Natural Language Processing Research

USC CS Annual Research Review
March 23, 2010

Kevin Knight

Research Associate Professor
Computer Science Department
USC

Senior Research Scientist and Fellow
Information Sciences Institute (ISI)
USC

What is Natural Language Processing?

- Challenges:
  - Understand what people say!
  - Generate utterances that make sense!

- Valuable applications:
  - Language Translation
  - Dictation
  - Question answering
  - Dialogue
  - Summarization
  - Creative language generation

Natural Language Research at USC

Computer Science Faculty

- Institute for Creative Technologies (ICT)
- Information Sciences Institute (ISI)
- Electrical Engineering Department (EE)

M. Arbib
D. Chiang
A. Gordon
J. Gratch
J. Hobbs
E. Hovy
K. Knight
A. Leuski
D. Marcu
S. Marsella
S. Narayanan
P. Pantel
F. Sha
W. Swartout
D. Traum

Natural Language Research at USC

Computer Science Faculty

M. Arbib
D. Chiang
A. Gordon
J. Gratch
J. Hobbs
E. Hovy
K. Knight
A. Leuski
D. Marcu
S. Marsella
S. Narayanan
P. Pantel
F. Sha
W. Swartout
D. Traum

Text and gesture (**Stacy Marsella**)

Origins of language (**Michael Arbib**)

New research faculty in 2010 (**Liang Huang** and **Kenji Sagae**)

Natural Language Research at USC

Computer Science Faculty

M. Arbib
D. Chiang
A. Gordon
J. Gratch
J. Hobbs
E. Hovy
K. Knight
A. Leuski
D. Marcu
S. Marsella
S. Narayanan
P. Pantel
F. Sha
W. Swartout
D. Traum

See **Andrew Gordon**'s text mining research in last week's Economist magazine.

Natural Language Research at USC

Computer Science Faculty

M. Arbib
D. Chiang
A. Gordon
J. Gratch
J. Hobbs
E. Hovy
K. Knight
A. Leuski
D. Marcu
S. Marsella
S. Narayanan
P. Pantel
F. Sha
W. Swartout
D. Traum

Shri Narayanan won a 2009 Best Paper Award for "Toward Detecting Emotions in Spoken Dialogs."

David Chiang won a 2009 Best Paper Award for "11,001 New Features for Statistical Machine Translation."

Natural Language Research at USC

Computer Science Faculty

M. Arbib
D. Chiang
A. Gordon
J. Gratch
J. Hobbs
E. Hovy
K. Knight
A. Leuski
D. Marcu
S. Marsella
S. Narayanan
P. Pantel
F. Sha
W. Swartout
D. Traum

Daniel Marcu co-founded an language translation software company that employs 110 people.

Jerry Hobbs was President of the Association for Computational Linguistics (ACL).

Eduard Hovy was also President of the ACL.

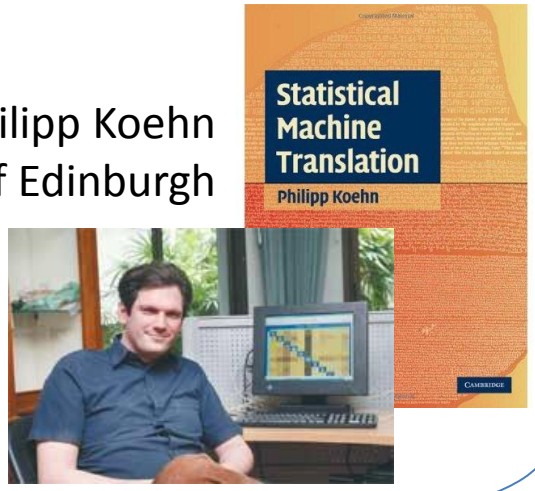
Even a **boy from Louisiana** can be President of ACL.

Computer Science Facu

M. Arbib
D. Chiang
A. Gordon
J. Gratch
J. Hobbs
E. Hovy
K. Knight
A. Leuski
D. Marcu
S. Marsella
S. Narayanan
P. Pantel
F. Sha
W. Swartout
D. Traum

PhD Alumni

Prof. Philipp Koehn
University of Edinburgh



Computer Science Faculty

M. Arbib
D. Chiang
A. Gordon
J. Gratch
J. Hobbs
E. Hovy
K. Knight
A. Leuski
D. Marcu
S. Marsella
S. Narayanan
P. Pantel
F. Sha
W. Swartout
D. Traum

Institute for Creative Technologies (ICT)

- Virtual humans
- Spoken dialogue
  - Understanding
  - Generation
  - Dialogue tracking
  - Emotion

Electrical Engineering (EE)

Prof. Shri Narayanan lab

- Virtual Sick Call
- Enriched **Speech Translation** Systems
- User-Centric Mixed-Initiative Spoken **Dialog Systems**

DIALOGUE

- Modeling **Emotive Improvisation** in Theater Performance
- Virtual Human Museum Guides
- Exploring Emotional Vocal Productions with **MRI**

EMOTION

- Human-Robot Interaction and **Socially Assistive Robotics**
- Social Communication Training in Children with **Autism**
- Early Assessment of Academic Standards

EDUCATION

- Human-Like Speech Processing
- Dynamics of **Vocal Tract Shaping**
- Prosody and **Articulatory Dynamics** in Spoken Language

DYNAMICS

Information Sciences Institute (ISI)

- Automatic Language Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

Chinese/ English

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Kowane mutum na da hakkin ya sami yancin yin tunani da na sanin yakamata da na bin addini; saboda haka yana da yancin sake addini ko ra'ayin da ya bada gaskiya gare shi, da kuma yancin nuna addininsa ko ra'ayinsa, shi daya ko a cikin taro kuma a fili ko a boye ta hanyar koyarwa ko yin ibada, ko bauta wa abin da ya bada gaskiya gare shi da yin abubuwan da abin da yake bauta wa din ya nuna masa.

Hausa/ English

Everyone has the right to freedom of thought, conscience and religion; this right includes freedom to change his religion or belief, and freedom, either alone or in community with others and in public or private, to manifest his religion or belief in teaching, practice, worship and observance.

Why People Get into Automatic Language Translation

- Passion about understanding how human language works
  - What makes one sequence of words grammatical, and another not?
- Interest in languages
  - What's the difference between English and Chinese?
- Desire to change the world
  - How will the world be different when the language barrier disappears?

Why It's Hard

- Each word has tons of meanings
  - I'll **get** a cup of coffee → ?
  - I didn't **get** that joke → ?
  - I **get** up at 8am → ?
  - I **get** nervous → ?
  - Yeah, I **get** around … → ?
- Each word has zillions of contexts
- Word order is different

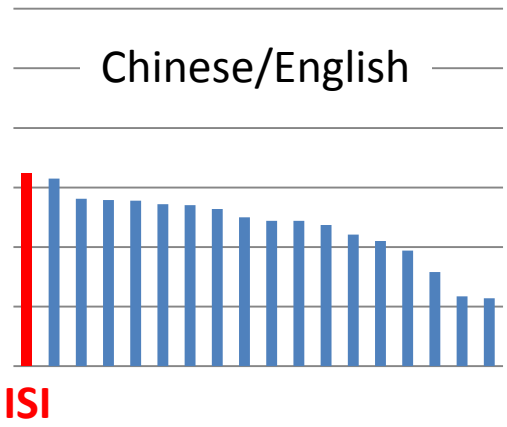
Why It's Hard

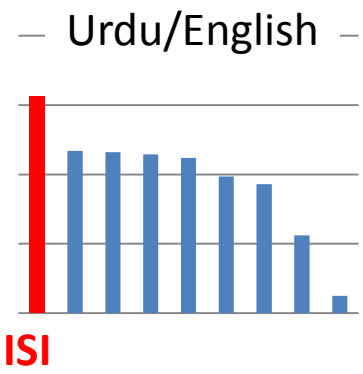
- Output must be a grammatical, sensible, never-before-uttered sentence!
- Computers **consume** lots of human language
  - Google, Yahoo, Bing …
  - Speech recognizers …
- More challenging to also **produce** human language
  - What makes one sequence of words grammatical, and another not?

Information Sciences Institute (ISI)

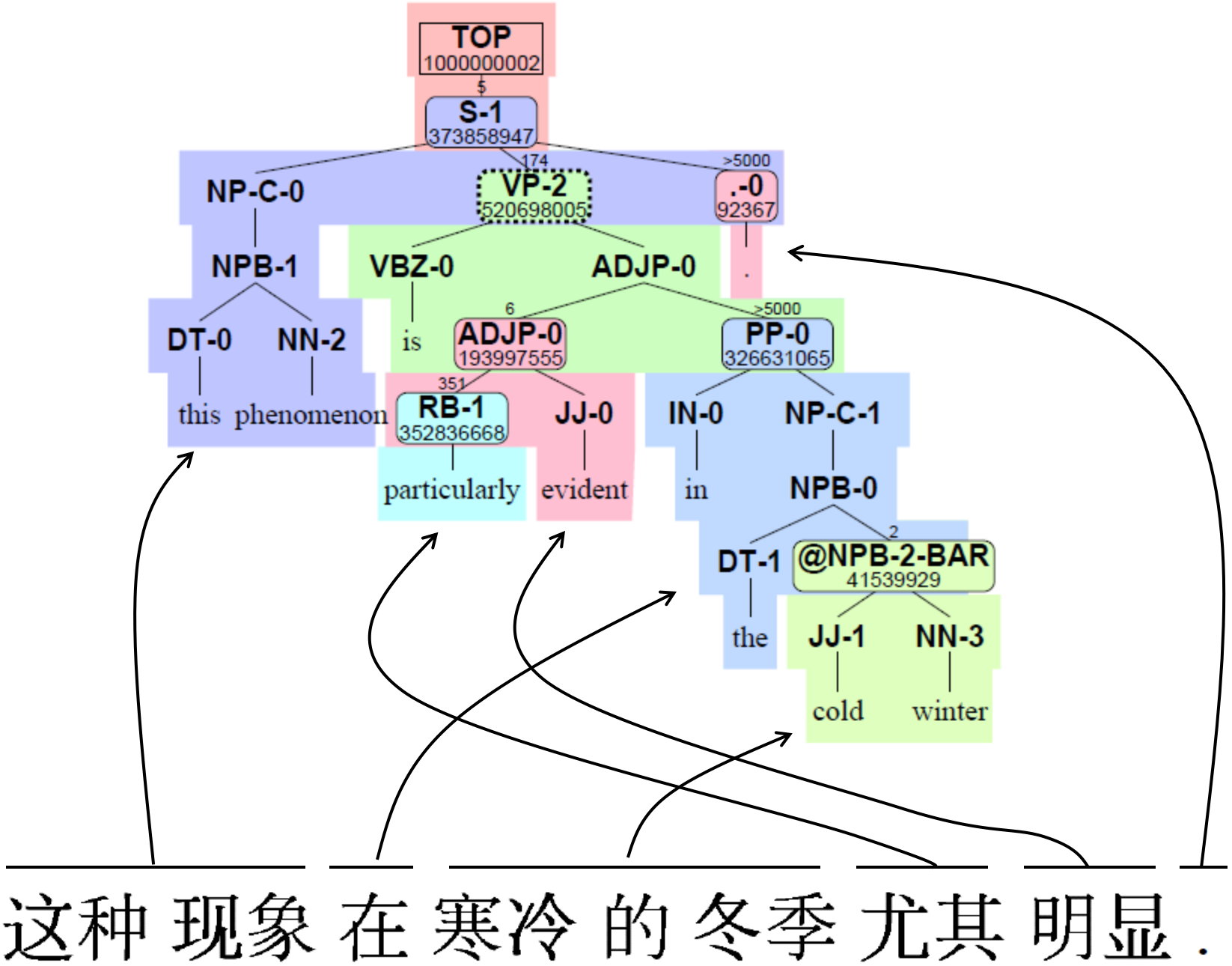
- Automatic Language Translation
  - Grand challenge machine learning problem
  - Billions of words of human translations to learn from
  - Marry language structure and text statistics

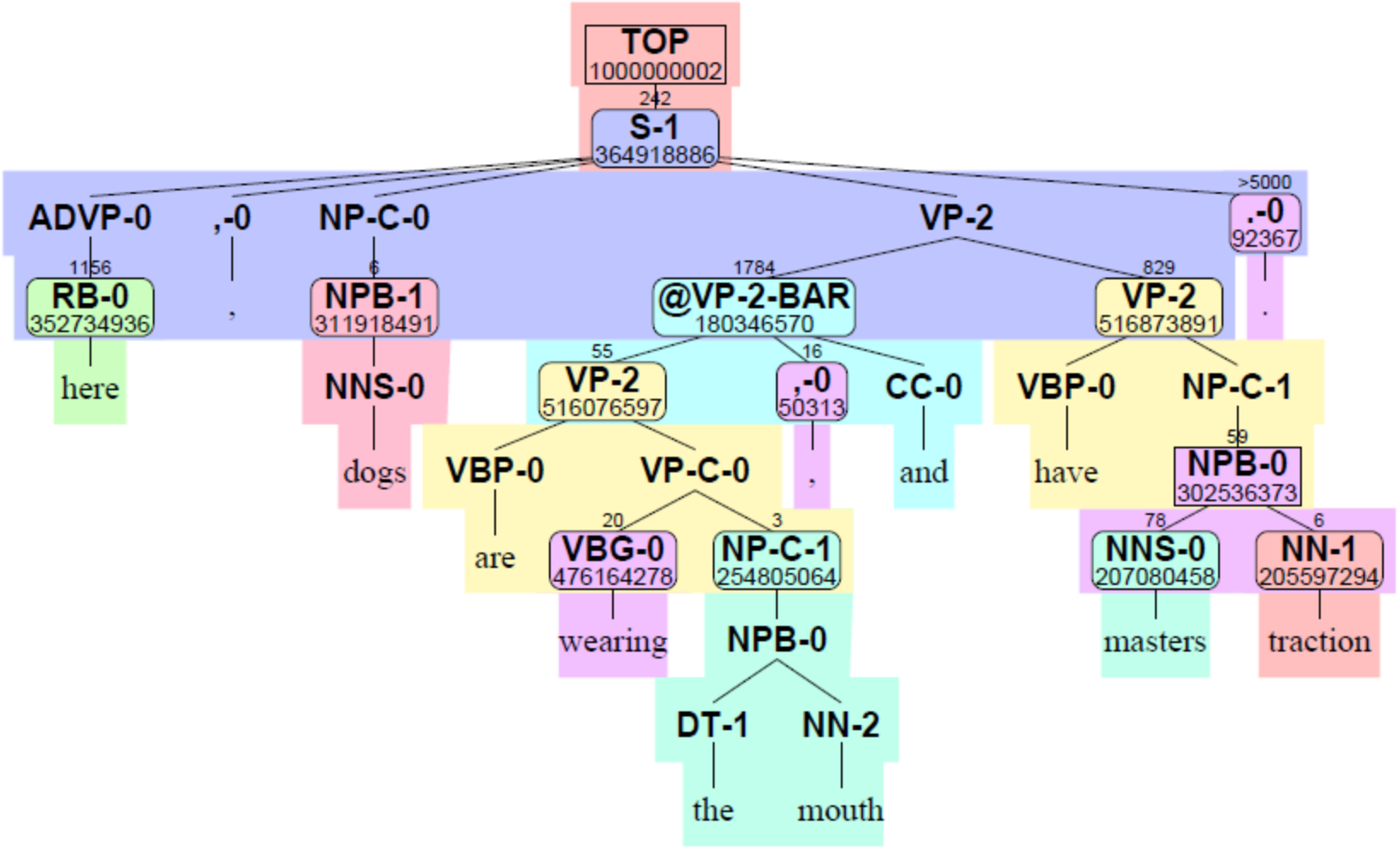
Annual international bake-off (2009)





这种现象在寒冷的冬季尤其明显.

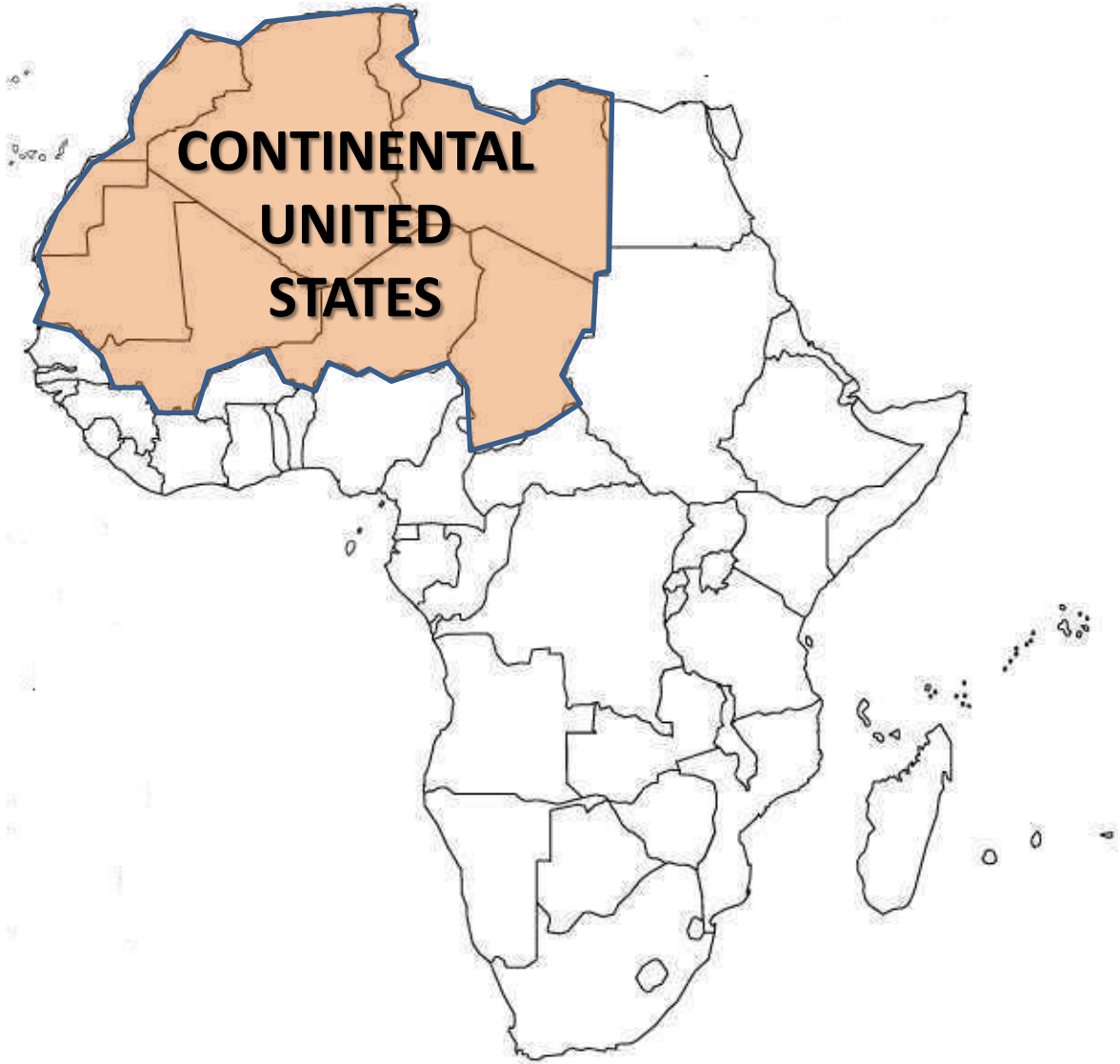

TOP
1000000002
5
S-1
373858947
174
>5000
NP-C-0
VP-2
520698005
.-0
92367
NPB-1
VBZ-0
ADJP-0
.
6
>5000
DT-0
NN-2
is
ADJP-0
193997555
PP-0
326631065
351
this
phenomenon
RB-1
352836668
JJ-0
IN-0
NP-C-1
particularly
evident
in
NPB-0
2
DT-1
@NPB-2-BAR
41539929
the
JJ-1
NN-3
cold
winter
这种 现象 在 寒冷 的 冬季 尤其 明显 .



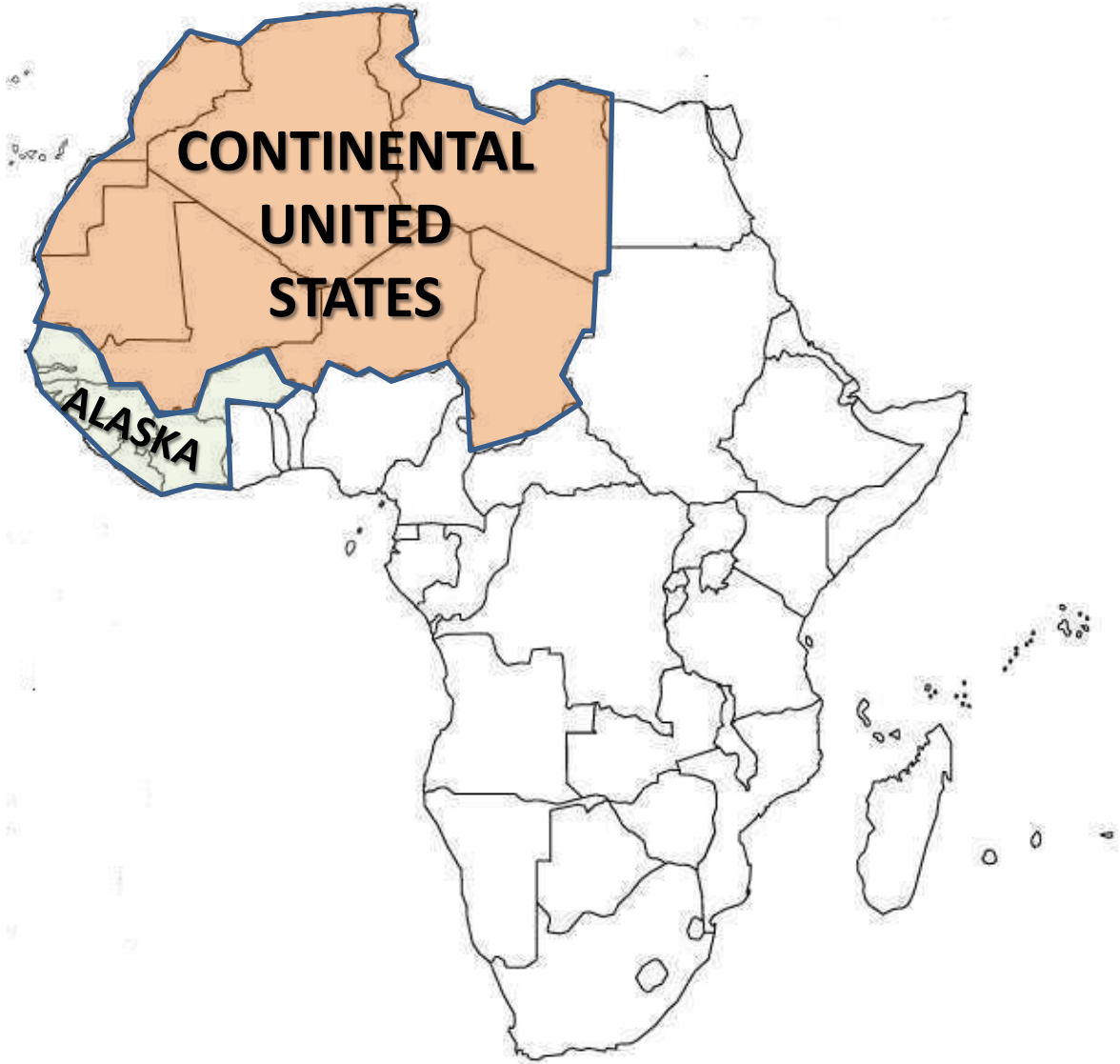
在这里，狗都配戴嘴套，并有主人牵引.

African Languages

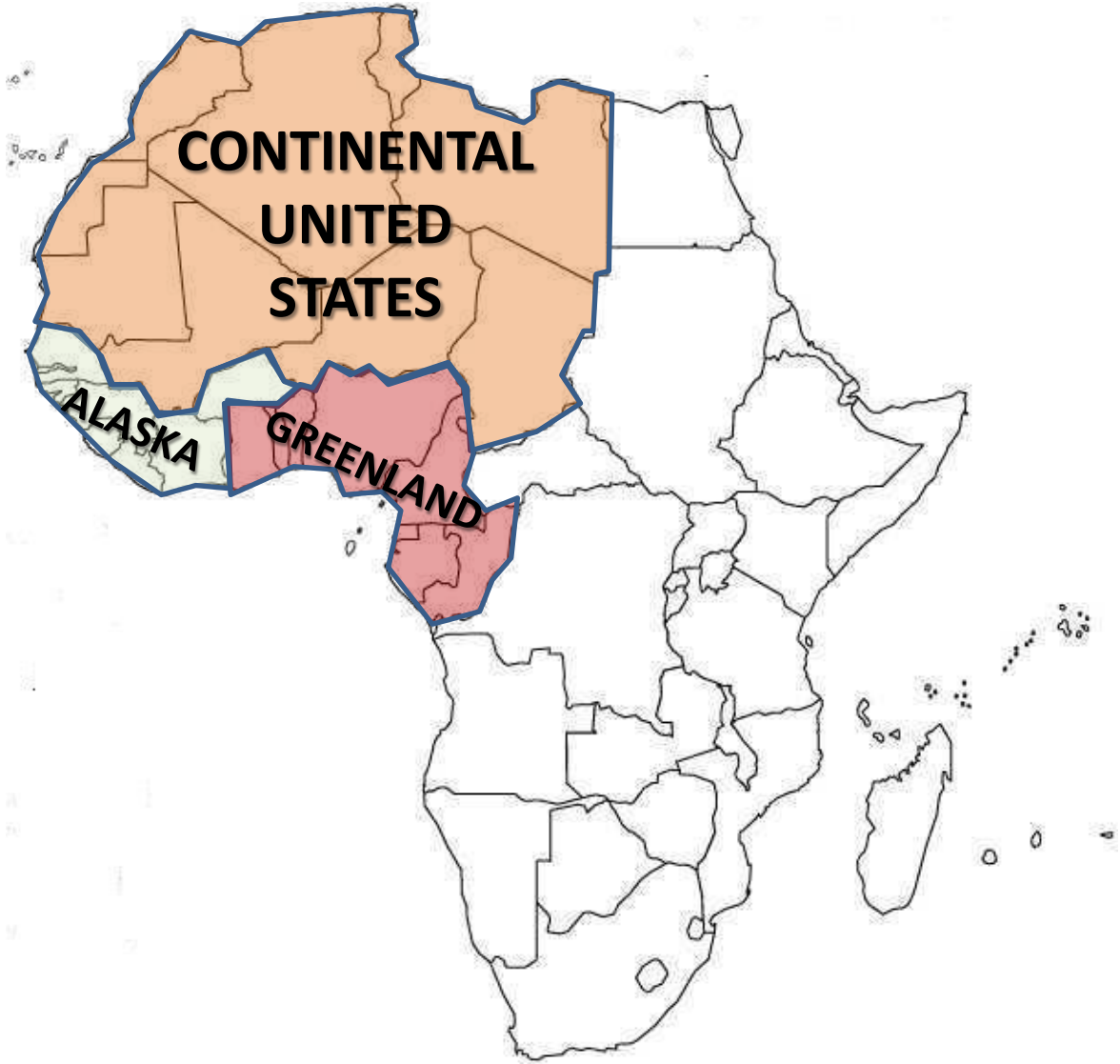
African Languages



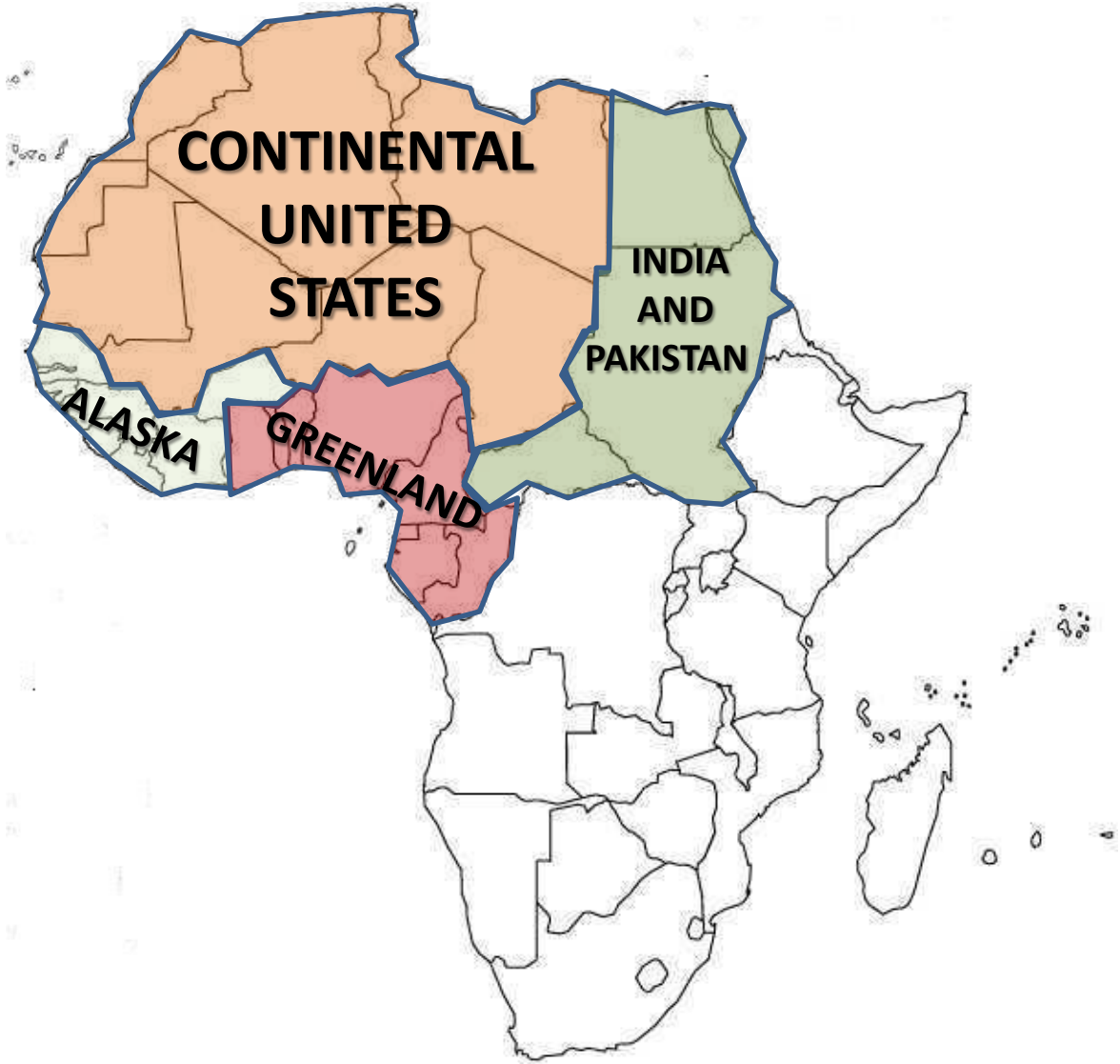
African Languages



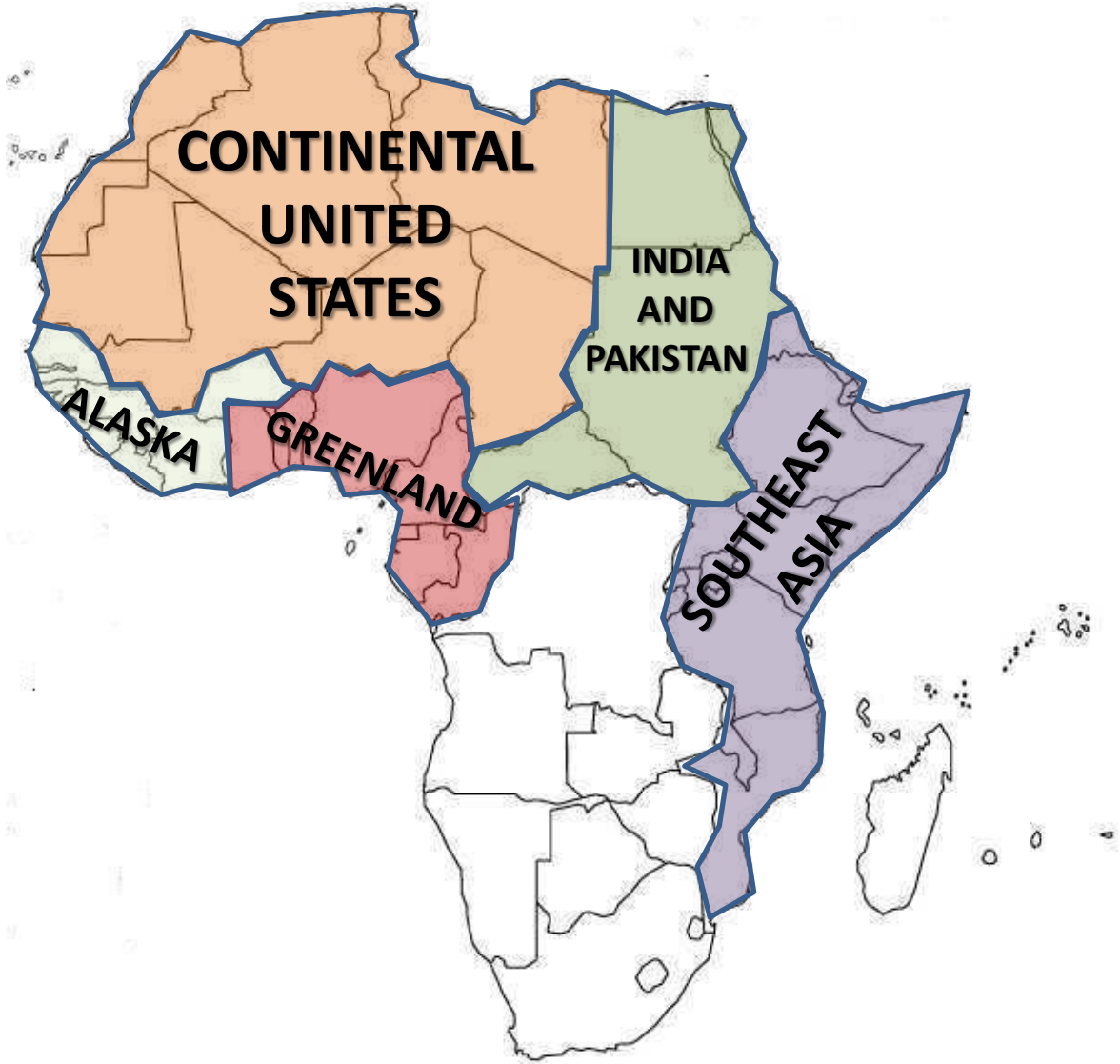
African Languages



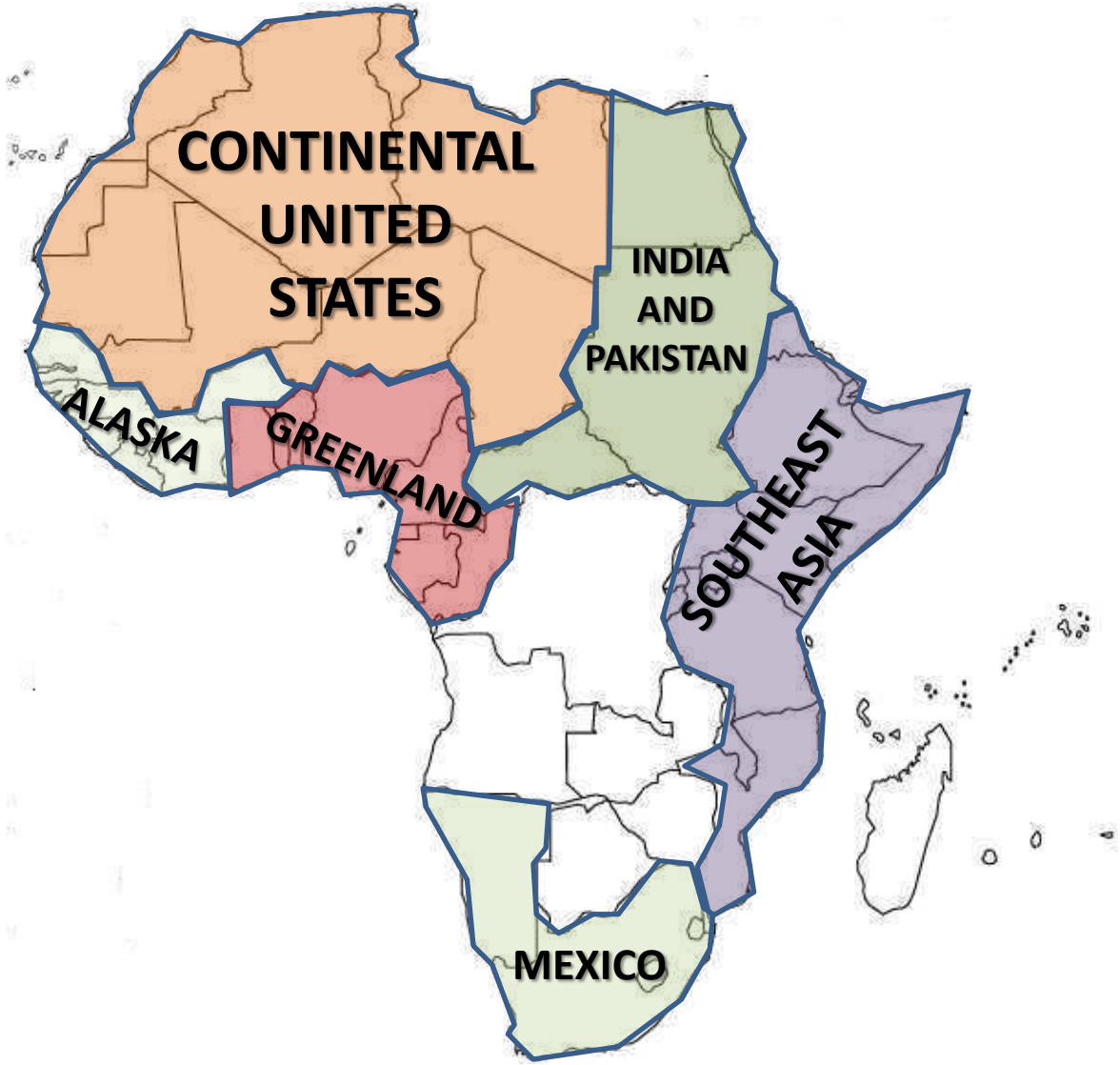
African Languages



African Languages



African Languages

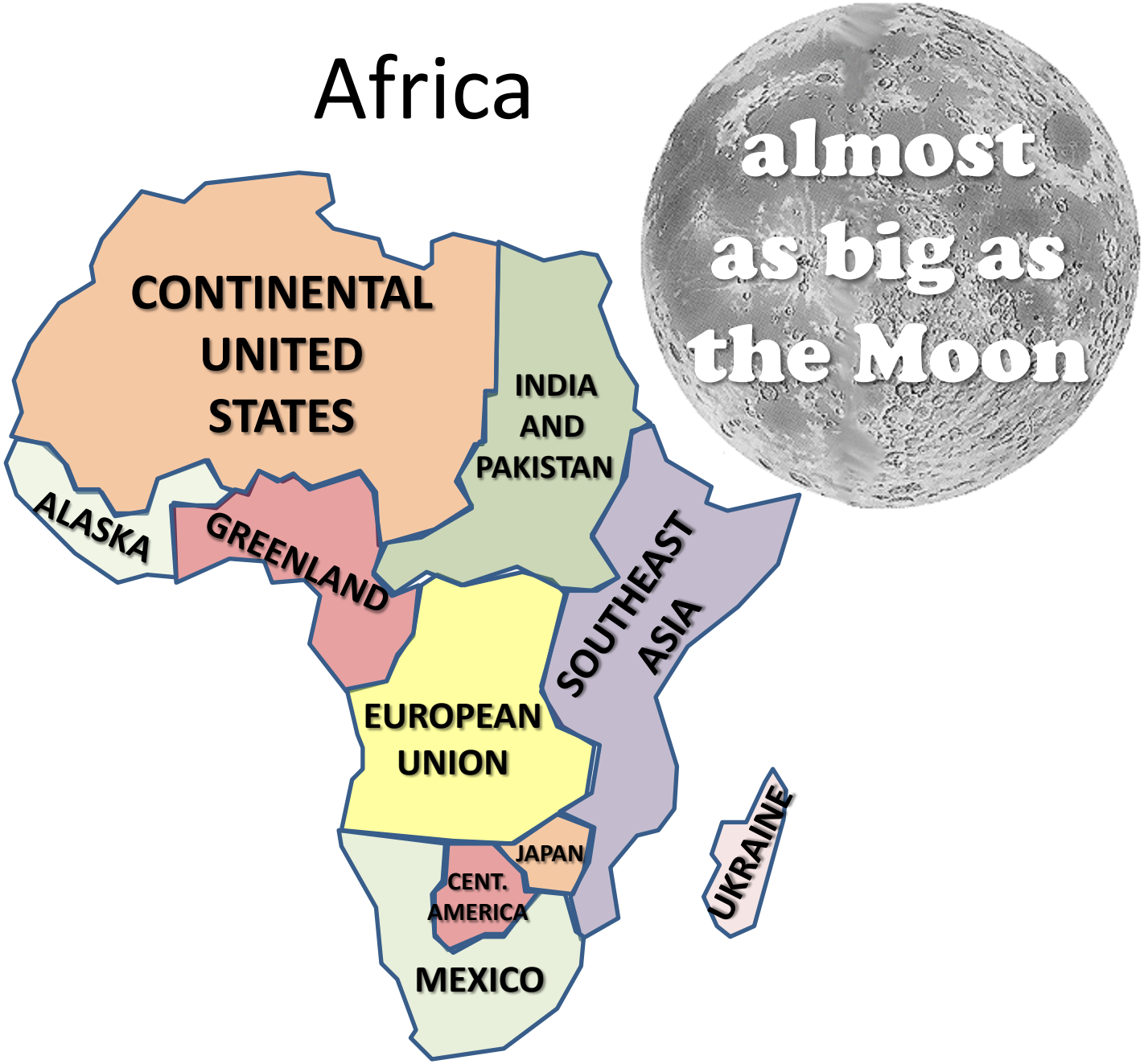


African Languages

African Languages

African Languages

Africa


almost as big as the Moon
CONTINENTAL UNITED STATES
INDIA AND PAKISTAN
ALASKA
GREENLAND
SOUTHEAST ASIA
EUROPEAN UNION
UKRAINE
JAPAN
CENT. AMERICA
MEXICO

Africa

Zero languages spoken

Africa

Zero languages spoken

1000+ languages spoken, 40+ by 1m+ speakers

Information Sciences Institute (ISI)

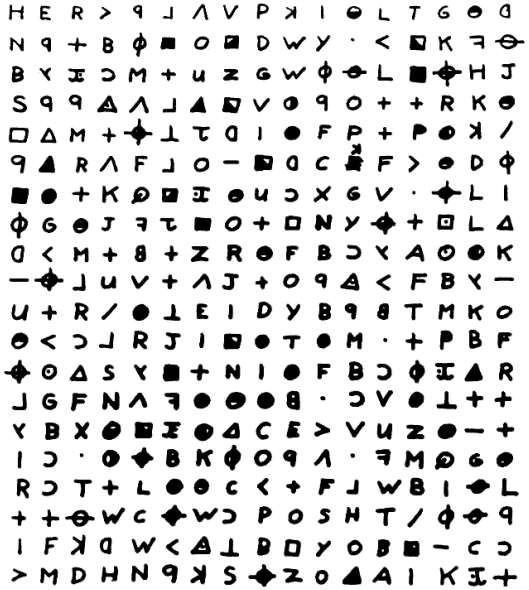
- Learning by Reading
  - Can a machine learn about a topic by reading articles?
- Biomedical Text Extraction
  - How to turn a scientific article into a formal representation of the experiment?
- Phylogeny of Languages
- Automatic Decipherment

Decipherment

Zodiac 408-letter cipher
(solved in 1969)

Ugaritic tablets
(ISI with MIT)

Voynich Manuscript, 240pp,
(undeciphered)

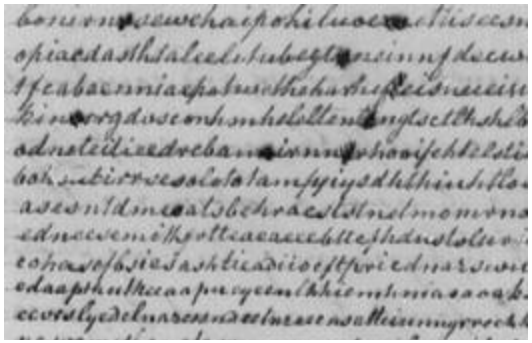


Zodiac 340-letter cipher
(undeciphered)

عربي ودولي

الحكومة العراقية تلغي ترخيص

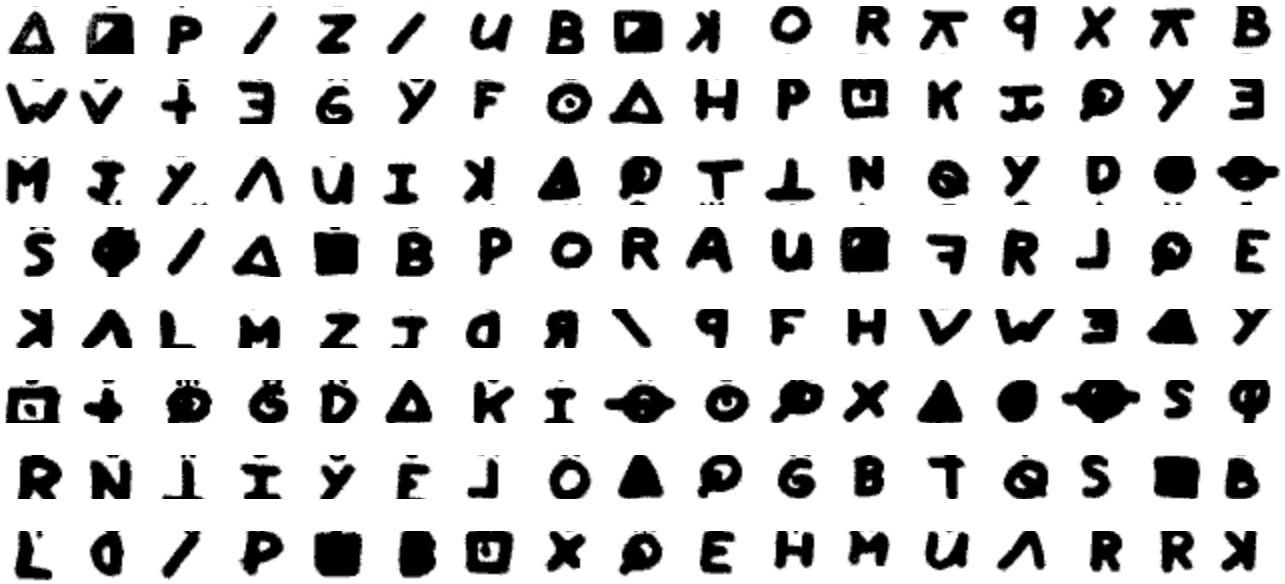
Foreign language as a code
for English (Weaver 1949)



Thomas Jefferson cipher
(solved, see WSJ 7/6/09)

Zodiac Serial Killer

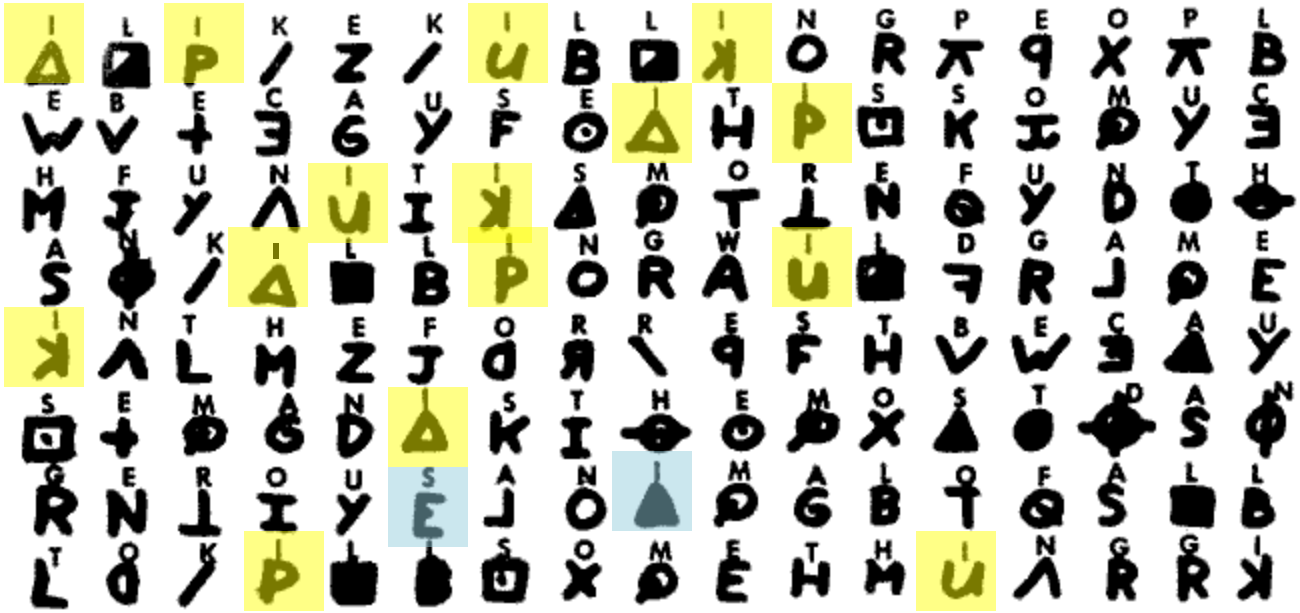
408-letter cipher, solved by husband and wife over breakfast:



(plus two more sections)

Zodiac Serial Killer

408-letter cipher, solved by husband and wife over breakfast:



(plus two more sections)

Homophonic Cipher

Plain	Cipher
A	1 3 G l S
B	V
C	e
D	@ f
E	+ E I N P W Z
F	J Q
G	R
H	= M
I	2 k P U
J	
K	/
L	4 7 B
M	q
N	^ $ D O
O	; d T X
P	:
Q	
R	! \ r
S	1 3 6 F K
T	* H I L N
U	Y
V	]
W	A 2
X	t
Y	5
Z	

‘We felt that the word “KILL” or “KILLING” would appear in his code, and the word “I”, because he had an ego.’ -- Donald Harden

Probabilistic Approach

$$P(\text{cipher}) = \sum_{\text{english}} P(\text{english}) * P(\text{cipher} \mid \text{english})$$

Program assigns a score
to any sequence of English letters.
High score = good English!
We have to supply this.

Probabilistic substitution table
used by Zodiac killer.
If plaintext letter is “E”, probability
of enciphering as “q” is 0.043.
**Hidden! Learning program sets this
to maximize P(cipher).**

Probabilistic Approach

$$P(\text{cipher}) = \sum_{\text{english}} P(\text{english}) * P(\text{cipher} \mid \text{english})$$

Program assigns a score
to any sequence of English letters.
High score = good English!
We have to supply this.

Probabilistic substitution table
used by Zodiac killer.
If plaintext letter is "E", probability
of enciphering as "q" is 0.043.
Hidden! Learning program sets this to maximize P(cipher).

Does maximizing P(cipher) yield more accurate substitution tables?

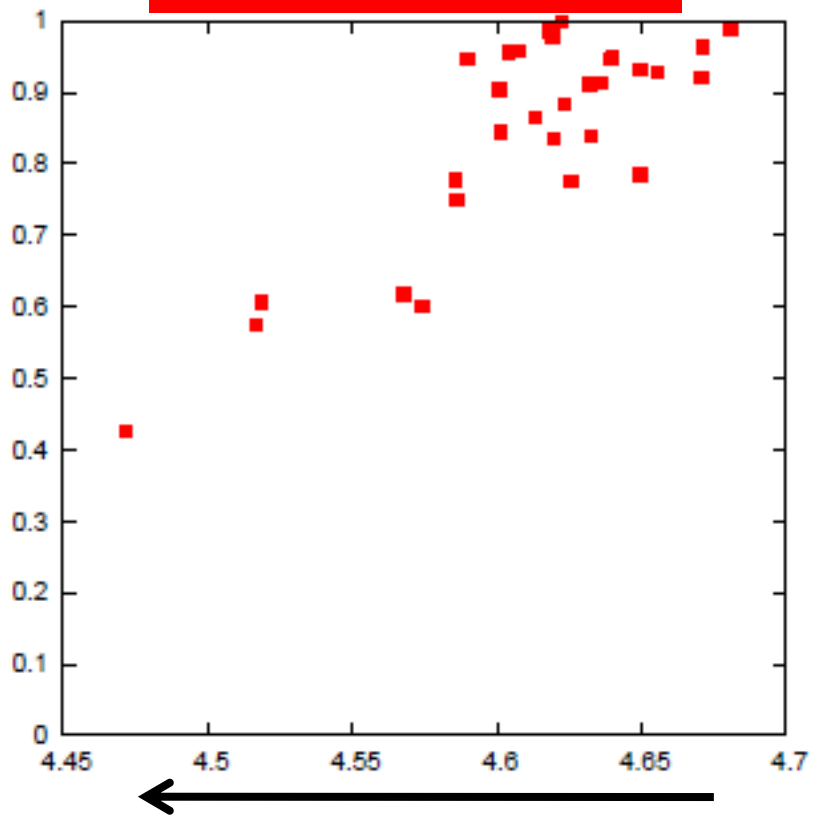
How to efficiently search for the table that maximizes P(cipher)?

How much English do you have to know to solve a 408-letter homophonic cipher?

Does maximizing P(cipher) yield more accurate substitution tables?

Each square is a different substitution table

Fortuitous result:
Minimize Decipherment (Task) Error

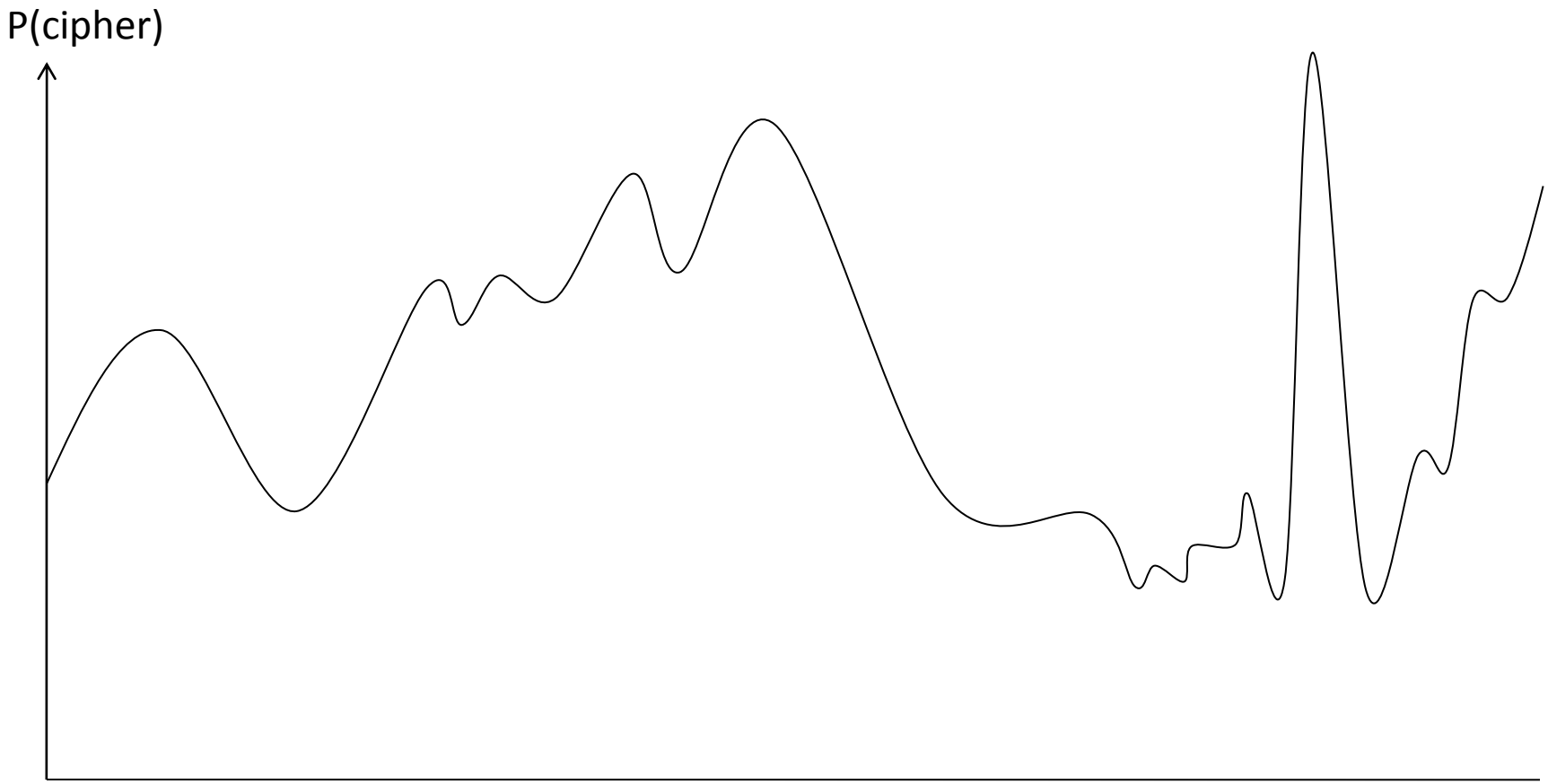


What computer does:
Maximize P(cipher)

all
hail
maximum
likelihood

How to efficiently search for the table that maximizes P(cipher)?

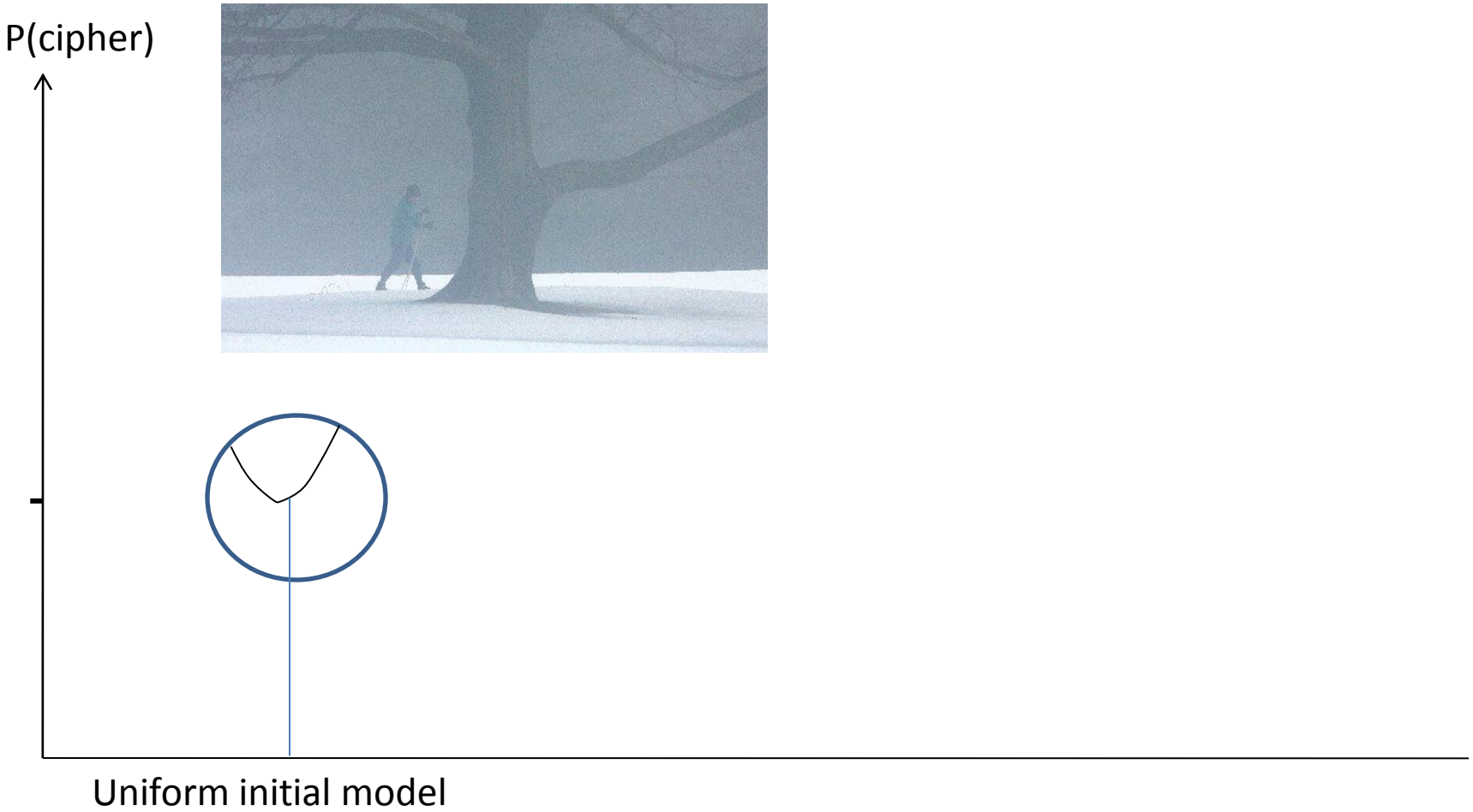
Search Space



All substitution tables
(fancifully laid out in one dimension instead of 26x64)

How to efficiently search for the table that maximizes P(cipher)?

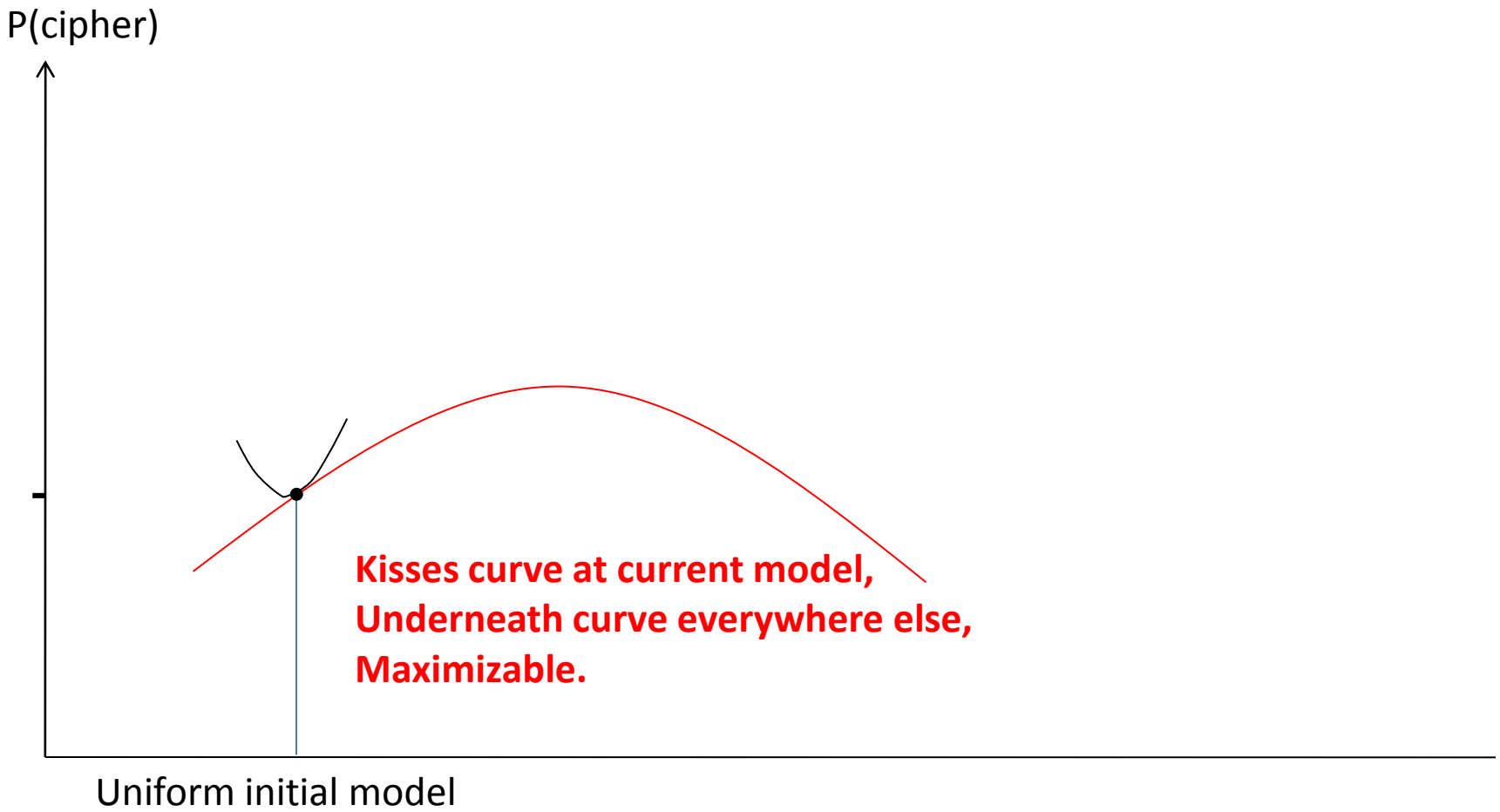
Search Space



All substitution tables
(fancifully laid out in one dimension instead of 26x64)

How to efficiently search for the table that maximizes P(cipher)?

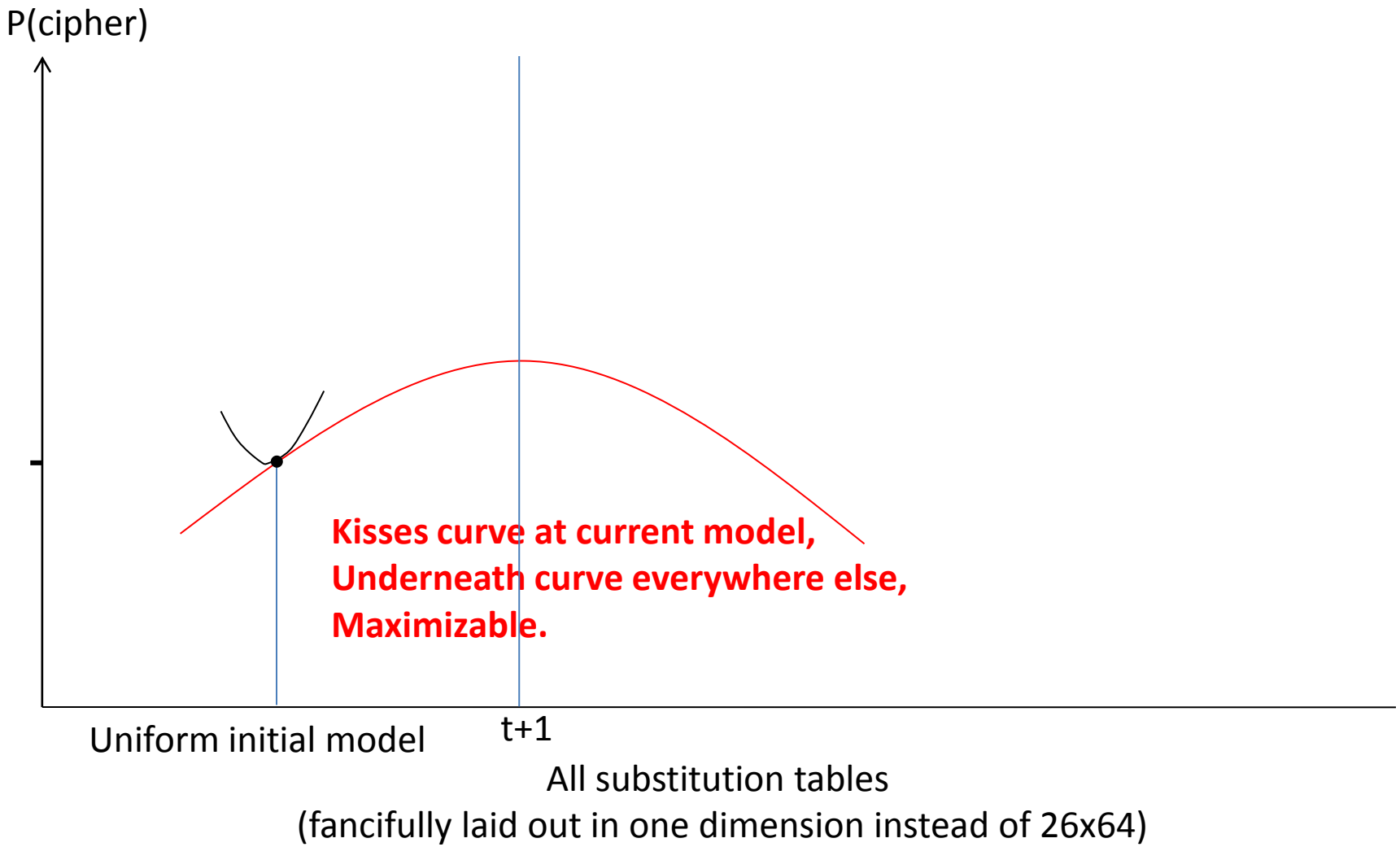
EM Algorithm



All substitution tables
(fancifully laid out in one dimension instead of 26x64)

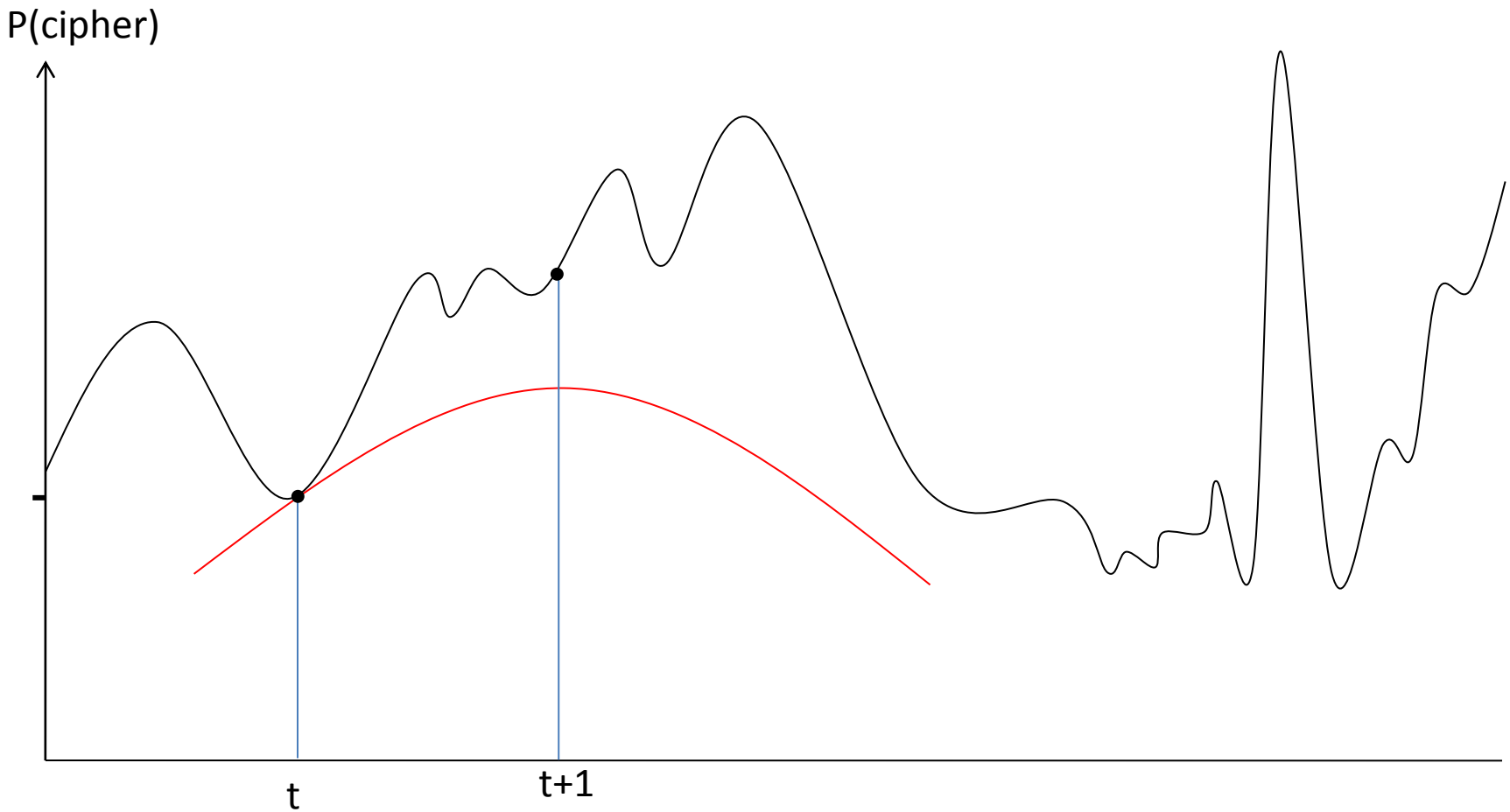
How to efficiently search for the table that maximizes P(cipher)?

EM Algorithm



How to efficiently search for the table that maximizes P(cipher)?

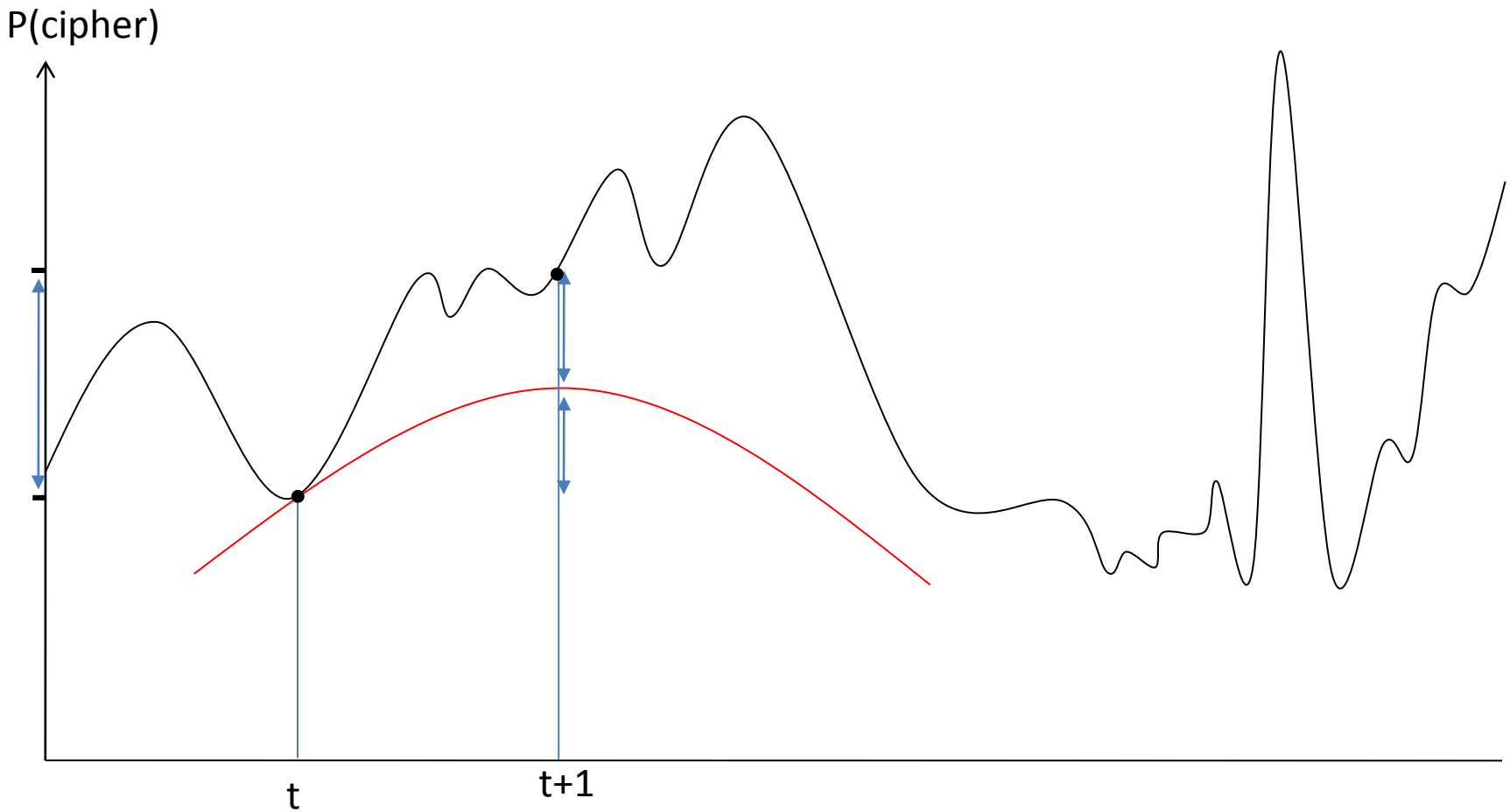
EM Algorithm



All substitution tables
(fancifully laid out in one dimension instead of 26x64)

How to efficiently search for the table that maximizes P(cipher)?

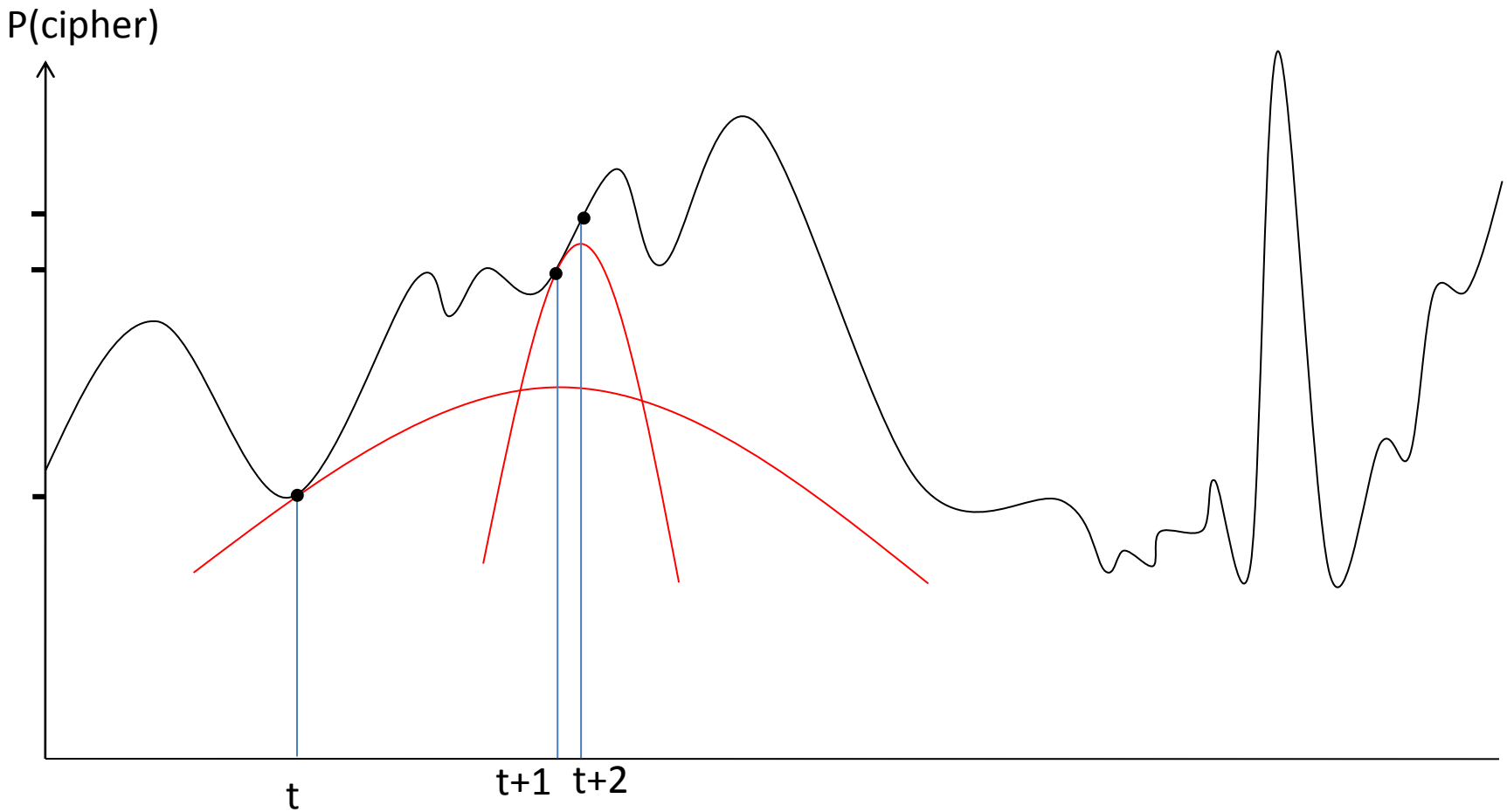
EM Algorithm



All substitution tables
(fancifully laid out in one dimension instead of 26x64)

How to efficiently search for the table that maximizes P(cipher)?

EM Algorithm



All substitution tables
(fancifully laid out in one dimension instead of 26x64)

How much English do you have to know to solve a 408-letter homophonic cipher?

Experiments on Zodiac 408

English knowledge	EM starts	Crib?	Decipherment error
3-gram letter statistics P(s | t t)	1	-	99.7 – everything wrong!
	100	-	71.2
	100	KILLING	24.6
1-gram word statistics P(pittsburgh)	1	-	98.4
	20	-	65.6
	20	ILIKEKILLINGPEOPLE	4.1 – almost perfect

Secret weapons:

- EM
- Bayesian reasoning
- Finite-state transducer toolkits
- Modeling

Natural Language Research at USC

Computer Science Faculty

- Institute for Creative Technologies (ICT)
- Information Sciences Institute (ISI)
- Electrical Engineering Department (EE)

M. Arbib
D. Chiang
A. Gordon
J. Gratch
J. Hobbs
E. Hovy
K. Knight
A. Leuski
D. Marcu
S. Marsella
S. Narayanan
P. Pantel
F. Sha
W. Swartout
D. Traum

thanks