

Machine Learning

In the context of a large enterprise

Arjun Viswanathan

(Head, Rates Big Data, Citibank Global Markets Limited)

Contents

- A step by step guide to best practices
- The Four Pillars of Good Machine Learning Implementations
- The Twelve Steps towards more productivity
- A.C.R.O.N.Y.M: an easy to remember framework.

(Gotcha!) The real Contents :

20-25 Minutes :

- Intro , who I am , what I do , why I like Matlab.
- Useful heuristics for getting value out of ML in a corporate setting.
- Some personal opinions on machine learning good practices
- (the fun bit) : 2 mini ML / Visualisation projects.
- (Final slide...) What does this all mean for humanity?

5-10 Minutes :

- Questions/discussion.

I am speaking here as a private individual. Any opinions expressed are my own.

This talk does not hold proprietary Citi data or Client data. All datasets used are public.

When I say “ML” I mean “Machine Learning” but Matlab could work just as well.

Intro

- History
- What I Do now
- Triple mandate: Use Machine Learning & all our data to:
 - 1) Make us better (ie more revenue in a compliant way)
 - 2) **Keep our employees happy and fulfilled as we enter this new world**
 - 3) Raise the profile of Citi (& Rates in particular) in the ML / AI space

“Seek to augment our people, not replace them”

Useful heuristics

- ML applied to business strategy can be **staggeringly** more effective than micro projects
- Low Hanging Fruit. Know when to stop. 80% impact in 5% of time
- **Understandable models are key** (human + ML beats either one)
- The Right Corporate Structure
- Many projects, each 3-4 prime movers, AI / ML does heavy lifting

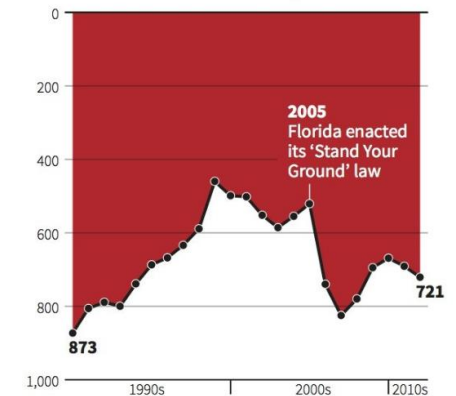
Useful heuristics

- Demystify. Get your people playing with new tools.
- Collaborate & share. Make friends via your cool tools.
- Going external? **Pay for a framework. Not a solution.**
- Beware of “gatekeepers” . Tech is for everyone, not a select few.
- Everyone has same info (where possible)
- Truth via data. For an example of the opposite...

<http://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2?IR=T>

Gun deaths in Florida

Number of murders committed using firearms



Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

REUTERS

The original graphic was from Reuters and routinely features in lists of the most egregious data blunders

Useful heuristics (contd)

- Always seek to empower your people. Bottom-up vs top down model
- Wrong: “It will take my job away.” Correct: “It will take your *old* job away”
- **Productivity is \$\$\$, not wallclock execution time**
- The ability to rapidly prototype is key. Fail fast etc
- The bottleneck : getting info into people’s hearts & minds faster
- Humans are visual creatures. Vision -> Emotion -> Understanding
- Graph-Network visualisations. Color. A good font is like a nice accent.
- No 3d pie charts ever (misleading perspective tricks)

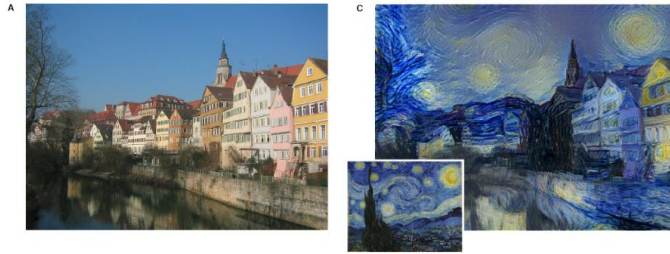
W.R.T. Machine Learning

- PLEASE try your own / build your own wherever possible
- DON'T let anyone tell you what is Easy, or Hard, or Impossible.
- But “Deep Learning is Easy, Try Something Harder Instead”*
- (Useful) ML is : **Convex Optimization**, Linear Algebra, Visualisation.
- 3 Courses, free as air : Andrew Ng’s ML, Geoff Hinton’s NN, Boyd’s Convex Optimization.
- Gamify. Try Kaggle! set up an internal problemsolving leaderboard.
- Matlab is great for all of the above. Mathworks is very user-engaged
- Some pet peeves! distribution/publication is a pain...

* <http://www.inference.vc/deep-learning-is-easy/> Ferenc Huszar

Miniproject 1 : Artistic style transfer

- 25 Aug 15: "A Neural Algorithm of Artistic Style" Gatys, Ecker, Bethge



- Around the same time the bizarre DeepDream images came out
- Lots of development in this space. E.g Fake Rembrandt
- DNNs. Still very slow (**hours**). Visual quality ... Variable
- Low Hanging Fruit- can we get 80% of the impact in 5% of the time?

Do you REALLY need a DNN for this problem? Probably not if you just want a subtle texture transfer..

Algorithm (quick, definitely can be improved)

Given Style image S and Target Image T , we want $U = \text{“}T \text{ in the style of } S\text{”}$

- 1) qtdecomp on T to get tiles $t_1 \dots t_n$
- 2) For each tile t_i , normxcorr2 to find best match s_i in S
- 3) imhistmatch (color-correct) s_i to t_i , paste into U

At this point, it's already pretty good ... But those seams are annoying. Intelligently removing the seams was the most tedious part-

- 4) For each color plane, get gradient around the seams.
- 5) Blur *along* the gradient if the gradient is small . Blur *orthogonal* to the gradient if the gradient is large.

Example

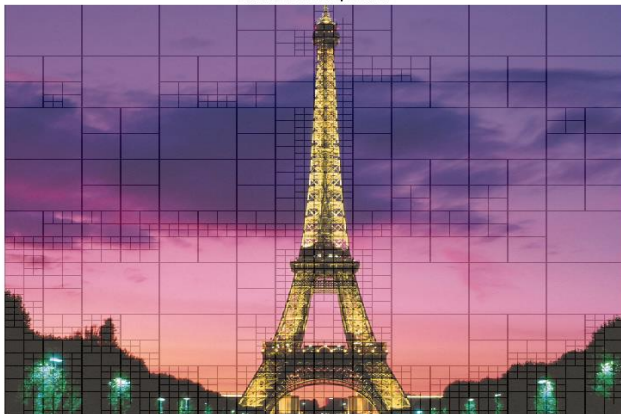
Style source:



Applied to:



Quadtree decomposition

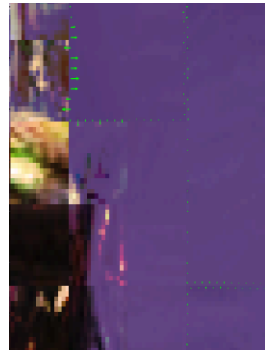


Raw result



Original size 900x1200, resized to 2048x2048 for the quadtree step.

Final Result : (note the subtle brushstrokes!)



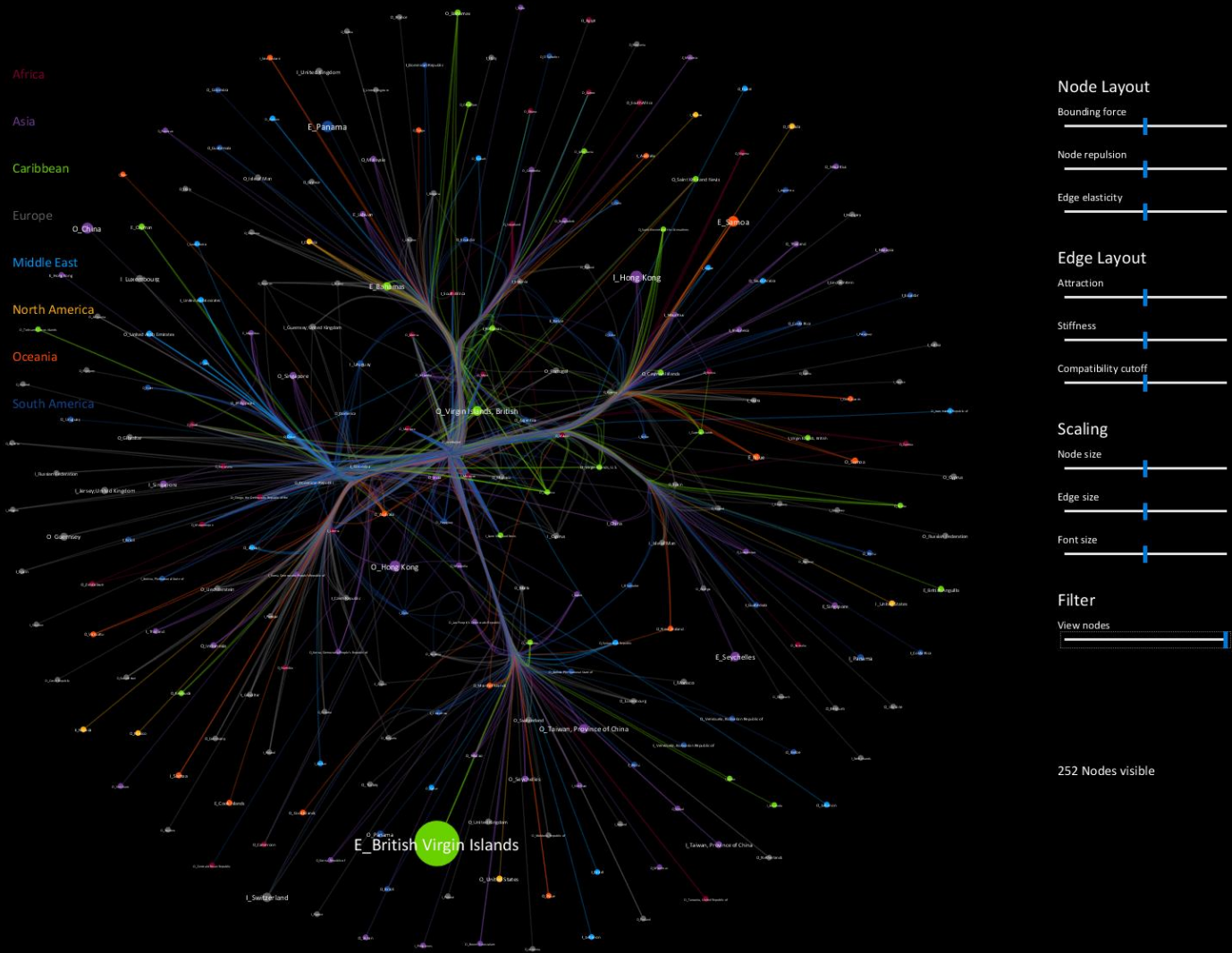
Takes about 30sec in total on an i7 / GTX980, 8 min on a laptop with no GPU. Not bad.

Miniproject 2 :The Panama Papers

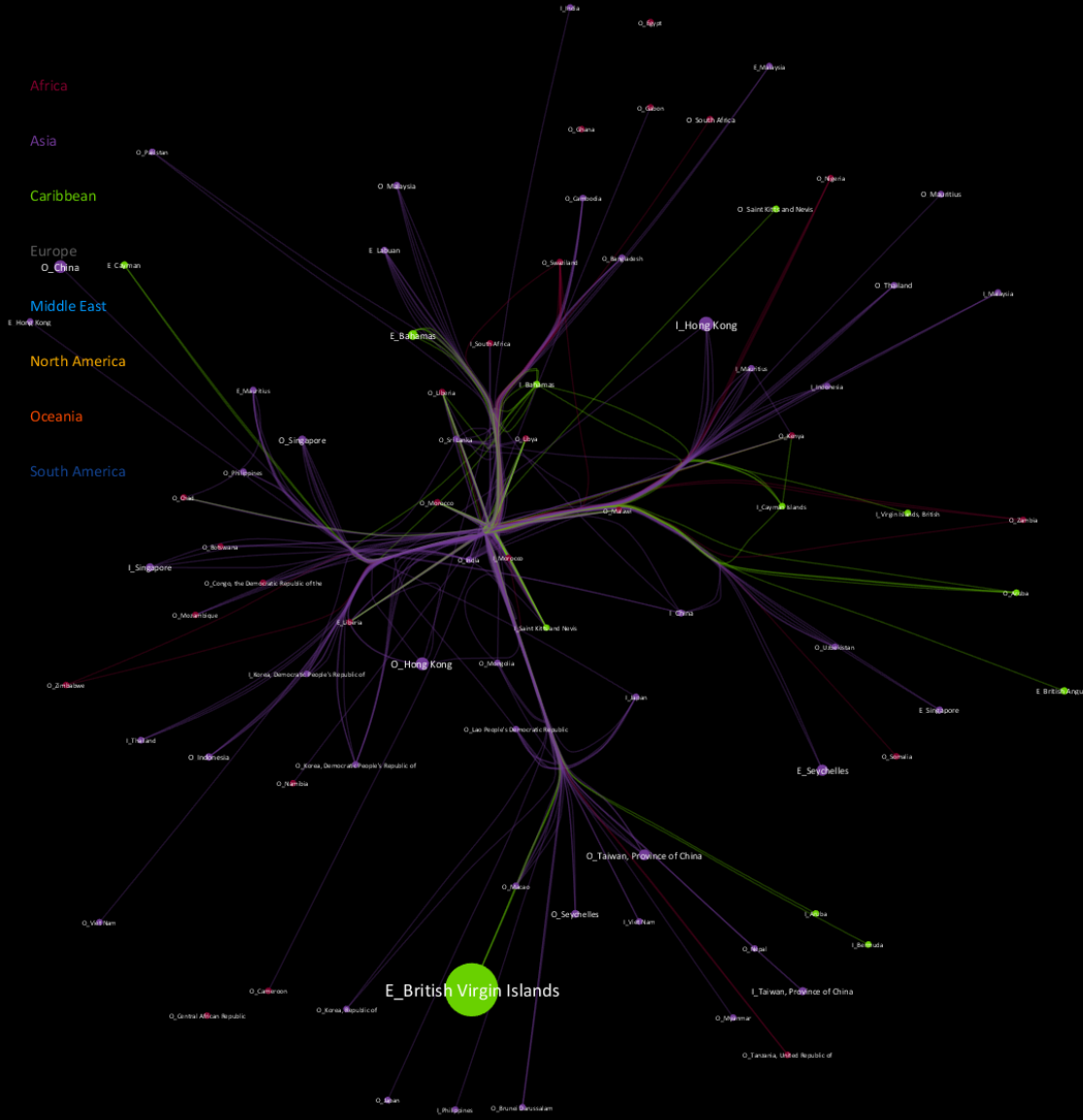
- <https://offshoreleaks.icij.org/pages/database> *ICIJ DISCLAIMER*

“There are legitimate uses for offshore companies and trusts. We do not intend to suggest or imply that any persons, companies or other entities included in the ICIJ Offshore Leaks Database have broken the law or otherwise acted improperly. Many people and entities have the same or similar names. We suggest you confirm the identities of any individuals or entities located in the database based on addresses or other identifiable information. If you find an error in the database please [get in touch with us](#).”

- 11mm + papers, 200k+ nodes, millions of edges & counting...
- Can we (**automatically**) simplify, visualise, and get some insight?
- Autoheuristic chooses Type, Country, Continent as best aggregation fields
- Colors, layouts, etc all automatic.
- Let's see what the network looks like (no names!)



Who knew Matlab could look so good! Thank you Yair Altman , Jenny Owen, David Sampson!



Node Layout

Bounding force

Node repulsion

Edge elasticity

Edge Layout

Attraction

Stiffness

Compatibility cutoff

Scaling

Node size

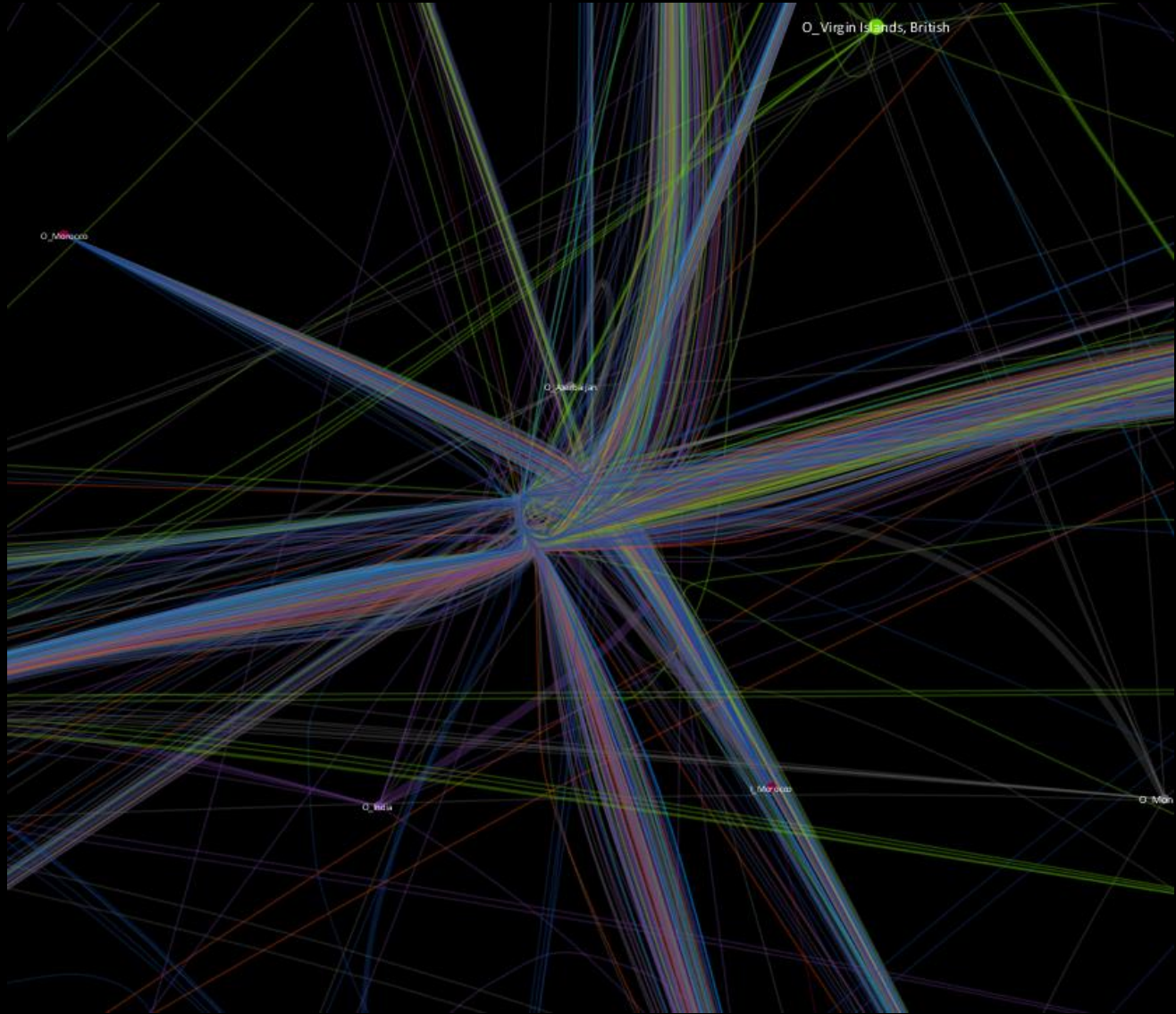
Edge size

Font size

Filter

View nodes

82 Nodes visible



O_Virgin Islands, British

O_Morocco

O_Saudi Arabia

O_Bahrain

O_Morocco

O_Milan

Hypothesizing a hidden, unknown node ...

Why I am excited to be alive today (philosophical digression)

- VR.
- Human-AI hybridization / Transhumanism / we are already hybrids
- Immortality ? ... Assuming we don't **** things up
- Taking our place in the universe. The next 1000 years.
- “Earth is the cradle of humanity. But we cannot live in a cradle forever.”
- Your ideas here...