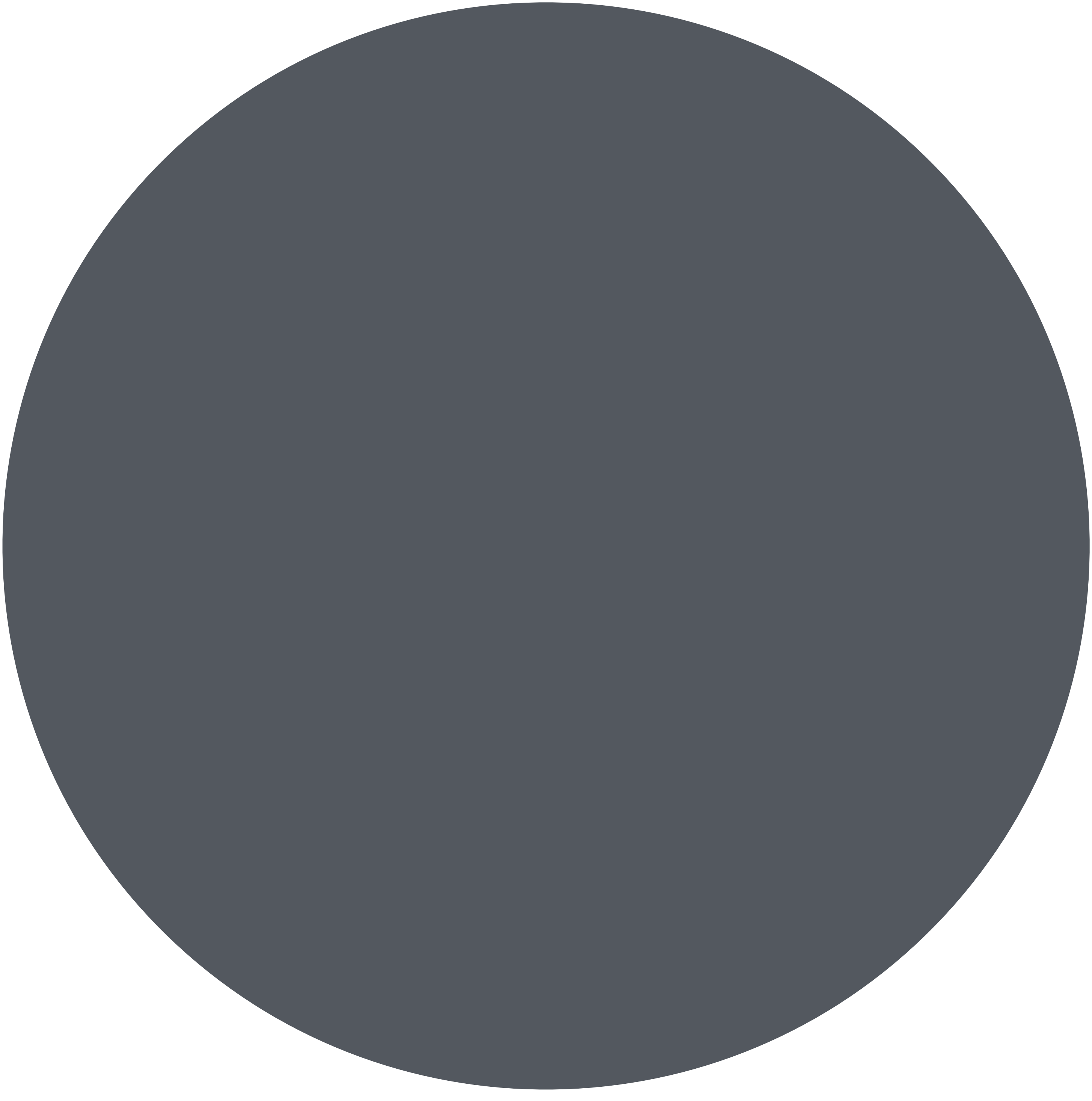
ADVERSARIAL MACHINE LEARNING

TIMELINE

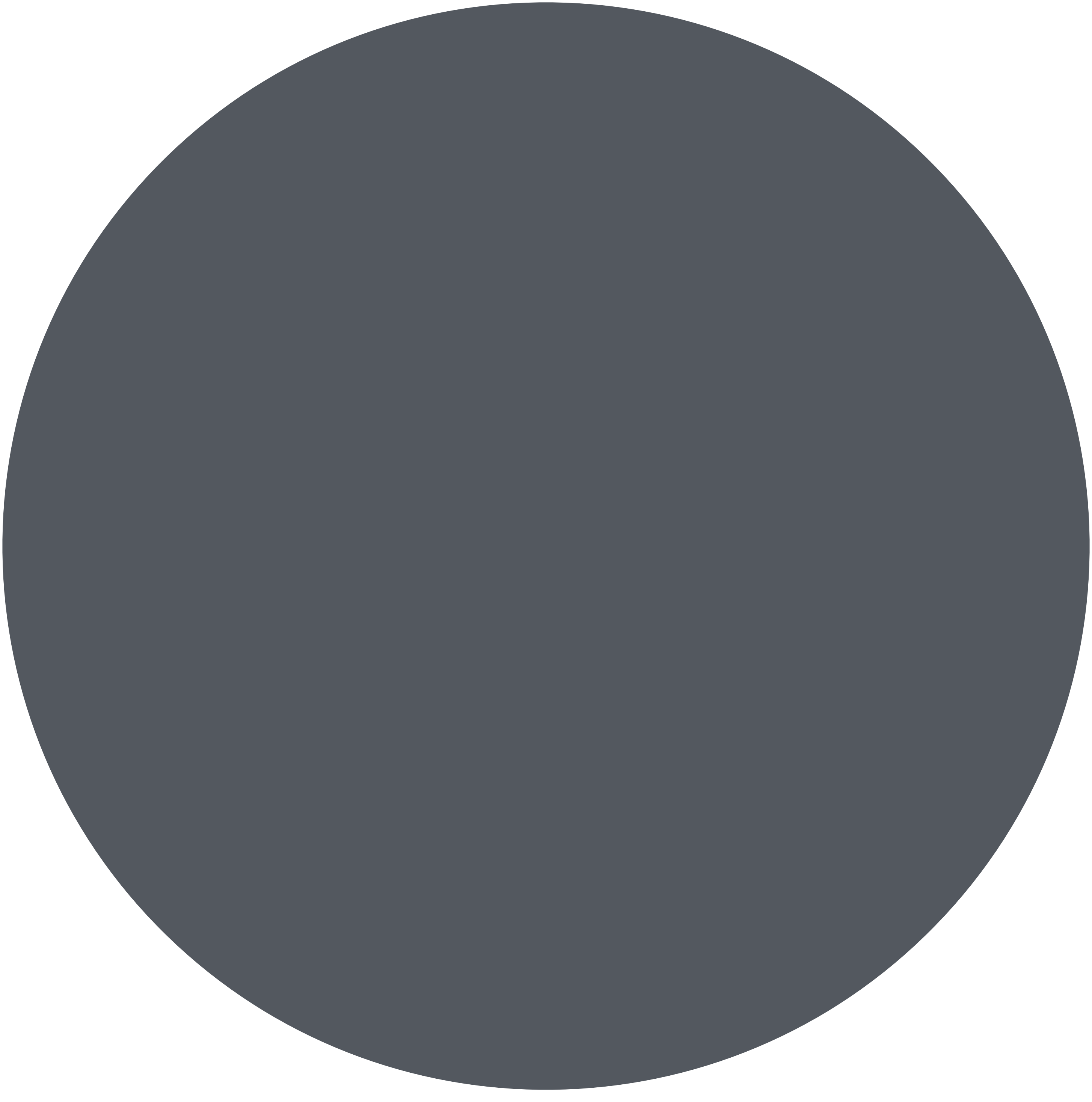
29



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

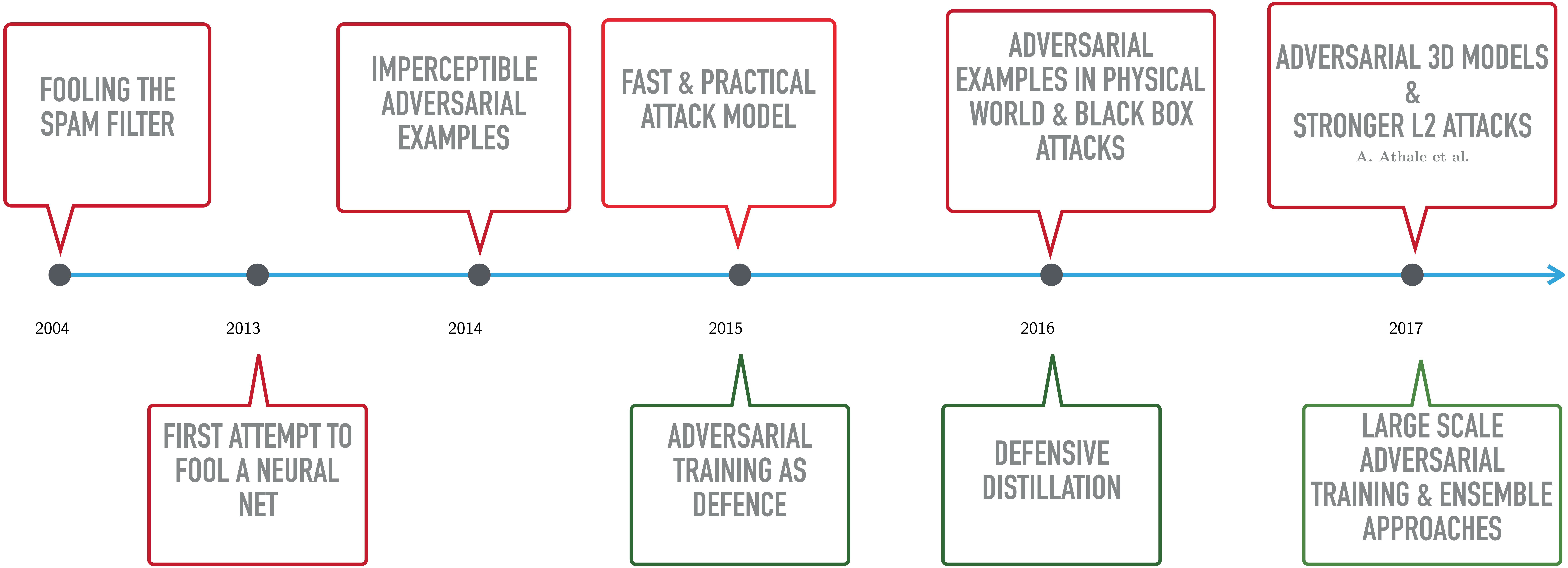
A. Athale et al.

FAST & PRACTICAL ATTACK MODEL

DEFENCE STRATEGIES

- Varies based on goal of adversary
- For multi class targeted or non-targeted attack defence means making classifier robust against adversarial perturbations (e.g. Self Driving Cars)
- For anomaly detection etc scenario defence means detecting adversarial examples

TIMELINE



2004

FOOLING THE SPAM FILTER

2013

FIRST ATTEMPT TO FOOL A NEURAL NET

2014

IMPERCEPTIBLE ADVERSARIAL EXAMPLES

2015

FAST & PRACTICAL ATTACK MODEL

ADVERSARIAL TRAINING AS DEFENCE

2016

ADVERSARIAL EXAMPLES IN PHYSICAL WORLD & BLACK BOX ATTACKS

DEFENSIVE DISTILLATION

2017

ADVERSARIAL 3D MODELS & STRONGER L2 ATTACKS

A. Athale et al.

LARGE SCALE ADVERSARIAL TRAINING & ENSEMBLE APPROACHES