GRAPH DB'S & APPLICATIONS

DENIS VRDOLJAK | GUNNAR KLEEMANN

UC BERKELEY SCHOOL OF INFORMATION
BERKELEY DATA SCIENCE GROUP, LLC

Presentation Road Map

- Intro
- Background
- Examples
- Our Work
- Graph Databases

Intro

Background

Examples

Our Work

Graph Databases

About Us: BDSG

Berkeley Data Science Group

Founded by UC Berkeley Data Science instructors and alumni with the goal of bringing Berkeley data science projects to market and commercializing Berkeley Data Science research.

About Us: the Speakers

Denis Vrdoljak

denis@bds.group

dvrdolja@cisco.com

Gunnar Kleemann, PhD

gunnar@bds.group

gunnarkl@berkeley.edu

Why Graphs?

Why Graph Databases?

GRAPH DB'S OPTIMIZED FOR RELATIONSHIPS

- Graph Databases store data in tables/rows/columns, just like a traditional RDBMS
- First Class Citizen is a relationship, not an entity
- Graph DB's are optimized for graph traversals
- This also makes them slow at data retrieval
- But, they're a LOT faster at traversing the nodes of a graph!

Intro

Background

Examples

Our Work

Graph Databases

WHAT IS GRAPH THEORY?

Dates back to 1736: Seven Bridges of Königsberg, by Leonhard Euler

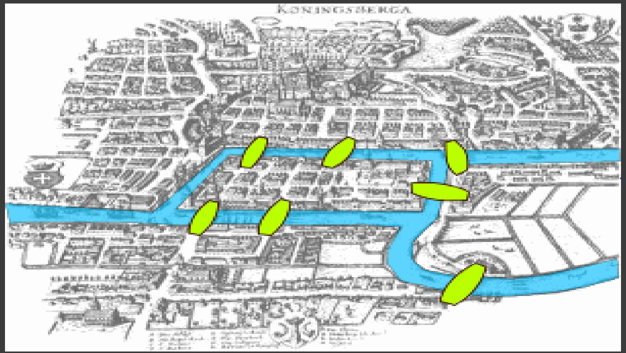
Laid down the original groundwork for what became Graph Theory

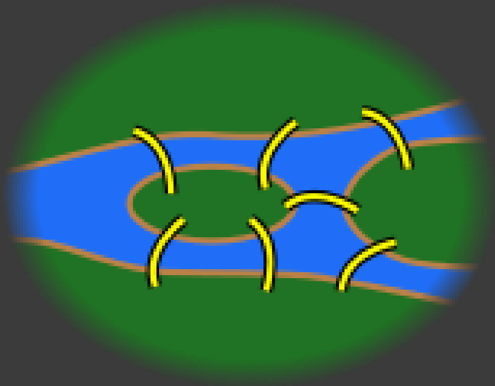
Also evolved into modern day Network Analysis (or Network Graph Analysis) and Social Network Analysis (SNA)

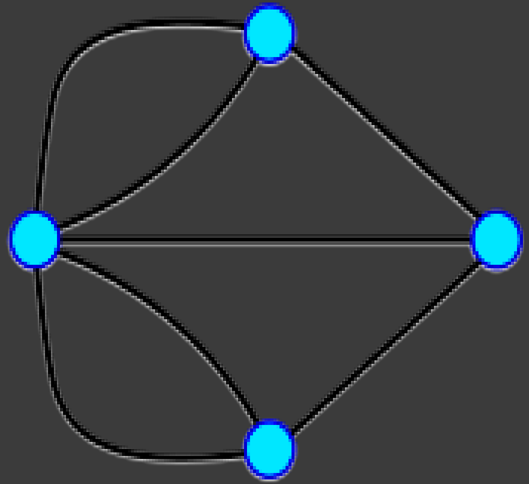
THE SEVEN BRIDGES OF KONIGSBERG

Problem Statement: Can you visit each part of the town, using each bridge only once?

LEONHARDT EULER CAME UP WITH A NEW WAY OF THINKING ABOUT THE PROBLEM, AND IN TURN BECAME THE FATHER OF MODERN GRAPH THEORY







RECOMMENDER SYSTEMS: TRADITIONAL VS. GRAPH

Traditional:

- At Scale
- Production Deployments

Graph-Based:

- Exploratory
- Highly Contextual
- Known Rules

Intro
Background
Examples
Our Work
Graph Databases

Applications of Graph Theory

- Social network analysis
- Map / GPS algorithms - shortest distance between two points, etc.
- AI algorithms
- Search engine algorithms

EXAMPLES OF GRAPH APPLICATIONS

- 9/11 Terrorist Network
- London Phone Network
- Enron Emails
- Panama Papers

Intro

Background

Examples

Our Work

Graph Databases

Keyword Rec:

an HR Keyword Assistant

CONSIDER A TYPICAL JOB REQUISITION

Job Title

- topic information

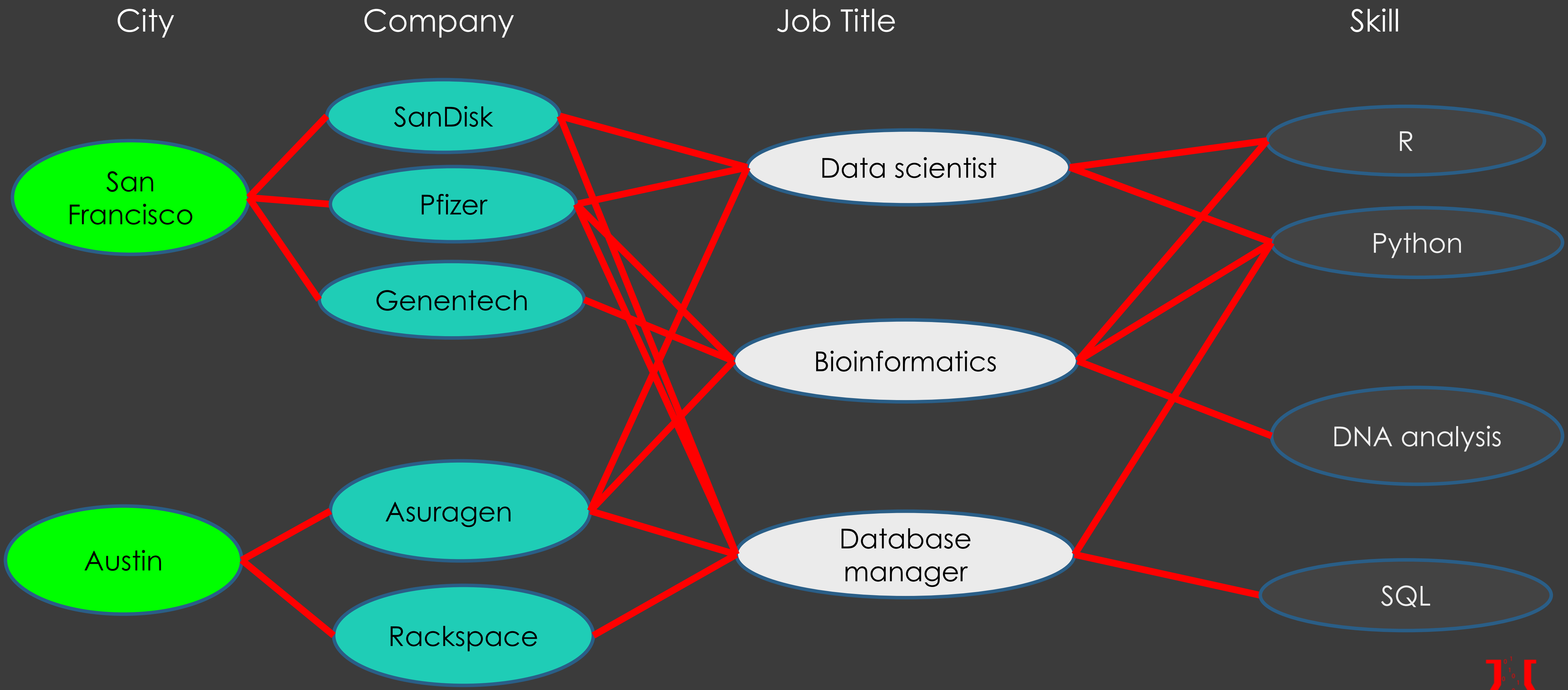
Company and institute

- workplace
- geographic location

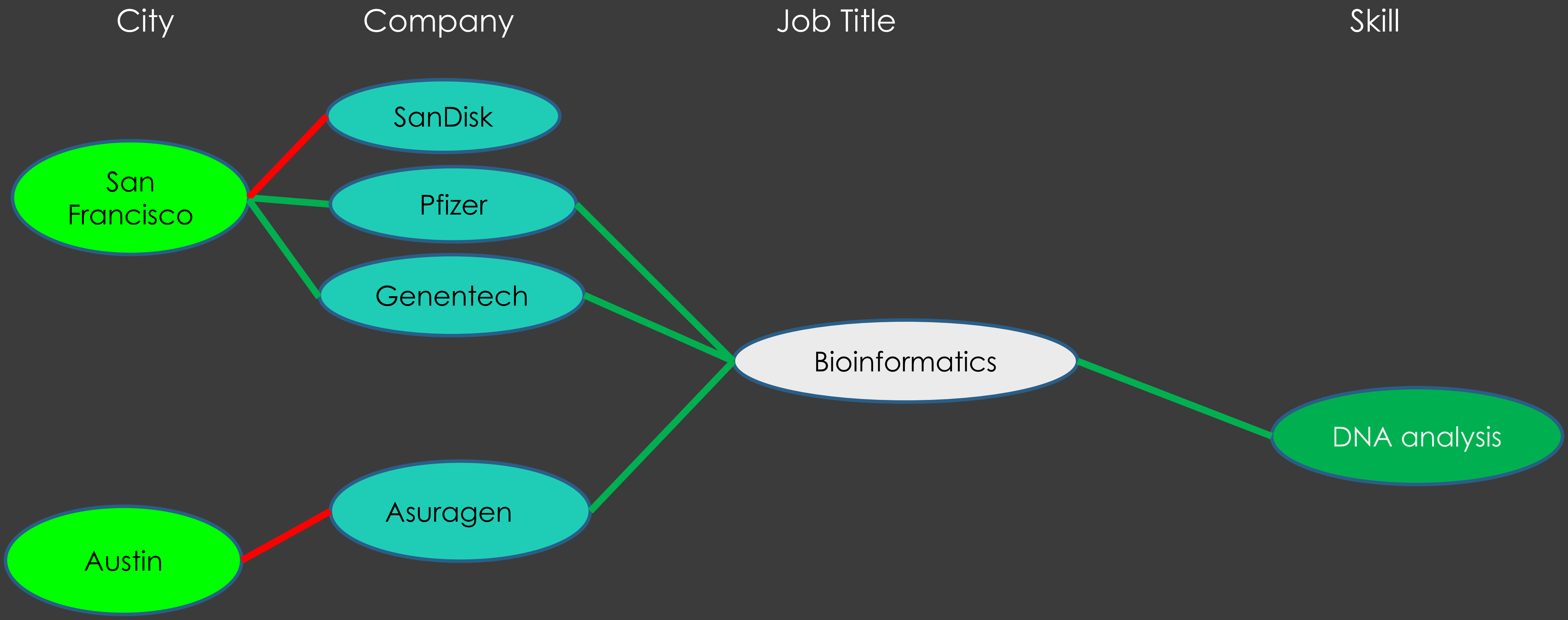
Keywords

- skills
- experience/focus

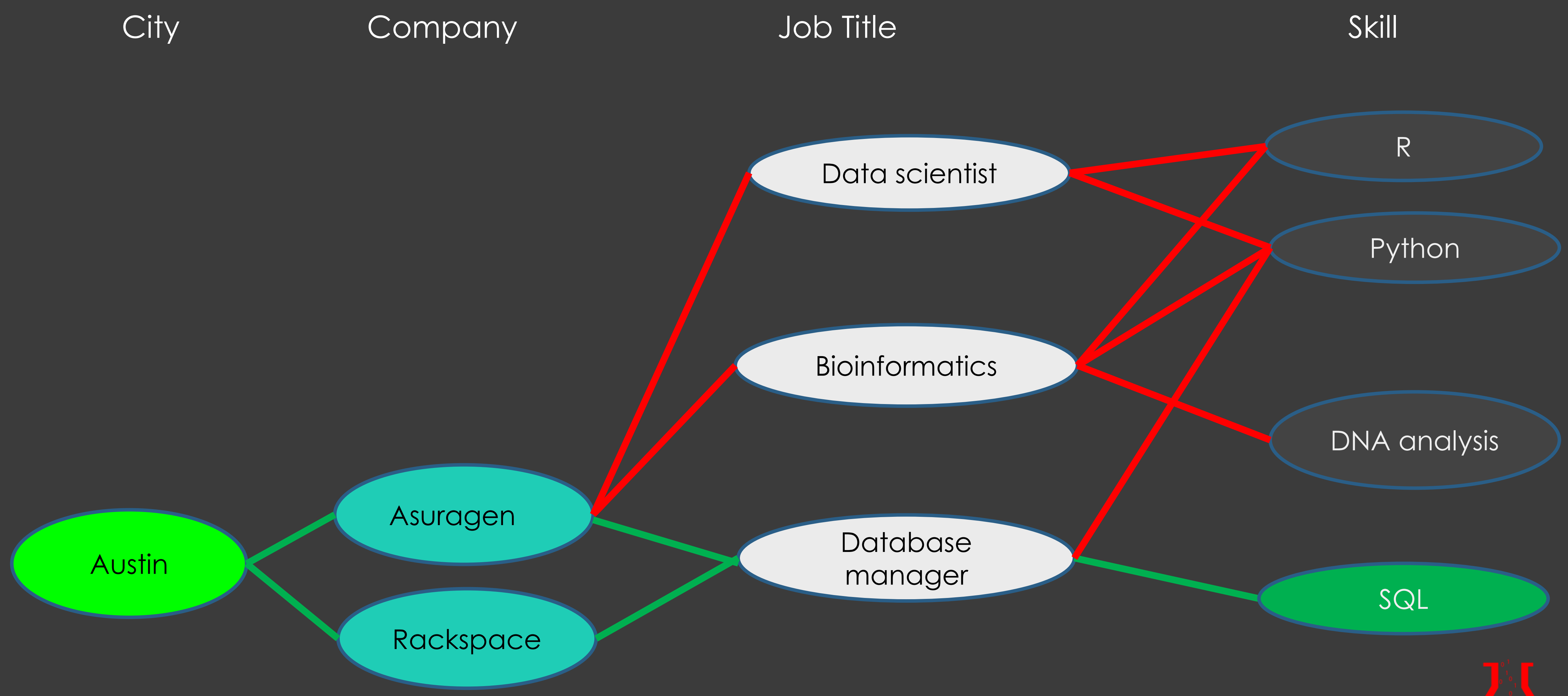
JOB REQUISITION ANALYSIS BY SKILLS



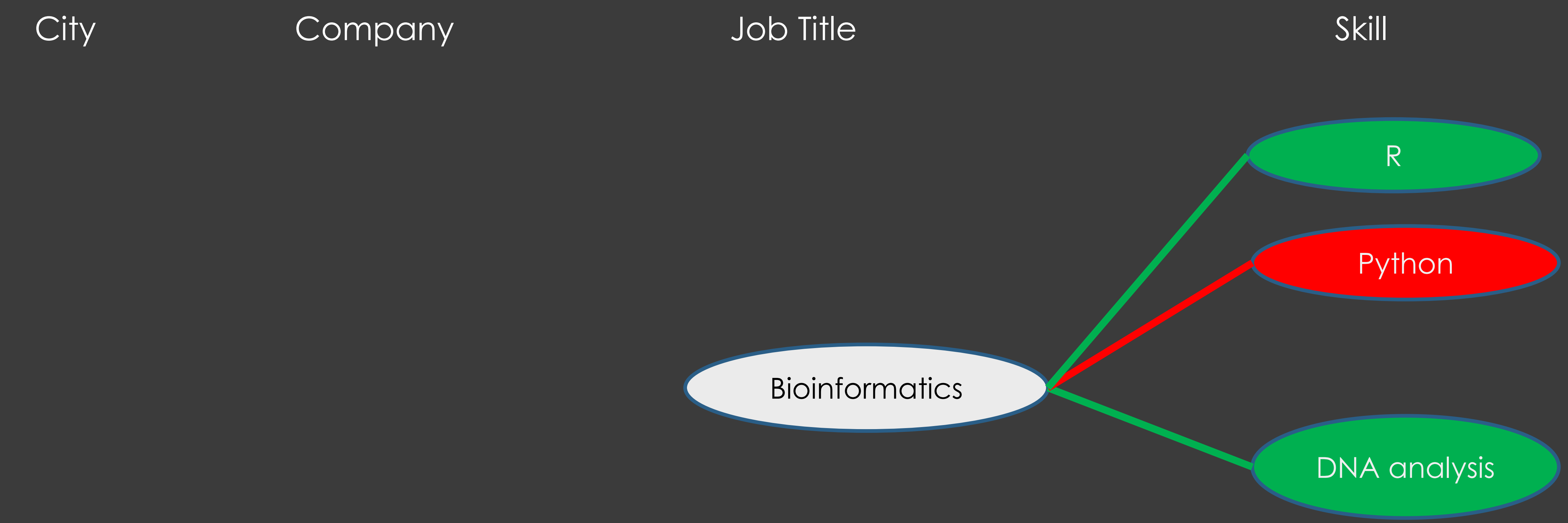
LOOKING AT A SKILL CAN SUGGEST A GEOGRAPHIC HUB



LOOKING AT A REGION WE MIGHT SUGGEST CRUCIAL SKILLS



WE CAN SUGGEST DEFICIENT SKILLS



THE FINAL API

```
relatedskills('Python','SV',5)
```

```
[('python', 328),
 ('similar', 137),
 ('unix', 136),
 ('programming experience', 21),
 ('language', 15)]
```

```
relatedskills('Python','NY',5)
```

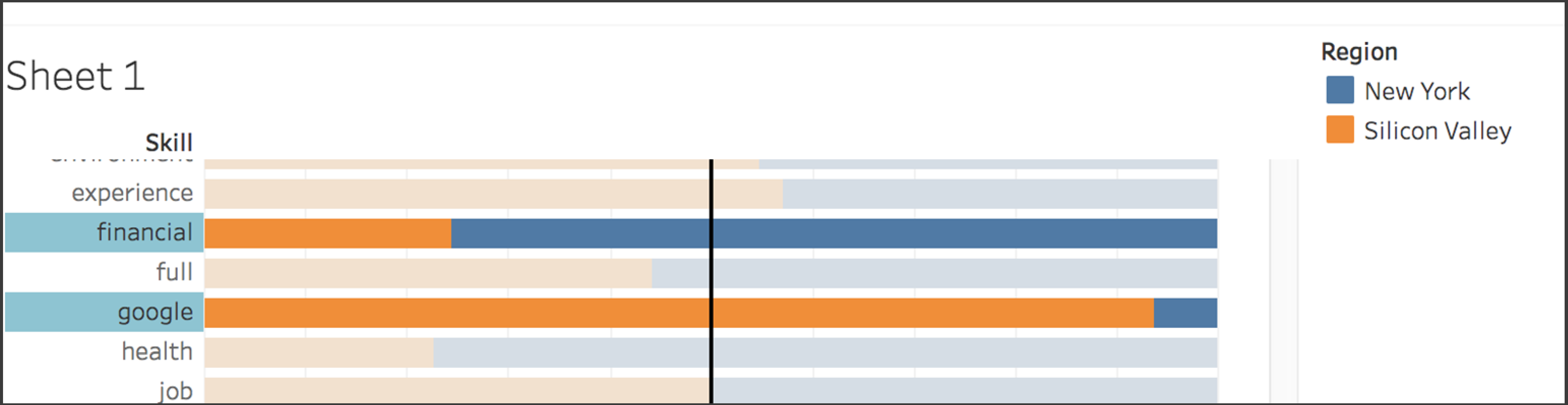
```
[('plus', 636),
 ('programming experience', 615),
 ('python', 383),
 ('sql', 324),
 ('r', 320)]
```

SF vs NY:

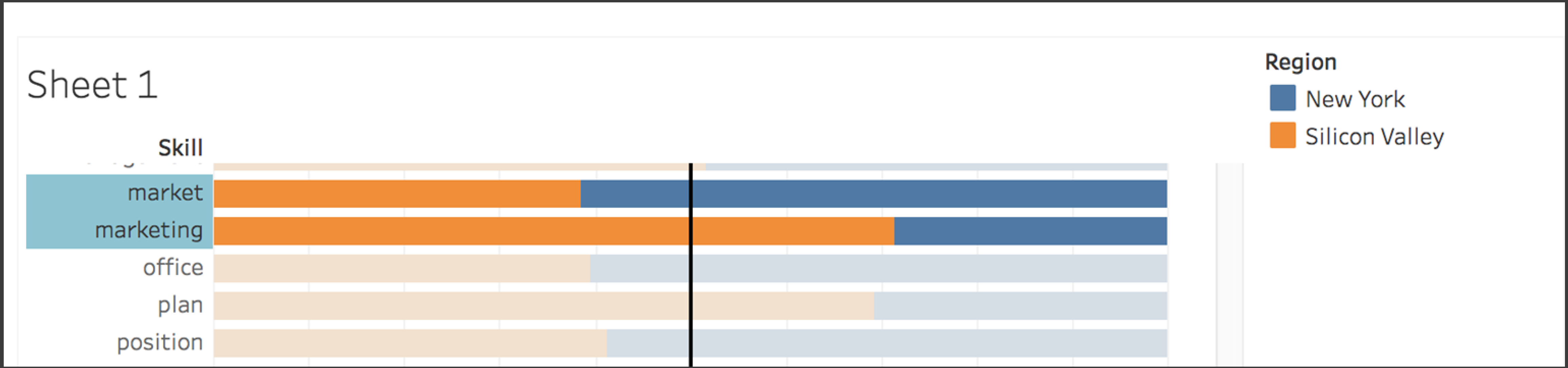
Who has better data scientists?

https://public.tableau.com/profile/denis.vrdoljak#!/vizhome/SVvsNY_JobSkillsPercentCompare/Sheet3?publish=yes

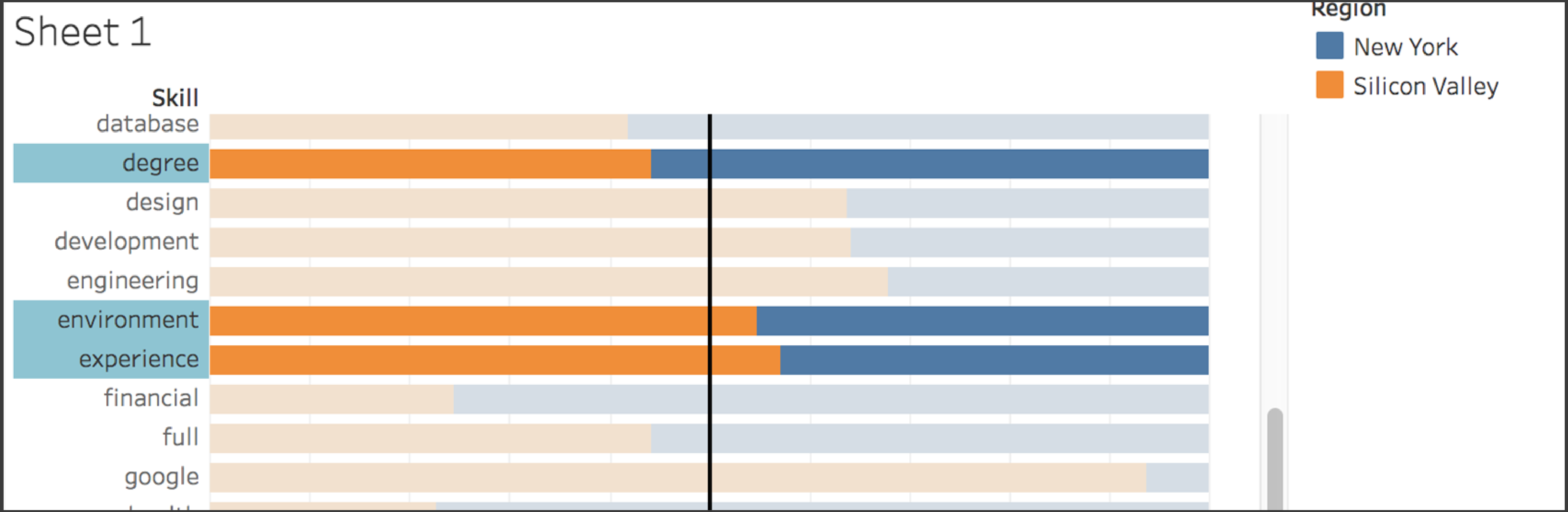
Who has Better Data Scientists?



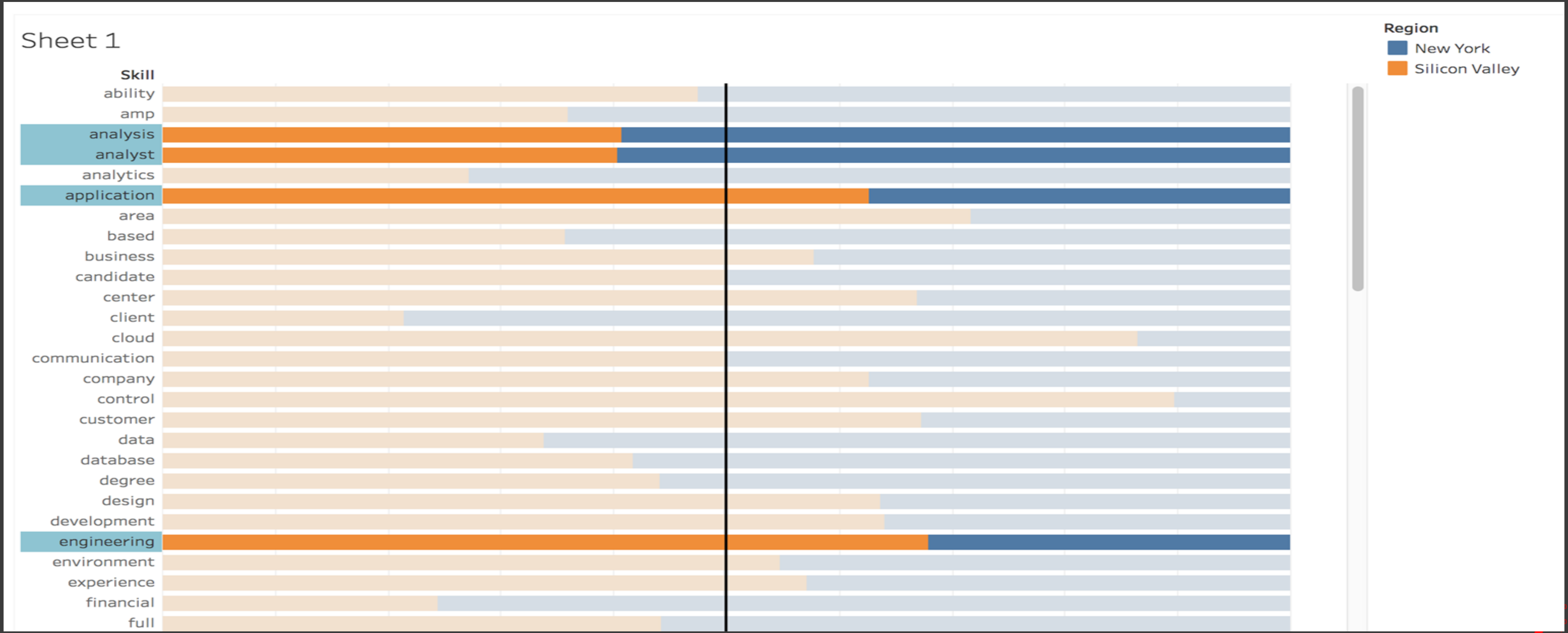
Who has Better Data Scientists?



Who has Better Data Scientists?



Who has Better Data Scientists?



…now let’s take a look back to our very first project, and what we’ve done with it since!

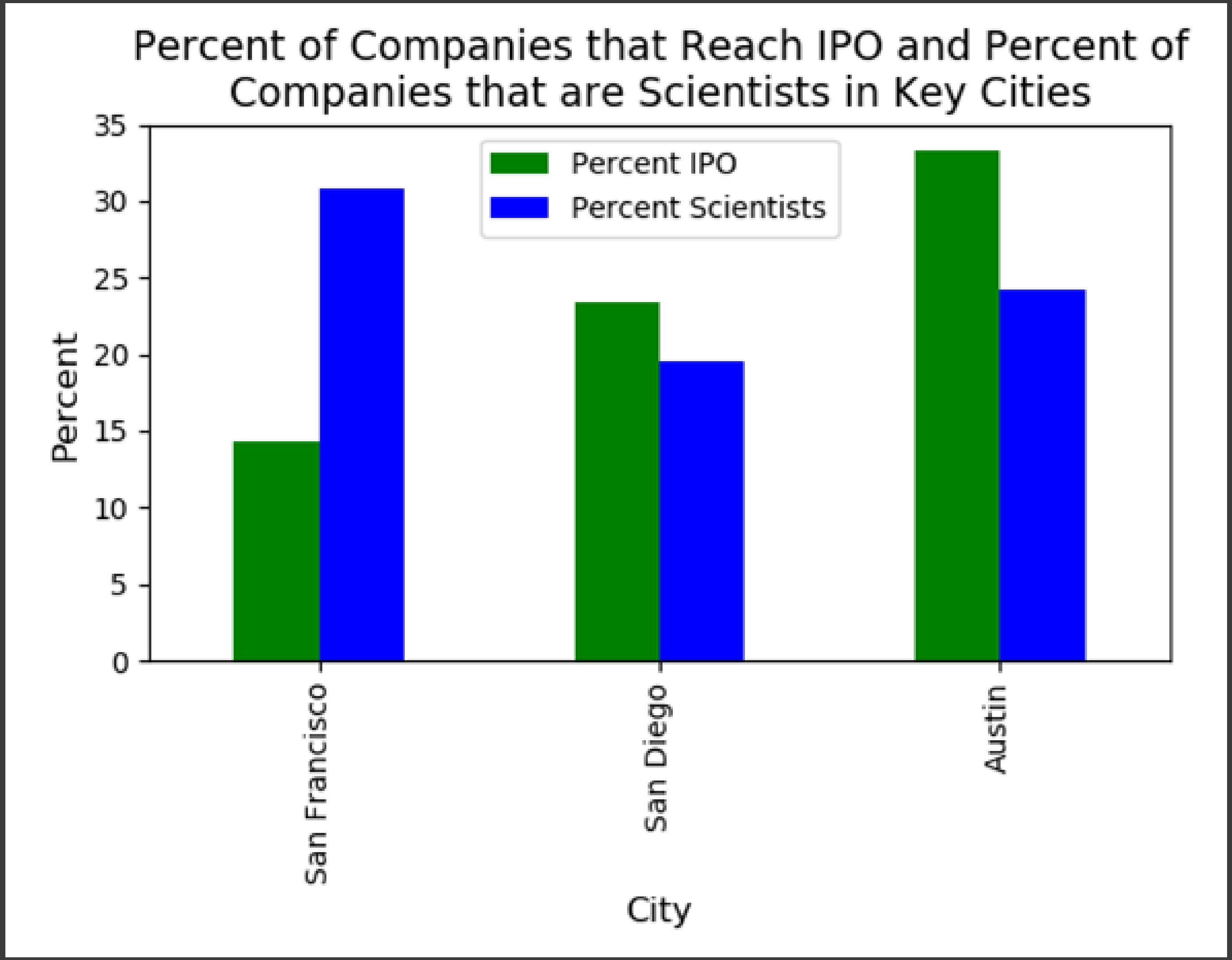
BioRevs:

Predicting Biotech IPO Rates through Collaboration Networks

CONDUCTING A JOB SEARCH IN BIOTECH

YOU COULD IDENTIFY BIOTECHNOLOGY HUBS BY COUNTING COMPANIES

BIOTECH HUBS CAN BE PROFILED BY IPO AND % SCIENTISTS



Percent IPO = % Companies reaching IPO in < 6000 days

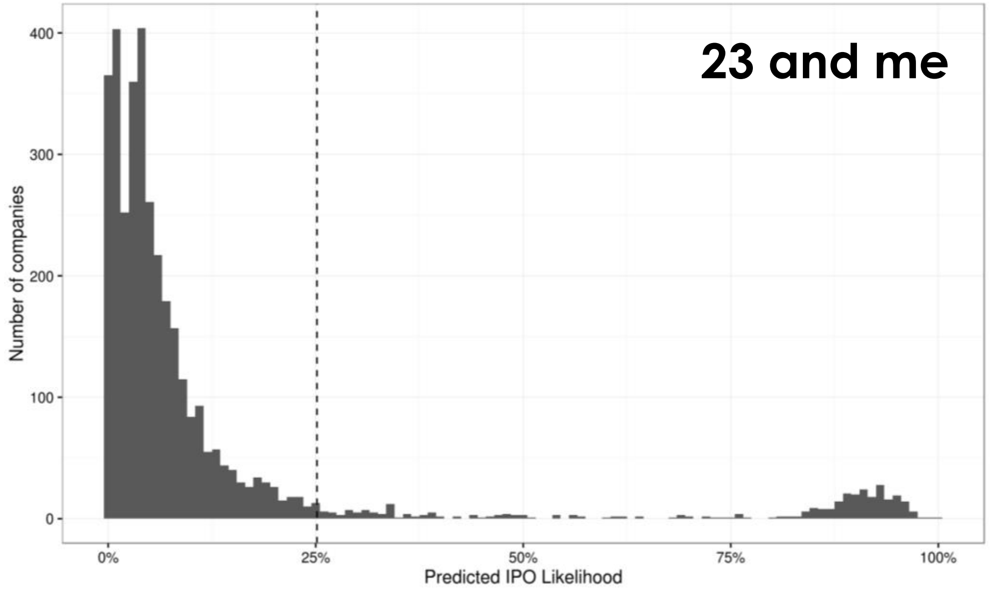
National average is ~16% reach IPO.

National average is ~21% Scientists.

BASED ON TRADITIONAL MACHINE LEARNING ANALYSIS WE BUILT BIOREVS

Success Likelihood

IPO Likehood:25.2%

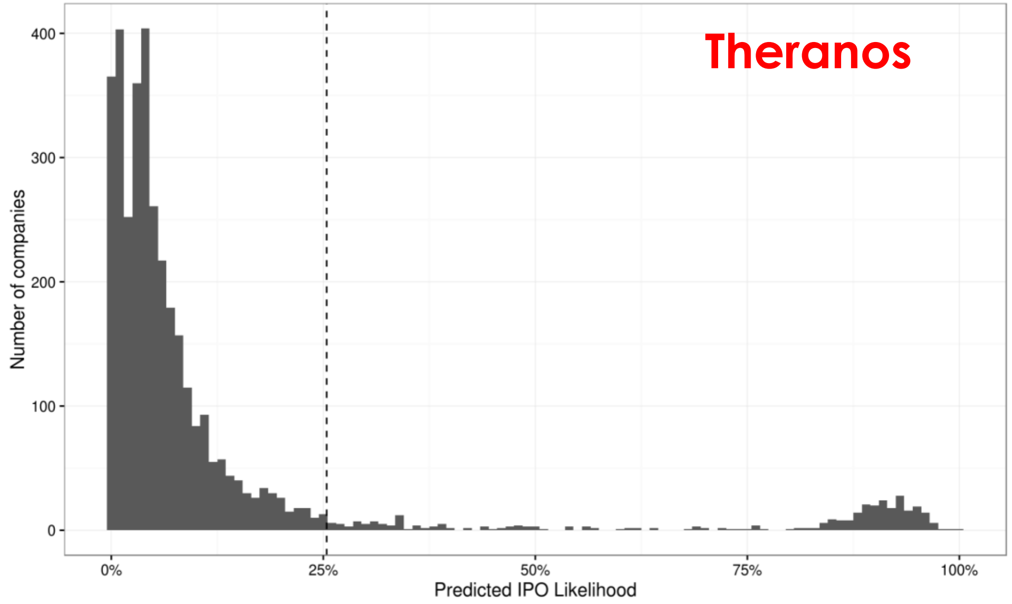


Expected Valuation

Forecasted IPO Amount: $521,812,444

Success Likelihood

IPO Likehood:25.4%



Expected Valuation

Forecasted IPO Amount: $133,861,306

* Data up to 2013

WE CAN USE A GRAPH TO GET MORE

PUBLICATIONS CAN BE PARSED INTO IMPORTANT DATA

Journal -

- Discipline
- Audience
- Impact

ANALYSIS

nature genetics

Understanding multicellular function and disease with human tissue-specific networks

Casey S Greene[1–3,13], Arjun Krishnan[4,13], Aaron K Wong[5,13], Emanuela Ricciotti[6,7], Rene A Zelaya[1], Daniel S Himmelstein[8], Ran Zhang[9], Boris M Hartmann[10], Elena Zaslavsky[10], Stuart C Sealfon[10], Daniel I Chasman[11], Garret A FitzGerald[6,7], Kara Dolinski[4], Tilo Grosser[6,7] & Olga G Troyanskaya[4,5,12]

Tissue and cell-type identity lie at the core of human physiology and disease. Understanding the genetic underpinnings of complex tissues and individual cell lineages is crucial for developing improved diagnostics and therapeutics. We present genome-wide functional interaction networks for 144 human tissues and cell types developed using a data-driven Bayesian methodology that integrates thousands of diverse experiments spanning tissue and disease states. Tissue-specific networks predict lineage-specific responses to perturbation, identify the changing functional roles of genes across tissues and illuminate relationships among diseases. We introduce NetWAS, which combines genes with nominally significant genome-wide association study (GWAS) *P* values and tissue-specific networks to identify disease-gene associations more accurately than GWAS alone. Our webserver, GIANT, provides an interface to human tissue networks through multi-gene queries, network visualization, analysis tools including NetWAS and downloadable networks. GIANT enables systematic exploration of the landscape of interacting genes that shape specialized cellular functions across more than a hundred human tissues and cell types.

The precise actions of genes are frequently dependent on their tissue context, and human diseases result from the disordered interplay of tissue- and cell lineage–specific processes[1–4]. These factors combine to make the understanding of tissue-specific gene functions, disease pathophysiology and gene-disease associations particularly challenging. Projects such as the Encyclopedia of DNA Elements (ENCODE)[5] and The Cancer Genome Atlas (TCGA)[6] provide comprehensive genomic profiles for cell lines and cancers, but the challenge of understanding human tissues and cell lineages in the multicellular context of a whole organism remains[7]. Integrative methods that infer functional gene interaction networks can capture the interplay of pathways, but existing networks lack tissue specificity[8].

Although direct assay of tissue-specific features remains infeasible in many normal human tissues, computational methods can infer these features from large data compendia. We recently found that even samples measuring mixed cell lineages contain extractable information related to lineage-specific expression[9]. In addition to tissue specificity, we[10–13] and others[14–17] have shown that heterogeneous genomic data contain functional information, for example, of gene expression regulation by protein-DNA, protein-RNA, protein-protein and metabolite-protein interactions. Here we develop and evaluate methods that simultaneously extract functional and tissue or cell-type signals to construct accurate maps of both where and how proteins act.

We build genome-scale functional maps of hum by integrating a collection of data sets thousands of experiments contained in more ,000 distinct publications. To integrate these automatically assess each data set for its relevan of 144 tissue- and cell lineage–specific functional texts. The resulting functional maps provide a detailed portrait of protein function and interactions in specific human tissues and cell lineages ranging from B lymphocytes to the renal glomerulus and the whole brain. This approach allows us to profile the specialized function of genes in a high-throughput manner, even in tissues and cell lineages for which no or few tissue-specific data exist.

In contrast with tissue-naive networks, which assume that the func-

[1]Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA. [2]Dartmouth-Hitchcock Norris Cotton Cancer Center, Lebanon, New Hampshire, USA. [3]Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, New Hampshire, USA. [4]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA. [5]Department of Computer Science, Princeton University, Princeton, New Jersey, USA. [6]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [7]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [8]Biology and Medical Informatics, University of California, San Francisco, San Francisco, California, USA. [9]Department of Molecular Biology, Princeton University, Princeton,

Title -

- topic information

Collaborator list

- professional relationships

Company and institute

- workplace
- geographic location

PUBMED DATABASE
THE WORLD'S BIOMEDICAL RESEARCH

24,000,000

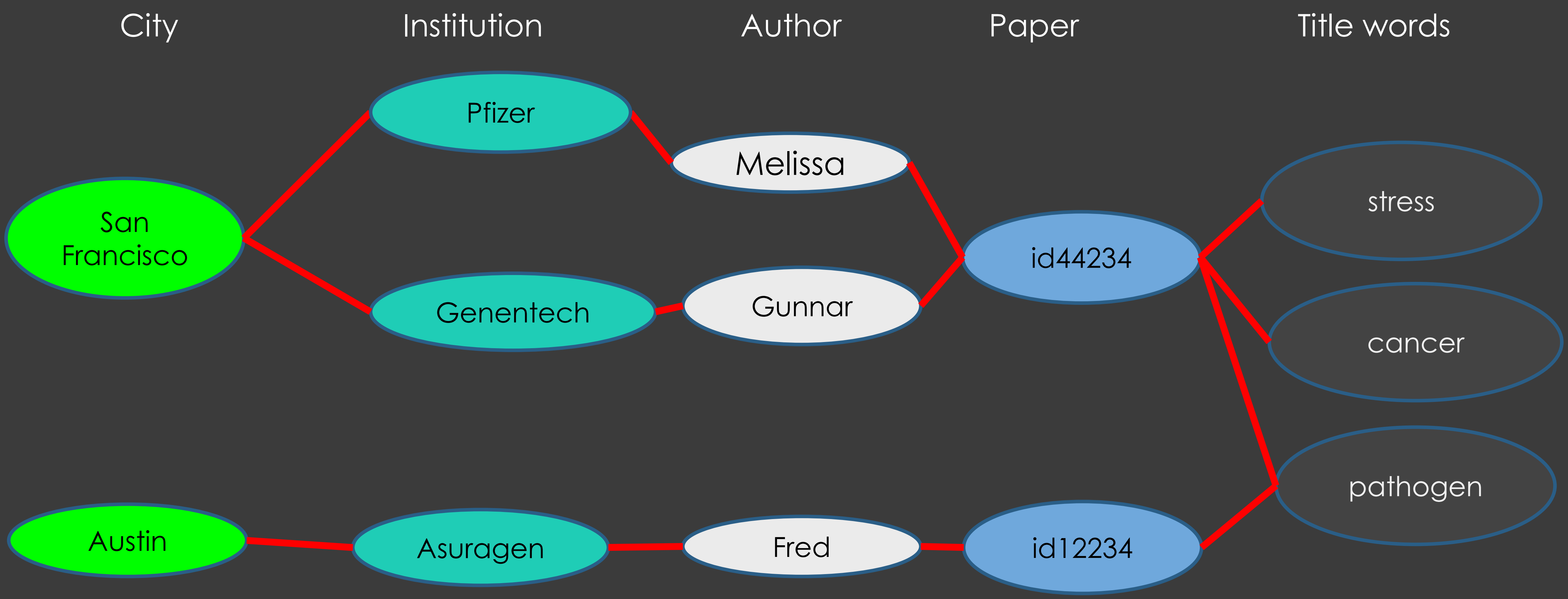
FULL PUBMED DATASET (ALL BIOMEDICAL LITERATURE)

~ 1.5 MILLION OPEN SOURCE

PUBMED CENTRAL (BIOTECH-OPEN ACCESS SUBSET)

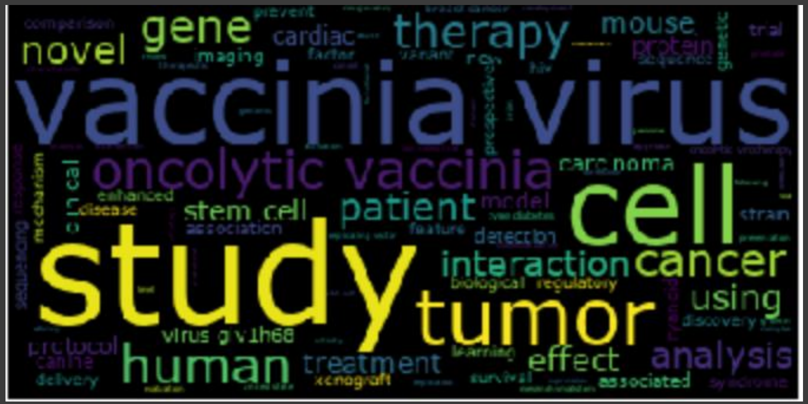
DATA INCLUDE: PUBLICATION TITLES, AUTHORS, AND OTHER METADATA

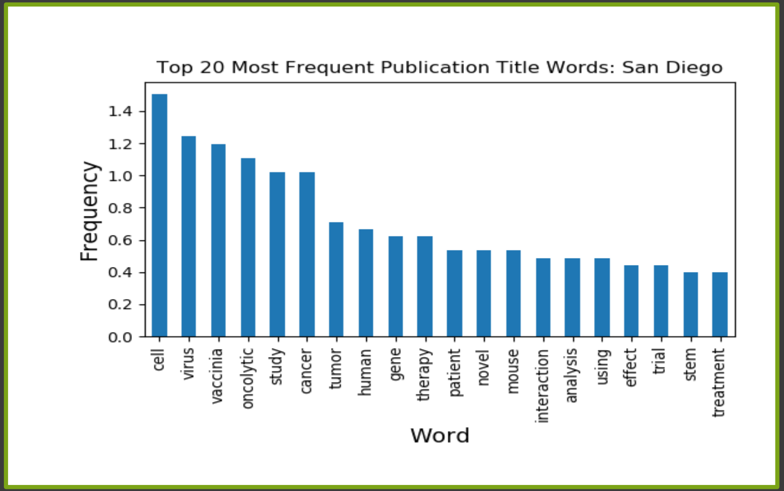
PUBLICATION AND COMPANY DATA IN A GRAPH



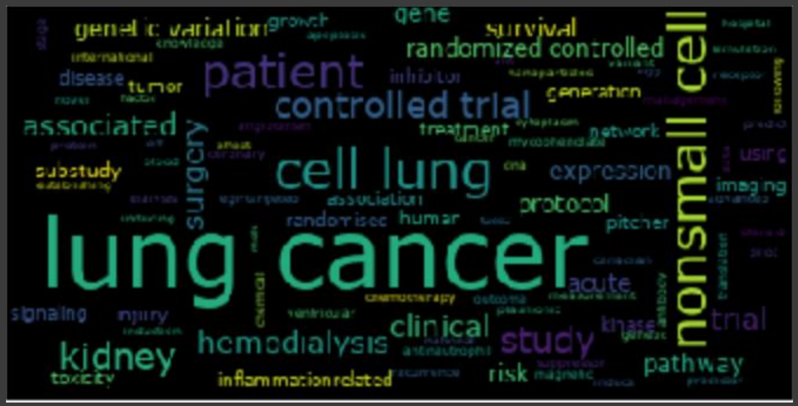
CITIES DIFFER IN SCIENTIFIC EXPERTISE

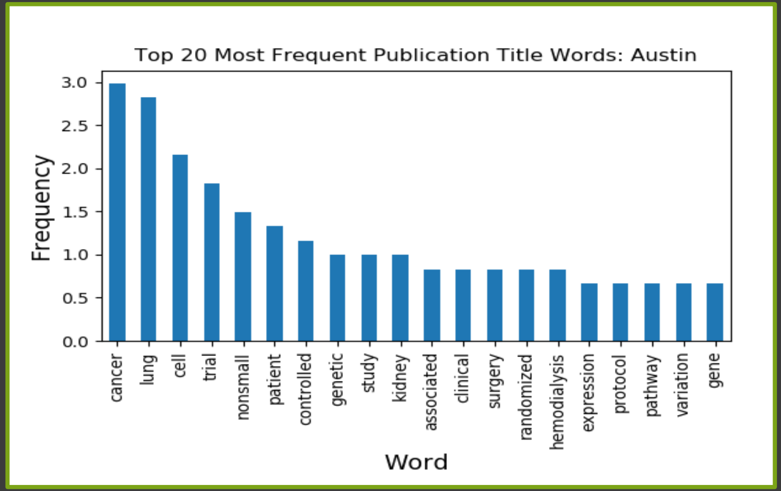
San Diego



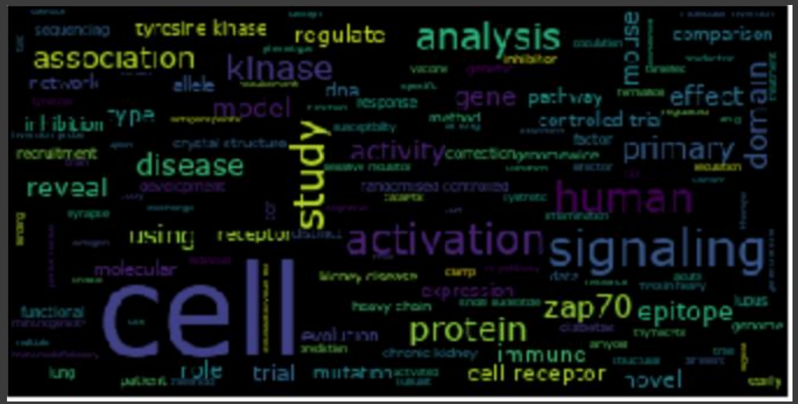


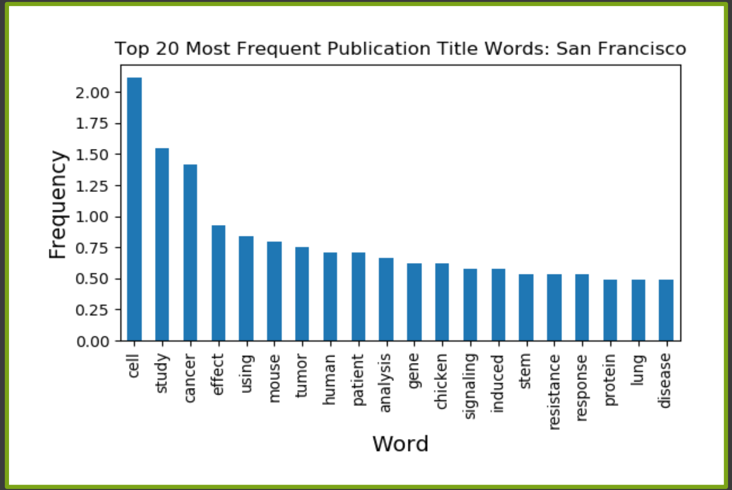
Austin





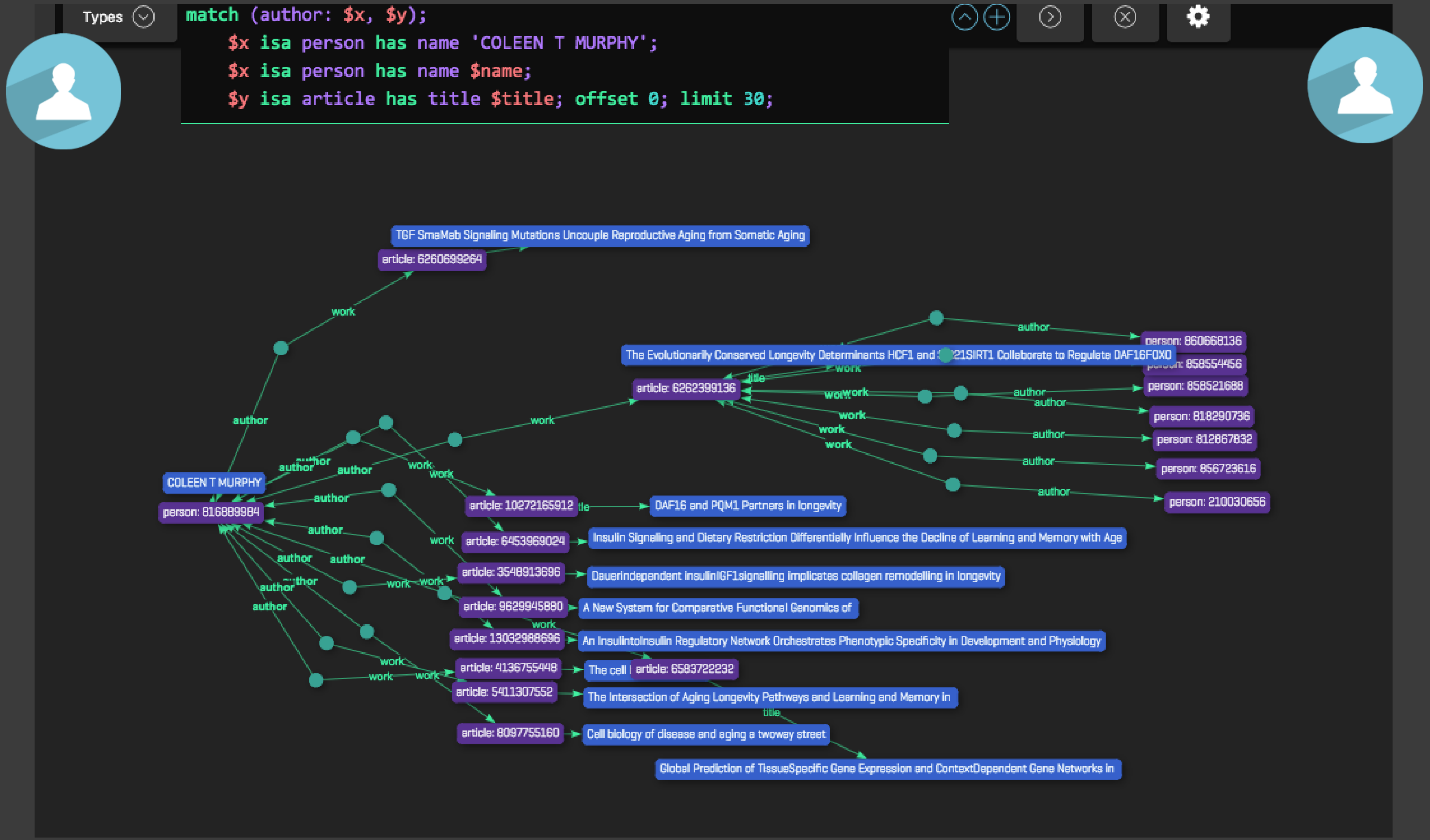
San Francisco





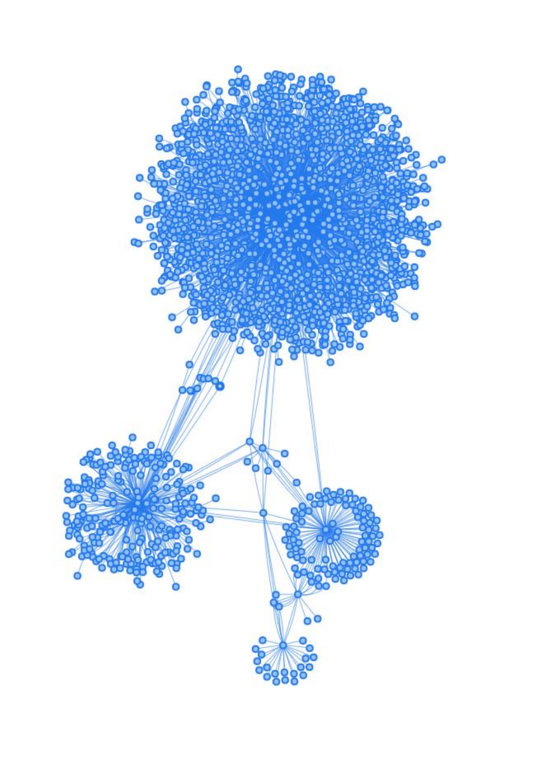
QUANTIFICATION OF SCIENCE NETWORKS WITH COLLABORATION GRAPHS

WE CAN GET THE COLLABORATION NETWORK TOPOLOGY FROM THE GRAPH

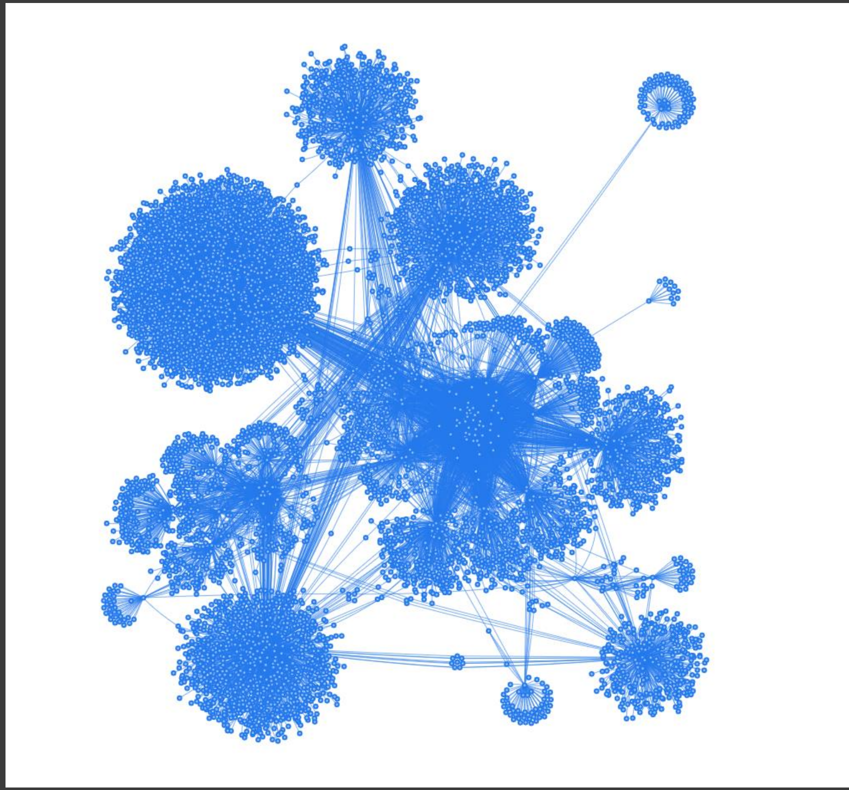


SOME TYPICAL PUBLICATION PATTERNS

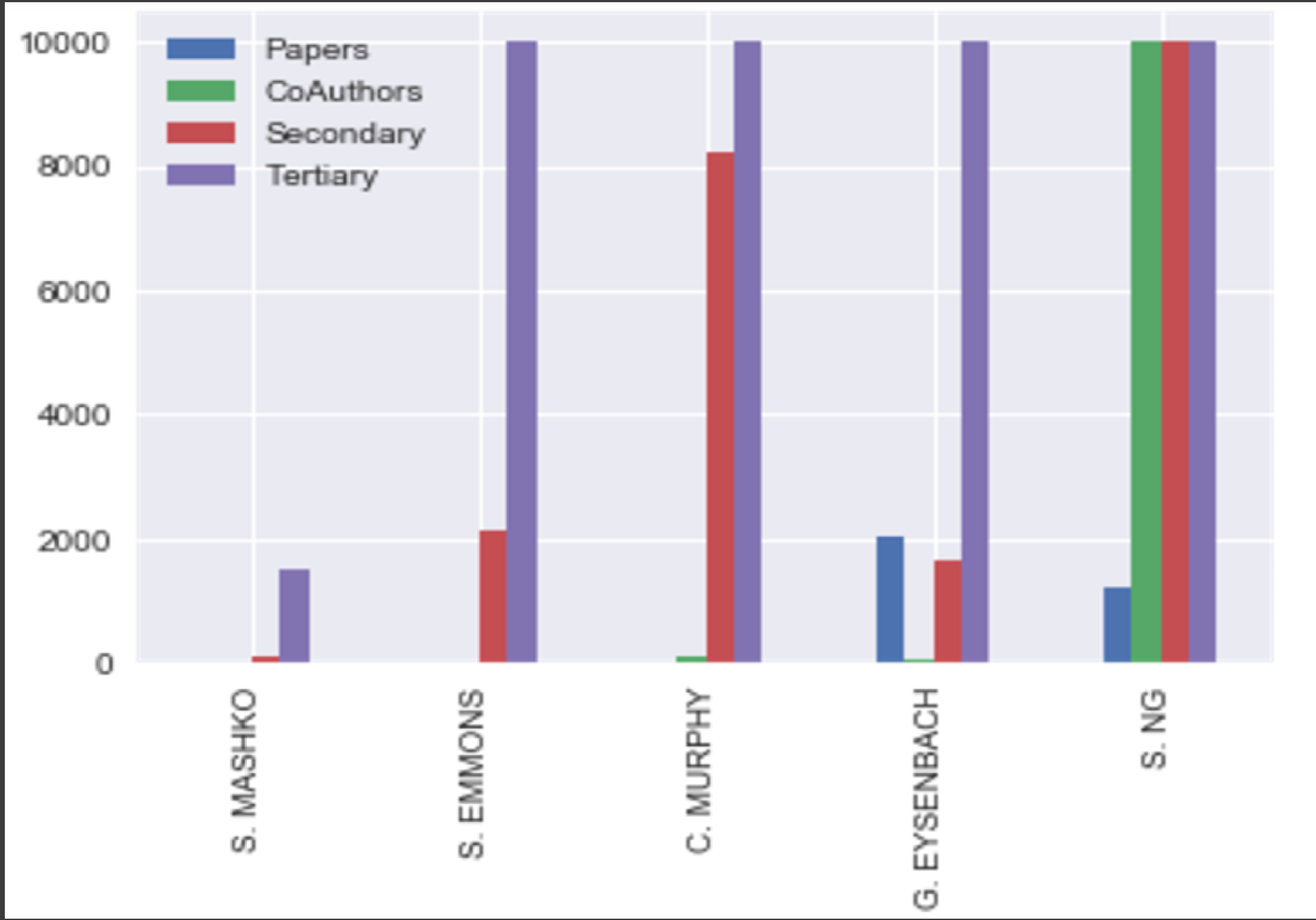
SCOTT EMMONS' 2ND DEGREE COLLABORATION NETWORK



COLEEN MURPHY'S 2ND DEGREE COLLABORATION NETWORK



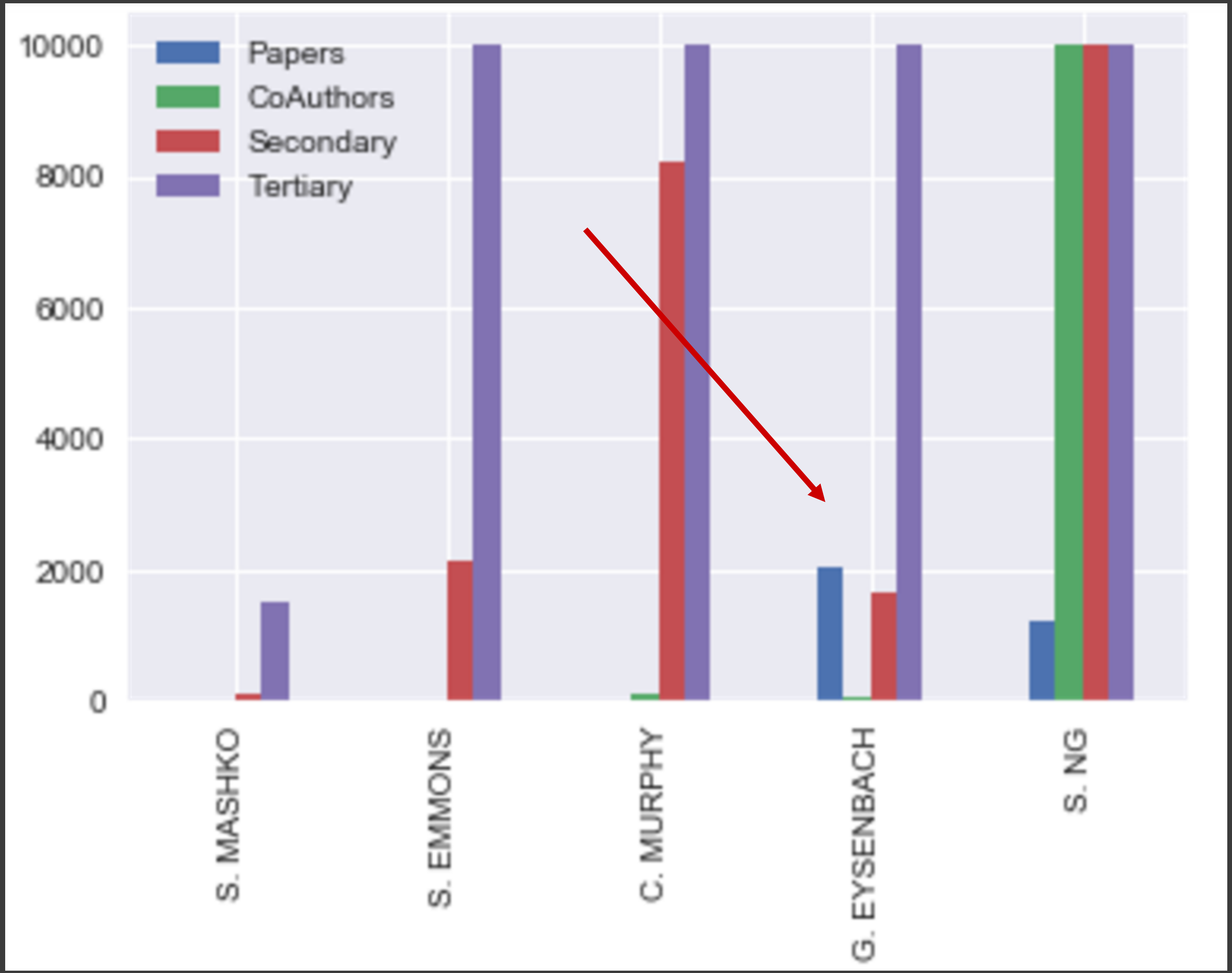
INCREASING NODE COUNT WITH DISTANCE FROM AUTHOR



PAPERS < COAUTHORS < SECONDARY < TERTIARY

WE CAN SEE AN OUTLIER!

PAPERS <
COAUTHORS <
SECONDARY <
TERTIARY



A BIG PUBLISHER WITH 2055 LINKS!

"ONE OF THE MOST PRODUCTIVE RESEARCHERS, EDITORS, AND PUBLISHERS IN THE ONLINE HEALTH FIELD."

IN 2004 RECEIVED THE JANSSEN-CILAG FUTURE AWARD, REFERRED TO AS THE GERMAN "HEALTH CARE NOBEL PRIZE".

FOUNDER OF AN ACADEMIC FIELD!

ASSOCIATION BETWEEN SEARCH ENGINE QUERIES AND INFLUENZA INCIDENCE,

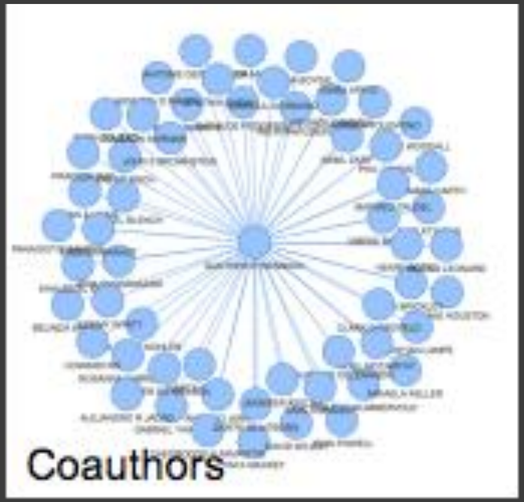
HE COINED THE TERMS "INFOVEILLANCE" AND "INFODEMIOLOGY" FOR THESE KINDS OF APPROACHES.

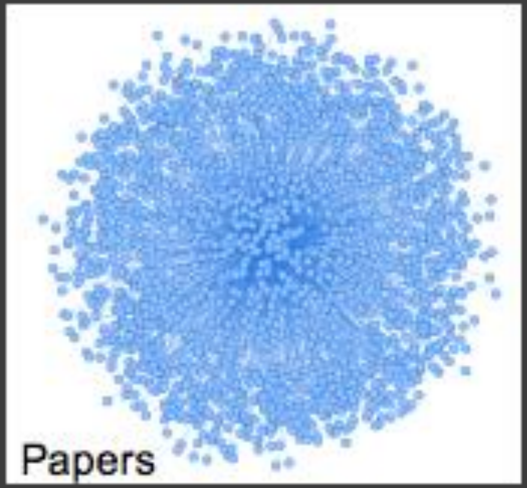
FEW CO-AUTHORS AND MANY PAPERS SUSPICIOUS PATTERN

BUT HE IS ONLY CITED FOR 120 PAPERS AND 40 BOOK CHAPTERS.

SO WHAT ARE THOSE OTHER 1900 LINKS ?

PROBABLY EDITING JOBS (FALSE POSITIVES)





Intro
Background
Examples
Our Work

Graph Databases

SOME KEY GRAPH DATABASES

- NEO4J (WELL SUPPORTED)
- TITAN/JANUS GRAPH (DISTRIBUTED BACKEND)
- AGENSGRAPH (POSTGRES COMPATIBLE)
- GRAKN (KNOWLEDGE GRAPH)

NEO4J

- Industry standard features
- Large userbase and developer community
- Built from the ground up as a graph database

https://neo4j.com/

TITAN/JANUS GRAPH

- Apache project
- Early adopter of distributed backend
- Elastic scalability
- Integration with Tinkerpop graph stack
- Multiple user access
- Real time updates

https://www.predictiveanalyticstoday.com/titan/

AGENSGRAPH

- HIGHLY PERFORMANT GRAPH DATABASE
- HYBRID DATABASE BUILT ON POSTGRESQL
- SQL AND CYPHER IN THE SAME QUERY

http://bitnine.net/agensgraph/?ck
attempt=1

GRAKN.AI

- Knowledge representation in graphs for AI purposes
  - Nodes represent "objects", and edges are relationships between them.
- SQL-type query language, Graql, used to quickly and intuitively make queries in the knowledge graph
- Steadily growing technology

https://grakn.ai/

Thank You!

- Denis Vrdoljak, MIDS
Managing Director, BDSG
Marketing Analyst/Data Scientist, Cisco
- denis@bds.group
- dvrdolja@cisco.com

Gunnar Kleemann, PhD, MIDS
Senior Data Scientist, BDSG
Data Science Instructor, UC Berkeley
- gunnar@bds.group
- gunnarkl@berkeley.edu