

CHICAGO

INTERNATIONAL  
SOFTWARE DEVELOPMENT  
CONFERENCE 2015

goto;  
conference

# Computing Professional Identity for the Economic Graph

*Vitaly Gordon*

# Agenda

**1**

Introduction

**2**

LinkedIn's Vision

**3**

Computing Professional Identity

**4**

Selected Topics

**5**

Summary



**1**

# Introduction

**2**

LinkedIn's Vision

**3**

Computing Professional Identity

**4**

Selected Topics

**5**

Summary



## About me



**Vitaly Gordon**



## About me



## About me



**Technion**  
Israel Institute of Technology

# About me



## About me



# About me



## About me



**@bigdatasc**



**/in/vitalygordon**

# What's in it for you?

What's in it for you?

- 1. You will get a better understanding of what a data scientists does**



What's in it for you?

- 1. You will get a better understanding of what a data scientists does**
- 2. You will learn about how hard cleaning data can be**

What's in it for you?

- 1. You will get a better understanding of what a data scientists does**
- 2. You will learn about how hard cleaning data can be**
- 3. You will learn why LinkedIn needs endorsements**

# What's in it for me?

# What's in it for me?



# What's in it for me?



**@bigdatasc**

**1**

Introduction

**2**

# LinkedIn's Vision

**3**

Computing Professional Identity

**4**

Selected Topics

**5**

Summary



**Create economic opportunity for  
every professional in the world**



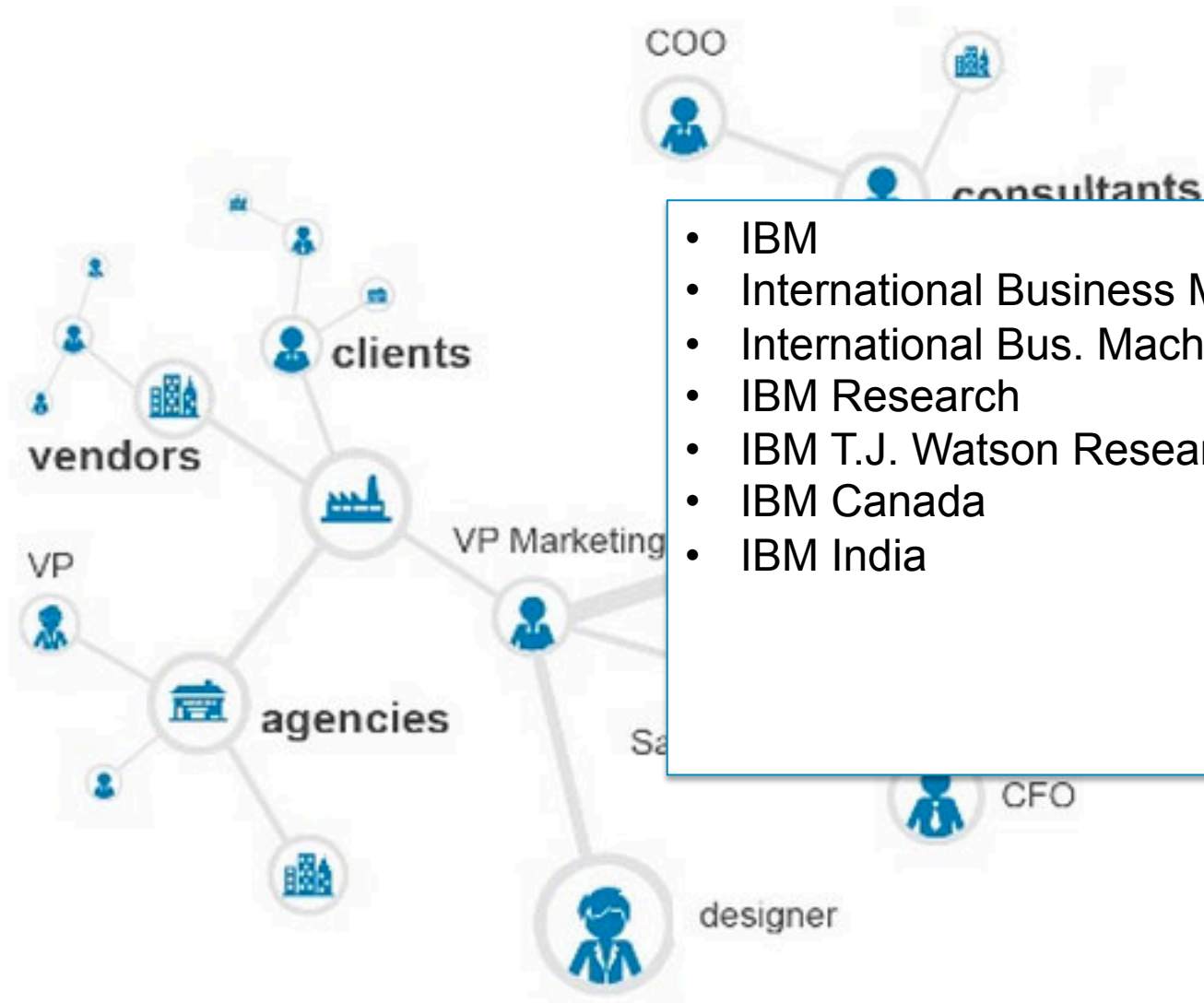
in

# THE ECONOMIC GRAPH



- CEO
- Chief Executive Officer
- CEO and Founder
- CEO & Co-founder
- President and CEO
- Owner

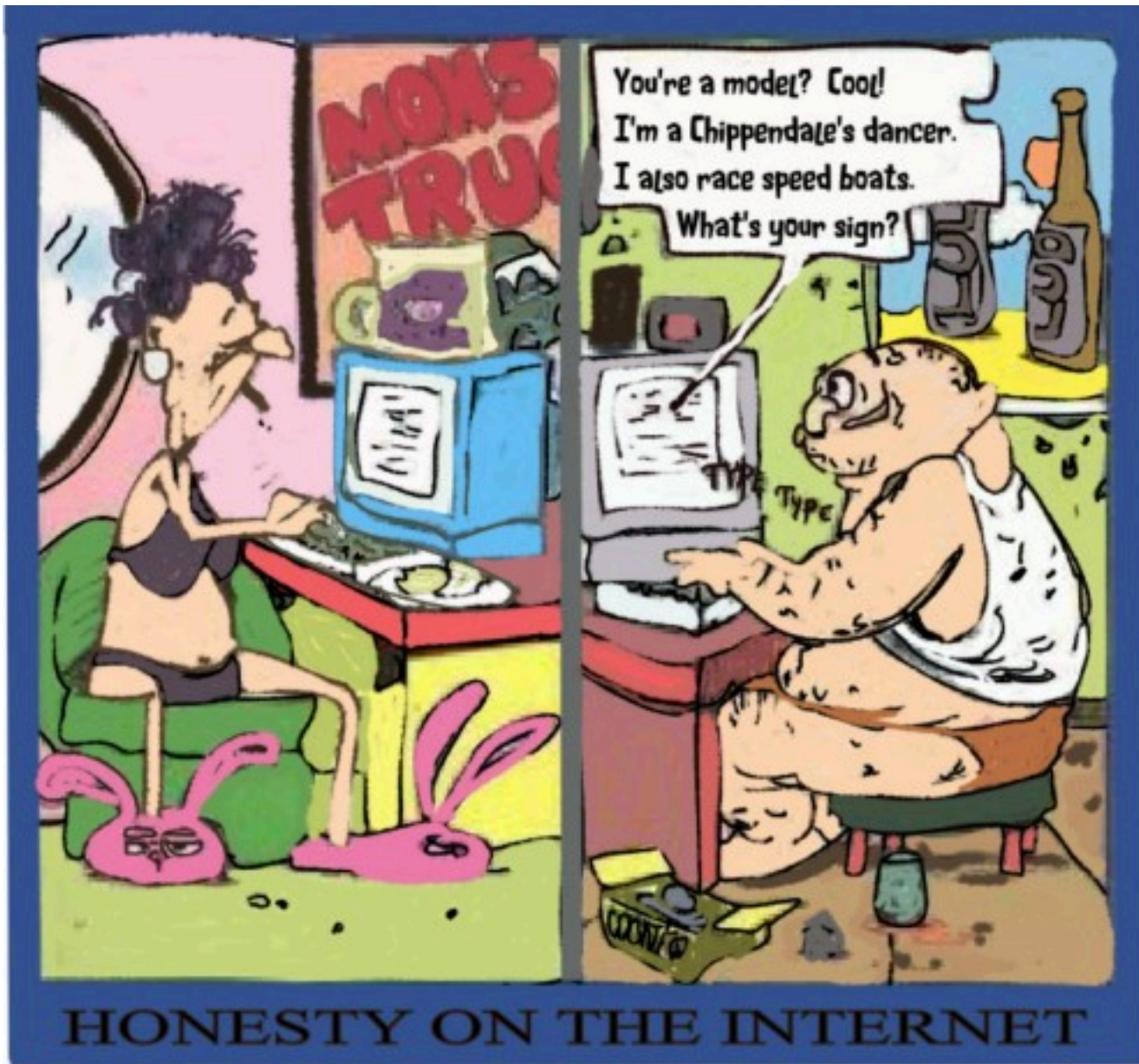




- IBM
- International Business Machines
- International Bus. Machines
- IBM Research
- IBM T.J. Watson Research Center
- IBM Canada
- IBM India

- UCLA
- University of California, Los Angeles
- UC Los Angeles
- The Anderson School of Management





**1**

Introduction

**2**

LinkedIn's Vision

**3**

**Computing Professional Identity**

**4**

Selected Topics

**5**

Summary



# Why Do We Need Identity Standardization?

# Why Do We Need Identity Standardization?





# Why Do We Need Identity Standardization?

## OR

If you would like to broaden your search to find profiles which include one or more terms you can separate those terms with the upper-case word OR.

- "Pitney Bowes" OR "Hewlett-Packard"
- Helpdesk OR "Help Desk" OR "Technical Support"
- "Vice President" OR VP OR "V.P." OR SVP OR EVP
- J2EE OR "Java Enterprise Edition" OR JEE OR JEE5
- "account executive" OR "account exec" OR "account manager" OR "sales executive" OR "sales manager" OR "sales representative"



**UNSTANDARDIZED DATA**

**MAKES KITTY SAD**

# Text Based Solution

- Applies acronym expansion (e.g. vp -> vice president)
- Applies abbreviation expansion (e.g. sr. -> senior)
- Select the most common standard titles
- Selects standard sub strings (e.g. software engineer and tech lead in search -> [software engineer, tech lead])

# Problems with a Text Based Approach

Software Engineer

Senior Software  
Engineer



# Problems with a Text Based Approach



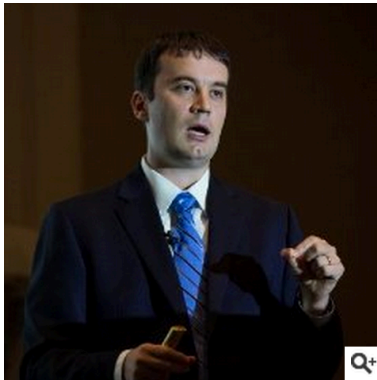
Software Engineer   Software Developer   Programmer



# Problems with a Text Based Approach







## Vitaly Gordon

Director, Data Science at Salesforce

San Francisco Bay Area | Computer Software

Current      Salesforce, Scalding  
Previous     LinkedIn, LinkedIn, LivePerson  
Education   Technion-Machon Technologi Le' Israel

[Complete your profile](#)

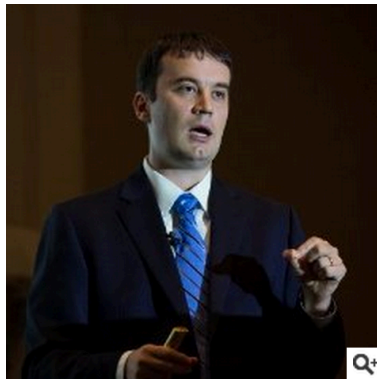
[Edit](#)



PREMIUM

500+  
connections





## Vitaly Gordon

Director, Data Science at Salesforce

San Francisco Bay Area | Computer Software

Current Salesforce, Scalding  
Previous LinkedIn, LinkedIn, LivePerson  
Education Technion-Machon Technologi Le' Israel

[Complete your profile](#)

[Edit](#)

PREMIUM

500+  
connections

81 Data Mining

74 Hadoop

67 Machine Learning

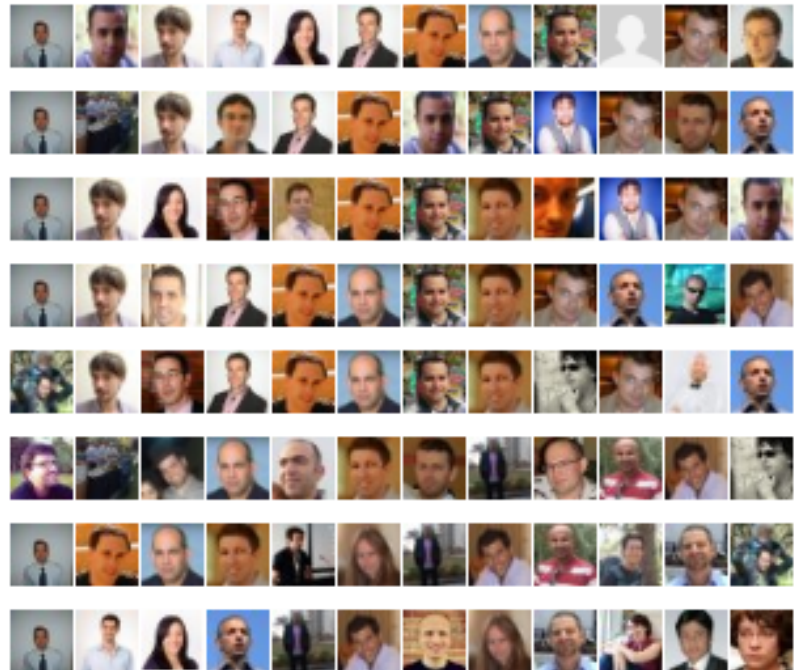
32 Java

27 Algorithms

19 Python

16 MapReduce

15 Data Science







## Daniel Tunkelang

Head of Search Quality at LinkedIn

Mountain View, California | Internet

Current LinkedIn, Karat, Etsy  
Previous LinkedIn, Google, Endeca  
Education Carnegie Mellon University

Send a message

Endorse

1st PREMIUM

500+  
connections

99+

Karaoke



99+

Information Retrieval



99+

Machine Learning



99+

Data Mining



99+

Text Mining



99+

Big Data



99+

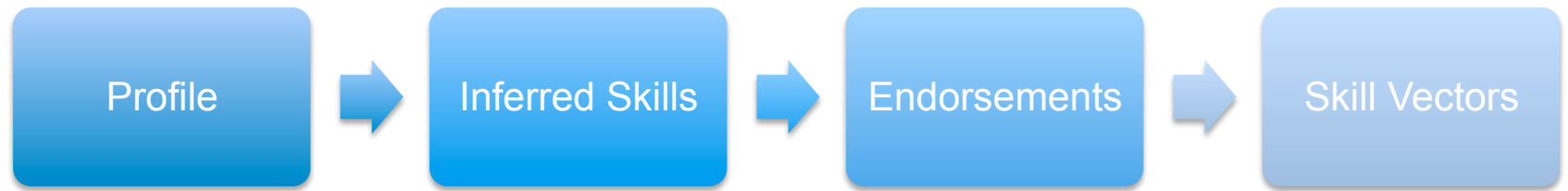
Search

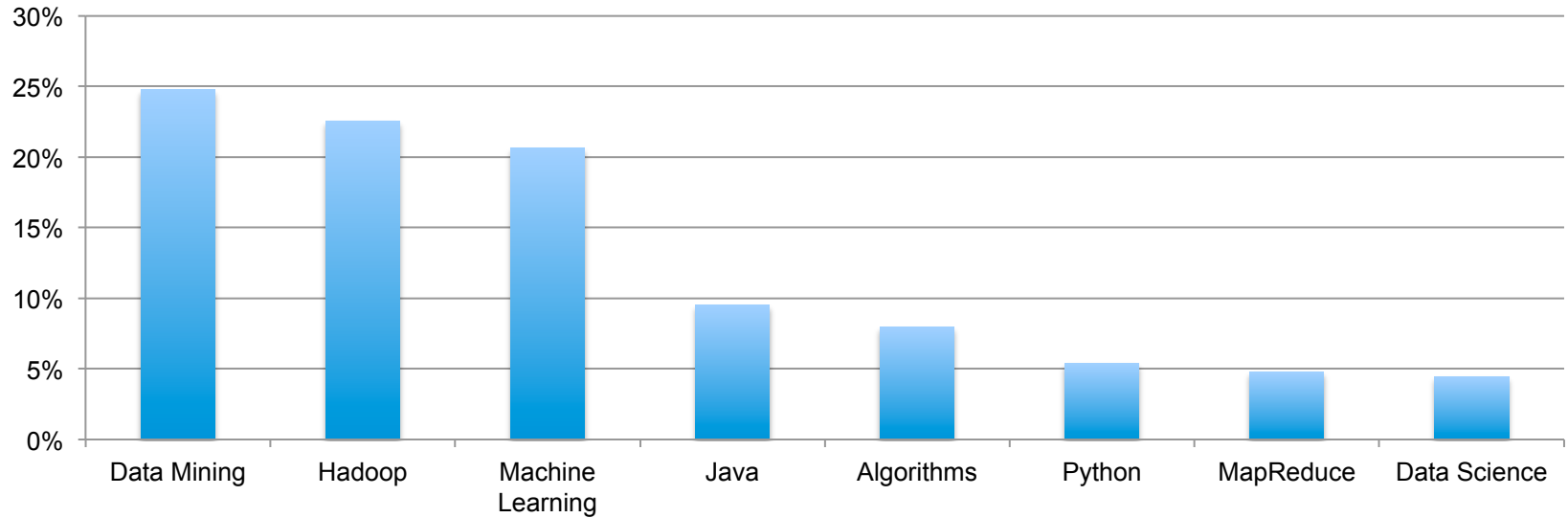
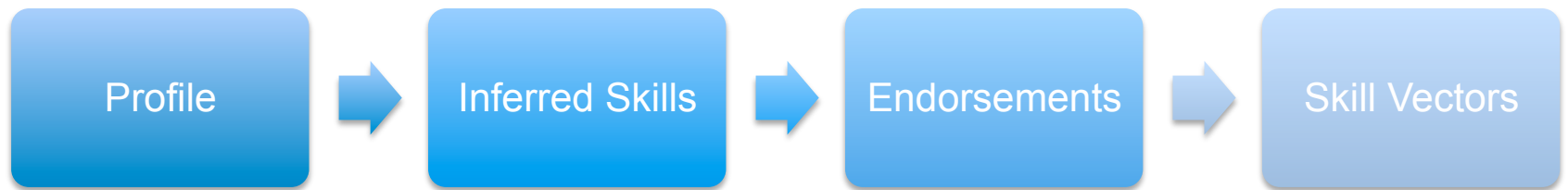


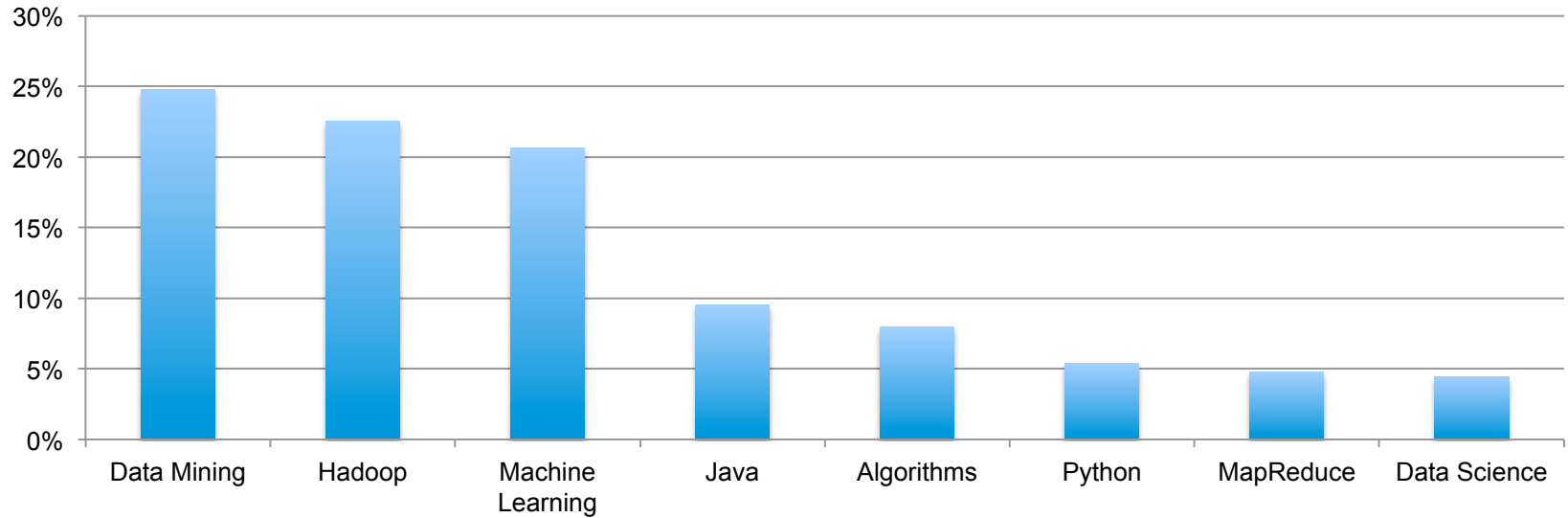
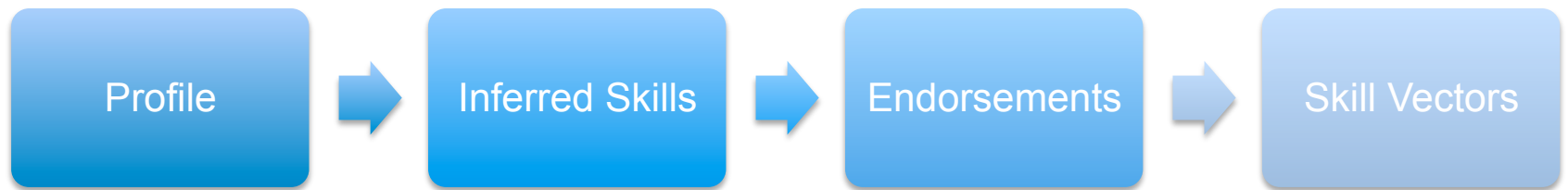
93

Computer Science

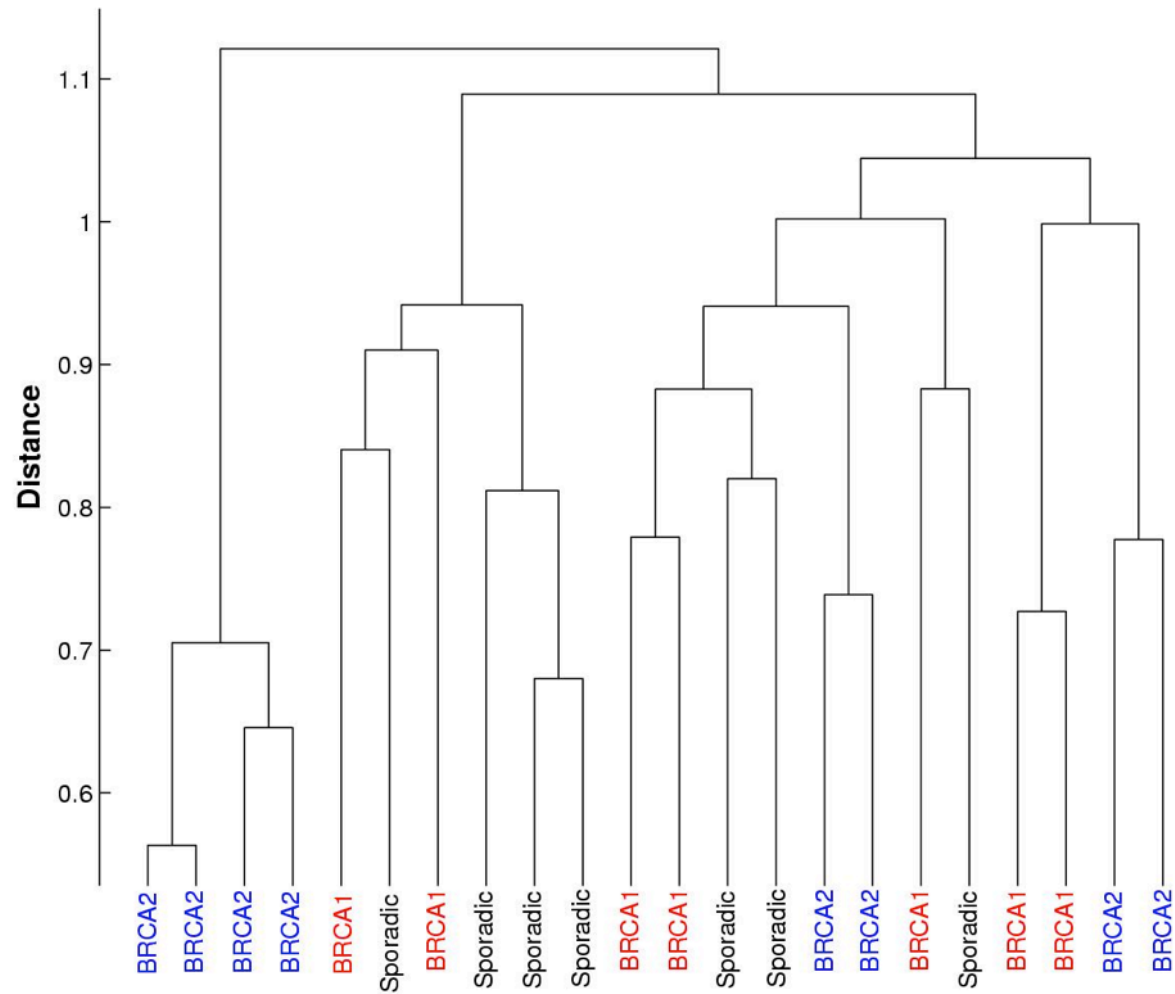




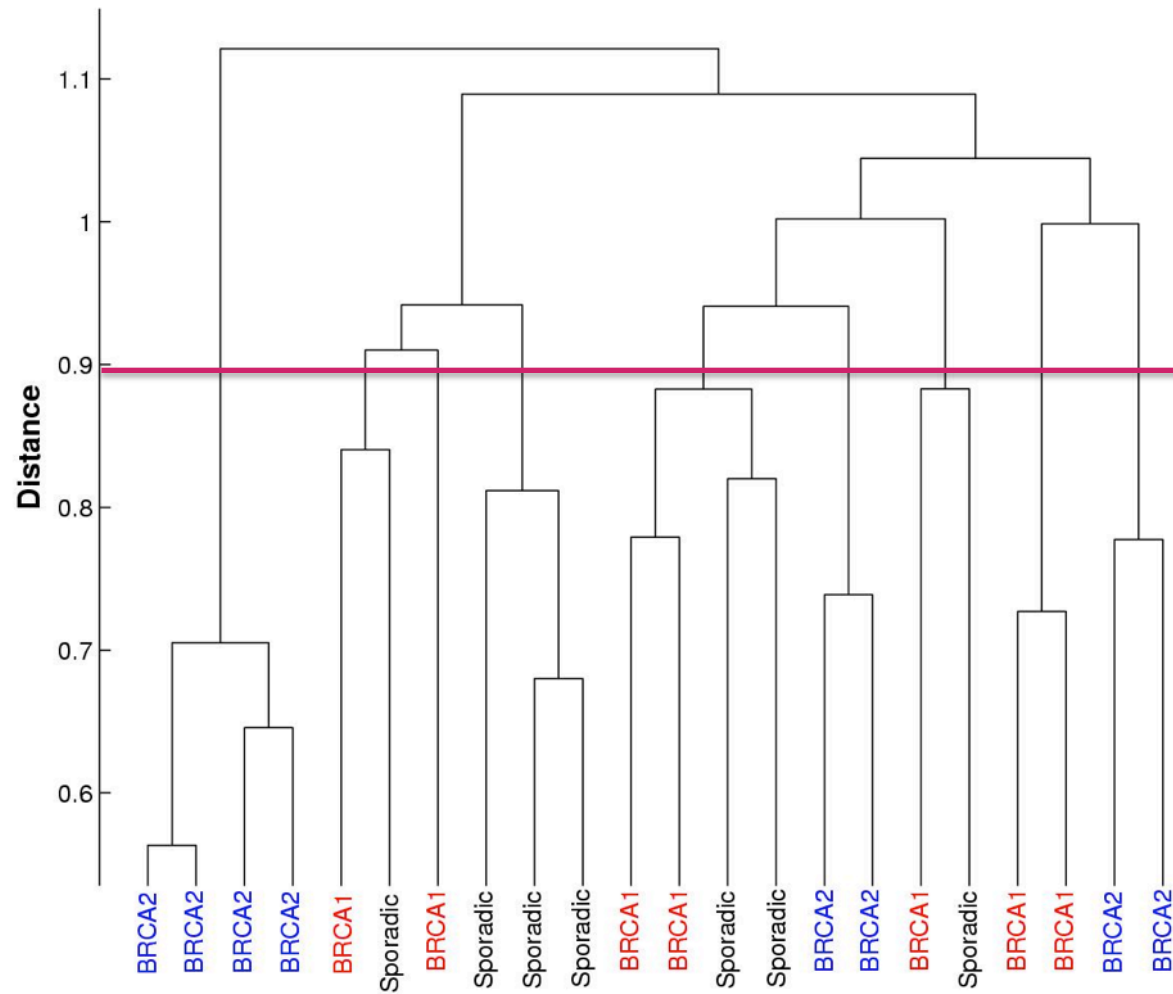




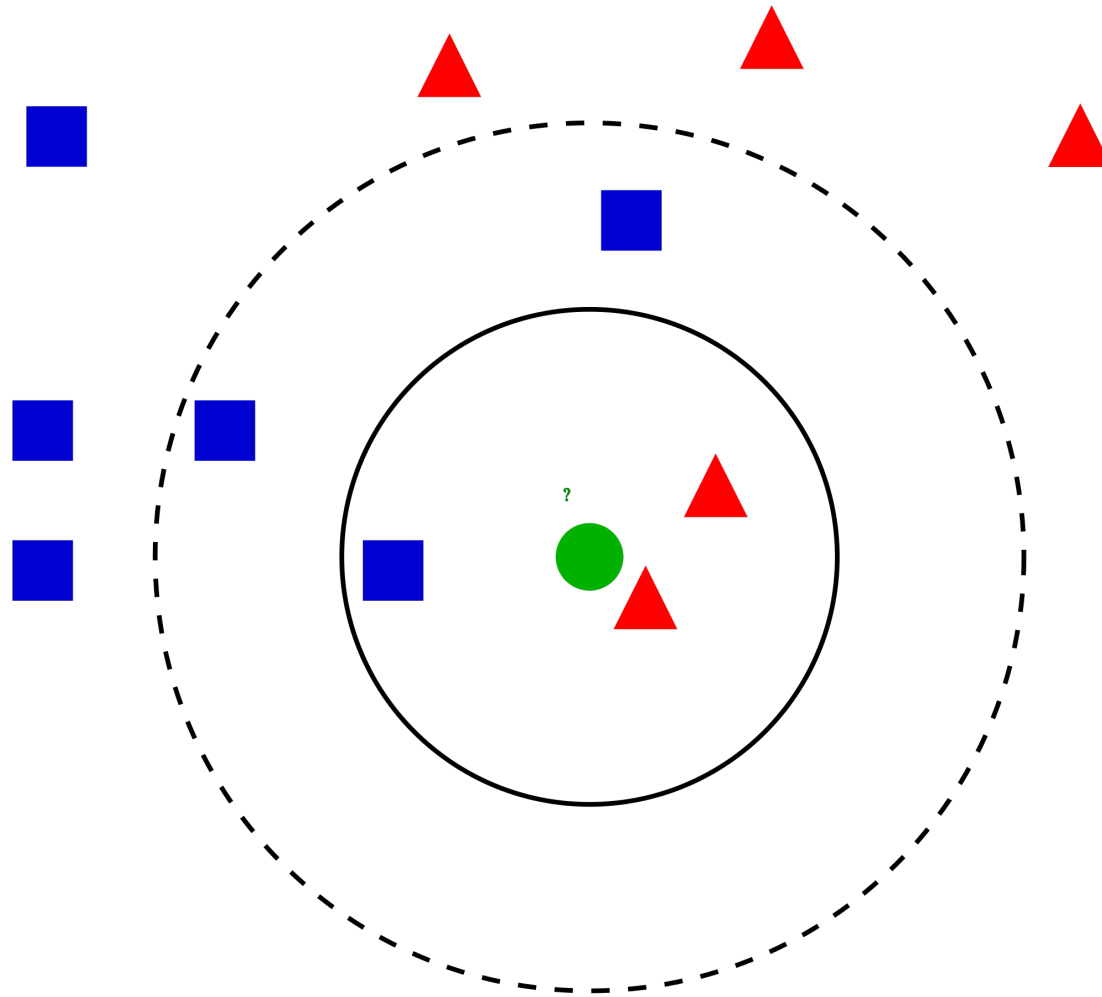
# Ontology Creation



# Ontology Creation



# Classification



**1**

Introduction

**2**

LinkedIn's Vision

**3**

Computing Professional Identity

**4**

**Selected Topics**

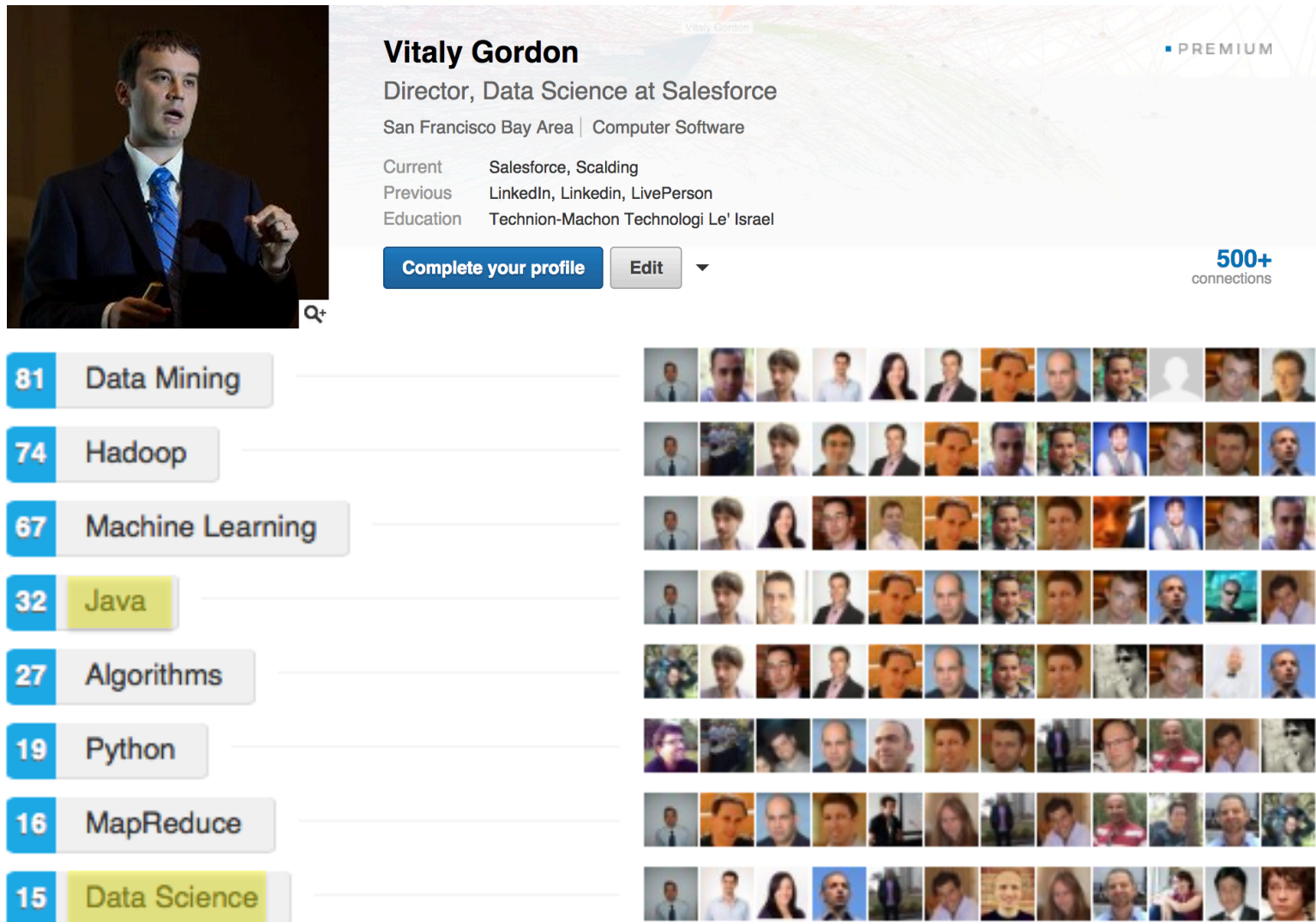
**5**

Summary





# Normalization



**Vitaly Gordon**  
Director, Data Science at Salesforce  
San Francisco Bay Area | Computer Software

Current: Salesforce, Scalding  
Previous: LinkedIn, LinkedIn, LivePerson  
Education: Technion-Machon Technologi Le' Israel

[Complete your profile](#) [Edit](#)

**500+** connections

**Skills and Endorsements:**

- 81 Data Mining
- 74 Hadoop
- 67 Machine Learning
- 32 Java
- 27 Algorithms
- 19 Python
- 16 MapReduce
- 15 Data Science

The grid of connections consists of 7 rows of 10 small profile pictures each, totaling 70 connections.

# Normalization

$$w_{x,y} = tf_{x,y} \times \log \left( \frac{N}{df_x} \right)$$

## TF-IDF

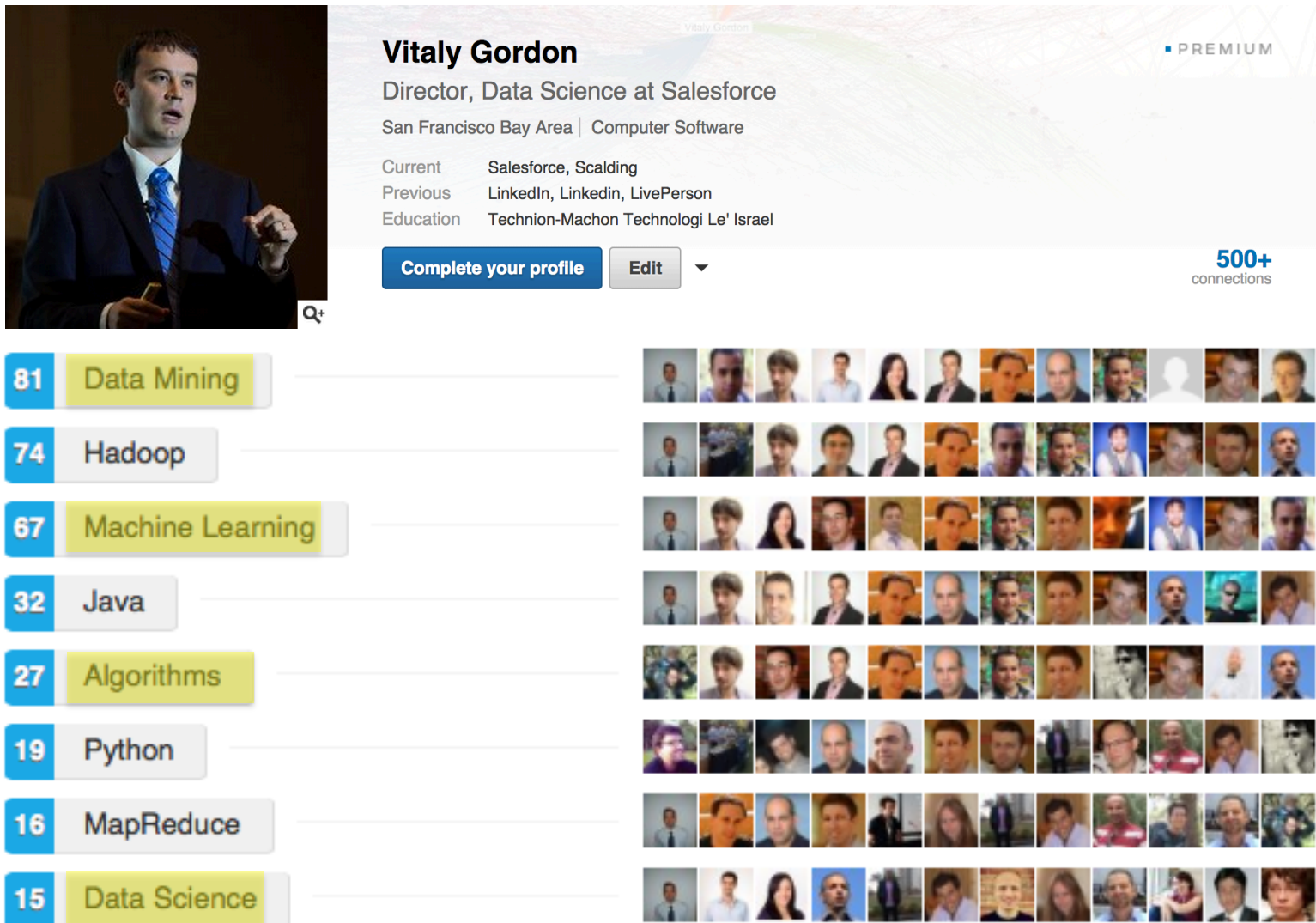
Term  $x$  within document  $y$

$tf_{x,y}$  = frequency of  $x$  in  $y$

$df_x$  = number of documents containing  $x$

$N$  = total number of documents

# Clustering



**Vitaly Gordon**  
Director, Data Science at Salesforce  
San Francisco Bay Area | Computer Software

Current: Salesforce, Scalding  
Previous: LinkedIn, LinkedIn, LivePerson  
Education: Technion-Machon Technologi Le' Israel

[Complete your profile](#) [Edit](#) ▼

500+ connections

**Skills and Endorsements:**

- 81 Data Mining
- 74 Hadoop
- 67 Machine Learning
- 32 Java
- 27 Algorithms
- 19 Python
- 16 MapReduce
- 15 Data Science

The visualization shows a network of connections, with a dense cluster of 15 connections sharing the skill 'Data Science'. Other clusters are visible for 'Machine Learning', 'Algorithms', and 'Python'.

# Clustering

## Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

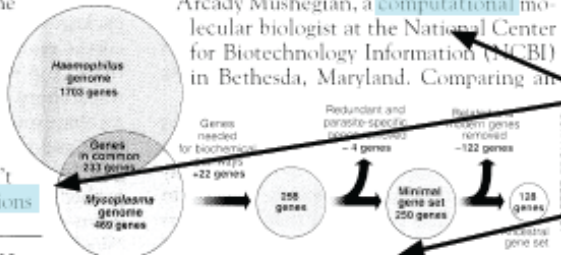
## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a Uppsala University in Sweden. She arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic** numbers game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

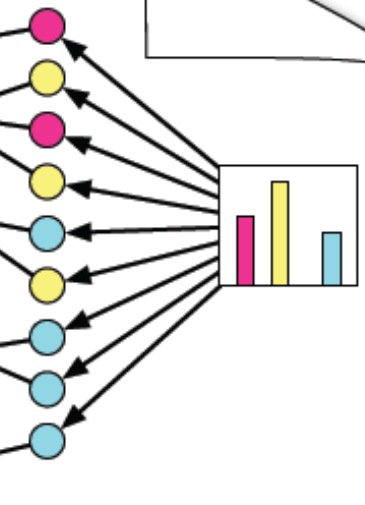


\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

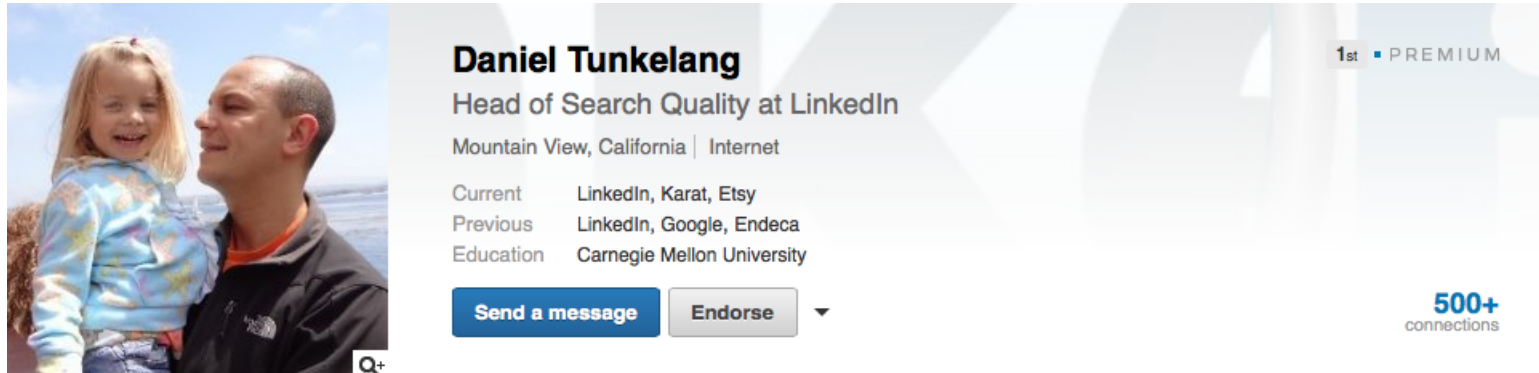
## Topic proportions and assignments



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics



# Anomaly Detection



**Daniel Tunkelang**  
Head of Search Quality at LinkedIn  
Mountain View, California | Internet

Current LinkedIn, Karat, Etsy  
Previous LinkedIn, Google, Endeca  
Education Carnegie Mellon University

1st PREMIUM

Send a message Endorse

500+ connections

Q+

99+ Karaoke +

99+ Information Retrieval +

99+ Machine Learning +

99+ Data Mining +

99+ Text Mining +

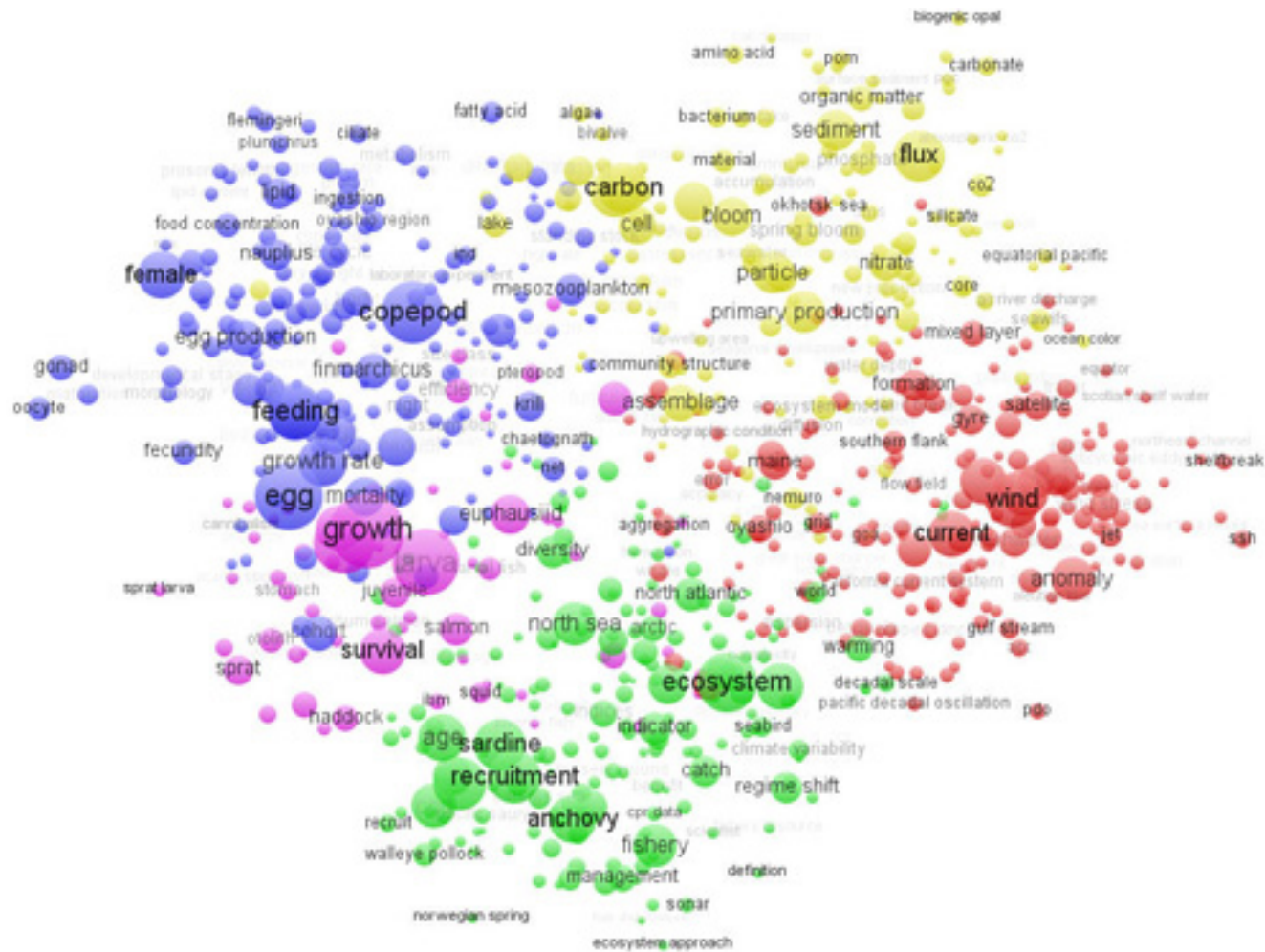
99+ Big Data +

99+ Search +

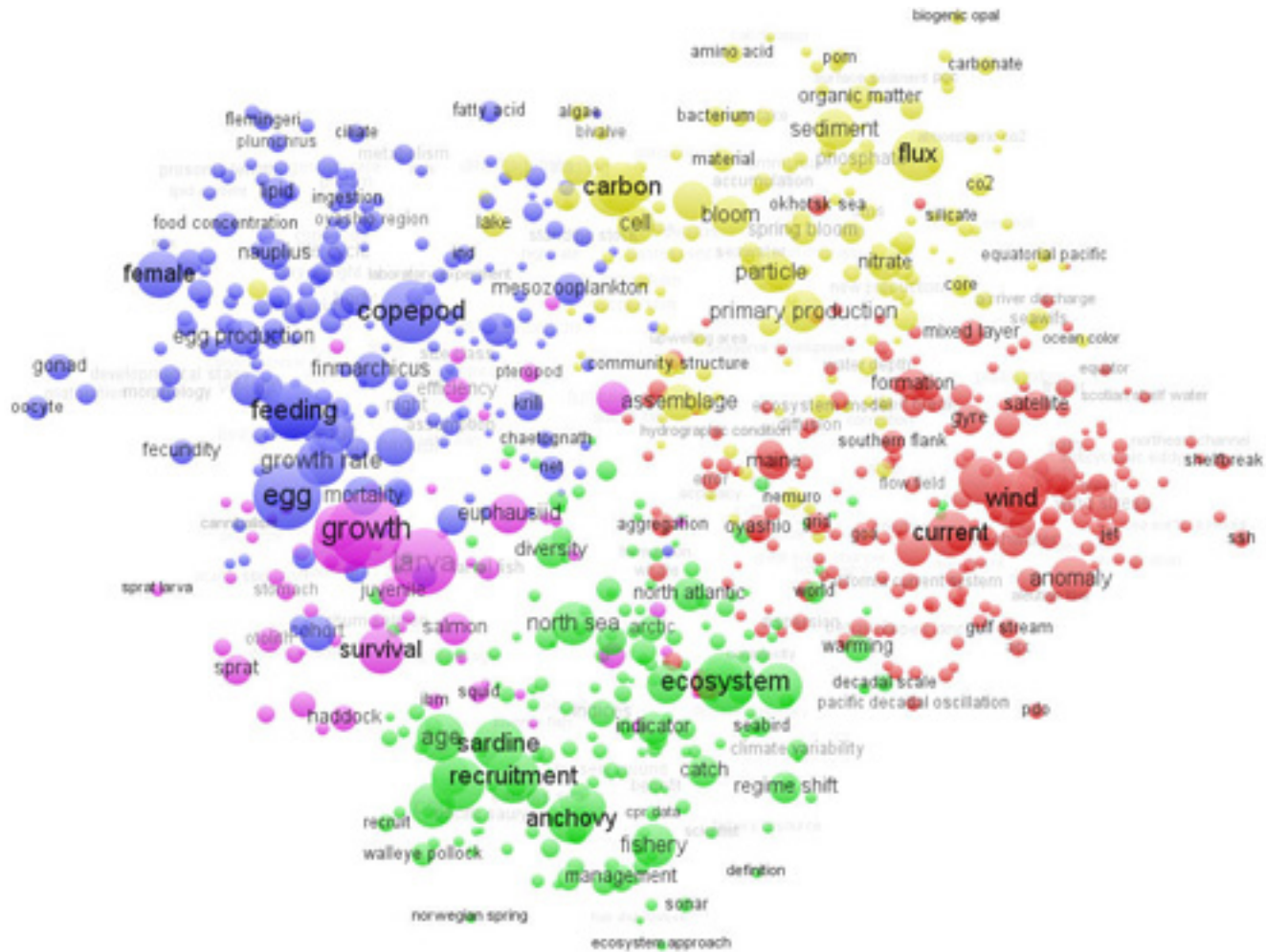
93 Computer Science +



# Anomaly Detection



# Anomaly Detection



<http://www.slideshare.net/tdunning/strata-2014-anomaly-detection>



# Evaluation

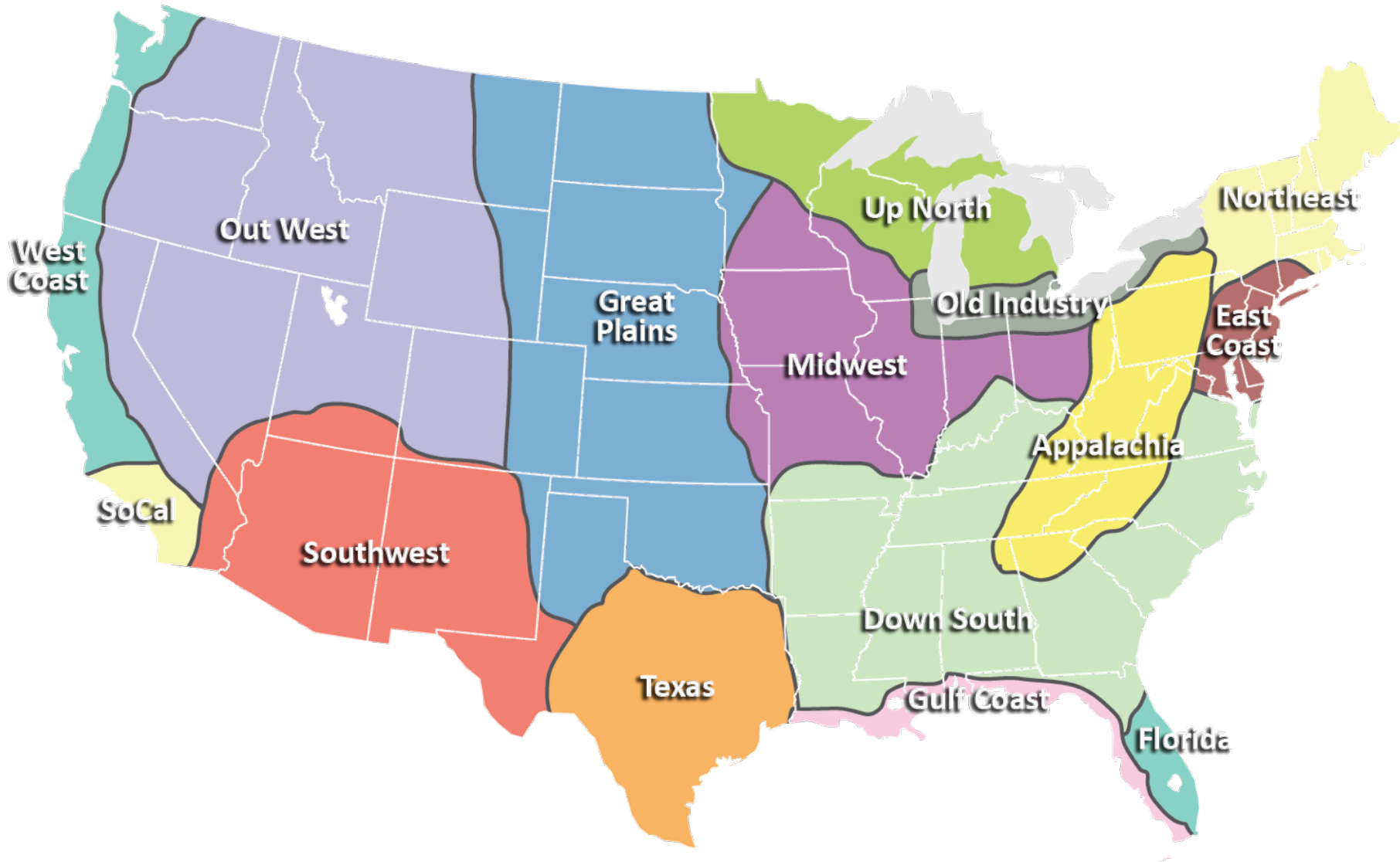




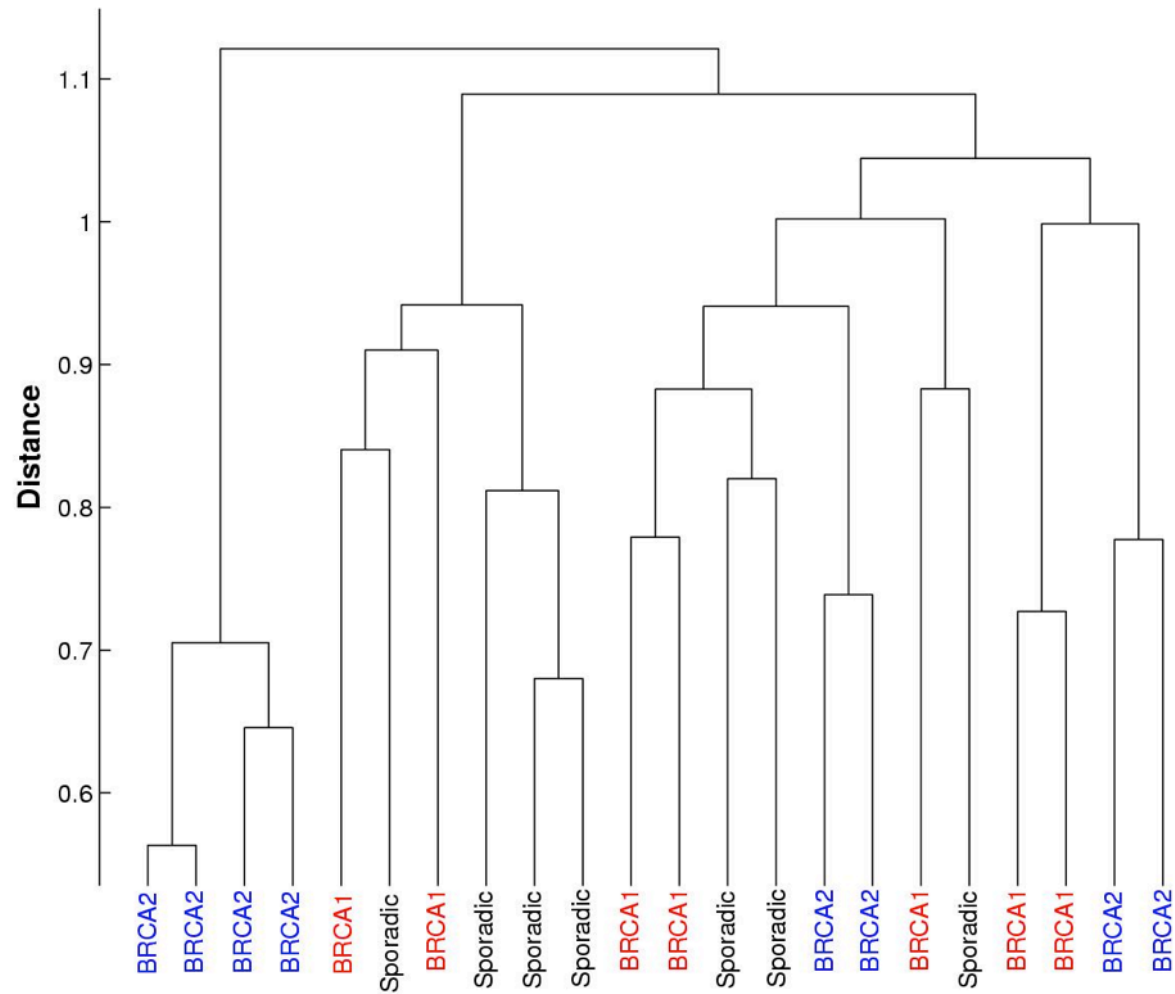
# Evaluation



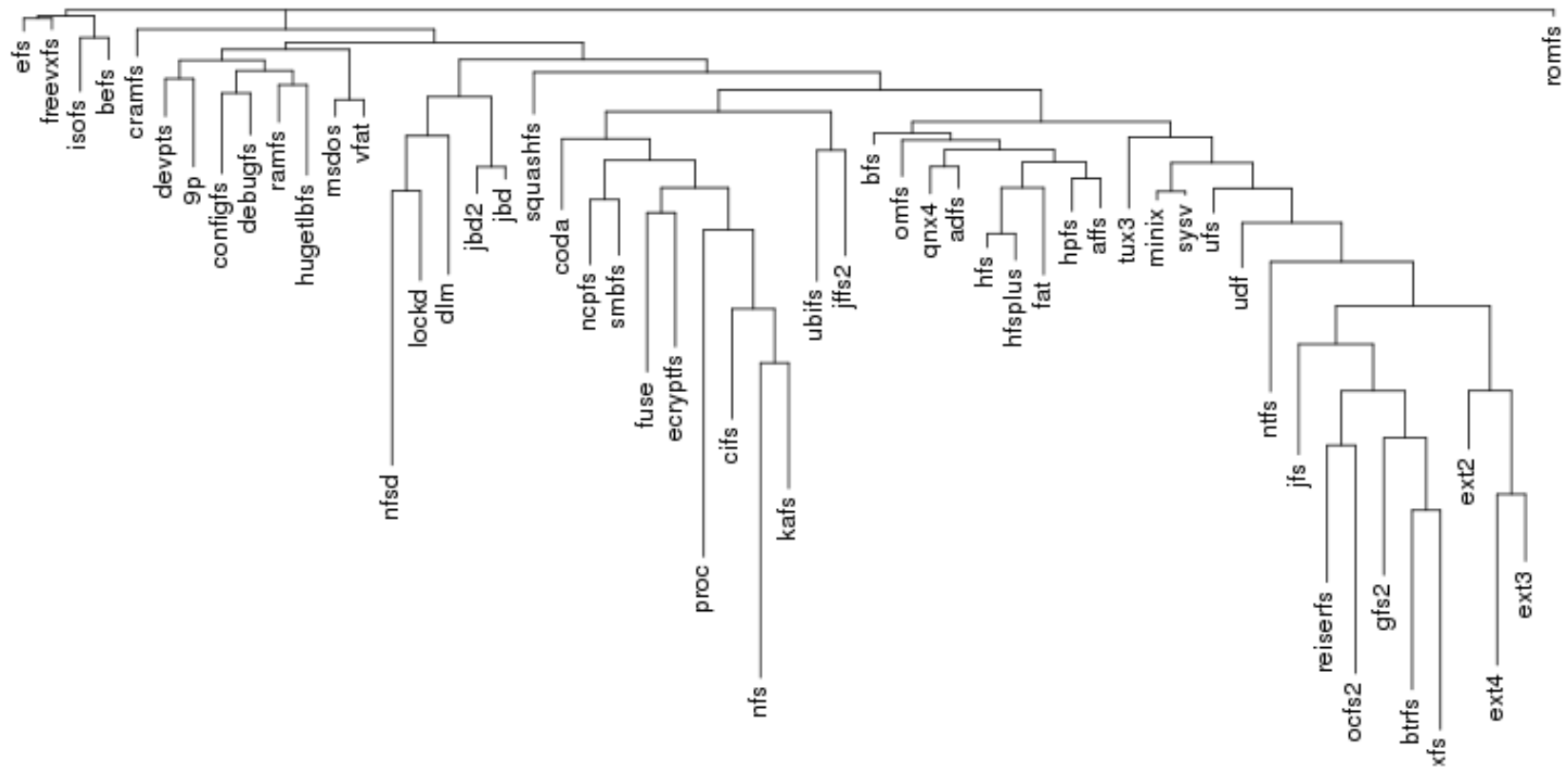
# Evaluation



# Hierarchical Clustering



# Hierarchical Clustering











**1**

Introduction

**2**

LinkedIn's Vision

**3**

Computing Professional Identity

**4**

Selected Topics

**5**

**Summary**





## Summary

- 1. User generated content from 300M members, creates 300M problems**



## Summary

- 1. User generated content from 300M members, creates 300M problems**
- 2. Data cleaning is so much more than filtering out empty values**

## Summary

- 1. User generated content from 300M members, creates 300M problems**
- 2. Data cleaning is so much more than filtering out empty values**
- 3. Try to be creative and work around difficult language problems**

CHICAGO

INTERNATIONAL  
SOFTWARE DEVELOPMENT  
CONFERENCE 2015

goto;  
conference

# Questions?

*Please remember to evaluate via the GOTO  
Guide App*

@bigdatasc

/in/vitalygordon

 follow us @gotochgo

Conference: May 11-12 / Workshops: 13-14