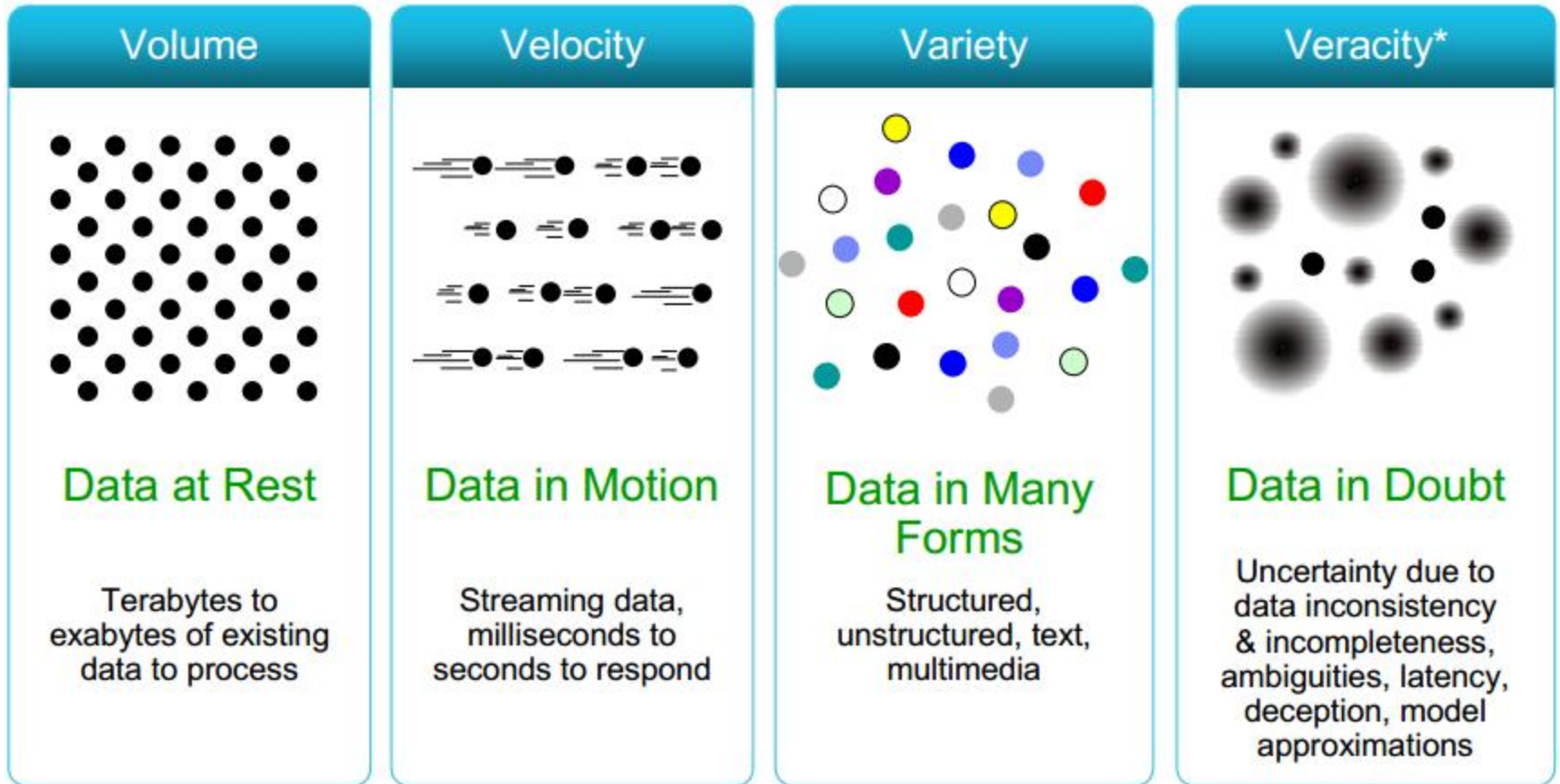# *Demystifying Data Science*

Natasha Balac, Ph.D.
Predictive Analytics Center of Excellence, Director
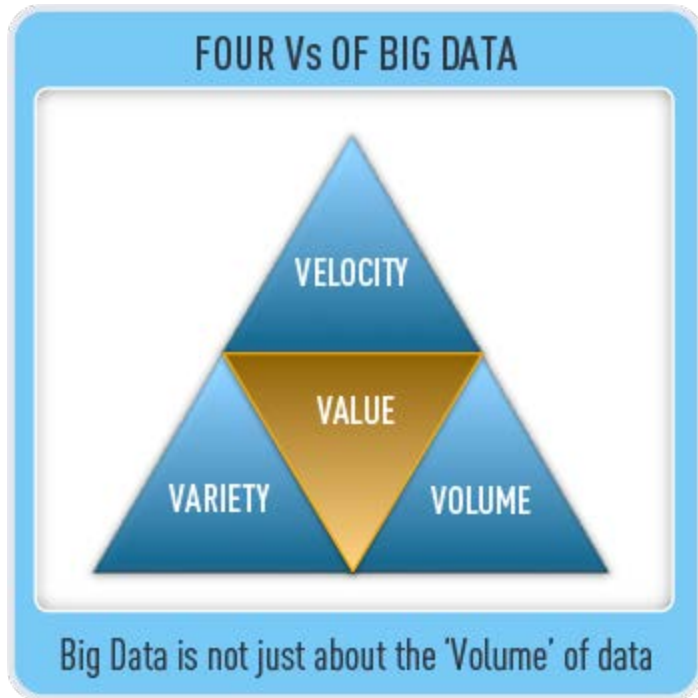## San Diego Supercomputer Center
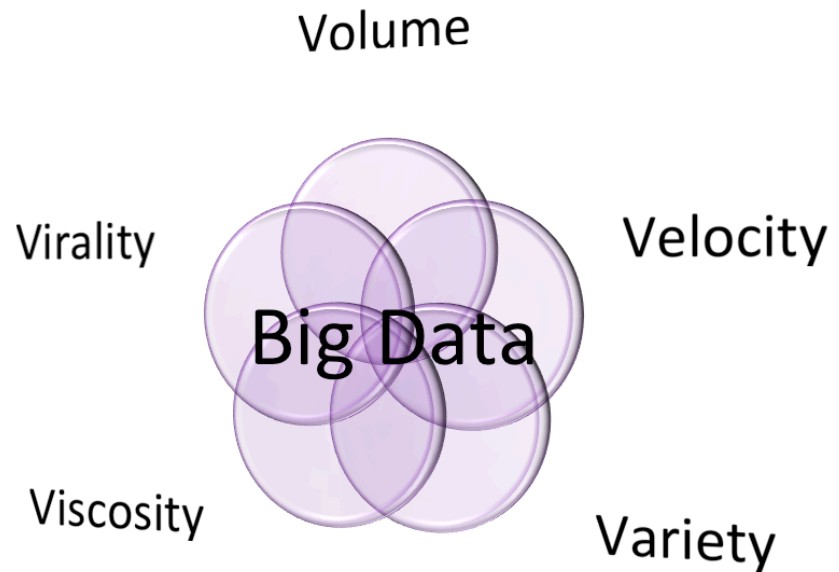## University of California, San Diego

4 V's of Big Data

Volume	Velocity	Variety	Veracity*
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

IBM, 2012

Growing Vs of Big Data

FOUR Vs OF BIG DATA

VELOCITY

VALUE

VARIETY · VOLUME

Big Data is not just about the 'Volume' of data

Zymr, Inc.

Volume

Virality

Big Data

Velocity

Viscosity

Variety

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

SDSC SAN DIEGO SUPERCOMPUTER CENTER

at the UNIVERSITY OF CALIFORNIA; SAN DIEGO
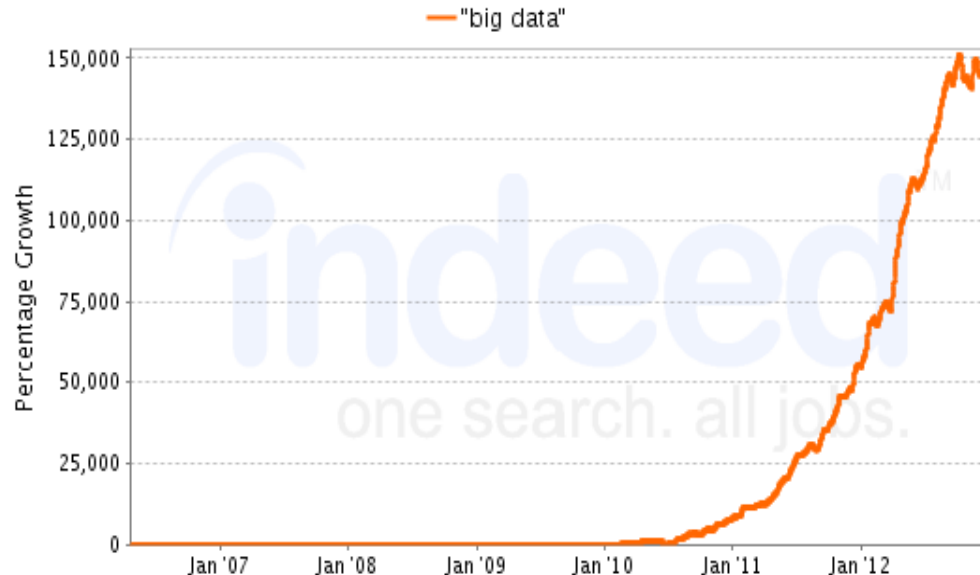
UCSD

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:
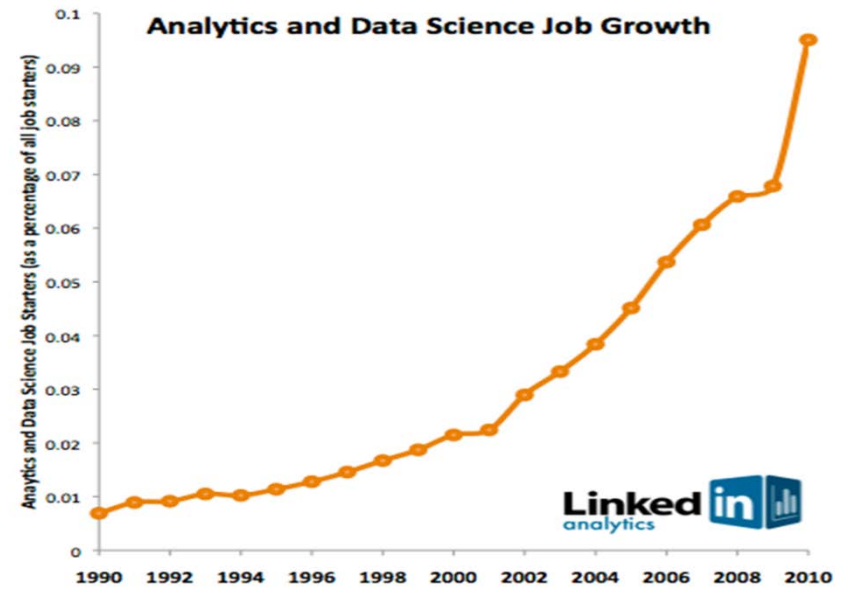# *The hot new gig in tech*

- **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
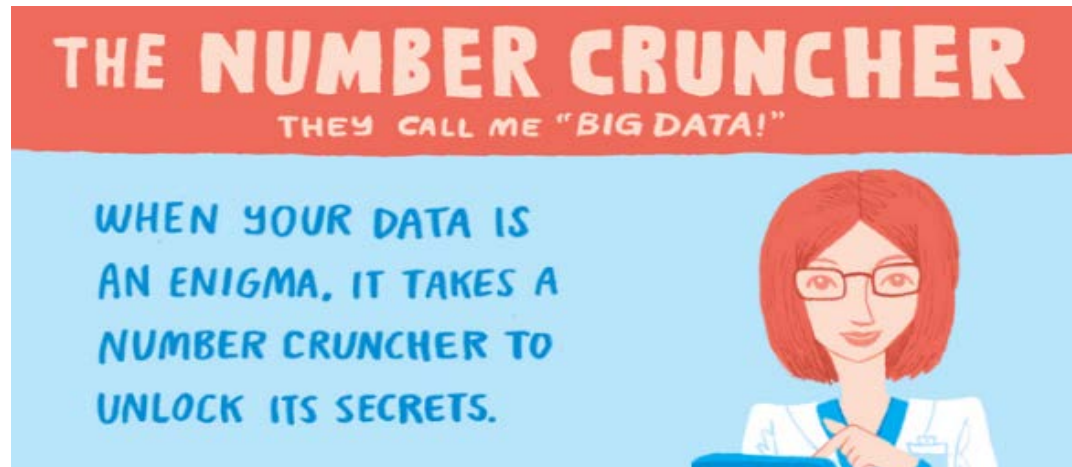
Data Science Job Growth

By 2018 shortage of 140-190,000 predictive analysts and 1.5M managers / analysts in the US

SDSC SAN DIEGO SUPERCOMPUTER CENTER

at the UNIVERSITY OF CALIFORNIA; SAN DIEGO

UCSD

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 NUMBER CRUNCHER
THEY CALL ME "BIG DATA!"

WHEN YOUR DATA IS AN ENIGMA, IT TAKES A NUMBER CRUNCHER TO UNLOCK ITS SECRETS.

- **The new "Cool Kids on Campus"**

Rising Part of the Culture and Enterprises

- ## **Movies and bestsellers**
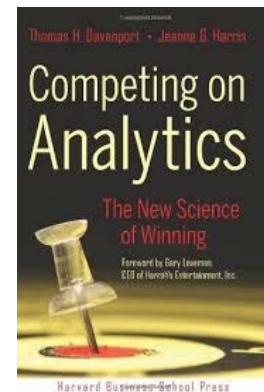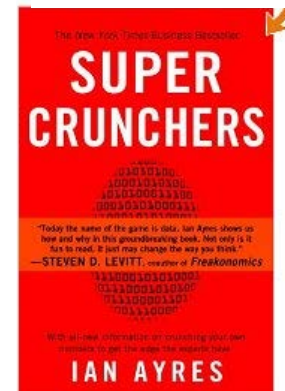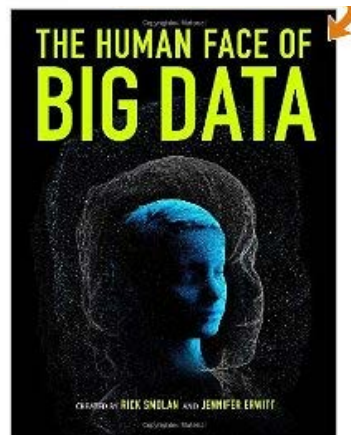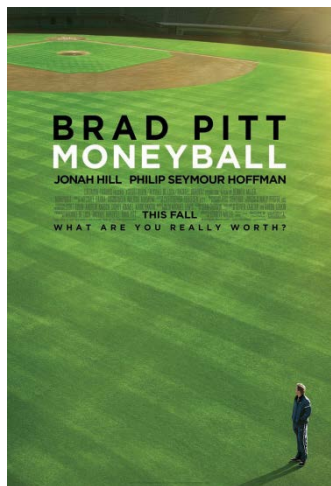
- ## **Enterprises competing**
  - Chief Data Officer now Chief Data Scientist
  - Data Driven Decisions strategic, tactical, operational

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

By Brendan Tierney, 2012

Business Strategy

Domain Knowledge

Statistics

Pattern Recognition

Neurocomputing

Visualisations

Business Analyse

Communications

Machine Learning

AI

Data Science

Databases & Data Processing

KDD

Problem Solving

Data Mining

Presentation

Inquisitiveness
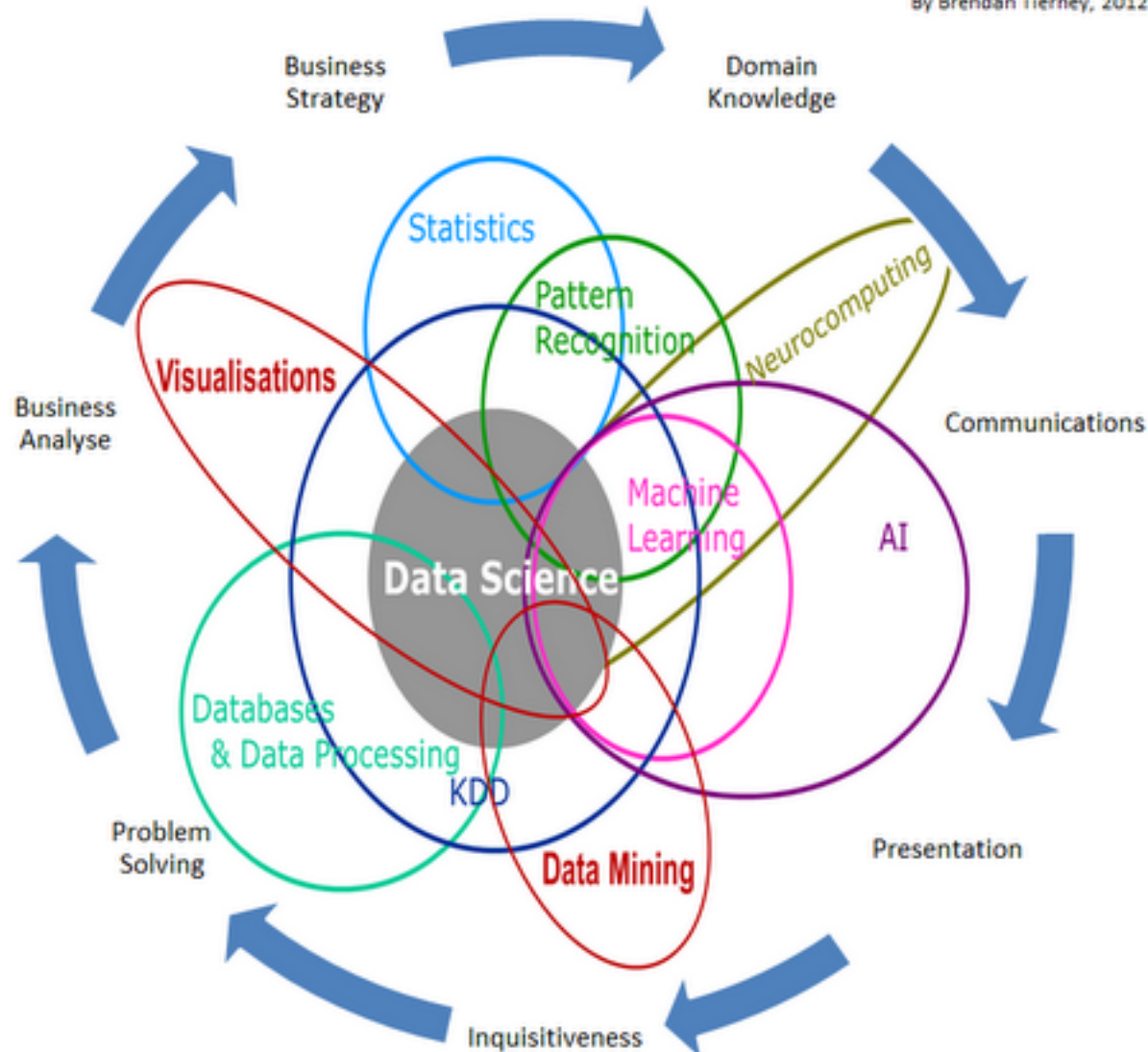
Successful Data Scientist Characteristics

- **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.
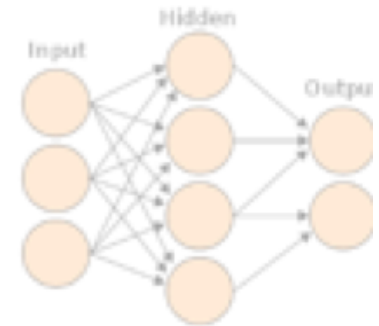
Data Scientist Qualities
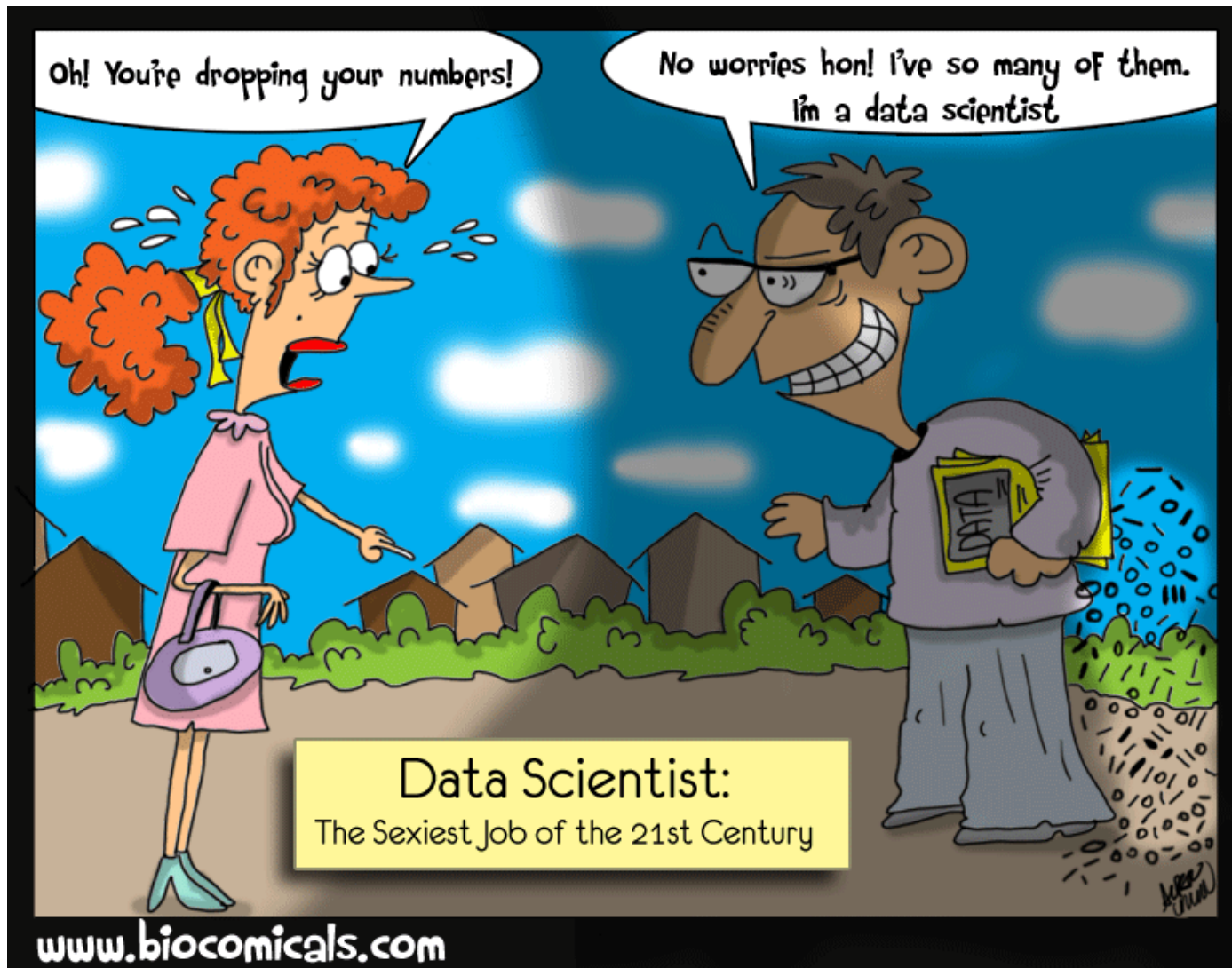
Data Scientist Qualities

The Data Scientist

- A New Role Exists – the **Data Scientist**
  - One Part Scientist/Statistician
  - Two Parts Sleuth/Artist
  - One Part Programmer
  - Focused on *data* not models
- Working with **analysts** to create business value

Confidential Think Big Analytics

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

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

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

Strata Survey Skills

Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

Strata survey, 2013

Skills and Self—ID Top Factors
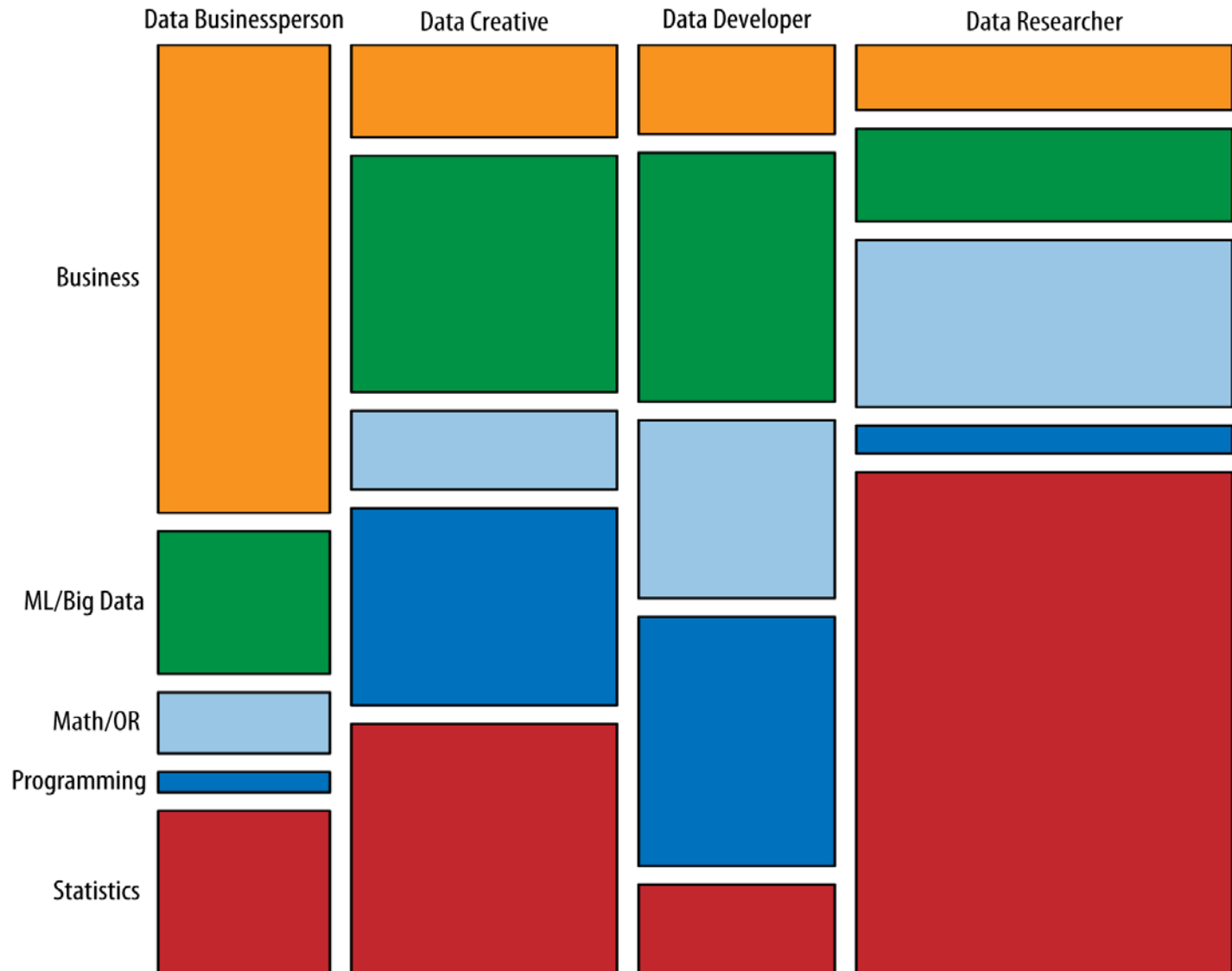
Range of Skills

- **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

SAN DIEGO SUPERCOMPUTER CENTER

at the **UNIVERSITY OF CALIFORNIA; SAN DIEGO**
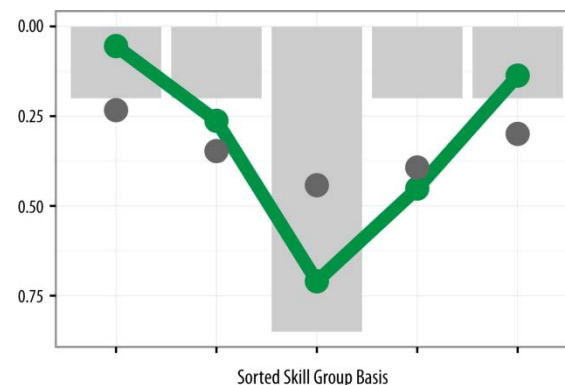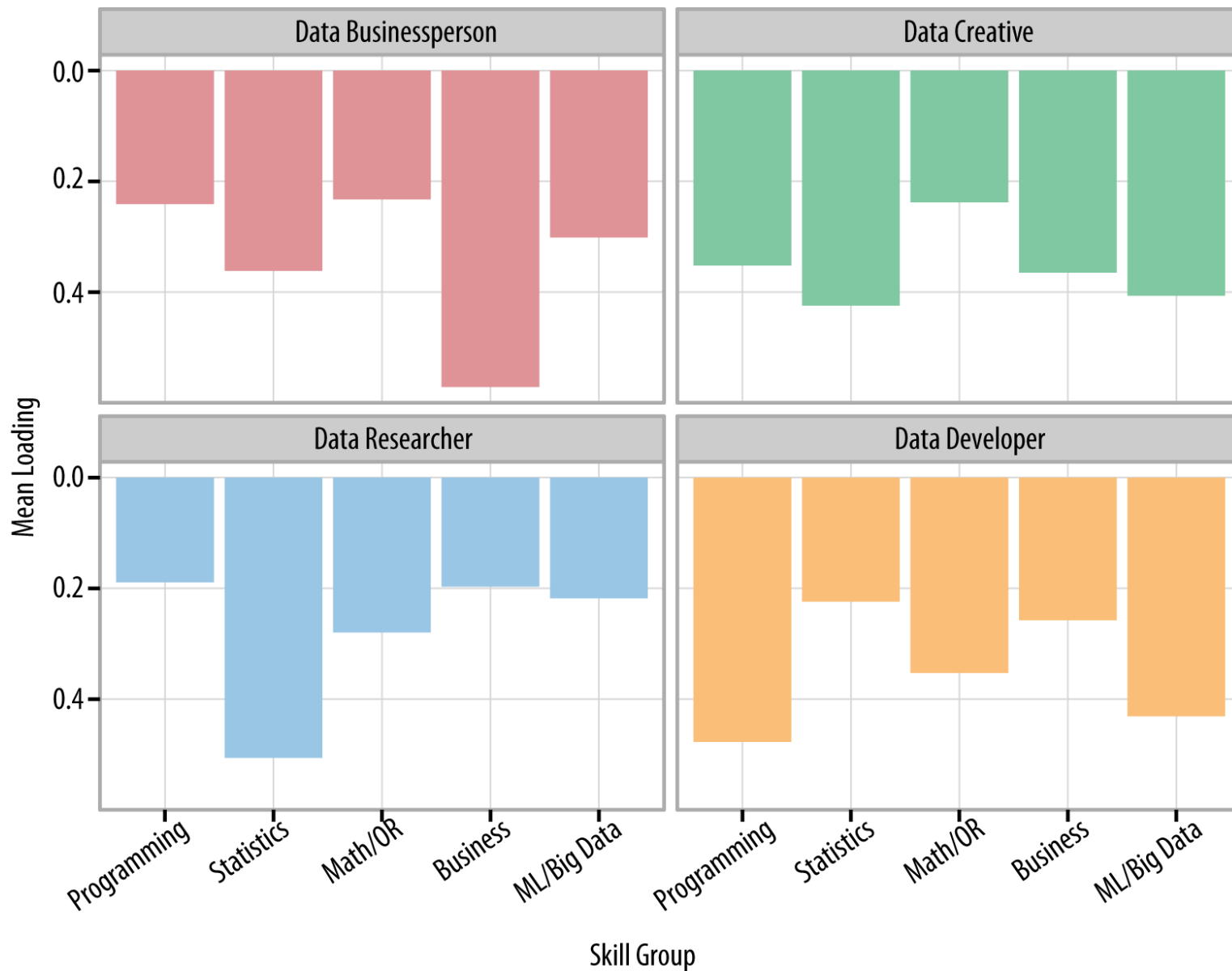
Learning and Training Opportunities

- **Many MS, MAS, Courses, Training, Workshops, Boot camps, etc.**

- **Coursera – Introduction to Data Science**
  - 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

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?

Number of Year	Responses
< 1 year (6)	2%
1-2 years (33)	12%
2-4 years (85)	31%
5-8 years (91)	33%
> 8 years (35)	13%
Not sure (28)	10%

KDnuggets survey [278 votes total]

SDSC SAN DIEGO SUPERCOMPUTER CENTER

at the UNIVERSITY OF CALIFORNIA; SAN DIEGO

UCSD

How long does it take for a beginner to become a good data scientist per Region?

| Region (Count)	Avg Years to become a good data scientist
AU/NZ (9)	6.9 years
E. Europe (19)	5.9 years
US/Canada (143)	4.9 years
W. Europe (60)	4.9 years
Asia (25)	4.9 years
Africa/Middle East (9)	4.4 years
Latin America (12)	3.9 years

Forbes: LinkedIn's Monica Rogati On "What Is A Data Scientist?"

- **Data-science curriculum at a school, what would the syllabus look like?**

- **Learning by doing is an essential component!**
- **A data science curriculum would need a strong lab or practical, applied component**
- **That would be central; that is what data scientists do, day in and day out**

Salary comparison

- **Burtch Works study**
- **Men are the majority in analytics and consistently paid higher than women**
- **Women - minority in the analytics community (STEM)**
- **33% of entry-level candidates are women, reduces to 12% at the senior executive level**
- **Women are paid less, making between 90 and 98% of men's salaries depending on job level**
- **Women in Big Data fare better than the nationwide average across all job fields (77%)**
- **The Burtch Works Study was conducted on a sample of 2,845 professionals nationwide and at all job levels**

Data Scientist Gender Salary Comparison
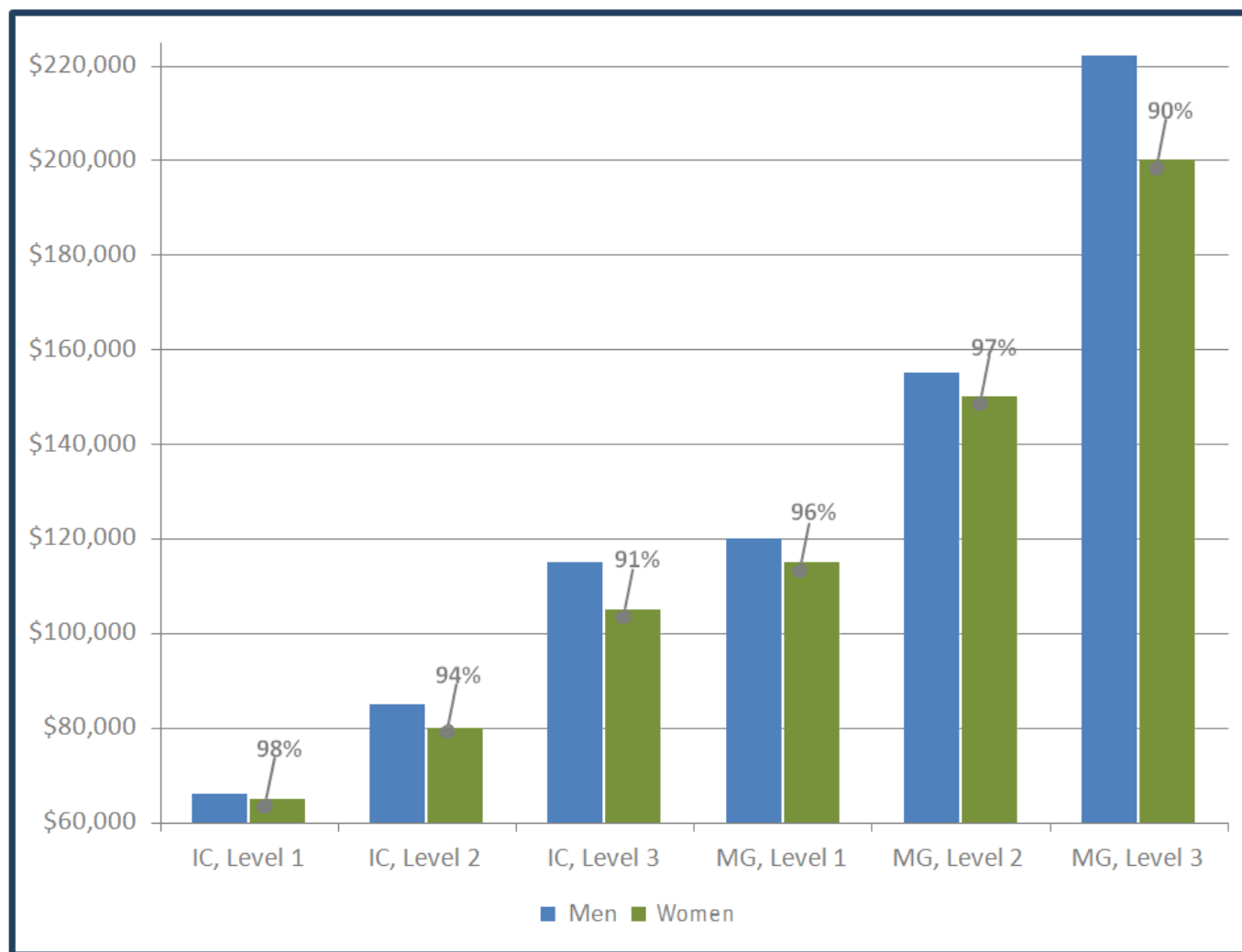
Key to a Great Data Scientist

Τεχηνιχαλ σκιλλσ (Προγραμμινγ, Στατιστιχσ, Ματη)

    + Χομμιτμεντ

        +Χρεατιϖιτψ

            + Ιντυιτιον

                +Πρεσεντατιον Σκιλλσ

                    +Βυσινεσσ Σαϖϖψ

    = Γρεατ Δατα Σχιεντιστ!

SDSC

UCSD

Key to a Great Data Scientist

Technical skills (Coding, Statistics, Math)

+ Commitment

+Creativity

+ Intuition

+Presentation Skills

+Business Savvy

= Great Data Scientist!

PACE Education

- **Data mining Boot Camps**
  - Boot Camp 1
    - September 12-13, 2013
  - Boot Camp 2
    - October 17- 18, 2013
- **Tech Talks - 3rd Wednesday**
- **Workshops – one day "latest and greatest"**
- **Bookclub, White papers, "Tool-off"**
- **"Bring your own data"**
- **Summer Institute**
- **Data Science Institute**

SDSC SAN DIEGO SUPERCOMPUTER CENTER

at the UNIVERSITY OF CALIFORNIA; SAN DIEGO

UCSD

KEEP CALM AND ANALYZE BIG DATA

SDSC **SAN DIEGO SUPERCOMPUTER CENTER**

[Mike Gualtieri's blog](#)

at the **UNIVERSITY OF CALIFORNIA; SAN DIEGO**

UCSD

Thank you!

SAN DIEGO SUPERCOMPUTER CENTER

at the UNIVERSITY OF CALIFORNIA; SAN DIEGO

Questions?

- **www.sdsc.edu**

- **For further information, contact Natasha Balac (nbalac@ucsd.edu)**