

# Data Science & Multi-Stakeholder Partnerships

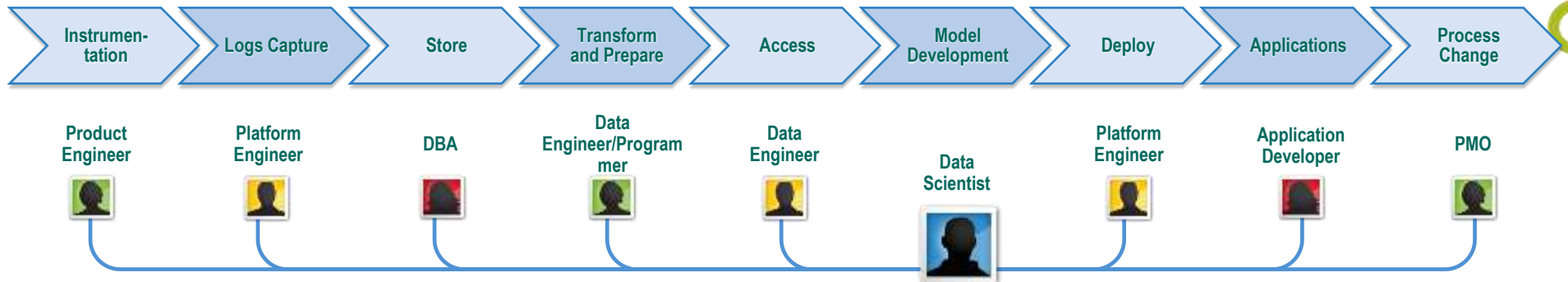
Annika Jimenez  
Global Head of Data Science Services

Big Data Senior Steering Group | April 11, 2013

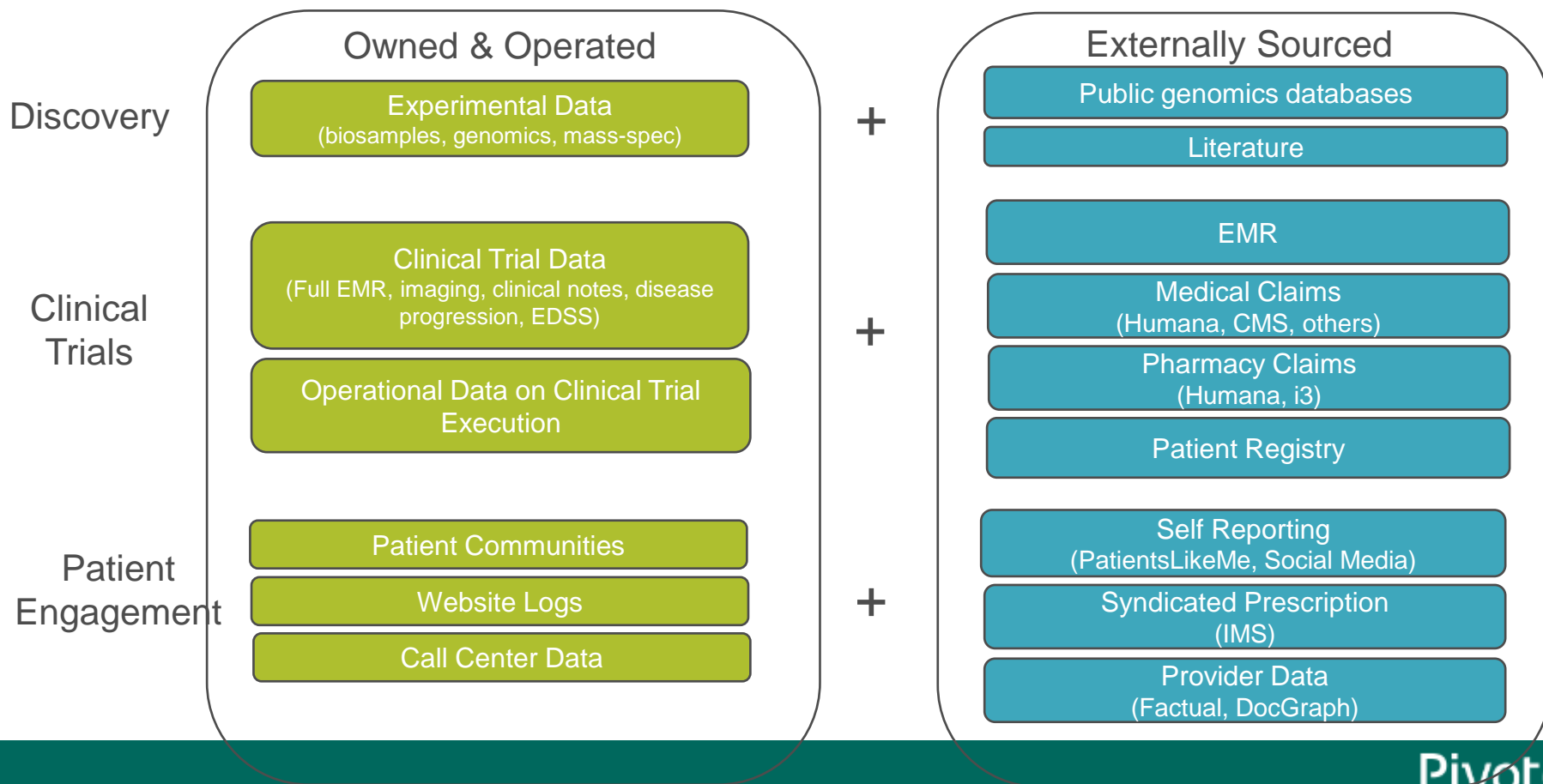
# DATA SCIENCE



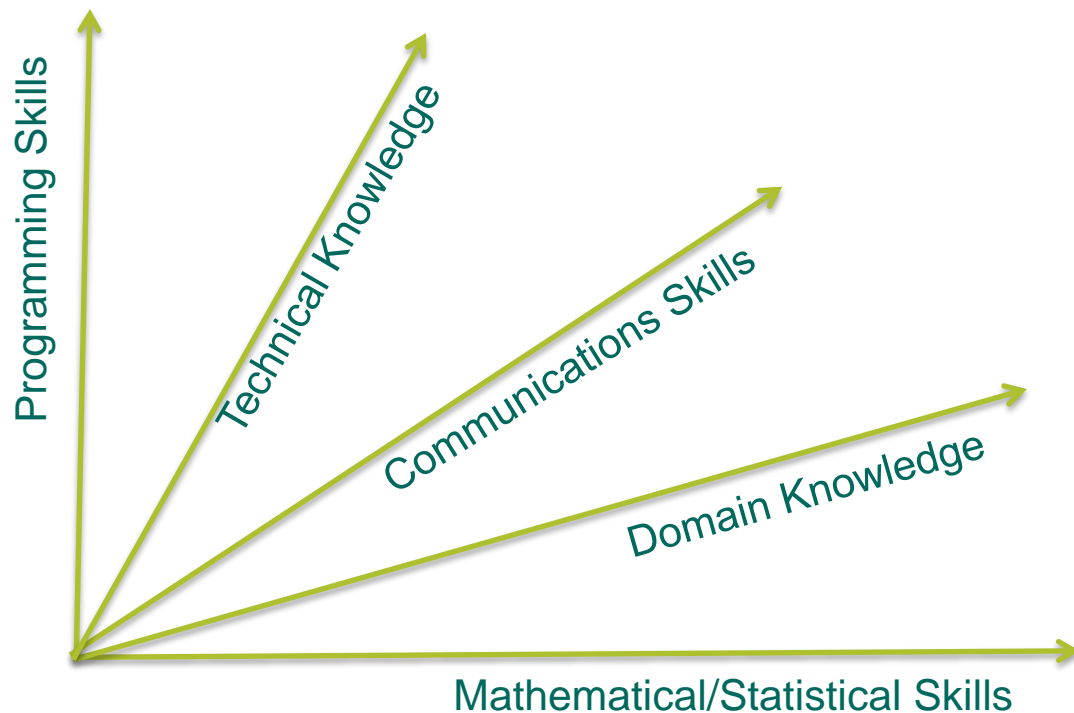
# Data Science Value Chain



# Sample Scenario – Data Sources



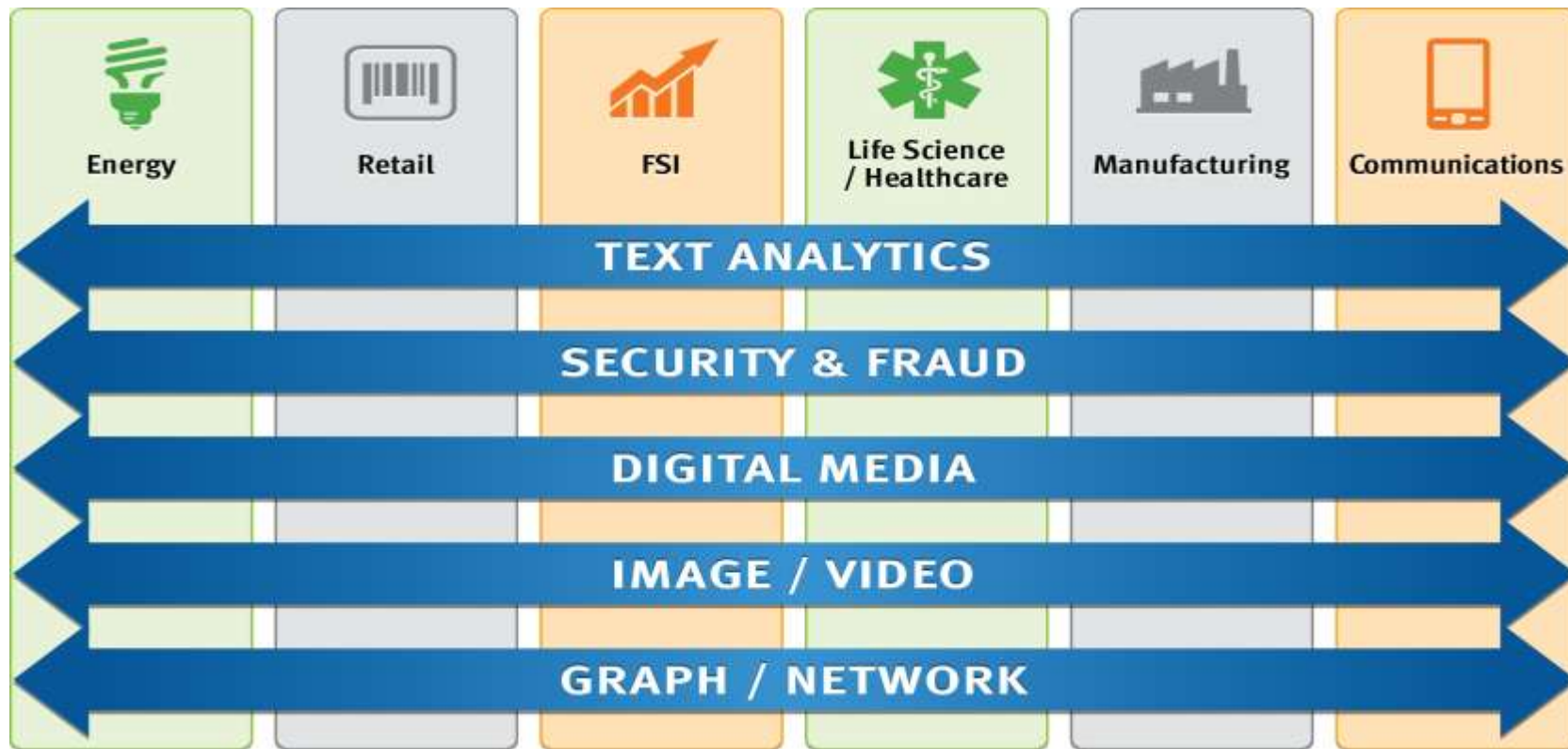
# What Is A “Data Scientist”?



# Pivotal Data Science Dream Team

- **Derek Lin** – Network Security, Fraud Detection, Speech and Language Processing, (Principal Scientist at RSA, M.S. in Signal Processing, USC)
- **Hulya Farinas** – Optimization, Resource Allocation in Healthcare (Modeler at M-Factor, IBM, Ph.D. in Operations Research, University of Florida)
- **Kaushik Das** – Mathematical Modeling in Energy, Retail and Telco (Director of Analytics at M-Factor, M.S. in Mineral Engineering, UC Berkeley)
- **Rahel Jhirad** – Quantitative Modeling and Risk Management in Trading and Finance (Global Risk Management at Saloman Brothers, Morgan Stanley, Ph.D. in Economics, Princeton, M.S. in Mathematics, Courant Institute)
- **Sarah Aerni** – Genomics and Machine Learning (Ph.D. in Biomedical Informatics, Stanford)
- **Mariann Micsinai** – Next Generation Sequencing (Market Risk Management Associate at Lehman Brothers, Ph.D. in Computational Biology, NYU and Yale)
- **Emily Kawaler** – Clinical Informatics and Machine Learning (M.S. in Computer Sciences, University of Wisconsin-Madison)
- **Joseph Zadeh** – IT/Network Traffic and Financial Modeling (Ph.D. in Mathematics, Purdue)
- **Victor Fang** – Imaging and Graph Analytics, Machine Learning (Sr. Scientist at Riverain Medical, Ph.D. in Computer Sciences, University of Cincinnati)
- **Anirudh Kondaveeti** – Trajectory Data Mining and Machine Learning (Ph.D. in Computing & Dec. Systems Eng, Arizona State University)
- **Hong Ooi** – Insurance and Finance Risk Modeling (Statistician at ANZ, Ph.D. in Statistics, Australian National University)
- **Michael Brand** – Text, Speech and Video Research for Retail, Finance and Gaming (Chief Scientist at Verint Systems, M.S. in Applied Mathematics, Weizmann Institute)
- **Kee Siong Ng** – Data Mining in Healthcare (Sr. Data Miner at Medicare Australia, Ph.D. in Computer Science, and Postdoctoral Fellow, Australian National University)
- **Noah Zimmerman** – Statistics and Immunology (Ph.D. in Biomedical Informatics, Stanford)
- **Noelle Sio** – Digital Media Analytics and Mathematical Modeling (Sr. Analyst at eHarmony, Fox Interactive Media (Myspace), M.S. in Applied Mathematics, Cal Poly Pomona)
- **Jin Yu** – Stochastic Optimization, Robust Statistics in Machine Learning, Computer Vision (Research Associate at U of Adelaide, Ph.D. in Machine Learning, Australian National University)
- **Rashmi Raghu** – Computational Methods and Analysis (Ph.D. in Mechanical Engineering, Stanford)
- **Woo Jung** – Bayesian Inference and Demand Analysis (Sr. Statistician at M-Factor, M.S. in Statistics, Stanford)
- **Jarrod Vawdrey** – Marketing Analytics & SAS (Analytics Consultant at Aspen Marketing, B.S. in Mathematics, Kennesaw State University)
- **Niels Kasch** – Text Analytics and NLP (Ph.D. in Computer Science, UMBC)
- **Vivek Ramamurthy** – Online Learning, Stochastic Modeling, Convex Optimization (Ph.D. in Operations Research, UC Berkeley)
- **Srivatsan Ramanujam** – NLP and Text Mining (Natural Language Scientist at Sony, Salesforce.com, M.S. in Computer Sciences, UT Austin)
- **Alexander Kagoshima** – Time Series, Statistics and Machine Learning (M.S. in Economics/Computer Science, TU Berlin)

# Pivotal Data Science Knowledge Development



# Data Science Curricula

- Online:
  - “Introduction to Data Science”  
University of Washington  
<https://www.coursera.org/course/datasci>
  - “Machine Learning”  
Stanford University  
<https://www.coursera.org/course/ml>
  - “Introduction to Databases”  
Stanford University  
<https://www.coursera.org/course/db>
  - “Introduction to Artificial Intelligence”  
Stanford University  
<http://www.udacity.com/overview/Course/cs271/CourseRev/1>
- Syracuse: Certificate of Advanced Study in Data Science  
<http://ischool.syr.edu/future/cas/datascience.aspx>
- Northwestern: M.S. Analytics  
<http://www.analytics.northwestern.edu/>
- North Carolina State University, Institute for Advanced Analytics: M.S. Analytics  
<http://analytics.ncsu.edu/>
- Columbia Institute of Data Science

Lots More!

<http://datascience101.wordpress.com/2012/04/09/colleges-with-data-science-degrees/>



# Data Science and Big Data Analytics Course and Certification

## Course Overview

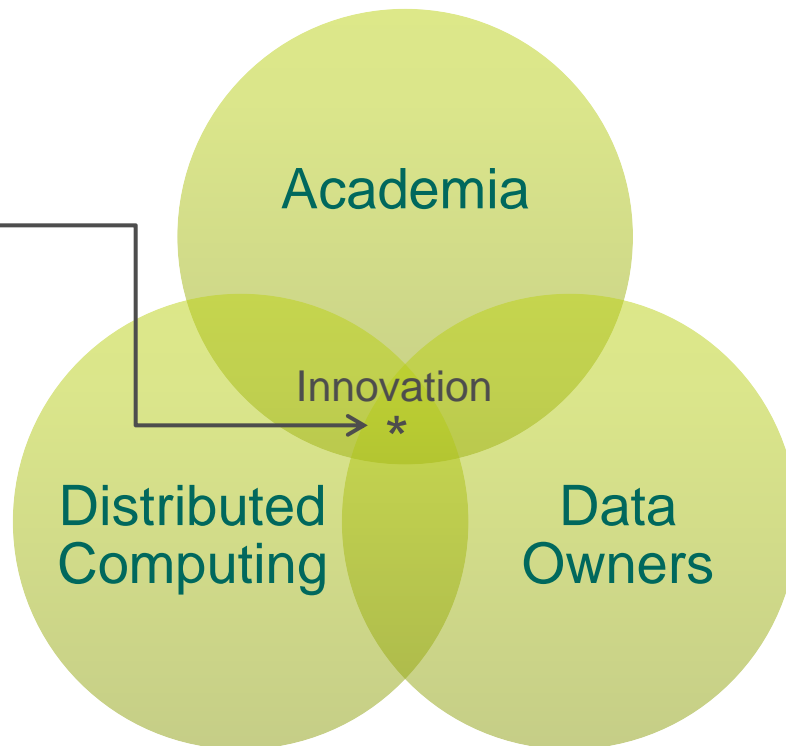


## Details

- “Open” curriculum
- Practitioner’s approach
- Enables immediate participation on analytics projects
- Prepares for EMC Proven Professional Data Science Associate Certification

# Hard Problem Solving

- Genomics
- Video
- Energy



# RNA Sequencing

## Customer

Translational Pathology Department of a University

## Business Problem

Reducing time spent on processing and analyzing RNA sequences

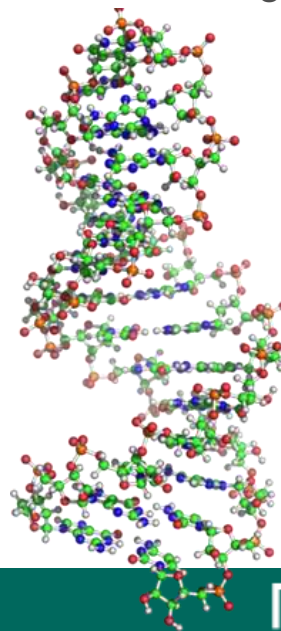
## Challenges

The center identified three bottleneck steps in their RNA sequencing pipeline: Alignment, splice junction detection, and gene level expression

## Solution

Algorithms implemented in UAP are three times faster than the client's existing platform.

In addition to improving the algorithm run time, we identified modifications to the method capable of improving prediction on short exons.



# Pivotal Industry Collaboration

- Analytics Workbench
  - 1000-node Hadoop cluster
  - Publicly available data sets
  - Largest test-bed for Hadoop, Mahout, etc.
  - Universities and non-academia
  - **Mission:** Provide a collaborative, community-focused, open and innovative platform for rapid discovery and demonstration of solutions to the world's biggest data challenges.

# Pivotal

A NEW PLATFORM FOR A NEW ERA