EMC WORLD 2013
LEAD YOUR TRANSFORMATION

Building Data Science Teams

David Dietrich
Advisory Technical Education Consultant
EMC Education Services

@imdaviddietrich

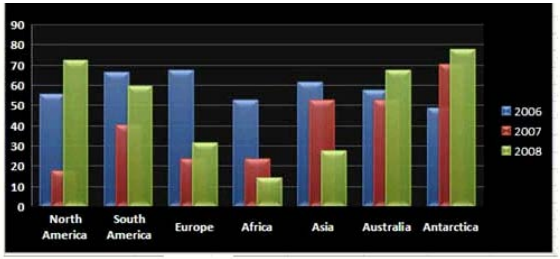
EMC²

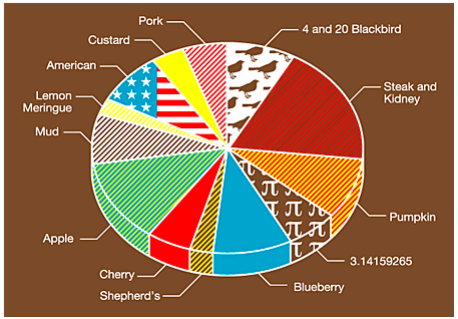
Roadmap Information Disclaimer

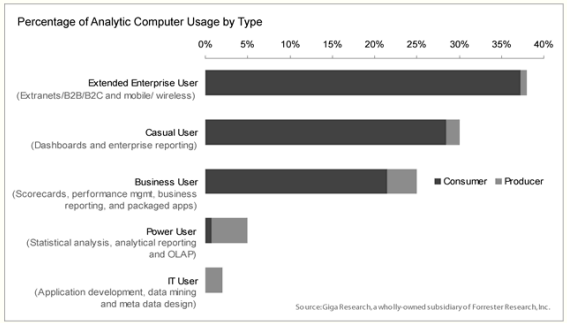
- EMC makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, “Roadmap Information”).
- Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.
- Roadmap information is EMC Restricted Confidential and is provided under the terms, conditions and restrictions defined in the EMC Non-Disclosure Agreement in place with your organization.

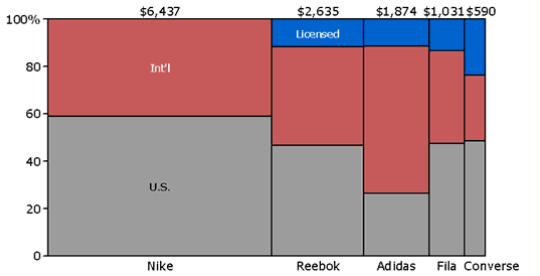
Do You Need A Data Science Team For This?

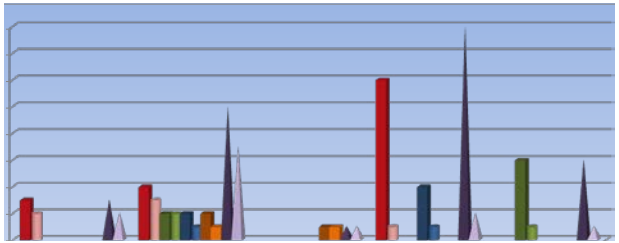
Creating Reports, Dashboards, and Databases…





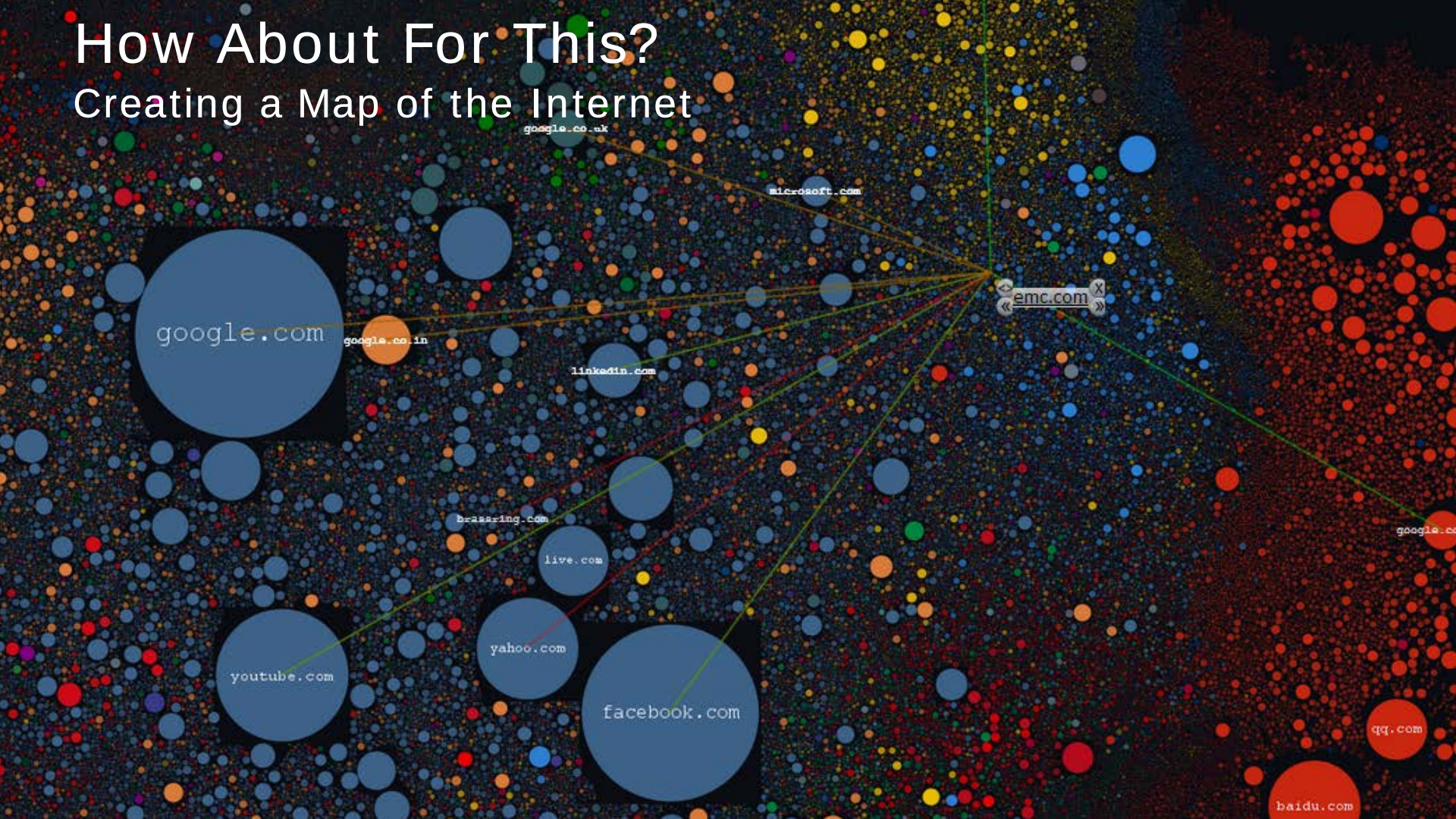






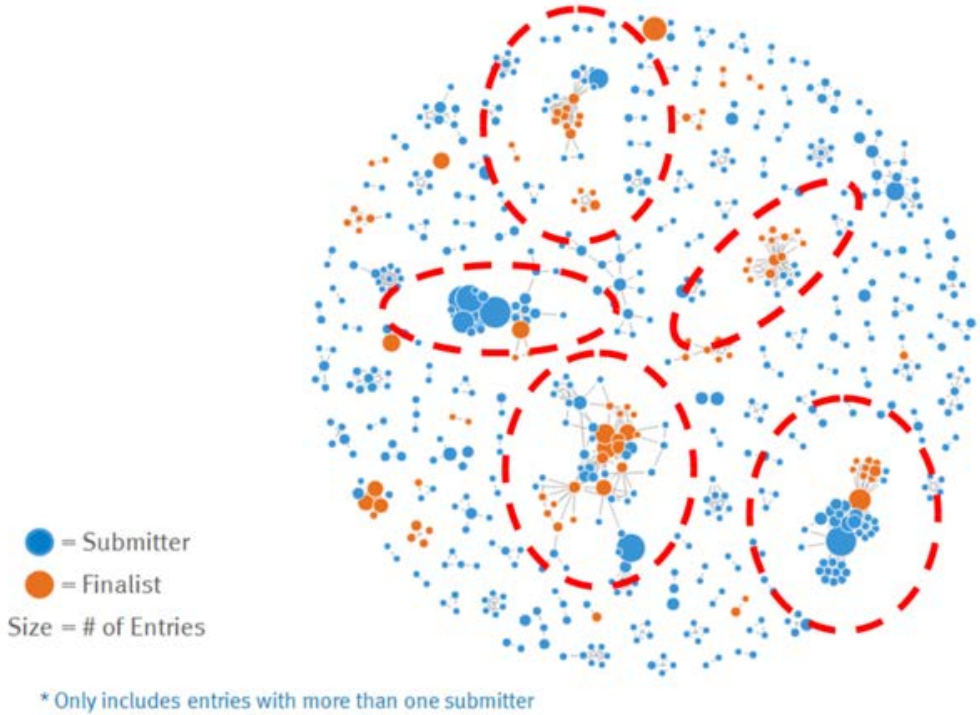
How About For This?

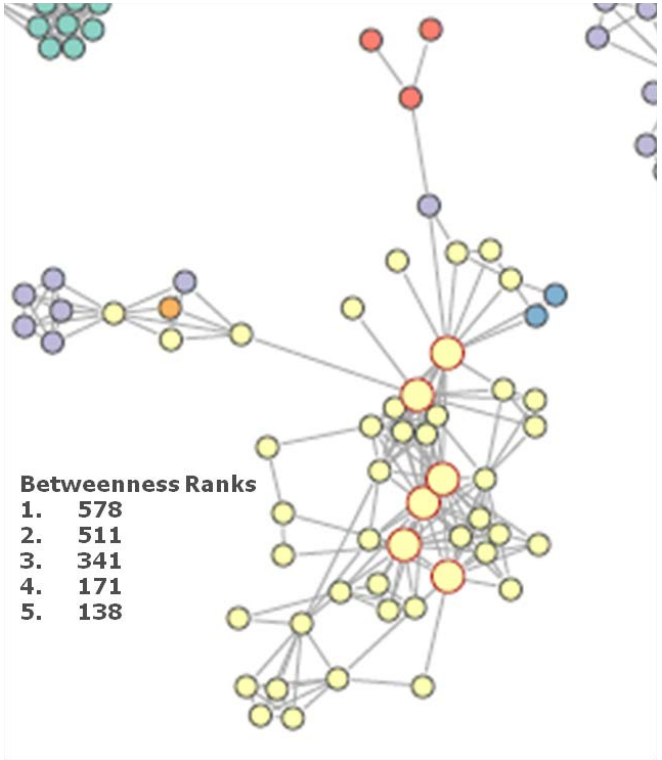
Creating a Map of the Internet



Example: Output From a Data Science Team

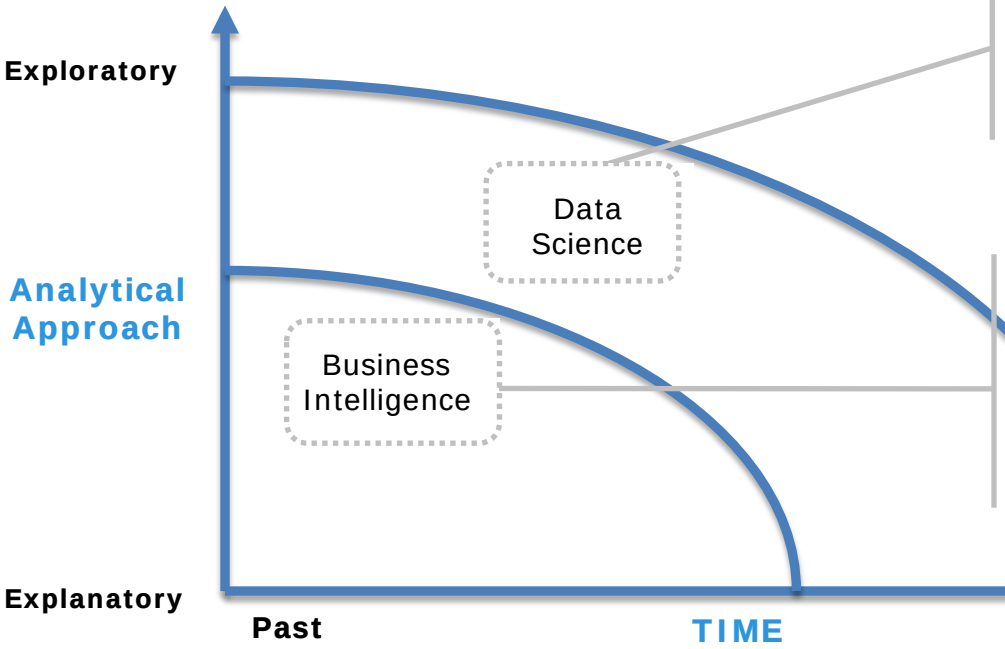
Mapping The Spread of Innovation Ideas Using Social Graphs





Big Data Requires New Approaches to Analytics

Business Intelligence Versus Data Science



Predictive Analytics and Data Mining (Data Science)	
Typical Techniques and Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large data sets
Common Questions	• What if…..? • What's the optimal scenario for our business? • What will happen next? What if these trends continue? Why is this happening?

Business Intelligence	
Typical Techniques and Data Types	• Standard and ad hoc reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable data sets
Common Questions	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations?

“Companies Are Always Looking To Reinvent Themselves….But It’s A Mistake To Treat Data Science Teams Like Any Old Product Group.

To Build Teams That Create Great Data Products, You Have To Find People With The Skills And The Curiosity To Ask The Big Questions.”

- DJ Patil, Data Scientist in Residence at Greylock Partners

Framework for Developing Data Science Teams

Data Science Team

Data Scientist

BI Analyst

Project Sponsor

Project Manager

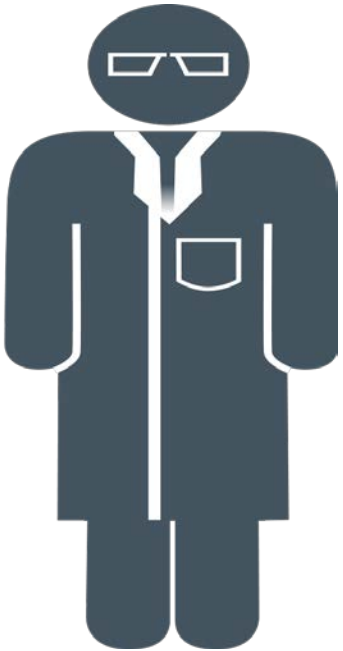
Business User

Data Engineer

DBA

Data Science Teams

Data Scientist: An Emerging Career



Harvard Business Review

SPOTLIGHT ON BIG DATA

Data Scientist:
The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

EMC²

Comparing Two Data Analysts

Traditional BI Analyst

ACME Healthcare

John

Sample Tasks

- Report Regional Sales For Last Quarter
- Perform Customer Feedback Surveys
- Identify Average Cost Per Supplier

Data Scientist

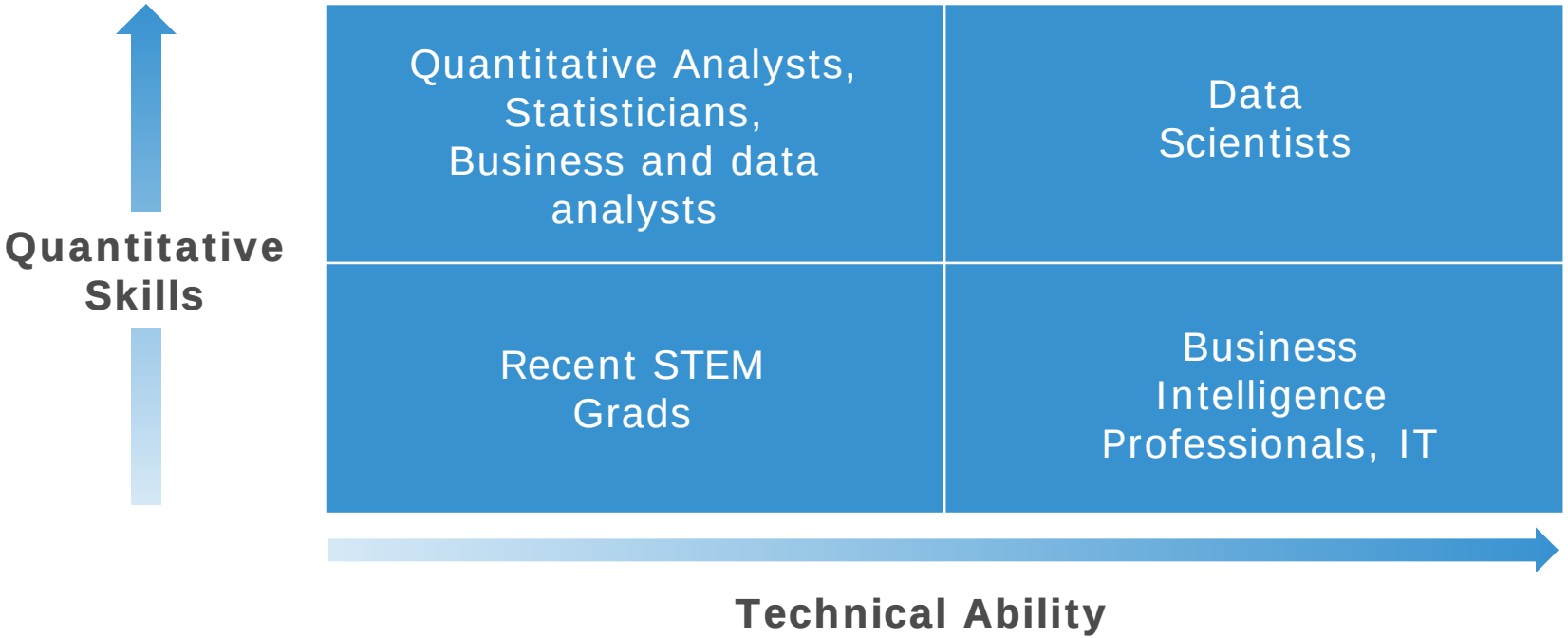
ACME Healthcare

Janet

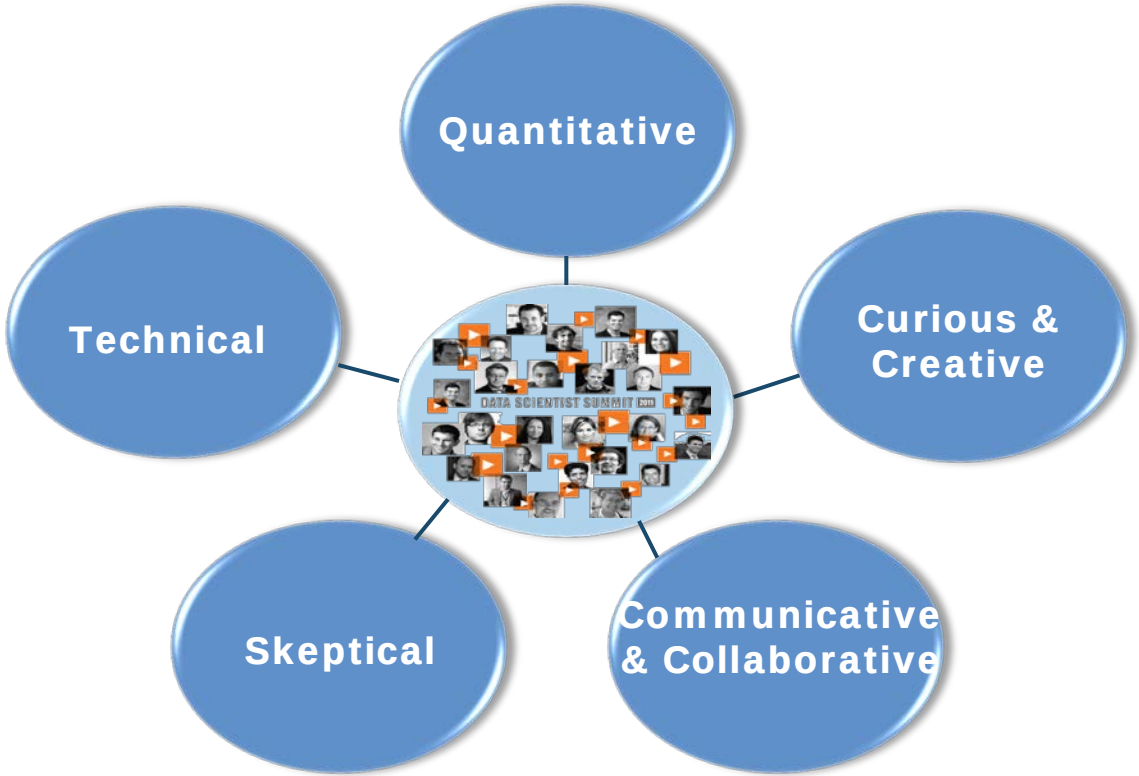
Sample Tasks

- Predict Regional Sales For Next Quarter
- Discover Customer Opinions Via Social Media
- Identify Ways to Maximize Sales Campaign ROI

Skills Matrix, Based on Recent Students



Profile of a Data Scientist



Interpreting the Resume of a Senior Data Scientist

Data Scientist Job Description

Responsibilities:
- Work with business owners to map business requirements into technical solutions
- Analyze and extract relevant information from large ar… key revenue-driven features
- Perform ad-hoc statistical and data mining analyses
- Design and implement scalable and repeatable solutions, and establish scalable, efficient, aut… …ta analys…
- Work clos… …am to dr…
- Design mu… …hypothes…

Qualification…
- A proven passion for generating insights from data, wit… higher-level trends in data growth, open-source platforms, and public data sets
- Experience with statistical languages and packages, inc… and/or Mahout
- Experience working with relational databases and/or d… and their query interfaces, such as SQL, MapReduce, Had…
- Strong communication skills, with ability to communicate at all levels of the organization
- Masters/PhD degree in mathematics, statistics, computer science or a similar quantitative field
- Experience in designing and implementing scalable data mining solutions
- Preferably experience with additional programming languages, including Python, Java, and C/C++
- Ability to travel as-needed to meet with customers

Statistics

Programming

Data Mining

Advanced STEM Degrees

Sample Data Scientist Resume

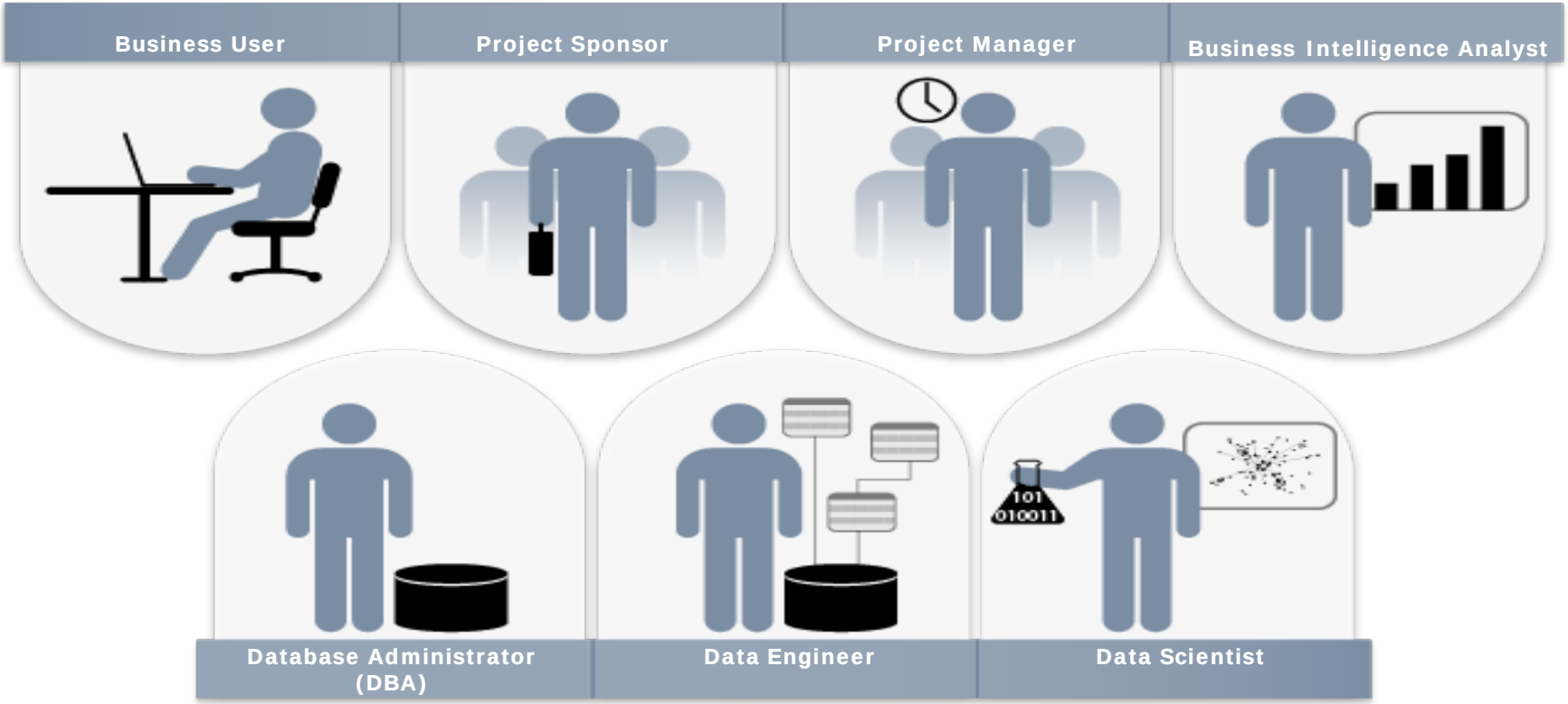
John Smith
john.smith@email.com

Skills
R, SAS, Java, data mining, statistics, ontology, bioinformatics, human-computer interaction, research

Experience
2009—Present, Senior Data Scientist, *ABC Analytics*
2007—2009, Founder&CEO, *Genome*
Genome specializes in consumer health information. The main product is InherithHealth, a tool for acquisition of family medical histories that provides familial disease risk assessment.
2005—2007, Knowledge Engineer, *ScienceExperts.com*
Managed technical outsourcing efforts. Developed criterion and evaluated engineering outsourcing agencies and individuals …
2004—2006, Research Scientist, *University of Washington*
Developed rigorous statistical and computational models for addressing primary shortcomings of observational data analysis in the context of disease risk and drug response.
2000—2004, Research Developer, *Nat'l Inst. of Standards and Technology*
Designed and implemented prototypes. Evaluated tools for representing rules of autonomous on-road navigation.

Education
Ph.D, Biomedical Informatics, *University of Washington*, 2011
Dissertation: Detection of Protein–protein Interaction in Living Cells by Flow Cytometry

Successful Analytic Projects Require Breadth of Roles



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

Developing Data Science Capabilities

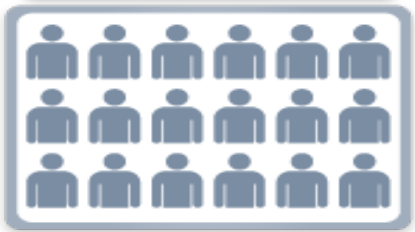
Four Approaches to Developing Data Science Capabilities

Transforming

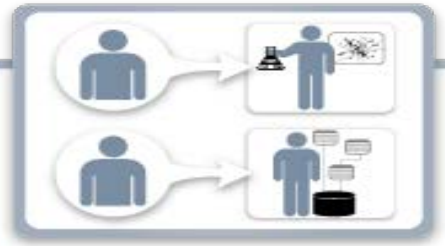
Creating

As a Service

Crowdsourcing



Approaches to Developing Data Science Capabilities:
Transforming Teams



Transforming And Realignment With Minimal Change To The Current Organizational Structure

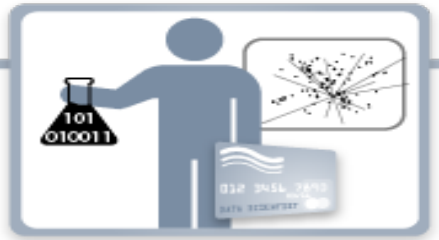
- **Industries Requiring Deep Domain Knowledge** (Such As Genetics And DNA Sequencing)
- Established Companies Who Wish To Introduce Data Science Into Their Business
- Companies Who Wish To Enrich The In-house Skill Sets

Approaches to Developing Data Science Capabilities: Transforming Teams

Developing A New Team From Scratch

- **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

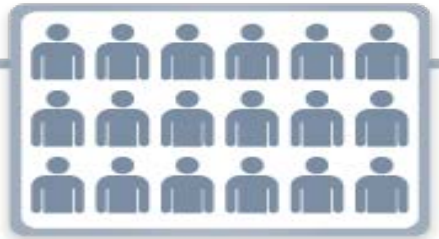
Approaches to Developing Data Science Capabilities:
Data Science as a Service



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

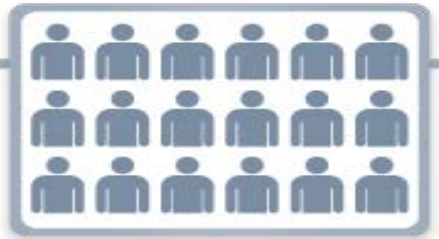


Outsource Data Science Project To Distributed Groups Of People

- 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)



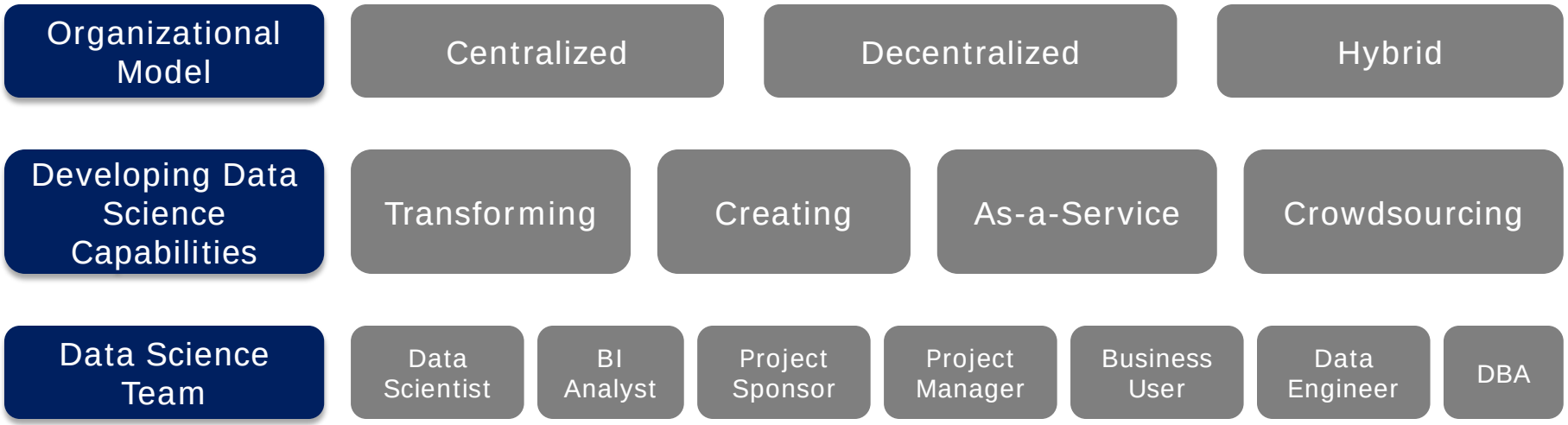
Outsource Data Science Projects To Distributed Groups Of People

- **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

Benefits and Drawbacks of the Four Approaches

Transforming	Creating	DSaaS	Crowdsourcing
• **Strong Domain Knowledge** • Knowledge of Business Processes • New Talent Raises Level of Team Performance • Gradually Increases the Quality of Service	• **Control Over Skill-sets** • More Flexibility • High Quality of Service	• **Able to Scale on Demand** • May Get Better Service Levels Than In-house • Learn From Outside Experts	• **Leverage Wisdom of the Crowds** • Diverse Perspectives • Lower Cost • Fast Results
• **Risk of homogeneous thinking** • May Struggle With Quality of Service • Some Team Members May Resist Change	• **Hiring and Knowledge Transfer Are Time-consuming** • Time Required to Find and Hire Right Team Members	• **Provider May Not Understand Company's Unique Processes** • Difficult to Bring Expertise Back In-house • Decreasing Quality of Service Over Time	• **No SLA; value not guaranteed** • Difficult to design the "Open" Problem • Difficult For Domain Intensive Tasks • Crowd Failure May Happen (Adds Cost)

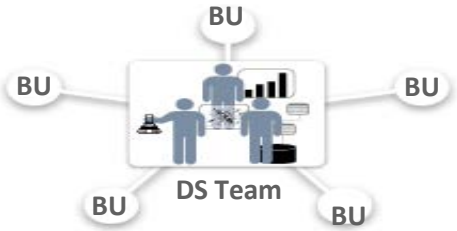
Framework for Developing Data Science Teams



Organizational Model

Organizational Models for Data Science Teams

Centralized



The Data Science Team Functions As A Hub And Spoke Model, In Which They Are A Central Provider Of Analytics To Multiple Business Units

Decentralized

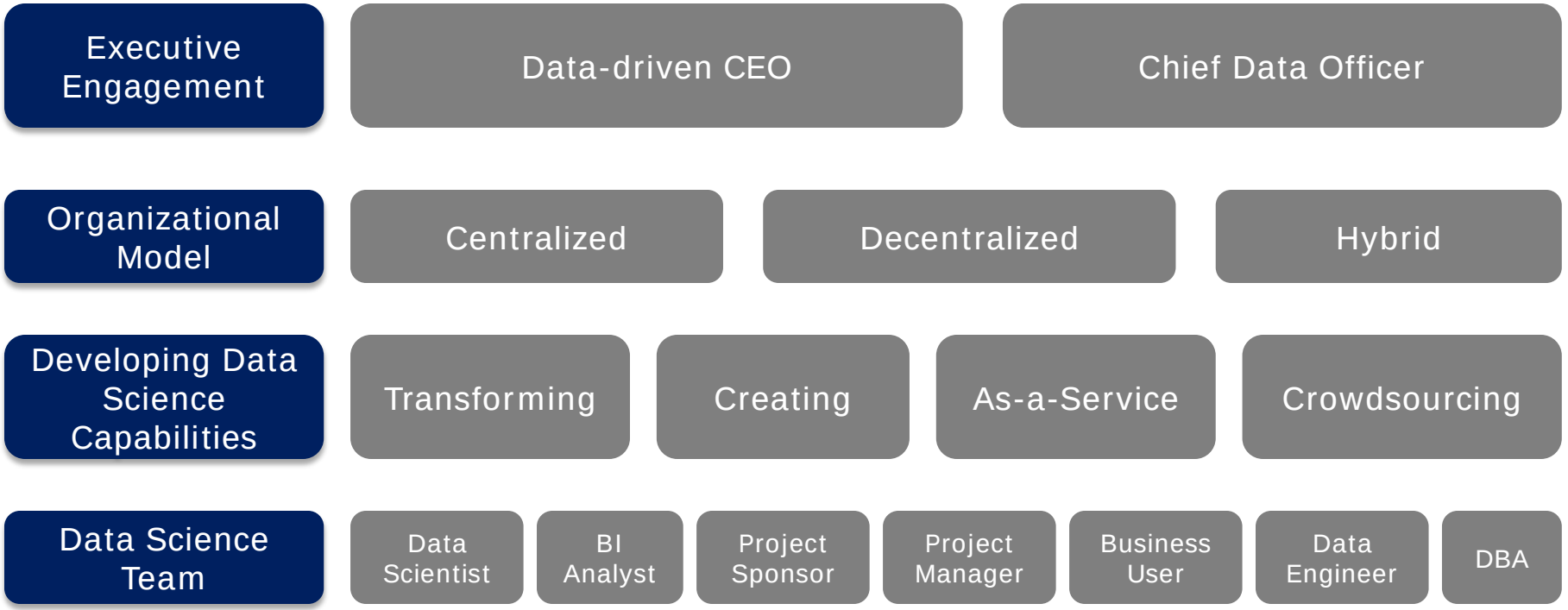
Each Business Unit Has Its Own Data Science Capabilities

Hybrid

There Is A Centralized Data Science Team, But Business Units Also Have Data Science Capabilities

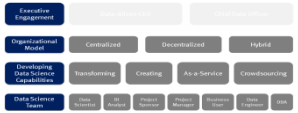
Regardless Of Which Approach, They All Need Executive Sponsorship To Succeed

Framework for Developing Data Science Teams



Executive Engagement

Analytics Requires Executive Level Engagement



"Executive Sponsorship Is So Vital To Analytical Competition…"

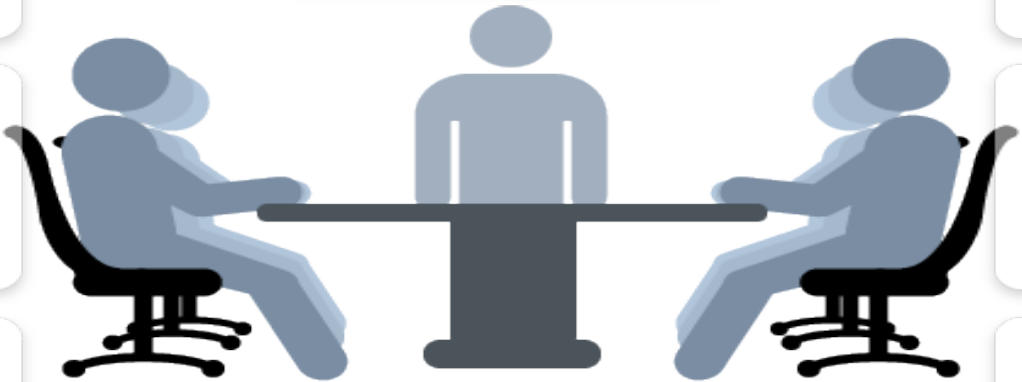
-- Tom Davenport (Competing on Analytics)

Chief Strategy Officer
Simulate Outcomes for Acquiring Our Top 3 Competitors

Chief Product Officer
Conduct Social Media Analyses to Identify Customer Opinions

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

CEO



Executive Boardroom

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

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

Chief Operating Officer
Mine Customer Opinions and Competitor Behaviors to Predict Inventory Demands

EMC²

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

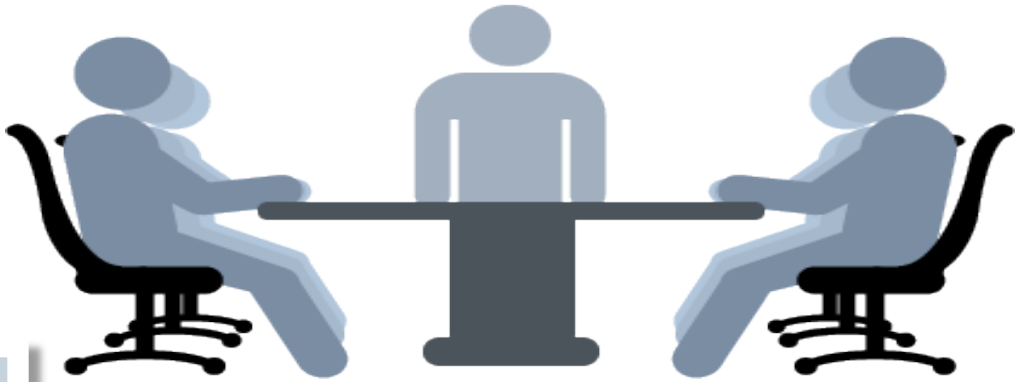
Procter & Gamble Business Sphere

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

Executive-level Advisor On Data Analytics



Executive Boardroom

> ***25% of organizations will have a Chief Data Officer by 2015.***
>
> *-- Gartner Blog Network*

Two New EMC Data Science Courses for Business Transformation

Business Leaders — 90 min

New

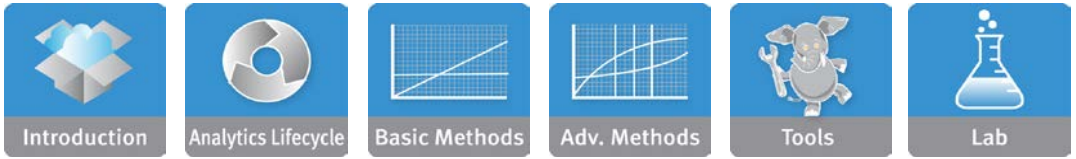
Introducing Data Science and Big Data Analytics for Business Transformation

Heads of Data Science Teams — 1 day

New

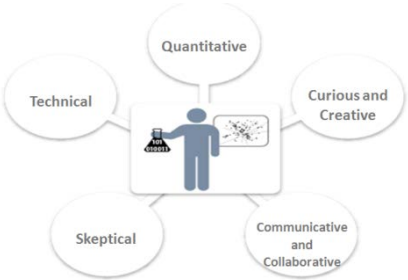
Data Science and Big Data Analytics for Business Transformation

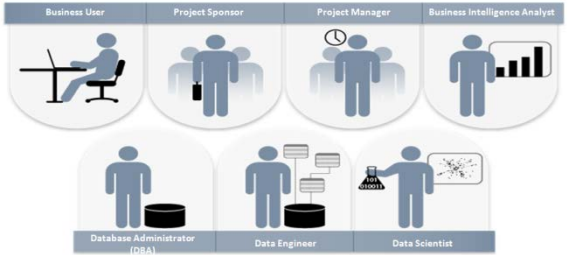
Aspiring Data Scientists — 5 days



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?

Visit the EMC Global Services Booth, #110

Additional Resources:

1. **EMC Education Services curriculum on Data Science and Big Data Analytics for Business Transformation**:
   http://education.emc.com/guest/campaign/data_science.aspx

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

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


EMC²