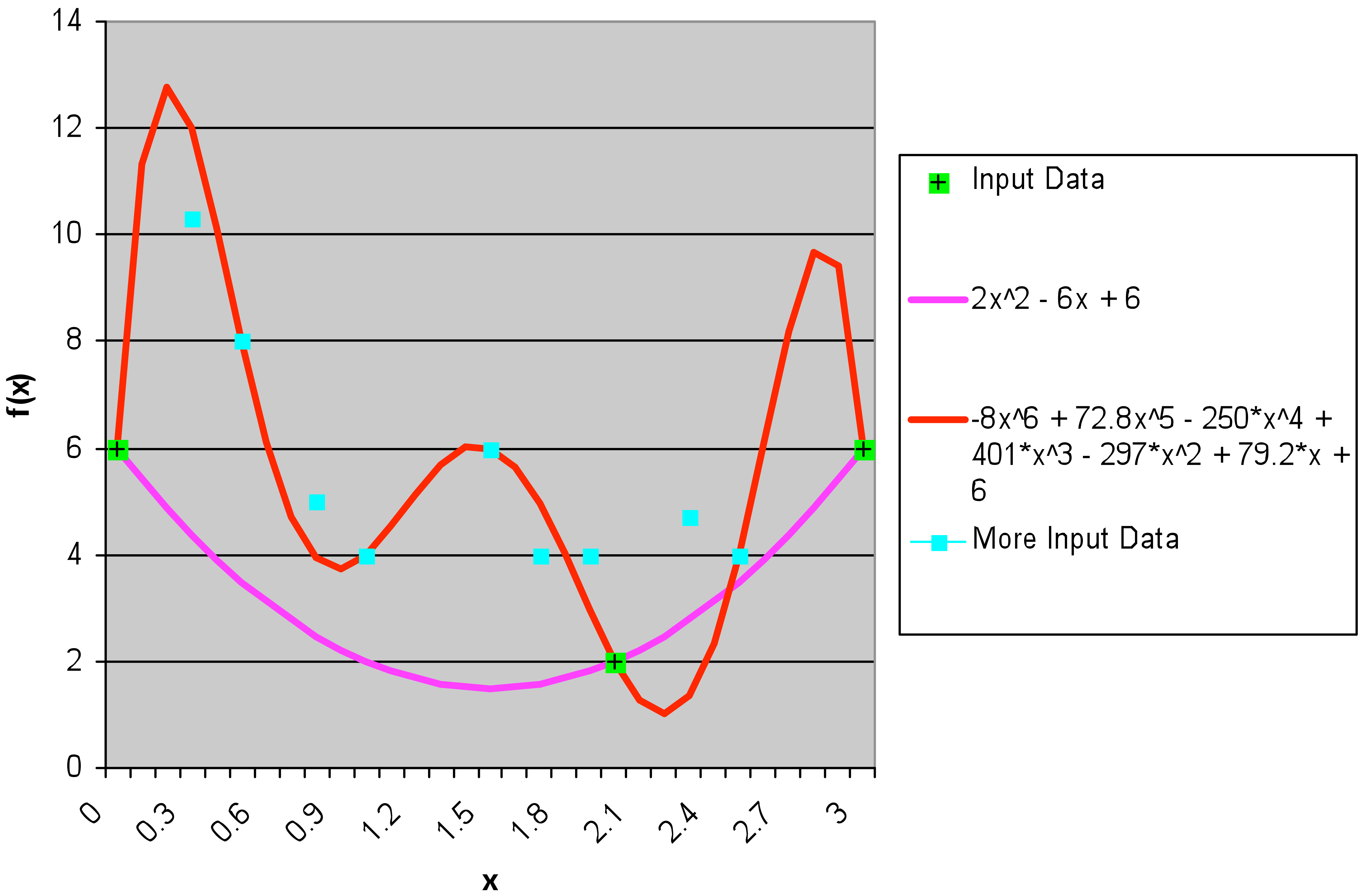
- But class of *probabilistic* context-free languages *is* learnable in the limit!! *(Horning, 1969)*

- If Mom has to output sentences randomly with the appropriate probabilities,
  - she's unable to be too evil
  - there are then perfect Babies that are guaranteed to converge to an appropriate probabilistic CFG
- I.e., from hearing a finite number of sentences, Baby can correctly converge on a grammar that predicts an infinite number of sentences.
  - Baby is generalizing! Just like real babies!

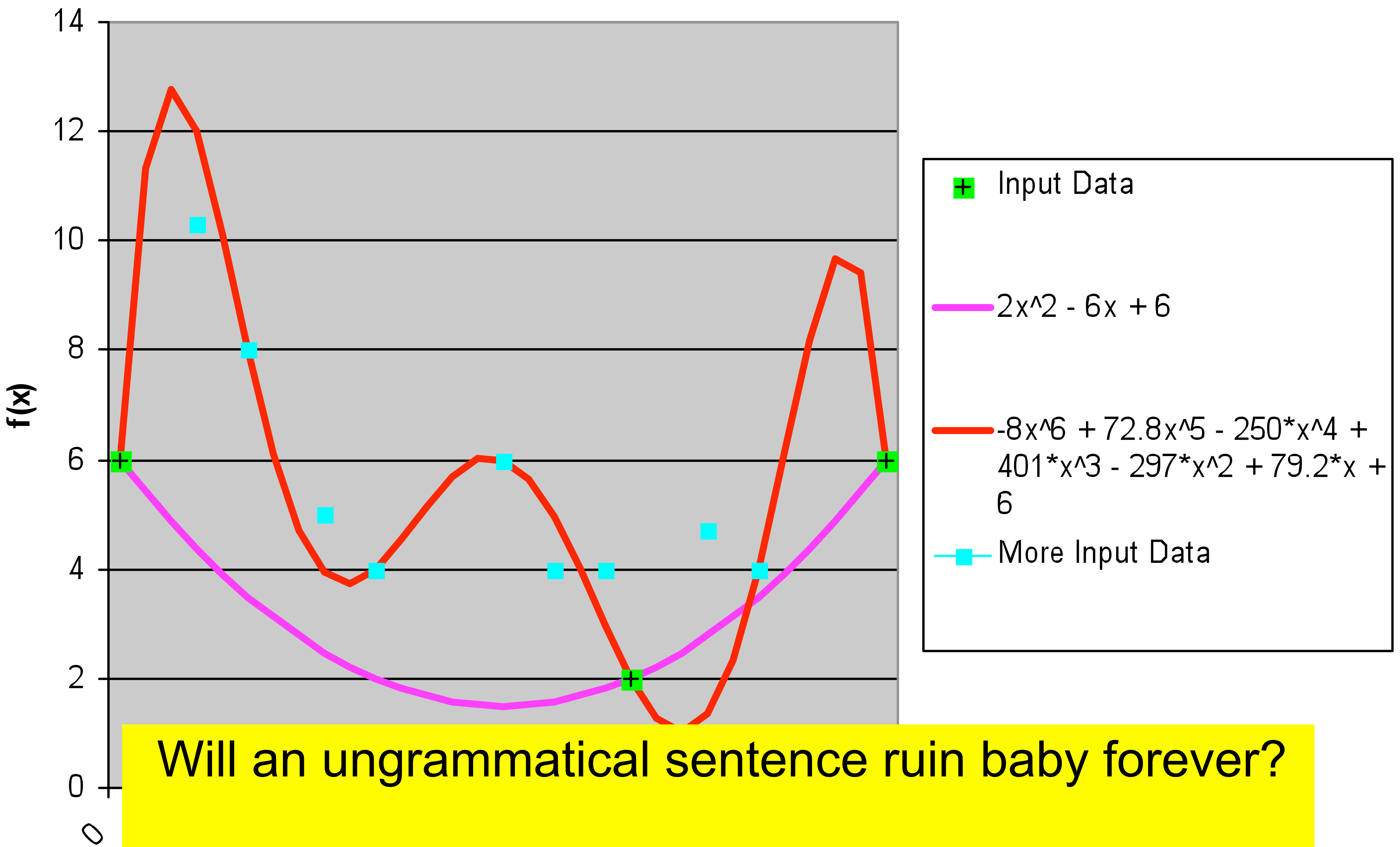
Perfect fit to perfect, incomplete data



Imperfect fit to noisy data

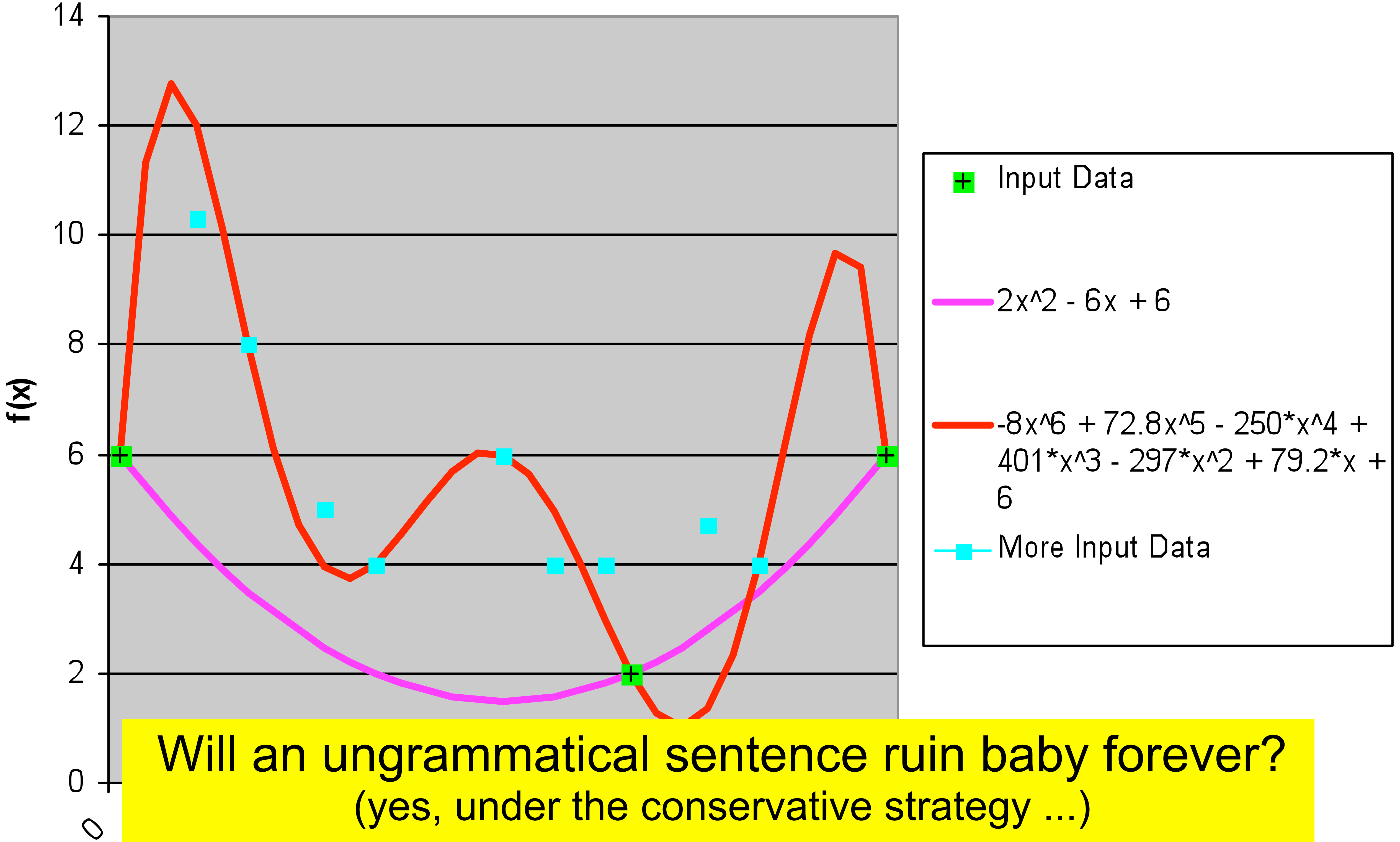


Imperfect fit to noisy data



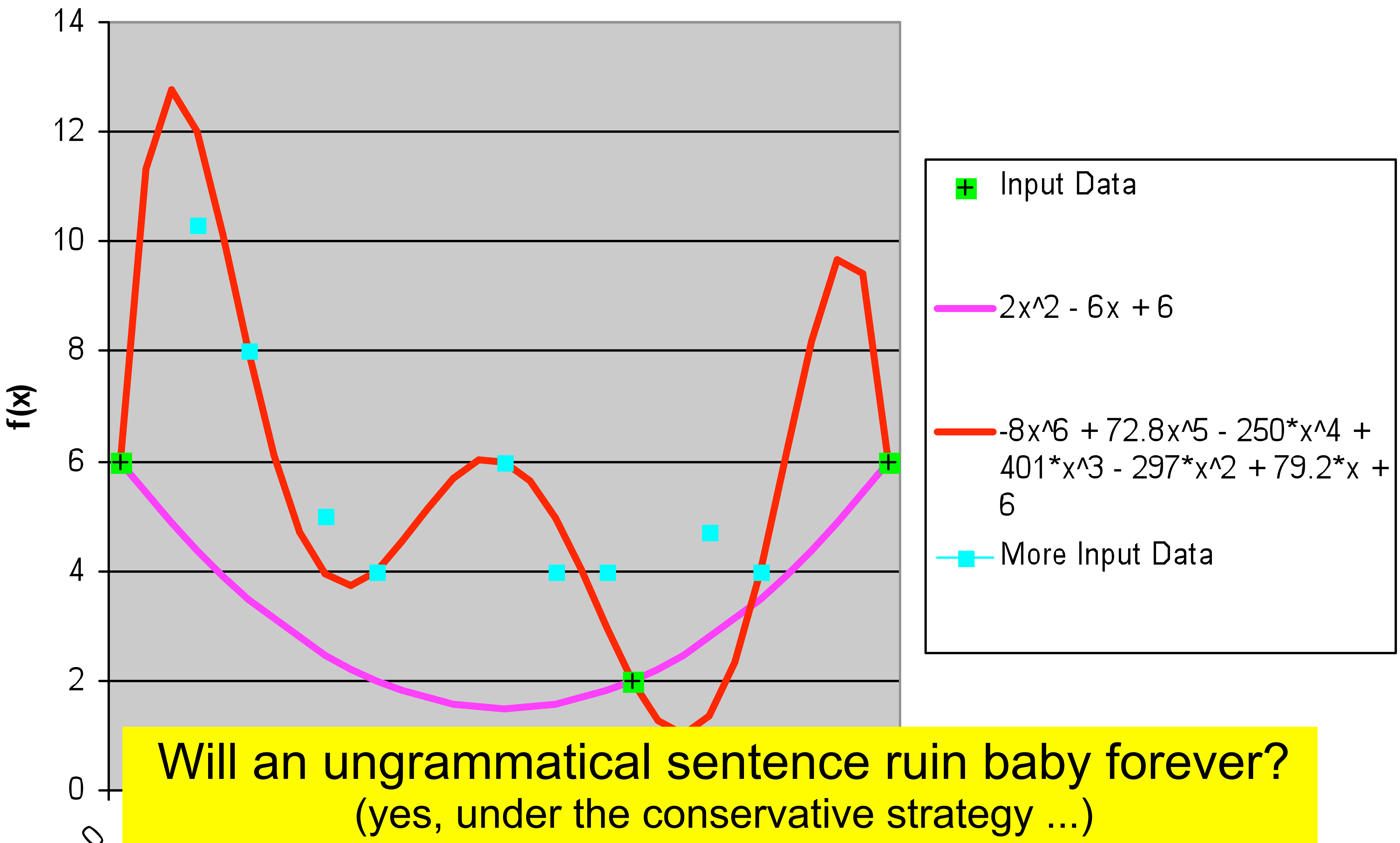
Will an ungrammatical sentence ruin baby forever?

Imperfect fit to noisy data



Will an ungrammatical sentence ruin baby forever?
(yes, under the conservative strategy ...)

Imperfect fit to noisy data

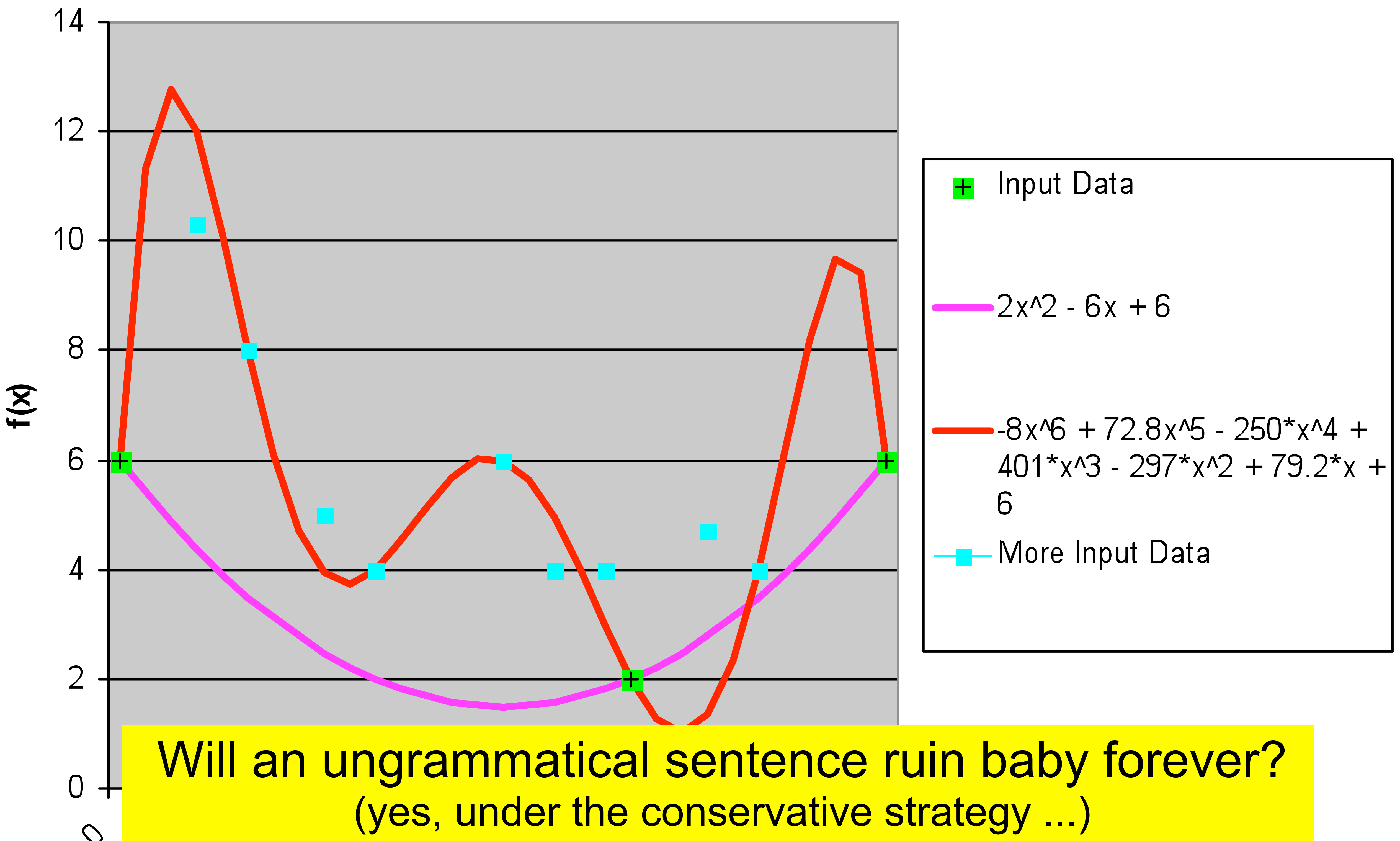


Will an ungrammatical sentence ruin baby forever?

(yes, under the conservative strategy ...)

Or can baby figure out which data to (partly) ignore?

Imperfect fit to noisy data



Will an ungrammatical sentence ruin baby forever?
(yes, under the conservative strategy ...)
Or can baby figure out which data to (partly) ignore?
Statistics can help again ... how?

Frequencies and Probabilities in Natural Languages

Chris Manning and others

Models for language

- Human languages are the prototypical example of a symbolic system
- From the beginning, logics and logical reasoning were invented for handling natural language understanding
- Logics and formal languages have a language-like form that draws from and meshes well with natural languages
- Where are the numbers?

Dominant answer in linguistic theory: Nowhere

Chomsky again (1969: 57; also 1956, 1957, etc.):

- "It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term."

Probabilistic models wrongly mix in world knowledge

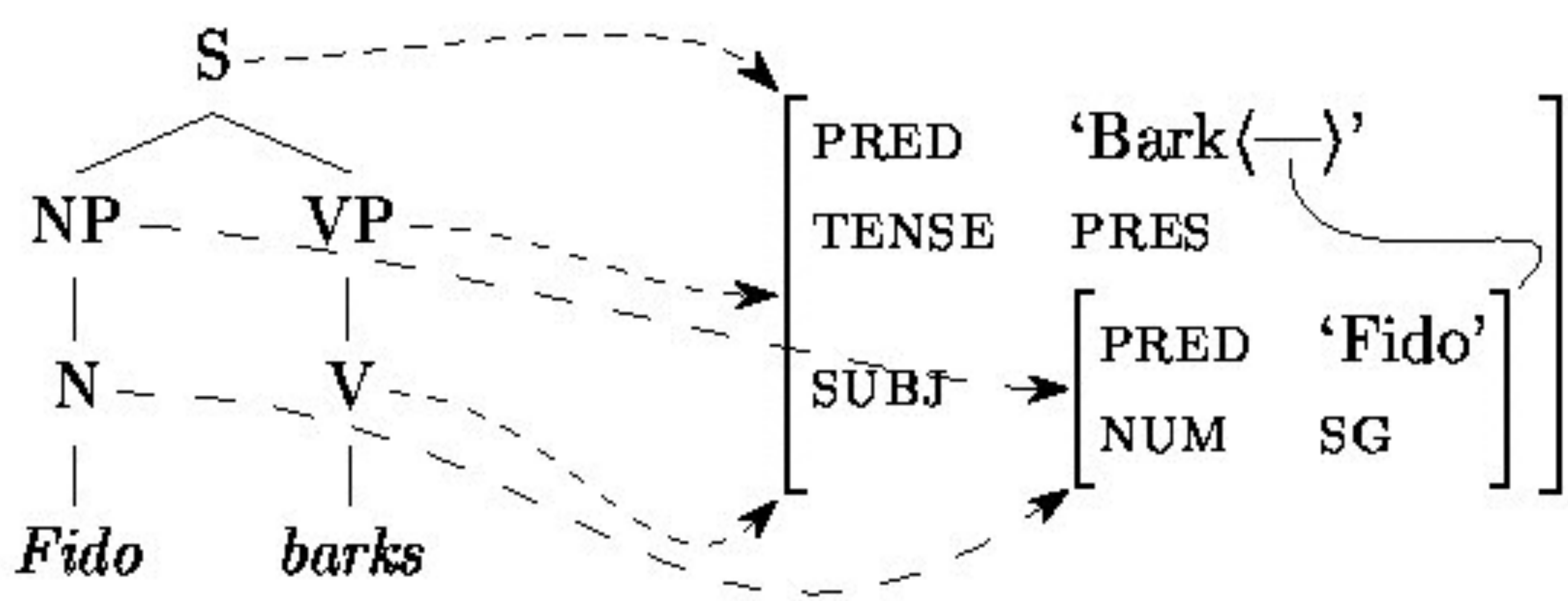
- New York vs. Dayton, Ohio

They don't model grammaticality [also, Tesnière 1959]

- Colorless green ideas sleep furiously
- Furiously sleep ideas green colorless
- [But see Pereira 2005]

Categorical linguistic theories (GB, Minimalism, LFG, HPSG, CG, …)

- Systems of variously rules, principles, and representations is used to describe an infinite set of grammatical sentences of the language
- Other sentences are deemed ungrammatical
- Word strings are given a (hidden) structure



The need for frequencies / probability distributions

The motivation comes from two sides:

- Categorical linguistic theories claim too much:
  - They place a hard categorical boundary of grammaticality, where really there is a fuzzy edge, determined by many conflicting constraints and issues of conventionality vs. human creativity
- Categorical linguistic theories explain too little:
  - They say nothing at all about the soft constraints which explain how people choose to say things
    - Something that language educators, computational NLP people – and historical linguists and sociolinguists dealing with real language – usually want to know about

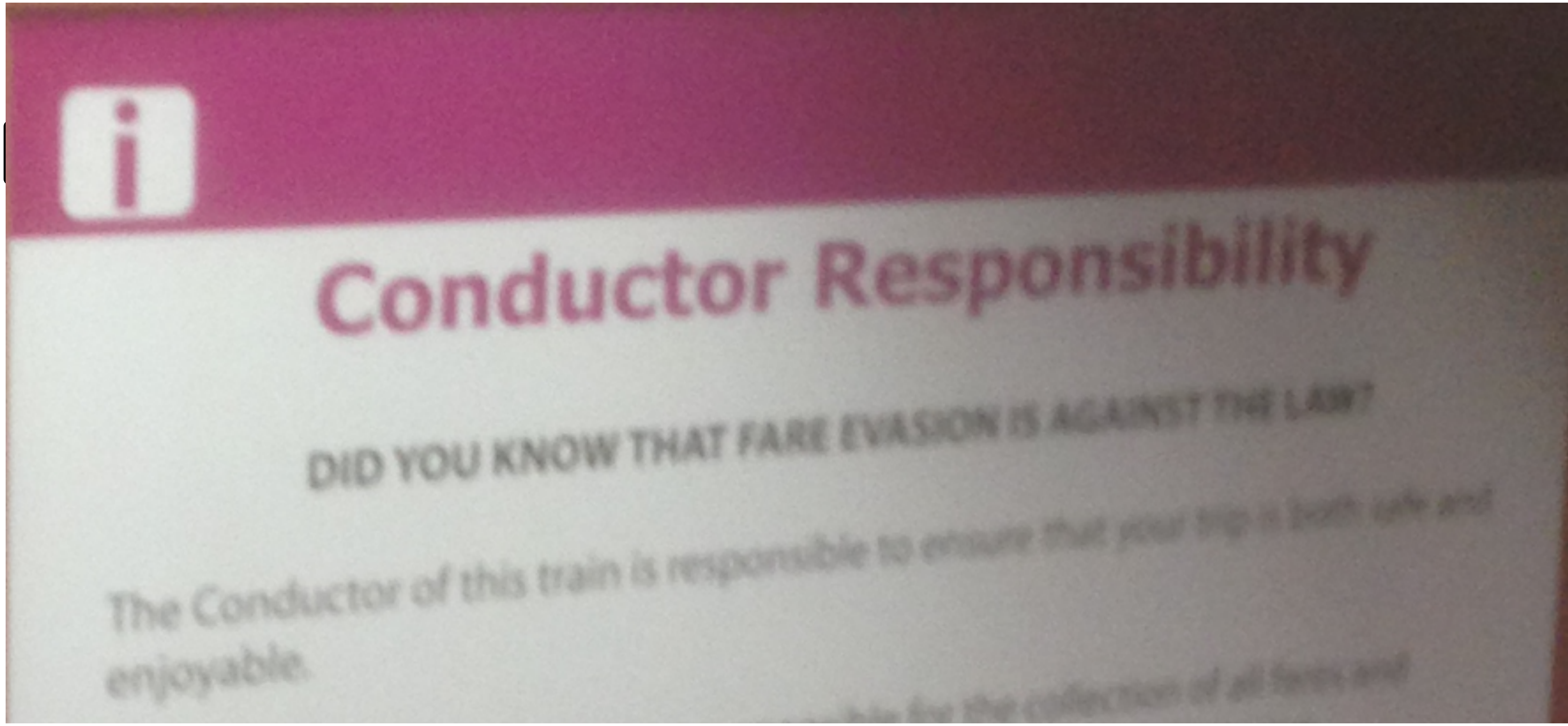
1. The hard constraints of categorical grammars

- Sentences must satisfy all the rules of the grammar
  - One group specifies the arguments that different verbs take – lexical subcategorization information
    - Some verbs must take objects: *Kim devoured
      [* means ungrammatical]
    - Others do not: *Kim's lip quivered the straw
    - Others take various forms of sentential complements
- In NLP systems, ungrammatical sentences don't parse
- But the problem with this model was noticed early on:
  - "All grammars leak." (Sapir 1921: 38)

Example: verbal clausal subcategorization frames

- Some verbs take various types of sentential complements, given as subcategorization frames:
  - regard: __ NP[acc] as {NP, AdjP}
  - consider: __ NP[acc] {AdjP, NP, VP[inf]}
  - think: __ CP[that]; __ NP[acc] NP
- **Problem:** in context, language is used more flexibly than this model suggests
  - Most such subcategorization ‘facts’ are **wrong**

Subcat on the MBTA



?The Conductor of this train is responsible to ensure that your trip is both safe and enjoyable.

…responsible for ensuring…

?…responsible that it be ensured that …

Standard subcategorization rules (Pollard and Sag 1994)

- We consider Kim to be an acceptable candidate
- We consider Kim an acceptable candidate
- We consider Kim quite acceptable
- We consider Kim among the most acceptable candidates
- *We consider Kim as an acceptable candidate
- *We consider Kim as quite acceptable
- *We consider Kim as among the most acceptable candidates
- ?*We consider Kim as being among the most acceptable candidates

Subcategorization facts from The New York Times

Consider as:

- The boys consider her as family and she participates in everything we do.
- Greenspan said, “I don't consider it as something that gives me great concern.
- “We consider that as part of the job,” Keep said.
- Although the Raiders missed the playoffs for the second time in the past three seasons, he said he considers them as having championship potential.
- Culturally, the Croats consider themselves as belonging to the “civilized” West, ...

More subcategorization facts: regard

Pollard and Sag (1994):

- *We regard Kim to be an acceptable candidate
- We regard Kim as an acceptable candidate

The New York Times:

- As 70 to 80 percent of the cost of blood tests, like prescriptions, is paid for by the state, neither physicians nor patients regard expense to be a consideration.
- Conservatives argue that the Bible regards homosexuality to be a sin.

More subcategorization facts: turn out and end up

Pollard and Sag (1994):

- Kim turned out political
- *Kim turned out doing all the work

The New York Times:

- But it turned out having a greater impact than any of us dreamed.

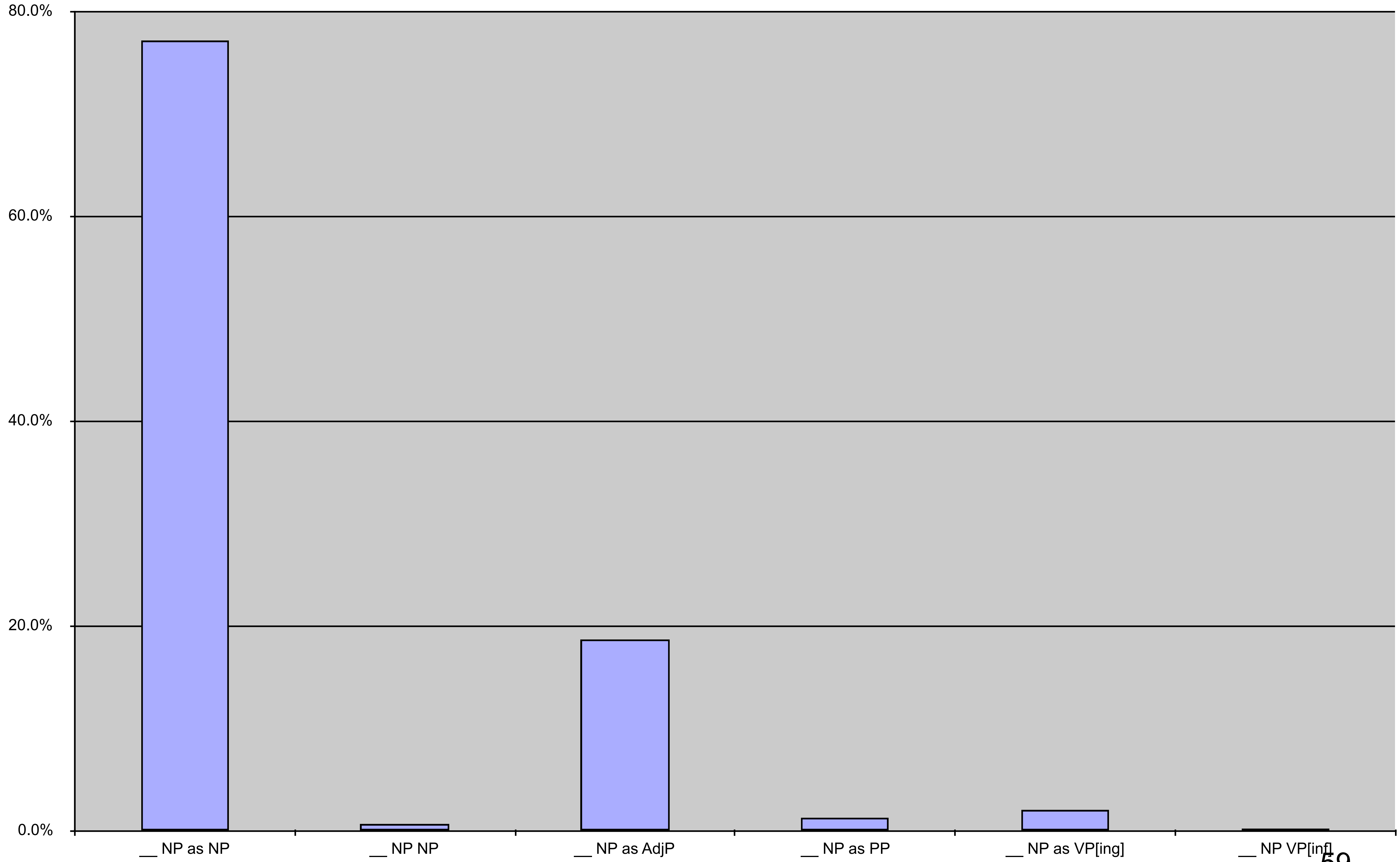
Pollard and Sag (1994):

- Kim ended up political
- *Kim ended up sent more and more leaflets

The New York Times:

- On the big night, Horatio ended up flattened on the ground like a fried egg with the yolk broken.

Probability mass functions: subcategorization of regard



Leakage leads to change

- People continually stretch the ‘rules’ of grammar to meet new communicative needs, to better align grammar and meaning, etc.
- As a result language slowly changes
  - **while:** used to be only a noun (That takes a while); now mainly used as a subordinate clause introducer (While you were out)
  - **e-mail:** started as a mass noun like mail (most junk e-mail is annoying); it’s moving to be a count noun (filling the role of e-letter): I just got an interesting email about that.

Blurring of categories: "Marginal prepositions"

- An example of blurring in syntactic category during linguistic change is so-called 'marginal prepositions' in English, which are moving from being participles to prepositions
- Some still clearly maintain a verbal existence, like following, concerning, considering; for some it is marginal, like according, excepting; for others their verbal character is completely lost, such as during [cf. endure], pending, notwithstanding.

Verb (VBG) ? Preposition IN

As verbal participle, understood subject agrees with noun:

- They moved slowly, toward the main gate, following the wall
- Repeat the instructions following the asterisk

A temporal use with a controlling noun becomes common:

- This continued most of the week following that ill-starred trip to church

Prep. uses (meaning is after, no controlling noun) appear

- He bled profusely following circumcision
- Following a telephone call, a little earlier, Winter had said …

Mapping the recent change of following: participle → prep.

- Fowler (1926): "there is a continual change going on by which certain participles or adjectives acquire the character of prepositions or adverbs, no longer needing the prop of a noun to cling to … [we see] a development caught in the act"
- Fowler (1926) -- no mention of following in particular
- Fowler [Gowers] (1948): "Following is not a preposition. It is the participle of the verb follow and must have a noun to agree with"
- Fowler [Gowers] (1954): generally condemns temporal usage, but says it can be justified in certain circumstances

2. Explaining more: What do people say?

- What people do say has two parts:
  - Contingent facts about the world
    - People in Minnesota have talked a lot about snow falling, not stocks falling, lately
  - The way speakers choose to express ideas using the resources of their language
    - People don't often put that-clauses pre-verbally:
      - That we will have to revise this program is almost certain
- The latter is properly part of people's Knowledge of Language—i.e., part of linguistics.

What do people say?

- Simply delimiting a set of grammatical sentences provides only a very weak description of a language, and of the ways people choose to express ideas in it
- Probability densities over sentences and sentence structures can give a much richer view of language structure and use
- In particular, we find that the same soft generalizations and tendencies of one language often appear as (apparently) categorical constraints in other languages
- A syntactic theory should be able to uniformly capture these constraints, rather than only recognizing them when they are categorical

Example: Bresnan, Dingare & Manning

- Project modeling English diathesis alternations (active/passive, locative inversion, etc.)
- In some languages passives are categorically restricted by person considerations:
  - In Lummi (Salishan, Washington state), 1/2 person must be the subject if other argument is 3rd person. There is variation if both arguments are 3rd person. (Jelinek and Demers 1983) [cf. also Navajo, etc.]
    - *That example was provided by me
    - *He likes me
    - ?I am liked by him

Bresnan, Dingare & Manning

- In English, there is no such categorical constraint, but we can still see it at work as a soft constraint.
- Collected data from verbs with an agent and patient argument (canonical transitives) from treebanked portions of the Switchboard corpus of conversational American English, analyzing for person and act/pass

	Active	Passive
1/2 Ag, 1/2 Pt	158	0 (0.0%)
1/2 Ag, 3 Pt	5120	1 (0.0%)
3 Ag, 1/2 Pt	552	16 (2.8%)
3 Ag, 3 Pt	3307	46 (1.4%)

Bresnan, Dingare & Manning

- While person is only a small part of the picture in determining the choice of active/passive in English (information structure, genre, etc. is more important), there is nonetheless a highly significant (X^2 $p < 0.0001$) effect of person on active/passive choice
- The exact same hard constraint of Lummi appears as a soft constraint in English
- This behavior is predicted by the universal hierarchies within a stochastic OT model (which extends existing OT approaches to valence – Aissen 1999, Lødrup 1999)
- Conversely linguistic model predicts that no "anti-English" [which is just the opposite] exists

Syntactic Matching

Roger Levy

Conclusions

- There are many phenomena in language that cry out for non-categorical and probabilistic modeling and explanation
- Probabilistic models can be applied on top of one's favorite sophisticated linguistic representations!
- Frequency evidence can enrich linguistic theory by revealing soft constraints at work in language use

What Next?

- Courses you could take
  - Machine Learning
  - Information Retrieval
  - Data Mining
  - Special Topics

What Next?

- People you could talk to
  - Lu Wang
  - Byron Wallace
  - Jay Aslam
  - Tim Bickmore
- People in network science, the social sciences, the humanities, and linguistics working on language data

nothing but the
the history of the
at if ever we did
rties would be so
r the defeat of an
he defeated party
deserved calamity,
var in which we
ever likely to
continent more
when it began,
stance of a purely
e to have even a
for the purpose of
ppression, and not
retation of some
st as independent
vn. If it be pos-
nd America, it is
e of doing so. It
s what we want.
ength is what we
etween two great
revent, or greatly
us, hopeless, and
ivil war in charac-
isies of national

, it would seem
good opportunity
more a cordial
t least if we may
the other side of

events, unless the accounts from many quarters as to General Schenck's instructions are utterly belied, the new American Ambassador will bring us quite *reasonable,* though not perhaps wholly admissible demands,—demands which we certainly ought to consider most gravely, and of which we should do well to yield frankly and freely all that we should ourselves feel called upon, in the same circumstances, to press. If we do so, General Schenck's mission may make England safer and stronger than she has ever been since the close of the Civil War in 1865, and will give her a reputation for moderation and candour as well.

ENGLISH PUBLIC OPINION ON THE WAR.

SOME of the philosophers should turn their attention from the subject of spectroscopic investigations and the invention of electrometers, galvanometers, hygrometers, and so forth, to the far more difficult problem of inventing a mode of measuring the intensity and diffusion of political wishes and convictions. No task at present is more difficult for a Statesman than this. There are, indeed, all sorts of shades of difference between the character of really prevalent and preponderant public opinions, of which no man, however acute, ever forms more than a purely conjectural impression, and of which, nevertheless, any respectably-accurate measure would be a matter of the highest political importance. For instance, there is at times a public opinion on one side of a question which is very widely diffused, but of very slight intensity,—which, in fact, amounts to nothing more than a wish in a particular direction without a will, and still more without any intention of submitting to a considerable sacrifice rather than not carry out the will into action. Again, there is such a thing as