Demystifying The Data Scientist

Natasha Balac, Ph.D.
Predictive Analytics Center of Excellence,
Director
San Diego Supercomputer Center
University of California, San Diego
**Brief History of SDSC**

- **1985-1997**: NSF national supercomputer center; managed by General Atomics
- **1997-2007**: NSF PACI program leadership center; managed by UCSD
  - PACI: Partnerships for Advanced Computational Infrastructure
- **2007-2009**: Internal transition to support more diversified research computing
  - still NSF national “resource provider”
- **2009-future**: Multi-constituency cyberinfrastructure (CI) center
  - provide data-intensive CI resources, services, and expertise for campus, state, and nation
- **Approaching $1B in lifetime contract and grant activity**
PACE
Predictive Analytics Center of Excellence

Closing the gap between
Government, Industry and Academia
PACE: Closing the gap between Government, Industry and Academia

PACE is a non-profit, public educational organization
- To promote, educate and innovate in the area of Predictive Analytics
- To leverage predictive analytics to improve the education and well being of the global population and economy
- To develop and promote a new, multi-level curriculum to broaden participation in the field of predictive analytics
Develop Standards and Methodology

- CRISP-DM
- PMML
- Data Mining Standards:
  - Data Warehouse, Web, API and Grid standards
- Benchmarking
- NIST Big Data WG

Inform, Educate and Train

Bridge the Industry and Academia Gap

Provide Predictive Analytics Services

Data Mining Repository of Very Large Data Sets

Foster Research and Collaboration

Predictive Analysis Center of Excellence

Develop Standards and Methodology

High Performance Scalable Data Mining
A “data-intensive” supercomputer based on SSD flash memory and virtual shared memory

- Emphasizes MEM and IO over FLOPS
- Scalable Data Mining
- Visualizations
Foster Research and Collaboration

- Fraud Detection
- Modeling user behaviors
- Smart Grid Analytics
- Solar powered system modeling
- Microgrid anomaly detection
- Battery Storage Analytics
- Sport Analytics
- Genomics
UCSD Smart Grid

- UCSD Smart Grid sensor network data set
  - 45MW peak micro grid; daily population of over 54,000 people
  - Self-generate 92% of its own annual electricity load

- Smart Grid data – over 100,000 measurements/sec
  - Sensor and environmental/weather data
    - Large amount of multivariate and heterogeneous data streaming from complex sensor networks

- Predictive Analytics throughout the Microgrid
Solar Power Integration Through Predictive Analytics

Solar + EV integration

Predictive analytics

Business & Societal change

Battery Storage Analytics
Sustainable San Diego Partnership

- Clean Tech San Diego, OSIsoft, SDG&E and UC San Diego Common data infrastructure connects physical assets: electrical, gas, water, waste, buildings, transportation & traffic
- Platform to securely transfer high volumes of Big Data from multiple, distributed measurement units
- Crowd-sourced Big Data in a cyber-secure, private cloud
- Predictive analytics on real-time time-series data

White House Big Data Event: “Data to Knowledge to Action” – Launch Partners Award
**Big Data from Microgrids**

- Complexities introduced by the large amount of multivariate and heterogeneous data streaming from complex sensor networks
- Extremely large, complex sensor networks, enabling a novel feature reduction method that scales well
Number of V’s of Big Data

**Volume**
- Data at Rest
  - Terabytes to exabytes of existing data to process

**Velocity**
- Data in Motion
  - Streaming data, milliseconds to seconds to respond

**Variety**
- Data in Many Forms
  - Structured, unstructured, text, multimedia

**Veracity**
- Data in Doubt
  - Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

IBM, 2012
What Happens in an **Internet Minute**?

- **639,800 GB of global IP data transferred**
- **20** New victims of identity theft
- **204 million** Emails sent
- **20 million** Photo views
- **47,000** App downloads
- **583,000** In sales
- **1,300** New mobile users
- **100+** New LinkedIn accounts
- **277,000** Logins
- **2+ million** Search queries
- **320+** New Twitter accounts
- **61,141** Hours of music
- **3,000** Photo uploads
- **1,3 million** Video views
- **6** New Wikipedia articles published
- **135** Botnet infections
- **30** Hours of video uploaded
- **100,000** New tweets
- **20 million** Products sold

And **Future Growth is Staggering**

- **Today, the number of networked devices = global population**
- **By 2015, the number of networked devices = 2x global population**
- **In 2015, it would take you 5 years to view all video crossing IP networks each second**

---

**SAN DIEGO SUPERCOMPUTER CENTER**
at the **UNIVERSITY OF CALIFORNIA; SAN DIEGO**
Where is your DATA?

Research says digital data will grow to 2.75 zettabytes in 2012 and rocket toward nearly 8 zettabytes by 2015.

How are we creating, replicating, saving, mining, and analyzing all of this data? What does our data driven reality of today tell us about the future?
BIG DATA

SAN DIEGO SUPERCOMPUTER CENTER
at the UNIVERSITY OF CALIFORNIA; SAN DIEGO
What Is The Value In Big Data?
What to do with big data?
ERIC SALL’s list

• Big Data Exploration
  • To get an overall understanding of what is there

• 360 degree view of the customer
  • Combine both internally available and external information to gain a deeper understanding of the customer

• Monitoring Cyber-security and fraud in real time

• Operational Analysis
  • Leveraging machine generated data to improve business effectiveness

• Data Warehouse Augmentation
  • Enhancing warehouse solution with new information models and architecture
Big Data – Big Training

• “Data Scientist”
  • The “Hot new gig in town”
    • O’Reilly report
  • Data Scientist: The Sexiest Job of the 21st Century
    • Harvard Business Review, October 2012
    • The next sexy job in next 10 years will be statistician” – Hal Varian, Google Chief Economist
    • Geek Chic – Wall Street Journal – new cool kids on campus
  • The future belongs to the companies and people that turn data into products

• “The human expertise to capture and analyze big data is both the most expensive and the most constraining factor for most organizations pursuing big data initiatives” – Thomas Davenport

• New curriculum – Boot camps, Certificates, Data Science Institute, ‘14 MAS
Data scientist: The hot new gig in tech

• Article in Fortune
  • “The unemployment rate in the U.S. continues to be abysmal (9.1% in July), but the tech world has spawned a new kind of highly skilled, nerdy-cool job that companies are scrambling to fill: data scientist”

• McKinsey Global Institute “Big data Report”
  • By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions
Data Science Job Growth

By 2018 shortage of 140-190,000 predictive analysts and 1.5M managers / analysts in the US
What is Data Science?
Who are Data Scientists?

• The buzz and the rise of Analytics culture
• Big Data and Data science
• Defining Data Science
  • Who?
  • What?
  • Where?
  • How?

• The new “Cool Kids on Campus”
What is Data Science?
Who are Data Scientists?
Rising Part of the Culture and Enterprises

• Movies and bestsellers

• Enterprises competing
  • Chief Data Officer now Chief Data Scientist
  • Data Driven Decisions strategic, tactical, operational
Past and Present

- Traditional approaches have been for DM experts: “White-coat PhD statisticians”
  - DM tools also fairly expensive
- Today: approach is designed for those with some Database/Analytics skills
  - DM built into DB, easy to use GUI, Workflows
  - Many jobs available from statistical analyst to data scientist – but many people are lacking education and experience!
  - Hands on experience is extremely valuable
Data Science Is Multidisciplinary

Adapted from B. Tierney, 2013
Successful Data Scientist Characteristics

• Intellectual curiosity, Intuition
  • Find needle in a haystack
  • Ask the right questions – value to the business

• Communication and engagements

• Presentation skills
  • Let the data speak but tell a story
  • Story teller – drive business value not just data insights

• Creativity
  • Guide further investigation

• Business Savvy
  • Discovering patterns that identify risks and opportunities
  • Measure
Big Data And Beyond Fueling Data Science

- Beyond Relational - click streams, machine data
- Beyond Structure - raw complex data that defines metadata and structure
- Beyond Warehouse - Hadoop, Mapreduce, NoSQL
- Advanced analytics - beyond BI, dashboards and warehouses
- Expanded views
  - Behavioral data, social media data, multiple sources
- Sample sizes explode
- Correlation vs. Causality
- Deep learning
Data Scientist Job Description

Amazon’s Shopper Marketing & Insights team focuses on serving the advertisers and our overall ad business to provide strategic media planning, customer insights, targeting recommendations, and measurement and optimization of advertising.

We are hiring outstanding Data Scientists who will use innovative statistical and machine learning approaches to drive advertising optimization and contribute to the creation of scalable insights. The ideal candidate should have one hand on the white-board writing equations and one hand on the keyboard writing code.
Data Scientist Qualities

Oh! You're dropping your numbers!

No worries hon! I've so many of them. I'm a data scientist.

Data Scientist:
The Sexiest Job of the 21st Century
Data Scientist Qualities

The Data Scientist

A New Role Exist – the **Data Scientist**

- One part Scientist/Statistician
- Two parts Sleuth/Artist
- One Part Programmer
- Working with business leaders and process owners to create business value
Analyzing the Analyzers

• O’Reilly Strata Survey – Harris, Murphy & Vaisman, 2013
• Based on how data scientists think about themselves and their work, not
  • Years of experience,
  • Academic degrees, favorite tools
  • Titles, pay scales, org charts
• Identified four Data Scientist clusters
Data Scientist Self-ID

<table>
<thead>
<tr>
<th>Data Developer</th>
<th>Developer</th>
<th>Engineer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Researcher</td>
<td>Researcher</td>
<td>Scientist</td>
</tr>
<tr>
<td>Data Creative</td>
<td>Jack of All Trades</td>
<td>Artist</td>
</tr>
<tr>
<td>Data Businessperson</td>
<td>Leader</td>
<td>Businessperson</td>
</tr>
</tbody>
</table>

O’Reilly Strata Survey suggested Self-ID Group, along with the self-ID categories most strongly associated with each Group
# Strata Survey Skills

<table>
<thead>
<tr>
<th>Business</th>
<th>ML / Big Data</th>
<th>Math / OR</th>
<th>Programming</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product Development</td>
<td>Unstructured Data</td>
<td>Optimization</td>
<td>Systems Administration</td>
<td>Visualization</td>
</tr>
<tr>
<td>Business</td>
<td>Structured Data</td>
<td>Math</td>
<td>Back End Programming</td>
<td>Temporal Statistics</td>
</tr>
<tr>
<td></td>
<td>Machine Learning</td>
<td>Graphical Models</td>
<td>Front End Programming</td>
<td>Surveys and Marketing</td>
</tr>
<tr>
<td></td>
<td>Big and Distributed Data</td>
<td>Bayesian / Monte Carlo Statistics</td>
<td></td>
<td>Spatial Statistics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Algorithms</td>
<td></td>
<td>Science</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simulation</td>
<td></td>
<td>Data Manipulation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Classical Statistics</td>
</tr>
</tbody>
</table>
Range of Skills

• Engineering analogy

• Define Analyst by the breath of skills
  • most successful data scientists are those with substantial, deep expertise in at least one aspect of data science, be it statistics, big data, or business communication

• T-Shaped Skills
  • Data science is an inherently collaborative and creative
  • Skills terminology
  • ID career path development

Strata survey, 2013
Learning and Training Opportunities

• Many MS, MAS, Courses, Training, Workshops, Certificates, Boot camps, etc.

• Introduction to Data Science Example
  • Part 1: Data Manipulation at scale
    • Databases and the relational algebra
    • Parallel databases, parallel query processing, in-database analytics, MapReduce, Hadoop, relationship to databases, algorithms, extensions, languages
    • Key-value stores and NoSQL; Entity resolution, record linkage
  • Part 2: Analytics, Predictive Analytics, Text mining
  • Part 3: Communicating Results
    • Visualization, data products, visual data analytics
    • Provenance, privacy, ethics, governance

https://www.coursera.org/course/datasci
From Theory To Practice

• Internships – large organizations
• Mentoring
• Organization Integration
• Data Science Team
  • direct access to both raw data and decision-makers
  • diversity of skills to make best use of that access
  • Avoid silos – rotated among internal teams
• Career Paths – to manage or not to manage?
**How Long Does It Take For a Beginner to Become a Good Data Scientist?**

<table>
<thead>
<tr>
<th>Number of Year</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1 year (6)</td>
<td>2%</td>
</tr>
<tr>
<td>1-2 years (33)</td>
<td>12%</td>
</tr>
<tr>
<td>2-4 years (85)</td>
<td>31%</td>
</tr>
<tr>
<td>5-8 years (91)</td>
<td>33%</td>
</tr>
<tr>
<td>&gt; 8 years (35)</td>
<td>13%</td>
</tr>
<tr>
<td>Not sure (28)</td>
<td>10%</td>
</tr>
</tbody>
</table>

KDnuggets survey [278 votes total]
How long does it take for a beginner to become a good data scientist per Region?

<table>
<thead>
<tr>
<th>Region (Count)</th>
<th>Avg Years to become a good data scientist</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU/NZ (9)</td>
<td>6.9 years</td>
</tr>
<tr>
<td>E. Europe (19)</td>
<td>5.9 years</td>
</tr>
<tr>
<td>US/Canada (143)</td>
<td>4.9 years</td>
</tr>
<tr>
<td>W. Europe (60)</td>
<td>4.9 years</td>
</tr>
<tr>
<td>Asia (25)</td>
<td>4.9 years</td>
</tr>
<tr>
<td>Africa/Middle East (9)</td>
<td>4.4 years</td>
</tr>
<tr>
<td>Latin America (12)</td>
<td>3.9 years</td>
</tr>
</tbody>
</table>
Key to a Great Data Scientist

Τεχνικά προόρισμα (Προγραμματική, Στατιστική, Μαθηματικά)
+ Χαρακτηριστικά
+ Χρεατισμός
+ Ειδυλλιακός
+ Εμπειρία
+ Εμπειρία Σκηνής
+ Βιομηχανικό Συνασπισμός
= Μεγάλος Δεδομένος Φυσικός!
Key to a Great Data Scientist

Technical skills (Coding, Statistics, Math)
  + Commitment
  + Creativity
  + Intuition
  + Presentation Skills
  + Business Savvy

= Great Data Scientist!
PACE Education

Educating the new generation of Data Scientist

• Data mining Boot Camps
  • Boot Camp 1
    • February, 2014
  • Boot Camp 2
    • March 2014
  • On-site training

• Tech Talks - 3rd Wednesday

• Workshops – one day “latest and greatest”
  • Hadoop, PMML

• Institute for Data Science and Engineering

• MAS in Data Science in Fall ‘14
KEEP CALM AND USE DATA WISELY
Questions?

- [www.sdsc.edu](http://www.sdsc.edu)
- For further information, contact Natasha Balac (nbalac@ucsd.edu)

Thank you!