

Data-Centric Machine Learning in Natural Hazards Engineering

Nenad Bijelić
March 28th, 2022



Introduction

Introduction



Source: http://www.krugerpark.co.za/africa_lion.html

Introduction



Big data = hangry*

* hangry = hungry + angry

Introduction



Source: <https://www.rd.com/list/cutest-cat-breeds/>



Source: <https://www.pinterest.se/pin/341077371781851528/>

Big data = hangry*

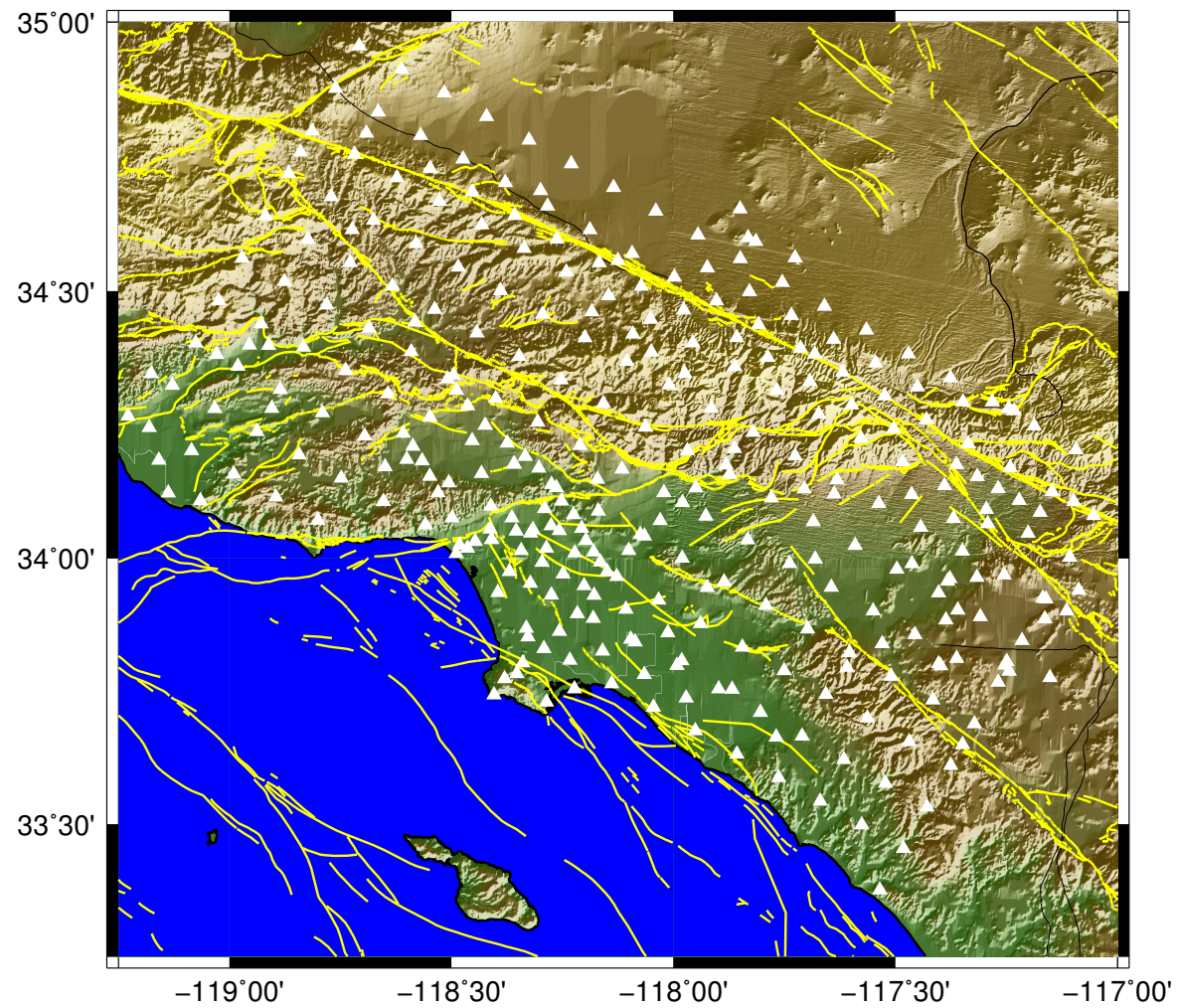
Reduce data requirements
through domain knowledge

soft kitty, warm kitty,
little ball of fur...

* hangry = hungry + angry

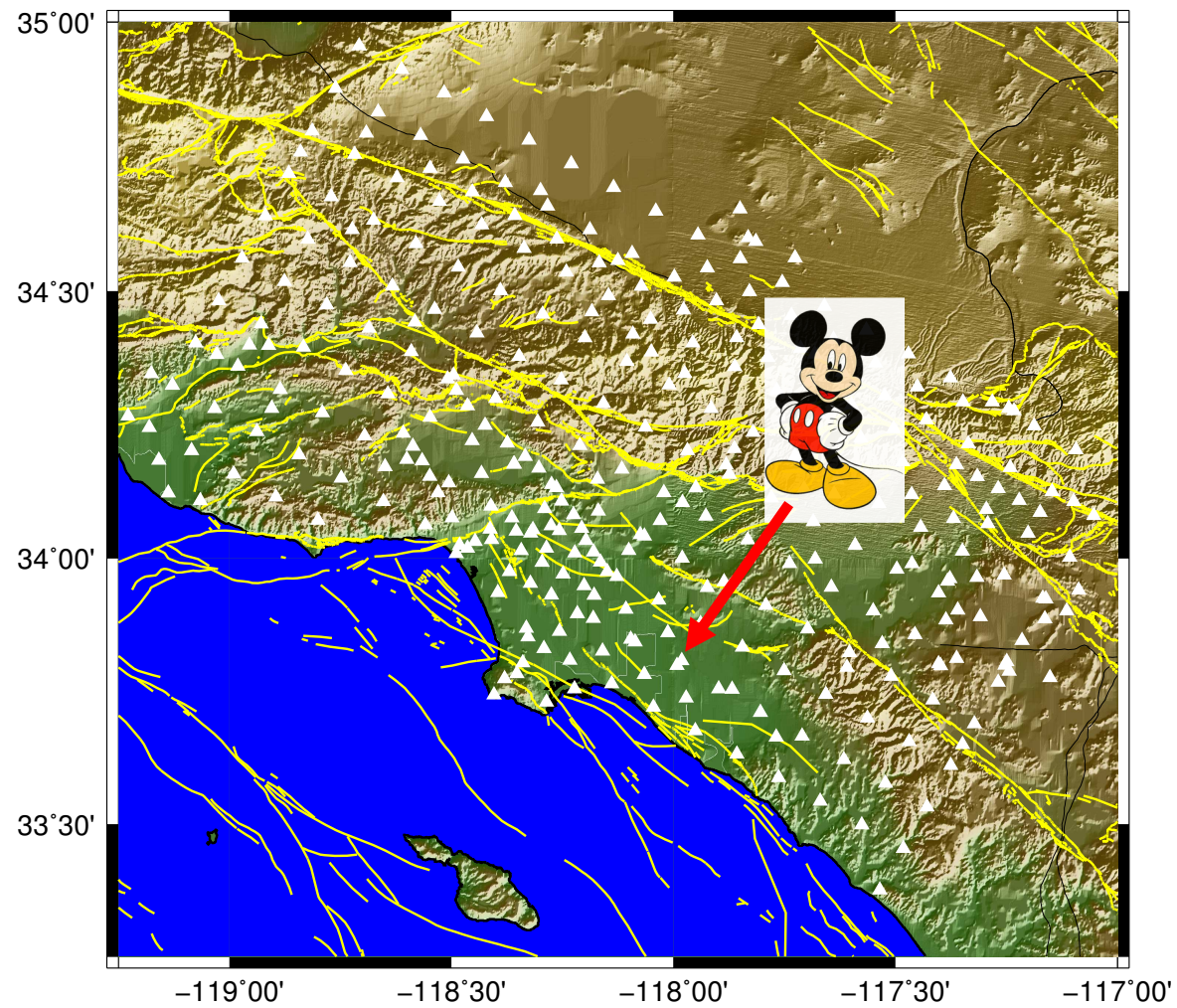
Regional collapse risk due to earthquakes

Southern California



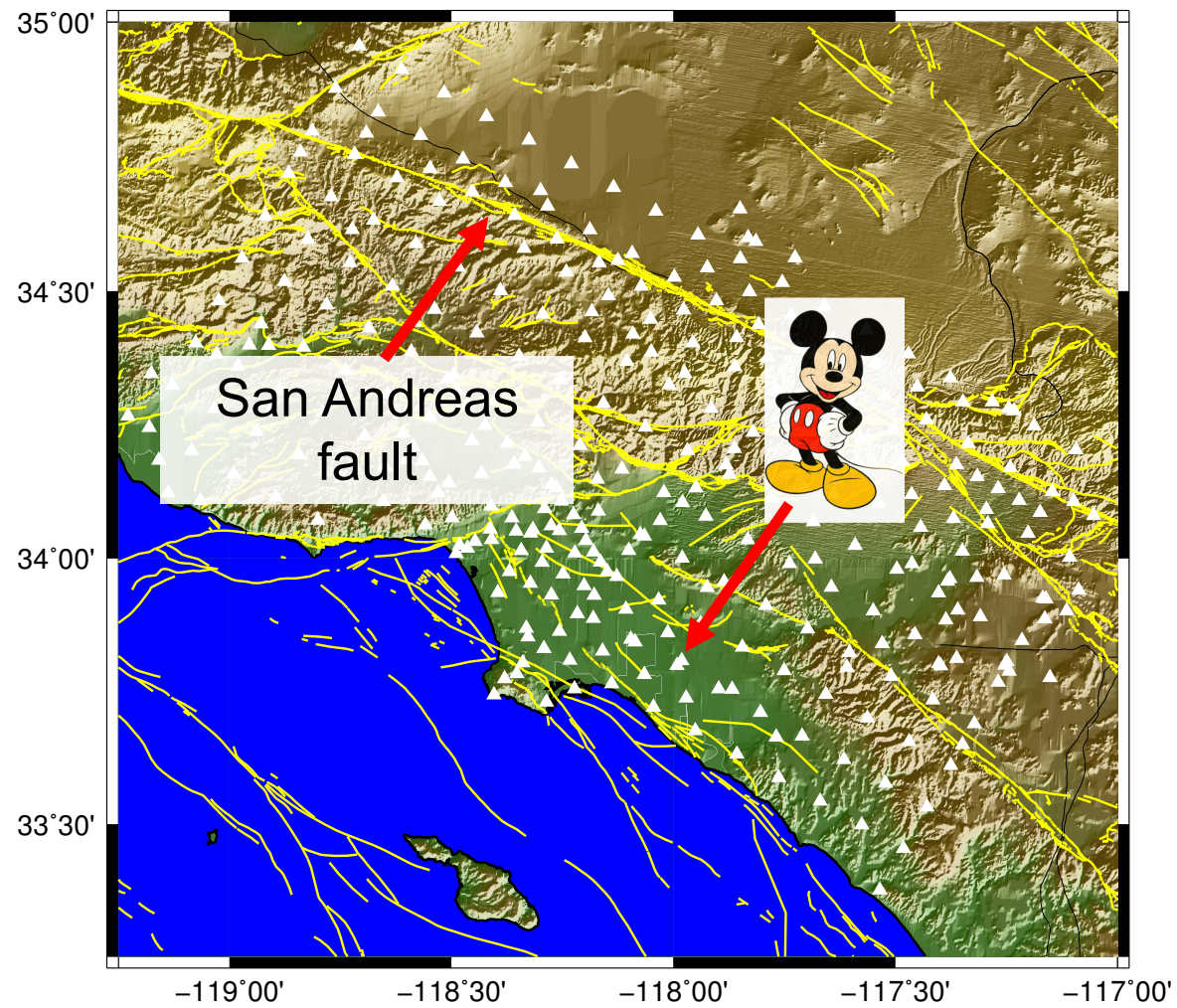
Regional collapse risk due to earthquakes

Southern California



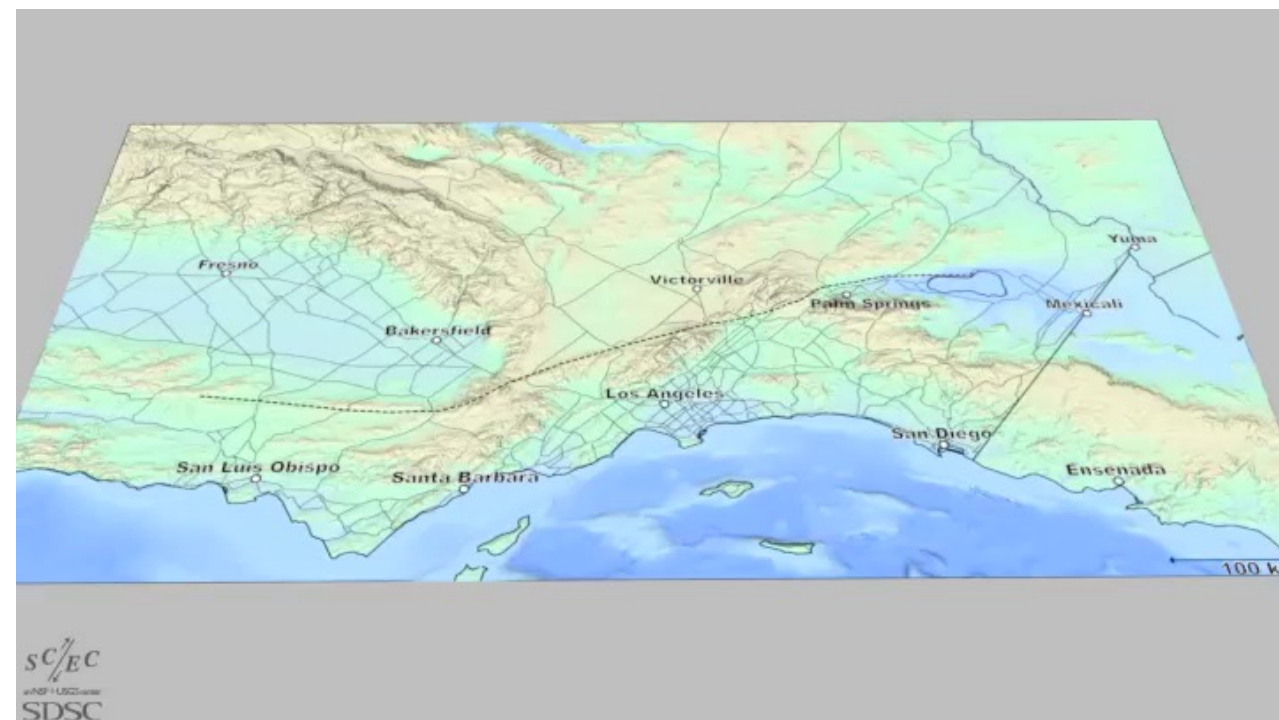
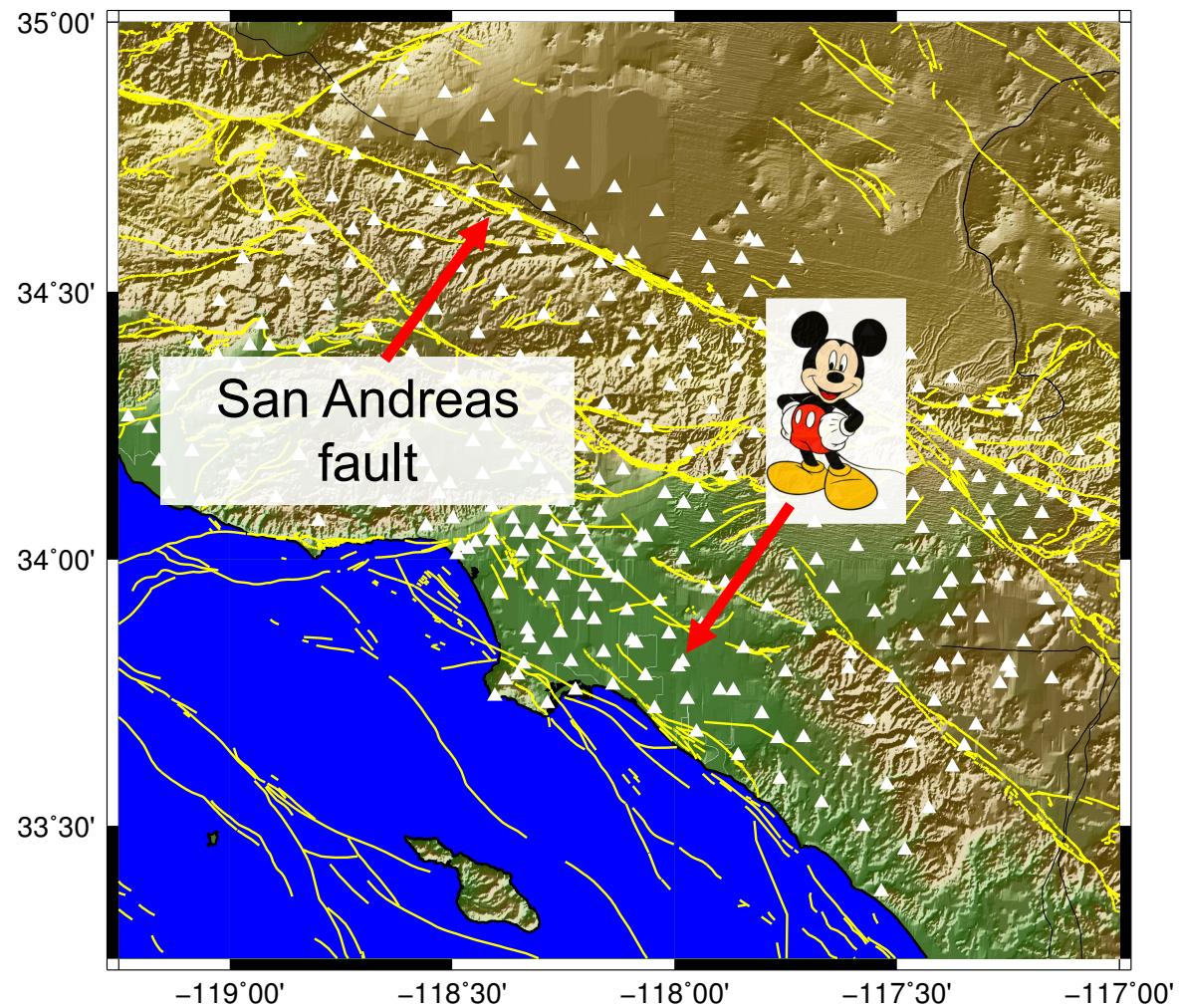
Regional collapse risk due to earthquakes

Southern California



Regional collapse risk due to earthquakes

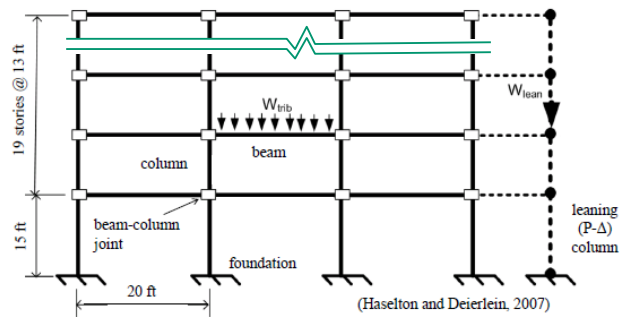
Southern California



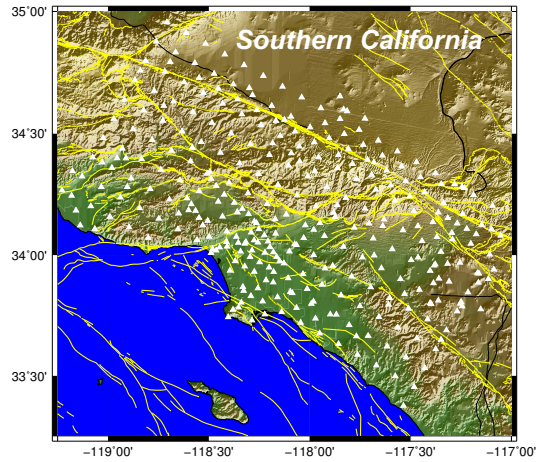
(*animation credit:* Kim Olsen, Yifeng Cui, Amit Chourasia)

Big Data to the rescue?

Collapse risk in the Los Angeles basin

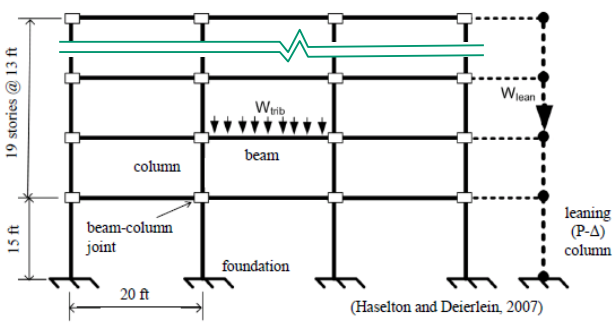


20-story RC moment frame, $T_1 = 2.60s$



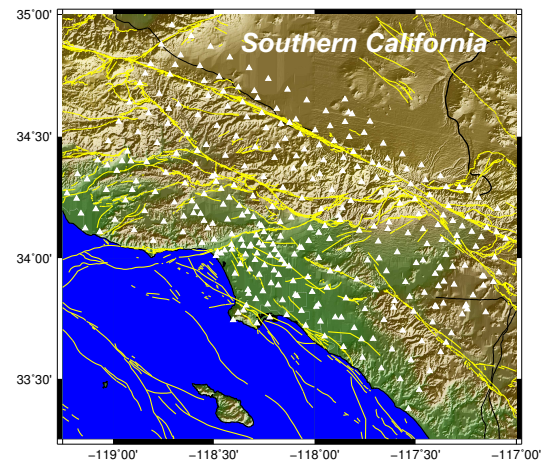
Big Data to the rescue?

Collapse risk in the Los Angeles basin



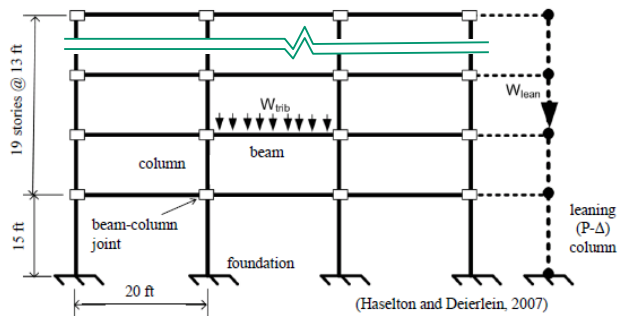
~1,900,000 nonlinear dynamic analyses

20-story RC moment frame, $T_1 = 2.60s$



Big Data to the rescue?

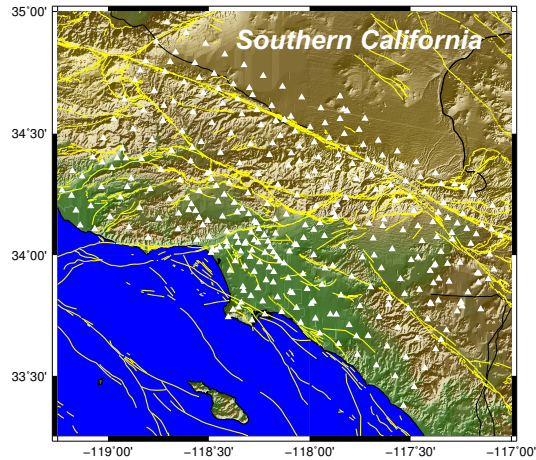
Collapse risk in the Los Angeles basin



20-story RC moment frame, $T_1 = 2.60s$

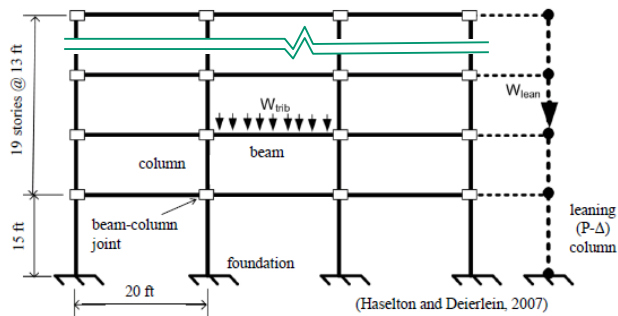
~1,900,000 nonlinear dynamic analyses

support vector machine



Big Data to the rescue?

Collapse risk in the Los Angeles basin



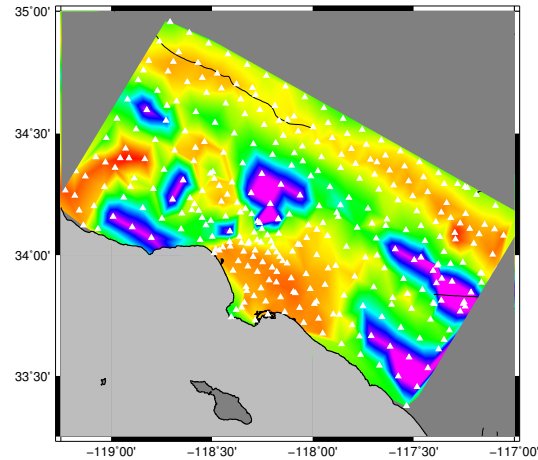
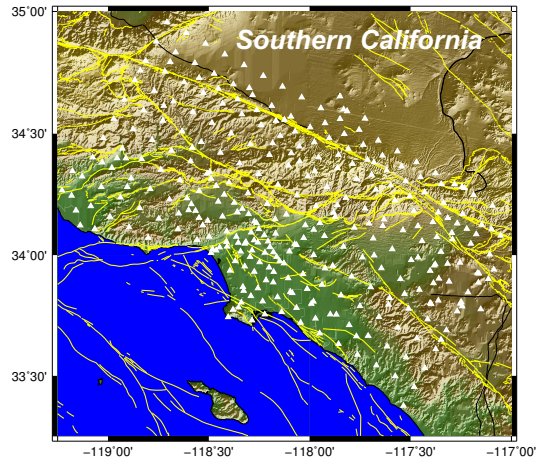
20-story RC moment frame, $T_1 = 2.60s$

~1,900,000 nonlinear dynamic analyses

support vector machine

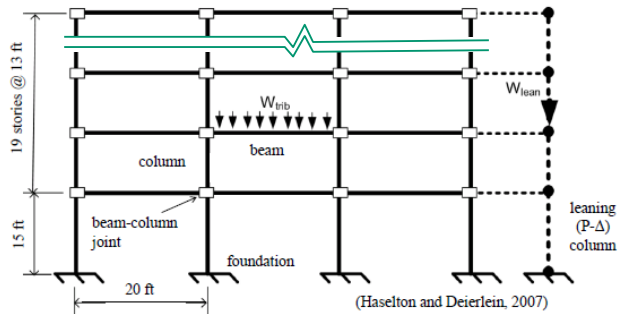
$P(\text{collapse})$ in 50 years

~0%  ~3%



Big Data to the rescue?

Collapse risk in the Los Angeles basin



20-story RC moment frame, $T_1 = 2.60s$

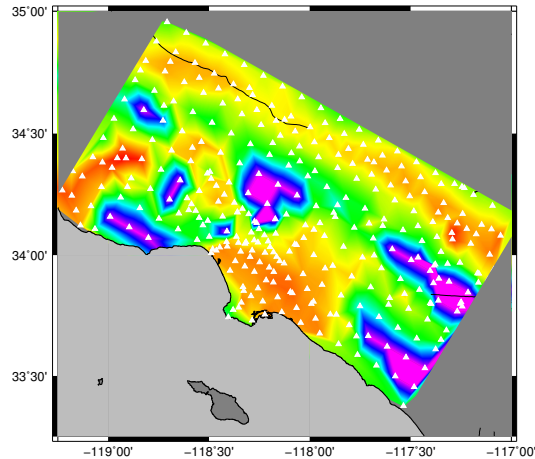
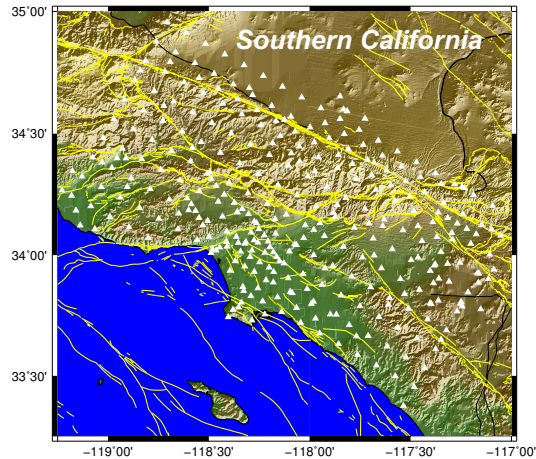
~1,900,000 nonlinear dynamic analyses

analyses

support vector machine

$P(\text{collapse})$ in 50 years

~0%  ~3%



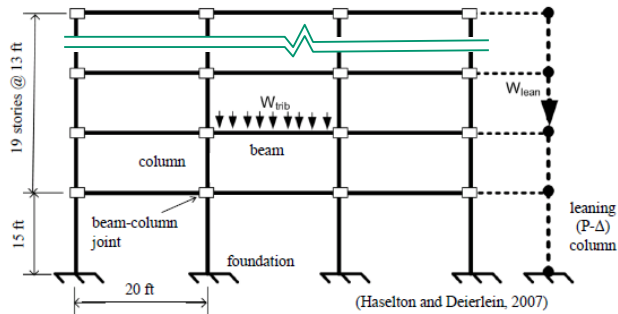
Analysis time:

336 sites with ~500,000 earthquakes per site

Parallel computation (20 node cluster): 5 – 10 years

Big Data to the rescue?

Collapse risk in the Los Angeles basin



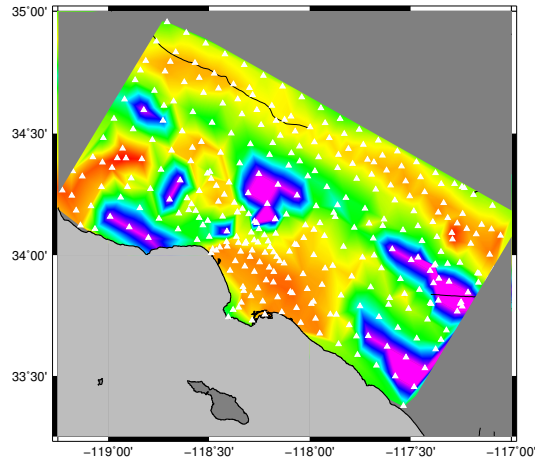
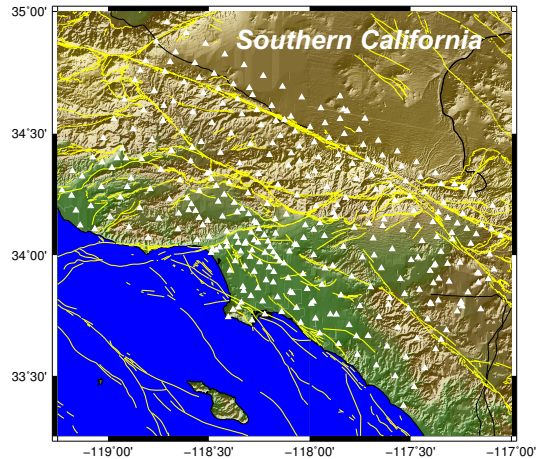
20-story RC moment frame, $T_1 = 2.60s$

~1,900,000 nonlinear dynamic analyses

support vector machine

$P(\text{collapse})$ in 50 years

~0%  ~3%



Analysis time:

336 sites with ~500,000 earthquakes per site

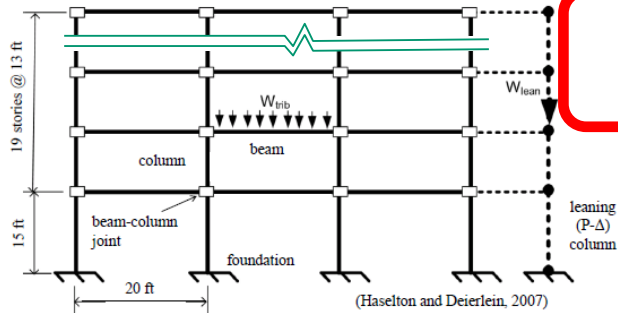
Parallel computation (20 node cluster): 5 – 10 years

ML classifier: 3 minutes on my mac (error ~15 – 30%)

Bijelić, Lin, Deierlein (2019, 2020)

Big Data to the rescue? No – small data and engineering!

Collapse risk in the Los Angeles basin



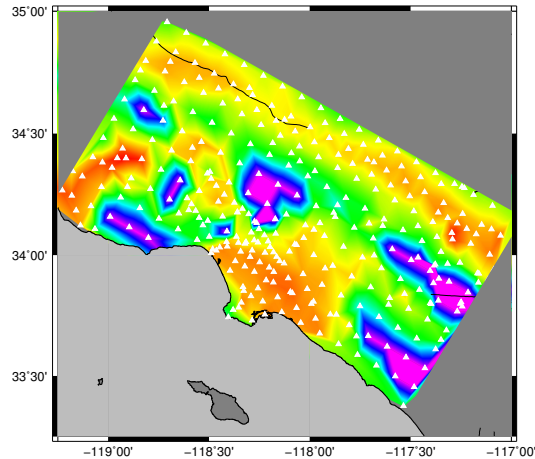
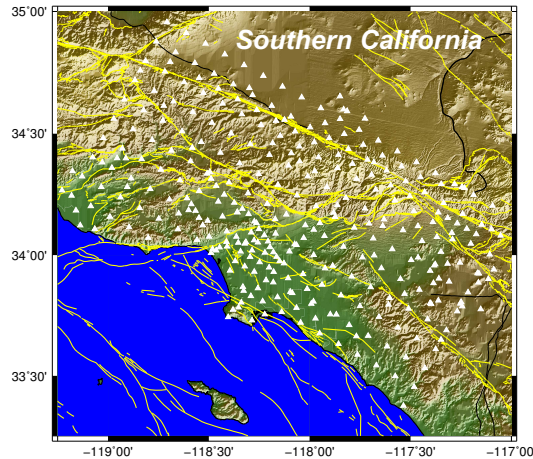
20-story RC moment frame, $T_1 = 2.60s$

~1,900,000 nonlinear dynamic analyses

support vector machine

P(collapse) in 50 years

~0% ~3%



Analysis time:

336 sites with ~500,000 earthquakes per site

Parallel computation (20 node cluster): 5 – 10 years

ML classifier: 3 minutes on my mac (error ~15 – 30%)

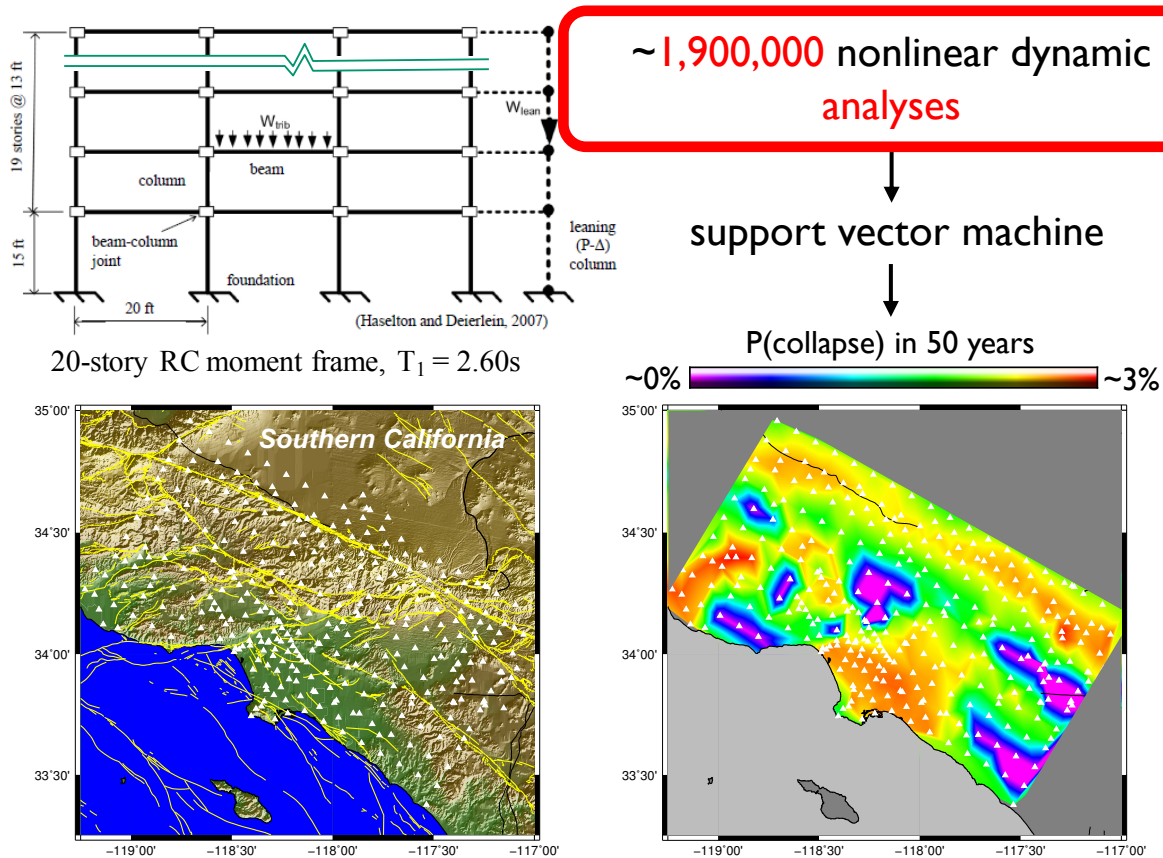
Bijelić, Lin, Deierlein (2019, 2020)

Number of analyses required
by the design code

3, 7 or 11

Big Data to the rescue? No – small data and engineering!

Collapse risk in the Los Angeles basin



Analysis time:

336 sites with ~500,000 earthquakes per site

Parallel computation (20 node cluster): 5 – 10 years

ML classifier: 3 minutes on my mac (error ~15 – 30%)

Bijelić, Lin, Deierlein (2019, 2020)

Data Engineering for the Built Environment

Facts of life:

- (1) we typically do not have big data
- (2) we have or can simulate 'small data'
- (3) we have domain knowledge

Number of analyses required
by the design code

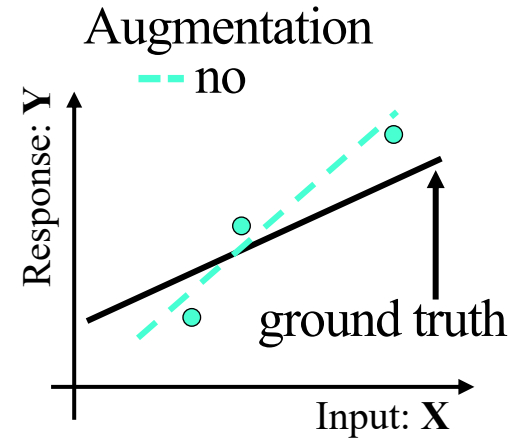
3, 7 or 11

Domain-specific data engineering – seismic collapse risk

‘small data’

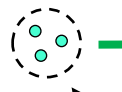


Initial
training data



Domain-specific data engineering – seismic collapse risk

‘small data’



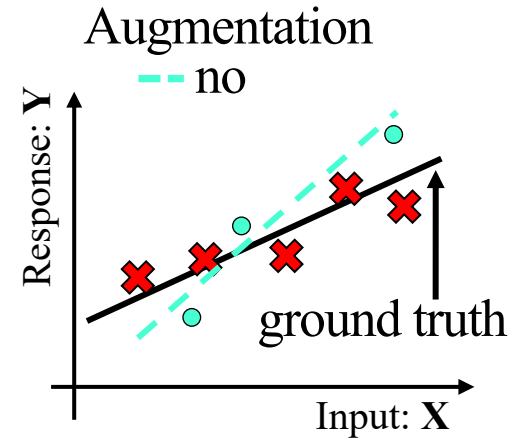
Initial
training data

domain-specific
augmentation

‘good big data’

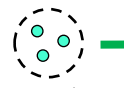


Augmented
training data



Domain-specific data engineering – seismic collapse risk

‘small data’



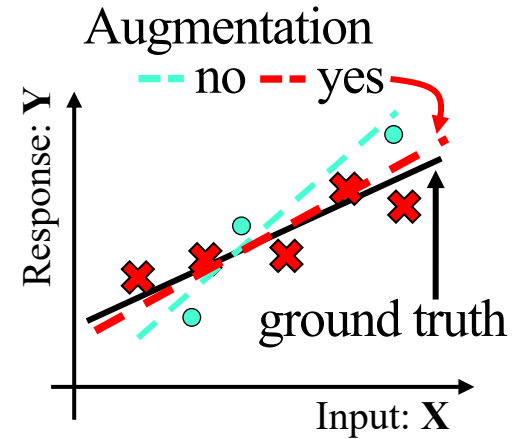
Initial
training data

domain-specific
augmentation

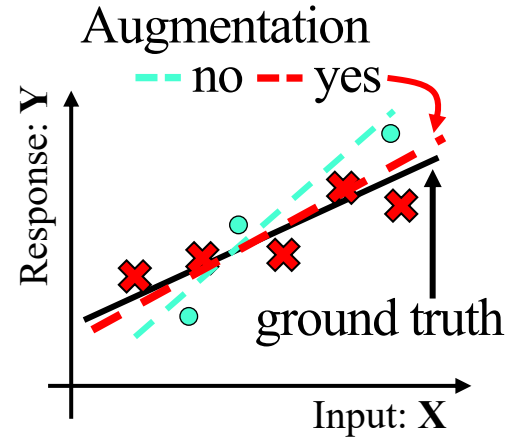
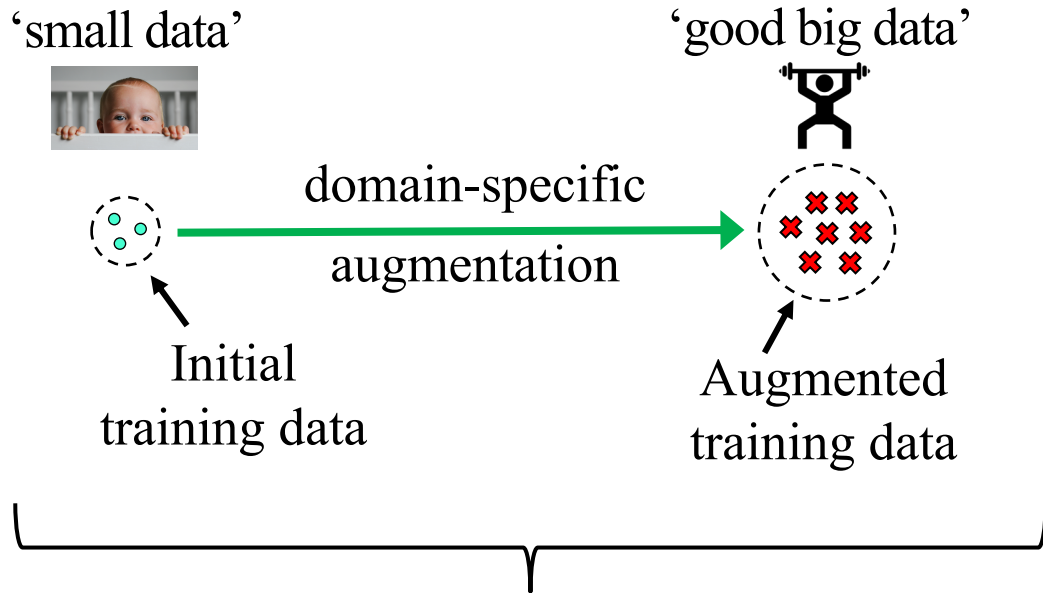
‘good big data’



Augmented
training data



Domain-specific data engineering – seismic collapse risk



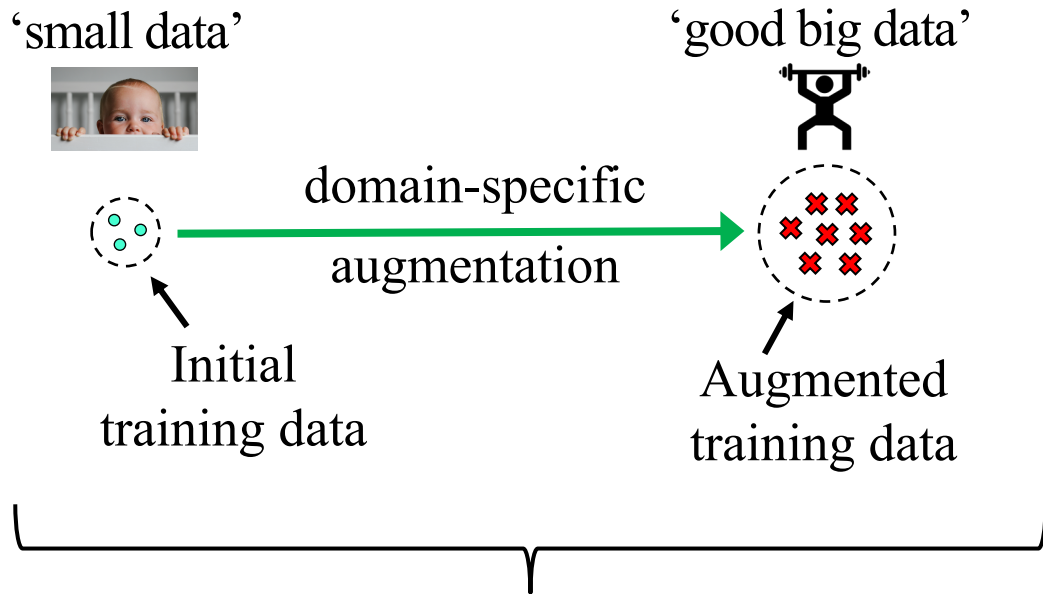
Automated Collapse Data Constructor

(AC/DC)




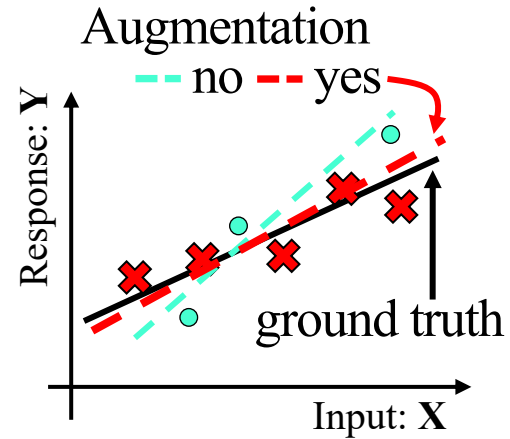
Bijelić, Lignos, Alahi (2022, in review)

Domain-specific data engineering – seismic collapse risk



Automated Collapse Data Constructor

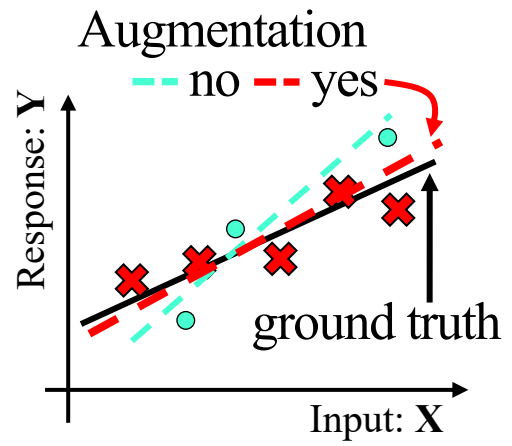
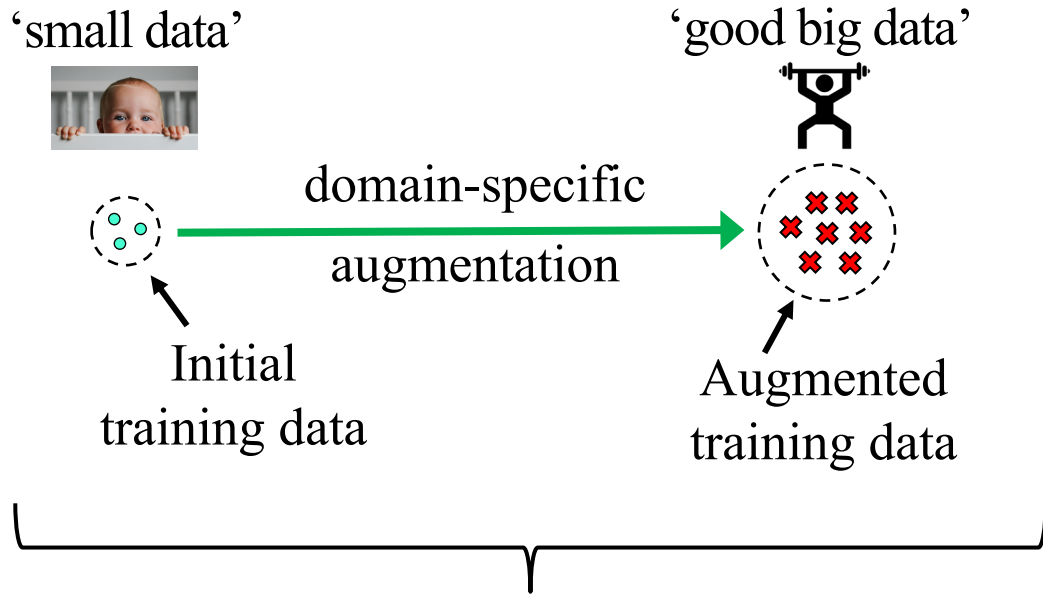
(AC/DC) 




Training data	Error in risk:
100%	~ -20%
20% + AC/DC	~ 3-5%

Bijelić, Lignos, Alahi (2022, in review)

Domain-specific data engineering – seismic collapse risk

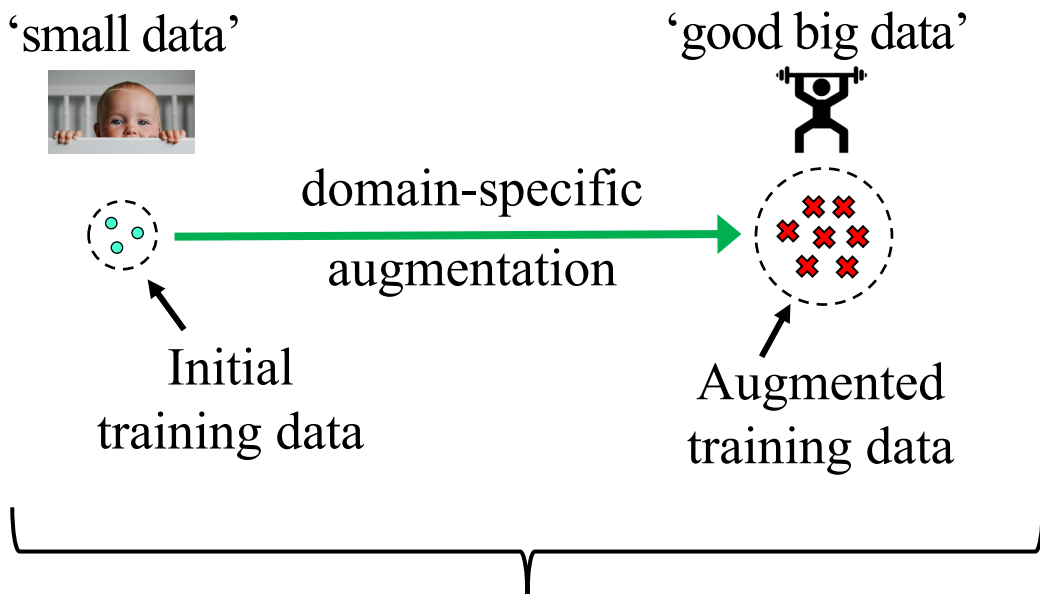


Training data	Error in risk:
100%	~ -20%
20% + AC/DC	~ 3-5%

Automated Collapse Data Constructor
(AC/DC) 

Bijelić, Lignos, Alahi (2022, in review)

Domain-specific data engineering – seismic collapse risk

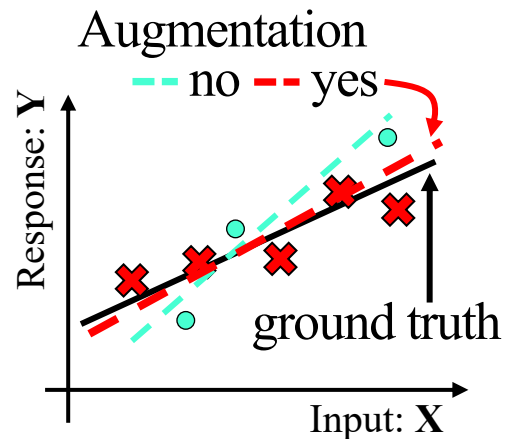


Automated Collapse Data Constructor

(AC/DC)



Bijelić, Lignos, Alahi (2022, in review)



Training data

Error in risk:

100%

~ -20%

20% + AC/DC

~ 3-5%

From Model-centric to Data-centric
AI by Andrew Ng:



<https://www.youtube.com/watch?v=06-AZXmwHjo>

Say no to bias – if linear regression can do it, then everyone can!



designed by  freepik

"<https://www.freepik.com/vectors/flower>"



Image of the Respect @ EPFL campaign - Etienne Etienne Agency

Thank you – let's talk!

Questions?

Nenad Bijelić
nenad.bijelic@epfl.ch