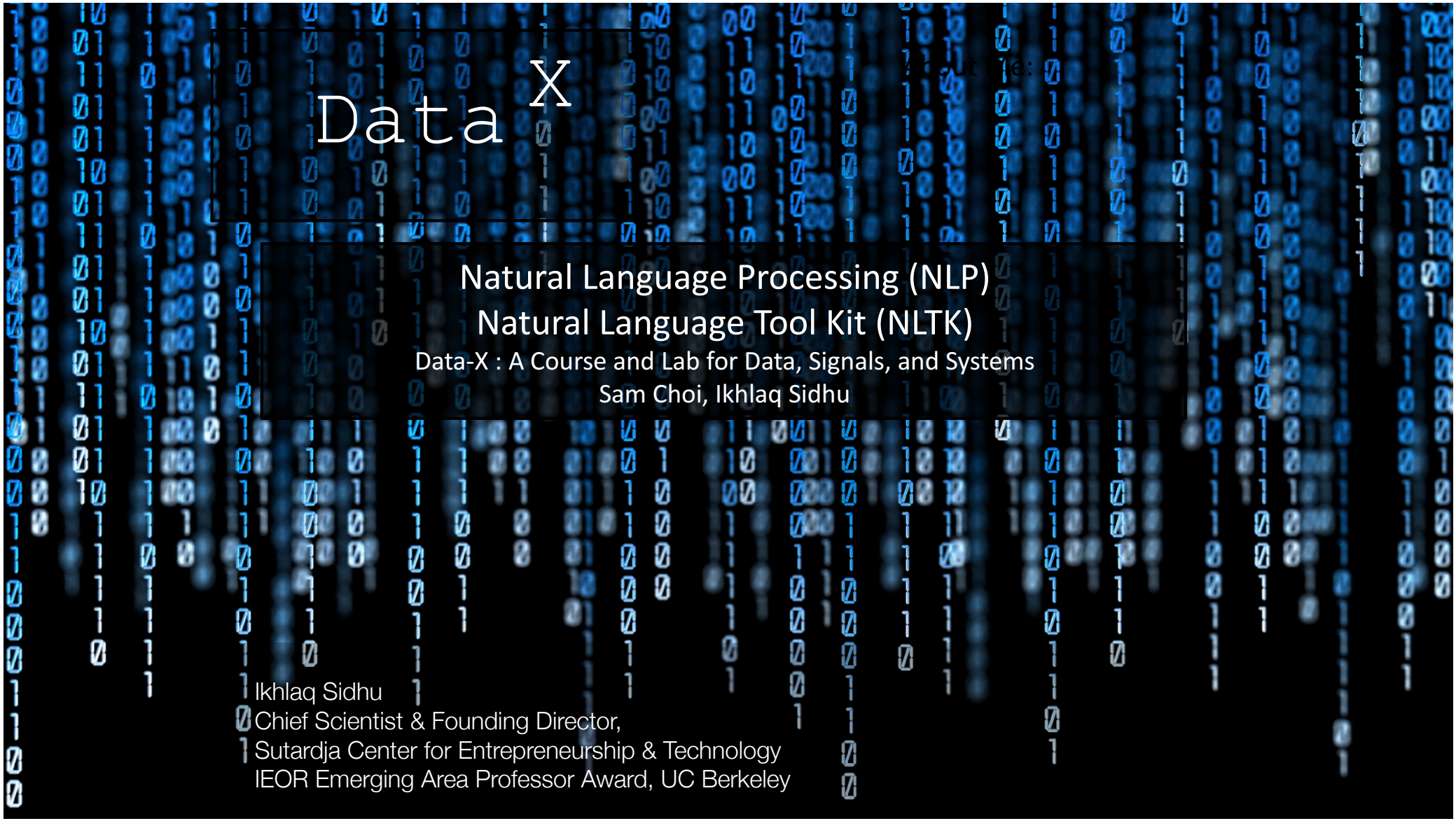

Data X
Natural Language Processing (NLP)
Natural Language Tool Kit (NLTK)
Data-X : A Course and Lab for Data, Signals, and Systems
Sam Choi, Ikhlaq Sidhu
Ikhlaq Sidhu
Chief Scientist & Founding Director,
Sutardja Center for Entrepreneurship & Technology
IEOR Emerging Area Professor Award, UC Berkeley

NLP: Main Idea

- What is NLP?
  - Natural Language Processing
- Core Questions
  - How can we get a computer to understand speech and writing?
  - How can we get a computer to speak/write like a person?

NLP Applications

- How can we use NLP to our advantage?
- High-Level Applications
  - Language Translation (Google Translate)
  - Speech detection (Siri, SoundHound)
  - Sentiment Analysis (Kensho)
  - Plagiarism detector (turnitin)
  - Grammar/Spelling checking (gmail, microsoft word)
  - Construction/Generation (chat bots)

NLP Applications

- Lower Level Applications
  - Co-reference
    - Multiple words refer to the same subject
    - Ex: Ikhlaq, professor, he
  - Classification
    - Labeling input based on type/class
  - Morphological
    - Identifying different forms of a word
    - Ex: open, opened, opens, opening

NLTK

- What is NLTK?
  - Natural Language Toolkit
- Features
  - Sentence & word tokenization
  - Part of speech tagging
  - Chunking & named entity recognition
  - Text classification
- Resources
  - Corpora, large sets of organized data
  - Sources include: WSJ, twitter, Project Gutenberg, etc.

NLTK: Getting Started

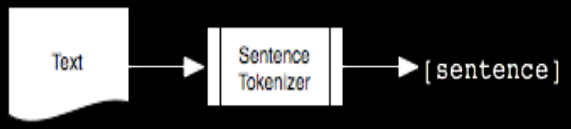
- Install Python
  - https://www.python.org/downloads/
- Install NLTK
  - http://www.nltk.org/install.html
- Download Corpora (NLTK Data)
  - http://www.nltk.org/data.html

Using NLTK

- Basic Functions
  - words()
    - Partitions a text file into a list where each element in a word
  - sents()
    - Partions a text file into lists of words – each list a sentence
  - sent_tokenize
    - Organize text into a list of sentences
  - word_tokenize
    - organize text into a list of words
  - pos_tag
    - tag part of speech for each word in a list

Sentence Tokenization

```
>>> from nltk.tokenize import sent_tokenize
>>> sent_tokenize("Hello SF Python. This is NLTK.")
['Hello SF Python.', 'This is NLTK.']

>>> sent_tokenize("Hello, Mr. Anderson. We missed you!")
['Hello, Mr. Anderson.', 'We missed you!']
```



Jacob Perkins

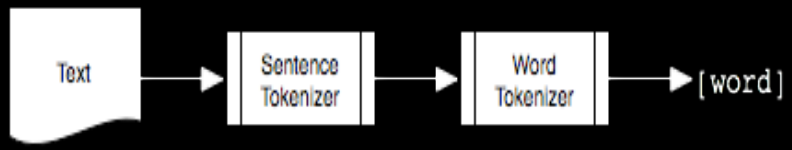
Using NLTK: sent_tokenize()

- sent_tokenize()
  - Takes a single string as input
  - Returns the string as a list of sentences

```
>>> from nltk.tokenize import sent_tokenize

>>> bio = “Hello world! My name is Ikhlaq Sidhu.”

>>> sent_tokenize(bio)
['Hello world!', 'My name is Ikhlaq Sidhu.']
```

Word Tokenization

```
>>> from nltk.tokenize import word_tokenize
>>> word_tokenize('This is NLTK.')
['This', 'is', 'NLTK', '.']
```



Jacob Perkins

Using NLTK: word_tokenize()

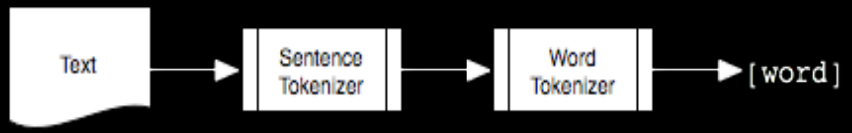
- word_tokenize()
  - Takes a single string as input
  - Returns the string as a list of words

```
>>> from nltk.tokenize import word_tokenize

>>> bio = “Hello world! My name is Ikhlaq Sidhu.”

>>> word_tokenize(bio)
['Hello', 'world', '!', 'My', 'name', 'is', 'Ikhlaq',
'Sidhu', '.']
```

What's a Word?

```
>>> word_tokenize("What's up?")
['What', "'s", 'up', '?']
>>> from nltk.tokenize import wordpunct_tokenize
>>> wordpunct_tokenize("What's up?")
['What', "'", 's', 'up', '?']
```

Learn More: http://text-processing.com/demo/tokenize/

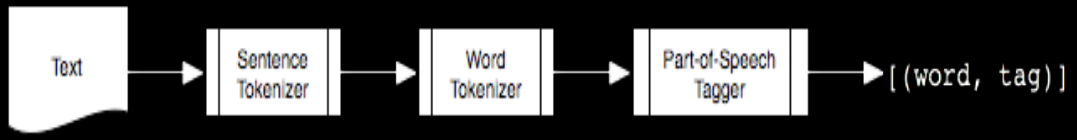


Jacob Perkins

Part-of-Speech Tagging

```
>>> words = word_tokenize("And now for something completely
different")
>>> from nltk.tag import pos_tag
>>> pos_tag(words)
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something', 'NN'), ('completely',
'RB'), ('different', 'JJ')]
```

Tags List:
http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html



Jacob Perkins

Using NLTK: pos_tag()

- pos_tag()
  - Takes a list of words as input
  - Returns a

```
>>> from nltk import pos_tag

>>> bio = “Hi my name is Ikhlaq”

>>> pos_tag(word_tokenize(bio))
[('Hi', 'NNP'), ('my', 'PRP$'), ('name', 'NN'), ('is',
'VBZ'), ('Ikhlaq', 'NNP')]
```

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

NLTK Data

- Organized into collections of written texts (corpora)
- Examples of NLTK Corpora
  - gutenberg (Project Gutenberg selections)
  - shakespeare (selection of Shakespeare's plays)
  - twitter_samples (samples of tweets)
  - brown (Brown University's collection of published works)
  - cmudict (Carnegie Mellon's dictionary of words/pronunciations)

End of Section