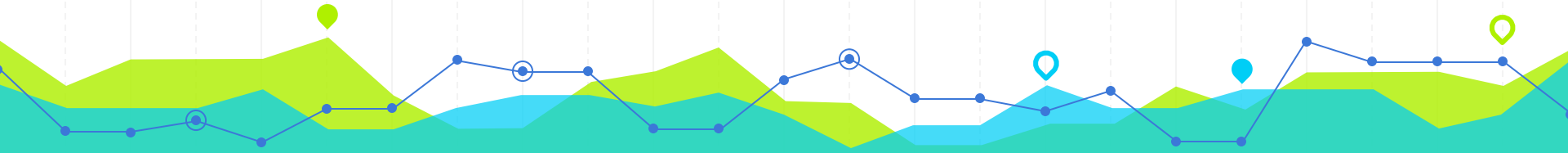
Biorevolutions: Machine Learning and Graph Analysis Illuminate Biotechnology Startup Success



Denis Vrdoljak | Gunnar Kleemann MIDS PhD | Kiersten Henderson PhD

UC Berkeley School of information
Berkeley Data Science Group, LLC
Capital Data Corporation of Austin

Berkeley SCHOOL OF INFORMATION

BDSG

Predict startup success with data science

What variables can be collected before a Series A round to predict a successful exit for a biotech startup, as defined by investor ROI?

- Rationale: it's important for stakeholders to have a quantitative framework to guide their investment decisions.
- Reasoning through analogy and past experiences is valuable, but a quantitative anchor helps identify biases and blind spots

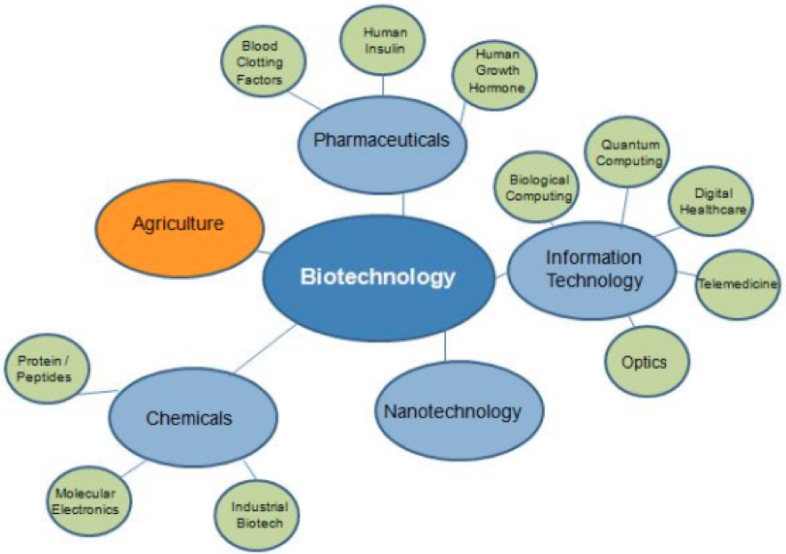


Image Source: http://info.azfb.org/blog/bid/384226/You-Accept-Biotechnology-in-Pharmaceuticals-Why-Not-in-Agriculture

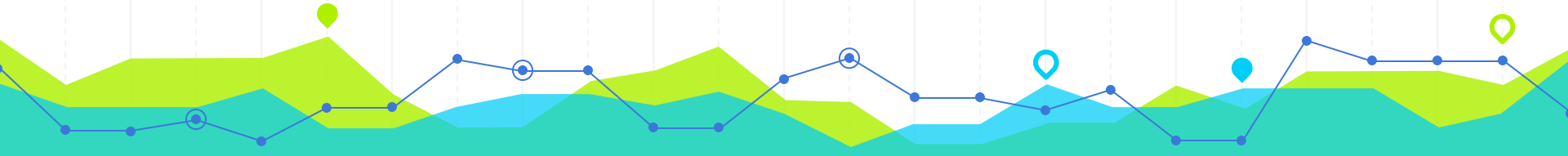
Business Case/Pain Point

Invested Quarterly: $1.5B
Cumulatively Invested (2014): $62B
Gross Returns: $106B

Only 11% reach a liquidation event

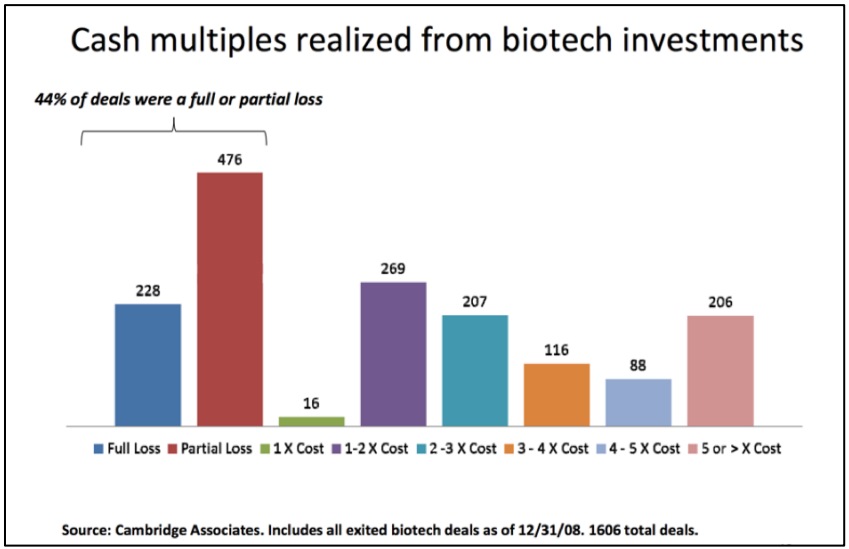
Sources: Forbes (2014) and PWC (2014)

Background



Problem Statement: Which Biotechs are Profitable?

“investing in a small or even medium sized biotech company is a lot like gambling in Las Vegas. You are either going to win big or lose big” - R. Langreth (Forbes online, 2010)



Source: Cockburn and Lerner The Cost of Capital for Early-Stage Biotechnology Ventures
http://buchpedersen.com/wp-content/uploads/2014/03/The-Cost-of-Capital-for-Early-Stage-Biotechnology-Ventures_CockburnLerner.pdf

Will a biotech startup survive

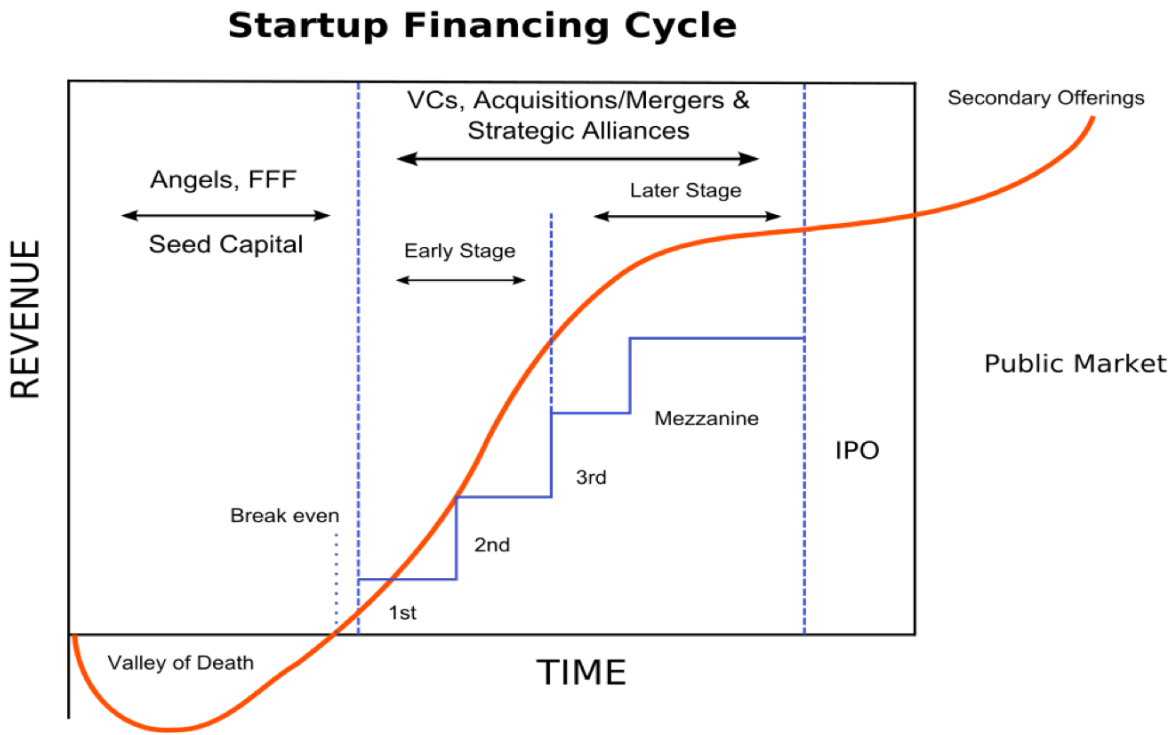
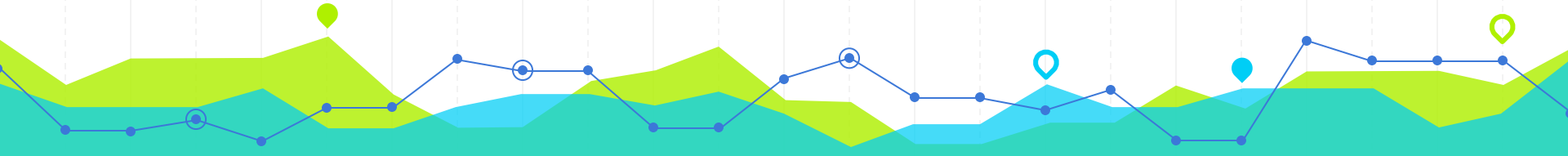


Image Source: http://www.persistentchange.com/startup/implications-financing-biotechnology-startup/

Methods/Techniques



WE HYPOTHESIZE THE RECIPE FOR SUCCESS

The Company

Current Investors
Value of idea
Complementary teams
Team dynamic

The Staff

Experience/Reputation
Staff composition
Prior scientific record
Prior company experience

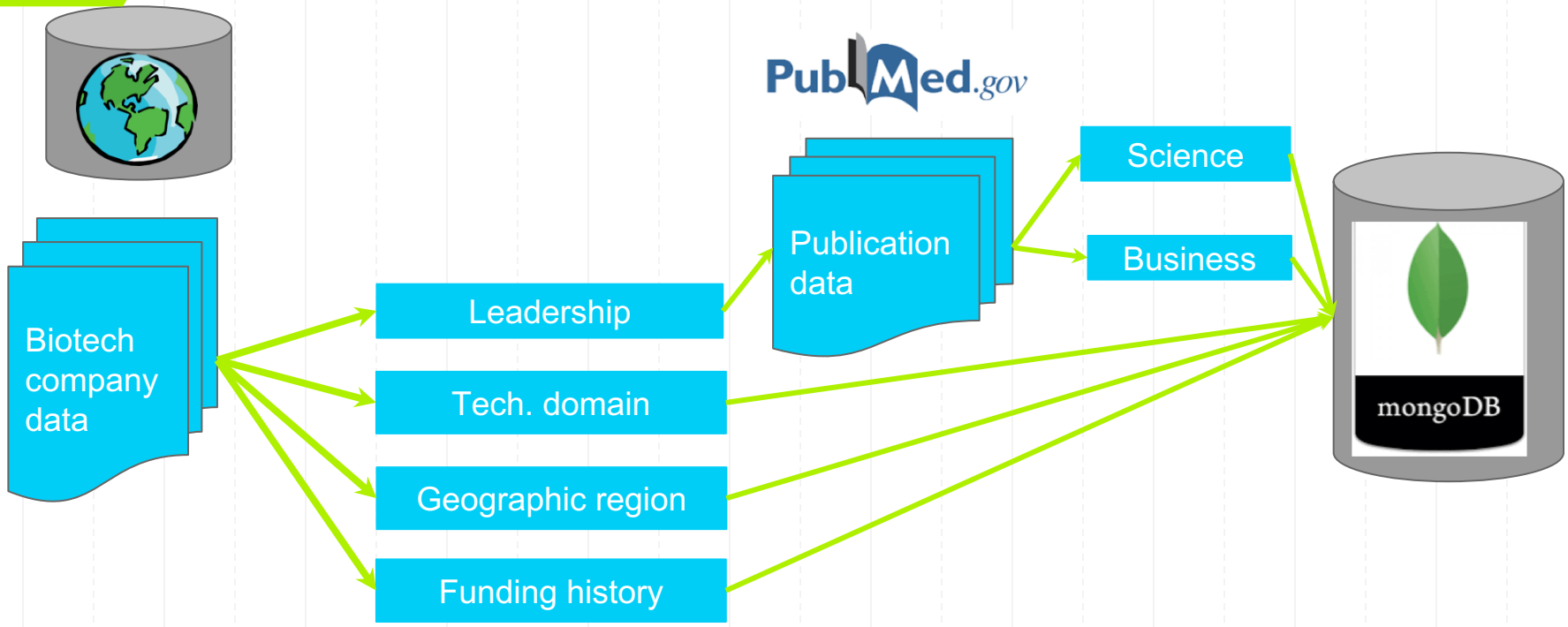
The Market

Market Favorability
Geographic region
Technology domain
Current economic climate

SUCCESS: Definition

- **Reaches liquidation event**
  - **IPO or merger & acquisition**
- Short time to liquidation
- Positive return on investment (ROI)
  - $ liquidation / $ startup

Collect

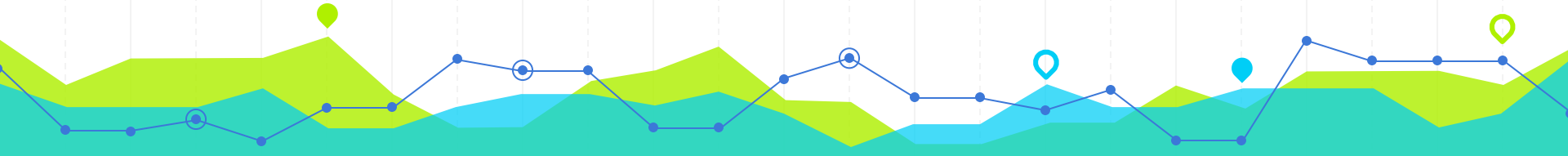

PubMed.gov
Biotech company data
Leadership
Tech. domain
Geographic region
Funding history
Publication data
Science
Business
mongoDB

Collect

Process

Results/Trends

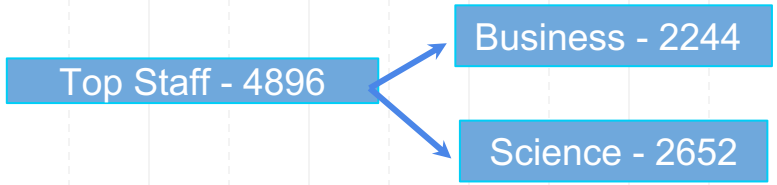
(is there a signal?)



Collect

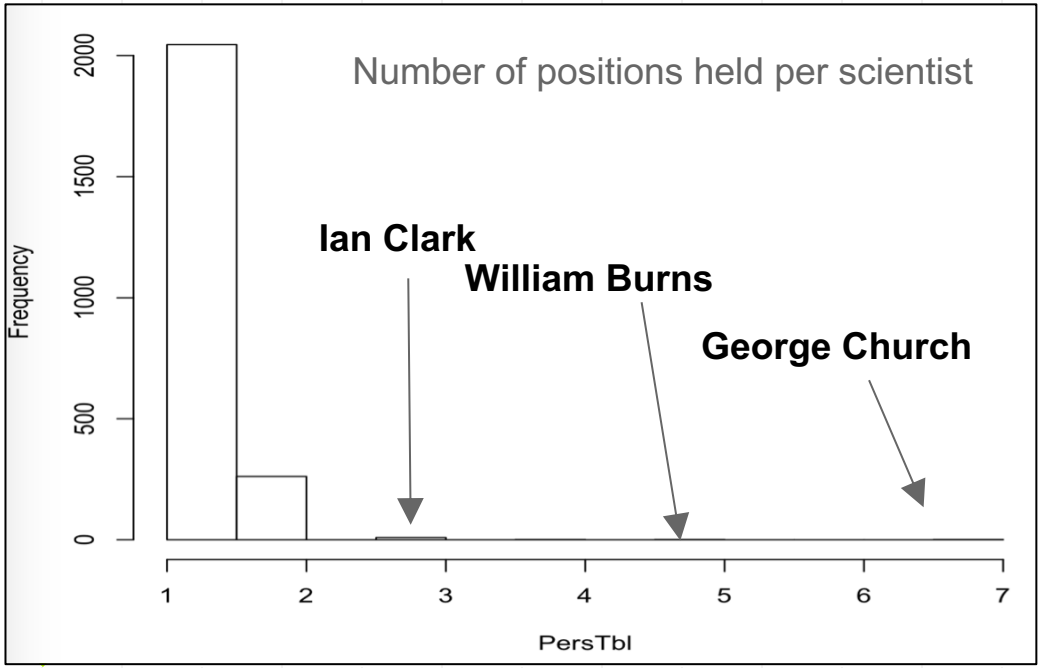
Process

Biotech and scientists



(in Pubmed)

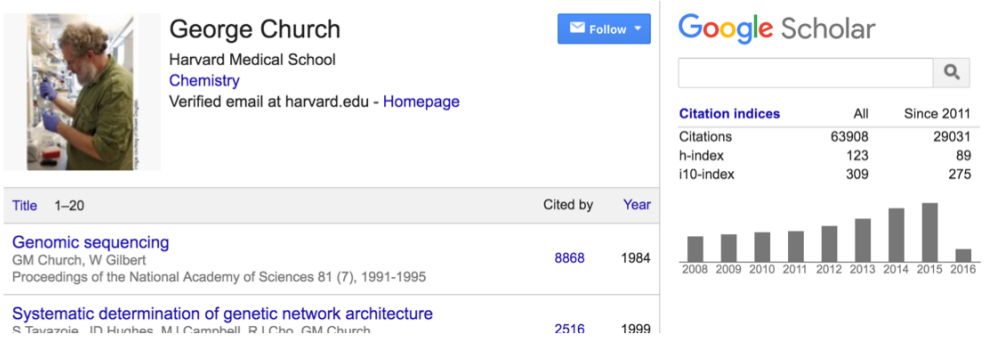
Scientific staff hold 1-7 positions



Collect

Process

>50% of staff have a publication footprint



George Church

Follow

Harvard Medical School

Chemistry

Verified email at harvard.edu - Homepage

Google Scholar

Citation indices	All	Since 2011
Citations	63908	29031
h-index	123	89
i10-index	309	275

2008 2009 2010 2011 2012 2013 2014 2015 2016

Title 1–20	Cited by	Year
Genomic sequencing GM Church, W Gilbert Proceedings of the National Academy of Sciences 81 (7), 1991-1995	8868	1984
Systematic determination of genetic network architecture S Tavazoie, JD Hughes, MJ Campbell, RJ Cho, GM Church	2516	1999

***STAR* 63,908 citations**

William Paul Burns

Follow

University of Ulster

Mobile Technologies, Ambient Assisted Living, Sensors, Wearables, Sel-Management

Verified email at email.ulster.ac.uk

167 citations - domain expert ?

Ian Clark

Follow

Professor of International Relations, University of Queensland

International Relations

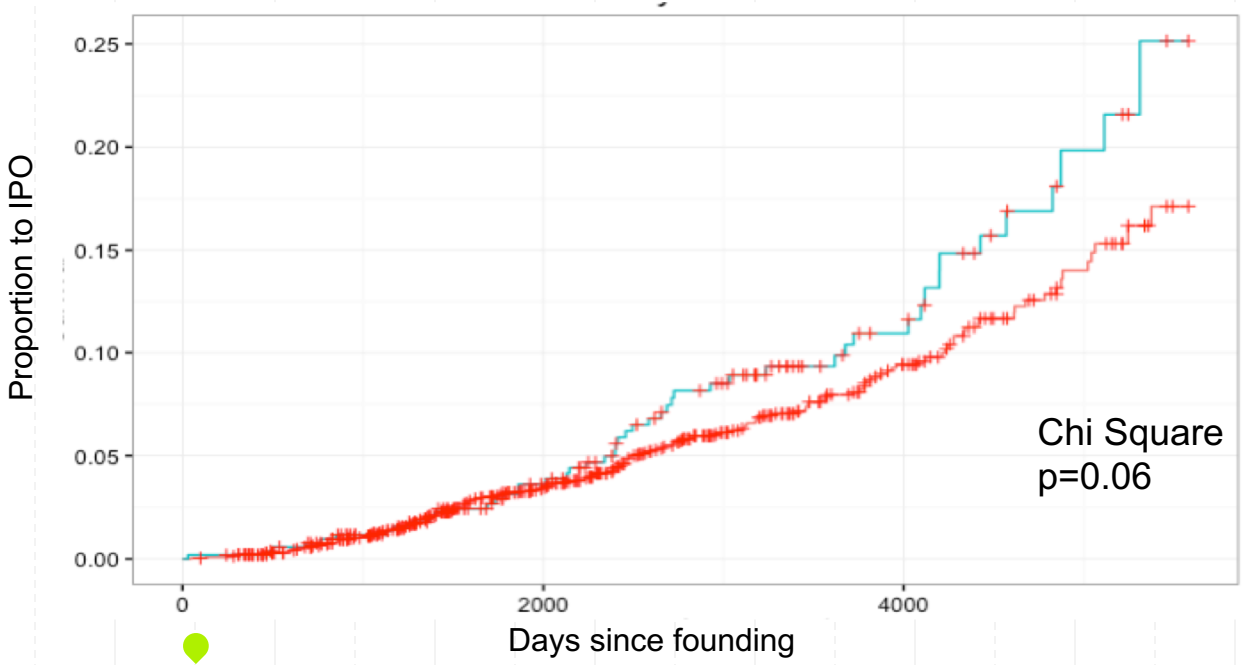
Verified email at uq.edu.au

3336 citations - domain expert ?

Collect

Process

Teams with more scientists IPO faster



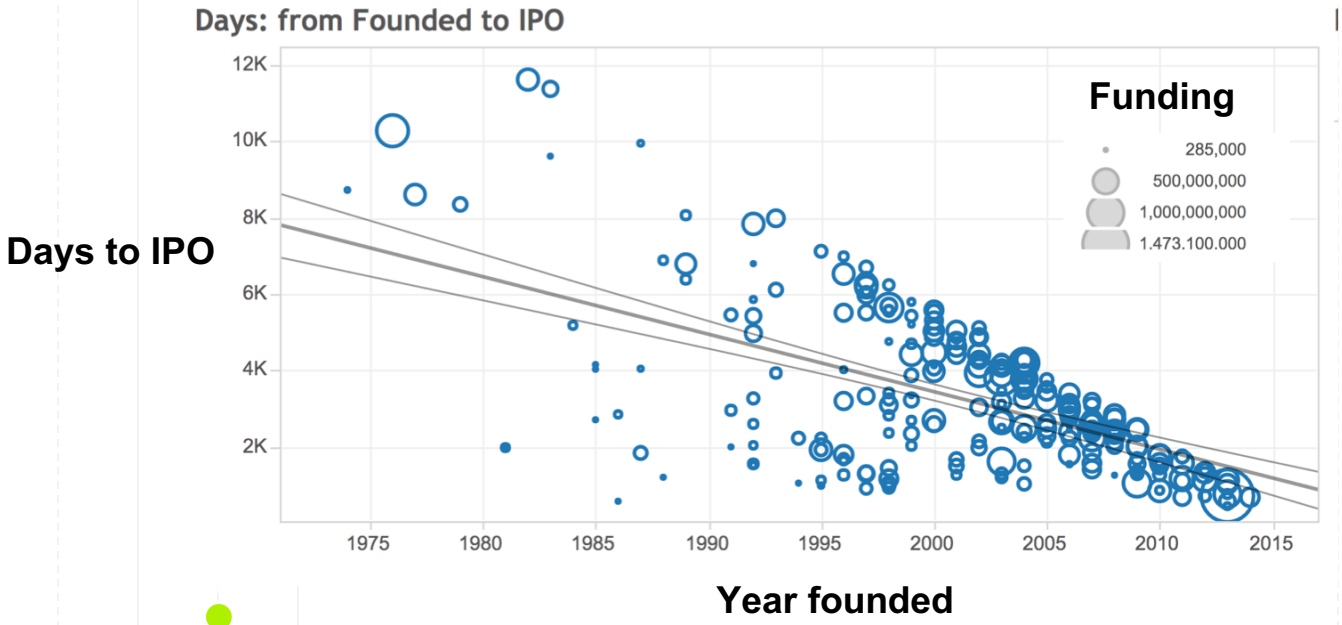
> 50% scientists

fewer or no scientists

Collect

Process

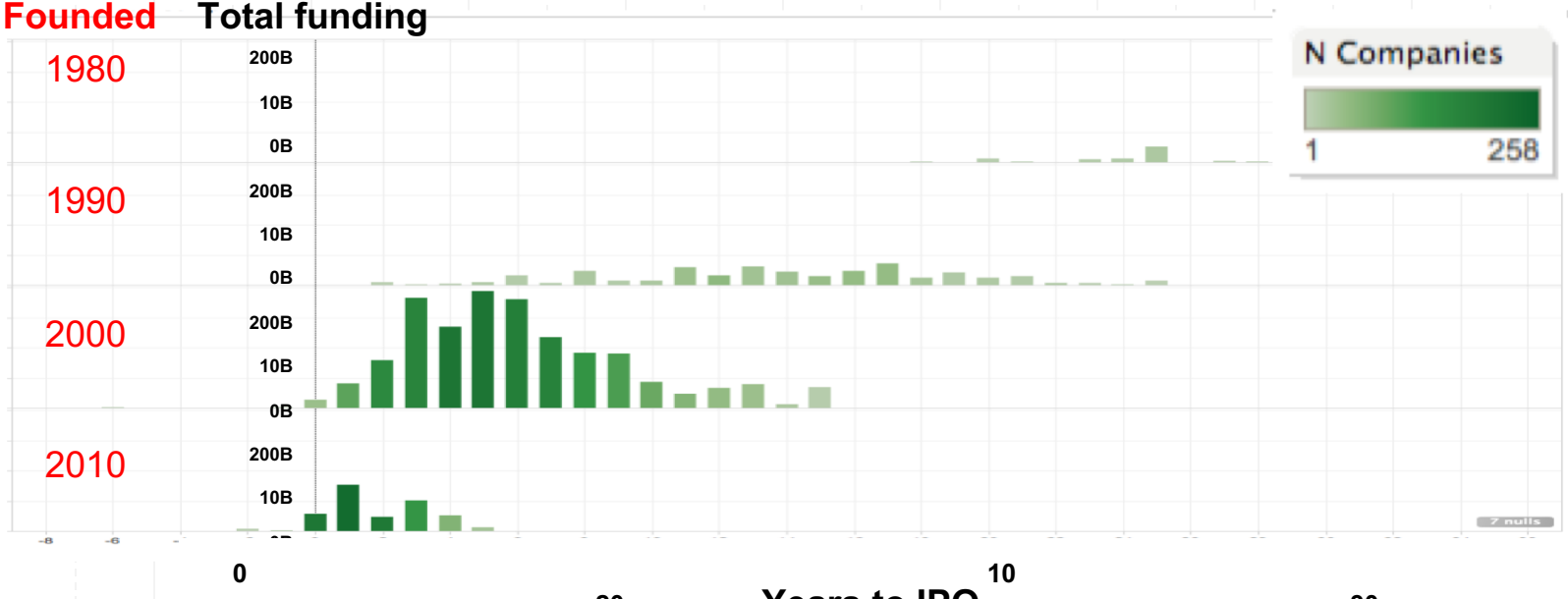
Biotech companies IPO faster every decade



Collect

Process

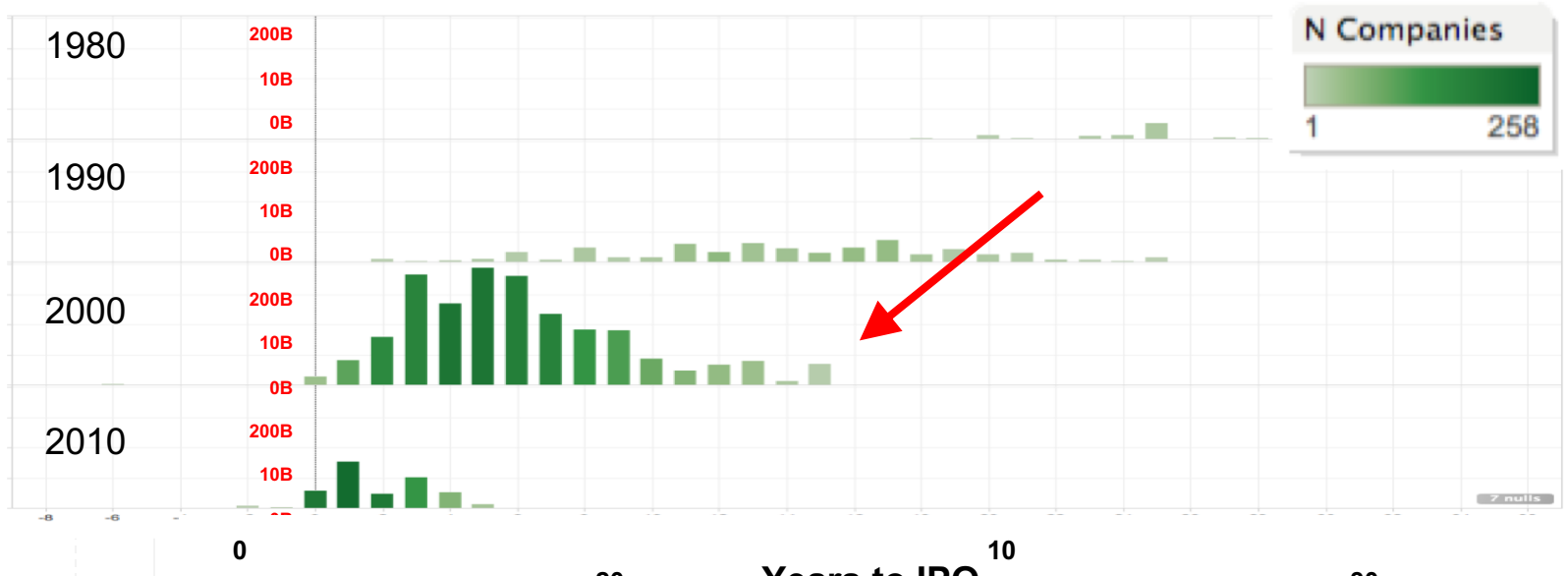
There is decade-level heterogeneity in days to IPO



Collect

Process

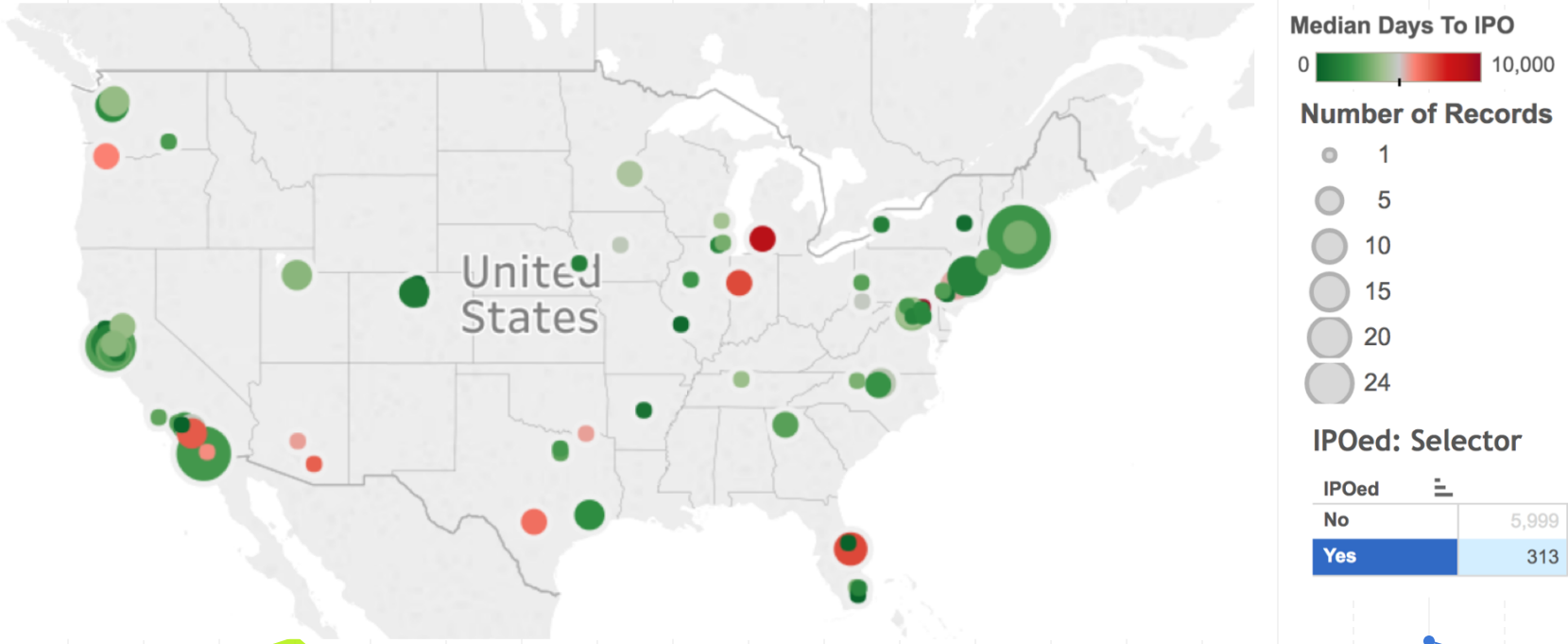
Funding and founding spiked 2000-2010



Collect

Process

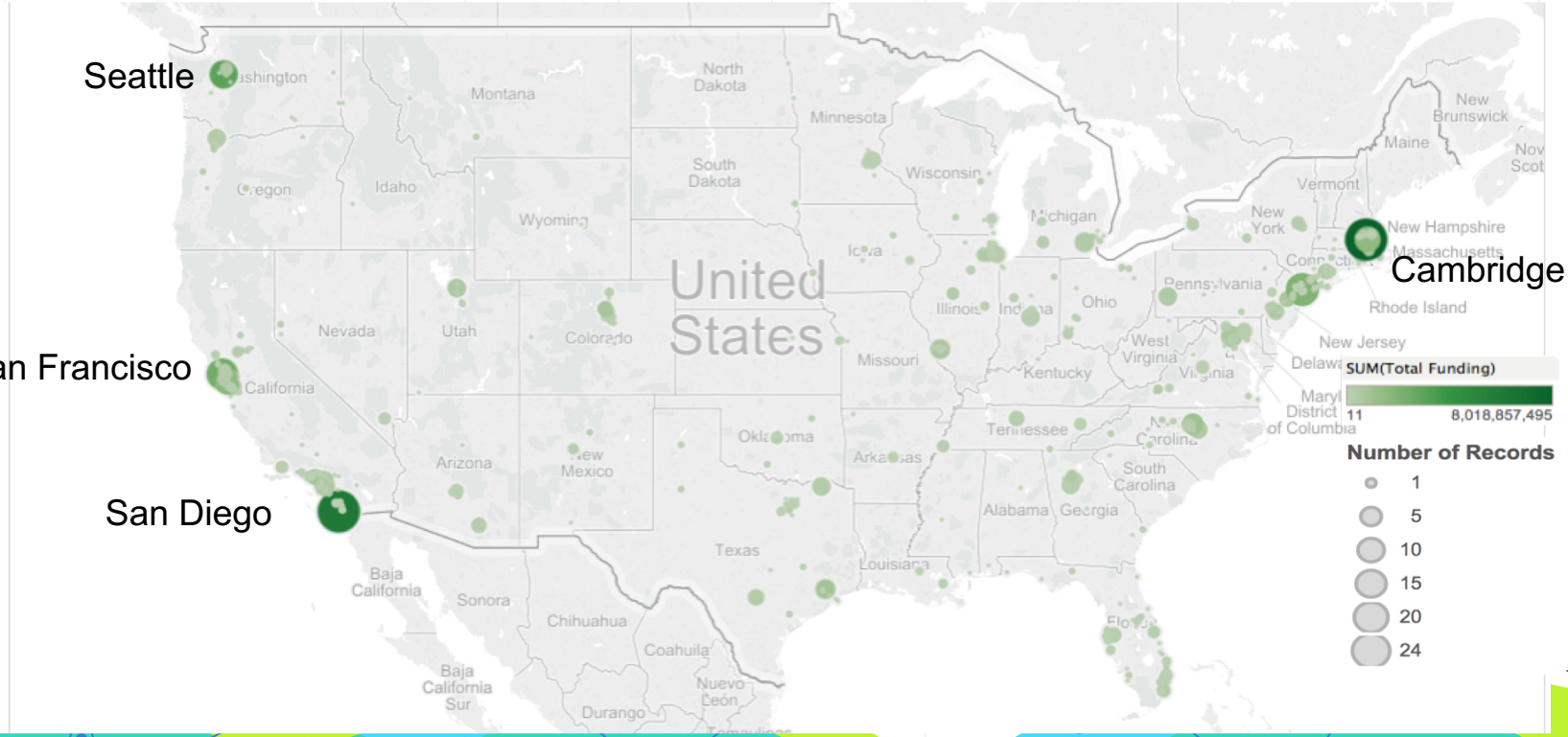
There are more companies on the coasts and they IPO faster



Collect

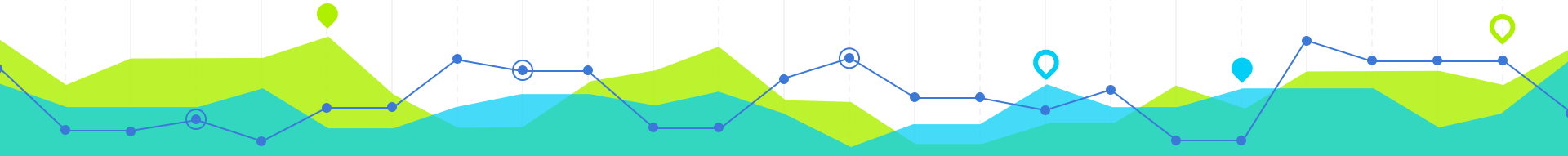
Process

Funding amounts are larger in key coastal cities



Collect Process Predict

Machine Learning



Model validation

To make the most use of our data, we'll run 10-fold cross-validation using a 10% holdout set to estimate model accuracy

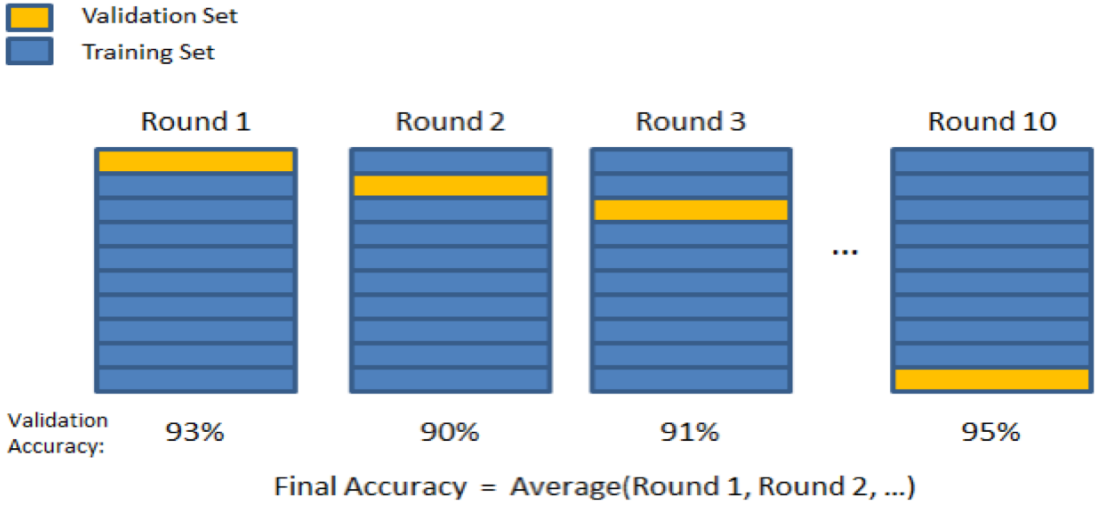


Image source: https://chrisjmccormick.files.wordpress.com/2013/07/10_fold_cv.png

We tested a range of machine learning models for performance

Regression based

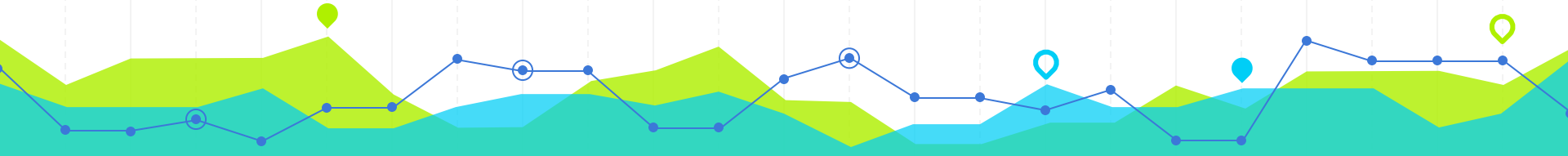
- LogReg - logistic regression
- PLS - partial least squares

Discriminant analysis

- SVM - support vector machine
- FDA - Flexible discriminant analysis

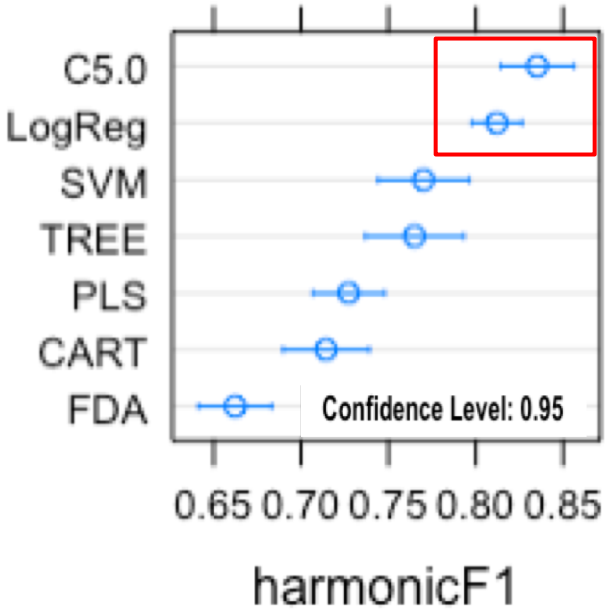
Decision trees

- C5.0
- TREE
- CART



Logistic regression and C5.0 outperform other models

10 Fold CV



F1 evaluates model performance in 2 ways:

1) Maximize detection of positive cases (recall)
2) Minimize detection of negative cases (precision)

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Recall* = True positive / Total positive cases

Precision^ = True positives/ Reported positive

* also Sensitivity or True Positive Rate

^ also Positive Predictive Value

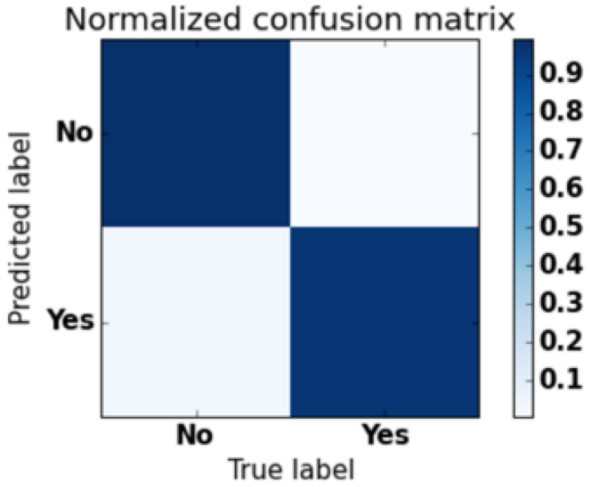
Source: https://en.wikipedia.org/wiki/F1_score

C5 performs well to predict liquidity

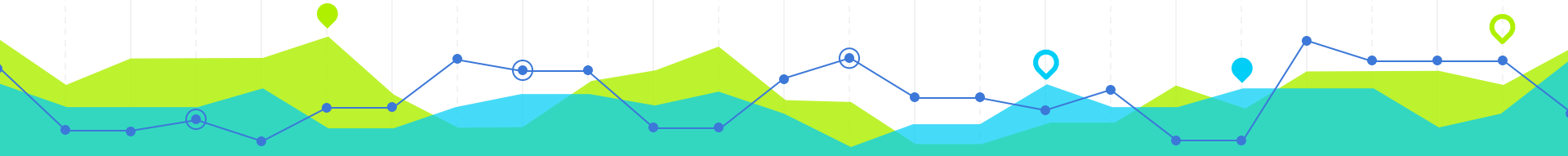
Test Set Results

C5.0

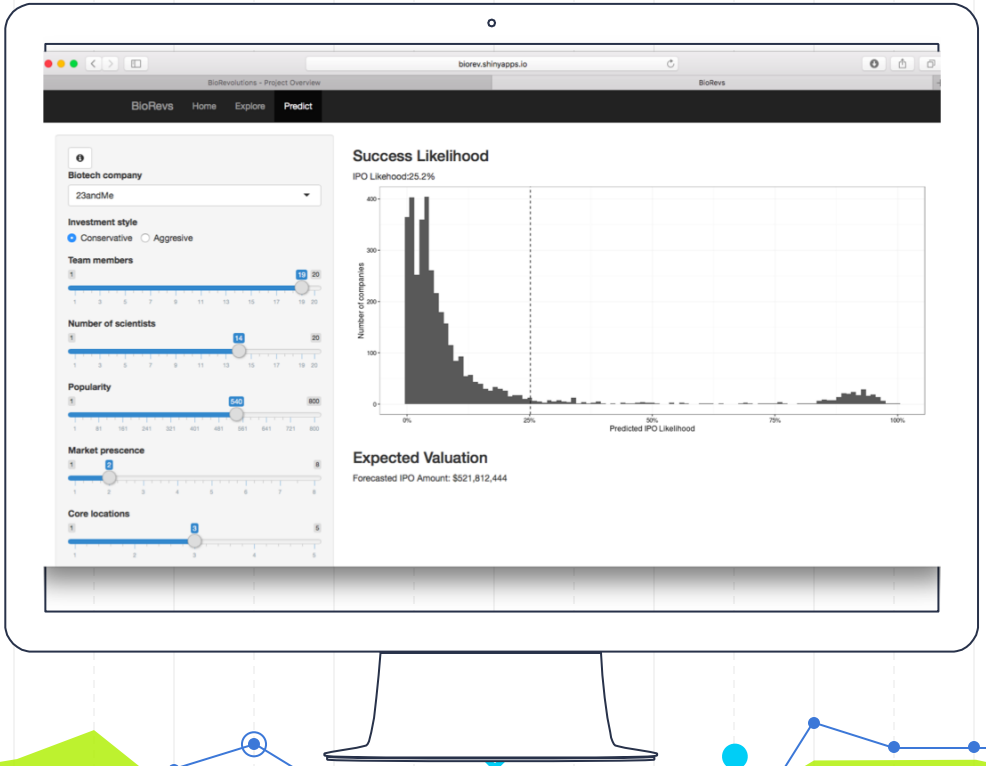
F1 = 0.94
Recall= 0.92
Precision = 0.97



Prediction explorer



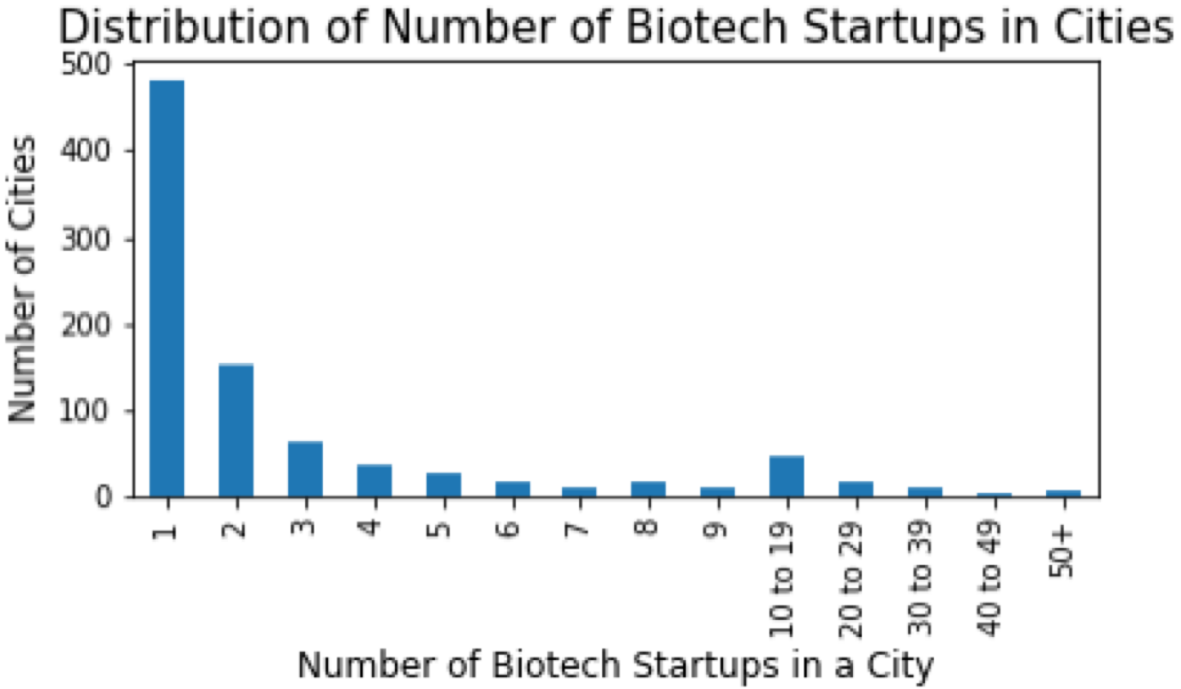
Collect → Process → Predict → Present



http://groups.ischool.berkeley.edu/biorevs/

Where should you look for your next job?

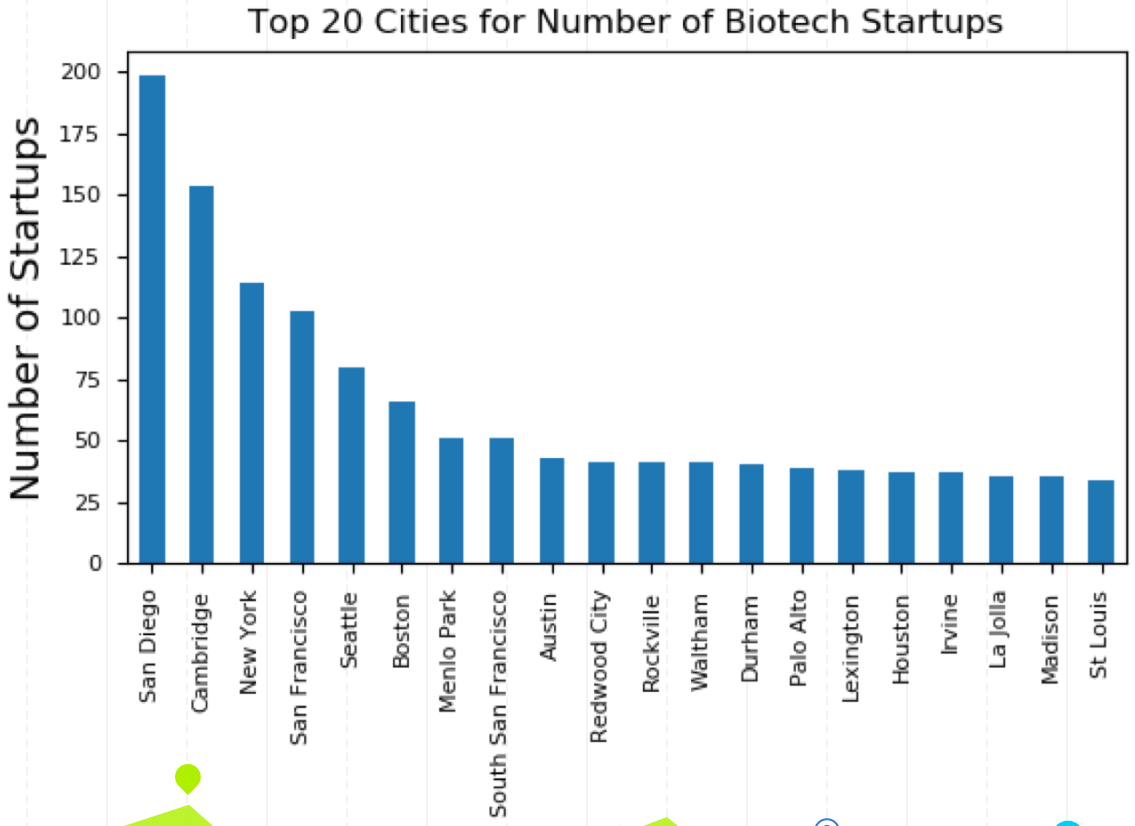
The Vast Majority of Cities have Few Biotech Startup Companies



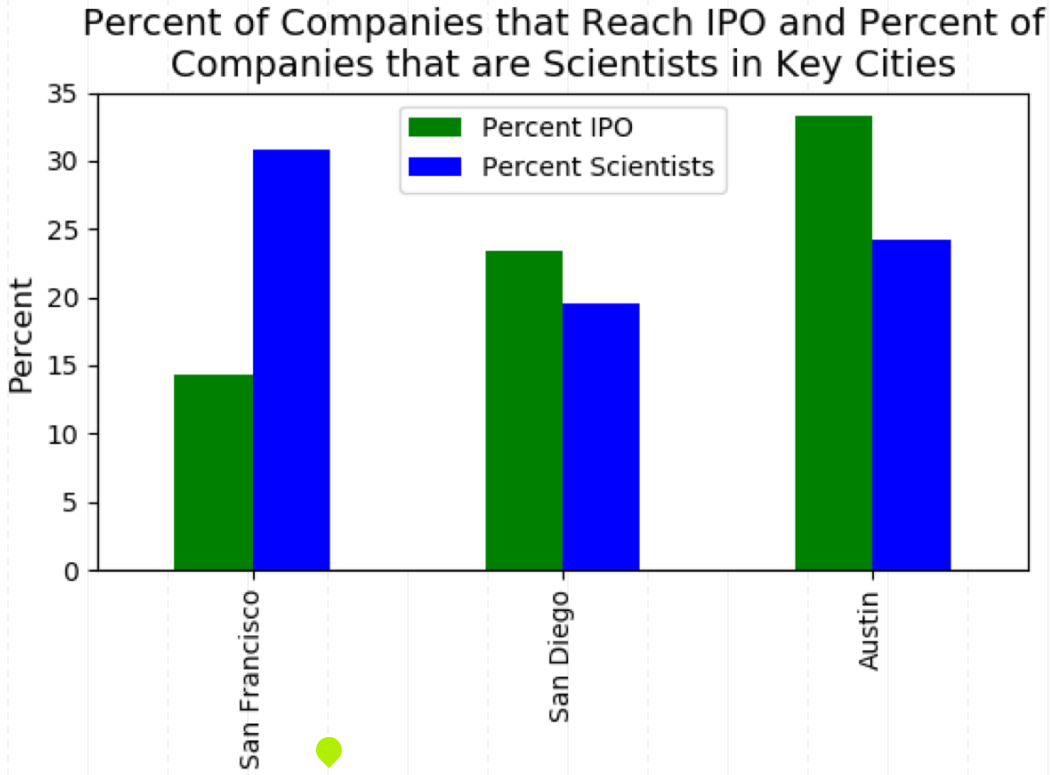
There are 899 US Cities that have at least one Biotech Company.

Most cities have very few companies.

We can focus on Some Biotech Hub Cities



Company Features in Different Cities

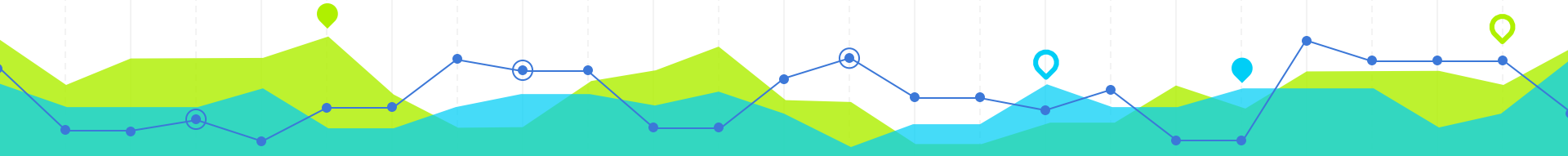


Percent IPO = % Companies reaching IPO in < 6000 days

National average is ~16% reach IPO.

National average is ~21% Scientists.

What Expertise do Scientists in these Cities Have?



We defined Scientific Expertise by Words Used in Publication Titles

“PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity”

“Mother-daughter asymmetry of pH underlies aging and rejuvenation in yeast”

We Analyzed Publication Titles Stored in the Pubmed Database

24,000,000

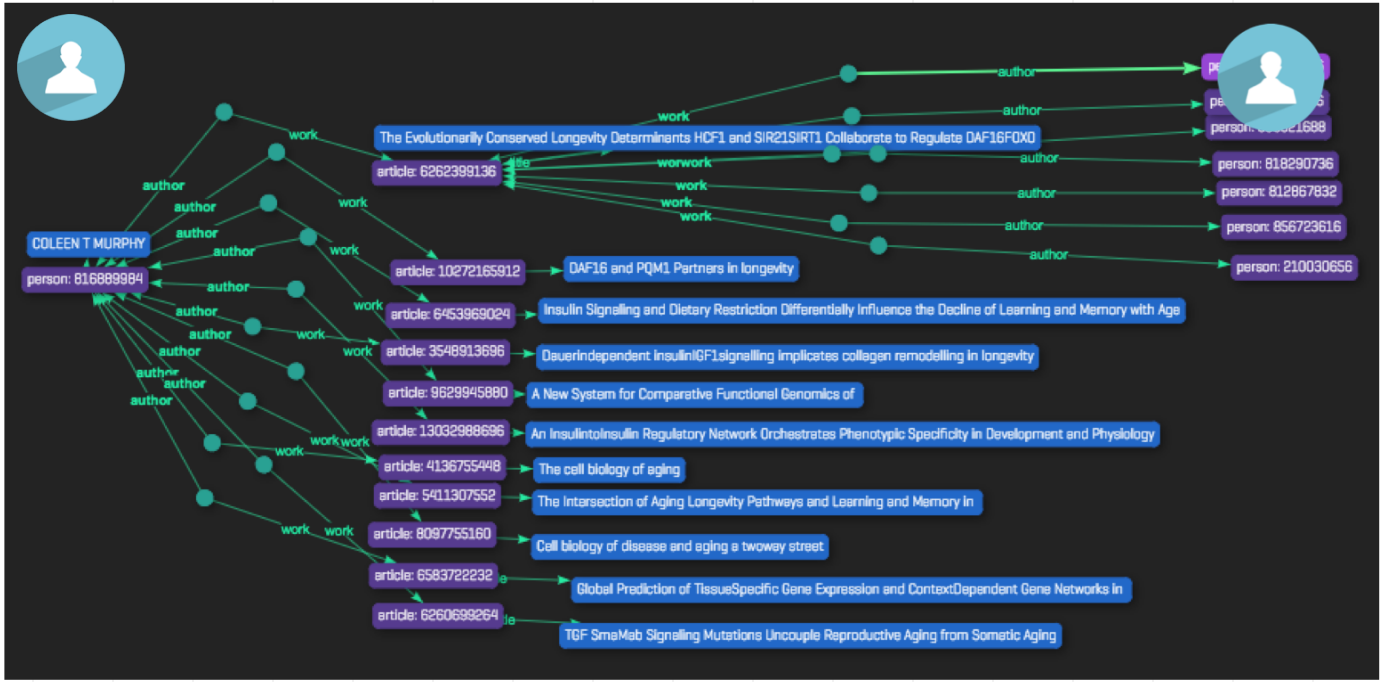
Full PubMed Dataset (all Biomedical literature)

~ 1.5 Million Open Source

PubMed Central (BioTech-open access subset)

Data include: Publication Titles, Authors, and other Metadata

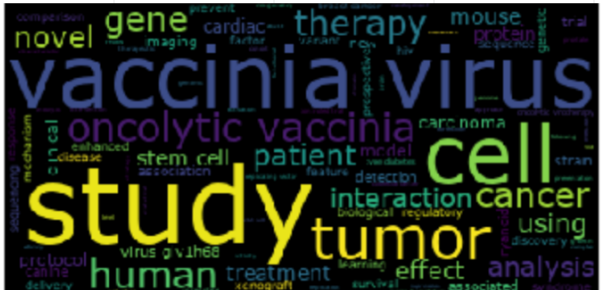
Learning about Scientists and Cities with a Knowledge Graph



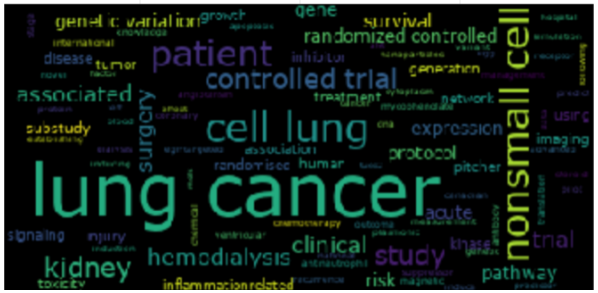
Start with Pubmed data in GRAKN.AI

Cities Differ in Scientific Expertise

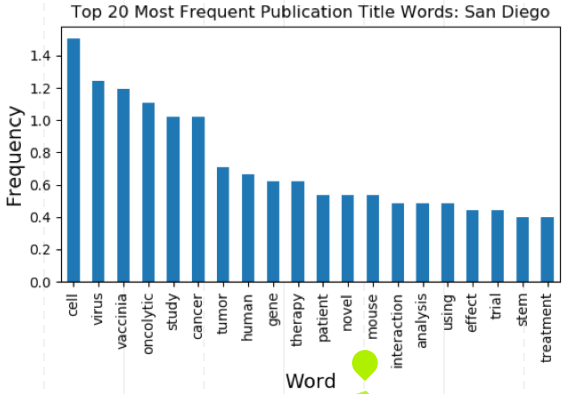
San Diego

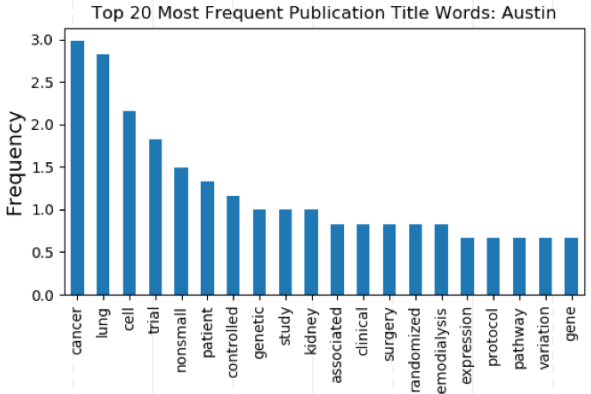


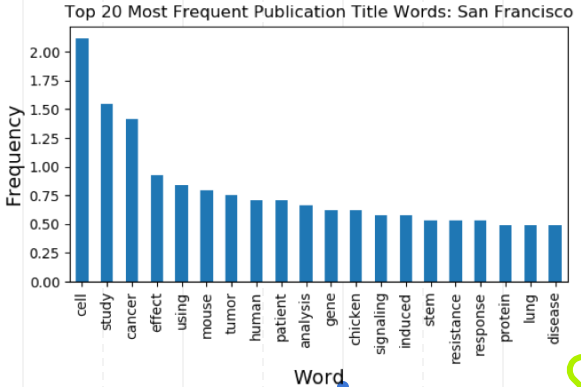
Austin



San Francisco







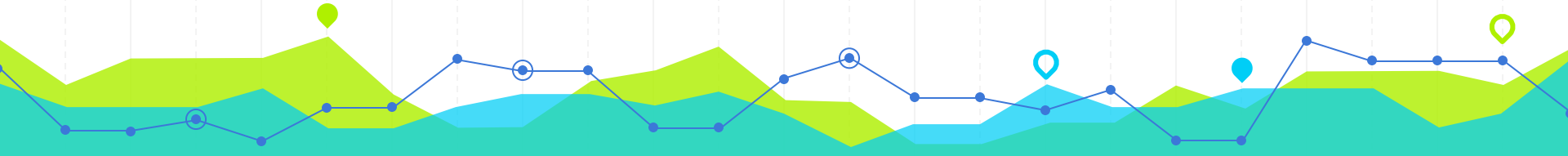
Scientists - Choose your own Adventure (wisely)!

For Cities of Interest, consider:

- # of Biotech Startups
- % Companies reaching IPO
- % Scientists at Companies
- Scientific Expertise

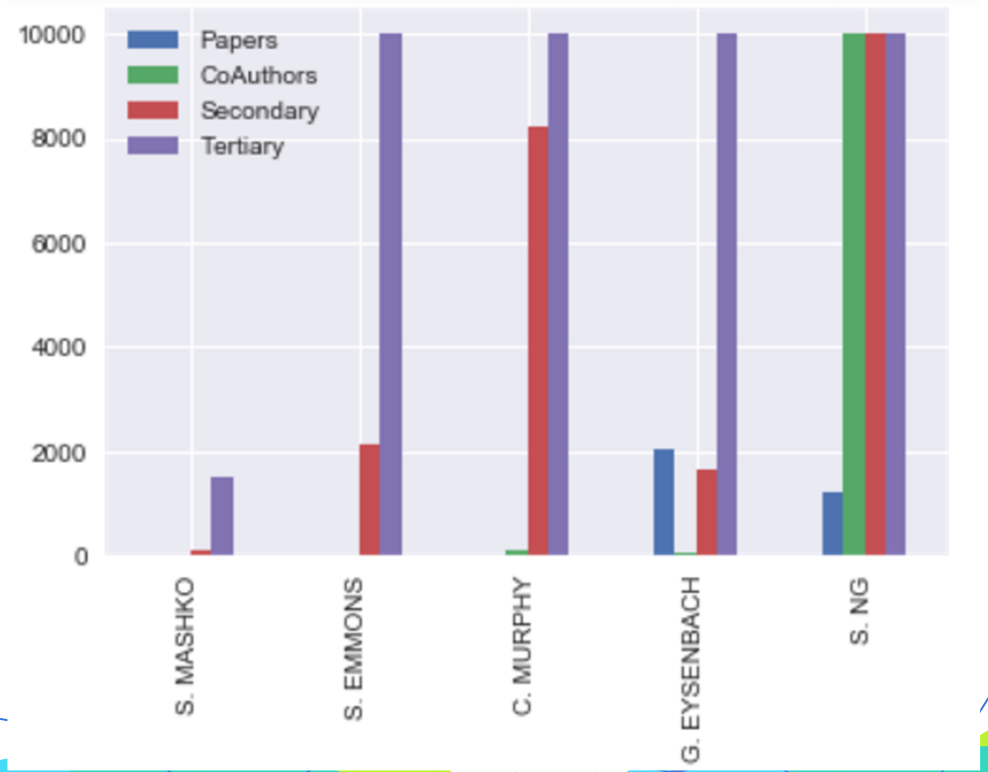
Ongoing development

Quantification of science networks

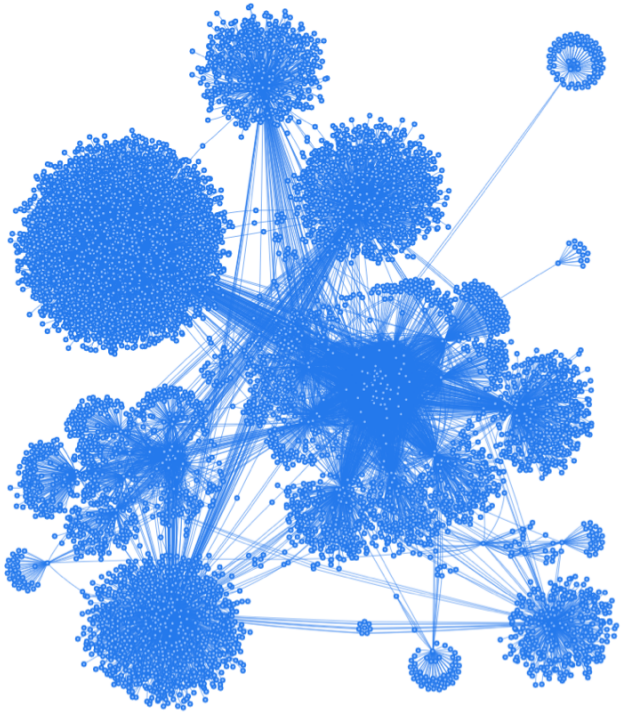


Graph traversal can be used to understand collaboration patterns

Papers and co-authors to three degrees



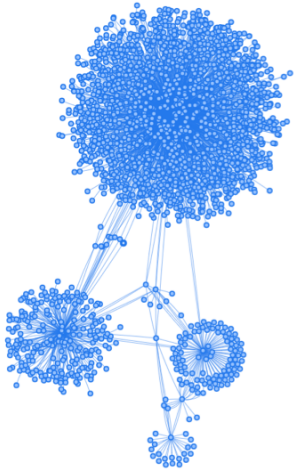
Collaborator networks differ by investigator



Coleen Murphy 2nd collaboration network

Collaborator networks differ by investigator

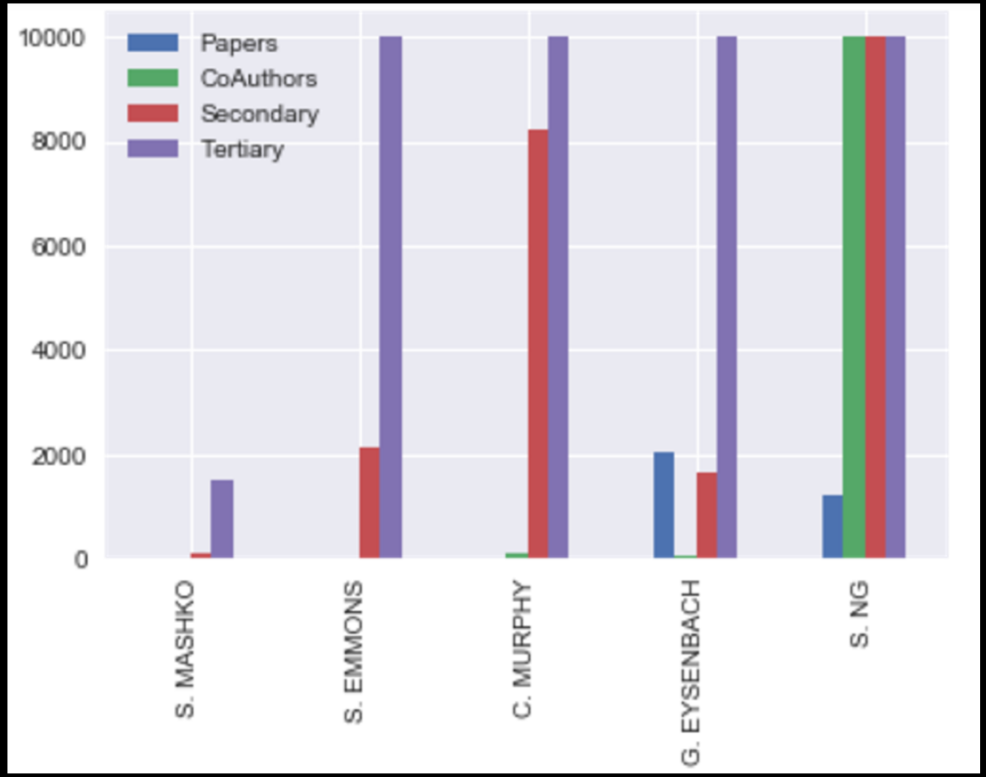
Scott Emmons 2nd degree collaboration network



Coleen Murphy 2nd collaboration network

Graph traversal-collaboration patterns

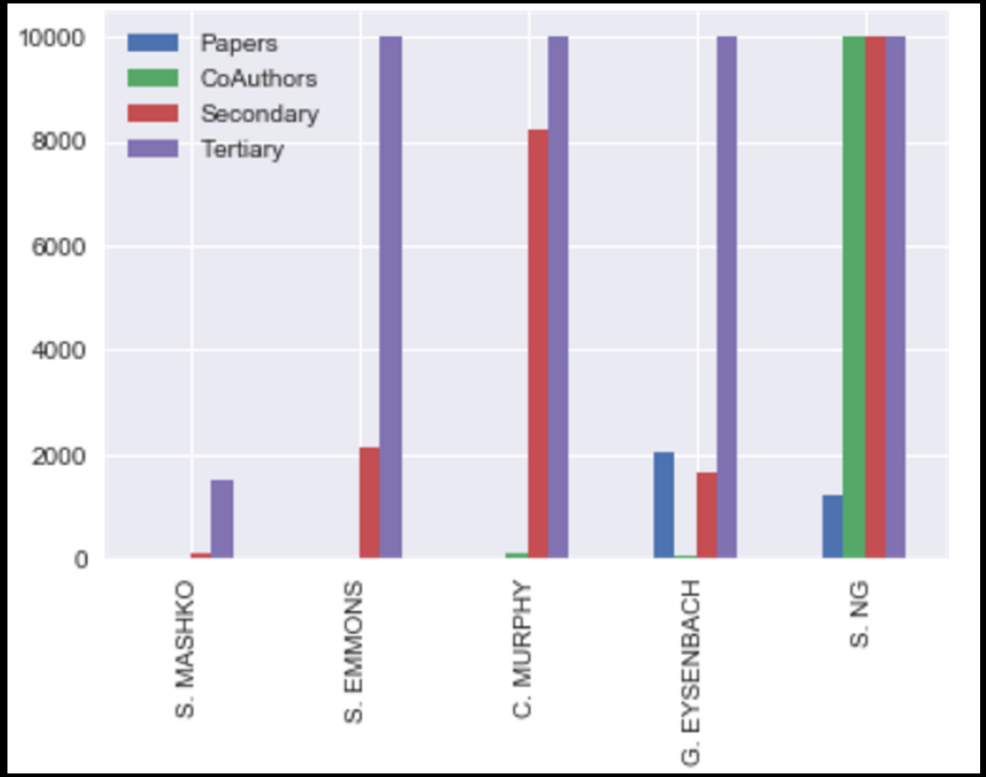
Papers or co-authors to three degrees



BD SG

Graph traversal-collaboration patterns

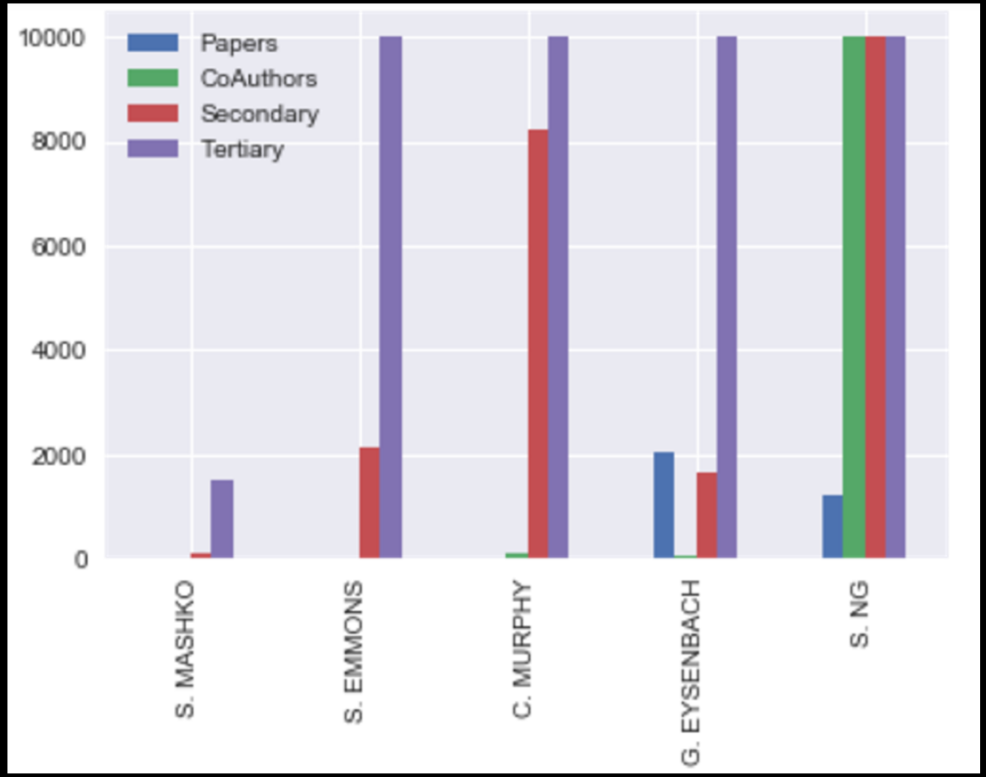
Papers or co-authors to three degrees



BD SG

Graph traversal-collaboration patterns

Papers or co-authors to three degrees



BD SG

Title words give an idea of expertise

- https://www.flickr.com/photos/utasel/6961732976

BD SG

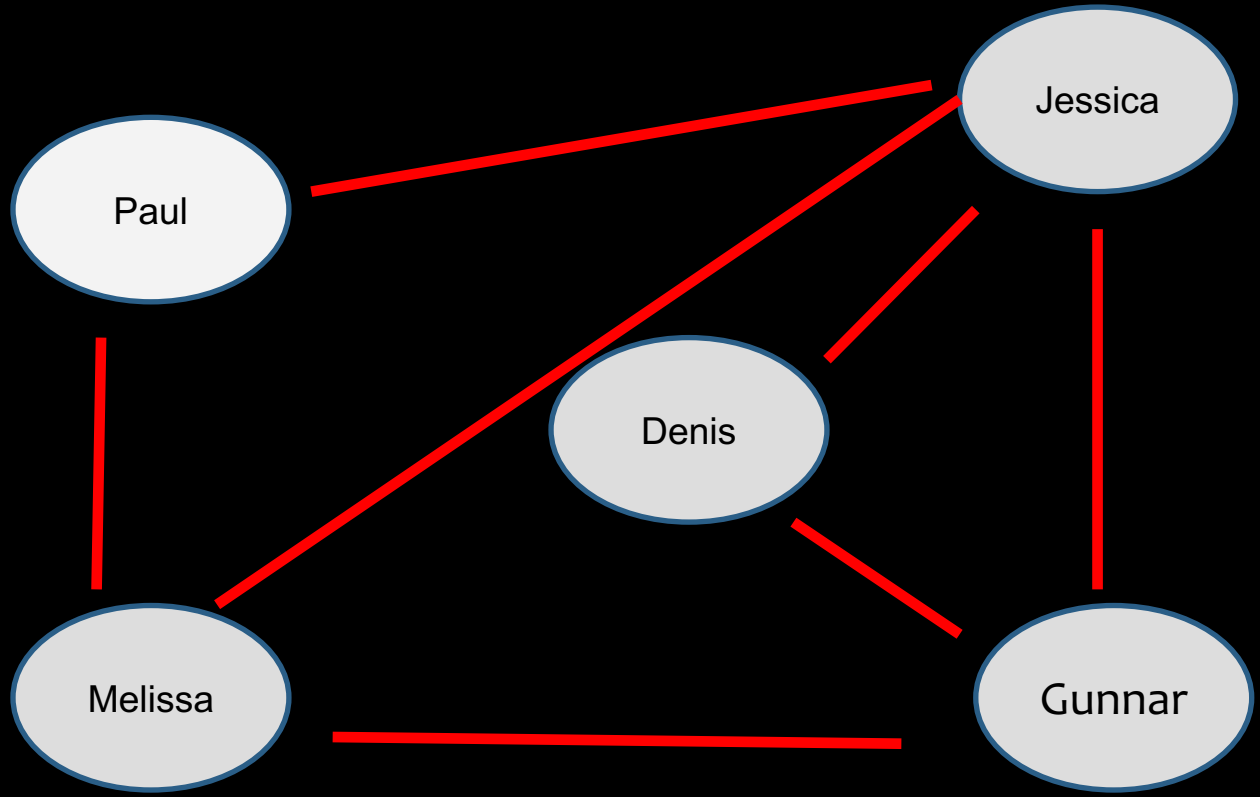
What is the expertise in their network

Secondary title cloud

Tertiary title cloud

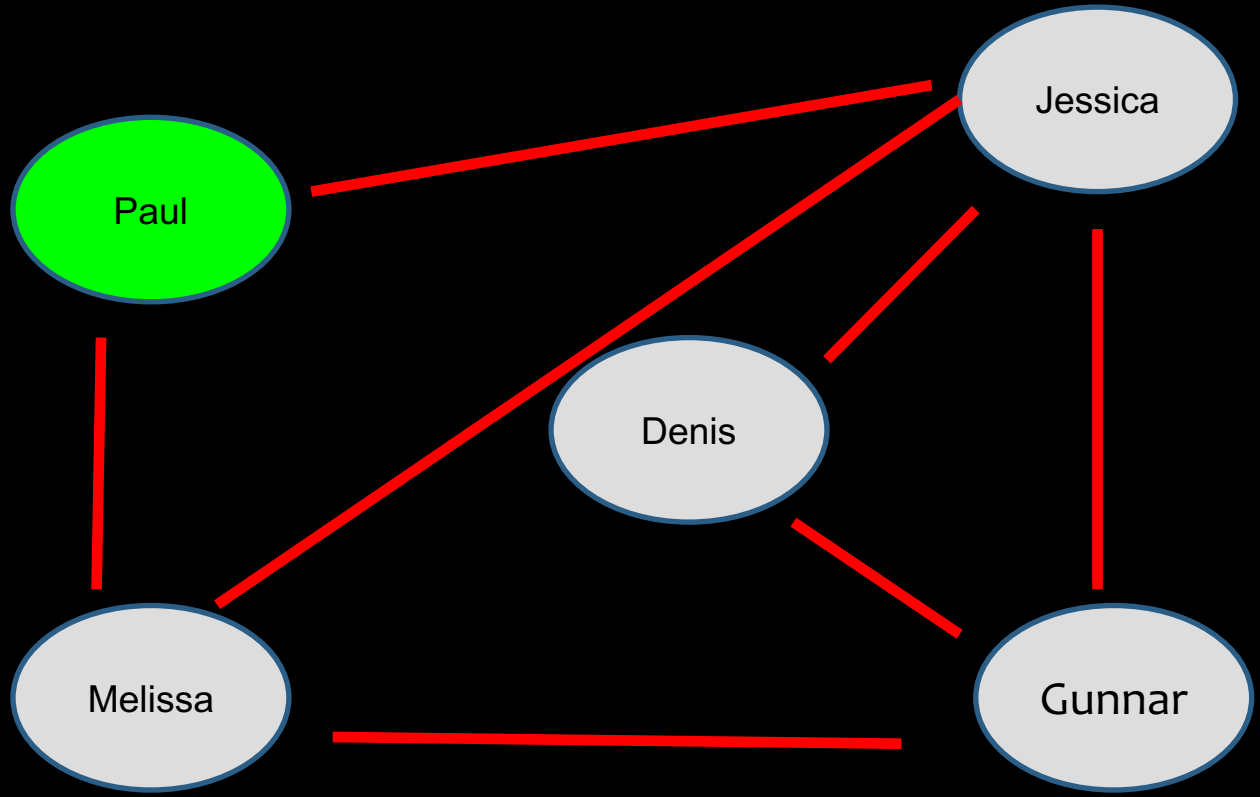
BDSG

Consider Paul in a simple social network



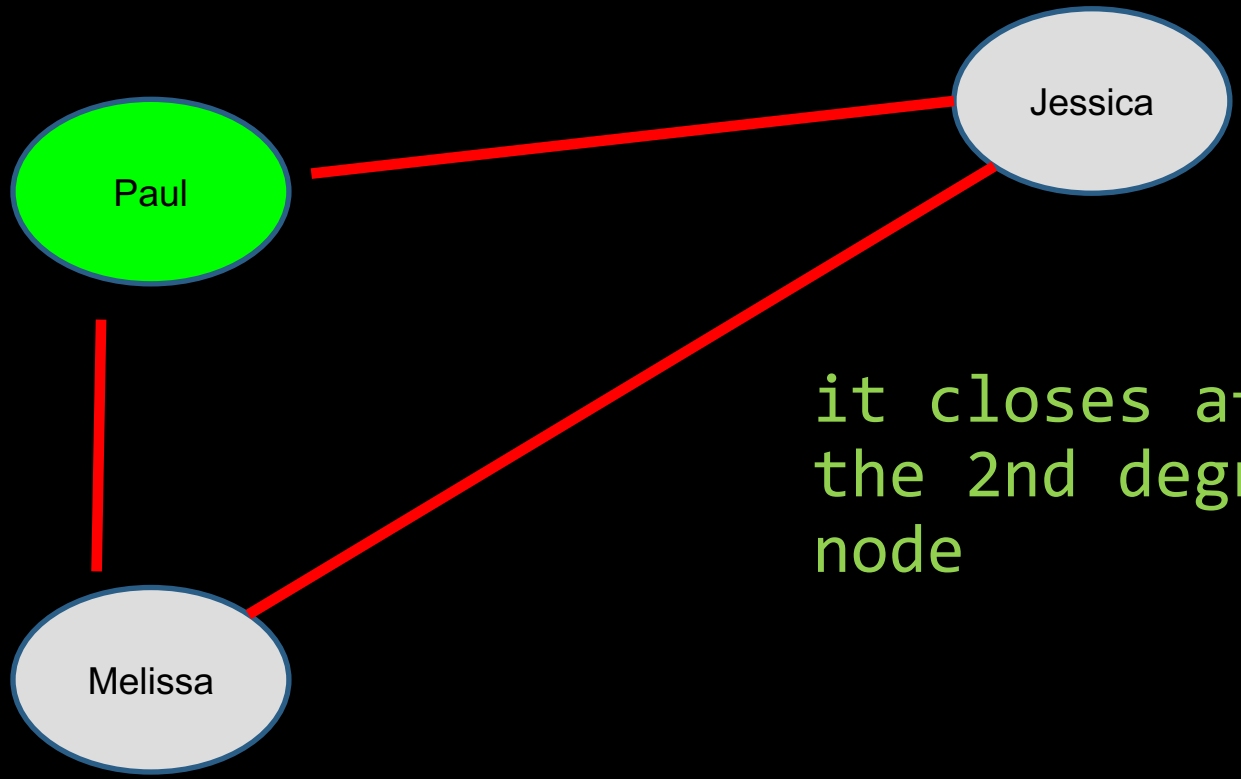
BD SG

Consider Paul in a simple social network



BD SG

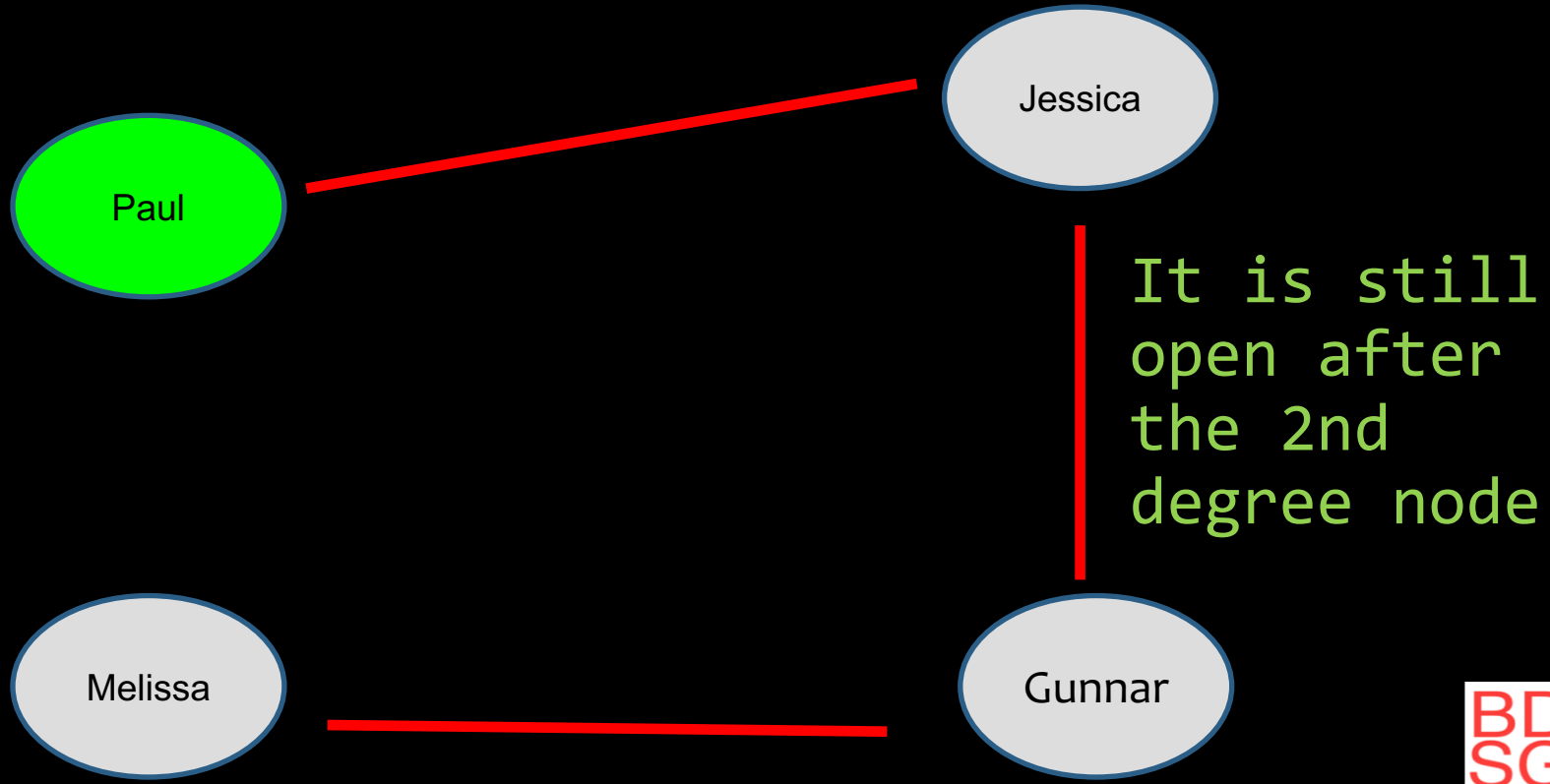
This relationship is a triangle



it closes after the 2nd degree node

BD SG

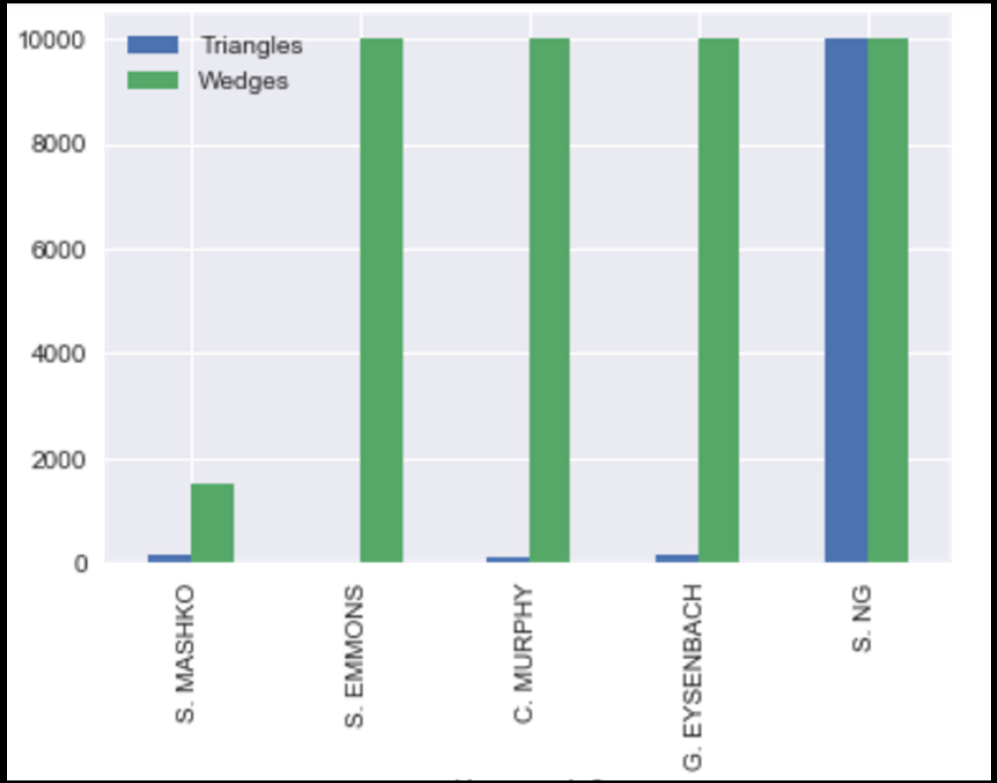
This relationship is a wedge



BD SG

Quantify networks with wedge/triangle ratio

Triangles vs Wedges



BD SG

THANKS!

Any questions?

http://bds.group

denis@bds.group

gunnar@bds.group

kiersten.henderson@austincapitaldata.com

Well defined Communities