ADVERSARIAL MACHINE LEARNING

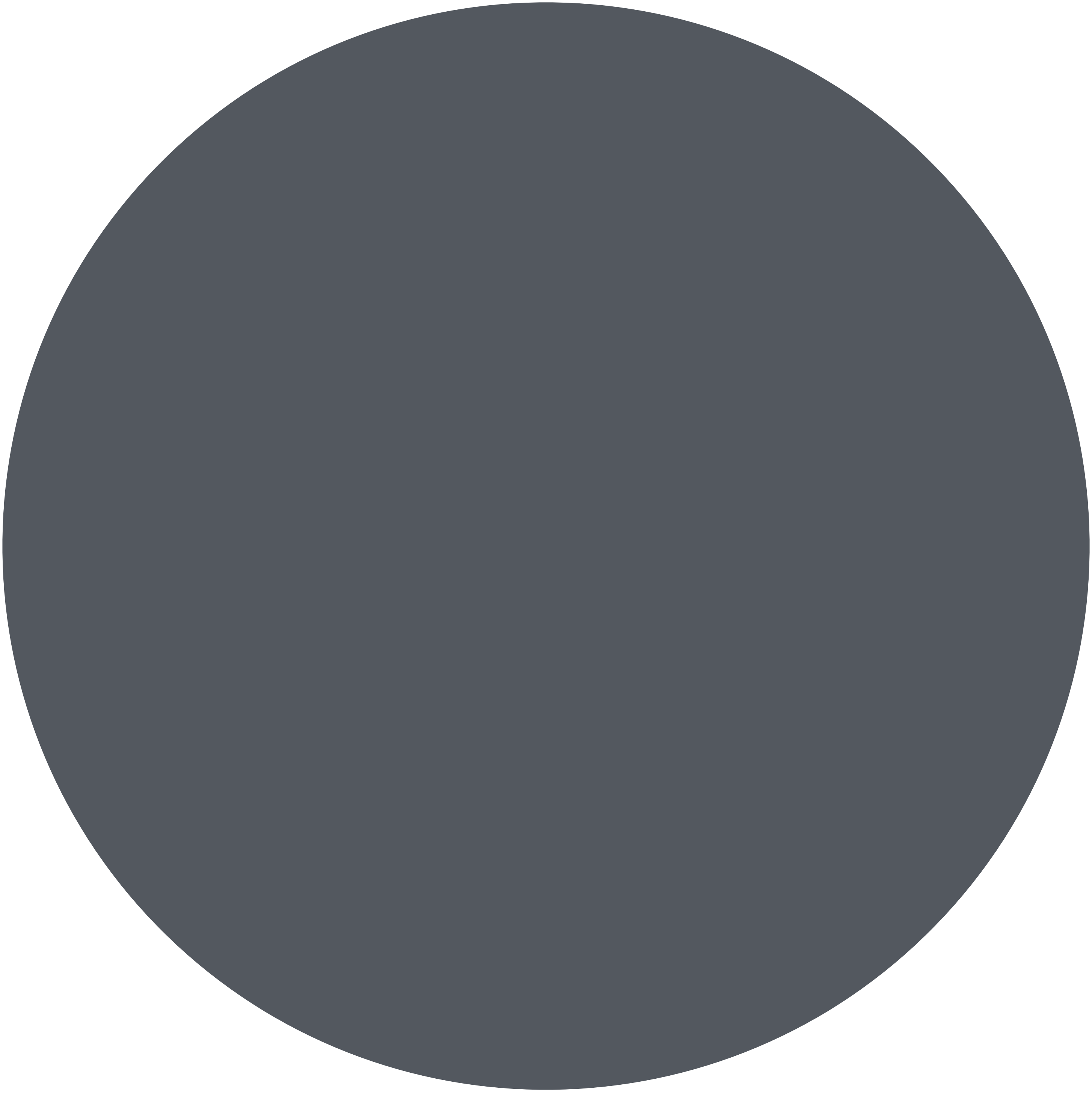
TIMELINE

11

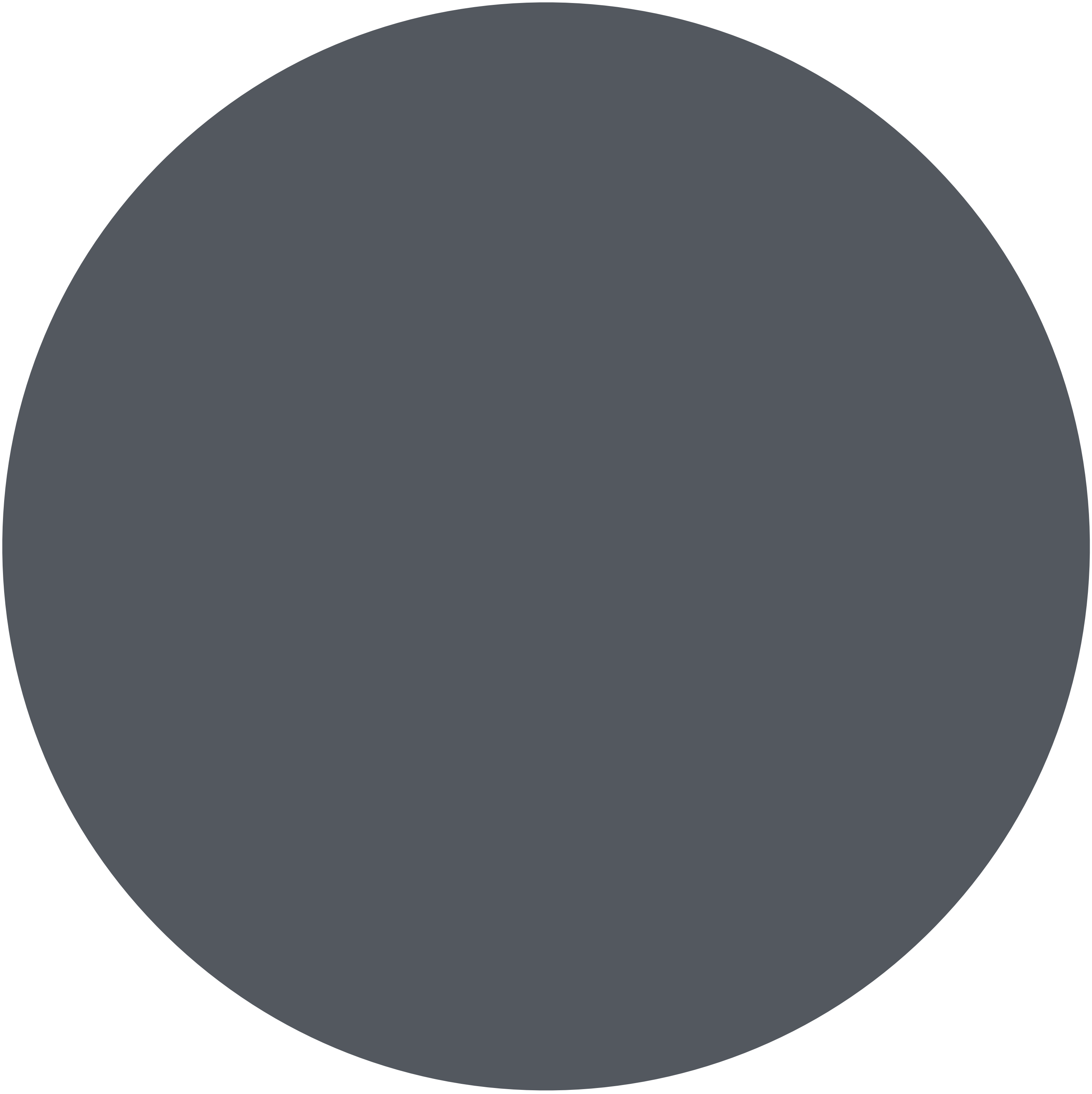
2004

2013

2014



2015



2016

FOOLING THE SPAM FILTER

FIRST ATTEMPT TO FOOL A NEURAL NET

IMPERCEPTIBLE ADVERSARIAL EXAMPLES

2017

ADVERSARIAL TRAINING AS DEFENCE

DEFENSIVE DISTILLATION

ADVERSARIAL EXAMPLES IN PHYSICAL WORLD & BLACK BOX ATTACKS

LARGE SCALE ADVERSARIAL TRAINING & ENSEMBLE APPROACHES

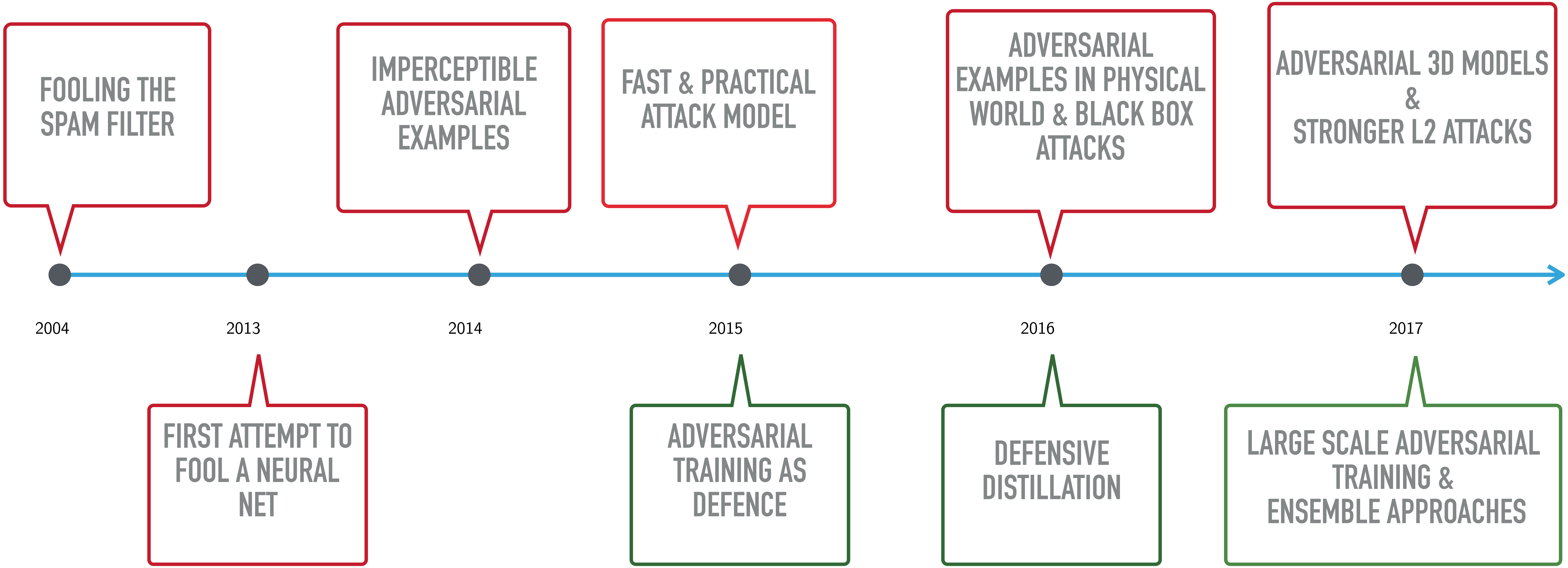
ADVERSARIAL 3D MODELS
&
STRONGER L2 ATTACKS

FAST & PRACTICAL ATTACK MODEL

HISTORY OF ADVERSARIAL MACHINE LEARNING

- Adversarial Classification by Dalvi et al. 2004 (Fooled Spam Filter)
- Evasion Attacks Against Machine Learning at Test Time by B. Biggio et al. (Fooled Neural Net)
- Intriguing properties of neural networks by C. Szegedy et al. 2014 (Imperceptible Adversarial Examples)
- Explaining and Harnessing Adversarial Examples by Goodfellow et al. 2015 (Fast Attack Model)
- Adversarial Examples in Physical World by Kurakin et al. 2016 (Demonstrates Adversary in Real World)
- Adversarial Machine Learning at Scale by Kurakin et al. 2017 (Adversarial Training)
- Practical Black Box Attacks by Papernot et al. 2016 (Practical Black Box Attacks)
- Evaluating the robustness of neural networks by Wagner et. al. 2017 (Strong L2 Attack)
- Ensemble approach for adversarial defence by Papernot et al. 2017 (Ensemble Defence)

TIMELINE



FOOLING THE SPAM FILTER
IMPERCEPTIBLE ADVERSARIAL EXAMPLES
FAST & PRACTICAL ATTACK MODEL
ADVERSARIAL EXAMPLES IN PHYSICAL WORLD & BLACK BOX ATTACKS
ADVERSARIAL 3D MODELS & STRONGER L2 ATTACKS
2004
2013
2014
2015
2016
2017
FIRST ATTEMPT TO FOOL A NEURAL NET
ADVERSARIAL TRAINING AS DEFENCE
DEFENSIVE DISTILLATION
LARGE SCALE ADVERSARIAL TRAINING & ENSEMBLE APPROACHES