GENERAL ASSEMBLY

DATA SCIENCE FOR MANAGERS

Stephen Merity
Machine Learning Consultant

Stephen Merity (@Smerity)

Currently consulting on data analytics & machine learning projects at numerous start-ups in Sydney and remotely

I’ve previously worked at Google, freelancer.com and others

BIT @ University of Sydney (First Class Honours + University Medal)

Also co-lecturing the post-graduate Computer & Network Security at the University of Sydney with Matt Barrie, CEO of Freelancer.com

Agenda

- What is Data Science?
- Attracting Data Scientists
- Integrating Data Scientists
- Practical Examples of Data Science & Management

(special note: slides *will* be available afterwards)

Focusing on what’s important

We won’t be discussing:

Maximum Entropy classifiers
Support Vector Machines
Neural Networks
Hadoop

Random Forest classifiers
Logistic Regression
Regularization
HBase

All of these are
tools & methods

They’re used to solve
business problems

Our Focus

As a manager, use your business expertise to:

locate these problems and opportunities

and then

provide direction and support for data scientists to solve them

DATA SCIENCE FOR MANAGERS

WHAT IS DATA SCIENCE?

(AND WHY SHOULD YOU CARE?)

Disclaimer: A muddled term

Data Science has no strict definition and covers many concepts

An evolution of the term “data mining” to entail what mining doesn’t.

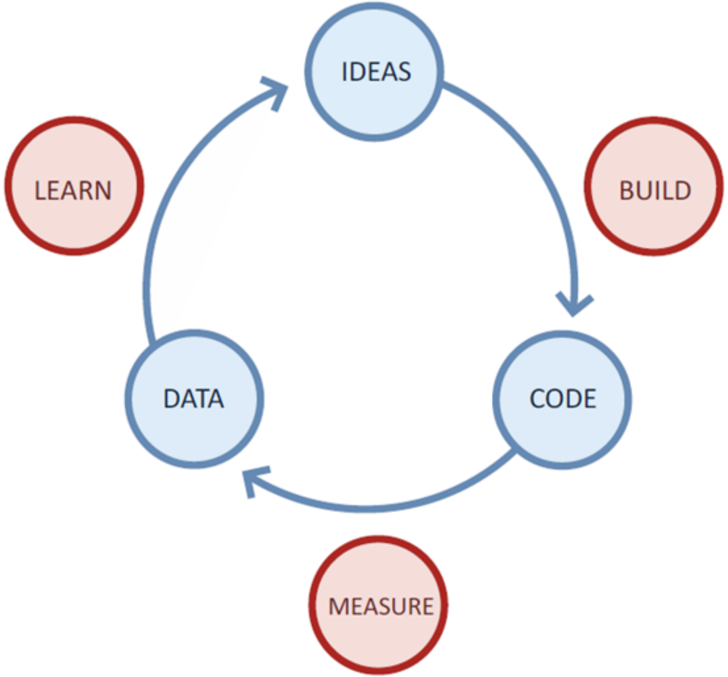
Data science has become more and more overloaded: covers many fields, used as a buzz word, used as a HR keyword.

(HBR: Data Scientist: The Sexiest Job of the 21st Century)

Data Science ⇒ A Loop of Three

1) (Ideas) Identify Business Problem or Opportunity
2) (Code) Acquiring & Handling Data
3) (Data) Interpretation using scientific principles

Data Science & The Lean Startup



http://theleanstartup.com/

Identifying Problem/Opportunity

The best data scientists not only understand the business they work for, they can help you understand your business better

Frame a business problem/opportunity in terms of a hypothesis: a question which can be definitively answered by acquiring the correct data

Identify the “leap of faith” assumption behind the idea

Acquiring & Handling Data

What important data don’t we have and how do we acquire it?

What data do we already have that isn’t properly stored or used?

Data sources are commonly too large for traditional data processing, unstructured and distributed amongst many data “silos”

Proprietary data and how it can be exploited is what fuels many modern companies

Science for Manipulating Big Data

Processing these datasets needs to be done efficiently & robustly

Efficiently: machine learning, Map Reduce, cloud computing, …

Robustly: scientific rigour, confidence intervals, proper sampling, …

With current technology & resources, almost any data source can be processed

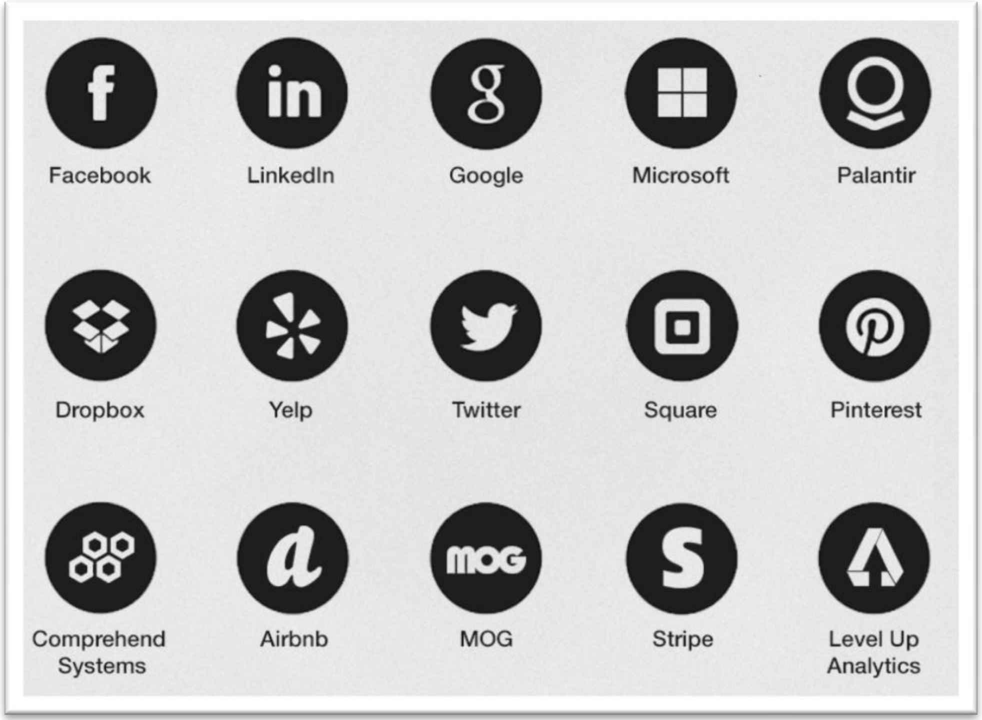
The real question is: what are you processing it for?

Generalists, Not Specialists

A data scientist is likely to use tools & techniques from many fields

As such, they need to be generalists: comfortable learning new concepts on the fly and working in new technology & novel problems.

Who hires Data Scientists?



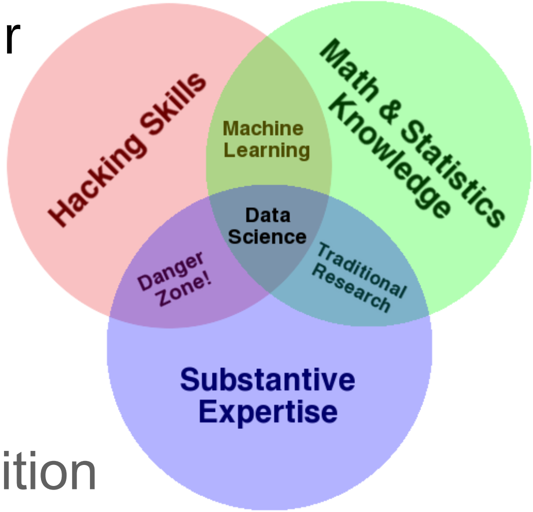
Actionable Insights

You won’t find a strict definition *anywhere*: it’s a loose & evolving term

Data scientists can describe the most likely ways a company’s data can create business value

Data Science = Practical Coding Skills + Scientific Rigour
+ Domain Understanding / Expertise

Data science aims to demonstrate that data drives
your business behaviour in a predictable manner



Drew Conway’s definition

ATTRACTING DATA SCIENTISTS

Finding the Right People: Skills

Top-notch quantitative analysis skills

Ability to produce production quality code

Self directed learner: willing to learn about both the technical field and your business domain

Finding the Right People: Mindset

Express complex ideas in simple terms

Interact productively with decision makers (engineering & management)

Excited by real world impact and the business decisions they drive

Right People ~= Data Scientists

The right candidate may not term themselves as a data scientist

Freelancer.com hires competent scientists in physics, robotics, computer science and other data-driven fields for Data Scientist positions

Insight Data Science (6 weeks – 100% placement rate)

Physics, Astrophysics, Mathematics, Statistics, Neuroscience, Bioinformatics, Engineering, Computer Science

HR & Data Scientists

Human resources needed a single keyword instead of…

Data Mining / Business Intelligence / Data Analytics / Machine Learning

As data science is multi-disciplinary, it became a catch-all term.

This could be a whole talk in itself, but:

Engineers tend to dislike HR, so use them with caution

HR & Data Scientists

HR are likely to discard and/or discourage promising candidates due to a fundamental lack of understanding:

This physics PhD doesn’t have 3+ years working with Hadoop?
That’s a no then.

Be willing to take on promising but “under qualified” candidates:

Numerous clients have converted undergraduate university students to key members of their data science team

Data Science is Competitive

Data scientists are being headhunted by everyone

How do you compete?

Convincing the Right People

The Data

Data Scientists commonly select fun & interesting over well paid

Promise of an interesting dataset to tackle

Real world impact & larger-scale recognition

Convincing the Right People

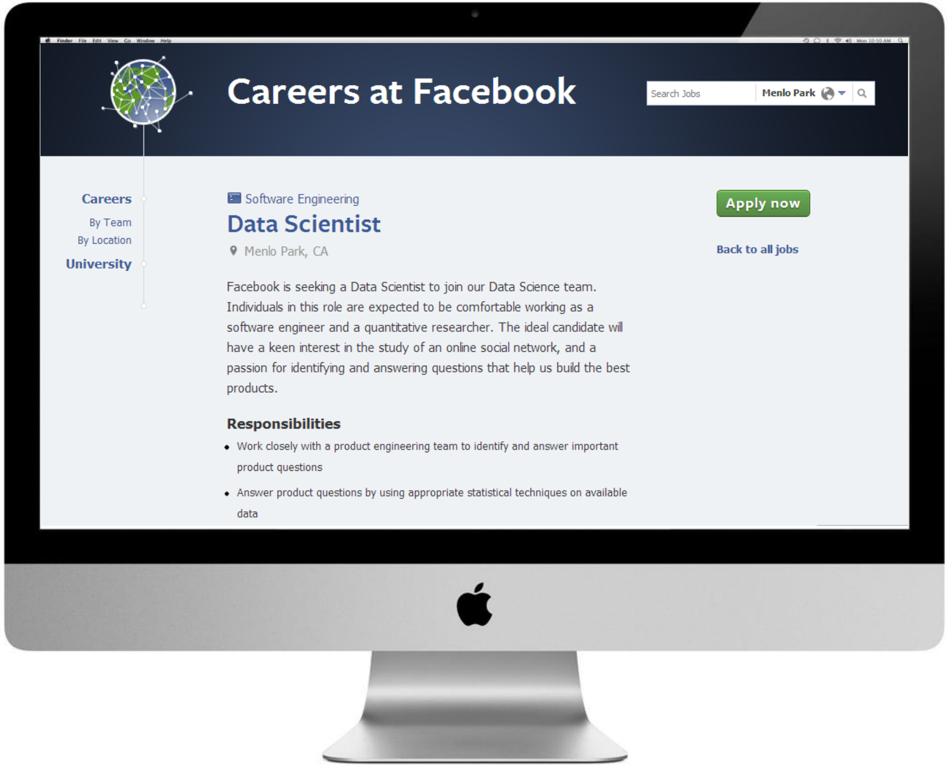
The Organisation

Engineering team are on-board and supportive

Confidence that their suggestions will be heard & acted upon

Promise of working with like-minded colleagues

Take Inspiration



Software Engineering

Data Scientist

Menlo Park, CA

Facebook is seeking a Data Scientist to join our Data Science team. Individuals in this role are expected to be comfortable working as a software engineer and a quantitative researcher. The ideal candidate will have a keen interest in the study of an online social network, and a passion for identifying and answering questions that help us build the best products.

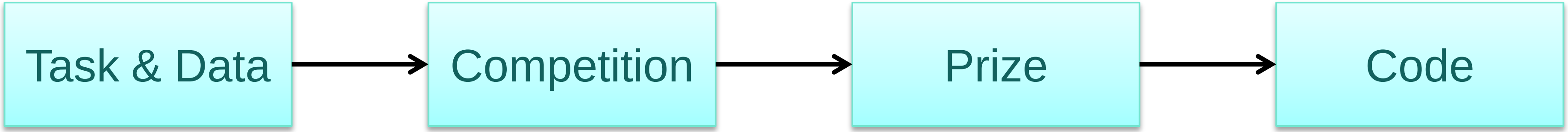
Responsibilities

- Work closely with a product engineering team to identify and answer important product questions
- Answer product questions by using appropriate statistical techniques on available data

https://www.facebook.com/careers/department?dept=data&req=a2KA0000000EaXuMAK

What is Kaggle?

Kaggle is the leading platform for predictive modeling competitions



Seek expert advise from numerous fields:

Physics, Bioinformatics, Computer Science, Mathematics, Statistics, …

Used by Facebook, NASA, RTA, StackOverflow, GE, Pfizer, HBR, …

Example Kaggle Competition: RTA

Task: Predict travel time on Sydney's M4
Data: Several years of historical travel time data for Sydney’s M4 freeway

Prize: $10,000 reward + Kaggle’s fee

(this can still be cheaper than a few days of a consultant’s time)

RTA provided data and domain knowledge, project handled by Kaggle

Example Kaggle Competition: RTA

3 month competition + $10k reward ⇒ 356 entries

Winning solution: Costa Rican PhD students in US and Canada

Later paid to collaborate further with the RTA

Issues with Kaggle

Solutions may be entirely useless for production (Netflix)

Optimising for the best result instead of the best understanding

Doesn’t ensure it’s a problem worth solving

Require skilled programmer / data scientist to create & clean initial data

Not “data scientists”: no full context, no feedback loop in your organisation

They’re “out of house” talent that may be un-hireable

Actionable Insights

Don’t just hunt for the word “data scientist”: be broad

Focus on self-directed learners with strong data-driven backgrounds

Provide a promising workplace (interesting problem and/or team)

Kaggle can be a recruiting tool or an approximation of a data scientist’s input: it is not, however, a replacement for an in-house data scientist

INTEGRATING DATA SCIENCE

Hiring a Data Scientist and putting them in your business does *not* make your business data driven

Two paths for data science

Data-assisted company:

Data science can be a division of engineering without tight integration to other divisions, running R&D style projects by request.

Data-driven company (definition from DJ Patil):

Data-driven organisation acquires, processes and leverages data in a timely manner to iterate on and develop new products, improve the company’s efficiency and navigate the competitive landscape.

Data-assisted companies

Data science becomes a new division in the engineering team

Tend to focus primarily on engineering projects

+ Easier sell for existing businesses

- Commonly has issues with broader impact on the company

Establishing a Data-Driven Culture

Companywide respect for measuring, testing & evaluating:

If you can’t measure it, you can’t fix it

Employees urged to base decisions on hard facts instead of intuition:

This applies to upper management as well

If hard facts aren’t available, work out the key steps in order to get them

Embracing a Data-Driven Culture

Freelancer.com successfully transitioned to a fully data-driven culture within two years of the CEO taking over the company.

This required the full buy-in of the CEO and CTO however.

This directive must come from upper management as it’s a fundamental change to the company.

Democratisation of Data

Data science and data analysis needs to be easy

Remove the bottleneck: inspire people, esp. non-technical, to use data

Easy-to-use data tools gets everyone thinking about data

The company dashboard makes companies proactive instead of reactive

Core Tenets of Data-Driven Culture

Instrument and collect as much data as you can:

If you don’t have it, you can’t use it

Measure and react. Are you products succeeding? If you don’t measure the results, how do you know?

Get people looking at and understanding the data, even the non-technical teams.

Foster curiosity as to why the data has changed or isn’t changing.

Actionable Insights

Minimum:

i. Establish data science team as part of engineering

ii. Ensure proper support from both engineering & upper management

Optimal:

i. Establish a data-driven culture for all employees

ii. Data scientists work hand in hand with management to support and push projects by setting business goals using data

DATA SCIENCE CASE STUDIES

Project Areas in Data Science

Decision science / business intelligence

Product & marketing analytics

Data products

Data engineering & infrastructure

HARRAH’S CASINO (1998)

SUMMARY

No relying on intuition or random bets but instead deep data analysis

1) Acquire rich repository of customer information
2) Slice and dice to develop marketing strategies
3) Identify core customers by predicting Lifetime Value (LTV)
4) Gather increasingly specific information about preferences
5) Appeal to those preferences and specific customers

KEY CHALLENGE/QUESTION

How well can you understand your business using data?

How does better understanding lead to higher profits?

If a casino is *not* relying on random bets & guesses, then neither should your company

“We use database marketing & decision-science-based analytical tools to understand customer incentives based on evidence instead of intuition”

A new strategy

Harrah replaced their COO with a Harvard Business School professor, Gary Loveman, who had experience working with numerous companies

The existing marketing unit was disbanded and replaced with a University of Chicago professor of mathematics, Richard Mirman

Mirman became Senior Vice President of Relationship Marketing

Idea

Build a function to understand how your business makes revenue

Understand the variables so they can be tackled individually

Harrah’s Casino focused on determining a customer’s lifetime value:

Frequency of visits

Total money gambled

Likelihood of returning

Triggers & responses

Demographic groups

Customer preferences

Harrah’s Strategy

Harrah’s strategy consisted of two elements:

+ Total Gold rewards program

+ Database Marketing / CRM

Total Gold Rewards Program

Used to track, retain, and reward guests nationwide

First attempt to consolidate their traditional data silos (individual casinos)

$30 million for development & $35 million for support

Database Marketing / CRM

Hotel system recorded customer’s stay, demographics & personal prefs

Slot machines recorded every game played, the time & bet size

Analysed the customer information for patterns & insights

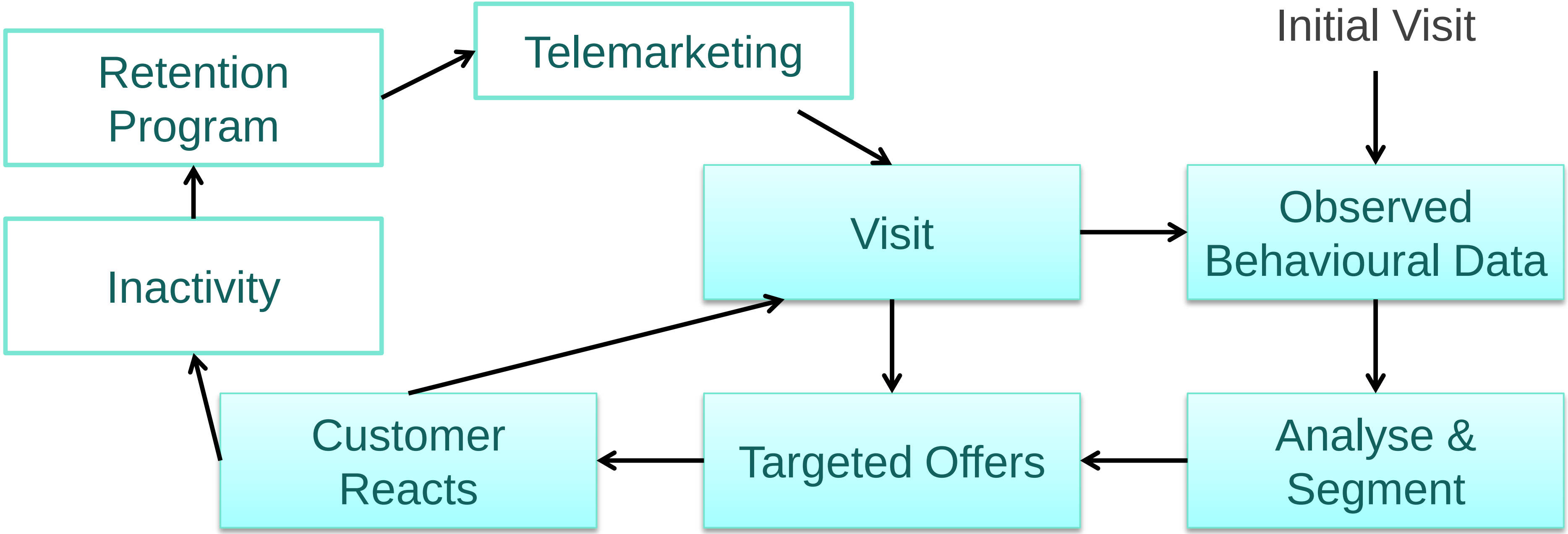
Database Marketing / CRM

Customised marketing initiatives broken down & tracked by customer segment

On an incoming call, customer details are loaded onto screen
This also provides suggestions as to upsells or offers that may work

The success of these offers was tracked, tweaked, and relaunched

Customer Retention Model



Initial Visit

Retention Program

Telemarketing

Visit

Observed Behavioural Data

Inactivity

Customer Reacts

Targeted Offers

Analyse & Segment

Focus on potential

Opportunity based marketing:

A customer may not look promising from the revenue observed at Harrah’s alone, but by careful analytical analysis they knew they were a great customer at Harrah’s competitors

Play $100 on Us, predict playing behaviour, targeted advertisements, analyse existing (Harrah’s) playing behaviour, difference = opportunity

Early A/B Testing

Group A ($125 package):

Free room, two steak meals, $30 of free chips at the casino

Group B ($60 package):

$60 of free chips at the casino

Result: free rooms were a waste, focus on free chips

Outcomes

Solidified customer loyalty by providing unique solutions to customers [Diamond/Platinum/Gold, locals, ...] (early concept of a “data product”)

Integrated data management system (and the patents on data processing / analysis they applied for) gave them a competitive edge

In two years (from 1997 to 1999), they’d doubled their net income

Outcomes

“The farther we get ahead and the more tests we run the more we learn. The more we understand our customers, the more substantial are the switching costs that we put in place, and the farther ahead we are of our competitors’ efforts.”

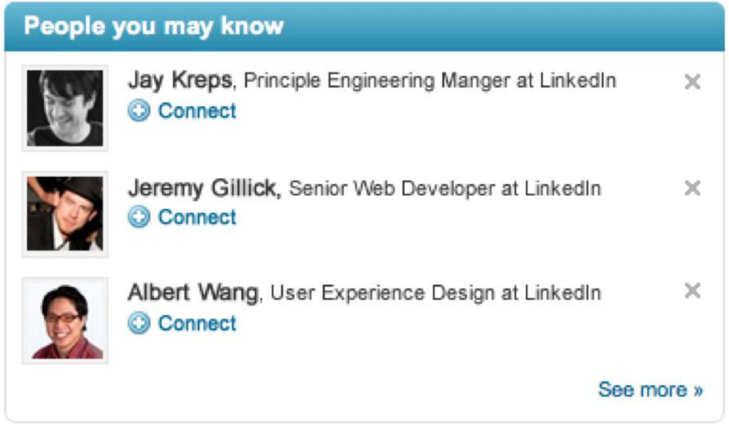
- Gary Loveman (Harvard Business School professor & Harrah’s COO)

LINKEDIN: PEOPLE YOU MAY KNOW (PYMK)

SUMMARY

Jonathan Goldman created PYMK for LinkedIn in 2006

PMYK is now a core growth mechanic of all social networks



People you may know

Jay Kreps, Principle Engineering Manger at LinkedIn

Connect

Jeremy Gillick, Senior Web Developer at LinkedIn

Connect

Albert Wang, User Experience Design at LinkedIn

Connect

See more »

KEY CHALLENGE/QUESTION

Multi-disciplinary: more about the person than the technique (PhD in Physics from Stanford)

Constrained by the engineering team: worked around them

Small, low cost experimentation

PYMK became critical for activation, retention and revenue

PYMK: Idea

PYMK seems a trivial idea now: not so at the time

The core questions to answer was how it would impact:

i. Long term user activity

ii. Long term site use

Was it worth the complexity of the implementation?

PYMK: (Hacky) Code

A hacked up version of PYMK was launched using the LinkedIn ad system

A. The complex calculations were performed for a few select users
B. Highly targeted (i.e. UserID = X) ads were created for PYMK
C. The click through rate of PYMK was tracked using the ad infrastructure

PYMK: Data

Result: the ad had click through rates 30% higher than LinkedIn’s own navigational elements

Provided proof to both engineering and management of the idea’s viability

PYMK: (Proper) Code

PMYK implementation was complex: using Hadoop, Project Voldemort (custom database based on Amazon’s DynamoDB), novel algorithms...

The tech is possible, the most important question to be asked was:

Will the positive impact justify the complexity?

(will the feature have a positive impact at all?)

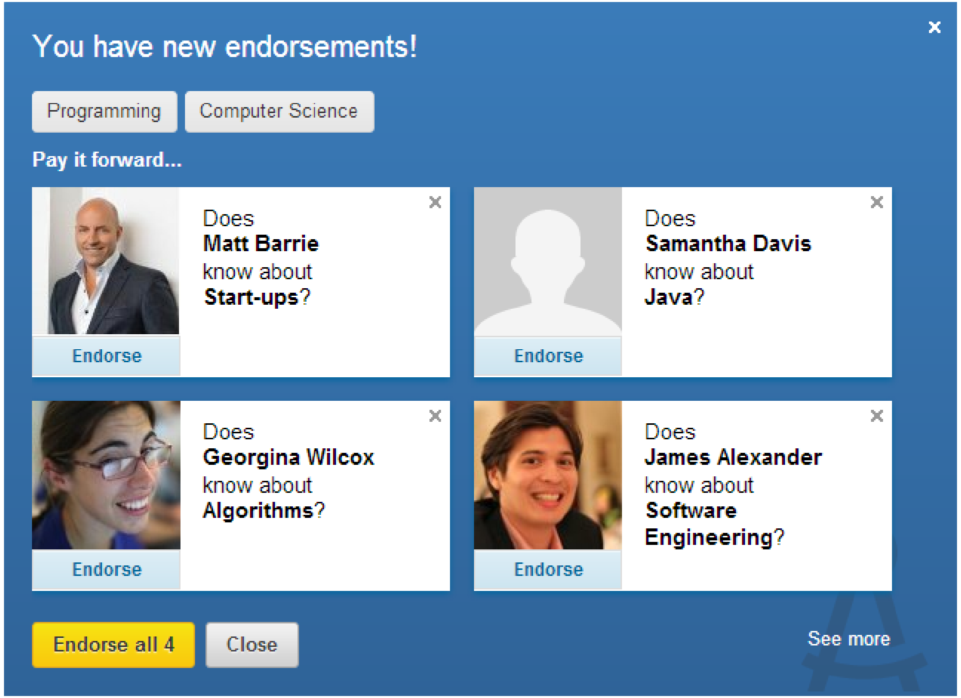
Answering this question quickly and cheaply is the core goal.

LinkedIn’s Focus: “Data Products”

Competitive advantage: proprietary data + innovative methods to extract insights

Facebook extended PYMK to monitor how long and how many friends are necessary before users are guaranteed to be engaged in the long term

Endorsements



You have new endorsements!

Pay it forward...

Does Matt Barrie know about Start-ups?

Does Samantha Davis know about Java?

Does Georgina Wilcox know about Algorithms?

Does James Alexander know about Software Engineering?

FREELANCER.COM

SUMMARY

Freelancer.com acquired by Matt Barrie in 2009

Existing database was poorly maintained: historical data not properly structured and current data not in a queryable format

The goal: understand the business through data. Identify strong customers, work out low value plans and services, focus on life time value (LTV) and cost per acquisition (CPA)

KEY CHALLENGE/QUESTION

Identifying key customer value

Data was badly structured due to being acquired from competitors

CEO & CTO ensured the the data science team had all the support they needed

Understanding the business through data: can predict a holiday through change in revenue from given regions

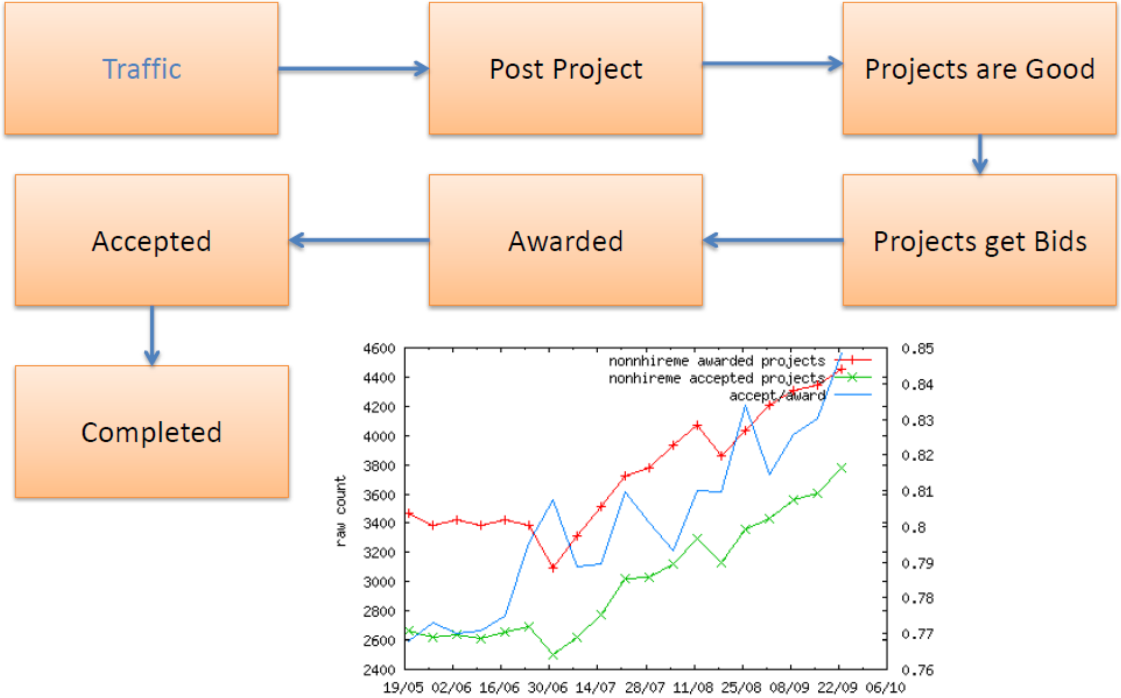
Understanding the business

Freelancer.com was founded in 2009 based upon an acquisition of GetAFreelancer.com

Existing database was poorly maintained: historical data not properly structured and data not in an easily queryable format

The goal: understand the business through data. Identify strong customers, work out low value plans and services, focus on LTV & CPA

Project Funnel Analysis



Traffic → Post Project → Projects are Good → Projects get Bids → Awarded → Accepted → Completed

Competitor Analysis

Understand your competitor’s moves before they do: replicate successful moves, avoid their pitfalls

More than just web stats, can be detailed statistics (new users per day, new projects per day, user attrition, ...)

Use knowledge from acquisitions to prove definitively your tracking accuracy

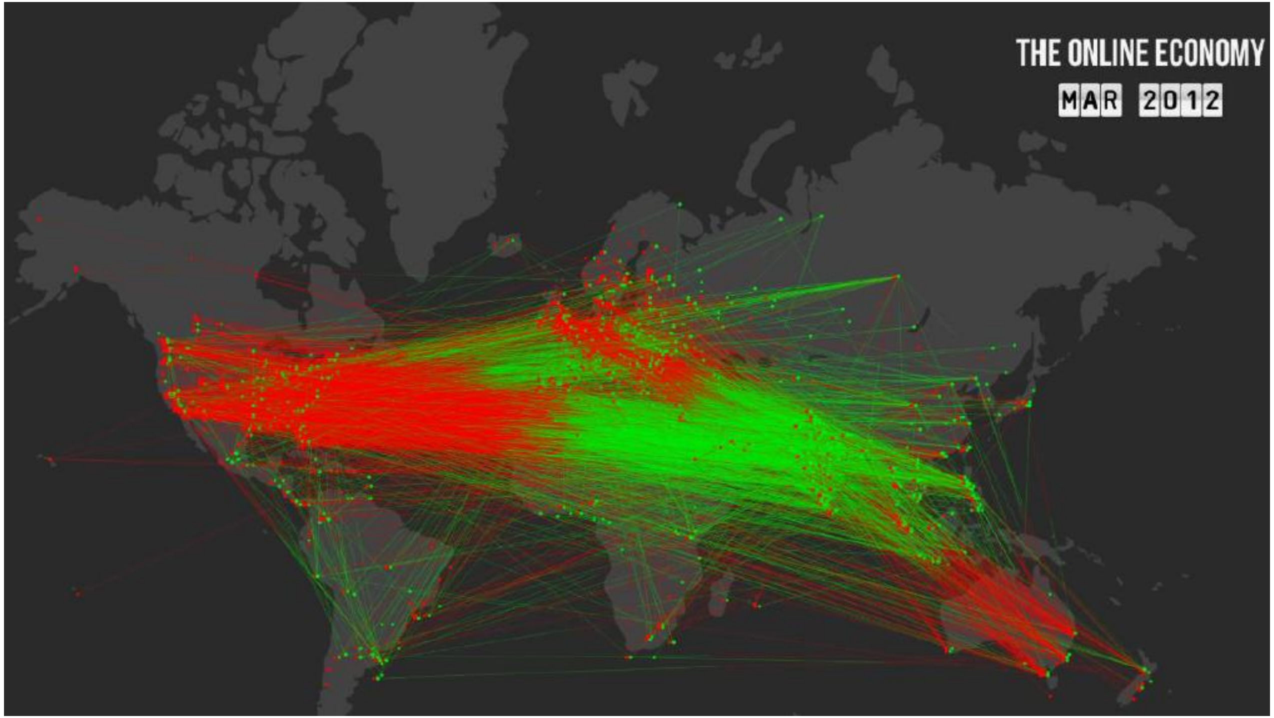
Freelancer.com’s Dashboard

Freelancer.com have an extensive internal dashboard: all employees are expected to look at it, understand it, and track their relevant metrics

Dashboard is updated nearly instantly with real-time figures: hourly, daily, weekly, yearly.

The CEO needs a calendar of global holidays. Why? The most common reason for drops in activity from specific countries.

Freelancer.com’s World Map



Data Science Recruitment

The CEO of Freelancer.com, Matt Barrie, runs two courses at the University of Sydney: Tech Venture Creation, Computer & Network Security

Both are funnels to pull bright students into his business, most of them to become data scientists

Actionable Insights

Data science needs to be supported by engineering & management to succeed

Do minimum work to test hypothesis, create full product on success

Democratise data using company dashboards

Find funnels in your business and put the time into understanding them

WRAPPING UP

For Your Business:

The interaction between data scientists and management is *crucial*

It is not enough to apply data science to your projects and strategies after the fact, it needs to become part of your company’s culture

Using and understanding your company’s proprietary data can lead to unique competitive advantages

Learning More

Refer to the documents in this presentation’s data pack

Follow / read articles written by:

i. DJ Patil (@dpatil) – Data Scientist at LinkedIn / Greylock

ii. Jeff Hammerbacher (@hackingdata) – Chief Scientist at Cloudera

Q&A

Data pack:

smerity.com/media/talks/data_science.pdf

smerity.com/media/talks/data_science_pack.zip

Stephen Merity (smerity@smerity.com)