Building Data Science Teams

David Dietrich
Advisory Technical Education Consultant
EMC Education Services

@imdaviddietrich

Predictive Analytics World Boston
September 30, 2013
Quick Profile

• Architect of EMC’s Data Science curriculum

• Co-author of 3 courses on Big Data and Data Science (with FOSS)

• Filed 9 patents on data science, data privacy, and cloud computing

• Advisor to universities on analytic programs (Babson, Harvard…)

Twitter: @imdavidddietrich
Today’s Discussion Questions

• When do you need to build a team? How do you know? What are the needs? What kind of people do you need on the team?

• Organizational model? Where should the data science team live within an organization?

• Is it always necessary to build a team? When to build, buy or partner?

• Who are the sponsors in large companies?
Do You Need A Data Science Team For This?
Creating Reports, Dashboards, and Databases...

Percentage of Analytic Computer Usage by Type

- Extended Enterprise User (Demand-side/B2B/B2C and retailer operations)
- Casual User (Dashboards and enterprise reporting)
- Business User (Scorecards, performance mgmt, business reporting, and packaged apps)
- Power User (Statistical analysis, analytical reporting and OLAP)
- IT User (Application development, data mining and web data mining)

Key:
- EMC NetWorker Installation, Configuration and Administration
- RecoverPoint System Deployment
- Symmetrix Configuration Management
- VMware vSphere: Install, Configure, Manage [V5.1]
- VNX Unified Storage Implementation

Graphs showing various data distribution and analysis.
How About For This?
Creating a Map of the Internet
Example: Output From a Data Science Team
Mapping The Spread of Innovation Ideas Using Social Graphs

Betweenness Ranks
1. 578
2. 511
3. 341
4. 171
5. 138

* Only includes entries with more than one submitter
Case Study Summary: EMC GINA

Initial Questions

• Can we find out how information spreads across a worldwide company?
• How can we figure out the main topics discussed in verbal conversations with universities in different countries?
• Which people are providing thought leadership and influence, and how would we know?

Findings

• Advanced analytics identified “hidden innovators” in China and Ireland – people with a disproportionate influence on ideas at EMC
• Ireland in particular had a high density of innovation award finalists, due to training and support of general management
Data Science Teams
Framework for Developing Data Science Teams

Data Science Team

- Data Scientist
- BI Analyst
- Project Sponsor
- Project Manager
- Business User
- Data Engineer
- DBA
Successful Analytic Projects Require Breadth of Roles

- Business User
- Project Sponsor
- Project Manager
- Business Intelligence Analyst
- Database Administrator (DBA)
- Data Engineer
- Data Scientist

© Copyright 2013 EMC Corporation. All rights reserved.
Data Science Team for **GINA Project**

- Volunteer Data Scientists, with graduate level backgrounds in machine learning, NLP, and complex network analysis
- IT dept provided DBA and Data Engineer skills
- Project sponsor (VP in CTO office) provided overall oversight (project mgmt, domain knowledge, business user perspective, and evangelism)
- Created analytic sandbox for team members to use and analyze data
Developing Data Science Capabilities
Framework for Developing Data Science Teams

Developing Data Science Capabilities

- Transforming
- Creating
- As-a-Service
- Crowdsourcing

Data Science Team

- Data Scientist
- BI Analyst
- Project Sponsor
- Project Manager
- Business User
- Data Engineer
- DBA
Four Approaches to Developing Data Science Capabilities

Transforming

Creating

As a Service

Crowdsourcing
Developing Data Science Capabilities for **GINA Project**

- Capabilities model a mix of Crowdsourcing and aaS ("volunteer aaS model")
- Team formed for a specific project, several Data Scientists working on different aspects
Framework for Developing Data Science Teams

Organizational Model
- Centralized
- Decentralized
- Hybrid

Developing Data Science Capabilities
- Transforming
- Creating
- As-a-Service
- Crowdsourcing

Data Science Team
- Data Scientist
- BI Analyst
- Project Sponsor
- Project Manager
- Business User
- Data Engineer
- DBA
Organizational Models for Data Science Teams

**Centralized**

The Data Science Team Functions As A Hub And Spoke Model; Central Provider Of Analytics To Multiple Business Units

**Decentralized**

Each Business Unit Has Its Own Data Science Capabilities

**Hybrid**

Centralized Data Science Team, But Business Units Also Have Data Science Capabilities

Regardless Of Which Approach, They All Need Executive Sponsorship To Succeed
Executive Engagement
Framework for Developing Data Science Teams

Executive Engagement
- Data-driven CEO
- Chief Data Officer

Organizational Model
- Centralized
- Decentralized
- Hybrid

Developing Data Science Capabilities
- Transforming
- Creating
- As-a-Service
- Crowdsourcing

Data Science Team
- Data Scientist
- BI Analyst
- Project Sponsor
- Project Manager
- Business User
- Data Engineer
- DBA
Analytics Requires Executive Level Engagement

“Executive Sponsorship Is So Vital To Analytical Competition...”
-- Tom Davenport (Competing on Analytics)

Chief Strategy Officer
Simulate Outcomes for Acquiring Our Top 3 Competitors

Chief Finance Officer
Use Time Series Analysis Over Historical Data to Predict KPIs to Project Earnings

Chief Product Officer
Conduct Social Media Analyses to Identify Customer Opinions

Chief Security Officer
Collect and Mine Log Data Within and Outside of the Company to Detect Unknown Threats

Chief Marketing Officer
Conduct Behavior Analyses to Predict If Customers Are Going to Churn

Chief Operating Officer
Mine Customer Opinions and Competitor Behaviors to Predict Inventory Demands
Executive Engagement for GINA Project

- VP in CTO office evangelized the project, provided support and resources
- Helped to raise awareness of project at executive levels of the company
EMC Courses on Data Science & Big Data Analytics

Business Leaders
- 90 min
- New
- Introducing Data Science and Big Data Analytics for Business Transformation

Heads of Data Science Teams
- 1 day
- New
- Data Science and Big Data Analytics for Business Transformation

Aspiring Data Scientists
- 5 days
- Data Science and Big Data Analytics
Closing Thoughts….

Now You Know How To Develop Data Science Teams...What Next?

- Determine How You Would Like To Develop Data Science Capabilities
- Hire People To Fill Out Your Data Science Team
- Consider Which Organizational Model Will Work Best For Your Situation
- Assess How Much Executive Engagement You Have Or Need
- Map Out Potential Projects -- Balance Quick Wins With Longer-term Wins
Questions?

Twitter: @imdavidddietrich

Additional Resources:

Blogs

• My Blog on Data Science & Big Data Analytics: http://infocus.emc.com/author/david_dietrich/

• Blog series on applying Data Analytics Lifecycle to measuring innovation data: http://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/

Courses

• 5-day course, Data Science & Big Data Analytics (for Data Scientists): http://bit.ly/15pxLLO

• 1-day course, Data Science & Big Data Analytics for Business Transformation (for leaders): http://bit.ly/uINB5U
Executive Engagement: Data-Driven CEO

“... If Your Organization Can Arrange It ... Have Someone In A Key Operational Role -- Business Unit Head, Chief Operations Officer, Even CEO -- To Be An Enthusiastic Advocate Of Matters Quantitative.”

-- Tom Davenport (HBR Blog Network)

Key Focus Areas of a Data-driven CEO:

• Strategic Data Planning
• Analytic Understanding
• Technology Awareness

Procter & Gamble Business Sphere
Executive Engagement: Chief Data Officer (CDO)

“... It's Time For Corporations To Embrace A New Functional Member Of The C-suite: The Chief Data Officer (CDO).”

-- Anthony Goldbloom and Merav Bloch, Kaggle

- Promote Data-driven Decision Making To Support Company’s Key Initiatives
- Ensure The Company Collects the Right Data
- Oversee and Drive Analytics Company-wide

25% of organizations will have a Chief Data Officer by 2015.

-- Gartner Blog Network

Executive-level Advisor On Data Analytics

Executive Boardroom
Approaches to Developing Data Science Capabilities: Transforming Teams

- **Industries Requiring Deep Domain Knowledge** (Such As Genetics And DNA Sequencing)
- Established Companies Who Wish To Introduce Data Science Into Their Business
- Companies Who Wish To Enrich The In-house Skill Sets

Transforming And Realignment With Minimal Change To The Current Organizational Structure
Approaches to Developing Data Science Capabilities: Creating Teams

- **Start-up Companies**
- Companies Who Wish To...
  - Increase Their Focus On Data Analytics
  - Start New Data Science Projects
- **Companies Where Data Is The Product**
- Deep Domain Knowledge Is Less Critical For The Analytics
Approaches to Developing Data Science Capabilities: Data Science as a Service

**Engaging Data Science as a Service (DSaaS)**

- **When To Engage DSaaS Providers**
  - Prefer Not To Change Existing Organizational Structure
  - When Creating Or Transforming Are Not Viable Options

- Consider Service-level Agreements (SLAs) When Determining Whether To Engage Internal Resources Or External Providers
Approaches to Developing Data Science Capabilities: Crowdsourcing Data Science

When To Crowdsource

- **The Problem Is “Open” In Nature**
- Willing To Accept Opinions From Distributed And Diverse Groups Of People
- There’s A Back-up Plan In Case Of “Crowd Failures”

Examples: Wikipedia, Netflix’s $1,000,000 Prize
Approaches to Developing Data Science Capabilities: Crowdsourcing Data Science (Cont’d)

- **Different Crowdsourcing Models**
  - Wisdom Of Crowds
  - Swarm Creativity (Collective Intelligence)

- **Crowdsourcing Platforms**
  - Kaggle.com, Innocentive.com
  - Amazon Mechanical Turk

- **Crowd Failures**: When The Turnout Of Crowdsourcing Is Unsatisfactory