The Explainability of AI

You Never Let Me Know What You're Thinking:
Why we need AI transparency and how to evaluate Deep Learning Models

Presented by:
Andrea Chan & Mark Donaldson

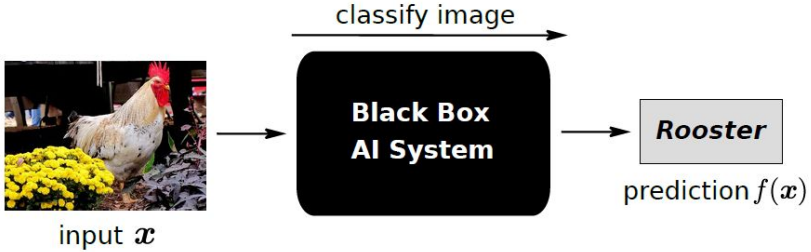
August 21, 2018

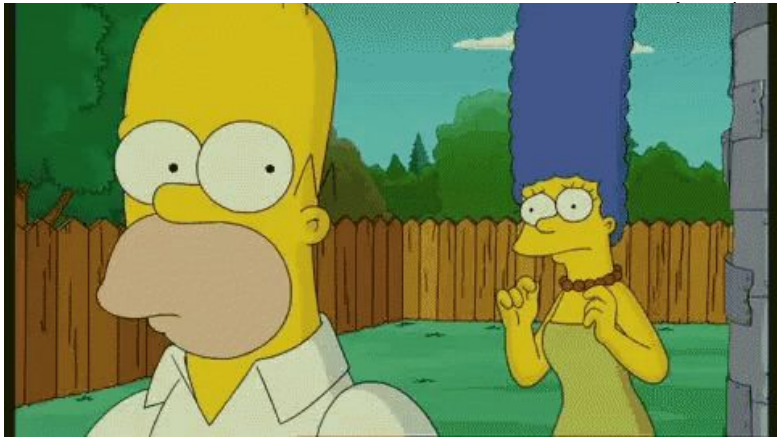
Time to expand some brains

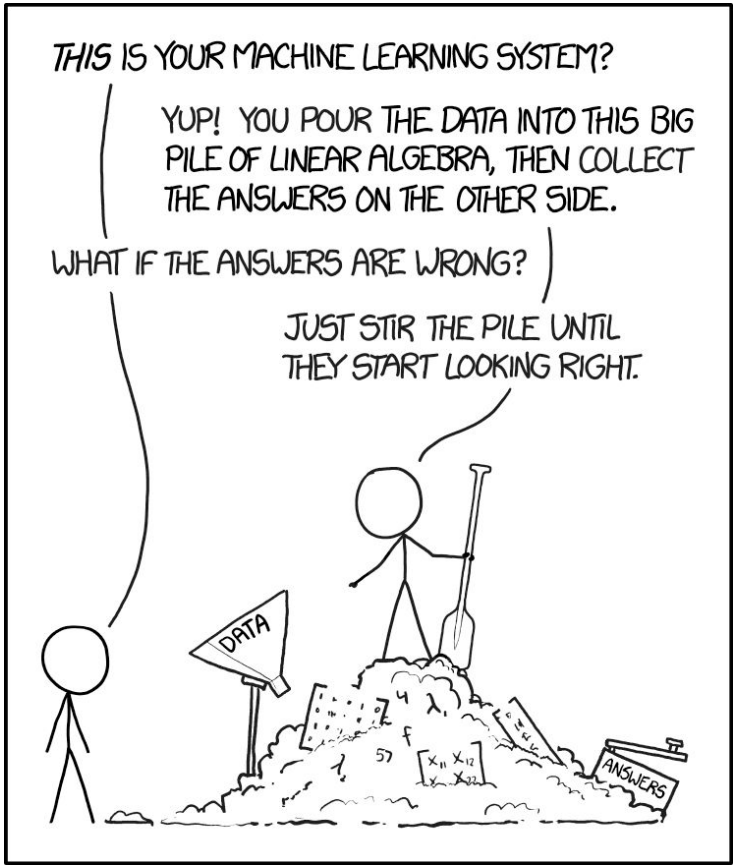
The Black Box Model

Deep Learning

- Nested, non-linear structure
- Not transparent and not easy to assess the process in which an algorithm processes input data
- Many hidden layers







Source: https://xkcd.com/1838/

AI’s role in… well… almost everything

- Getting new video recommendations on Youtube
- Finding discounts and deals on hotels for a holiday trip
- Applying and getting approved for a line of credit
- Your Facebook news feed
- And so much more...

Talking Tech with Elon Musk!
Marques Brownlee
Recommended for you
New

The New York Times

Fake Russian Facebook Accounts Bought $100,000 in Political Ads

By Scott Shane and Vindu Goel

Sept. 6, 2017

733

阅读简体中文版 · 閱讀繁體中文版

Matt Rourke/Associated Press

Providing new evidence of Russian interference in the 2016 election, Facebook disclosed on Wednesday that it had identified more than $100,000 worth of divisive ads on hot-button issues purchased by a shadowy Russian company linked to the Kremlin.

Why explainability?

- Verification of the system
  - High-stakes and high-risk decision-making would still need to be vetted by an expert
  - E.g. Medical applications: life-or-death decisions
- Improvement of the system
  - How can we improve on something we don't fundamentally understand/grasp?
  - Mismatched objectives or multi-objective trade-offs
- Learning from the system
  - Physicists, chemists and biologists would be interested in identifying hidden laws of nature
- Regulatory and compliance issues
  - AI systems to make decisions on legal individual rights
  - Arguably, even moral obligations beyond what is currently legal

Barriers to transparency

Going from data to information requires transparency, and researcher Burrell distinguishes 3 barriers:

1. **Intentional concealment** by corporations/institutions, where decision making procedures are kept from public scrutiny
2. **Gaps in technical literacy** - even if provided code, most people may not have the background to understand it
3. **Limited human capacity** to understand the scope of what machines are capable of. *"Mismatch between the mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of interpretation"*.

If #1 is solved by legislation, and #2 presumably you can have someone qualified to interpret, then you have #3: **it stands to reason that an algorithm can only be explained if the trained model can be articulated and understood by a human.**

Interpretability

Proposed by Zachary Lipton in *The Mythos of Model Interpretability*

Desiderata:

- Trust
- Causality
- Transferability
- Informativeness

General Data Protection Regulation

- GDPR applies to any company processing EU residents' personal data
- Maximum penalties of 20 million Euro or 4% of global revenue, whichever is greater

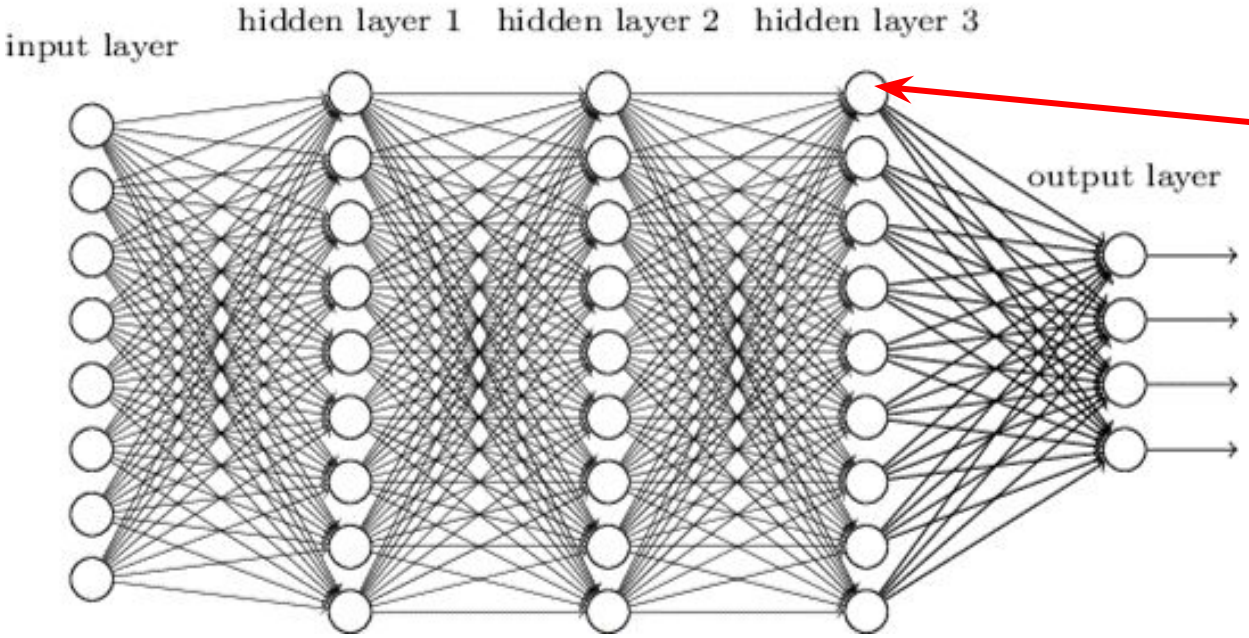
ARTICLE 13 & 14

(f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject

ARTICLE 22

The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

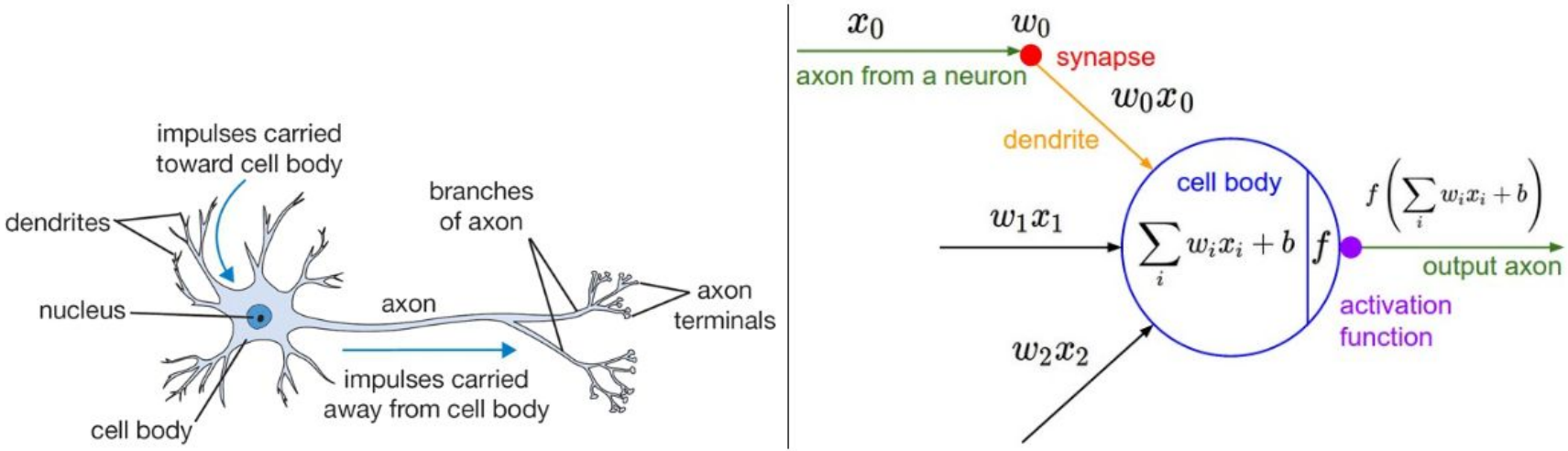
Neural Network Topology: DNN



Layers are composed of neurons

High Level Neuron & Mathematical Model

Biological and mathematical model of a DNN neuron:



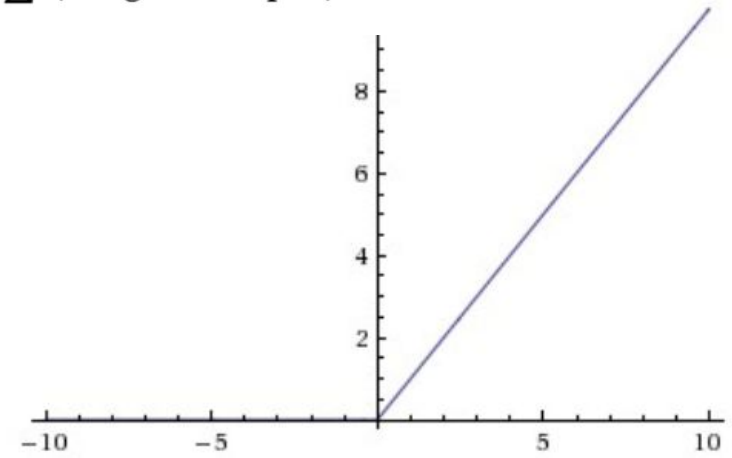
A cartoon drawing of a biological neuron (left) and its mathematical model (right).

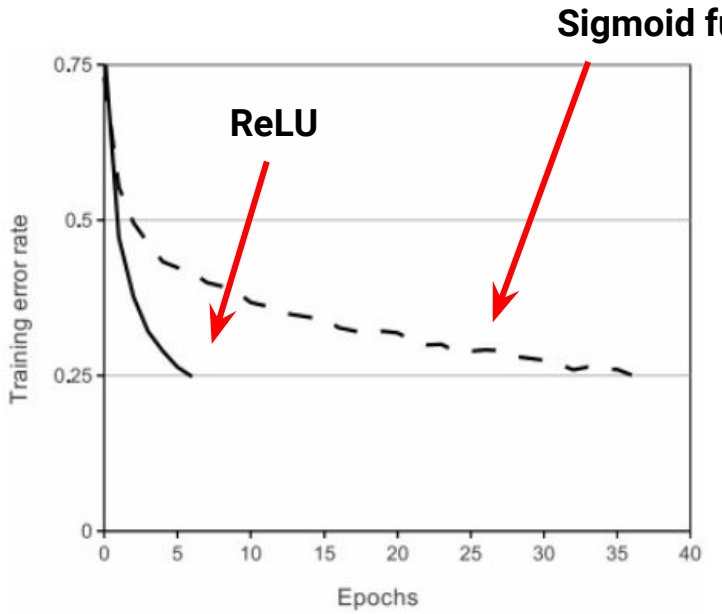
Activation Function Makes AI Learn

The use of the ReLU (Rectified Linear Units) Activation Function:

ReLU Activation Function (Eq. 1)

$$Y = \sum (weight * input) + bias$$





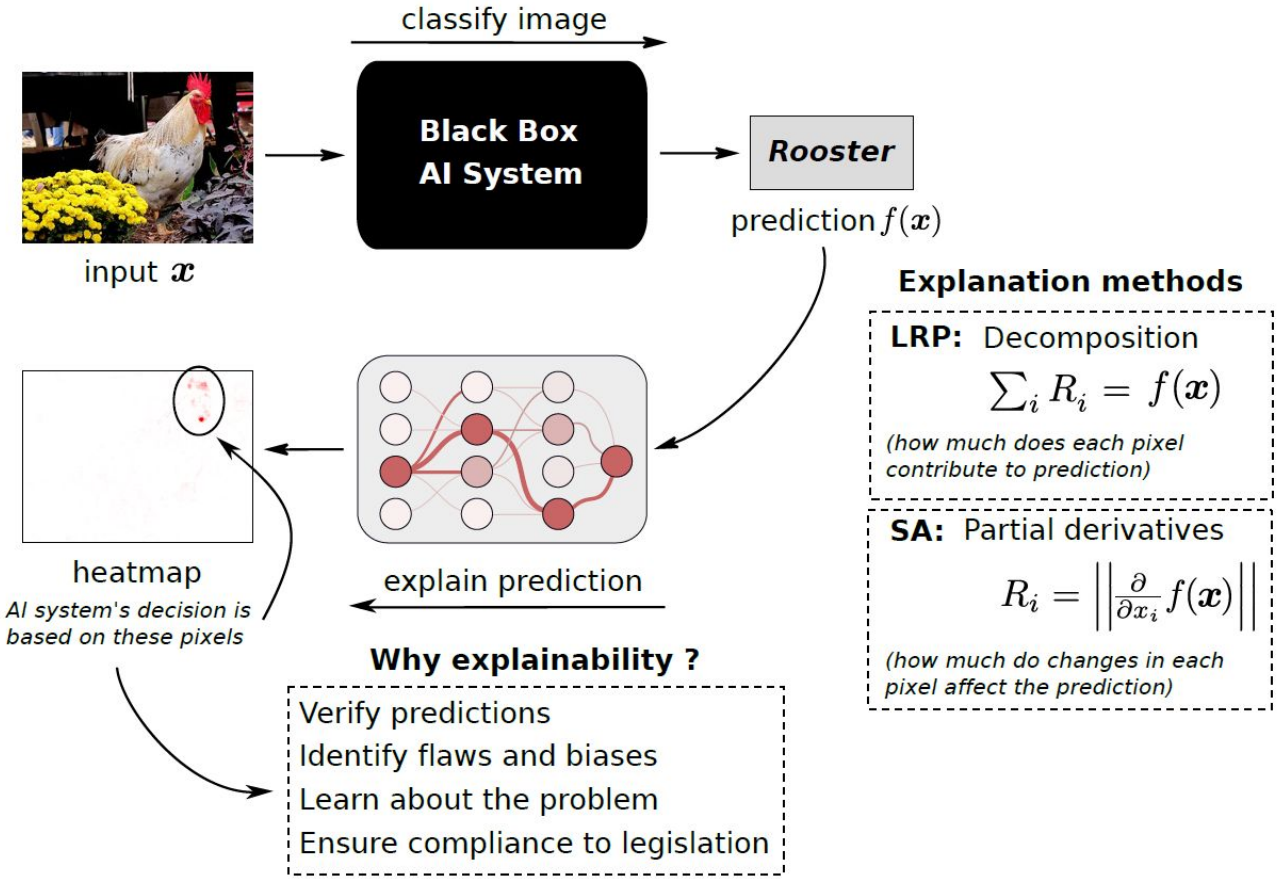


Figure 1 in paper

Evaluating Black Box Models: 2 Types

1. Sensitivity Analysis (SA)

2. Layer-wise Relevance Propagation (LRP)

There are others, but those won’t be covered in today’s discussion.

Sensitivity Analysis

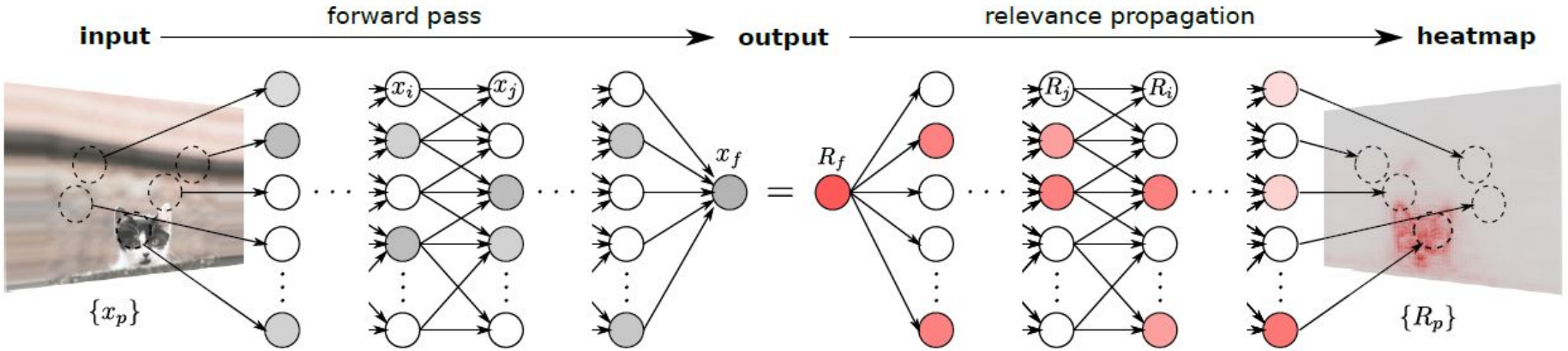
Sensitivity Analysis (SA), mathematically, is to quantify the importance of each input variable

Eq. 2

$$R_i = \left|\left| \frac{\partial}{\partial x_i} f(\mathbf{x}) \right|\right|.$$

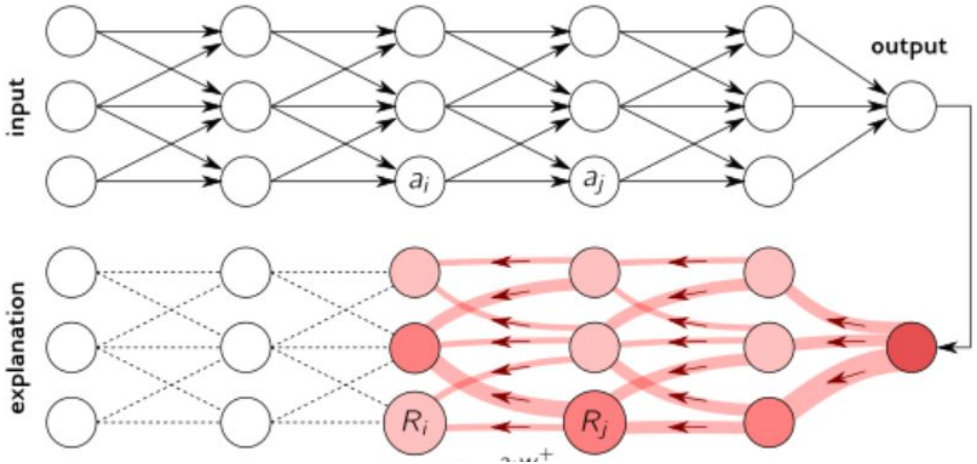
Layer-wise Relevance Propagation

An even *deeper* look inside a DNN….. So, you're a cat person…
Let's verify more clearly using Layer-wise Relevance Propagation (LRP)



Layer-wise Relevance Propagation

An even *deeper* look inside a DNN…..
Are you sure you saw what you said you saw? Let's check! Layer-wise Relevance Propagation (LRP)



Eq. 3

$$R_j = \sum_k \frac{x_j w_{jk}}{\sum_j x_j w_{jk} + \epsilon} R_k$$

Eq. 4

$$R_i = \sum_j \frac{x_i w_{ij}^{+}}{\sum_i x_i w_{ij}^{+}} R_j$$

Layer-wise Relevance Propagation

Inside a DNN: The “alpha-beta” alternative redistributes rule and what the math below means for LRP and “deep Taylor decomposition” of a DNN:

$$R_j = \sum_k \left(\alpha \cdot \frac{(x_j w_{jk})^+}{\sum_j (x_j w_{jk})^+} - \beta \cdot \frac{(x_j w_{jk})^-}{\sum_j (x_j w_{jk})^-} \right) R_k$$

Eq. 5

Heatmap Comparison

All together now…. Aaaaaaawwwww. So cute!

Image

Conclusion

What did we come away with?:

- LRP vs SA is better at extracting domain knowledge information of the data set input
- LRP allows us to more clearly measure and verify the DNN’s model validation
- Use of LRP allows us to create transparent black box machine learning models

Conclusion

Benefits and uses of LRP:

- Black Box transparency can be used for forensic investigation and analysis
- Transparency should be included into every DNN system that is used in critical outcome domains like:
  - self-driving cars
  - health sciences
  - consumer finance credit applications
  - bias outcome explanations and many more areas that are critical to humans, the environment, the Earth and beyond

References and related sources for this presentation:

→ bit.ly/Explainability

Toronto Deep Learning Series:

→ bit.ly/TDLSgroup

Thank you!

ANDREA CHAN

 hi@andreachan.com

 @helloitsdrea

 linkedin.com/in/andreahmchan/

MARK DONALDSON

 mark.donaldson@ryerson.ca

 @markdheilong

 linkedin.com/in/markdonaldson888/