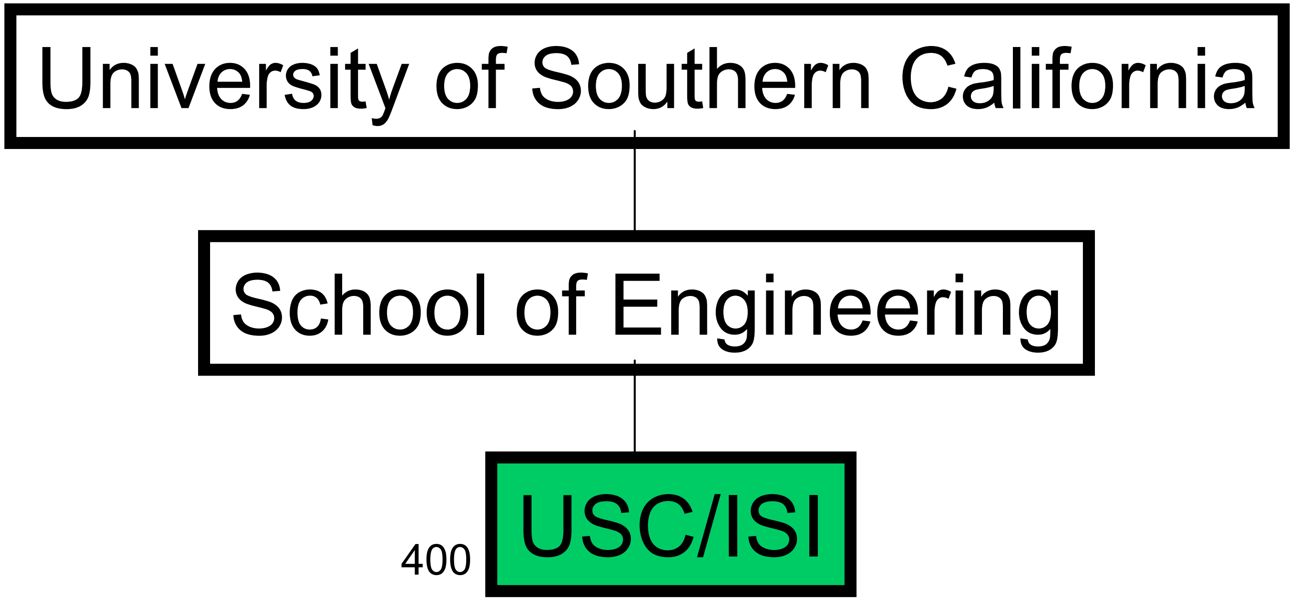
More EM Applications

Kevin Knight

CS562

Oct 17, 2006

Reference: "Results on Decipherment Problems," (Knight, Nair, Rathod, Yamada).
www.isi.edu/~knight.



University of Southern California

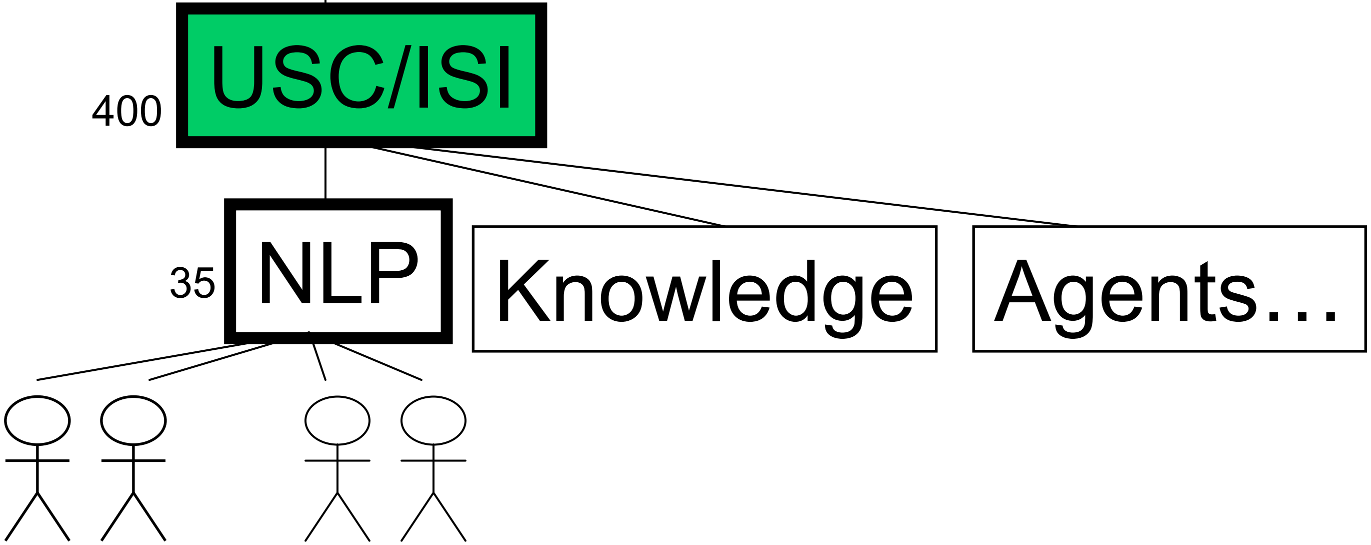
School of Engineering

400

USC/ISI

University of Southern California

School of Engineering



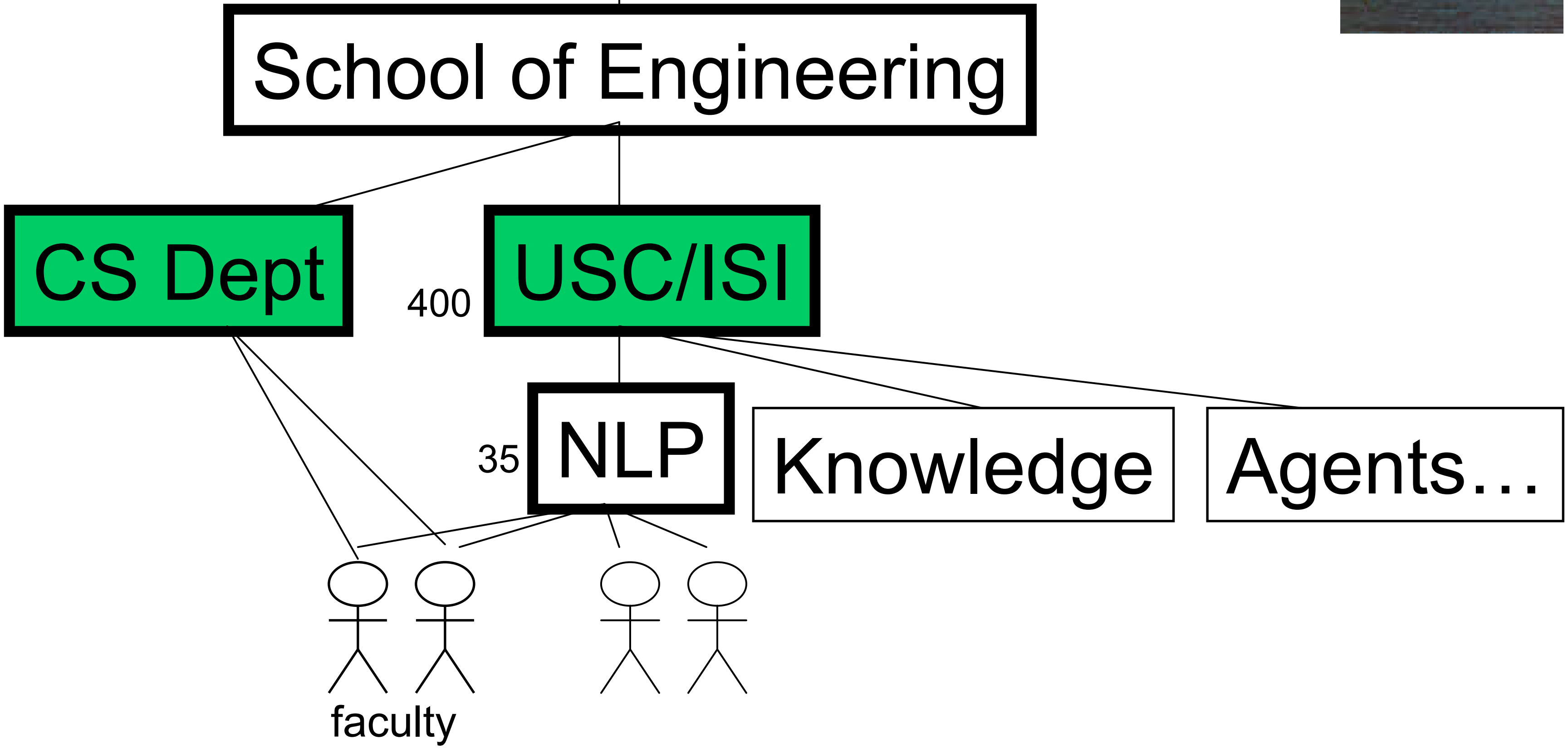
400 USC/ISI

35 NLP

Knowledge

Agents…

University of Southern California



School of Engineering

CS Dept

400

USC/ISI

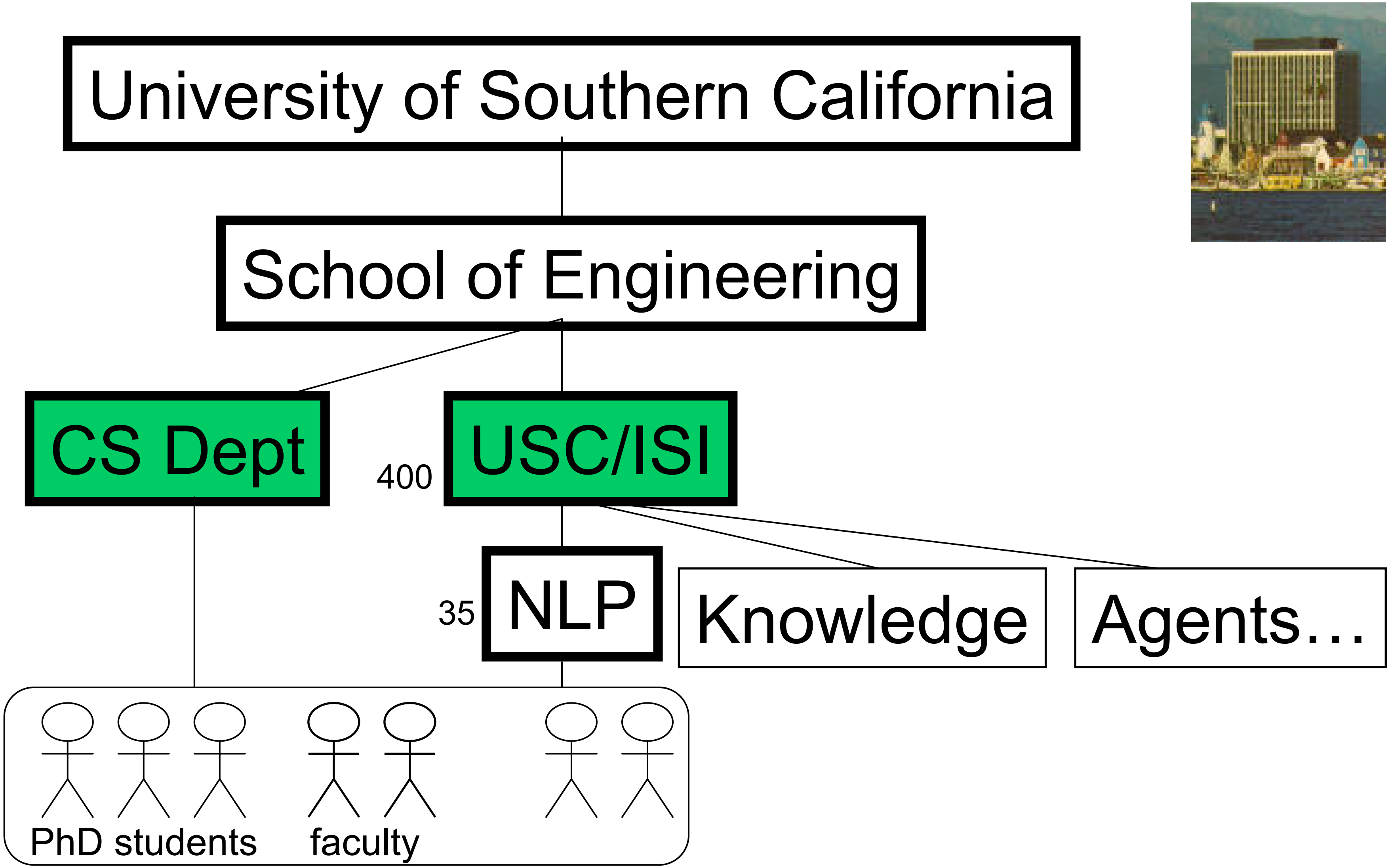
35

NLP

Knowledge

Agents…

faculty



University of Southern California

School of Engineering

CS Dept

400

USC/ISI

35

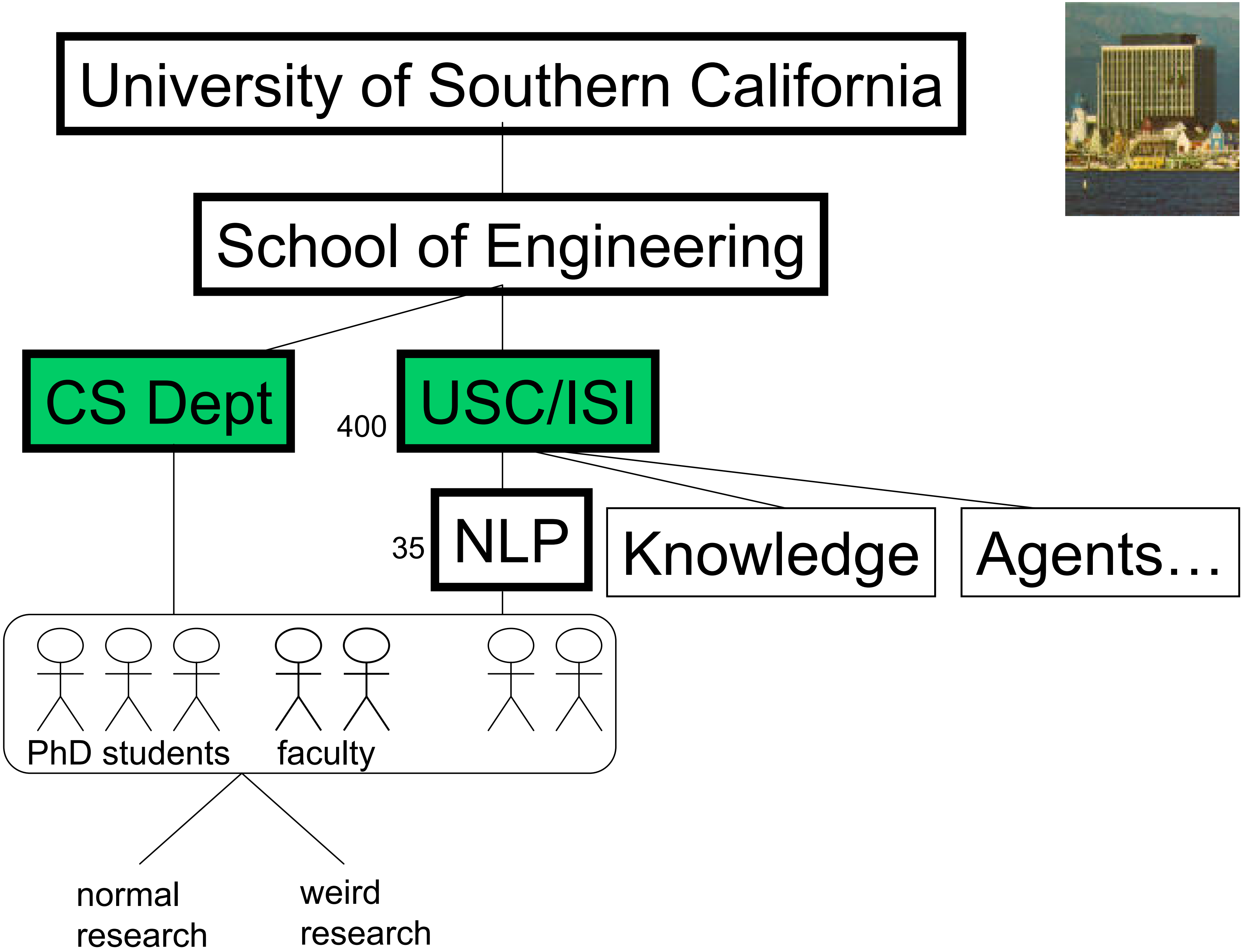
NLP

Knowledge

Agents…

PhD students

faculty



University of Southern California

School of Engineering

CS Dept

400

USC/ISI

35

NLP

Knowledge

Agents…

PhD students

faculty

normal research

weird research

Warren Weaver

ingcmpnqsnwf cv fpn owoktvcv

hu ihgzsnwfv rqcffnw cw owgcnwf

kowazoanv ...

Warren Weaver

```
 e    e  e        e
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```

Warren Weaver

```
 e    e  e      the
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```

Warren Weaver

```
 e   he  e      the
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e t
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```

Warren Weaver

```
 e   he  e   of the
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e t
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```

Warren Weaver

```
 e   he  e   of the      fof
ingcmpnqsnwf cv fpn owoktvcv
        e  f   o  e  o     oe t
hu ihgzsnwfv rqcffnw cw owgcnwf
       ef
kowazoanv ...
```

Warren Weaver

 e   he e   ~~of~~ the

ingcmpnqsnwf cv fpn owoktvcv

        e   e   e t

hu ihgzsnwfv rqcffnw cw owgcnwf

       e

kowazoanv ...

Warren Weaver

```
 e   he  e   is the      sis
ingcmpnqsnwf cv fpn owoktvcv
        e  s   i  e  i     ie t
hu ihgzsnwfv rqcffnw cw owgcnwf
       es
kowazoanv ...
```

Warren Weaver

decipherment is the analysis

ingcmpnqsnwf cv fpn owoktvcv

of documents written in ancient

hu ihgzsnwfv rqcffnw cw owgcnwf

languages ...

kowazoanv ...

Warren Weaver

Computational
Cryptography

↑

Can this be computerized?

"When I look at an article in Russian, I say: this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode." (1947)

↓

Statistical Machine Translation

Finite-State
String Automata

Tree
Automata

But that's another talk, now back to weird…

This Talk
Some Novel, Interesting Decipherment Problems

- Ciphertext: some observed sequence
- Plaintext: the “true” sequence behind the ciphertext, normally not obvious
- Deciphering: turning ciphertext into plaintext

- Outline
  - Basic mathematical approach, used in all applications
  - Decipherment application 1
  - Decipherment application 2
  - Decipherment application 3
  - Decipherment application 4
  - Decipherment application 5

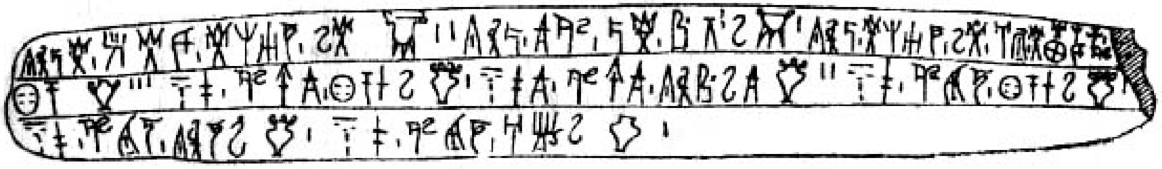
Classic Cryptanalysis

- Ciphertext: **XZPPT ETQPV ...**
- Plaintext: **HELLO WORLD ...**

- People can solve simple ciphers with pencil and eraser
- Computers solve them quite differently (we'll get to that)

Ancient Civilizations

- Ciphertext:

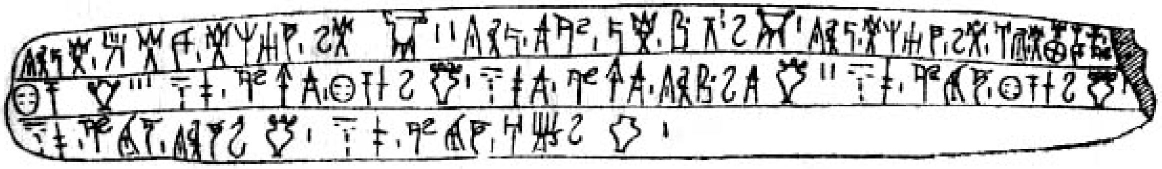


- Plaintext:

- Linear B, Mayan hieroglyphs, Egyptian hieroglyphs, Easter Island glyphs...

Ancient Civilizations

- Ciphertext:



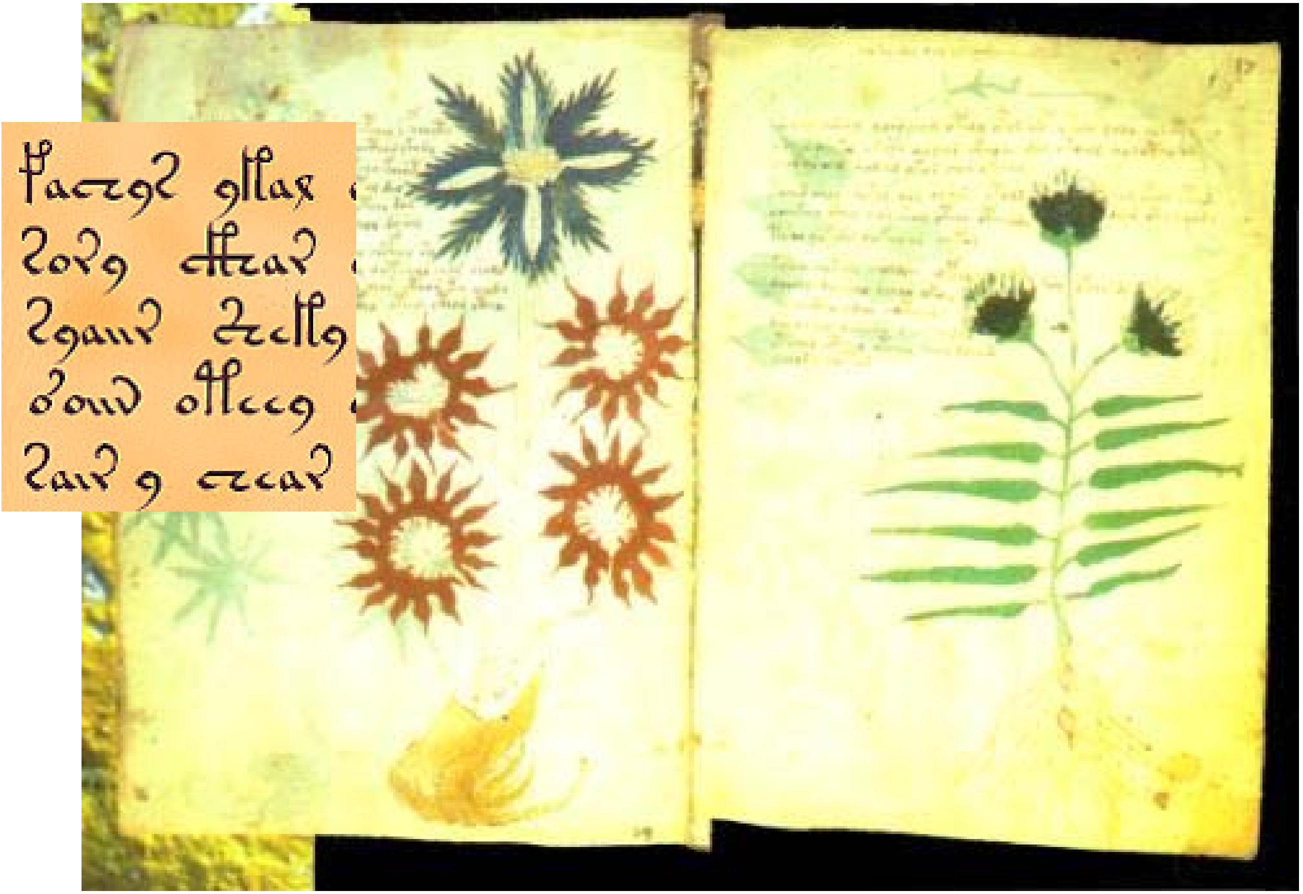
- Plaintext:

  A big vessel with 4 grips, Two big vessels with 3 grips,
  A small vessel with 4 grips, A small vessel with 3 grips, …

- Linear B, Mayan hieroglyphs, Egyptian hieroglyphs, Easter Island glyphs...

Medieval Studies: Voynich Manuscript

- Ciphertext:
  - 20k words
  - illustrated



- Plaintext:
  - unknown!

Romanization and Transliteration

- Ciphertext: アンジラ ナイト
- Plaintext: a n ji ra na i to

easy

“When I look at katakana, I say to myself, this is really English, but it has been encoded in some strange symbols…”

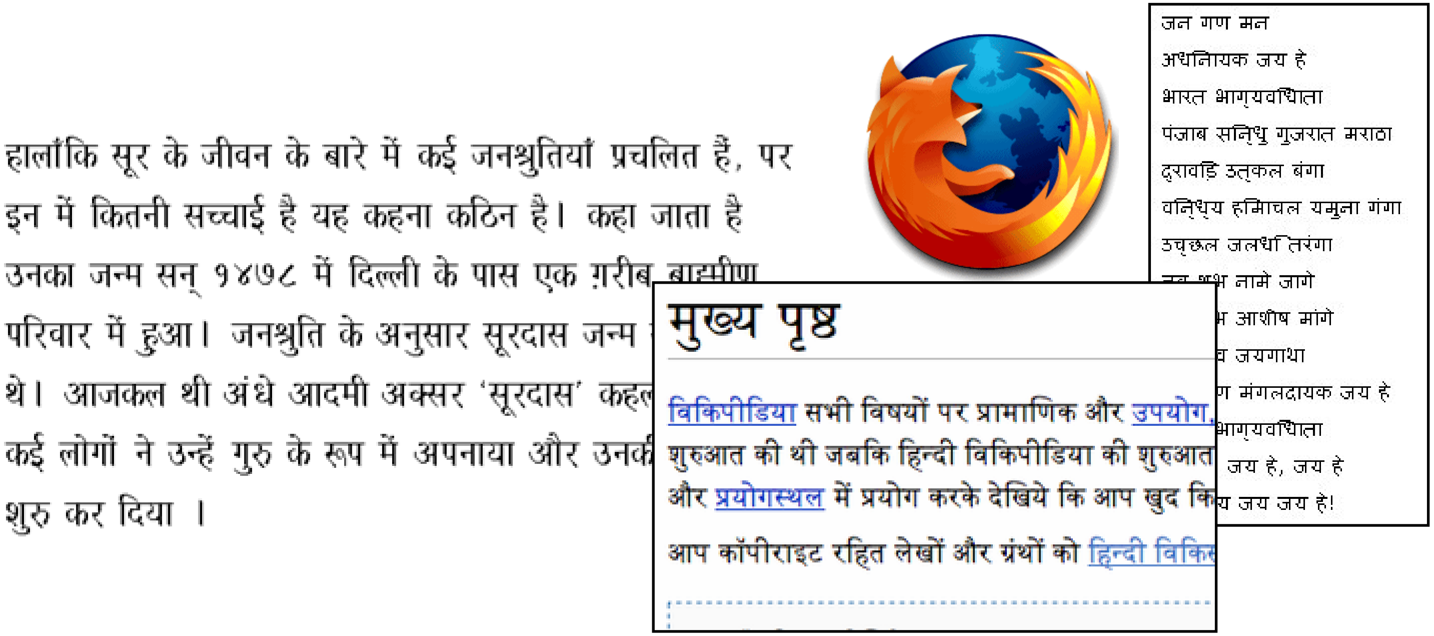
- Ciphertext: アンジラ ナイト
- Plaintext: Angela Knight

hard

[Knight & Graehl 98]

Character Code Conversion

- There are 1000s of languages and lots of character-encoding schemes
  - Spanish/Latin1, Spanish/UTF-8, …
  - Hindi/UTF-8, Hindi/DV-TTYOGESH, Hindi/KRISHNA, and dozens more (“surprise language experiment”)



हालांकि सूर के जीवन के बारे में कई जनश्रुतियां प्रचलित हैं, पर
इन में कितनी सच्चाई है यह कहना कठिन है। कहा जाता है
उनका जन्म सन् १४७८ में दिल्ली के पास एक ग़रीब ब्राह्मण
परिवार में हुआ। जनश्रुति के अनुसार सूरदास जन्म
थे। आजकल थी अंधे आदमी अक्सर ‘सूरदास’ कहल
कई लोगों ने उन्हें गुरु के रूप में अपनाया और उनक
शुरु कर दिया ।

मुख्य पृष्ठ

विकिपीडिया सभी विषयों पर प्रामाणिक और उपयोग,
शुरुआत की थी जबकि हिन्दी विकिपीडिया की शुरुआत
और प्रयोगस्थल में प्रयोग करके देखिये कि आप खुद कि
आप कॉपीराइट रहित लेखों और ग्रंथों को हिन्दी विकि

जन गण मन
अधनिायक जय हे
भारत भाग्यवधिाता
पंजाब सनि्धु गुजरात मराठा
द्रावड़ि उत्कल बंगा
वनि्ध्य हमिाचल यमुना गंगा
उच्छल जलधतिरंगा
भ नामे जागे
भ आशीष मांगे
य जयगाथा
ग मंगलदायक जय हे
भाग्यवधिाता
जय हे, जय हे
य जय जय हे!

Character Code Conversion

- Ciphertext:
  - `20 77 76 118 17 146 42 12 ...`

    (Hindi byte sequence in an unknown encoding system)

- Plaintext:
  - `15 122 101 98 97 32 8 65 42 ...`

    (Hindi byte sequence in UTF-8)

Deciphering Alien Messages from Space

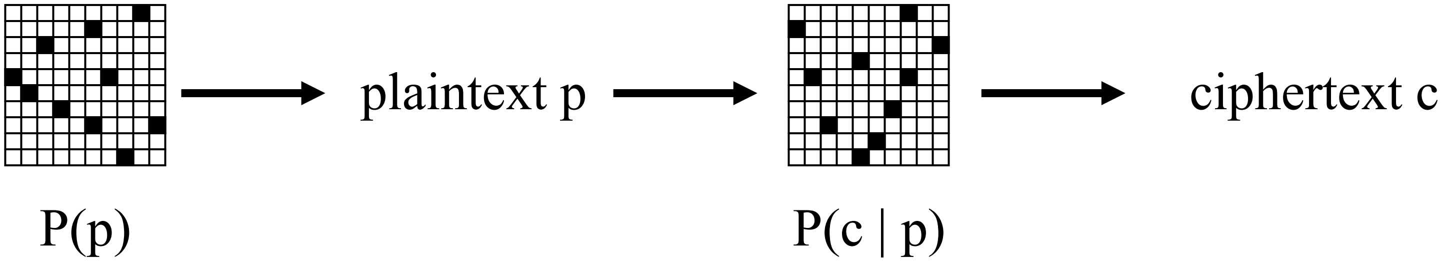
Jodie Foster, excellent actor

Deciphering Alien Messages from Home

Basic Approach

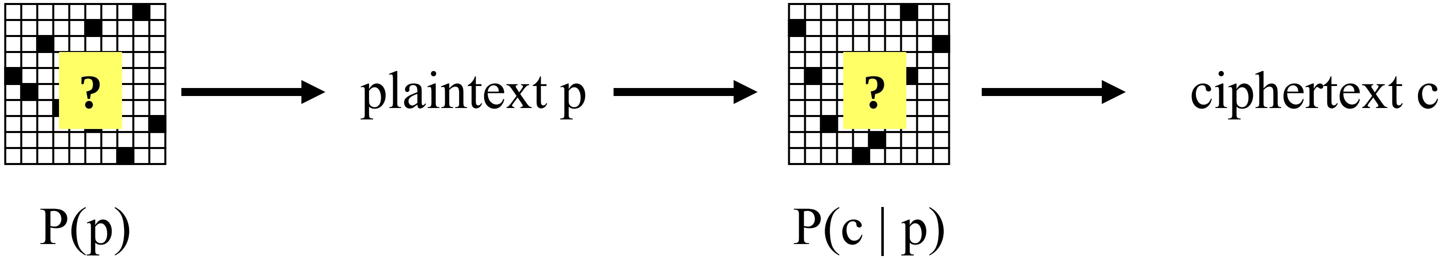
ciphertext c

Basic Approach



P(p) → plaintext p → P(c | p) → ciphertext c

Basic Approach



plaintext p

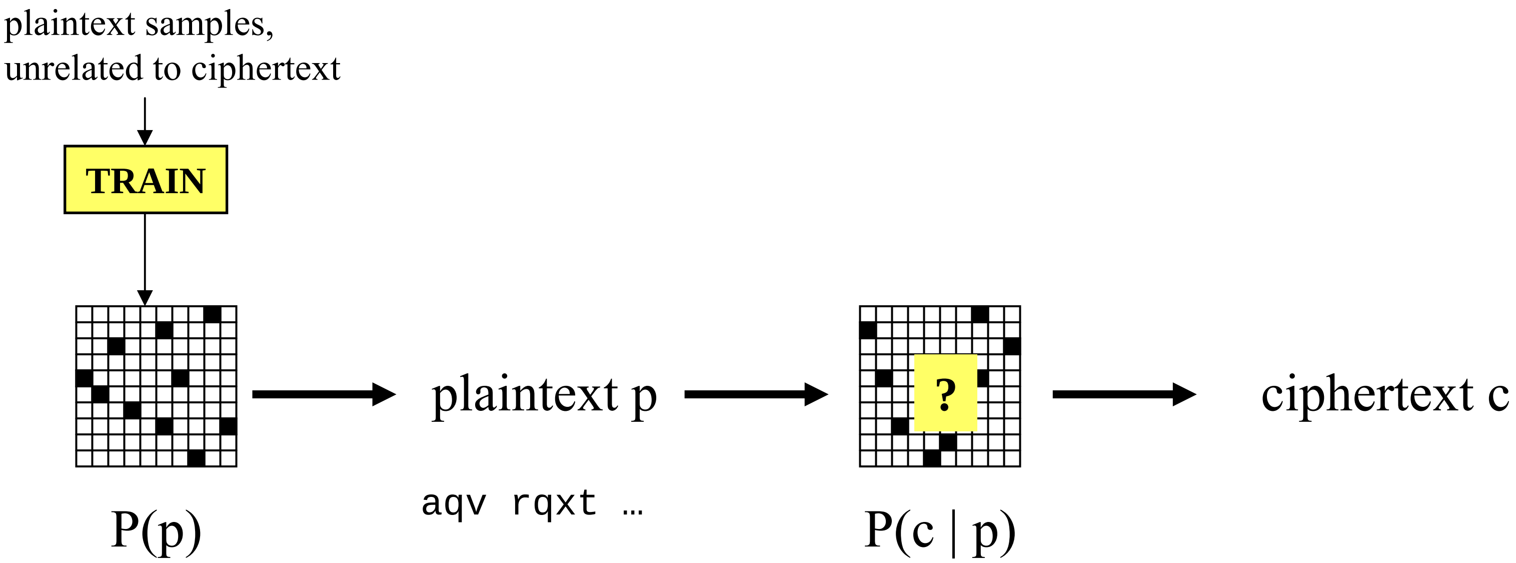
ciphertext c

P(p)

P(c | p)

General knowledge about the plaintext language will drive decipherment.

Basic Approach



plaintext samples,
unrelated to ciphertext

TRAIN

P(p)

plaintext p

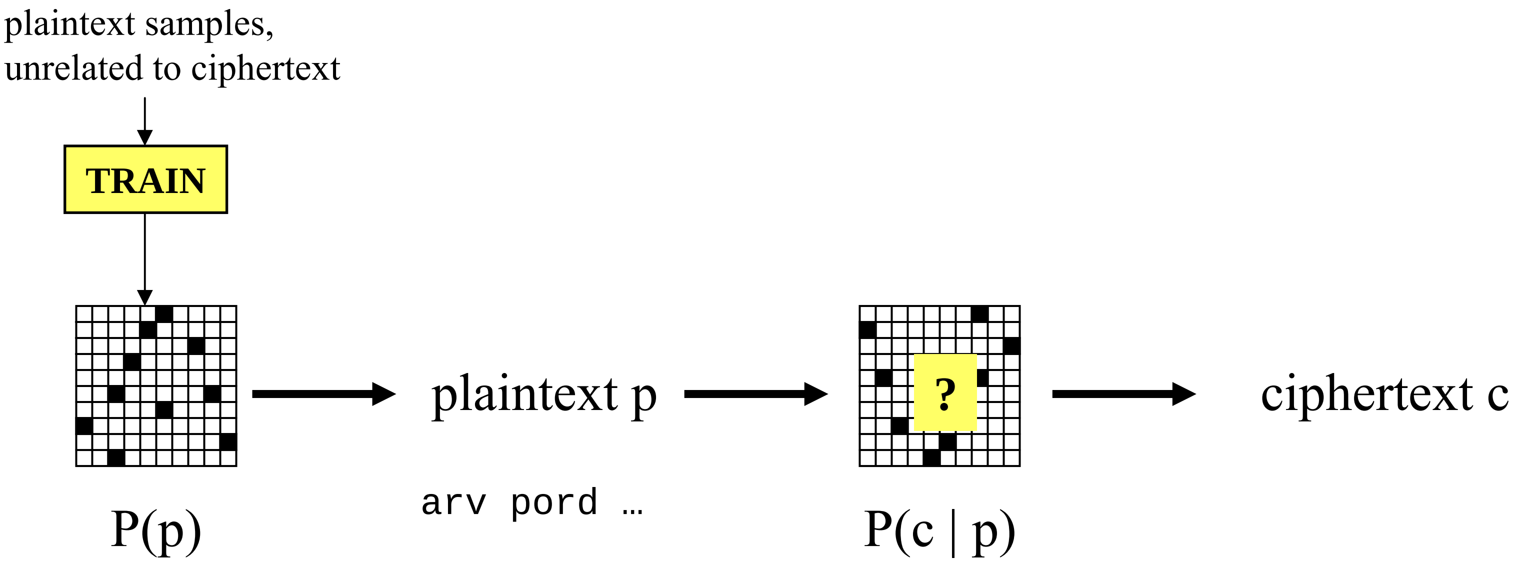
`aqv rqxt …`

?

P(c | p)

ciphertext c

Basic Approach



plaintext samples,
unrelated to ciphertext

TRAIN

P(p)

plaintext p

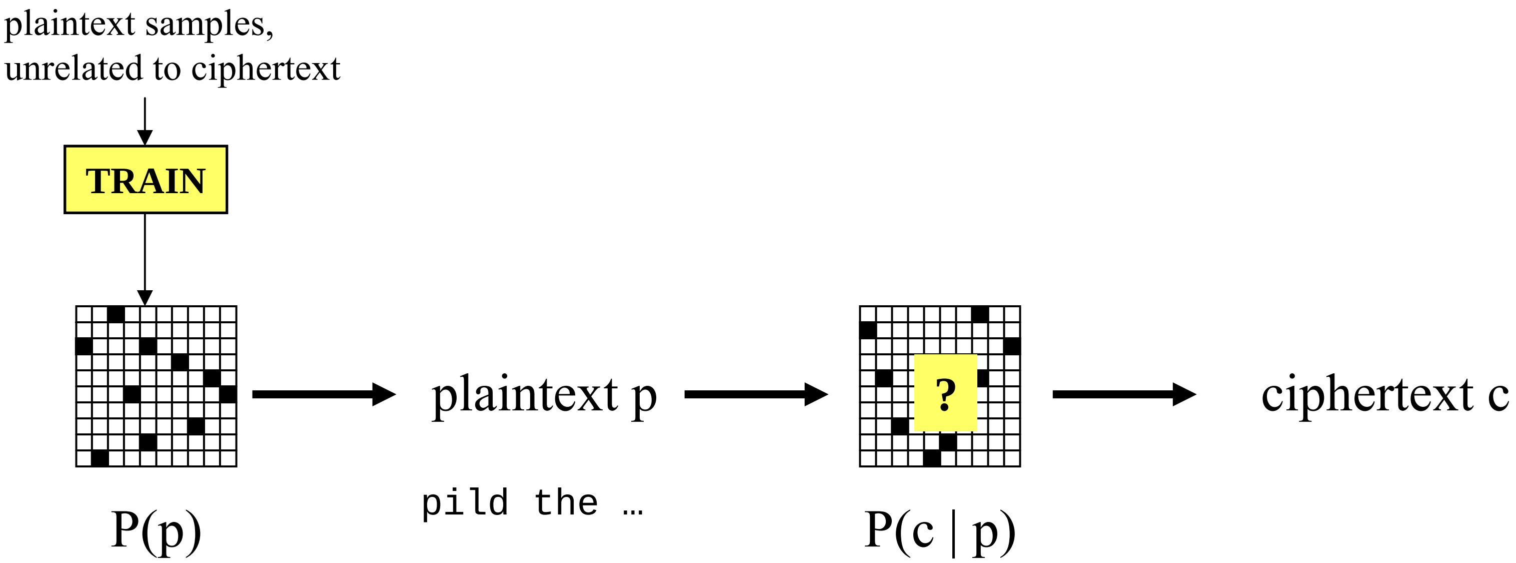
`arv pord …`

?

P(c | p)

ciphertext c

Basic Approach



plaintext samples,
unrelated to ciphertext

TRAIN

P(p)

plaintext p

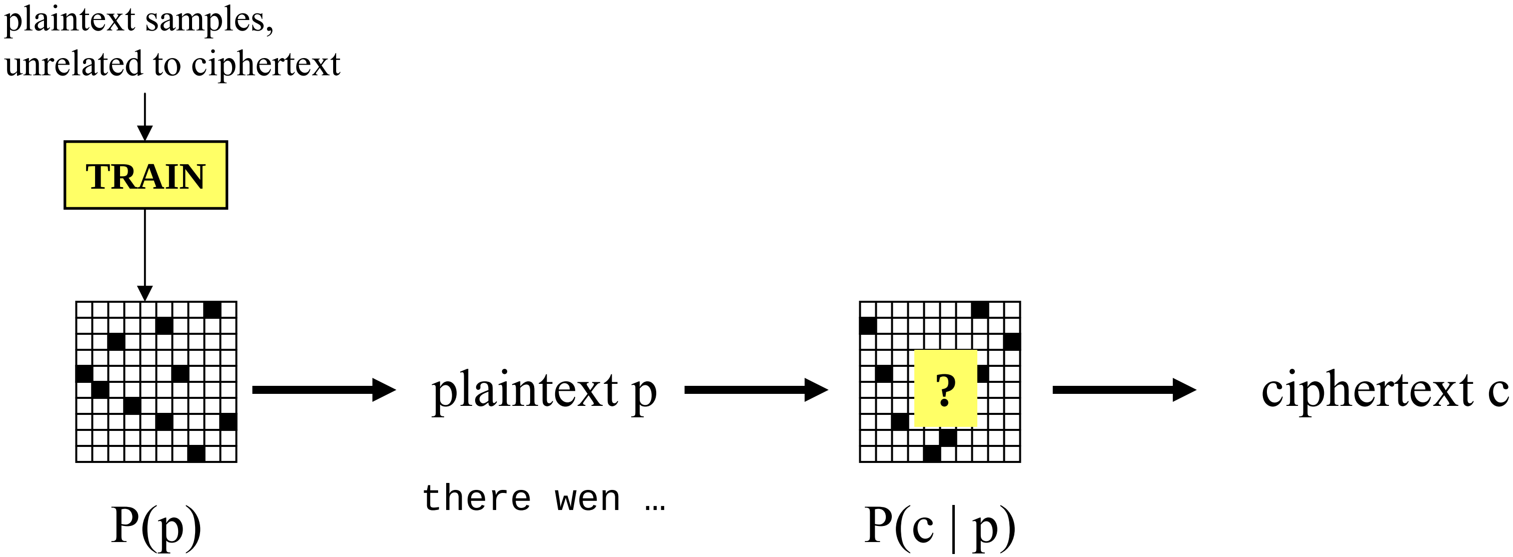
pild the …

?

P(c | p)

ciphertext c

Basic Approach



plaintext samples,
unrelated to ciphertext

TRAIN

P(p)

plaintext p

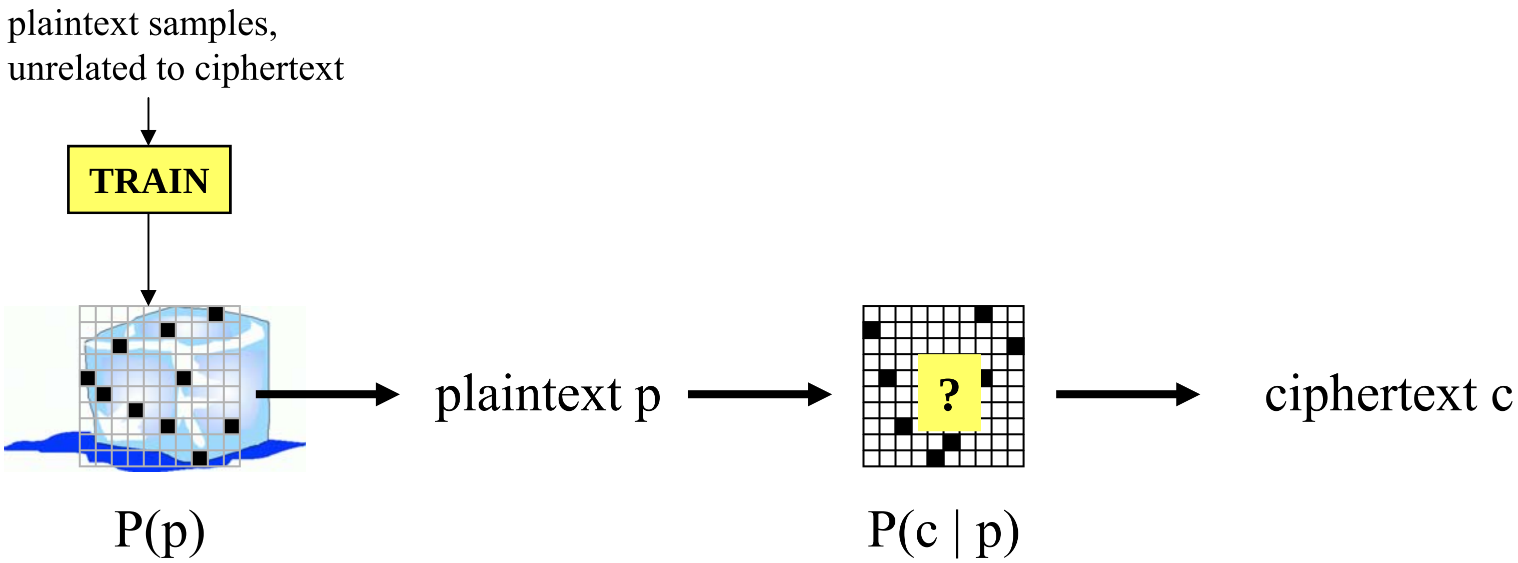
there wen …

?

P(c | p)

ciphertext c

Basic Approach



plaintext samples,
unrelated to ciphertext

TRAIN

plaintext p

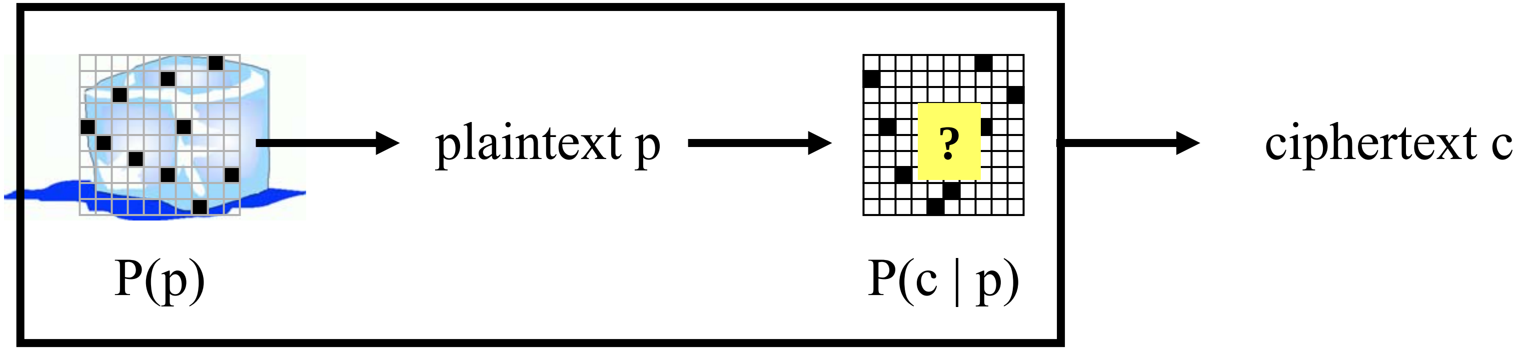
?

ciphertext c

P(p)

P(c | p)

Basic Approach

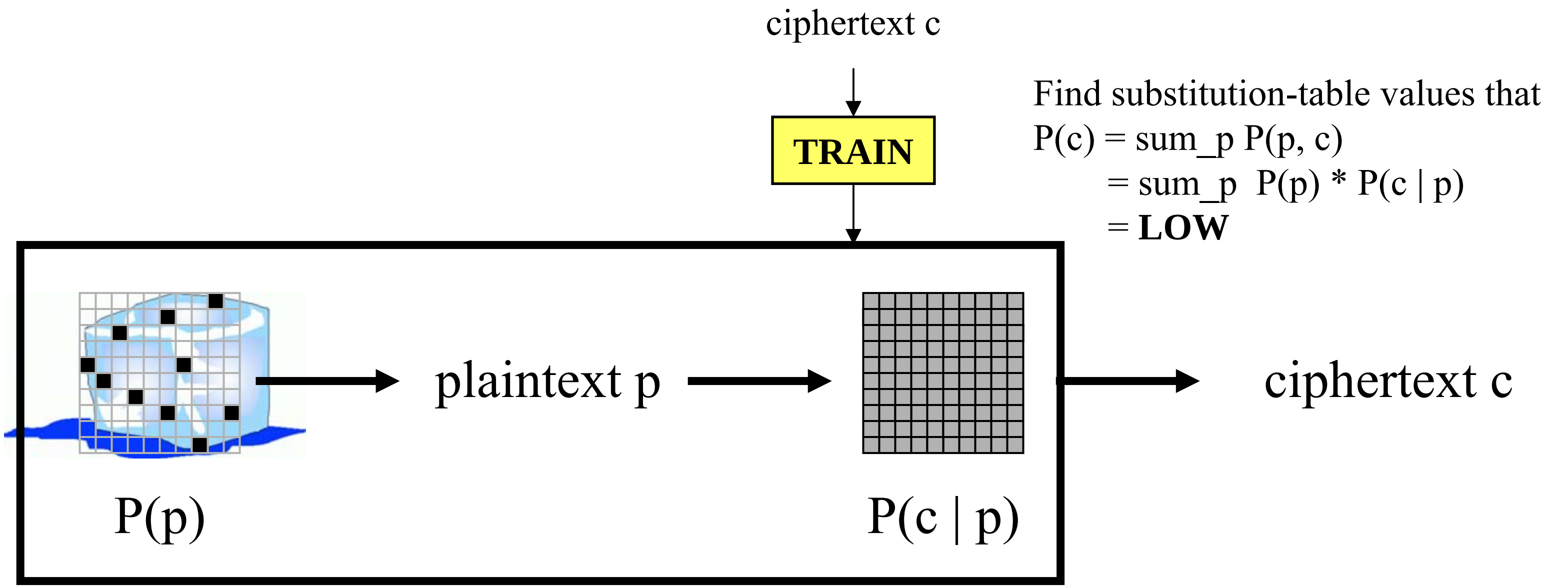


P(p) → plaintext p → P(c | p) → ciphertext c

This whole box is a laser gun that shoots out ciphertexts.

What substitution table would make it most likely to shoot out c?
Or, what substitution table, applied to c, would make it “plaintext-like”?

Basic Approach



ciphertext c

TRAIN

plaintext p

ciphertext c

P(p)

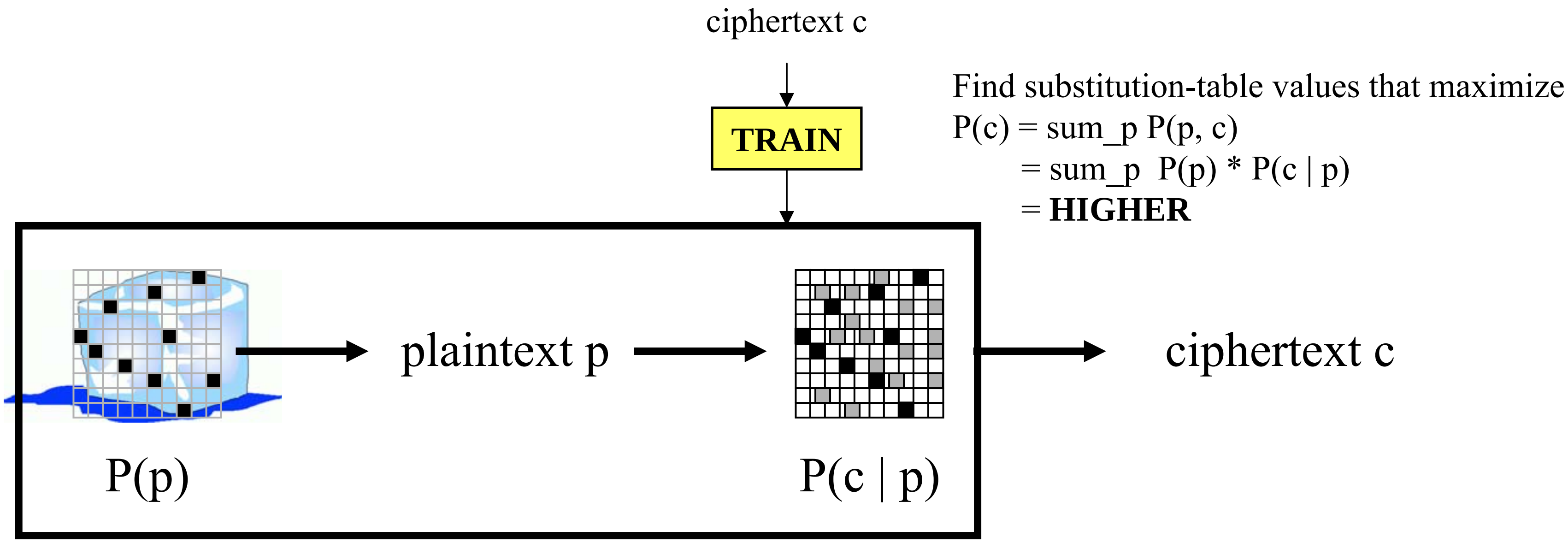
P(c | p)

Find substitution-table values that maximize
P(c) = sum_p P(p, c)
= sum_p P(p) * P(c | p)
= **LOW**

This whole box is a laser gun that shoots out ciphertexts.

What substitution table would make it most likely to shoot out c?
Or, what substitution table, applied to c, would make it “plaintext-like”?

Basic Approach



ciphertext c

TRAIN

Find substitution-table values that maximize
P(c) = sum_p P(p, c)
= sum_p P(p) * P(c | p)
= **HIGHER**

plaintext p

ciphertext c

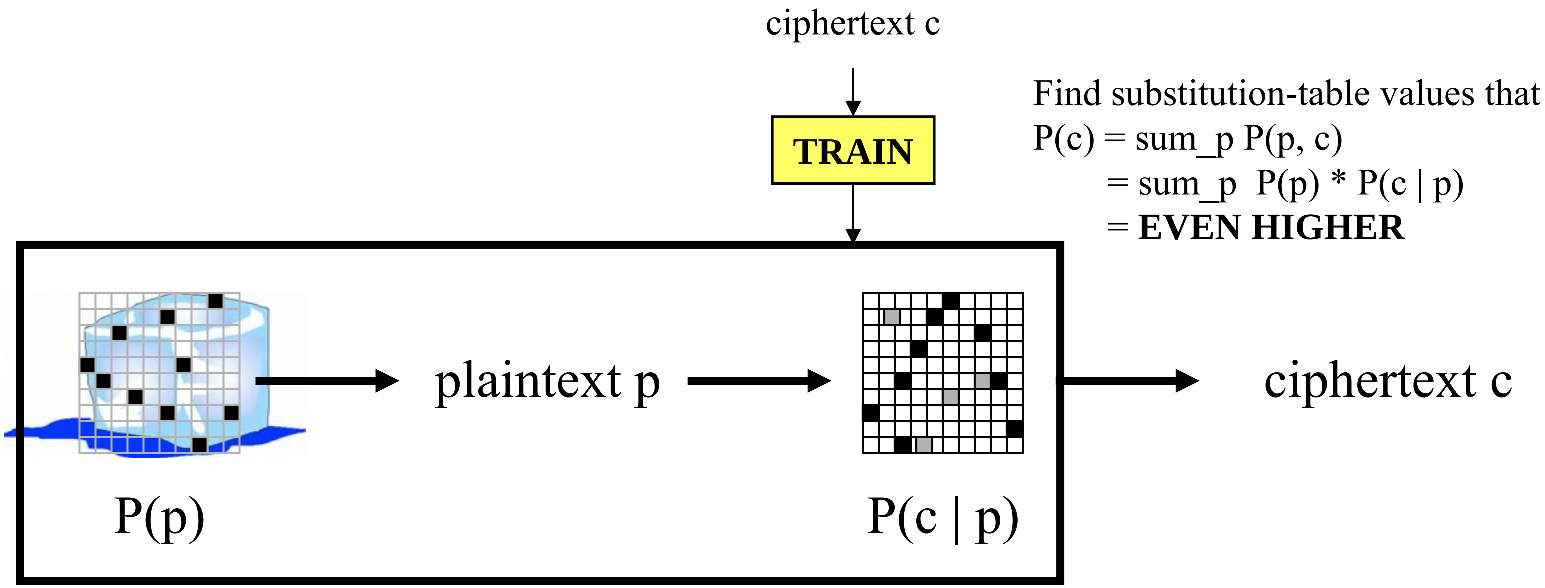
P(p)

P(c | p)

This whole box is a laser gun that shoots out ciphertexts.

What substitution table would make it most likely to shoot out c?
Or, what substitution table, applied to c, would make it “plaintext-like”?

Basic Approach

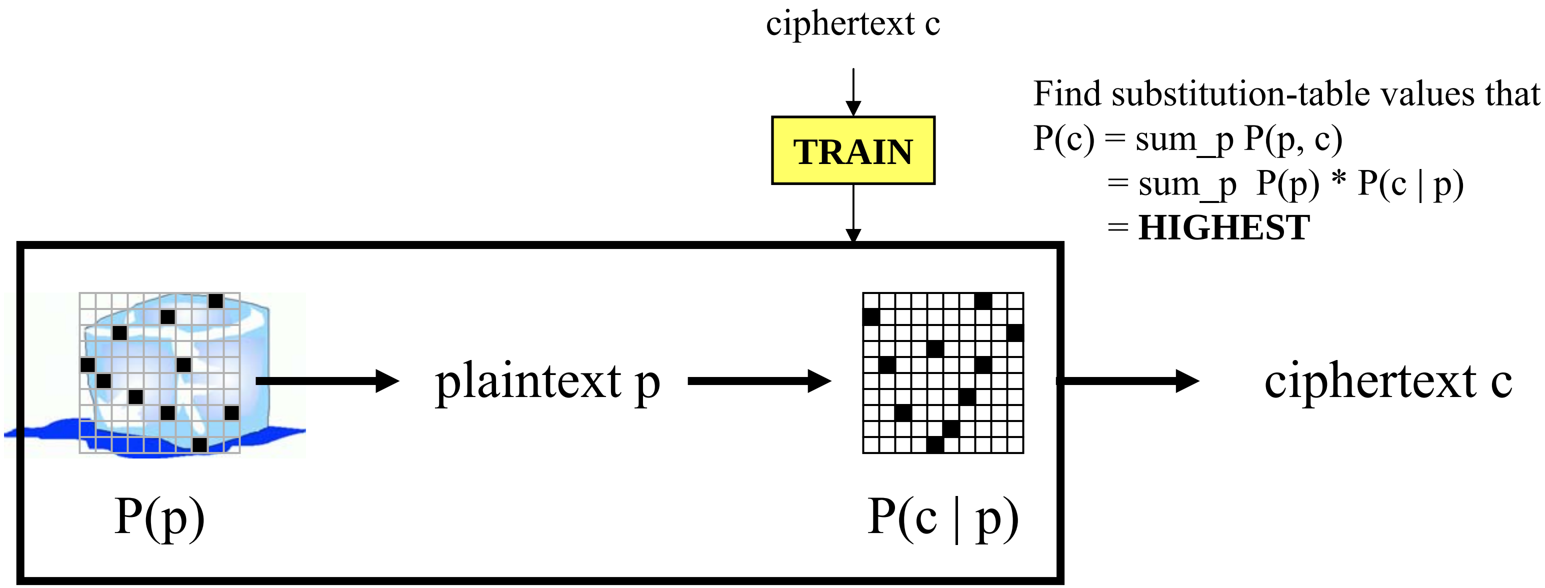


Find substitution-table values that maximize
P(c) = sum_p P(p, c)
= sum_p P(p) * P(c | p)
= **EVEN HIGHER**

This whole box is a laser gun that shoots out ciphertexts.

What substitution table would make it most likely to shoot out c?
Or, what substitution table, applied to c, would make it “plaintext-like”?

Basic Approach

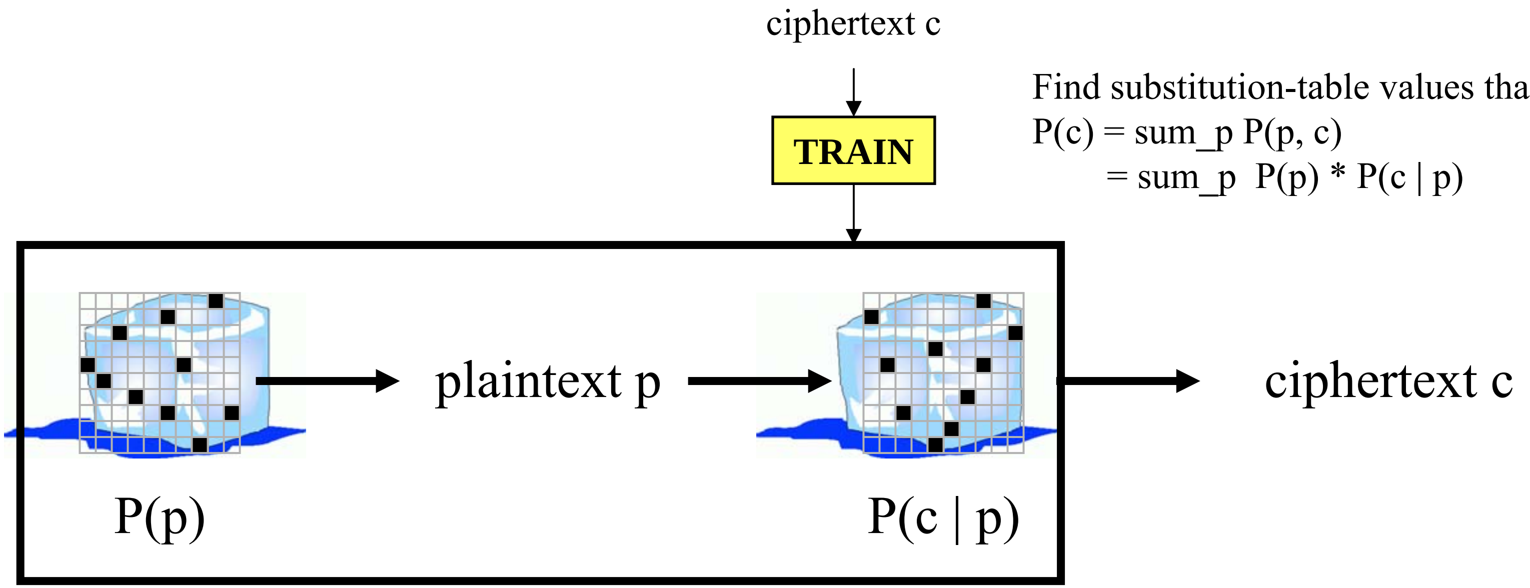


Find substitution-table values that maximize
P(c) = sum_p P(p, c)
= sum_p P(p) * P(c | p)
= **HIGHEST**

This whole box is a laser gun that shoots out ciphertexts.

What substitution table would make it most likely to shoot out c?
Or, what substitution table, applied to c, would make it “plaintext-like”?

Basic Approach



ciphertext c

TRAIN

Find substitution-table values that maximize
P(c) = sum_p P(p, c)
= sum_p P(p) * P(c | p)

plaintext p

ciphertext c

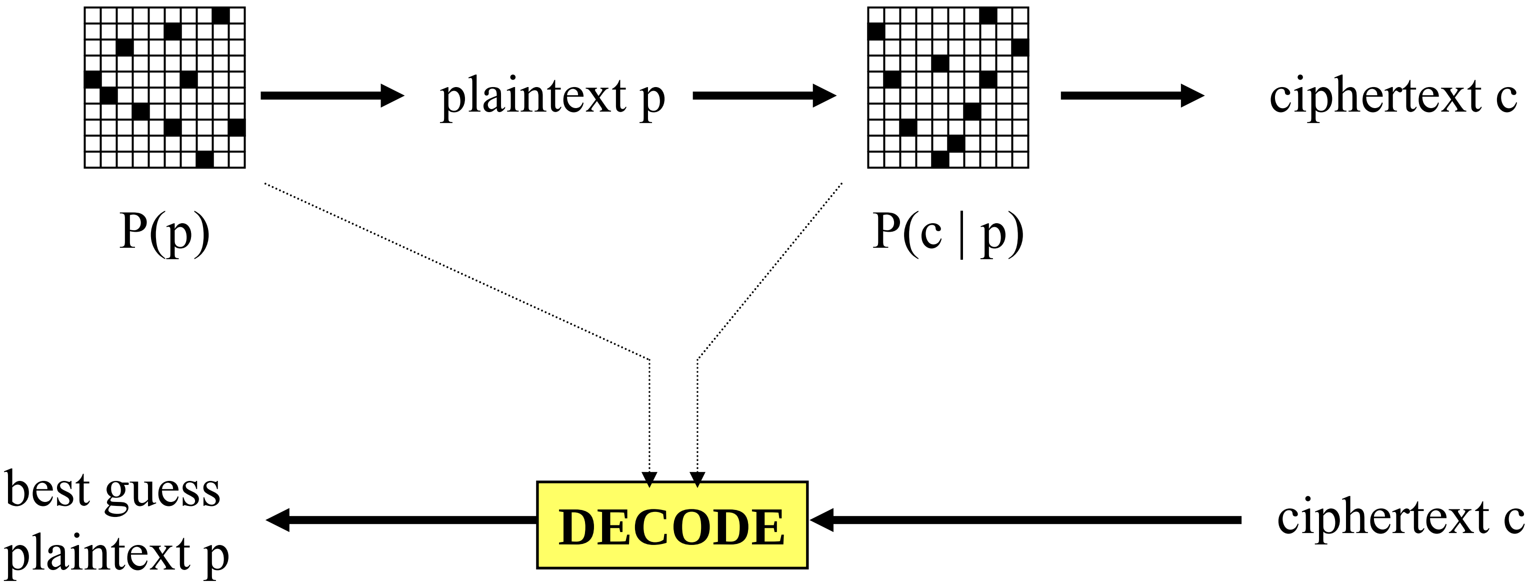
P(p)

P(c | p)

This whole box is a laser gun that shoots out ciphertexts.

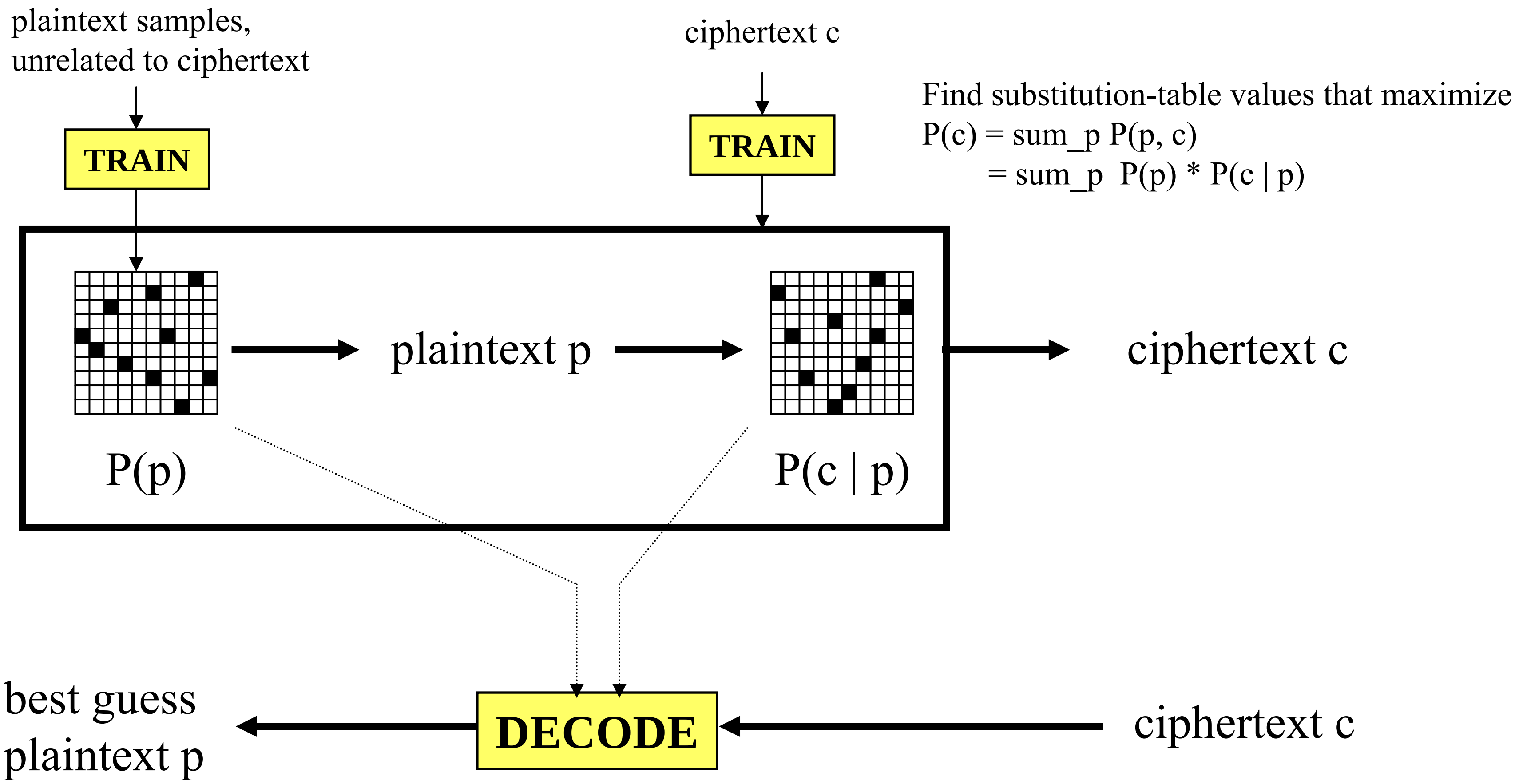
What substitution table would make it most likely to shoot out c?
Or, what substitution table, applied to c, would make it “plaintext-like”?

Basic Approach



Find plaintext p that maximizes
P(p | c) ~ P(p) * P(c | p)

Basic Approach



plaintext samples,
unrelated to ciphertext

TRAIN

ciphertext c

TRAIN

Find substitution-table values that maximize
P(c) = sum_p P(p, c)
= sum_p P(p) * P(c | p)

P(p)

plaintext p

P(c | p)

ciphertext c

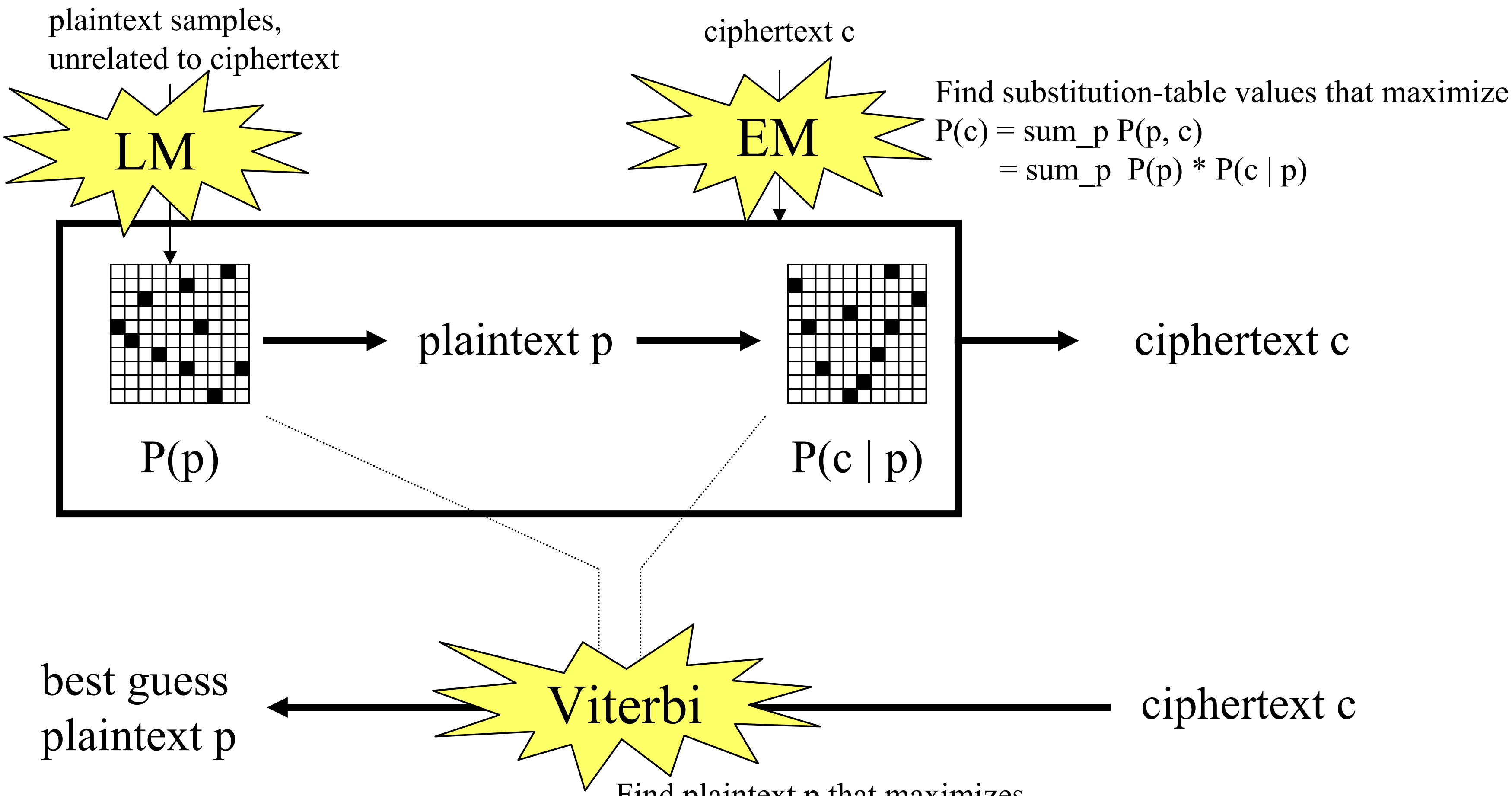
best guess
plaintext p

DECODE

ciphertext c

Find plaintext p that maximizes
P(p | c) ~ P(p) * P(c | p)

Basic Approach



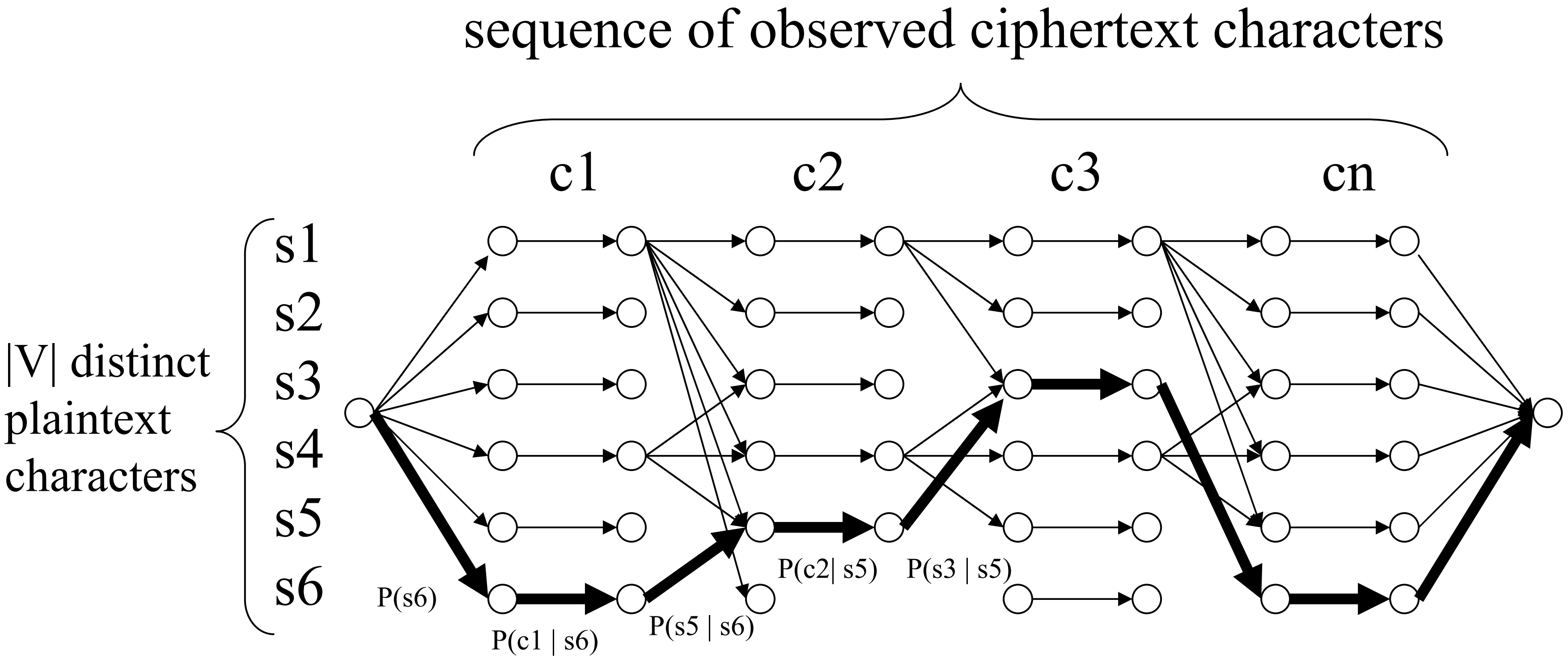
plaintext samples,
unrelated to ciphertext

ciphertext c

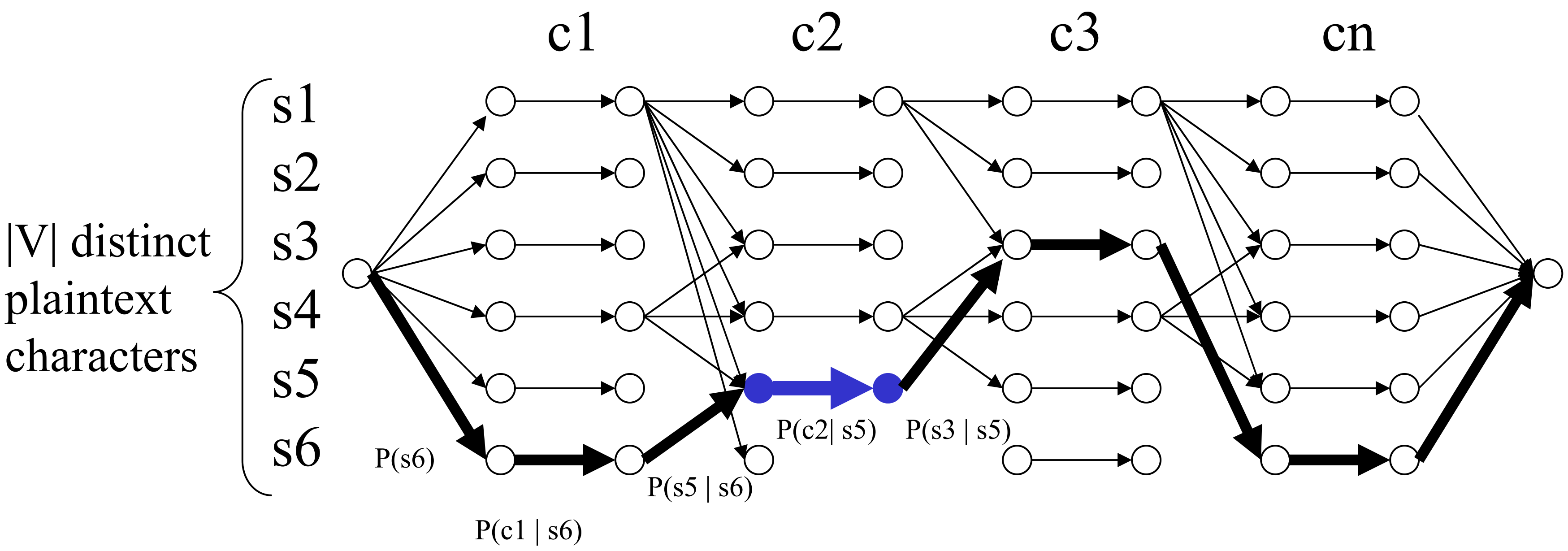
Find substitution-table values that maximize
P(c) = sum_p P(p, c)
= sum_p P(p) * P(c | p)

Find plaintext p that maximizes
P(p | c) ~ P(p) * P(c | p)

Viterbi Decoding [1967]


sequence of observed ciphertext characters
c1
c2
c3
cn
|V| distinct plaintext characters
s1
s2
s3
s4
s5
s6
P(s6)
P(c1 | s6)
P(s5 | s6)
P(c2| s5)
P(s3 | s5)

EM [Baum & Eagon 67]



Repeat:

1. Assign alpha[node] to each node:sum of path costs from start to node
2. Assign beta[node] to each node: sum of path costs from node to end
3. Collect counts for transitions between each node n1 and n2:

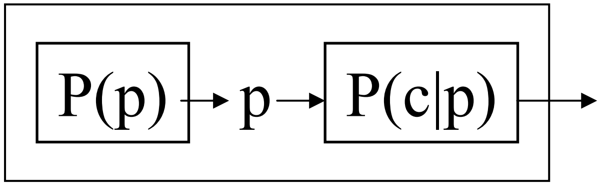
   count(ci, sj) += alpha[n1] * $P(c_j|s_i)$ * beta[n2] / beta[start]

4. Normalize counts into probabilities.

c = ciphertext
p = plaintext

Details

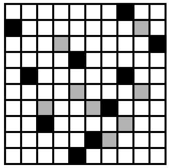
- Generative story
  - how did the observed c get here?
  - decision-oriented, probabilistic



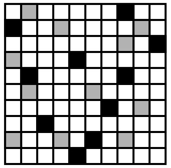
$P(p) = P(p1 \mid START) * P(p2 \mid p1) * \ldots$

$P(c|p) = P(c1 \mid p1) * P(c2 \mid p2) * \ldots$

- Parameters of the story
  - real-valued probs governing decisions



$P(p_i \mid p_j)$



$P(c_j \mid p_i)$

- Formula for P(c)

$$P(c) = \Sigma_p \; P(p) * P(c|p)$$

- Decoding
  - search for s to maximize P(p | c)

search problem!

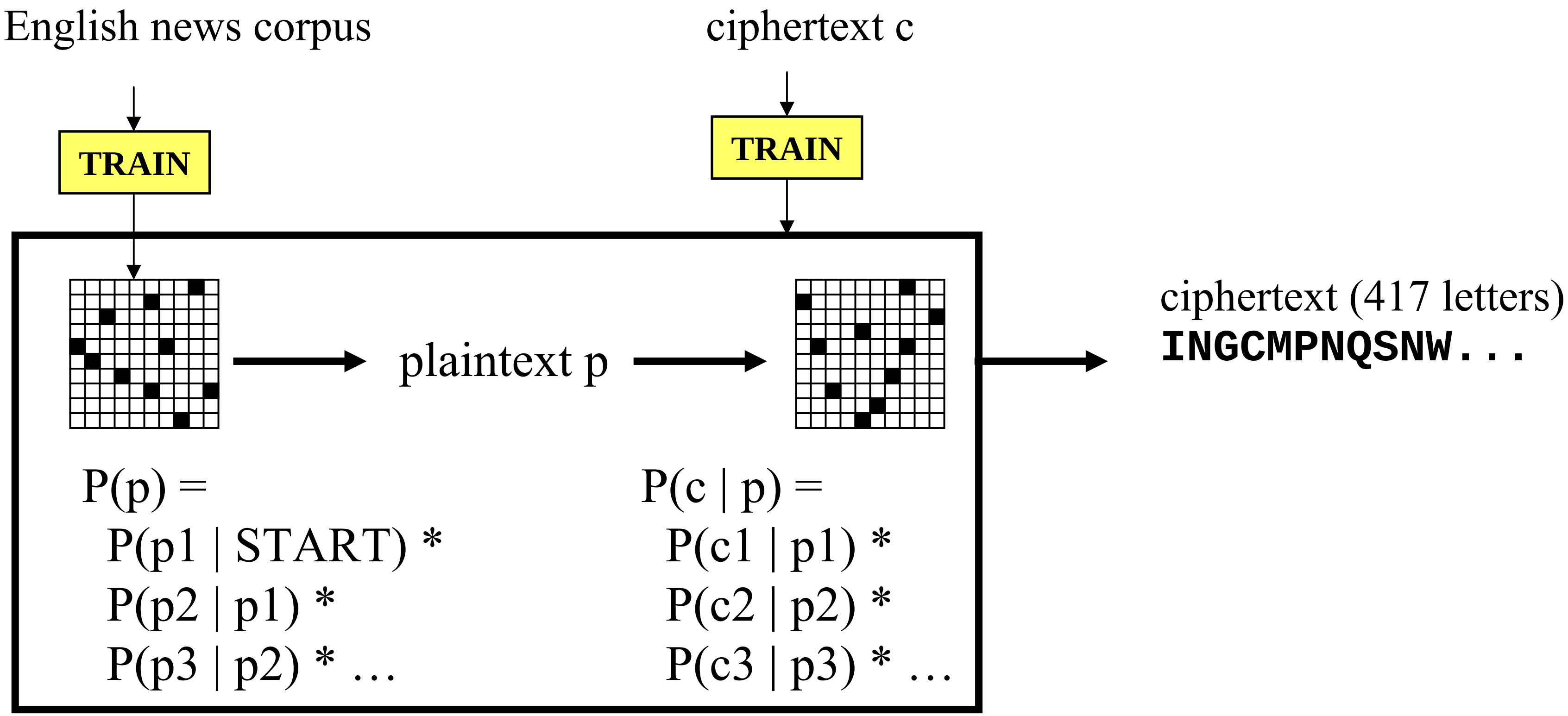
- Training
  - set parameters to maximize P(c)

search problem!

English Letter Substitution Cipher

ciphertext (417 letters)
INGCMPNQSNW...

English Letter Substitution Cipher



English news corpus

TRAIN

plaintext p

ciphertext c

TRAIN

ciphertext (417 letters)

INGCMPNQSNW...

P(p) =
  P(p1 | START) *
  P(p2 | p1) *
  P(p3 | p2) * …

P(c | p) =
  P(c1 | p1) *
  P(c2 | p2) *
  P(c3 | p3) * …

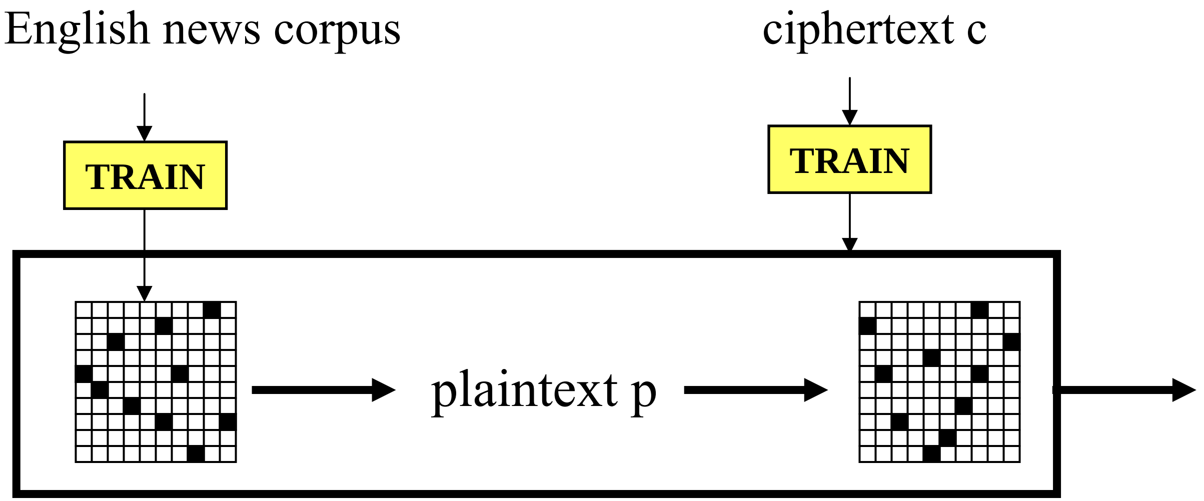
Highest probability decipherment:

```
wecitherkent is the analysis of wocoments pritten in ancient buncquges...
```

Reasonable conclusion:

EM training doesn’t work! Please, stop the madness…

English Letter Substitution Cipher



ciphertext (417 letters)
INGCMPNQSNW...

`wecitherkent is the analysis of wocoments pritten in ancient buncquges...`

	First try	68 errors (17%)
	Plaintext trigrams	57 errors
	More plaintext	32 errors
#1	Decoder maximize $P(p) \cdot P(c \mid p)^3$	15 errors [Knight & Yamada, 1999]
#2	Smooth P(p) model	10 errors
	Gather related web data, retrain P(p)	0 errors (0%)

`decipherment is the analysis of documents written in ancient languages...`

Character Code Conversion

“When I look at this byte sequence, I say to myself, this is really UTF-8 Hindi, but it has been encoded in some strange symbols…”

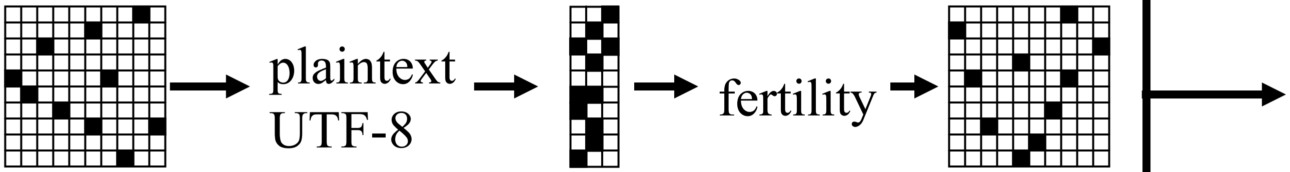
ciphertext (12k bytes)

```
13 5 14 . 16 2
25 26 2 25 . 17
2 3 . 15 2 8 …
```

(Hindi song lyrics)

Character Code Conversion

“When I look at this byte sequence, I say to myself, this is really UTF-8 Hindi, but it has been encoded in some strange symbols…”



P(p) =
  P(p1 | START) *
  P(p2 | p1) *
  P(p3 | p2) * …

P(f | p) =
  P(1 | p1) *
  P(2 | p2) *
  P(1 | p3) *

P(c | p) =
  P(c1 | p1) *
  P(c2 | p2) *
  P(c3 | p3) *

ciphertext (12k bytes)

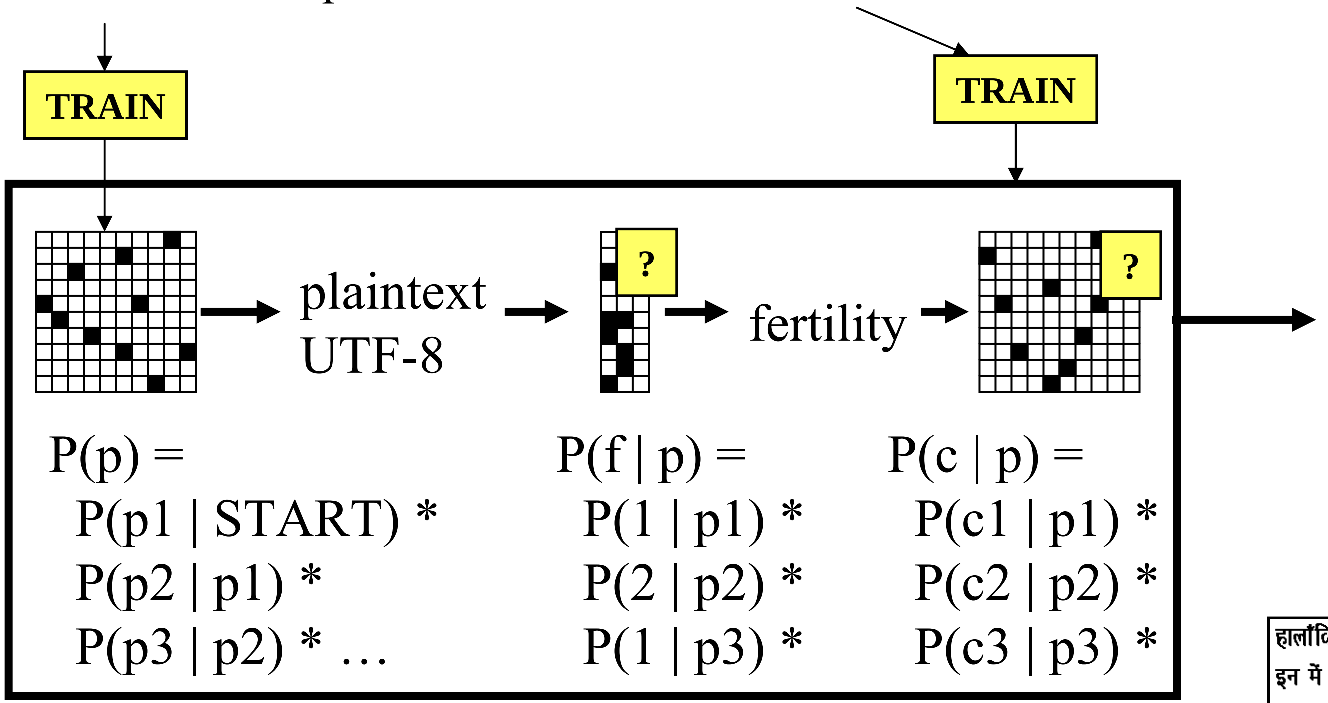
```
13 5 14 . 16 2
25 26 2 25 . 17
2 3 . 15 2 8 …
```

(Hindi song lyrics)

Character Code Conversion

Unrelated
Hindi UTF-8 Corpus

ciphertext c



P(p) =
P(p1 | START) *
P(p2 | p1) *
P(p3 | p2) * …

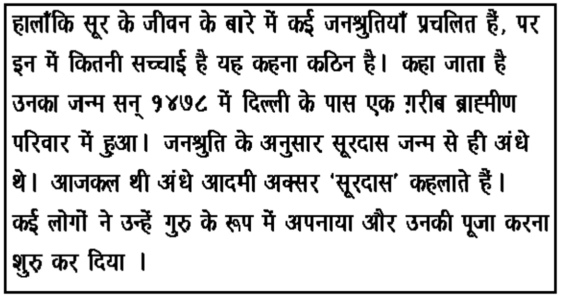
P(f | p) =
P(1 | p1) *
P(2 | p2) *
P(1 | p3) *

P(c | p) =
P(c1 | p1) *
P(c2 | p2) *
P(c3 | p3) *

ciphertext (12k bytes)

13 5 14 . 16 2
25 26 2 25 . 17
2 3 . 15 2 8 …

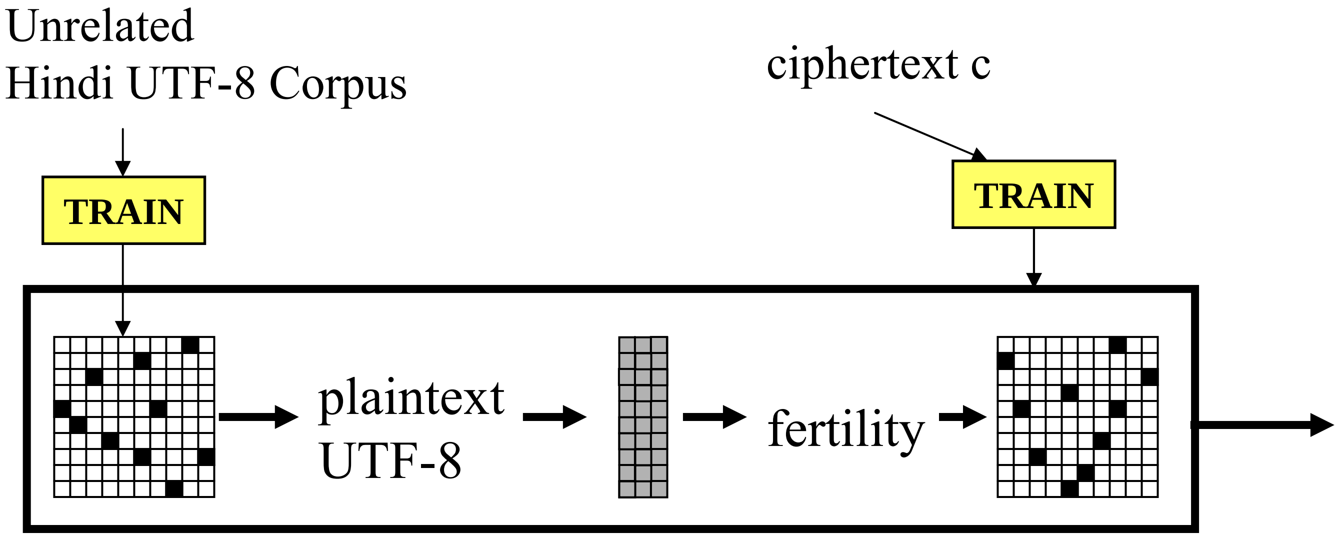
(Hindi song lyrics)



हालाँकि सूर के जीवन के बारे में कई जनश्रुतियाँ प्रचलित हैं, पर
इन में कितनी सच्चाई है यह कहना कठिन है। कहा जाता है
उनका जन्म सन् १४७८ में दिल्ली के पास एक ग़रीब ब्राह्मीण
परिवार में हुआ। जनश्रुति के अनुसार सूरदास जन्म से ही अंधे
थे। आजकल थी अंधे आदमी अक्सर 'सूरदास' कहलाते हैं।
कई लोगों ने उन्हें गुरु के रूप में अपनाया और उनकी पूजा करना
शुरु कर दिया ।

What's the correct plaintext?
Humans can't do it! (Deciphering is hard…)
We cheated: looked at the website display and re-typed in UTF-8.
(Gold standard only for 59 words = 201 UTF-8 characters)

Character Code Conversion



ciphertext (12k bytes)

13 5 14 . 16 2
25 26 2 25 . 17
2 3 . 15 2 8 …

Initial decipherment: (**161** / 201 errors)

Trigram P(p): (**127** / 201 errors)

#3 Fix uniform fertility parameters (don't allow training): (**93** / 201 errors, 15/59 words right)

6 35 . 12 28 49 10 28 . 3 4 6 . 1 10 3 . 29 4 8 20 4 …

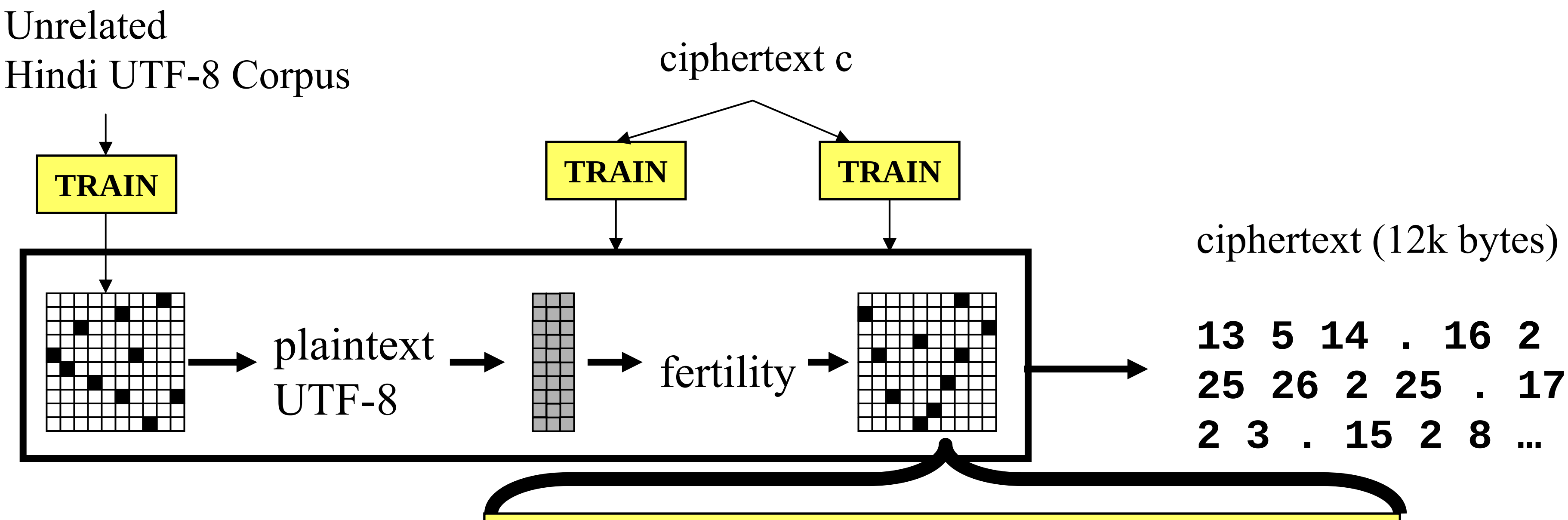
Word-based P(p), trained on top 5000 Hindi UTF-8 words: (**92** / 201 errors, 25/59 words right)

6 35 24 . 12 28 21 4 . 11 6 . 12 25 . 29 8 22 4 …

Correct answer:

6 35 24 . 12 28 21 28 . 3 4 6 . 1 25 . 29 8 20 4 …

Character Code Conversion



ciphertext (12k bytes)

```
13 5 14 . 16 2
25 26 2 25 . 17
2 3 . 15 2 8 …
```

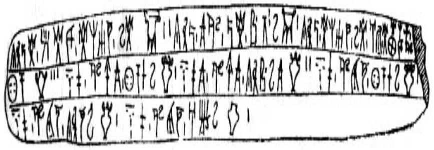
```
P(13 |  6) = 0.66  *      P( 8 | 24) = 0.48
P(32 |  6) = 0.19         P(14 | 24) = 0.33  *
P( 2 |  6) = 0.13         P(17 | 24) = 0.14
P(16 |  6) = 0.02         P(25 | 24) = 0.04

P( 5 | 35) = 0.61  *      P(16 | 12) = 0.58  *
P(14 | 35) = 0.25         P( 2 | 12) = 0.32  *
P( 2 | 35) = 0.15         P(31 | 12) = 0.03
```

First results on unsupervised character code conversion that we know of.
Semi-supervised (align parallel ciphertext/UTF-8 corpus) works fine.

Phonetic Decipherment

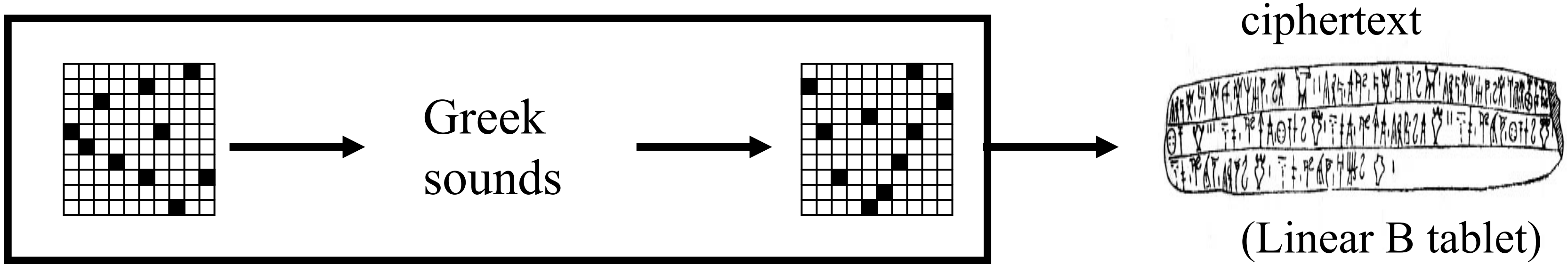
ciphertext



(Linear B tablet)

Phonetic Decipherment

“make the text speak”



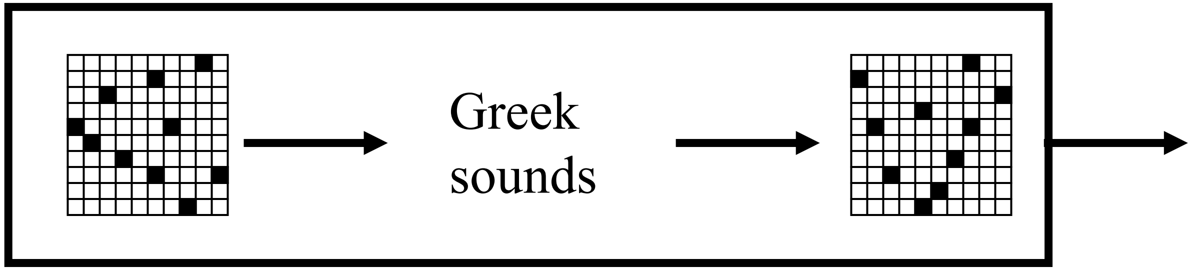
Greek sounds

ciphertext

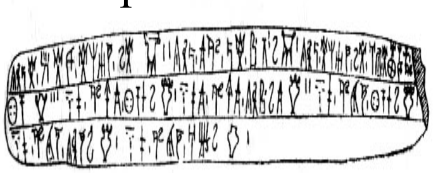
(Linear B tablet)

Phonetic Decipherment

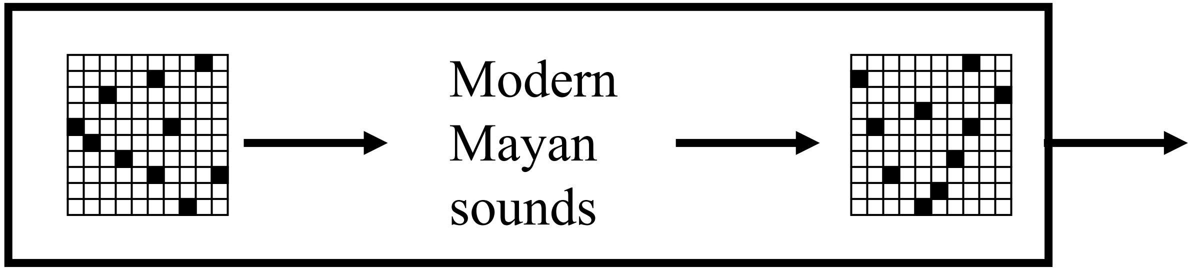
“make the text speak”


Greek sounds

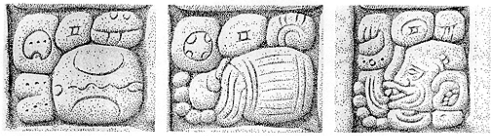
ciphertext



(Linear B tablet)


Modern Mayan sounds

ciphertext



(Mayan writing)

Phonetic Decipherment

ciphertext (6980 letters)

`primera parte`
`del ingenioso`
`hidalgo don …`

(Don Quixote)

32 letters:
ñ, á, é, í, ó, ú,
a, b, c, d, e, f, g,
h, i, j, k, l, m, n,
o, p, q, r, s, t, u
v, w, x, y, z

[Knight & Yamada, 1999]

Phonetic Decipherment

"When I look at these squiggles, I say to myself, this is really a sequence of Spanish phonemes, but it has been encoded in some strange symbols…"

ciphertext (6980 letters)

```
primera parte
del ingenioso
hidalgo don …
```

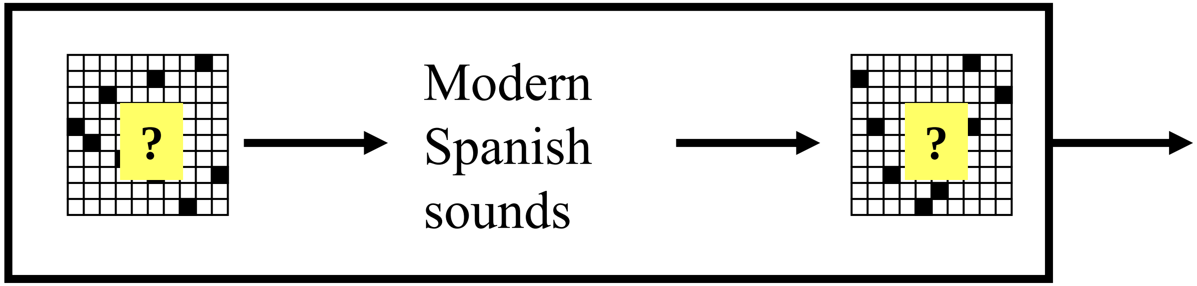
(Don Quixote)

32 letters:

ñ, á, é, í, ó, ú,
a, b, c, d, e, f, g,
h, i, j, k, l, m, n,
o, p, q, r, s, t, u
v, w, x, y, z

[Knight & Yamada, 1999]

Phonetic Decipherment



Modern Spanish sounds

ciphertext (6980 letters)

`primera parte`
`del ingenioso`
`hidalgo don …`

(Don Quixote)

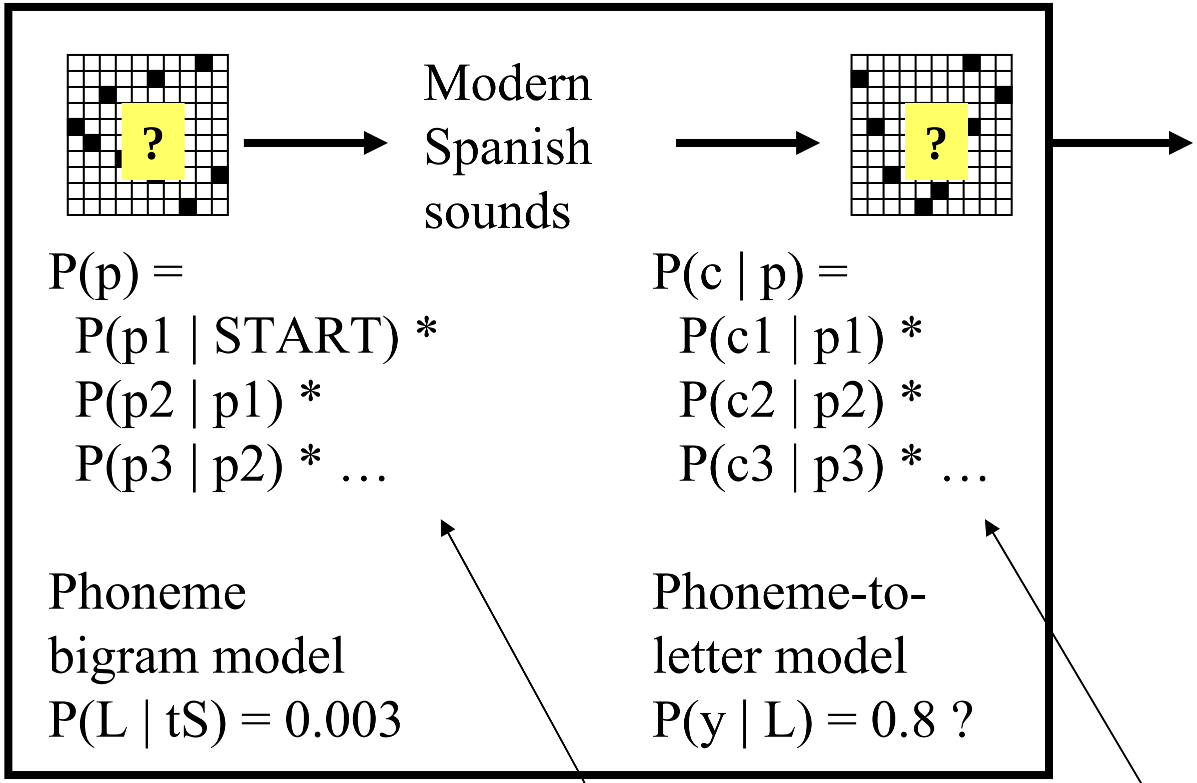
26 sounds:
B, D, G, J (canyon),
L (yarn), T (thin), a,
b, d, e, f, g, i, k, l,
m, n, o, p , r,
rr (trilled), s,
t, tS, u, x (hat)

⟷ ? ⟷

32 letters:
ñ, á, é, í, ó, ú,
a, b, c, d, e, f, g,
h, i, j, k, l, m, n,
o, p, q, r, s, t, u
v, w, x, y, z

[Knight & Yamada, 1999]

Phonetic Decipherment



Modern
Spanish
sounds

P(p) =
 P(p1 | START) *
 P(p2 | p1) *
 P(p3 | p2) * …

Phoneme
bigram model
P(L | tS) = 0.003

P(c | p) =
 P(c1 | p1) *
 P(c2 | p2) *
 P(c3 | p3) * …

Phoneme-to-
letter model
P(y | L) = 0.8 ?

ciphertext (6980 letters)

primera parte
del ingenioso
hidalgo don …

(Don Quixote)

Is this enough knowledge of the source language to drive phonetic decipherment?

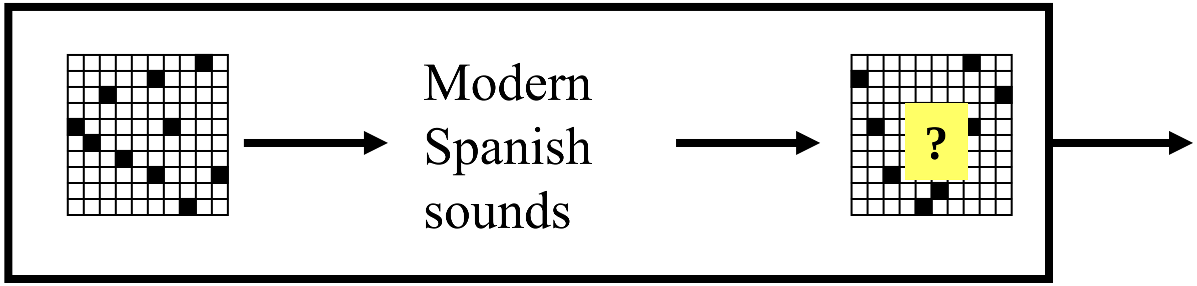
What about silent letters (h) and sounds written with 2 letters (ll)?

Ideal Phonetic Decipherment

sound	letter
B	b or v
D	d
G	g
J	ñ
L	l l or y
a	a or á
b	b or v
d	d
e	e or é
f	f
g	g
i	i or í
l	l
m	m
n	n
o	o or ó
p	p

sound	letter
r	r
t	t
tS	c h
u	u or ú
x	j
nothing	h
T (before a, o, u)	z
T (before e or I)	c or z
T (otherwise)	c
k (before e or I)	q u
k (before s)	x
k (otherwise)	c
rr (start of word)	r
rr (otherwise)	rr
s (after k)	nothing
s (otherwise)	s

Phonetic Decipherment



ciphertext (6980 letters)

`primera parte`
`del ingenioso`
`hidalgo don …`

(Don Quixote)

Decoder maximize P(p) * P(c | p)^3	805 errors
Smooth P(p) with lambdas	684
Use per-symbol lambdas	621
Trigram P(p)	492 (7%)

Correct:	**`primera parte del inxenioso  iDalGo don kixote…`**
Initial:	**`primera parte des intenioso liDasto don fuiLote…`**
Improved:	**`primera parte del inGenioso biDalGo don kixote…`**

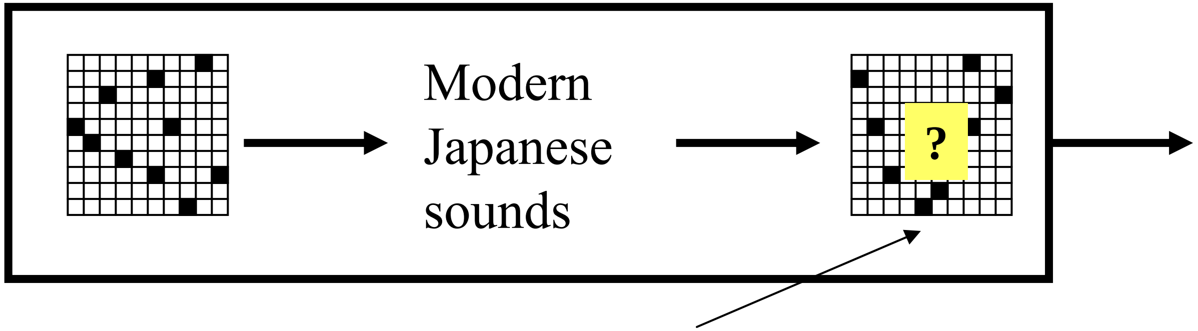
Deciphering Syllabic Writing

ciphertext (200 sentences)

アンジラナイト…

kana writing (roughly one symbol per syllable)

Deciphering Syllabic Writing



ciphertext (200 sentences)

アンジラナイト…

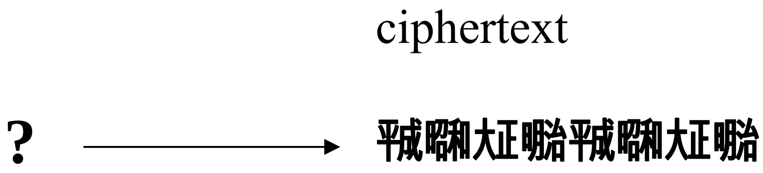
kana writing (roughly one symbol per syllable)

Transducer allows mapping any C, CV, C, or CSV sequence onto any written character.

Results:

Sentences of ciphertext	Phonetic accuracy
200	99.0 %
100	97.5
50	96.2
20	82.2
5	48.5

Deciphering Logographic Writing



ciphertext

? → 平成昭和大正明治平成昭和大正明治

Deciphering Chinese writing is hard.

Baseline (guess “de” for every character) = 3.2% syllable accuracy

Best result = 22% syllable accuracy

How to Decipher Unknown Script if Spoken Language is Also Unknown?

- One idea: build a *universal* model P(s) of human phoneme sequence production

- Human might generally say: K AH N AH R IY
- Human won’t generally say: R T R K L K

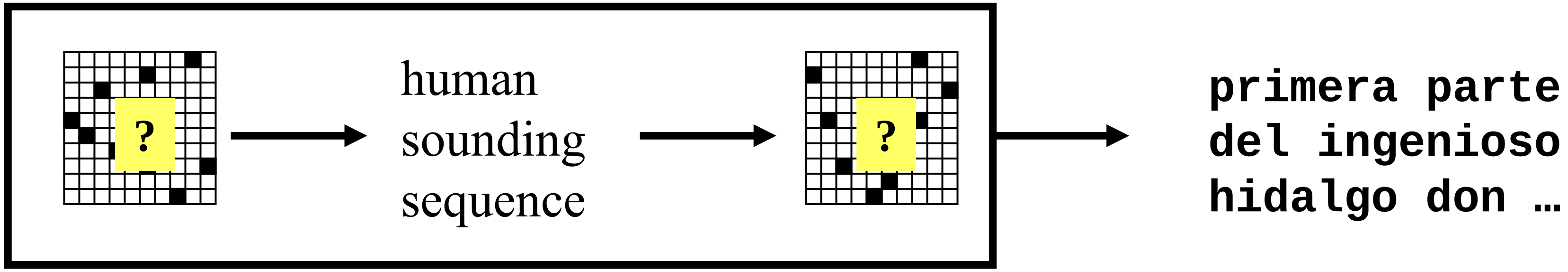
- Deciphering means finding a P(c | p) table such that there is a decoding with a good universal P(p) score

Universal Phonology

- Linguists know lots of stuff!
- Phoneme inventory
  - if z, then s
- Syllable inventory
  - all languages have CV (consonant-vowel) syllables
  - if VCC, then also VC
- Syllable sonority structure
  - {stdbptk} {mnrl} {V} {mnrl} {stdbptk}
  - dram, lomp, tra, ma, ? rdam, ? lopm, ? tba, ? mla
- Physiological preference constraints
  - tomp, tont, tongk, ? tomk, ? tonk, ? tongt, ? tonp

Universal Phonology

Task 1: Label each letter with a phoneme

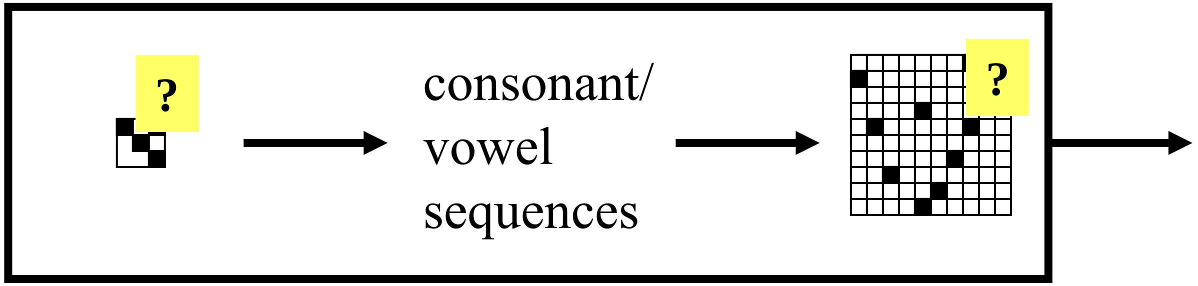


human
sounding
sequence

```
primera parte
del ingenioso
hidalgo don …
```

Universal Phonology

Task 2: Label each letter with a phoneme class: C or V



consonant/
vowel
sequences

`primera parte`
`del ingenioso`
`hidalgo don …`

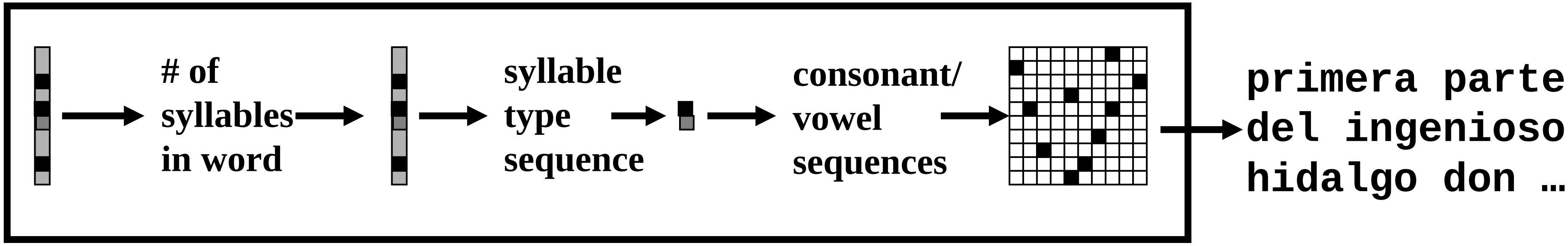
P(C | V C) = ?
P(V | V C) = ?
etc.

P(a | V) = ?
P(a | C) = ?
etc.

Input:	**`primera parte del ingenioso hidalgo don …`**
Output:	**`vvcvcvc vcvvc vcv cvvcvccvc vcvcvvc vcv …`**

Universal Phonology

Task 2: Label each letter with a phoneme class: C or V



# of syllables in word

syllable type sequence

consonant/ vowel sequences

primera parte del ingenioso hidalgo don …

P(1) = ?
P(2) = ?
etc.

Must fix uniform!

P(CV) = ?
P(V) = ?
P(CVC) = ?
+ 7 other types

P(V | V) = ?
P(VV | V) = ?

P(a | V) = ?
P(a | C) = ?
etc.

Input: **primera parte del ingenioso hidalgo don …**
Output: **ccvcvcv cvccv cvc vccvcvvcv cvcvccv cvc …**

P(CV)	= 0.45	P(VC)	= 0.09
P(V)	= 0.15	P(CVC)	= 0.22
P(CCV)	= 0.02	P(CCVC)	= 0.01

P(a | V)	= 0.27	P(a | C)	= 0.00
P(b | V)	= 0.00	P(b | C)	= 0.04
P(c | V)	= 0.00	P(c | C)	= 0.07

Unknown Source Language

- Another idea: brute force
- If we don’t know the spoken language, simply decode against all spoken languages:
  - Pre-collect P(p) for 300 languages
  - Train a P(c | p) using each P(p) in turn
  - See which decoding run assigns highest P(c)
- Hard to get phoneme sequences
- Can use text sequence as a substitute

UN Declaration of Human Rights

300+ words in many of world’s languages, UTF-8 encoding

No one shall be arbitrarily deprived of his property

Niemand se eiendom sal arbitrêr afgeneem word nie

Asnjeri nuk duhet të privohet arbitrarisht nga pasuria e tij

لا يجوز تجريد أحد من ملكه تعسفا

Janiw khitisa utaps oraqeps inaki aparkaspati

Arrazoirik gabe ez zaio inori bere jabegoa kenduko

Den ebet ne vo tennet e berc'hentiezh digantañ diouzh c'hoant

Никой не трябва да бъде произволно лишен от своята собственост

Ningú no serà privat arbitràriament de la seva propietat

任何人的财产不得任意剥夺。

Di a so prupiità ùn ni pò essa privu nimu di modu tirannicu

Nitko ne smije samovoljno biti lišen svoje imovine

Nikdo nesmí být svévolně zbaven svého majetku

Ingen må vilkårligt berøves sin ejendom

Niemand mag willekeurig van zijn eigendom worden beroofd

Nul ne peut être arbitrairement privé de sa propriété

Nimmen mei samar fan syn eigendom berôve wurde

Ninguín será privado arbitrariamente da súa propiedade

Niemand darf willkürlich seines Eigentums beraubt werden

Κανείς δεν μπορεί να στερηθεί αυθαίρετα την ιδιοκτησία του

Avavégui ndojepe'a va'erâi oimeháicha reinte imbáe teéva

Ba wanda za a kwace wa dukiyarsa ba tare da cikakken dalili ba

Senkit sem lehet tulajdonától önkényesen megfosztani

Engan má eftir geðþótta svipta eign sinni

Tak seorang pun boleh dirampas hartanya dengan semena-mena

Necuno essera private arbitrarimente de su proprietate

Ní féidir a mhaoin a bhaint go forlámhach de dhuine ar bith

Al neniu estu arbitre forprenita lia proprieto

Kelleltki ei tohi tema vara meelevaldselt ära võtta

Eingin skal hissini vera fyri ongartøku

Me kua ni dua e kovei vua na nona iyau

Keltään älköön mielivaltaisesti riistettäkö hänen omaisuuttaan

Unknown Source Language

- Input:

  cevzren cnegr qry vatravbfb uvqnytb qba dhvwbgr qr yn znapun …

- Languages with best P(c) after deciphering?

Unknown Source Language

- Input:

```
cevzren cnegr qry vatravbfb uvqnytb qba dhvwbgr qr yn znapun …
```

- Top 5 languages with best P(c) after deciphering:

-5.29120	spanish
-5.43346	galician
-5.44087	portuguese
-5.48023	kurdish
-5.49751	romanian

- Best-path decoding assuming plaintext is Spanish:

```
primera parte del ingenioso hidalgo don quijote de la mancha …
```

- Best-path decoding assuming plaintext is English:

```
wizaris asive bek u-gedundl pubscon bly whualve be ks asequs …
```

- Simultaneous language ID and decipherment

Consonantal Writing

- Input (known to be only consonants):

```
ceze ceg qy ataf uqyt qa dwg q y zapu …
```

- Languages best P(c) after deciphering?

Consonantal Writing

- Input (known to be only consonants):

```
ceze ceg qy ataf uqyt qa dwg q y zapu …
```

- Top 5 languages best P(c) after deciphering:

-2.66979	spanish
-2.67214	chinese
-2.69454	rhaeto-romance
-2.70965	fijian
-2.70979	galician

- Best-path decoding assuming plaintext is Spanish:

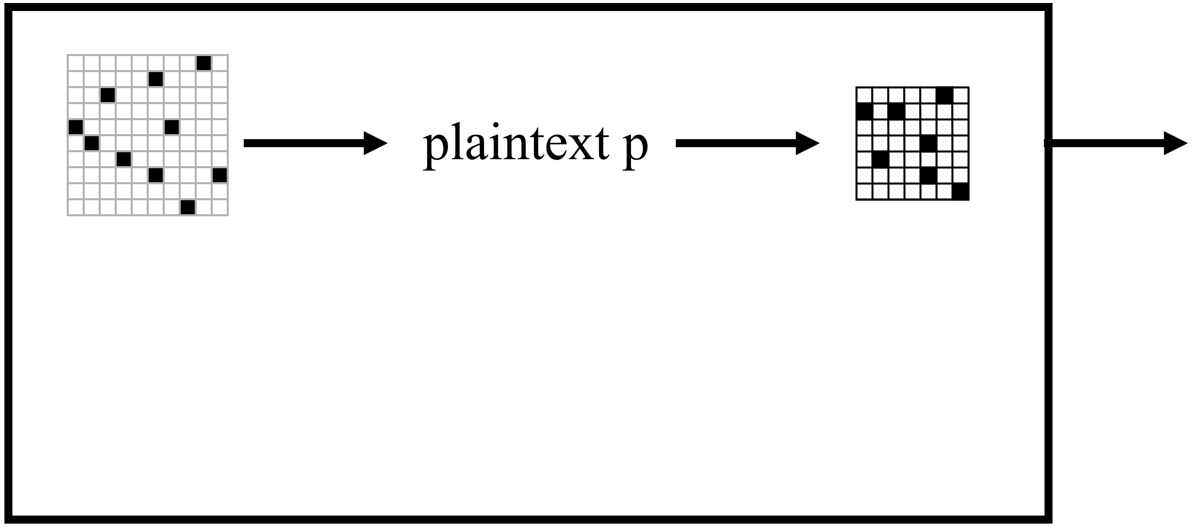
```
prmr prt dl ngns hdlg dn qvt d l mnch …
```

- Best-path decoding assuming plaintext is English:

```
ql-l qlv tn hghd btng th frv n n whmb …
```

Last Experiment: Word Substitution Cipher

“When I look at an article in Arabic, I say to myself, this is really English, but it has been encoded in some strange symbols!!! Let’s decode!!!”



ciphertext (1b words)

رفض رئيس السلطة الفلسطينية محمود عباس مجددا
تصريحات وزير الخارجية الإسرائيلي سيلفان شالوم التي
قال فيها إنه يتعين على إسرائيل إعادة النظر في انسحابها
من غزة، المقرر أن يتم الصيف المقبل إذا فازت حركة
المقاومة الإسلامية حماس في الانتخابات التشريعية.
وقال عباس في مؤتمر صحفي على هامش مشاركته في
القمة العربية-اللاتينية الأولى إنه يتعين على إسرائيل
احترام خيار الشعب الفلسطيني حتى لو فازت حماس
بالانتخابات، وأضاف "إذا نجحت حماس أو فتح سيكون
هذا خيار الشعب الفلسطيني، وعلى الجميع قبول هذا
الخيار بكل ترحاب".
من جانبه شجب رئيس الحكومة الفلسطينية أحمد قريع
الطابع الأحادي الجانب للانسحاب الإسرائيلي من غزة،
وأكد أن إسرائيل تريد مغادرة هذه الأراضي لتعزيز
سيطرتها على الضفة الغربية.
وقال قريع في كلمة له خلال مؤتمر نظمته وزارة
الأوقاف في رام الله "سينسحبون من غزة ولكننا لا
نعرف ما هو شكل هذا الانسحاب وماذا سيتركون، وما
هو مصير المعابر والحدود, وكل ذلك غامض لأنه قرار
أحادي الجان

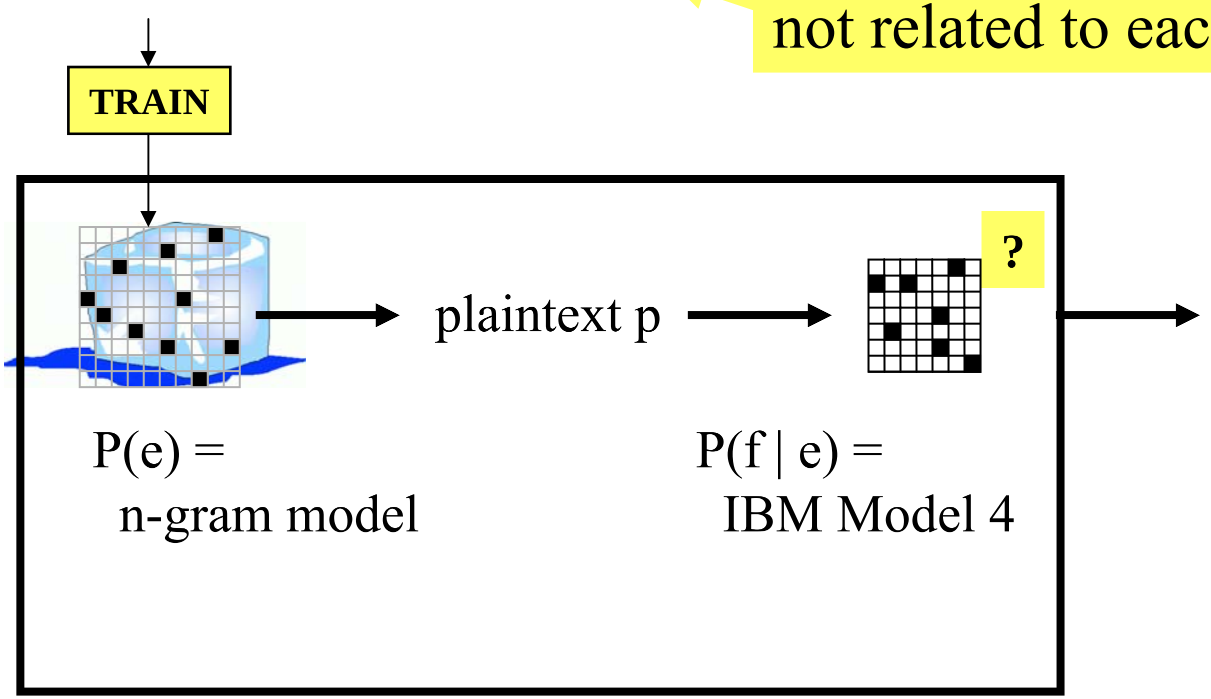
Last Experiment: Word Substitution Cipher

BAGHDAD, Iraq (CNN) -- Six bombings killed at least 54 Iraqis and wounded 96 others Wednesday, including 20 civilians who died as they lined up to join the Iraqi army in Hawija when a suicide bomber detonated explosives hidden under his clothing, Iraqi officials said.

That attack in the town about 130 miles (209 kilometers) north of Baghdad also wounded 30 Iraqis, said Iraqi army Lt. Col. Khalil al-Zawbai.

A car bombing in Saddam Hussein's ancestral homeland of Tikrit also killed 30 Iraqis and wounded another 40, Iraqi officials said. The Tikrit explosion…

Key Point: These texts are not related to each other.



TRAIN

plaintext p

?

P(e) =
n-gram model

P(f | e) =
IBM Model 4

ciphertext (1b words)

رفض رئيس السلطة الفلسطينية محمود عباس مجددا
تصريحات وزير الخارجية الإسرائيلي سيلفان شالوم التي
قال فيها إنه يتعين على إسرائيل إعادة النظر في انسحابها
من غزة، المقرر أن يتم الصيف المقبل إذا فازت حركة
المقاومة الإسلامية حماس في الانتخابات التشريعية.

وقال عباس في مؤتمر صحفي على هامش مشاركته في
القمة العربية-اللاتينية الأولى إنه يتعين على إسرائيل
احترام خيار الشعب الفلسطيني حتى لو فازت حماس
بالانتخابات، وأضاف "إذا نجحت حماس أو فتح سيكون
هذا خيار الشعب الفلسطيني، وعلى الجميع قبول هذا
الخيار بكل ترحاب".

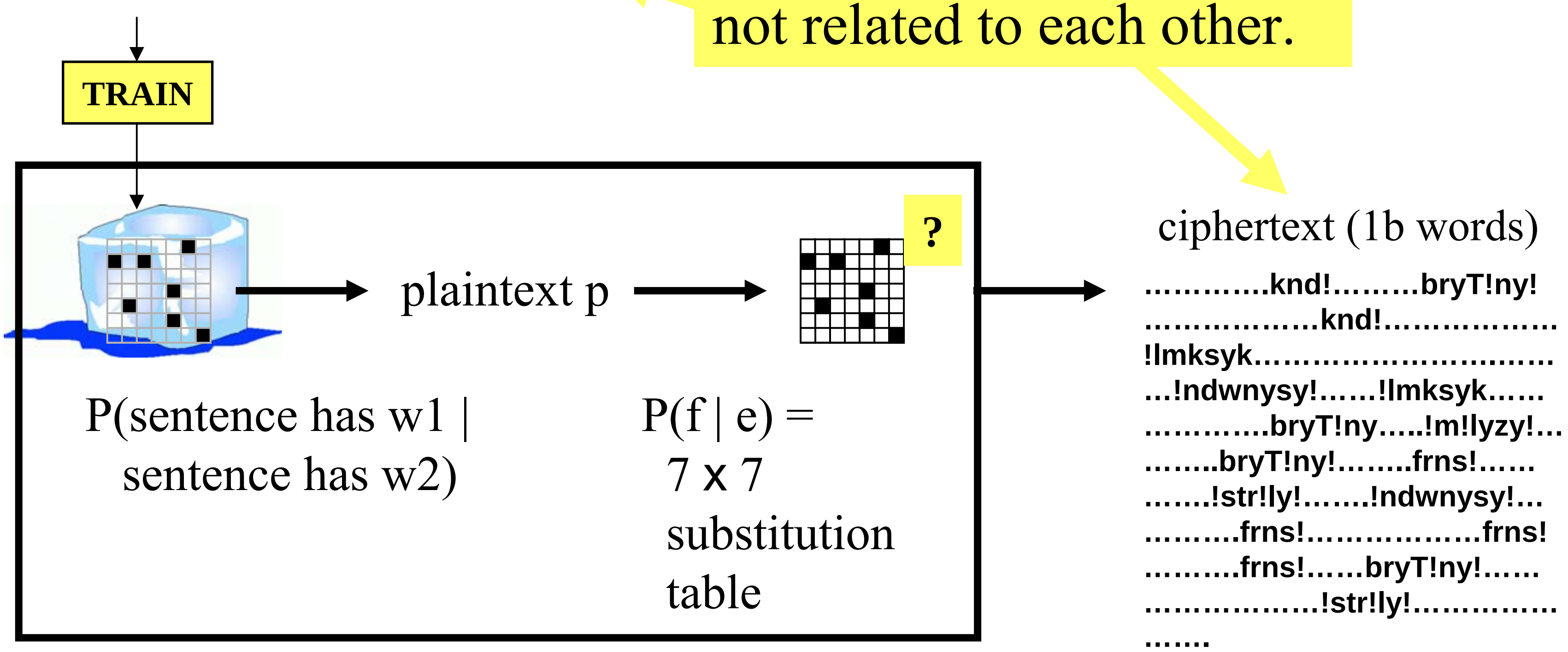
من جانبه شجب رئيس الحكومة الفلسطينية أحمد قريع
الطابع الأحادي الجانب للانسحاب الإسرائيلي من غزة،
وأكد أن إسرائيل تريد مغادرة هذه الأراضي لتعزيز
سيطرتها على الضفة الغربية.

وقال قريع في كلمة له خلال مؤتمر نظمته وزارة
الأوقاف في رام الله "سينسحبون من غزة ولكننا لا
نعرف ما هو شكل هذا الانسحاب وماذا سيتركون، وما
هو مصير المعابر والحدود, وكل ذلك غامض لأنه قرار
أحادي الجان

Word Substitution Cipher

…………France………Britain………………Canada…
…………Mexico………………Indonesia……Malaysia…
……………Britain……….Canada……….Australia…
…..Britain…………France……………Indonesia………
….Mexico………………Australia……………France…
…Britain……………………………………….

Key Point: These texts are not related to each other.



TRAIN

plaintext p

?

P(sentence has w1 | sentence has w2)

P(f | e) = 7 x 7 substitution table

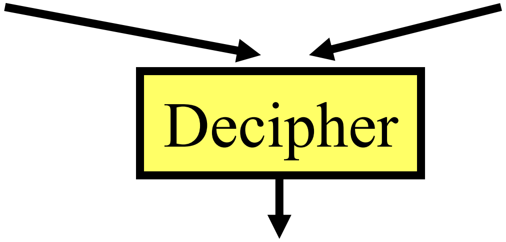
ciphertext (1b words)

…………..knd!………bryT!ny!
………………knd!………………
!lmksyk…………………………
…!ndwnysy!……!lmksyk……
…………bryT!ny…..!m!lyzy!…
……..bryT!ny!……..frns!……
…….!str!ly!…….!ndwnysy!…
……….frns!………………frns!
……….frns!……bryT!ny!……
………………!str!ly!……………
…….

Word Substitution Cipher

…………France………Britain………………Canada…
…………Mexico………………Indonesia……Malaysia…
……………Britain……….Canada……….Australia…
…..Britain…………France……………Indonesia………
….Mexico………………Australia……………France…
…Britain……………………………………….

…………knd!………bryT!ny!………………knd!…………
……!lmksyk……………………………!ndwnysy!……!lm
ksyk………………bryT!ny…..!m!lyzy!……….bryT!ny!
……..frns!…………!str!ly!…….!ndwnysy!…………frns
!………………frns!……….frns!……bryT!ny!……………
………!str!ly!………………….



Decipher

Fails: Every English word learns same mapping. Local minimum.

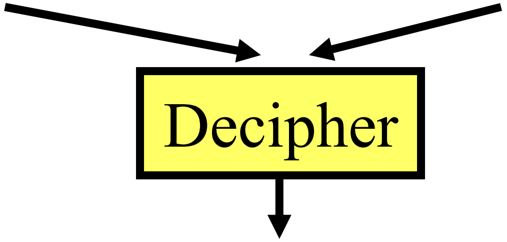
Pick random starting points for EM

# of random starts	Accuracy of learned table
1	57%
5	71%
40 or more	100%

Word Substitution Cipher

…………France………Britain……………..Canada…
…………Mexico……………..Indonesia……Malaysia…
……………Britain……….Canada……….Australia…
…..Britain…………France……………Indonesia………
….Mexico……………..Australia……………France…
…Britain……………………………………….

…………..knd!………bryT!ny!……………….knd!…………
……!lmksyk……………………………..!ndwnysy!……!lm
ksyk………………..bryT!ny…..!m!lyzy!……….bryT!ny!
……..frns!…………..!str!ly!…….!ndwnysy!…………..frns
!………………frns!……….frns!……bryT!ny!……………
………!str!ly!………………….



Decipher

Australia →	!str!ly! (0.93)	!ndwnysy! (0.03)	m!lyzy! (0.02)
Britain →	bryT!ny! (0.98)	!ndwnysy! (0.01)	!str!ly! (0.01)
Canada →	knd! (0.57)	frns! (0.33)	m!lyzy! (0.06)
France →	frns! (1.00)		
Indonesia →	!ndwnysy! (1.00)		
Malaysia →	m!lyzy! (0.93)	lmksyk (0.07)	
Mexico →	!lmksyk (0.91)	m!lyzy! (0.07)	

Summary of Results

English letter substitution cipher	100%
Hindi character code conversion	54%
Phonetic decipherment - alphabetic Spanish writing - syllabic Japanese writing	93-99%
Spanish CV assignment & syllable structure	100%
Simultaneous language ID and decipherment - alphabetic writing - consonantal writing	100%
Word substitution cipher - 7 words English - 7 words Arabic	100%

Summary of Suggested Techniques

- #0 It never works the first time.
- #1 Cube learned substitution probabilities before decoding.
- #2 Use well-smoothed plaintext model.
- #3 Use fixed uniform probabilities for non-central parameters.
- #4 Appeal to linguistic universals to constrain models.
- #5 Bootstrap bigger models from smaller ones to constrain models.
- #6 Use random restarts to avoid local minima.

Future Work

- Other decipherment problems
- Better results
- Will a computer make discoveries in linguistics?
  - it has happened in astronomy…
  - and chemistry…
- Archaeology, animal languages, …
  - where supervised training is not an option…

end