

03603351 วิทยาศาสตร์ข้อมูลเบื้องต้น

# Introduction to Data Science

ภาคการศึกษาที่ 1 ปี 2562



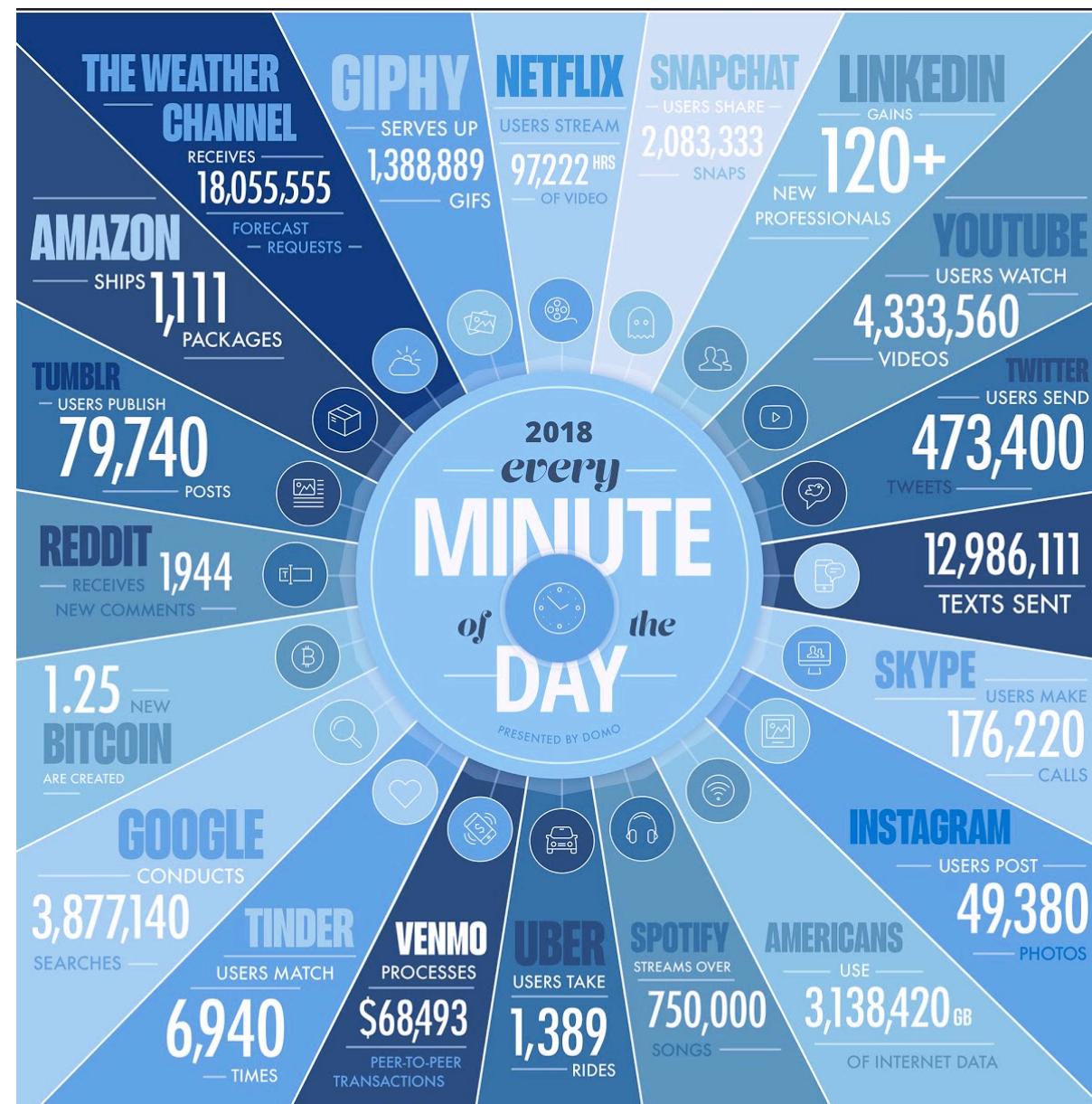
FACULTY OF ENGINEERING  
AT SRI RACHA  
DEPARTMENT OF COMPUTER ENGINEERING

ผศ.ดร. กุลวadee สมบูรณ์วิวัฒน์  
[kulwadee@eng.src.ku.ac.th](mailto:kulwadee@eng.src.ku.ac.th)

ภาควิชาวิศวกรรมคอมพิวเตอร์  
คณะวิศวกรรมศาสตร์ ศรีราชา  
มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา

# เรารวยในโลกที่เต็มไปด้วยข้อมูล

- Domo, Inc. ได้ทำการศึกษาปริมาณข้อมูลที่ถูกสร้างขึ้นบนโลกดิจิตอล ใน 1 นาที
- ผลการศึกษาพบว่า **ทุก 1 นาทีในปี 2018** มีการทวีตข้อความ **473,400** ทวีต  
มีการโพสต์รูปลงในอินสตาแกรม **49,380** รูป  
มีการค้นหาบนกูเกิล **3,877,140** ครั้ง  
มีผู้ใช้ดูวิดีโอบนยูทูป **4,333,560** วิดีโอ

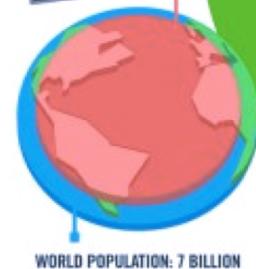


ເປັນຢຸດຂອງ BIG DATA

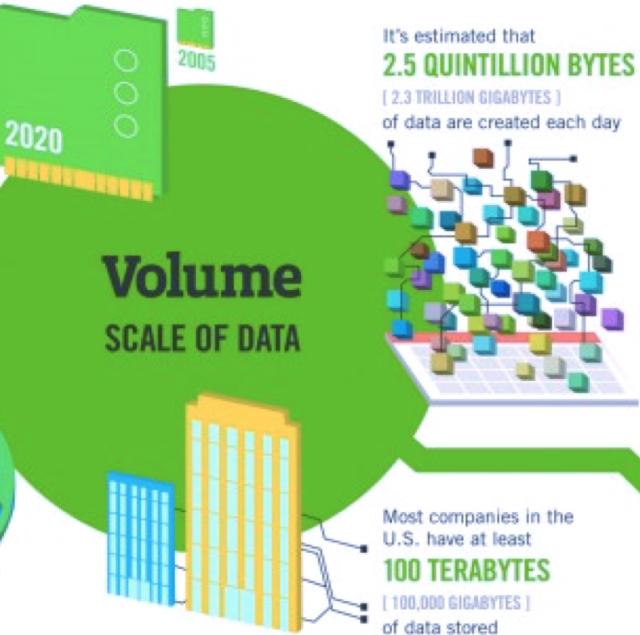
**40 ZETTABYTES**

[ 43 TRILLION GIGABYTES ]  
of data will be created by  
2020, an increase of 300  
times from 2005

 **6 BILLION**  
PEOPLE have cell  
phones



## Volume SCALE OF DATA



The New York Stock Exchange captures  
data from **1 TB OF TRADE INFORMATION** during each trading session.



## Velocity ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

**18.9 BILLION**  
**NETWORK CONNECTIONS**

– almost 2.5 connections per person on earth



# The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume**, **Velocity**, **Variety** and **Veracity**.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States.



As of 2011, the global size of data in healthcare was estimated to be

**150 EXABYTES**  
[ 161 BILLION GIGABYTES ]



## Variety DIFFERENT FORMS OF DATA

**30 BILLION**  
**Pieces of Content**

are shared on Facebook every month

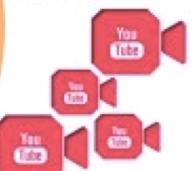


By 2014, it's anticipated there will be

**420 MILLION**  
**WEARABLE, WIRELESS**  
**HEALTH MONITORS**

**4 BILLION+**  
**HOURS OF VIDEO**

are watched on YouTube each month



**400 MILLION**  
**TWEETS** are sent per day by about 200 million monthly active users



## 1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around

**\$3.1 TRILLION A YEAR**



## Veracity UNCERTAINTY OF DATA

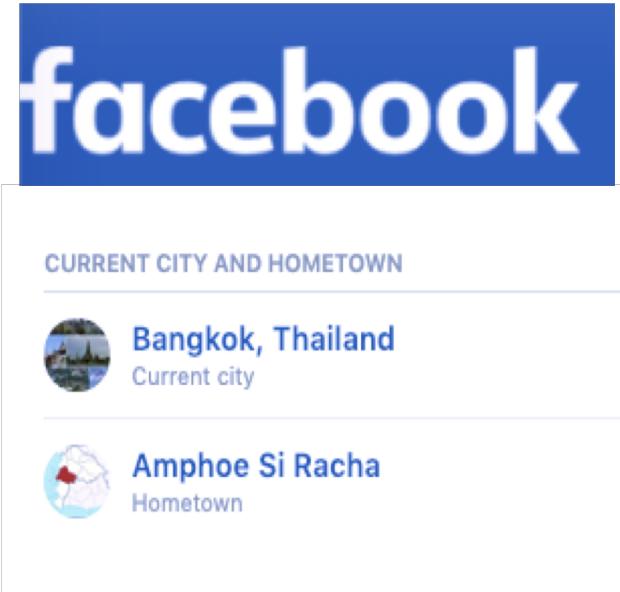
# What is Data Science ?

- วิทยาศาสตร์ข้อมูล (**data science**) คือ การค้นหารูปแบบที่มีนัยสำคัญ ซึ่งแฝงอยู่ในข้อมูล แล้วนำมาใช้ให้เกิดประโยชน์
- นักวิทยาศาสตร์ข้อมูล (**data scientist**) คือ สกัดความรู้เชิงลึกจากข้อมูลโดยประยุกต์ใช้ ทักษะทางด้านคณิตศาสตร์และสถิติ, การเขียนโปรแกรมและการเรียนรู้ของเครื่องจักร, และความรู้เฉพาะทาง

# ตัวอย่างการใช้วิทยาศาสตร์ข้อมูล

- ทีมนักวิทยาศาสตร์ข้อมูลของ Facebook ศึกษาฐานแบบการโยกย้ายถิ่นฐานของประชากรโลก (ที่เป็นสมาชิกของ Facebook) โดยใช้ข้อมูล hometown และ current city จากโปรไฟล์ของผู้ใช้งาน
- Amazon ใช้ข้อมูลการซื้อสินค้าของผู้ใช้งาน เพื่อสร้างระบบแนะนำสินค้า (recommender system)
- Grab วิเคราะห์ข้อมูลการเดินทางของผู้ใช้งาน เพื่อให้ระบบสามารถจับคู่ คนขับ กับ ผู้โดยสาร ได้รวดเร็วที่สุด
- การหา Key connectors ในเครือข่ายสังคม (social network)

# การวิเคราะห์ การโยกย้ายถิ่นฐานของประชากรโลก โดย Facebook Data Science Team



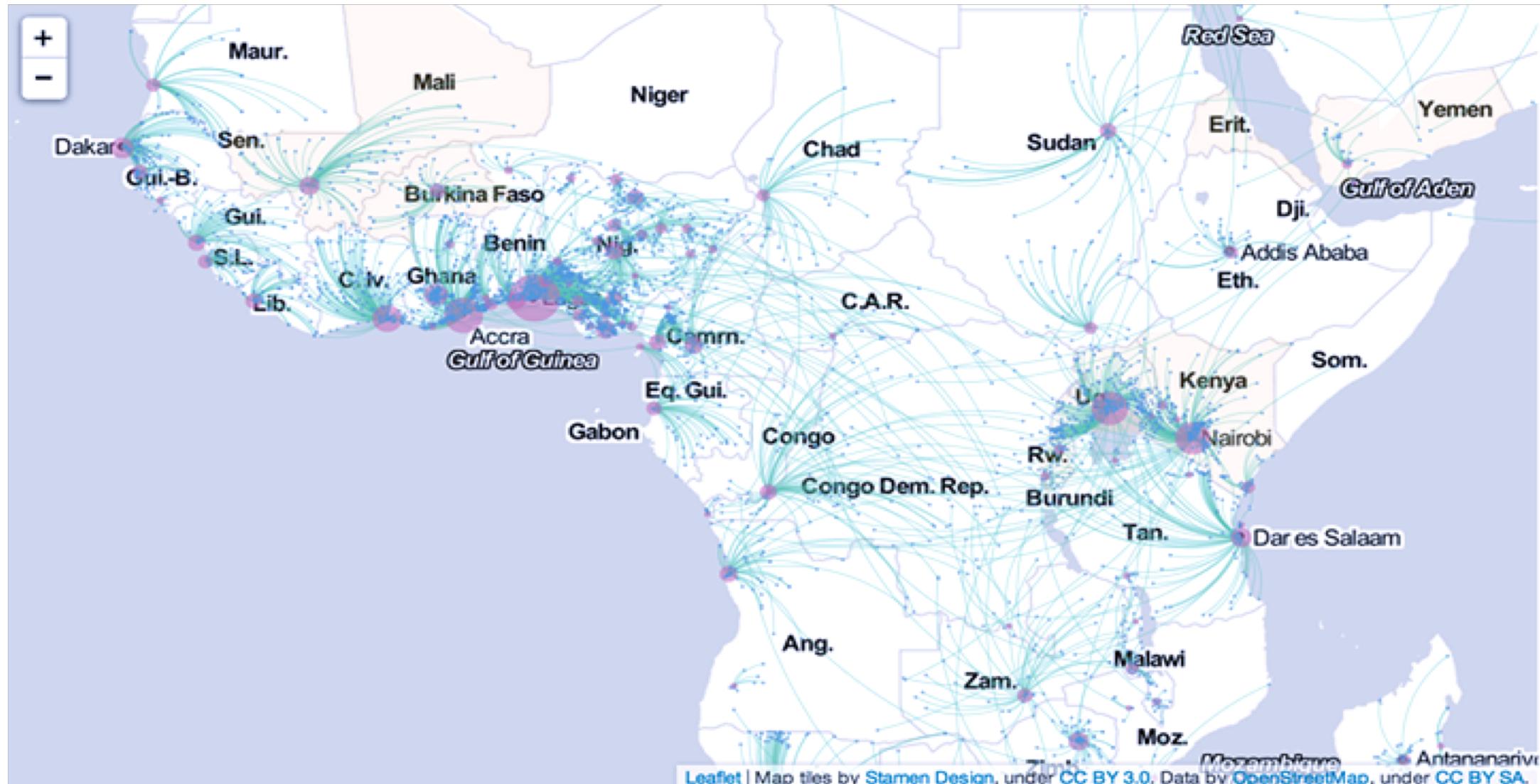
- ใช้ข้อมูล Hometown (บ้านเกิด) และ Current City (เมืองที่อาศัยอยู่ปัจจุบัน) วิเคราะห์หารูปแบบการโยกย้ายถิ่นฐานของประชากรโลก
- นิยามปัญหา: Coordinated migration คือการโยกย้ายจากบ้านเกิด  $h$  ไปยังเมืองปัจจุบัน  $c(h)$  ที่มีความเป็นไปได้มากที่สุด
$$h \Rightarrow c(h) \text{ ซึ่งมีค่า } P( c(h) | h ) \text{ มากริ่ง}$$

\* อ่านรายละเอียดเพิ่มเติมได้ที่ <https://goo.gl/uQu626>

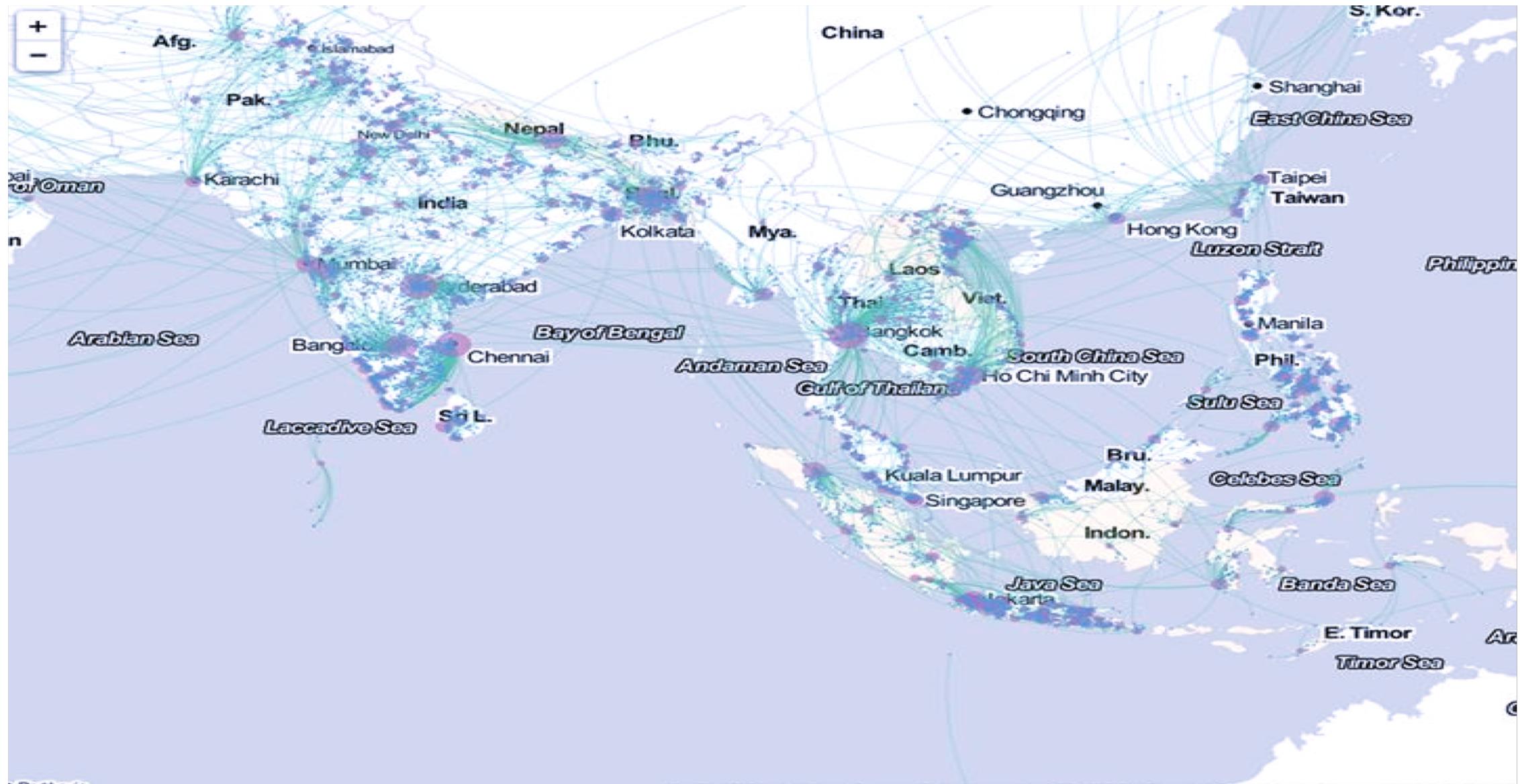
# ผลการวิเคราะห์ 1: เมืองปลายทางหลัก อยู่ในประเทศที่มีการเติบโตของสังคมเมืองสูง

Destination city	Country	Number of coordinated migration: n(h)	Urbanization growth between 2000 and 2012 (%)	Hometown countries of the coordinated migrations
Lagos	Nigeria	566	18.6	Nigeria (96%)
Istanbul	Turkey	387	11.7	Turkey (84%), Macedonia (4%), Bulgaria (3%)
Bogota	Colombia	370	4.8	Columbia (98%)
Bangkok	Thailand	322	10.7	Thailand (90%)
Accra	Ghanna	315	19.5	Ghanna (97%)
Hyderabad	India	307	14.4	India (98%)
Kampala	Uganda	280	32.4	Uganda (93%), Democratic Republic of the Congo (3%)
Lima	Peru	279	6.2	Peru (97%)
Chennai	India	278	14.4	India (98%)
London	Great Britain	270	1.4	Great Britain (94%)

## ผลการวิเคราะห์ 2: รูปแบบการโยกย้ายถิ่นฐานในแต่ละพื้นที่ แอฟริกาตะวันตก เป็นแบบ single hub

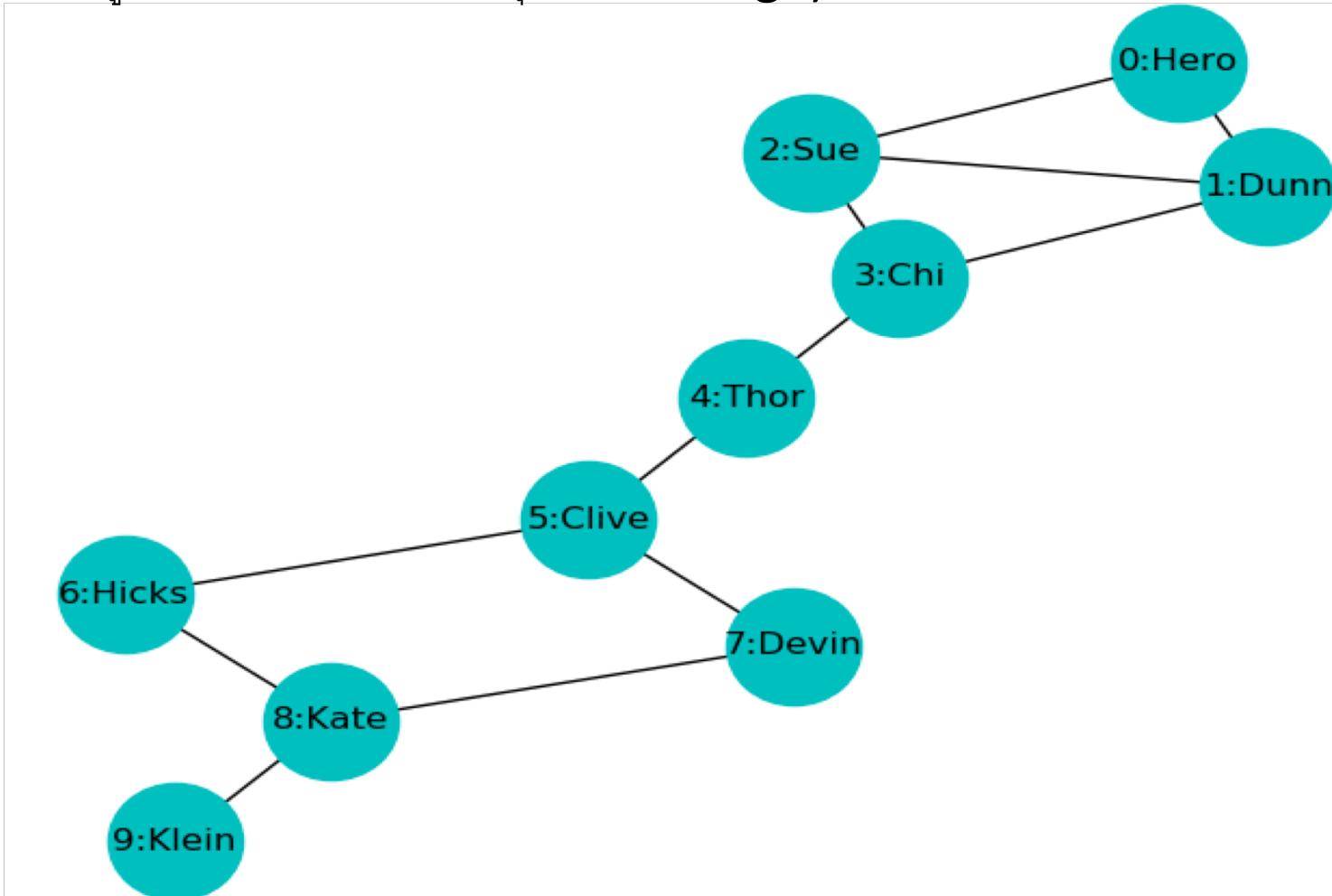


# ผลการวิเคราะห์ 3: รูปแบบการโยกย้ายถิ่นฐานในแต่ละพื้นที่ (เอเชีย) เป็นแบบ multiple hubs



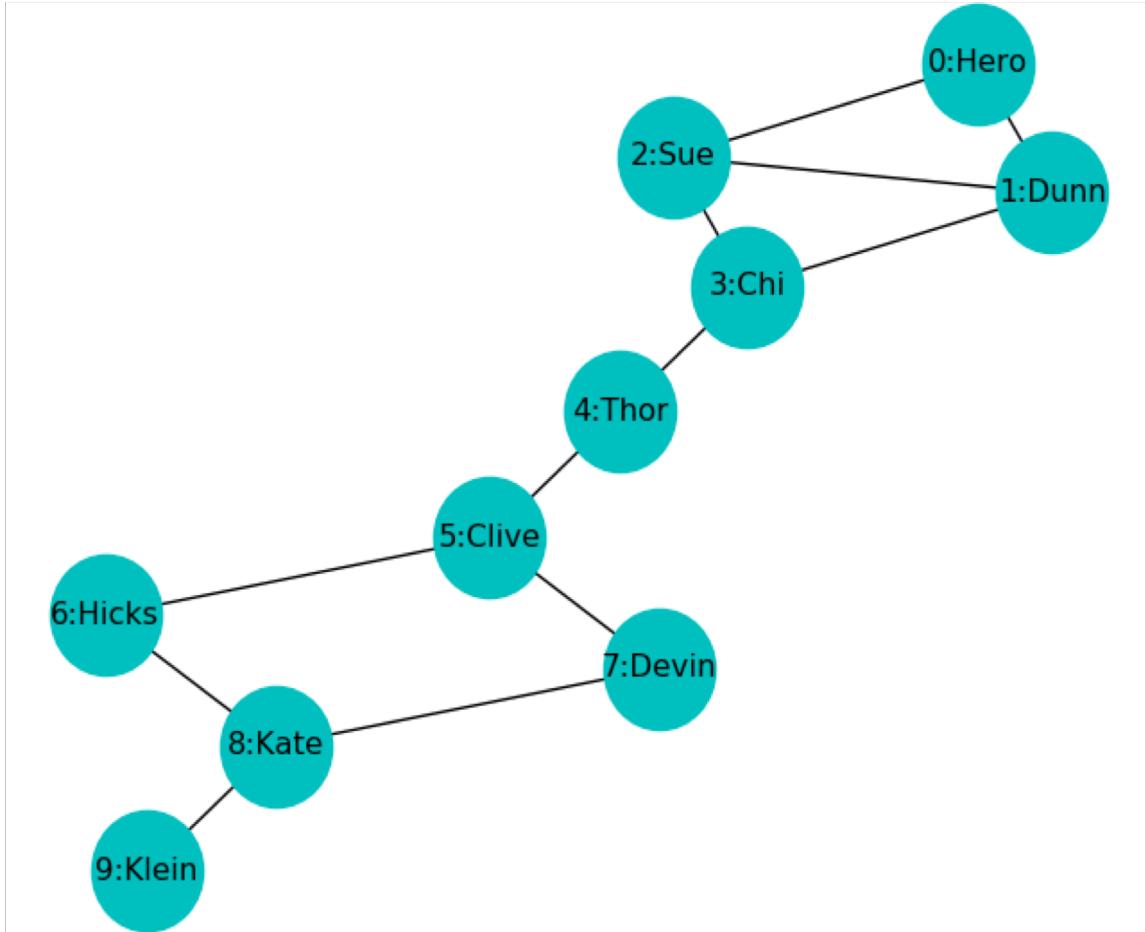
# หา Key connectors ใน Social network

- Key connectors = ผู้ใช้ที่มีจำนวนเพื่อนมากที่สุด (จำนวน edge)



# ໂຄງສ້າງຂໍ້ມູນສໍາຫຼັບເຄຣືອໜ່າຍ

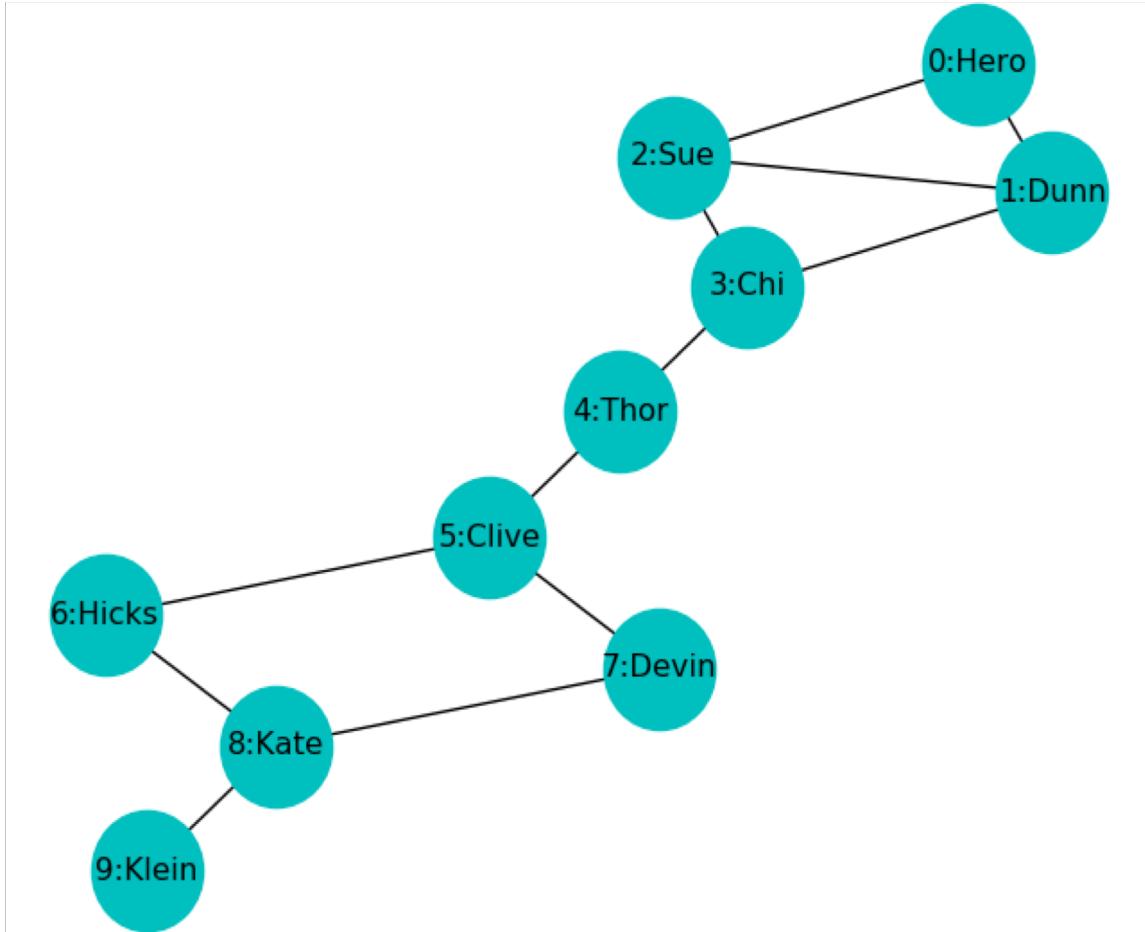
- ກາພ = ເຊຕຂອງ **nodes** ແລະ ເຊຕຂອງ **edges**



```
users = [  
    {"id": 0, "name": "Hero"},  
    {"id": 1, "name": "Dunn"},  
    {"id": 2, "name": "Sue"},  
    {"id": 3, "name": "Chi"},  
    {"id": 4, "name": "Thor"},  
    {"id": 5, "name": "Clive"},  
    {"id": 6, "name": "Hicks"},  
    {"id": 7, "name": "Devin"},  
    {"id": 8, "name": "Kate"},  
    {"id": 9, "name": "Klein"}]  
  
friendship_pairs = [(0, 1), (0, 2), (1, 2), (1, 3),  
                    (2, 3), (3, 4), (4, 5), (5, 6),  
                    (5, 7), (6, 8), (7, 8), (8, 9)]
```

# ໂຄງສ້າງຂໍ້ມູນລສຳຫວັບເຄຣືອໝ່າຍ

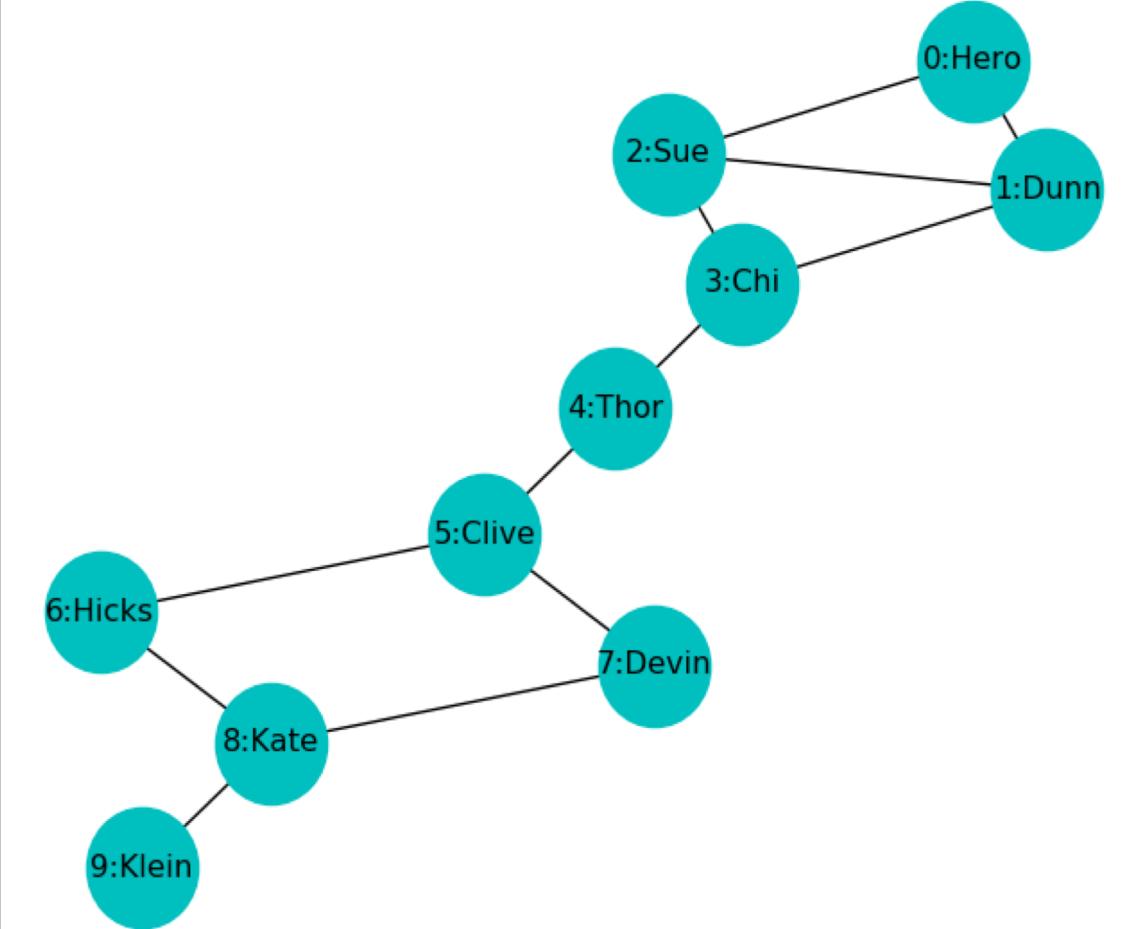
- Adjacency List



```
friendships = {user["id"]: [] for user in users}

for i, j in friendship_pairs:
    friendships[i].append(j)
    friendships[j].append(i)
```

# นับจำนวนเพื่อน (number of edges)



```
def number_of_friends(user):  
    """How many friends does _user_ have?"""  
    user_id = user["id"]  
    friend_ids = friendships[user_id]  
    return len(friend_ids)
```

# นับจำนวนเพื่อน (number of edges) – ต่อ

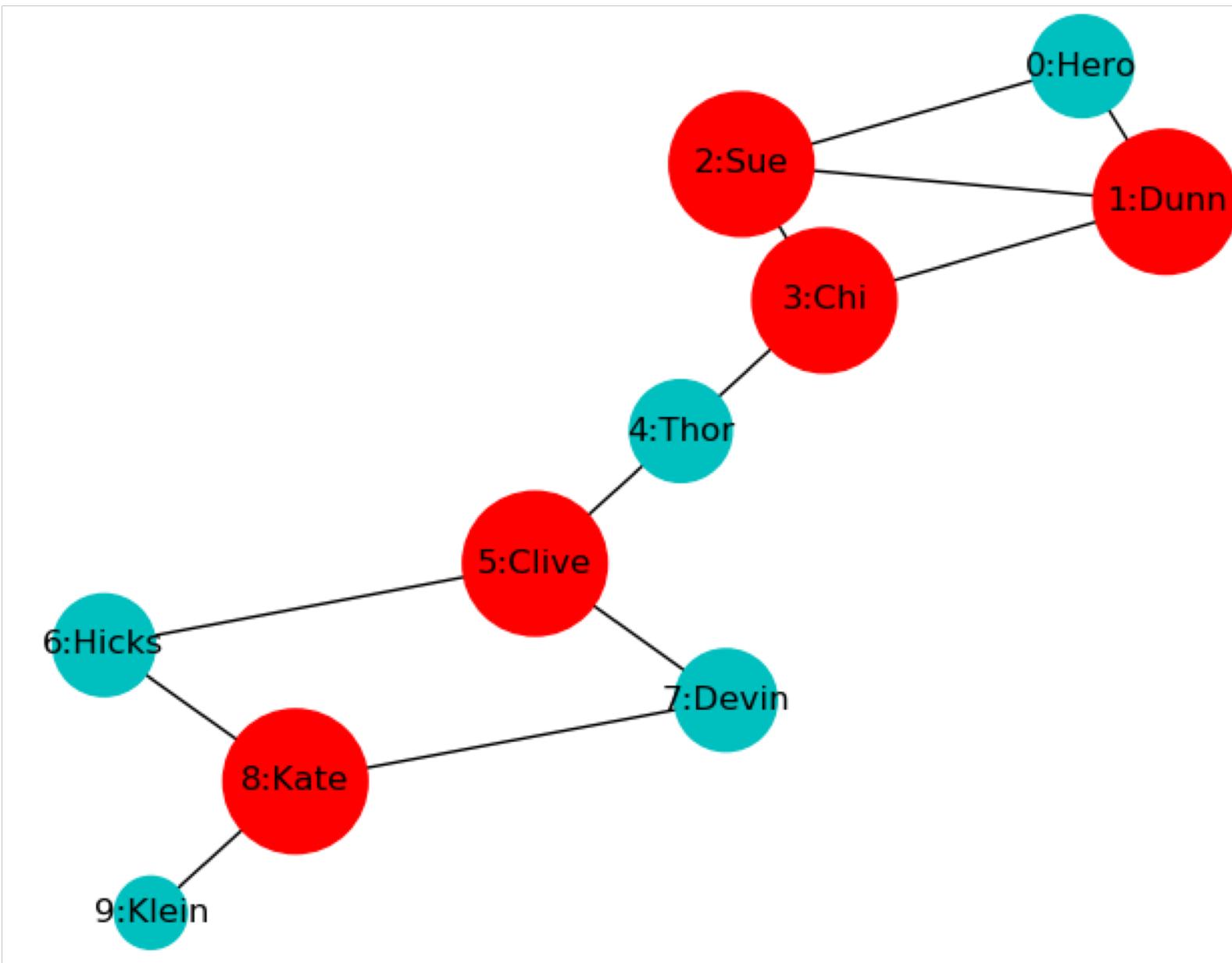
```
# หาจำนวนเพื่อนของ user แต่ละคน
num_friends_by_id = [(user["id"], number_of_friends(user))
                      for user in users]

# เรียงลำดับ user ตามจำนวนเพื่อนมากสุด ไป น้อยสุด
num_friends_by_id_sorted = sorted(num_friends_by_id,
                                    key=lambda id_and_friends: id_and_friends[1],
                                    reverse=True)

# แมป user id ไปเป็น user name
id2name = {user["id"]: "{0}:{1}".format(user["id"], user["name"])
            for user in users}

# แสดงผลลัพธ์
print("Users sorted by number of friends")
for (userid, numfriends) in num_friends_by_id_sorted:
    print("{0} {1} {2}".format(userid, id2name[userid], numfriends))
```

# Visualization – แสดงเครือข่ายเพื่อให้เห็นภาพว่า ใครคือผู้ใช้ที่เป็น Key connectors



# สรุป สิ่งที่ได้เรียนในวันนี้

- วิทยาศาสตร์ข้อมูล คือ การประยุกต์ใช้ คณิตศาสตร์ สถิติ โปรแกรมมิ่ง การเรียนรู้จากข้อมูล และ ความรู้เฉพาะทาง (domain knowledge) เพื่อสกัดรูปแบบหรือความรู้เชิงลึกจากชุดข้อมูล
- Big Data (Volume, Varieties, Velocity, Veracity) เป็นตัวเร่งให้เกิดความต้องการด้าน วิทยาศาสตร์ข้อมูลมากขึ้น
- ตัวอย่าง Data Science Projects
  - การโยกย้ายถิ่นฐานของประชากรโลก โดย Facebook Data Science Team
  - การหาตัวเชื่อมต่อหลัก (key connectors) ในเครือข่ายสังคม (social network)

```
import networkx as nx
import math
import matplotlib.pyplot as plt

users = [
    {"id": 0, "name": "Hero"}, {"id": 1, "name": "Dunn"}, {"id": 2, "name": "Sue"}, {"id": 3, "name": "Chi"}, {"id": 4, "name": "Thor"}, {"id": 5, "name": "Clive"}, {"id": 6, "name": "Hicks"}, {"id": 7, "name": "Devin"}, {"id": 8, "name": "Kate"}, {"id": 9, "name": "Klein"}]

friendship_pairs = [(0, 1), (0, 2), (1, 2), (1, 3), (2, 3), (3, 4),
                     (4, 5), (5, 6), (5, 7), (6, 8), (7, 8), (8, 9)]

friendships = {user["id"] : [] for user in users}
for i, j in friendship_pairs:
    friendships[i].append(j)
    friendships[j].append(i)

def number_of_friends(user):
    """How many friends does _user_ have?"""
    user_id = user["id"]
    friend_ids = friendships[user_id]
    return len(friend_ids)

num_friends_by_id = [(user["id"], number_of_friends(user))
                      for user in users]

num_friends_by_id_sorted = sorted(num_friends_by_id,
                                   key=lambda id_and_friends: id_and_friends[1],
                                   reverse=True)

id2name = {user["id"] : "{0}:{1}".format(user["id"], user["name"])
           for user in users}

print("Users sorted by number of friends")
for userid, numfriends in num_friends_by_id_sorted:
    print("{0} {1} {2}".format(userid, id2name[userid], numfriends))
```

```
# Graph Drawing
maxfriends = num_friends_by_id_sorted[0][1]
node_sizes = []
node_colors = []
for (uid, numfriends) in num_friends_by_id:
    node_sizes.append(math.pow(2, numfriends)*1000/maxfriends)
    if numfriends == maxfriends:
        node_colors.append('r')
    else:
        node_colors.append('c')

G=nx.Graph()
G.add_edges_from(friendship_pairs)
Gt = nx.relabel_nodes(G, id2name)
nx.draw_kamada_kawai(Gt, with_labels=True, node_size=node_sizes,
                      node_color=node_colors)
plt.savefig('key_connectors.png')
plt.show()
```