



DIGITALAMBIT

Who will win the 2016 Stanley
Cup?

Dagny & Cayla Evans

Agenda

- Introductions
- Project Overview
- Methodology
- Hockey Stats Complexity
- Results
- Lessons Learned

Who are we?

Cayla Evans

- Junior @ Bishop Ireton HS
- National bound hockey player
- No prior work experience

Dagny Evans

- Entrepreneur
- Expert in process management, project management and data analytics
- Degrees from AU and GW
- Advocate & supporter for WIT and young women pursuing STEM

Project Overview

In Scope

- Using big data techniques to predict who will win the 2016 Stanley Cup
- Leverage interest in sports to expose technology to Cayla

Out of Scope

- Not a hardcore statistics class
- No game-by-game stat collection or analysis

Tools & Sources

- R & R Studio
- Various websites
 - Helpful website lynda.com
 - nhl.com
 - stats.hockeyanalysis.com
 - the teams' personal website
- Excel/comma separated value text files
- Book: Practical Data Science in R (Nina Zumel & John Mount)
- Github – presentation, data files & R scripts posted (<https://github.com/dagnyevans/stanleycup>)

Methodology

1. Find & download the data
2. Combine disparate data sources
3. Cleanse data (spelling, cases)
4. Use Excel & R to analyze data
 1. Looking for data quality & correlations in stats to winners
5. Calculate mean of historical player stats as 2015-2016 stats
6. Aggregate player stats to team stats*
7. Train & test models against data sets

Project Details

- Data & R script walk-through
- Data Overview
 - History records: 4,352
 - Seasons: 5
 - Teams: 30
 - Players: 1,421

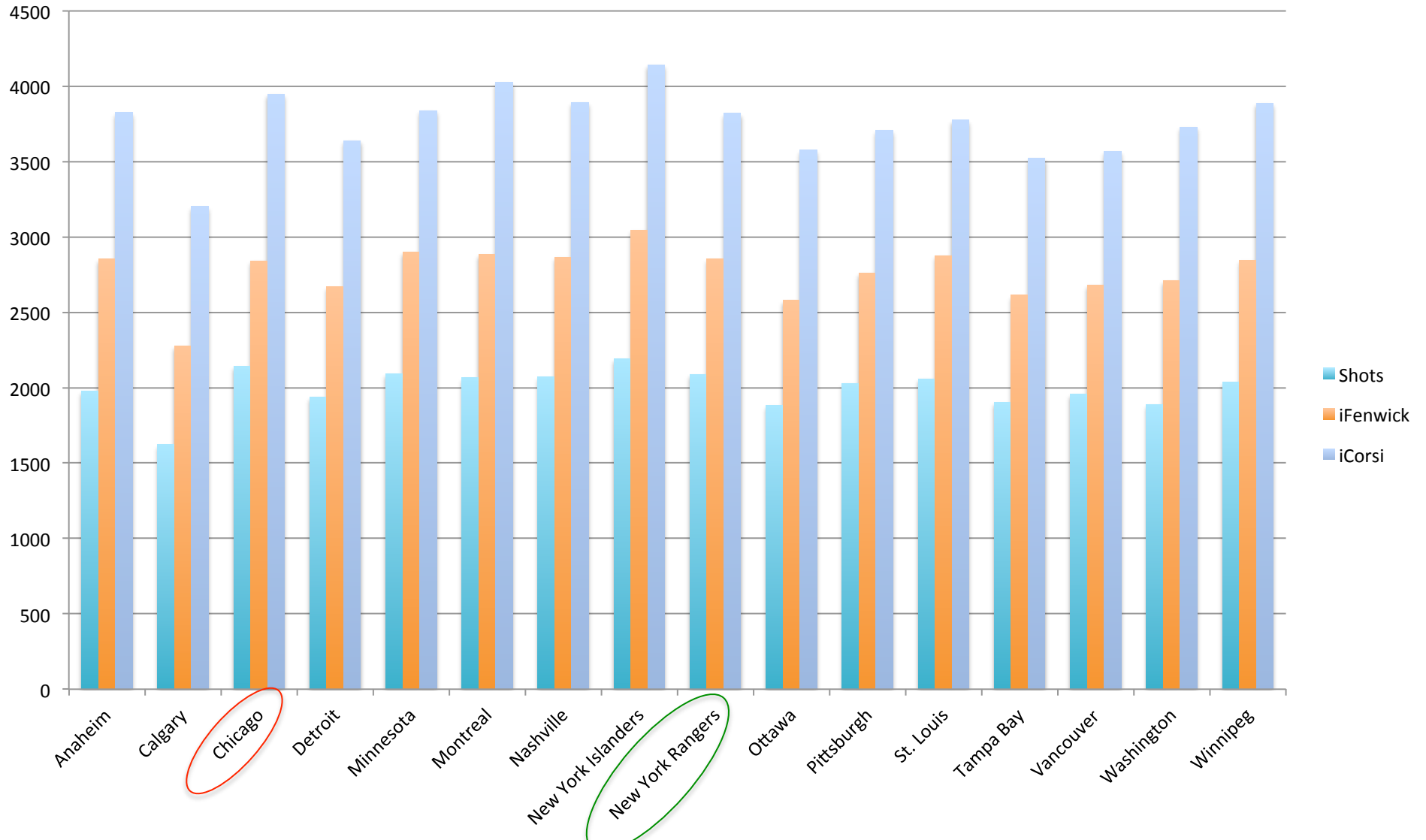


DIGITALAMBIT

Complexity in Hockey Stats

- History of Hockey Stats/Inherent complexity
 - Shots on goal is primary stat used in hockey
 - Still trying to figure out proper player stats
- Other factors
 - Best team does not always win
 - Humans have bad days
 - Performance of team is sum of player performance

2014-2015 Team Performance



How'd we do?

- Learned fundamentals of data analysis
- Learned R syntax for: loads, functions, merges, modeling, & analysis
- Cleansed and merged data to get to clean data set for modeling
- Used history to predict 2015-2016 player stats
- Ran models and correlations to forecast winner

On any given day, any team can win

Passing the torch

- Expand data set to include playoff participants and game by game player stats
- Try alternate models
- Share your work!

Reminder: data sets, script and powerpoint all available at: <https://github.com/dagnyevans/stanleycup>

Cayla's Lessons Learned

- Remember to save the work you do so that you do not have to repeat yourself
- Computers are stupid and will do exactly what you tell them to
- The data you start out with is not always the data you need
- Trial and error
- Map your project
- Take notes – process, progress and results



DIGITALAMBIT

Dagny's Lessons Learned

- Don't assume your intern knows everything you do
- Act -> Review -> Proceed -> Repeat
- Just because you have the tools, doesn't mean you can answer the question
- Clear, concise written reference & how-to instruction for r (or data science) are hard to find
- If you use an interesting subject to introduce tech ideas, you can engage (and teach) young people about tech

Contact Info

Dagny Evans
Digital Ambit

dagny@digitalambit.com
dagny@dagnyevans.com

@dagnyevans
@digitalambit

<https://github.com/dagnyevans/stanleycup>