Crowdsourcing in Data Science

José A. R. Fonollosa

Universitat Politècnica de Catalunya

jose.fonollosa@upc.edu

May 21, 2014

Crowdsourcing in Data Science. Index

- Open Innovation, Crowdsourcing, Co-creation
- Crowdsourcing in data science: data collection
- Crowdsourcing in data science: data processing
- GE Flight Quest 1. Arrival Time
- GE Flight Quest 2. Flight Optimization
- Crowdsourcing: Pros, Cons, And More

Open Innovation, Crowdsourcing, Co-creation

- Club of experts.
  - Preselected: e.g. USA Defense Advanced Research Projects Agency (DARPA)
  - Open: e.g. NIST evaluations (related or not with DARPA projects)
- Crowd of people.
  - Multiple simple tasks
  - Search for a brilliant idea from a hidden genius
  - The Wisdom of Crowds
- Coalition (competing partners)
  - Standard definition
  - Sharing databases
- Community (group with similar interests and goals)

Crowdsourcing in data science: data collection
Large number of very simple tasks.

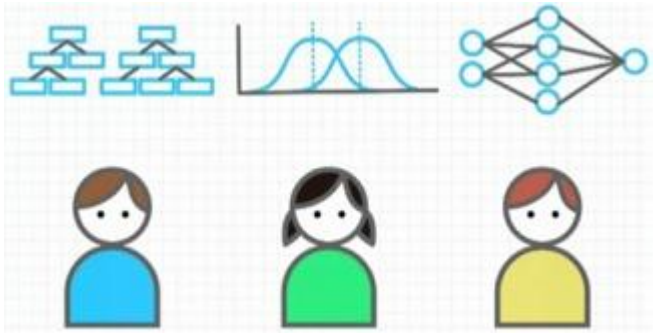
- Data collection, transcription and assessment
- Data categorization
- Sentiment analysis
- Gold standard
- Evaluation

Crowdsourcing in data science: data processing Singular, focused difficult problems

- Data prediction
- Data classification
- Feature extraction and models
- Optimization
- Data visualization

kaggle CodaLab ALPHA INNOCENTIVE® NINESIGMA℠ Accelerating the Innovation Cycle

kaggle



Crowd of experts. Data Science

- Most companies have huge amounts of corporate data and the need to make sense of it.
- It’s not easy to find the required talent or solution.
- Particular solutions might cost millions of dollars.
- Kaggle introduces innovative crowdsourcing ideas as the concept of “real-time” competition with an objective evaluation.
- A Gamification approach to research in data science

✈ Flight Quest 1

in partnetship with Alaska Airlines kaggle

- **Prediction of the runway and gate arrival time for each airplane**.
- Evaluation: Root-mean-square deviation (RMSE) between the predicteds and the truth (0.75×RMSE_Gate)+(0.25×RMSE_Runway)
- Data
  - Aircraft Situational Display to Industry data (ASDI): Flight plan, tracks, …
  - Flight history: scheduled and actual gate/runway departures and arrivals.
  - Weather: METAR, Airsigmet, FB wind, TAF
  - Air Traffic Control System Command Center (ATSCC)
- Results: Average errors of 4.2 and 3.2 minutes for gate and runway arrivals. (40% and 45% improvements)

Flight Quest 2

in partnetship with Alaska Airlines kaggle

- **Optimize flight paths so airlines can reduce cost, avoid bad weather, and get to their destinations on time**.
- Route is a list of waypoints: latitude, longitude, airspeed, altitude
- Cost function calculated by a flight simulator but using **predicted** weather and traffic conditions.
- Evaluation: flight cost (USA$) evaluated with **actual** weather and traffic conditions: fuel, delay, turbulent zones, changes of altitude.
- 3 phases: milestone (1 month), main and final.
- Updated simulator for the main/final phase. Restricted and turbulent zones.

Flight Quest 2

in partnetship with Alaska Airlines kaggle

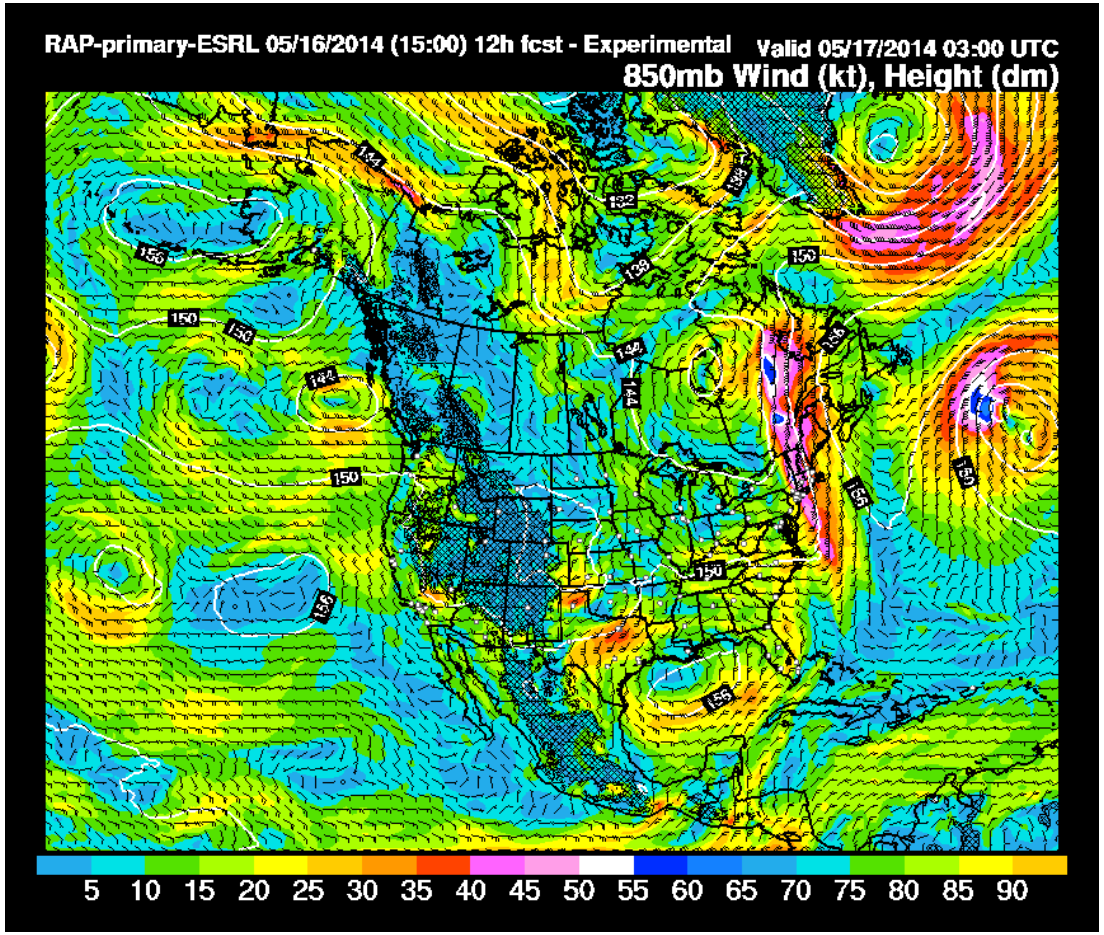
- **Provide a route for about 1000 real domestic USA flights that were on the air at a given time.** (for 14 different days)
- Aircraft position and altitude
- Destination airport
- Scheduled arrival time
- Number of passengers (standard and premium)
- Fuel cost
- Delay cost (crew, standard, premium)
- Turbulence ‘cost’ (USA$ for each minute)
- Restricted and turbulent zones.

Wind forecasts. NOAA Rapid Refresh (RAP)

NOAA NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
UNITED STATES DEPARTMENT OF COMMERCE

Earth System Research Laboratory
Rapid Refresh (RR)

- High-frequency updated (1h)
- Short-range weather model forecasts (each 1h out to 18h)
- 13-km horizontal resolution



Rules

- Code provided only to participate in the Competition.
- Entry must not include open source licensed under GPL, AGPL (copyleft) since derived works can only be distributed under the same license terms
- The competition is purely data driven and the final results are determined solely by Leaderboard ranking on a Final Evaluation Leaderboard
  - Milestone prize: $30,000
  - First prize: $100,000
  - Second prize: $50,000
  - Third prize: $40,000
  - Fourth prize: $30,000

Flight Quest 2

in partnetship with Alaska Airlines kaggle

To claim a prize:

- Provide the final model (agent) code, documentation
- Declaration of Eligibility
- Intellectual Property Assignment of Rights.
- Code scanned and/or reviewed for intellectual property risk and/or security vulnerability.

Results (GE press release)

The winning model proved to be up to **12 percent more efficient** when compared to data sets from actual flights.
If each flight worldwide reduced the distance by only 10 miles, airlines could reduce annual fuel consumption by 360 million gallons and save more than **$3 billion**

#	Team Name *	Score
1	jarfo *	11685.64
2	charango *	11714.16
3	Willem Mestrom *	11714.30
4	Murashka *	11714.73
5	Taki & Chris	11717.79
6	Doug Koch	11730.86
7	RouteFinder	11748.42
8	xing	11752.77
9	DerDasMachenMuss	11761.32
10	id	11762.46

Crowdsourcing: Pros, Cons, And More

- Fast, multiple solutions. Fixed cost.
- Different types of rewards: money, job, paper, points, reputation, …
- The reward, time period has to be proportional to the difficulty and novelty of the challenge.
- Competition instead of collaboration, but
  - Team mergers are allowed, Forum for beginners, baseline systems
  - Two or more phases
  - Top ranking solutions may be combined after the completion
- Confidentiality issues: data, software, innovative ideas, but
  - Errors detected, feedback
- Different types of competitions: industrial, scientific, educational.