Data Science Training Program

About Me

Name: Dae Won Kim

Major: Operations Research, M.Eng.

Senior Advisor, CDS

History:

- President
- Yelp subteam manager

Fun Facts:

1) I was a freshman in 2010
2) I was in the Korean army but used VBA

dk444@cornell.edu

What Is This Class?

- Focus on application
- Data scientist starter pack
- Learning to speak data science
- Understanding those buzzwords
- A gateway to becoming a CDS member

Remember JJL

Jared Junyoung Lim

Education Lead, CDS

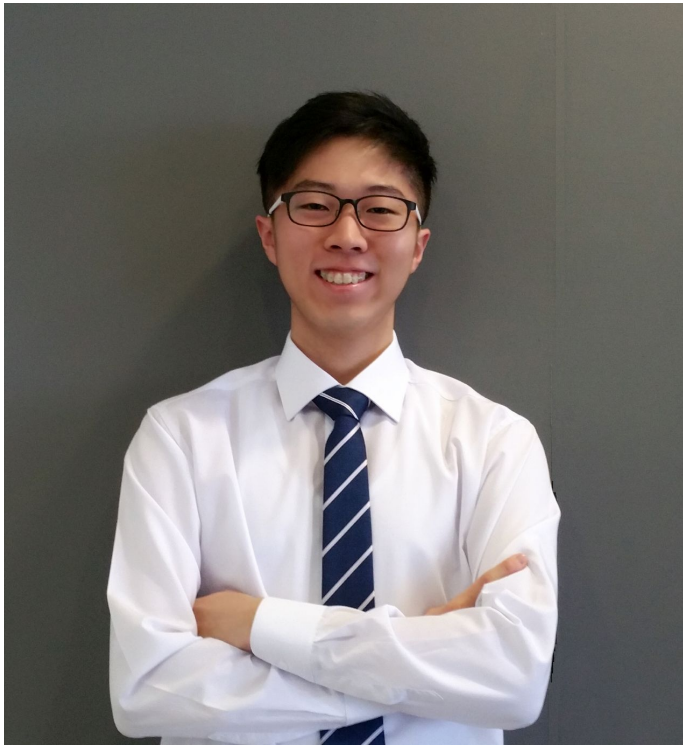
Instructor, INFO 1998

Computer Science ‘20

Fun Facts:

1) **No fun fact**
2) Does **not** tolerate **fun** and **facts**
3) There will be **no fun** in this class
4) #3 is a **fact**

jl3248@cornell.edu



Teaching Associates

Piazza Team		Office Hour Team	
Abby Beeler	arb379	Ann Zhang	az275
Kexin Zheng	kz73	Cameron Ibrahim	cai29
Shubhom Bhattacharya	sb2287	Ryan Kannanaikal	rk635

Course Logistics

11-Week Course

- Leaf 1: **Data Analysis** (1-3)
- Leaf 2: **Machine Learning** (4-11)

One Big Project

- Divided into **5 parts**

\+ Tiny miny little **quizzes** for **lecture 1 & 2**

Form a GROUP of 3-4 people ASAP

Course Logistics

Grading

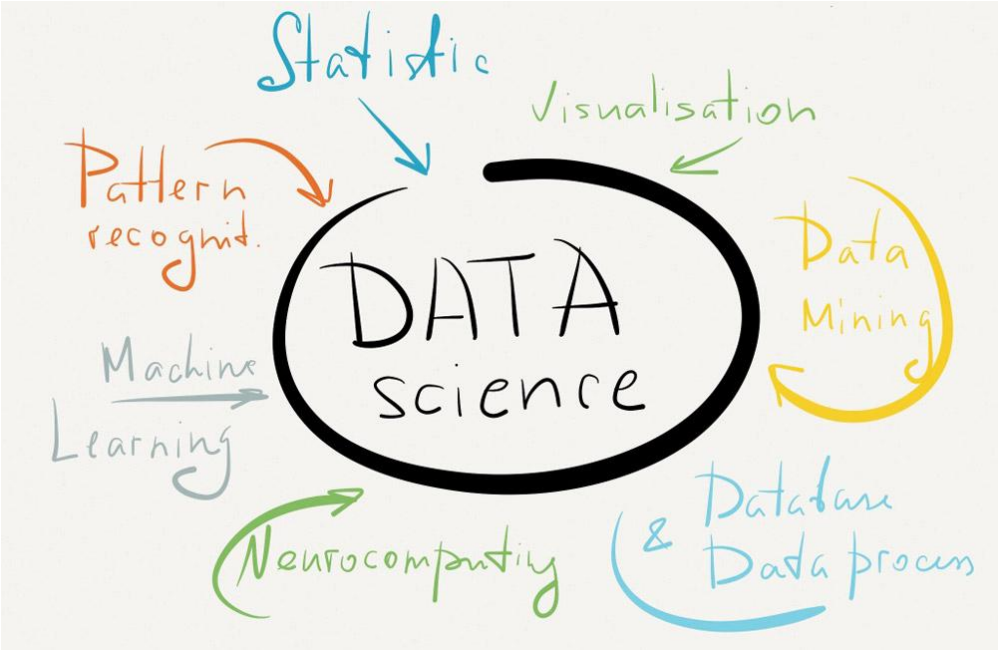
- **10%** Take-home Quiz 1
- **10%** Take-home Quiz 2
- **15%** Each of Project part A, B, C, D
- **20%** Project part E

70%

Every Assignment due Tuesday Midnight

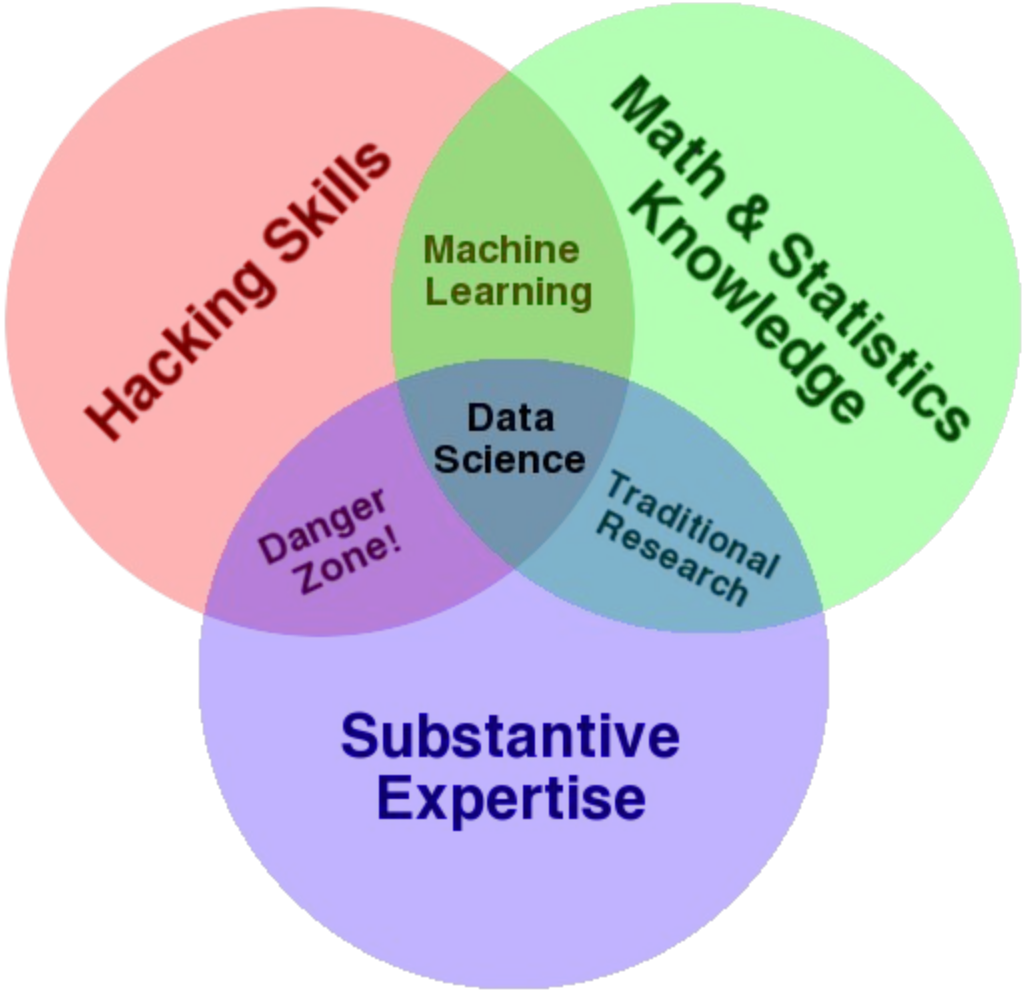
What is Data Science?

- Empirical Research
- Predictive Analytics
- Preventive Analytics
- Real-time Analysis
- Automation

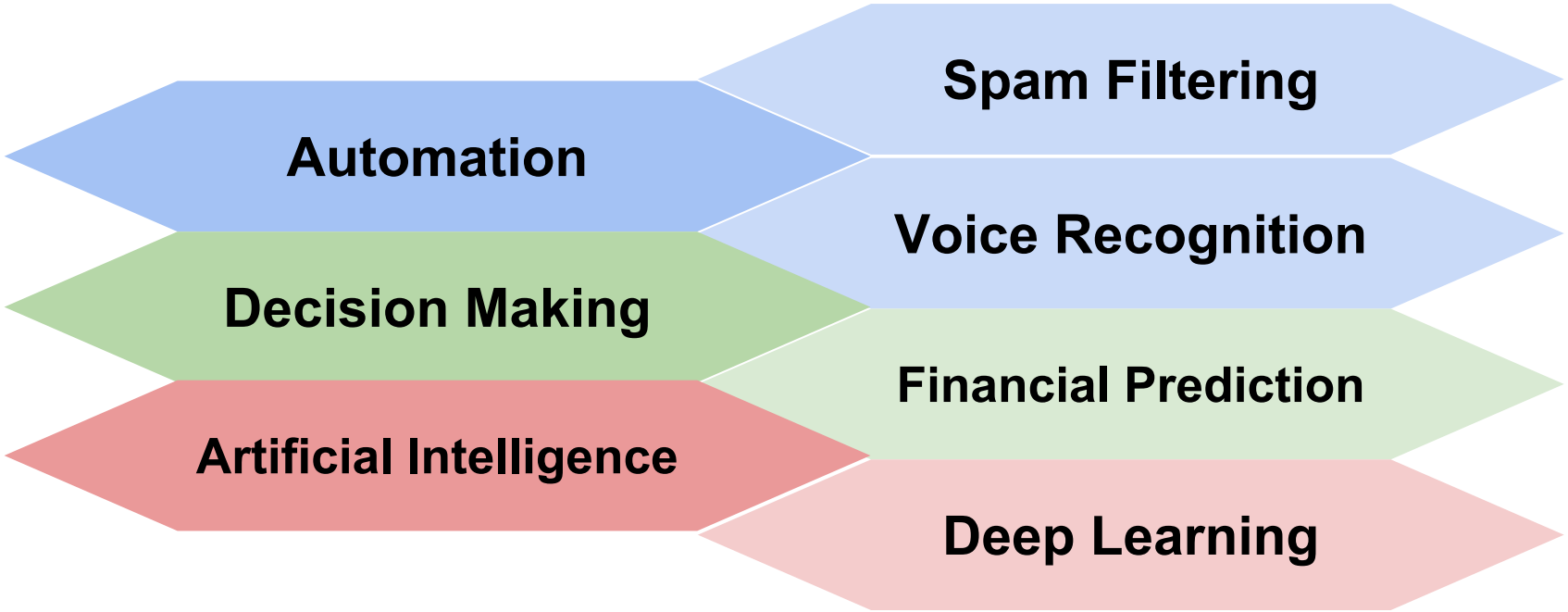


Data can be...

LARGE	**V**olume
fast	**V**elocity
unStRUcTUReD	**V**ariety

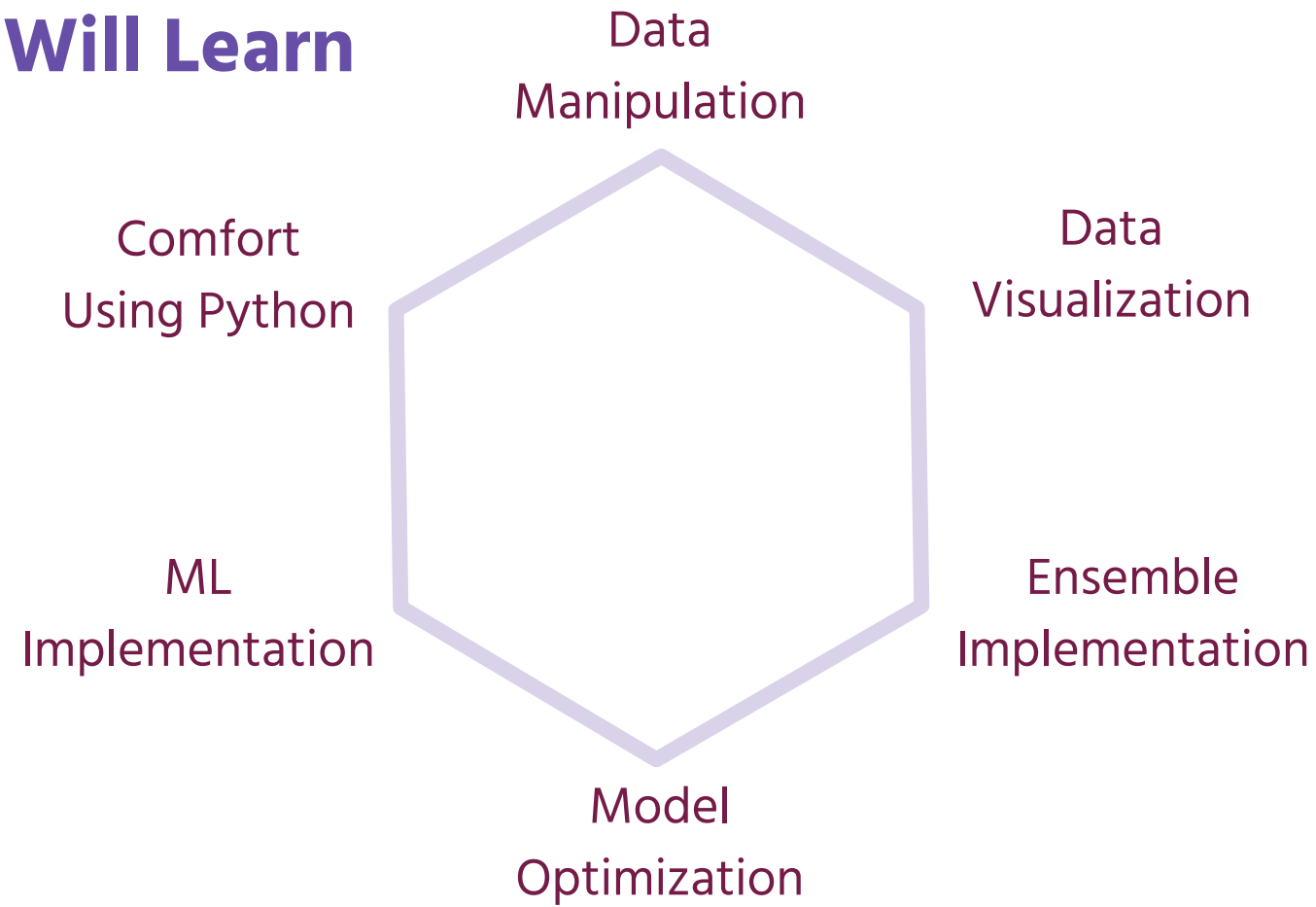

Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Applications



Data science in action

What You Will Learn



… Why?

Why Data Science?

Why Python?

Why Jupyter?

Why CDS?

Why Jupyter Notebooks?

- Document the process
  - Code
  - Visuals
- Intuitive
  - Supports Python, R, Julia, etc.
- Easy to share

jupyter New Lecture 2 Last Checkpoint: 08/03/2017 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Markdown CellToolbar

Lecture 2: Data Transformation

Now that we've picked up some basic tools for doing data science, we're ready to sharpen our data handling skills. As you might have already observed, data rarely comes in a neatly packaged “ready-to-use” format. We need to be able to manipulate datasets and shape them as we please so that we can run machine learning algorithms on them. Let's start with geting a little bit more comfortable with R.

Type *Markdown* and LaTeX: α^2

Writing Fast R

R is an excellent language for data science. However, R behaves very differently from commonly used object-oriented languages like Java and Python. Such differences can cause huge inefficiencies to unsuspecting beginners of R. Let's take a look at one of the most misunderstood concepts in R: the inefficiency of using explicit for-loops, as indicated below.

In [15]:
```
# Process time comparison of explicit for-loop with implicit loops.
vec <- c(1:1000000)

# explicit version
system.time({for(i in 1:1000000) {
    vec[i] <- vec[i] * 2
}})

# implicit version
system.time({vec <- vec * 2})
```

```
   user  system elapsed 
  0.872   0.003   0.876 
```

```
   user  system elapsed 
  0.003   0.000   0.003 
```

Language Wars

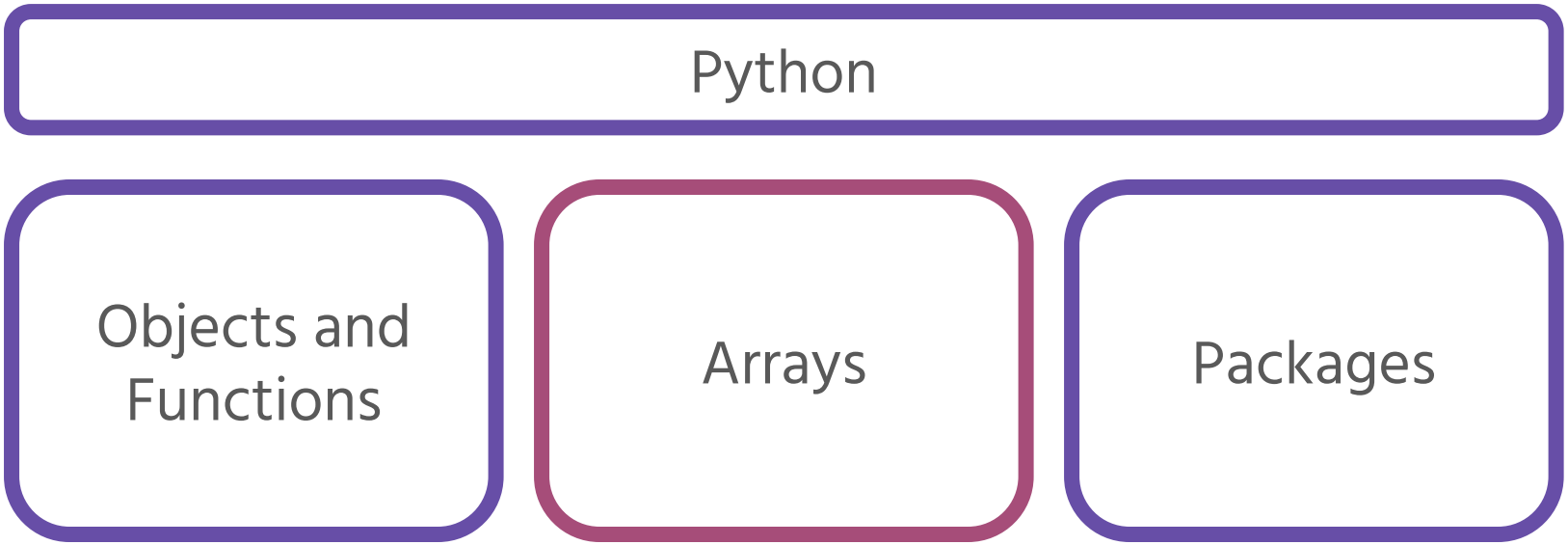
Why Python?

Easy to learn and **readable**.

Extendable and **compatible**.

Open source with a large **community**.

Python Overview



Python Data Types

Boolean	String	List	Numerical (Int, Float)
True/False	“Hello World”	[“hi, 12, True]	56 9000.1

Numpy Overview

Numpy

Arrays Improve Speed

Vectorization

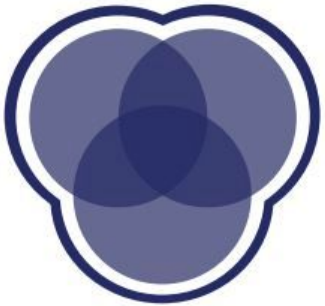
Built-in Functions

Coming Up

Your assignment: Jupyter Setup & Take-home Quiz 1 (released tonight)

Due: Next Tuesday Midnight **Submit Through:** CMS

Next week: LECTURE 2 - Data Manipulation with Pandas



CDS

Cornell Data Science

Helpful Links

Sign up for CMS here! bit.ly/cornellcdscms

^you must do this before submitting assignments^

Course website: datascienceis.life

See you next week!