Why Data Science Fails

The Key Ingredients for Highly-Successful Data Science Teams
Scott Ernst
Director of Data Science & Engineering

When I Work®
Optical Fluid Characterization

“Time-Resolved Photon Correlation Spectroscopy”

Shine a laser at a liquid to determine what it contains
Computational Astrophysics

“What happens when stars collide?”

“Magnetohydrodynamic Shock-wave Stability Simulations”
ML-Enhanced CG Animation & VFX

“Supervised & Unsupervised Behavioral Character Locomotion”

Can CG Characters Walk without Animators?
Application: Dinosaurs

How did they really move?
ML-Enhanced Video Production

“Rapid production of short, high-energy advertising videos”

How fast can a company make an awesome video?
Dinosaur Data Science

“Predict behavioral information from massive tracksite in Switzerland”
Why Data Science Fails

The Key Ingredients for Highly-Successful Data Science Teams
You’re doing the math wrong!
But First

Some Historical Perspective
Answer to the Ultimate Question of Life, the Universe, and Everything...
Web Development
Success & Failures

Amazon

Google

boo.com
One of the biggest failures at Boo was to assume that [web development] was not a technology issue. Up through launch and beyond, the [web] team was first reporting to business development and then to marketing.

- Tristan Louis, CTO
Joel Spolsky

Co-Founder & CEO of stackoverflow

Founder of Trello
The Joel Test
Yes/No Questions
For Assessing the Quality of a Software Team

1. Do you use source control?
2. Can you make a build in one step?
3. Do you make daily builds?
4. Do you have a bug database?
5. Do you fix bugs before writing new code?
6. Do you have an up-to-date schedule?
7. Do you have a spec?
8. Do programmers have quiet working conditions?
9. Do you use the best tools money can buy?
10. Do you have testers?
11. Do new candidates write code during their interview?
12. Do you do hallway usability testing?
Data Science
What is the “Joel Test” for Data Science?
Keeping in mind that Success is not the absence of complete failure
Pipeline & Process

Low-Friction + High-Value Analyses
Data Science Pipeline (Simplified)

1. COLLECT
2. PROCESS
3. RESEARCH
4. DEVELOP
5. DEPLOY
6. VALIDATE
“Garbage Plant”
Poor experiments & improper data collection produce garbage results

Controlled experimentation is the most critical tool in the data science arsenal. Teams that don’t make regular use of it are doing something wrong.

- Drew Harry, Data Science @ Twitch
“Silver Bullet”
Machine Learning (ML) will save us

1. COLLECT
2. PROCESS
3. RESEARCH
4. DEVELOP
5. DEPLOY
6. VALIDATE
“Debt for Deadlines”
Where would we find the time...
“Blindly Charging Ahead”
Are the models actually improving
“80% of what we call analytics is not analytics at all but just hard work”

- Werner Vogels, CTO @ Amazon.com
Diversity

Skills & Background
Highly Interdisciplinary Field

Requires expertise in a very wide range of skills

Data Science Skills
First Hire

How were they chosen?

Expertise

- Extreme
- Low

Data Science Skills
Selection Bias

A Data Scientist is someone like me...

Data Science Skills

Expertise

Extreme

Low
A Team Missing Skills

Important elements for success are lacking

Data Science Skills
If All You Have is a Hammer...
Integration

Working with the Business
Localized or Distributed?

Only the beginning
Team Collaboration

Solo Integration

Cooperative Integration
Consider Department Goals & Structure
Data Science is Diffusive Across Departments

Data Science Contributes Broadly
Tooling & Structure

Beyond the Basics
Basic Toolchain

ETL ➔ Storage ➔ Analyze
How do these Tools Help Overcome...

Entropy

Fragmented and Disparate Data
How do these Tools Help Overcome...

Verbal Documentation

Tribal Communication of Knowledge
How do these Tools Help Overcome...

Patchwork of Access
Data Landscape looks Different among Peers
How do these Tools Help Overcome...

Inconsistencies

Results Should not be Snowflakes
Governance

Data and results must be consistently managed for reliability and access
Reproducibility

All analyses should be **easily** reproducible at any time in the future.
A Maslow’s Hierarchy of Needs

ETL  "food"

Storage  "shelter"

Analyze  "water"
Review

Checking Work
Elements of Data Science

Mathematics

Advanced Analytics

Data Science

Traditional Research

Computer Science

Data Engineering

Domain Expertise
Test & Validate Everything

- Visualization/Presentation Review
- Code Review
- Assumption Testing
- Automated Unit/Functional Testing
Competent Programmer Hypothesis

Most program faults are syntactically small and can be corrected with a few keystrokes
Test Code!

Source Code:
```python
def add(a, b):
    return a + b
```

Mistaken Source Code:
```python
def add(a, b):
    return a * b
```

Successful Unit Test:
```python
def test_add(self):
    self.assertEqual(add(3, 2), 5)
```

Failed Unit Test:
```python
def test_add(self):
    self.assertEqual(add(3, 2), 5)
```
Collaboration

Data is a primary Business Concern
Data Afterthought

- Concept
- Planning
- Development
- Production
- Iteration
Data Driven

- Concept
- Planning
- Development
- Production
- Iteration
High Leverage Problems

Focus on Business Impact
Levels of Significance

Statistical

Effective

Business
Levels of Significance

Statistical  "Observable"

Effective  "Meaningful"

Business  "Actionable"
The Bad Way

Statistical “Observable”

Effective “Meaningful”

Business “Actionable”
The Good Way

Statistical  “Observable”

Effective  “Meaningful”

Business  “Actionable”
Communication

Data Flow Through Business
How Results are Consumed
Bottom Up
Direct to Decision Makers
Decision Level Data Partnership

“C-Level Buy-In”
So What is the “Joel Test” for Data Science?