# Table of contents

## 1. Data Science
- Origins
- People
- Work

## 2. Math Behind Data Science
- Experimentation
- Growth Normalization
- If Time

Be challenged at LinkedIn. We’re looking for superb analytical minds of all levels to expand our small team that will build some of the most innovative products at LinkedIn.

No specific technical skills are required (we’ll help you learn SQL, Python, and R). You should be extremely intelligent, have quantitative background, and be able to learn quickly and work independently. This is the perfect job for someone who’s really smart, driven, and extremely skilled at creatively solving problems. You’ll learn statistics, data mining, programming, and product design, but you’ve gotta start with what we can’t teach - intellectual sharpness and creativity.

Figure: LinkedIn Job Posting April 2008
Data Scientist – Growth Analytics at LinkedIn

Data Scientists on our team partner with product managers, engineers and a cross-functional team to drive LinkedIn membership growth and connectivity. We inform product strategy and product decisions by:

- Extracting and analyzing LinkedIn data to derive actionable insights.
- Formulating success metrics for completely novel products and creating dashboards/reports to monitor them.
- Designing and analyzing experiments to test new product ideas.
- Developing models and data-driven solutions that add material lift to principal performance metrics.

LinkedIn member data is amazingly rich and provides a fantastic opportunity for Data Scientists to explore and create, ultimately developing ways for members to improve their professional lives. You'll have the opportunity to work with some of the best data people anywhere in an environment which truly values data-driven decisions. Required qualifications include:

- BS/MS in a quantitative discipline: Statistics, Applied Mathematics, Operations Research, Computer Science, Engineering, Economics, etc.
- 1+ years experience working with large amounts of real data with SQL (Teradata, Oracle, or MySQL) and R, or other statistical package.
- 1+ years work experience programming in Java or Python - Pig experience desired.
- Proficiency in a Unix/Linux environment for automating processes with shell scripting.
- Able to translate business objectives into actionable analyses.
- Able to communicate findings clearly to both technical and non-technical audiences

Preferred Qualifications include:

- Experience with Consumer Internet products.
- Knowledge in one of the following areas is a strong plus: Viral Growth mechanisms, user acquisition in International markets, Search Engine Optimization (SEO)
- Expertise in applied statistics, understanding of controlled experiments.

Figure: LinkedIn Job Posting July 2015
Latest Data Science Job Rec - Applicants

Top skills and areas of expertise among applicants

- R
- SQL
- Matlab
- C++
- Data Mining
- Python
- Data Analysis
- Java
- Machine Learning
- Statistics

Education

- 55% Master's Degree
- 18% Bachelor's Degree
- 18% Doctor of Philosophy
- 9% Other

Top fields of study:
- Computer Science
- Statistics
- Economics

Figure: Applicants now have SQL, Python, and R. 702 applicants in 5 months.
Definition (Data Science as a Victim of Success)
When use of a skill demonstrates improvements in support and innovation, it is added to the next job rec.

Rule of thumb when hiring, does your favorite colleague pass your interview?
Goals

Invariant

- Use data to support colleagues: marketing, finance, engineering, ...
- Use data to innovate: products, strategies, performance, ...

Cherry on Top

- Do what it takes to drive company success.
Progress

1. Data Science
   - Origins
   - People
   - Work

2. Math Behind Data Science
   - Experimentation
   - Growth Normalization
   - If Time
LinkedIn Data

Connecting talent with opportunity at massive scale

- Members: 364M
- Companies: 3.5M
- Jobs: 1M
- Skills: 3B
- Schools: 24k
- Updates: B
Source of 125k Data Professionals

Backgrounds of Data Professionals

- Computer Science
- Economics
- Finance
- Information Technology
- Statistics
- Mathematics
- Psychology
- Marketing
- Accounting
- Physics
- Other

Figure: Incredibly diverse.
Data Professionals on LinkedIn

> 2k degree fields (after standardization)

16% are Unique Degrees:

- Oral Surgery
- Phytopathology
- Wedding Planning
- Ground Transportation
- Library Sciences
- Turfgrass Management
- Embryology
- Fire Fighting
- Stagecraft
- Art Conservation
Data Science Homogenization Trend

Backgrounds of Data Professionals

Top 10 Background %

Year
Uneven Growth of Top 10 Backgrounds

Share of Top 10 Backgrounds

- Computer science
- Marketing

Year

Proportion

2001  2005  2009  2013
Uneven Growth of Top 10 Backgrounds

Figure: Increased recruitment of economists and statisticians.
Industries of Data Professionals

- Information Technology
- Financial Services
- Health Care
- Oil & Energy
- Telecommunications
- Computer Software
- Insurance
- Government Administration
- Banking
- Research
- Other
Industry Diversification of Data Professionals

Industries of Data Professionals

Top 10 Background %

Year
Uneven Growth of Top 10 Industries

Share of Top 10 Industries

- Information Technology
- Telecommunications

Proportion vs. Year

Year:
- 2001
- 2005
- 2009
- 2013

Proportion:
- 0
- 5
- 10
- 15
- 20
- 25
- 30
- 35
Trends

- Homogenization of Sources of Data Professionals

- Diversification of Industry Destinations of Data Professionals
1 Data Science
   - Origins
   - People
   - Work

2 Math Behind Data Science
   - Experimentation
   - Growth Normalization
   - If Time
Figure: What portion of work data scientists do on a daily basis depends on product life cycle.
Figure: Content spreads along existing connection network.
**Follow Network**

**Demographics of your readers**

<table>
<thead>
<tr>
<th>Top industries</th>
<th>Top job titles</th>
<th>Top locations</th>
<th>Top traffic sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>16% Information Technology a…</td>
<td>16% Salesperson</td>
<td>30% San Francisco Bay Area</td>
<td>58% LinkedIn.com</td>
</tr>
<tr>
<td>13% Computer Software</td>
<td>9% Consultant</td>
<td>11% Greater New York City Area</td>
<td>35% LinkedIn Pulse</td>
</tr>
<tr>
<td>9% Internet</td>
<td>8% Laboratory Scientist</td>
<td>6% Greater Chicago Area</td>
<td>2% LinkedIn Profile</td>
</tr>
<tr>
<td>7% Financial Services</td>
<td>8% Marketing Specialist</td>
<td>6% London, United Kingdom</td>
<td>2% Groups</td>
</tr>
<tr>
<td>7% Hospital &amp; Health Care</td>
<td>7% Business / Corporate Str…</td>
<td>5% Toronto, Canada Area</td>
<td>1% Google Search</td>
</tr>
<tr>
<td>6% Marketing and Advertising</td>
<td>7% Project Manager</td>
<td>5% Greater Seattle Area</td>
<td>1% Homepage</td>
</tr>
<tr>
<td>5% Management Consulting</td>
<td>6% Human Resources Spec…</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% Human Resources</td>
<td>5% Software Developer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4% Higher Education</td>
<td>5% Corporate Trainer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4% Banking</td>
<td>4% Research / Graduate Ass…</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure:** Change the game. Increase readership and visibility via follows.
# Product Cycle - Follow Network

<table>
<thead>
<tr>
<th>Stage</th>
<th>Work</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideation</td>
<td>Explore how to make content go big. <em>Follows.</em></td>
<td>2 weeks</td>
</tr>
<tr>
<td>Design &amp; Spec</td>
<td>Define a <em>Follow</em> for security, PR, marketing, all teams possibly affected.</td>
<td>3 weeks</td>
</tr>
<tr>
<td>Development</td>
<td>Database engineering, rollback safe, experimental framework.</td>
<td>6 months</td>
</tr>
<tr>
<td>Test &amp; Iterate</td>
<td>Slow release experiment.</td>
<td>3 months</td>
</tr>
<tr>
<td>Release</td>
<td>Clean up code, outline fast follows</td>
<td>1 month</td>
</tr>
</tbody>
</table>

**Table:** Follow Network, slow and steady development cycle.
<table>
<thead>
<tr>
<th>Area of Data</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyze</td>
<td>Understand</td>
</tr>
<tr>
<td>Visualize</td>
<td>Communicate</td>
</tr>
<tr>
<td>Business Decisions</td>
<td>Orchestrate Action</td>
</tr>
<tr>
<td>Prototype Product</td>
<td>Demonstrate Usefulness</td>
</tr>
<tr>
<td>Refine Product</td>
<td>Maximize Usefulness</td>
</tr>
<tr>
<td>Design Experiment</td>
<td>Measure Changes</td>
</tr>
<tr>
<td>Analyze Experiment</td>
<td>Learn</td>
</tr>
<tr>
<td>Log</td>
<td>Save Everything</td>
</tr>
<tr>
<td>Process</td>
<td>Make Data Useable</td>
</tr>
<tr>
<td>Load to Server/DB</td>
<td>Make Data Accessible</td>
</tr>
</tbody>
</table>

**Table:** General data science stack.
Who does What

**Figure:** Depth v. breadth of different fields.
### Skills of Data Professionals

<table>
<thead>
<tr>
<th>Languages</th>
<th>Tools</th>
<th>Hard Skills</th>
<th>Soft Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQL</td>
<td>Microsoft (Office, Excel, SQL, Visio)</td>
<td>Research</td>
<td>Management</td>
</tr>
<tr>
<td>Java</td>
<td>Oracle</td>
<td>Statistics</td>
<td>Leadership</td>
</tr>
<tr>
<td>Matlab</td>
<td>SAS</td>
<td>ETL</td>
<td>Process Improvement</td>
</tr>
<tr>
<td>Javascript</td>
<td>SharePoint</td>
<td>Data Modeling</td>
<td>Customer Service</td>
</tr>
<tr>
<td>R</td>
<td>SAP</td>
<td>Software Dev</td>
<td>Software Docs</td>
</tr>
<tr>
<td>Python</td>
<td>Cisco</td>
<td>Data Mining</td>
<td>Strategy</td>
</tr>
<tr>
<td>C++</td>
<td>Salseforce</td>
<td>Forecasting</td>
<td>Public Speaking</td>
</tr>
<tr>
<td>XML</td>
<td>Six Sigma</td>
<td>Database Design</td>
<td>Team Leadership</td>
</tr>
</tbody>
</table>

Table: From LinkedIn’s 125k Data Professionals.
1. Data Science
   - Origins
   - People
   - Work

2. Math Behind Data Science
   - Experimentation
   - Growth Normalization
   - If Time
Traditional A/B Testing

Figure: Traditional ab testing. [Salesforce]

High Level
Randomly divides users into two groups for different treatments.
Social Influence

Figure: Users can communicate experiences in social networks.

Cross Over

Testing interaction features such as messaging, connections, and profile views inherently have cross cohort communication.
High Level

Partition network into relatively low intra communication groups.

Figure: See geographical bounds. [Ugander et al]
Elegant Solution

Downside
Costly to implement and assign elegant solution. Limited number of experiments can run simultaneously.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Actual Performance</th>
<th>Observed Performance</th>
<th>Observed Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>x</td>
<td>z</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>y</td>
<td>c \cdot z</td>
<td>c - 1</td>
</tr>
</tbody>
</table>

Table: What exists and is observed.
2 equations, 3 variables, can compute upper bound for $\frac{x}{y}$.
Elegant v. Brute Force Tradeoff

Bound

Actual impact \( a \) is bounded by observed impact \( c \) & viral coefficient \( V \):

\[
a = \frac{c - V}{1 - cV}
\]

Figure: Small impact for low viral products. [Andrews]
Alternative Brute Force

Control Interactions

Split on the interaction at the cost of inconsistent user experience. Benefit is test the impact of sending or receiving.

<table>
<thead>
<tr>
<th>Sender / Receiver</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Treatment</td>
<td>Control</td>
</tr>
<tr>
<td>B</td>
<td>Control</td>
<td>Control</td>
</tr>
</tbody>
</table>
Progress

1. Data Science
   - Origins
   - People
   - Work

2. Math Behind Data Science
   - Experimentation
   - Growth Normalization
   - If Time
Figure: Since 2008 Health Care has increased relationships with Recruiters.
Figure: Growth of relationships is dominated by LinkedIn’s growth.
Control Confounding Variables

Data quality and growth can dominate underlying trends.

- LinkedIn’s Network Growth is massive and diverse
- Venture Capitalists and Recruiters are hyper connectors

Figure: Stan Lee
Control for Growth and Behavioral Variables

Approach

Set as constants the number of users in an industry and how many connections they have. Then reconnect connections at random.

Figure: Break edges and reconnect randomly.
Expected Connections

Closed Form Solution

Reducible to pulling red and blue balls from a bag without replacement. The solution is the expectation of the Hypergeometric distribution.

\[ E[|\text{Edges(Health Care, I)}|] = \frac{|\text{Edges(Health Care)}||\text{Edges(I)}|}{\sum_{i,j} |\text{Edges}(i,j)| - |\text{Edges(Health Care)}|} \]
Figure: Given growth and behavioral patterns, we expect some industries to have a dramatic number of connections to health care professionals.
Significant Relations with Health Care Appear

Figure: Venture Capitalists and Recruiters are no longer in the top rankings.
Significant Relations with Health Care Appear

Relations Now Reflect the Larger Economy

- City programs have increased in-home and preventative care
- Many hospitals are named after Saints and affiliated with Religious Denominations
- Medical Devices and Pharmaceuticals have and have always had strong connection to Health Care

Figure: Industries with Significant Connections to Health Care

- Medical Devices
- Individual & Family Services
- Pharmaceuticals
- Religious Institutions
- Non-Profit Organization Management
- Public Safety
- Professional Training & Coaching
Significant Relations with Realtors

Connections with Realtors

Figure: Period of dramatic growth for real estate
Significant Relations with Realtors

Connections with Realtors

Figure: Period of economic change
Significant Relations with Construction

**Figure:** Symmetric relationship between real estate and construction. Construction workers migrate between real estate and oil and mining.
Construction Worker Migration

Figure: Construction workers connecting with Oil & Mining over Real Estate
Industry Migration - Mechanics

How?

- Is migration prompted by influential people?
- Is migration independent pockets of movement?
Figure: Median is 4 neighbors migrated before conversion
Figure: Size of bubble is proportional to size of complete cascade.

How?

Migration is largely independent, with some cascades.
1 Data Science
   • Origins
   • People
   • Work

2 Math Behind Data Science
   • Experimentation
   • Growth Normalization
   • If Time
Figure: Data growth is exponential. Rule of thumb is doubles every 4-8 months.
**Figure:** Linear algorithms are fast, predictable, and complete.
Figure: Every project involved at least 3 people.
Progress

1 Data Science
- Origins
- People
- Work

2 Math Behind Data Science
- Experimentation
- Growth Normalization
- If Time
MAP combines:

- **Precision** - Give me only what I want
- **Recall** - Give me everything I want

*Figure: (Precision, Recall) values with same MAP score.*
Figure: Two sets of (Precision, Recall) values with same MAP score.
Figure: Improve a search algorithm from point a with either small increase in Recall or a large increase in Precision.
Figure: Additional points.
MAP

Snake Oil

When Precision and Recall values are not balanced, MAP only responds to changes in the lower one.

North Star

When Precision and Recall values are balanced, promotes improvement of both Precision and Recall.