Error Correction through Language Processing

Anxiao (Andrew) Jiang, Yue Li, Jehoshua Bruck

Texas A&M University
and
California Institute of Technology

May 1, 2015

Another title for the talk ...

Is There A New Way To Correct Errors

Two fundamental approaches for correcting errors:

- Add external redundancy
- Use internal redundancy

- APPROACH I: Add External Redundancy.

  ECCs are already approaching channel capacity.

  That is Wonderful!

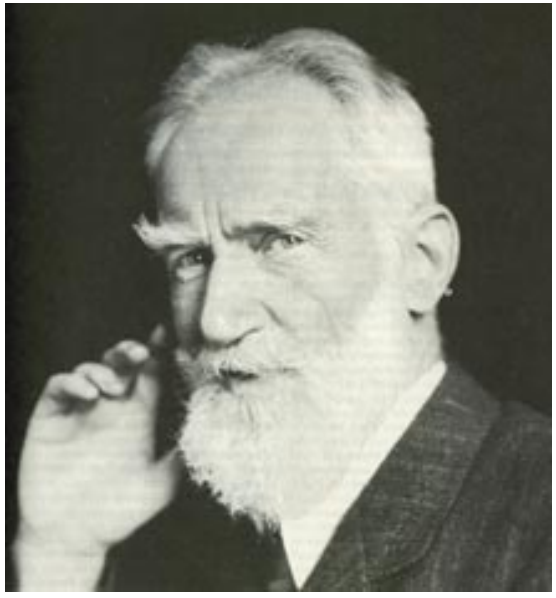
- APPROACH I: Add External Redundancy.

  ECCs are already approaching channel capacity.

  That is Wonderful!

No No No...

"There are two tragedies in life:
One is dream lost,
The other is dream realized."



George Bernard Shaw

- It is time to consider APPROACH II: Use internal redundancy.

Open Problem:

- How to use the redundancy inside data to correct errors,

Open Problem:

- How to use the redundancy inside data to correct errors,and combine it with ECCs?

Open Problem:

- How to use the redundancy inside data to correct errors,and combine it with ECCs?

Suggestion: Focus on "intelligent information", such as languages, images, videos...

Reason: Such data have plenty of redundancy, even after data compression.

It is related to: Joint Source-Channel Coding

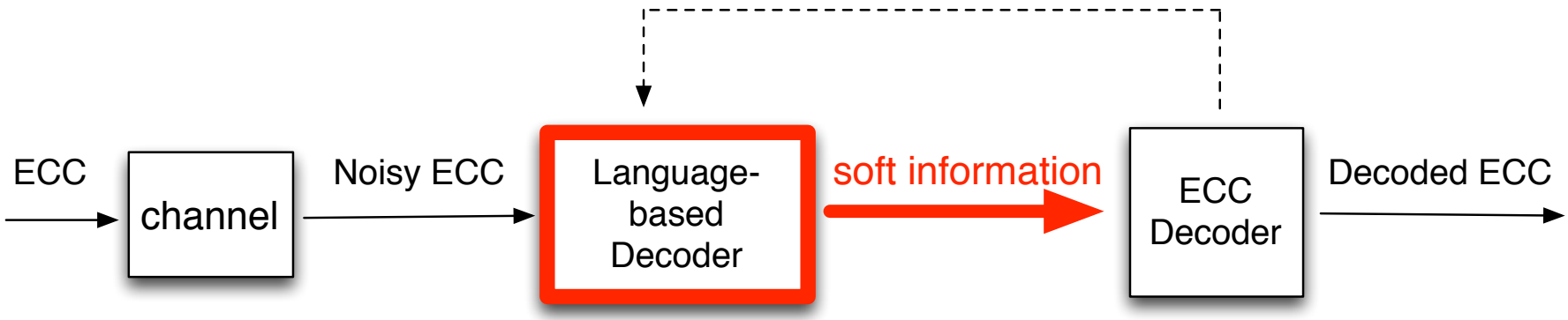
Distinct Features:

- No joint work: Focus on error correction
- No change to ECC data: Compatible with existing systems

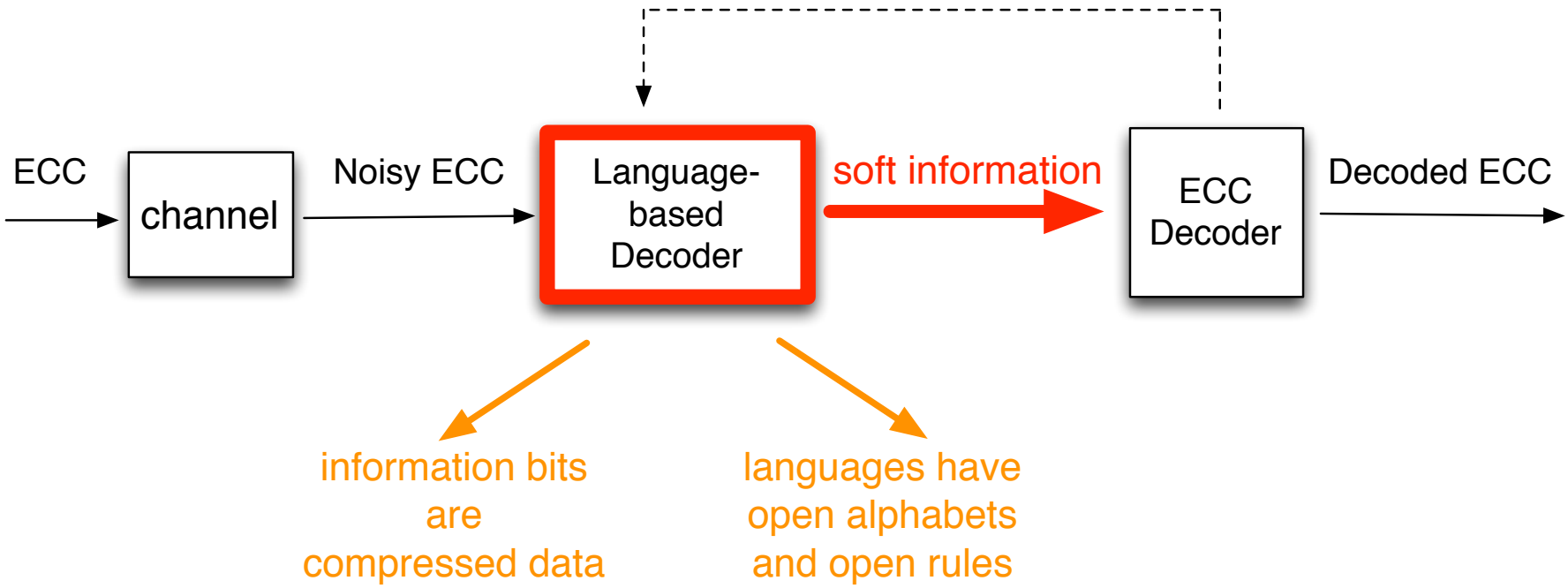
It is related to: Denoising

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J. Weinberger,
Universal Discrete Denoising: Known Channel,
in *IEEE Trans. Information Theory*, vol. 51, no. 1, pp. 5–28, January 2005. (IEEE Communications Society and Information Theory Society Joint Paper Award, 2006.)

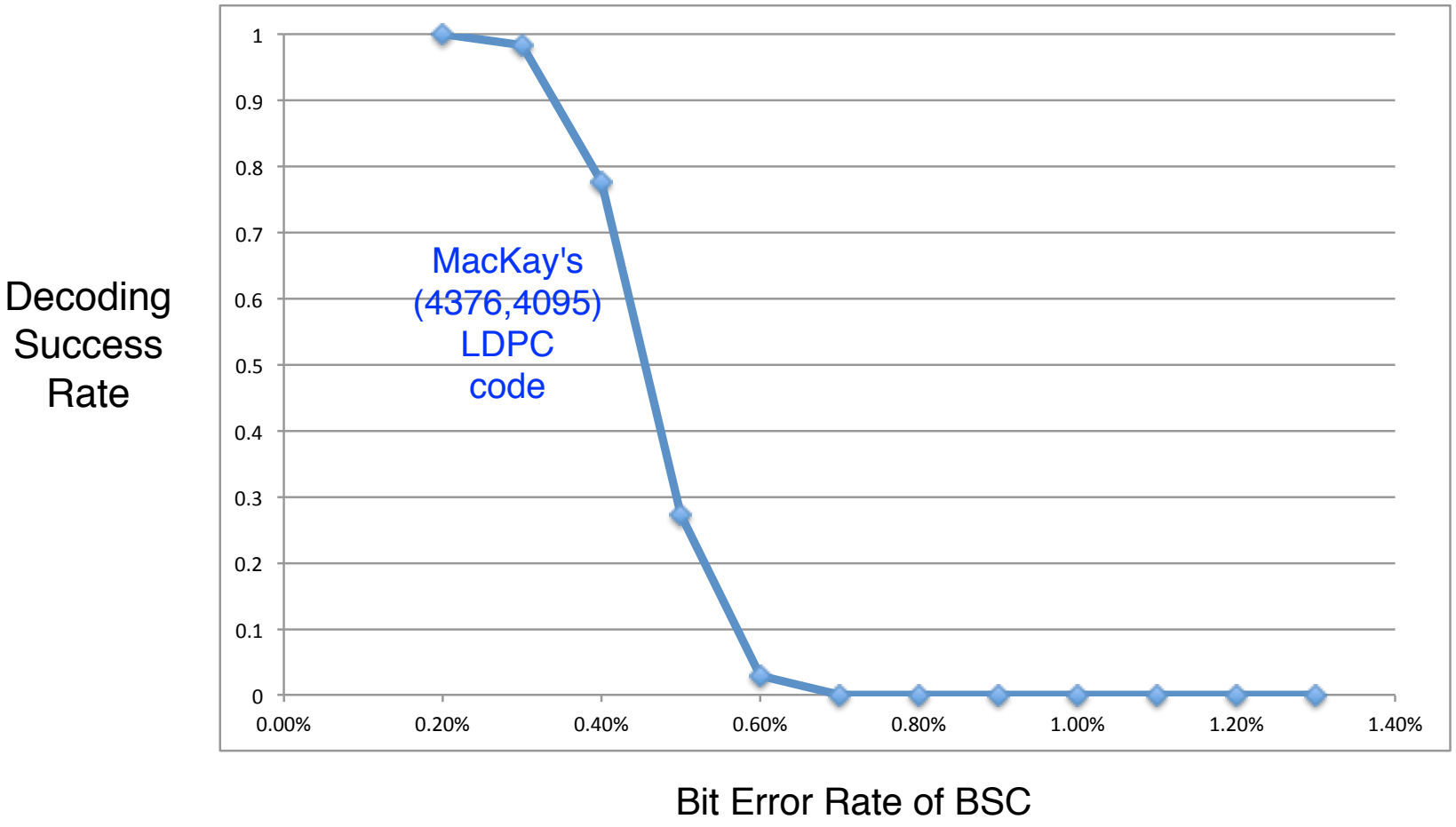
Error-Correction Framework



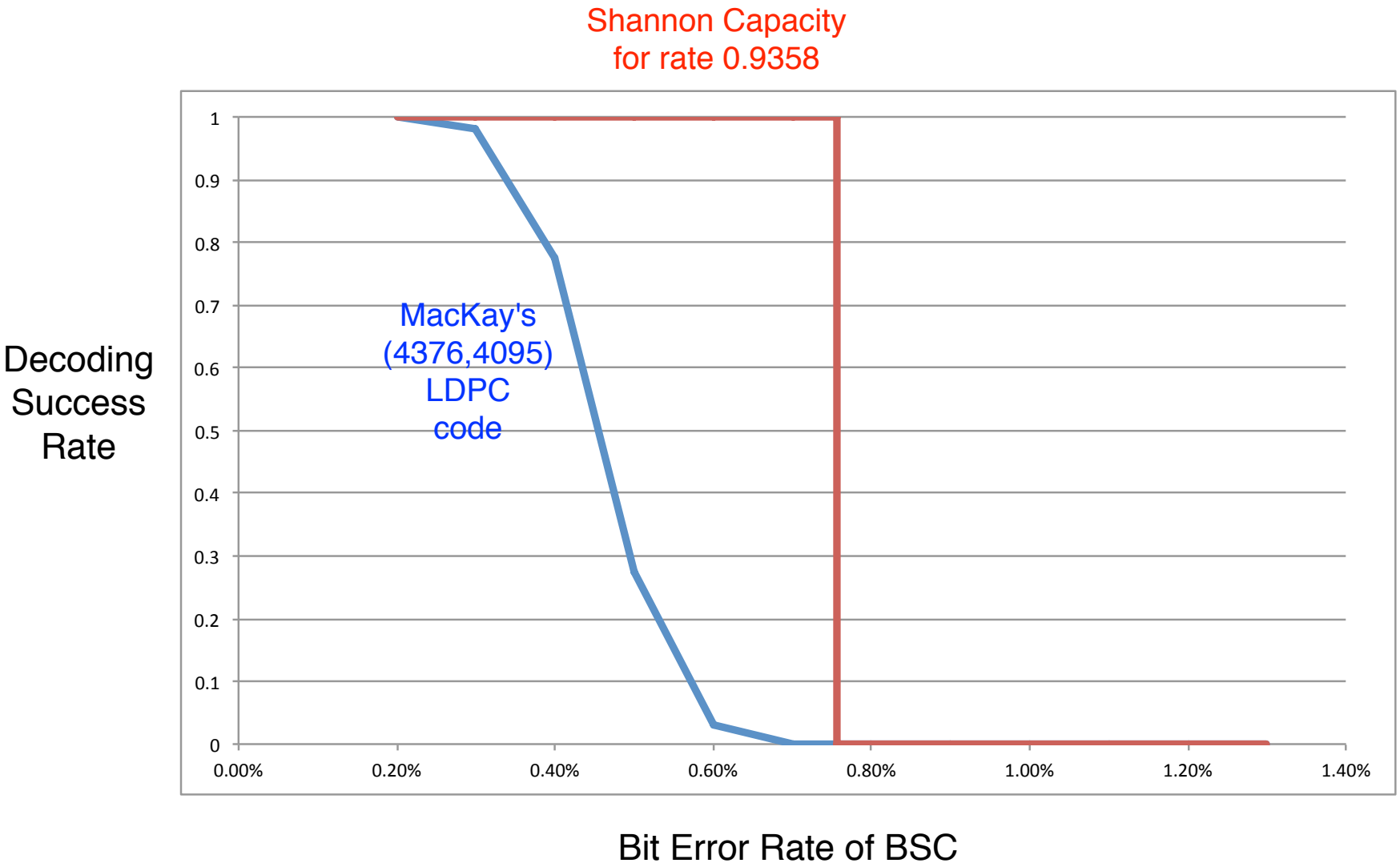
Error-Correction Framework



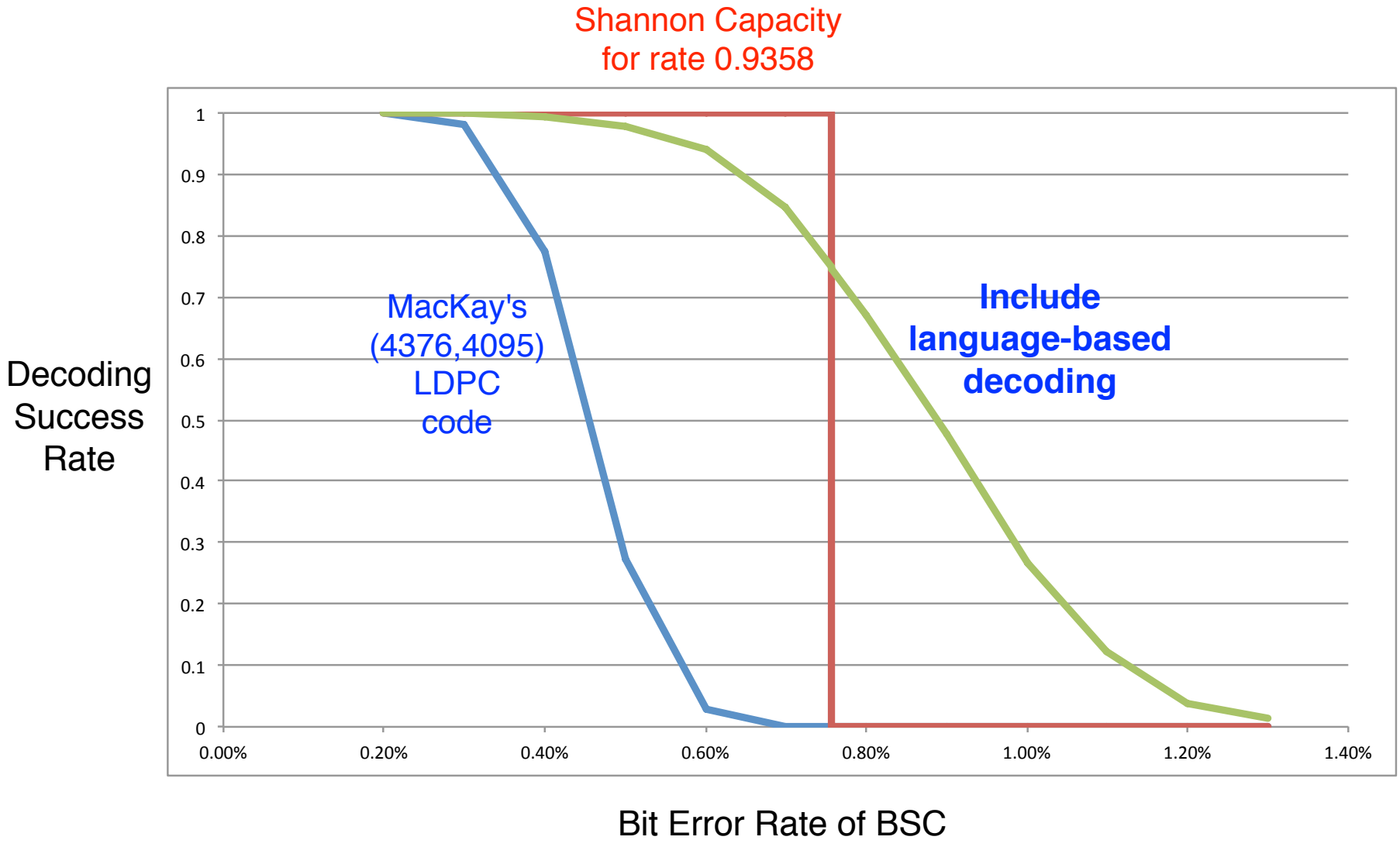
Does it work?


Decoding Success Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
MacKay's (4376,4095) LDPC code
0.00%
0.20%
0.40%
0.60%
0.80%
1.00%
1.20%
1.40%
Bit Error Rate of BSC

Does it work?


Shannon Capacity
for rate 0.9358
Decoding
Success
Rate
MacKay's
(4376,4095)
LDPC
code
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.00%
0.20%
0.40%
0.60%
0.80%
1.00%
1.20%
1.40%
Bit Error Rate of BSC

Does it work?


Shannon Capacity
for rate 0.9358
Decoding
Success
Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.00%
0.20%
0.40%
0.60%
0.80%
1.00%
1.20%
1.40%
MacKay's
(4376,4095)
LDPC
code
Include
language-based
decoding
Bit Error Rate of BSC

Does it work?

- It can be proved: If optimal codes (codes that reach Shannon capacity) are used, the required redundancy will be 3.52 times the redundancy used by the shown code.

The above decoding algorithm is based on word recognition.

Key observation:

- Valid texts are **extremely sparse**.

A Manual Example

Consider the well known paragraph by Shannon:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

A Manual Example

Compress it with an optimized Huffman code (for characters):

101101100001111100101101000001110010111101000001001111001010101000110000101110111001010000010001100111.....000

A Manual Example

Compress it with an optimized Huffman code (for characters):

101101100001111100101101000001110010111101000001001111001010101000110000101110111001010000010001100111.....000

Add i.i.d. errors to the codeword with BER 1%:

101101100001111100101101000001110000111101000001001111001010101000110000101110111001010000000001100111.....000

A Manual Example

Compress it with an optimized Huffman code (for characters):

101101100001111100101101000001110010111101000001001111001010101000110000101110111001010000010001100111.....000

Add i.i.d. errors to the codeword with BER 1%:

101101100001111100101101000001110000111101000001001111001010101000110000101110111001010000000001100111.....000

Comment: 1% is a large BER for storage systems.

A Manual Example

Decode with Huffman code:

The fuidamental proc nra of communication is that of reproducing at one point eithea exactly or approximatelorgaaessage selected at another point. Frecduently the messages have meaning; that is they epoer to y(nri,rrelated according to some system with certain pnsesica.sl c aih0ohal entities. These semantic aspects of communication are irrelevast to the engineering problem. The tignificant atpecs is that thloc a o iaaessa nrrffnledlected from a set of possible ne0nles. The system muszae designed to operquttidach possible selection, 0t just the one which wil)cohally be chosen since t, olfa rig iwn at the time of design.

A Manual Example

Decode with Huffman code:

The fuidamental proc nra of communication is that of reproducing at one point eithea exactly or approximatelorgaaessage selected at another point. Frecduently the messages have meaning; that is they epoer to y(nri,rrelated according to some system with certain pnsesica.sl c aih0ohal entities. These semantic aspects of communication are irrelevast to the engineering problem. The tignificant atpecs is that thloc a o iaaessa nrrffnledlected from a set of possible ne0nles. The system muszae designed to operquttidach possible selection, 0t just the one which wil)cohally be chosen since t, olfa rig iwn at the time of design.

Comment: Error propagation is not too bad.

A Manual Example

Decode with Huffman code:

The fuidamental proc nra of communication is that of reproducing at one point eithea exactly or approximatelorgaaessage selected at another point. Frecduently the messages have meaning; that is they epoer to y(nri,rrelated according to some system with certain pnsesica.sl c aih0ohal entities. These semantic aspects of communication are irrelevast to the engineering problem. The tignificant atpecs is that thloc a o iaaessa nrrffnledlected from a set of possible ne0nles. The system muszae designed to operquttidach possible selection, 0t just the one which wil)cohally be chosen since t, olfa rig iwn at the time of design.

Comment: Red segments are "recognizable", black segments are "noisy segments".

A Manual Example

Number of bits	Number of errors	Number of all solutions	Number of decodable solutions	Number of "valid" solutions	"Valid" solutions
138	3	438128	373033	4	(1) th actual message is on selected (2) th actual message is one selected (3) the actual message is on selected (4) the actual message is one selected

The fuidamental proc nra of communication is that of reproducing at one point eithea exactly or approximatelorgaaessage selected at another point. Frecduently the messages have meaning; that is they epoer to y(nri,rrelated according to some system with certain pnsesica.sl c aih0ohal entities. These semantic aspects of communication are irrelevast to the engineering problem. The tignificant atpecs is that thloc a o iaaessa nrrffnledlected from a set of possible ne0nles. The system muszae designed to operquttidach possible selection, 0t just the one which wil)cohally be chosen since t, olfa rig iwn at the time of design.

A Manual Example

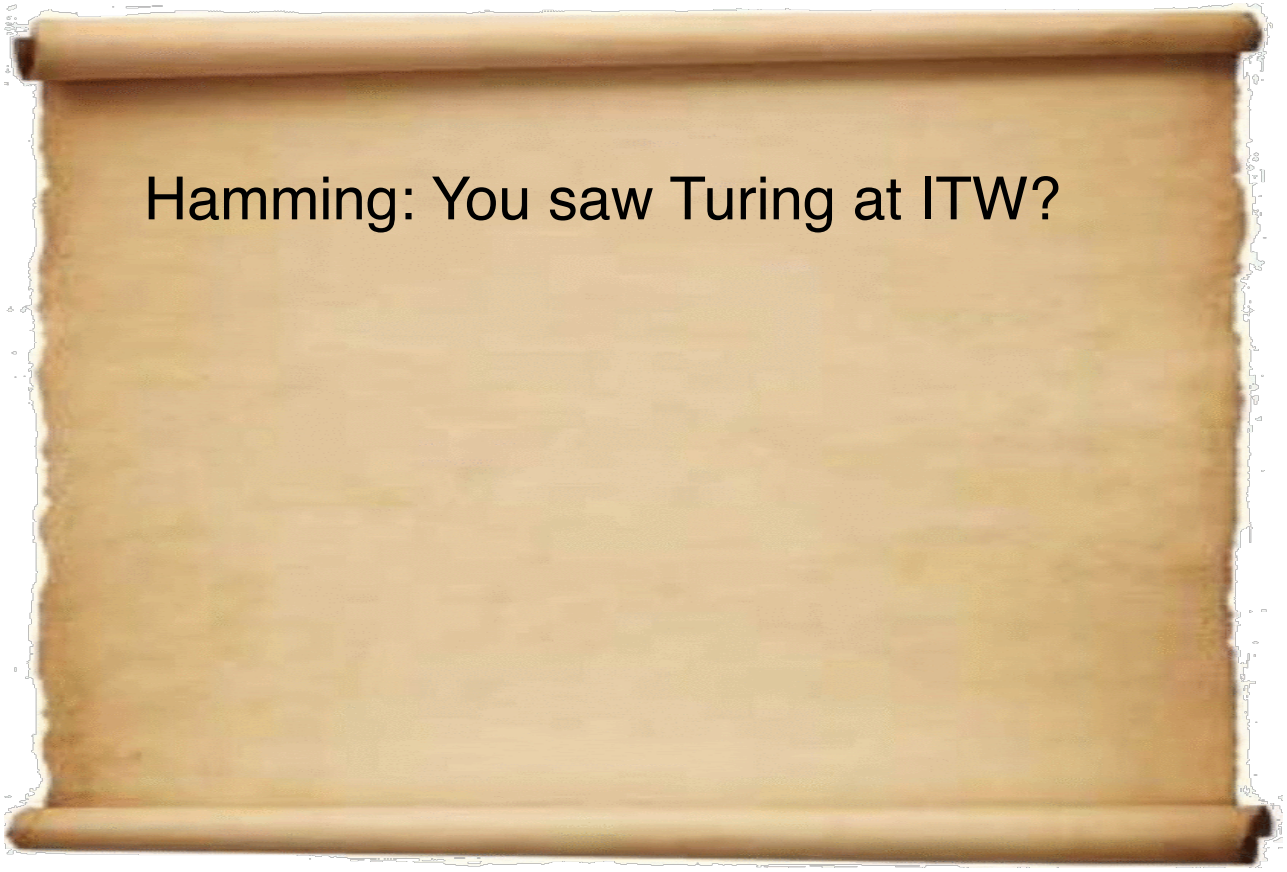
Observations:

- It is easy to recognize the correct answer for all 15 "noisy segments".
- Note that 1% is a large BER for storage systems. (Many storage systems use error-correcting codes designed for BER of 0.4% or less.) However, the noisy paragraph here has been corrected completely via human scanning without using any additional redundancy (as ECCs do).

The above decoding algorithm is based on word recognition.

More techniques at many levels:

- Grammar
- Context
- Topic
- Meaning
- Knowledge
- Logic
- ...


Hamming: You saw Turing at ITW?

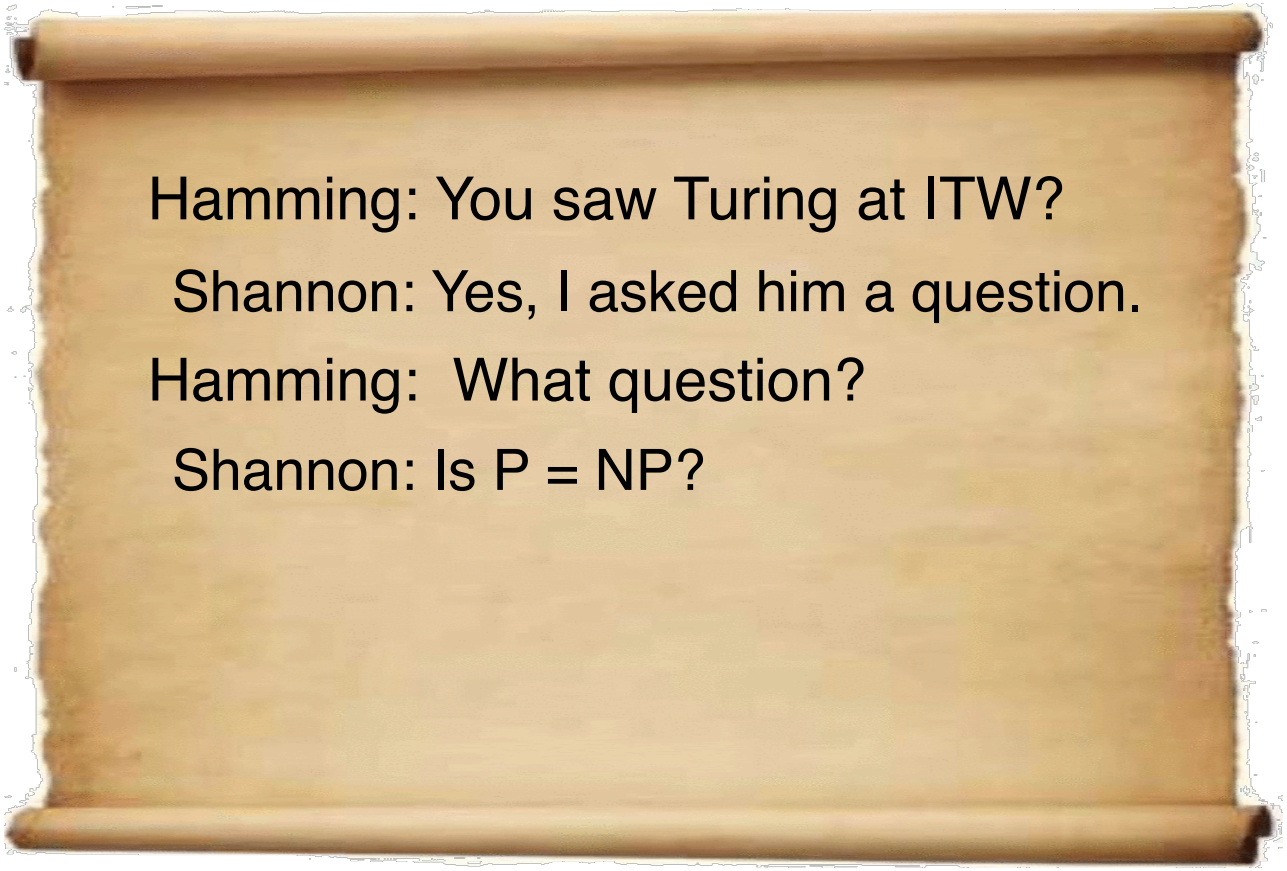
Hamming: You saw Turing at ITW?

Shannon: Yes, I asked him a question.

Hamming: You saw Turing at ITW?

Shannon: Yes, I asked him a question.

Hamming: What question?


Hamming: You saw Turing at ITW?
Shannon: Yes, I asked him a question.
Hamming: What question?
Shannon: Is P = NP?

Hamming: You saw Turing at ITW?

Shannon: Yes, I asked him a question.

Hamming: What question?

Shannon: Is P = NP?

Hamming: What did he say?

Hamming: You saw Turing at ITW?

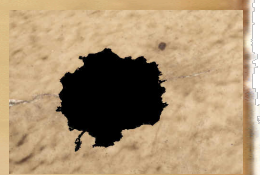
Shannon: Yes, I asked him a question.

Hamming: What question?

Shannon: Is P = NP?

Hamming: What did he say?

Shannon: He said the answer is



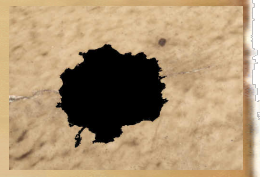
Hamming: You saw Turing at ITW?

Shannon: Yes, I asked him a question.

Hamming: What question?

Shannon: Is P = NP?

Hamming: What did he say?

Shannon: He said the answer is 

Hint for the erasure:

1. **yes**
2. **no**