THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Personal Thoughts

Guido Zuccon

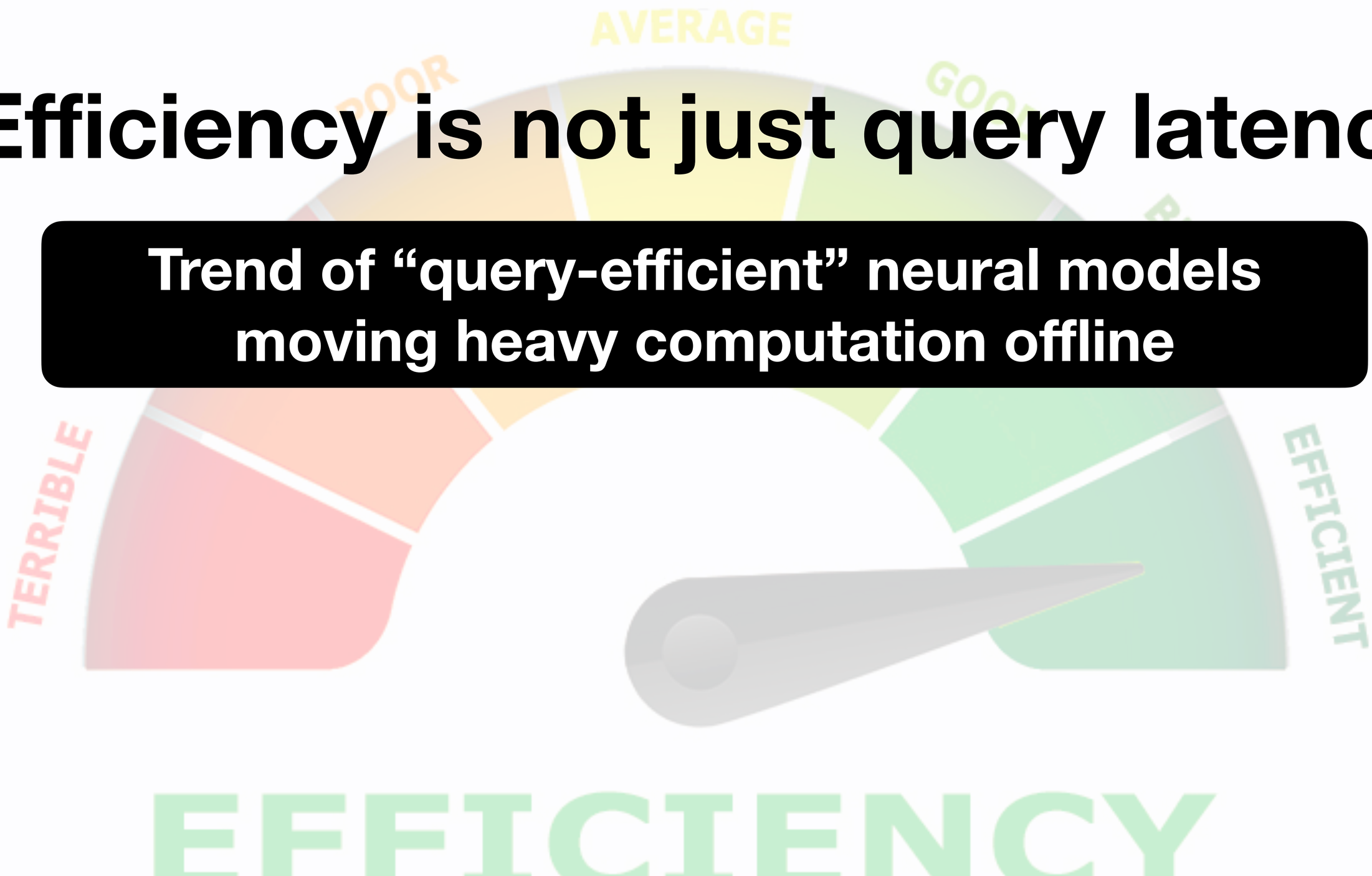
g.zuccon@uq.edu.au
ielab, The University of Queensland, Australia
www.ielab.io

ie
lab

Efficiency is not just query latency

Efficiency is not just query latency

Trend of “query-efficient” neural models moving heavy computation offline



Efficiency is not just query latency

Trend of “query-efficient” neural models moving heavy computation offline

This computation still costs (time, hardware, energy, emissions)

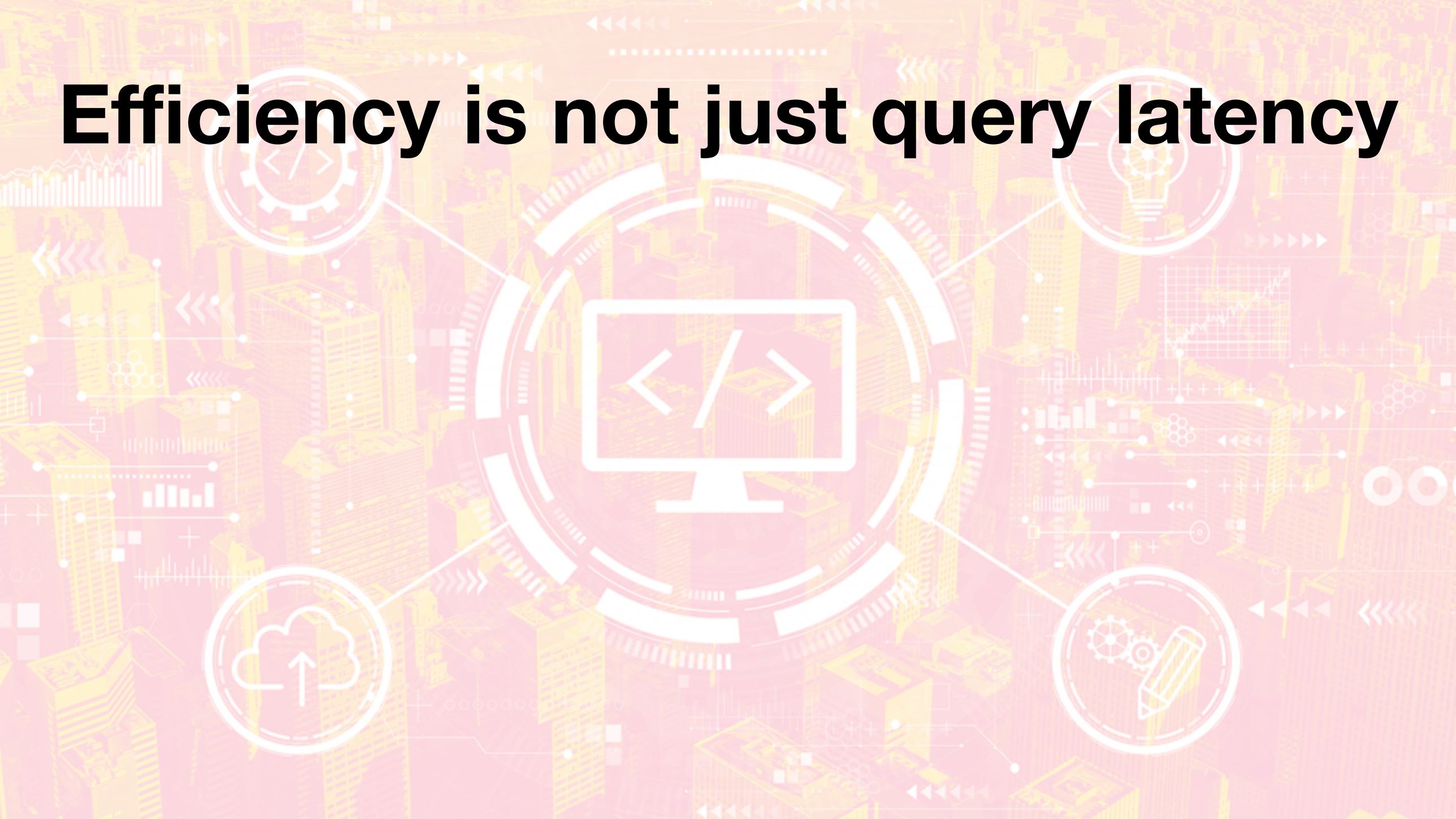
Efficiency is not just query latency

Trend of “query-efficient” neural models moving heavy computation offline

This computation still costs (time, hardware, energy, emissions)

It’s not a once off, as one often thinks

Efficiency is not just query latency



Efficiency is not just query latency

Trade offs: effectiveness vs. efficiency vs. space vs. architecture

Efficiency is not just query latency

Trade offs: effectiveness vs. efficiency vs. space vs. architecture

resource constrained systems

Efficiency is not just latency, energy

Efficiency is not just latency, energy

Data Efficiency

Efficiency is not just latency, energy

Data Efficiency

Learning with little data

Efficiency is not just latency, energy

Data Efficiency

Learning with little data

frugal models, federated learning, few-shot, zero-shot, prompt learning