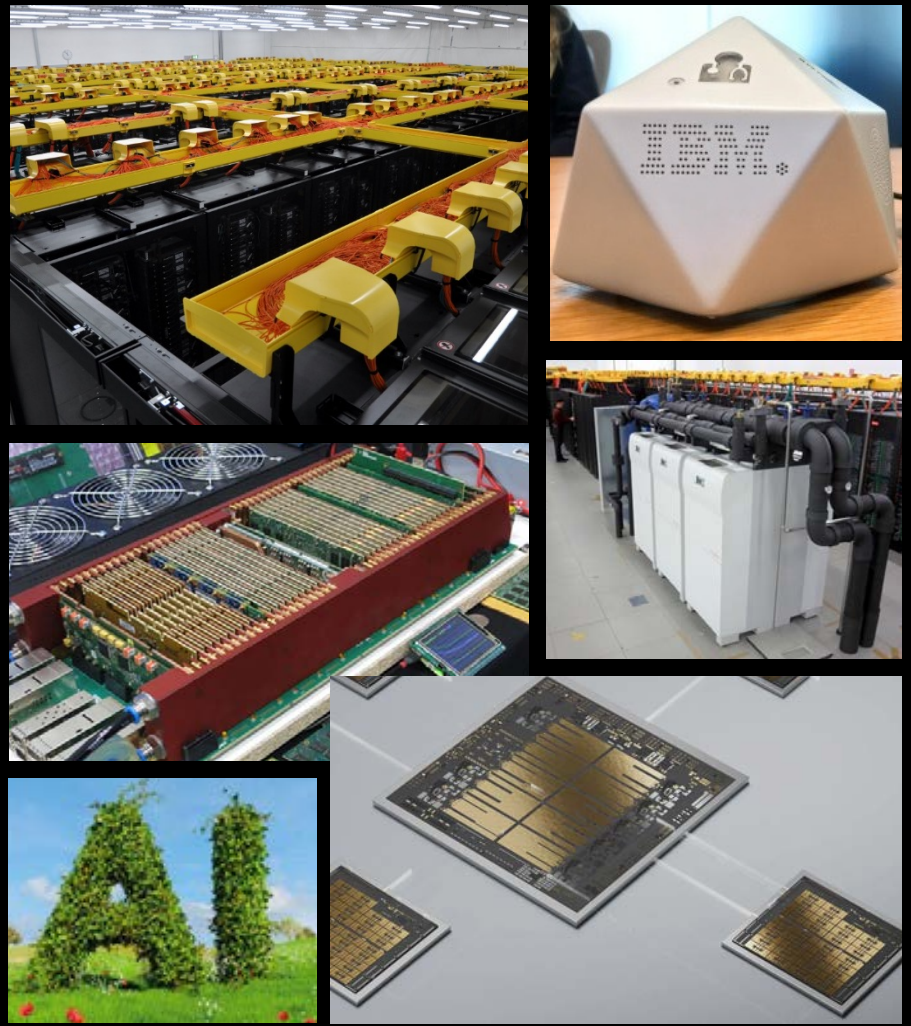
Künstliche Intelligenz: Grundlagen, Entwicklungen, und Anwendungen

Bruno Michel

IBM Gruppenleiter, Nachhaltigkeits-Experte und Berater
IEEE-Fellow, Mitglied US Nationale Akademie der Ingenieure (NAE)

Thermal Transformer (Gründer)
bmi@thermaltransformer.ch



Agenda: KI: Grundlagen, Entwicklungen, und Anwendungen

KI-Grundlagen und Entwicklungen

Teile und Löse Nachhaltigkeits-Problem von KI...

Cloud und Edge Rechenzentrums-Effizienz

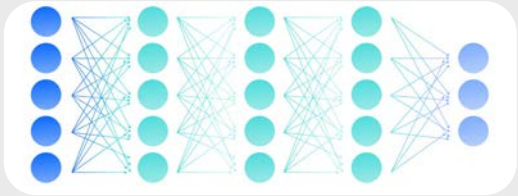
Trend Umkehr von «Roter KI» zu «Grüner KI»

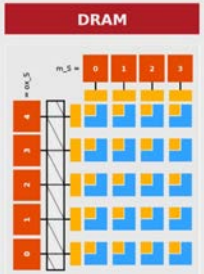
Entscheidungs-Effizienz and Lifecycle-Effizienz

Bedarf für universelle Messung (wie Energy Star)

Skalierbarkeit natürliche gegen künstliche Intelligenz

Wie geht es weiter?

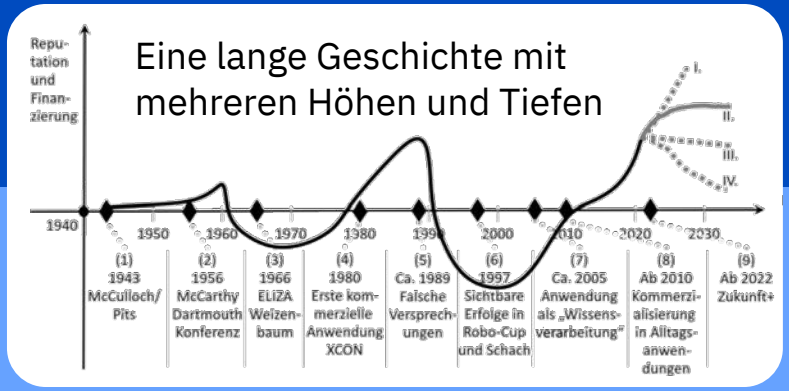




1950er

Künstliche Intelligenz (KI)

Menschenähnliche Intelligenz von Maschinen



1980er

Maschinelles Lernen (ML)

Lernen von historischen Daten

2000er

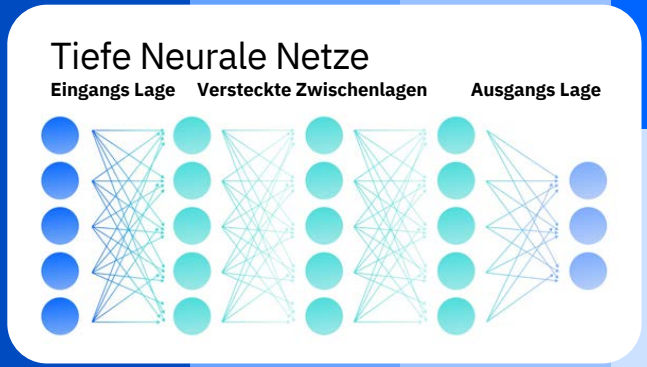
Deep Learning (DL)

ML-Modelle die das menschliche Gehirn nachahmen

2020er

Generative Künstliche Intelligenz (Gen KI)

Fundamentale Modelle die neuen Inhalte generieren können



2030er

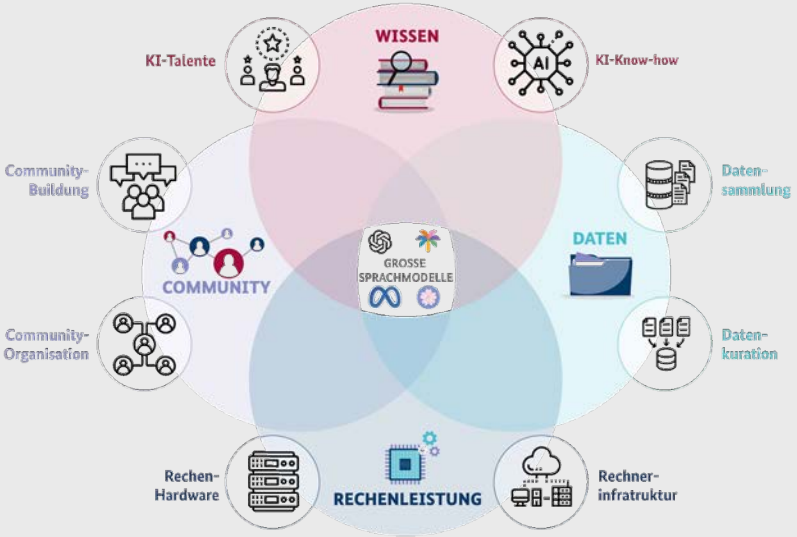
Effiziente Künstliche Intelligenz (Effi KI)

Künstliche Intelligenz die mit untrainierten Menschen konkurrieren kann (Grüne KI)

Künstliche Intelligenz

- Ermöglicht Computern menschliches Lernen, Verständnis, Problemlösung, Kreativität und Autonomie zu simulieren.
- Objekte sehen und identifizieren.
- Menschliche Sprache „verstehen“ und darauf reagieren.
- Aus neuen Informationen und Erfahrungen lernen.
- Anwendern und Experten detaillierte Empfehlungen geben.
- Agieren unabhängig und machen menschliche Intervention überflüssig (selbstfahrendes Auto).
- Schlagzeilen über generative KI (gen AI), die Originaltexte, Bilder, Videos erstellt.
- Generative KI basiert auf maschinellem Lernen (ML) und Deep Learning (DL).

Generative KI



- Beginnt mit unbeschrifteten, Internet-daten trainierten „großen Sprachmodellen“ für Text-, Bild-, Video-, oder Musikgenerierung.
- Besteht aus aus Milliarden von Parametern, und generiert autonom Antworten.
- Nutzer bewerten Ergebnisse und optimieren Modelle wöchentlich aber Grundmodelle nur alle 12–18 Monate.
- Rechenintensiv und teuer aber Open-Source (Llama-2, DeepSeek) senkt kosten.
- Feinabgestimmt mit gekennzeichneten Daten und Frage – Antwort Paaren.
- Reinforcement Learning mit Human Feedback (RLHF), verbessert Relevanz.
- Retrieval Augmented Generation (RAG) erweitert Quellen, für bessere Relevanz.

Vorteile und Nutzen Künstlicher Intelligenz

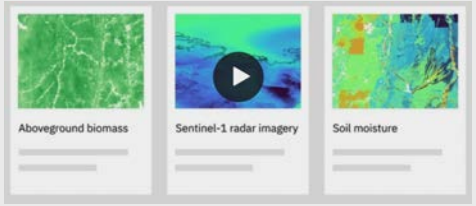
- Bessere Entscheidungsfindung für genauere Vorhersagen mit weniger Fehlern in repetitiven Prozessen.
- Automatisierung von Dateneingabe, Logistik, und Produktion bei gleichbleibender Leistung 24/7 und weniger Personal.
- Reduziertes Risiko - Selbstfahrende Autos verringern Risiko.
- Support: Chatbots liefern 24/7 schneller Antworten auf häufige (einfache) Fragen.
- Betrugserkennung: KI kennzeichnet Anomalien in Transaktionen.
- Personalisiertes Marketing steigert Umsatz. DL empfiehlt Produkte und Dienste.
- Personalwesen: KI optimiert Einstellungen, Papierkram, und Reaktionszeiten.
- Wiederholende Codierungen für Modernisierung von Legacy-Anwendungen.
- Vorausschauende Wartung: IoT Datenanalyse und Betriebstechnologie (OT) vermeiden Geräteausfälle und Lieferkettenprobleme.

KI Anwendungs-Beispiele von IBM

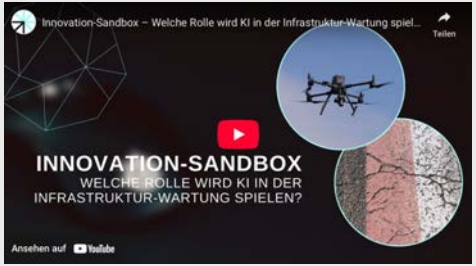
IBM WatsonX: Kundendienst, Personalwesen, Marketing, Finanzwesen, IT-Betrieb und Lieferketten-Optimierung.

«IBM Environmental Intelligence Suite» (EIS): Messung von Entwaldung, Biomasse (Zertifikate), Emissionen, Schneeräumung, Warnung vor Überflutungen und Waldbränden usw.

Vorausschauende Wartung im Bau: Korrosion von Brücken, Pisten, Trassees mittels Analyse von Drohnenbildern



Materialforschung: Schnellere Entwicklung von Materialien für Medizin, Chemie, und Dekarbonisierung.



«Basismodelle» für nachhaltigen Ausbau von elektrischen Netzen.

Risiken von Künstlicher Intelligenz

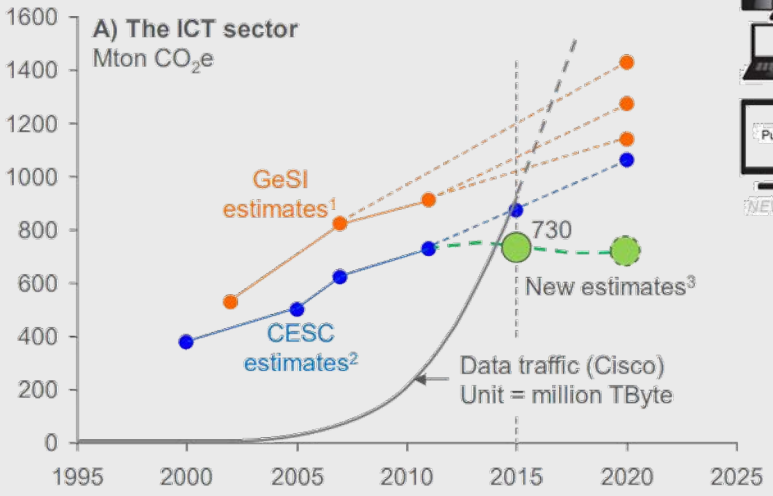
- ~~Elon Musk sieht 20% Risiko, dass KI die Menschheit auslöscht*~~
- Datenrisiken: Datenvergiftung / Manipulation / Verzerrung oder Cyberangriffe.
- Schutz der Datenintegrität, Sicherheit, und Verfügbarkeit im KI-Lebenszyklus.
- Modellrisiken: Gauner nutzen KI-Modelle für Diebstahl oder Manipulation.
- Angriff auf Modell-Integrität, durch Manipulation von Architektur, oder Gewichten.
- Betriebsrisiken: Modelldrift, und Störungen führen zu Ausfällen und Sicherheitslücken.
- Ethik: Datenschutzverletzungen und voreingenommene Ergebnisse.
- Warum uns KI nicht dominieren wird (TED-Vortrag)

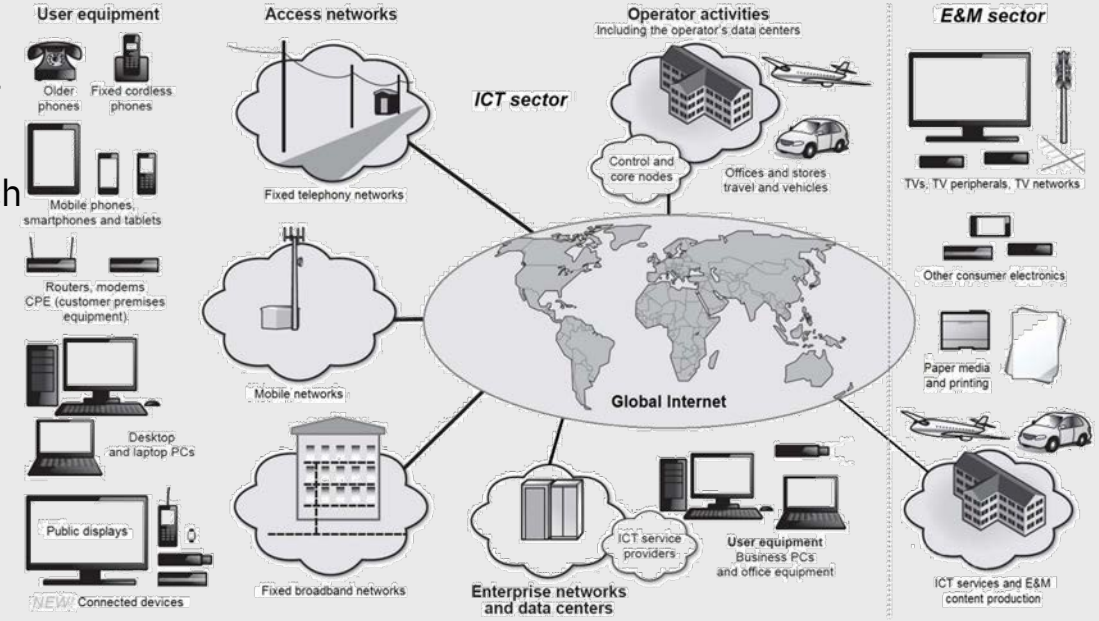
Was limitiert KI?

Weltweite CO2 Emission heute = 37 Gt, ICT ~1 Gt = ~3%

Schelles Wachstum von KI-Energieverbrauch

The Energy and Carbon Footprint of the Global ICT and E&M Sectors 2010–2015, J. Malmodin and D. Lundén, Sustainability 2018, 10, 3027.





Schnell wachsender Kohlenstoff Fussabdruck von Cloud Rechenzentren ...

Daten Transport Energie ist ein grosser Teil

KI-Energie verdoppelt sich alle 3-4 Monate

Aufteilung des Problems zur besseren Lösung

$CFP = E_{IT} \times R_{CPF}$

$R_{CFP} = \{(E_{IT} + E_{DC}) \times CI - \alpha \times E_{reuse}\} / E_{IT}$ R_{CFP}: >2.0 … -1.0

$PUE = (E_{IT} + E_{DC}) / E_{IT}$ PUE: >2.0 … 1.05

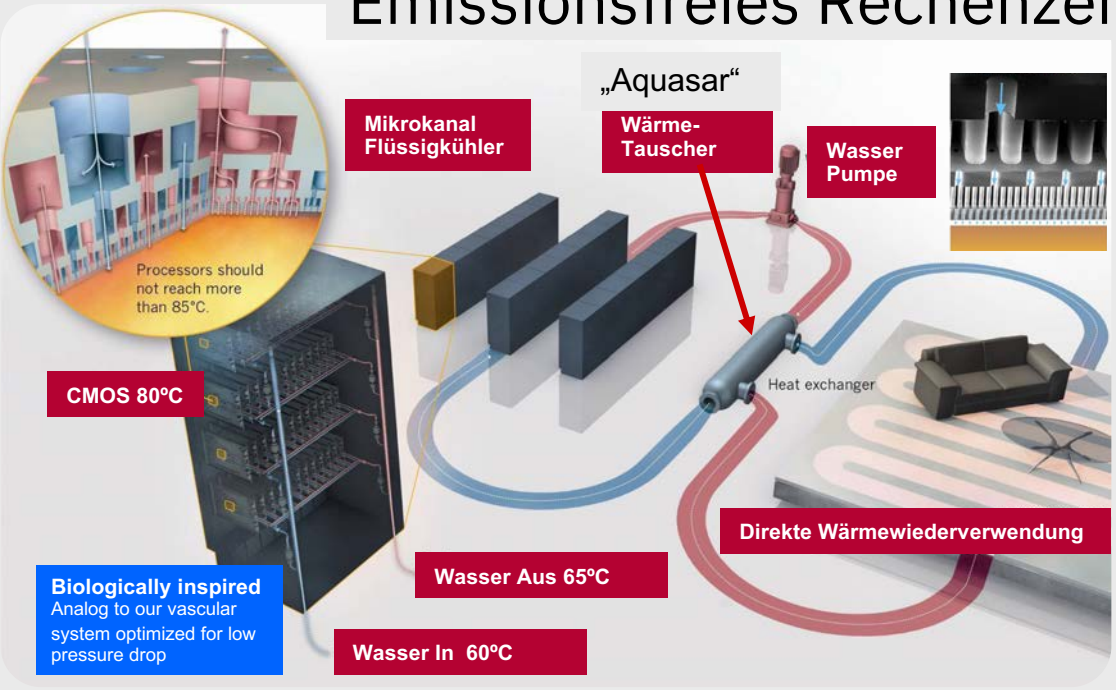
$ERE = (E_{IT} + E_{DC} - E_{reuse}) / E_{IT}$ ERE: >2.0 … 0.0

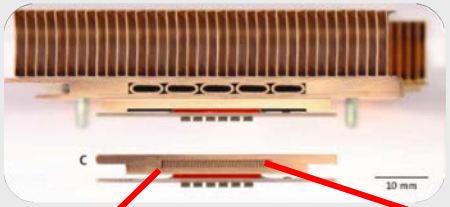
$\alpha \times E_{reuse}$ (physikalischer Offset) α: 1..0 E_{reuse}: 0 … $E_{IT} + E_{DC}$

$E_{IT} = E_{comm} + E_{comp} = E_{KI} + E_{other}$

$E_{comp} = E_{move} + E_{proc}$ ➔ weniger als 5% der ganzen Energie

Emissionsfreies Rechenzentrum





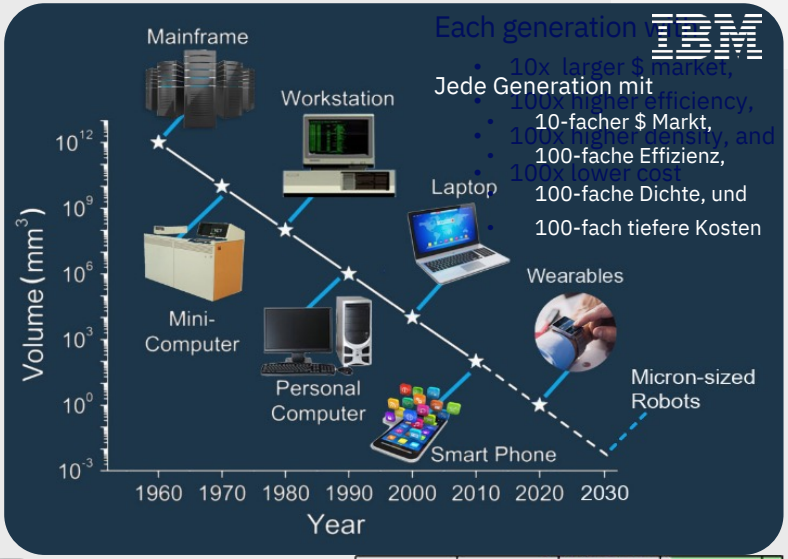
E_{reuse}

- **Chip-Kühlung verbessert Effizienz UND CO2-Fussabdruck**
  - Kühlung mit **ΔT = 20** anstatt 75ºC, **spart 50% Energie**
  - **Widerverwendung:** ~2000 Häuser mit 10 MW Rechenzentrum
- **CO2-Fussabdruck Reduktion in allen Klimas**
  - Heisse Klimas: Freie Kühlung, Entsalzung
- **Europa: 5000 Fernheiz-Systeme**
  - Verteilung 6% des thermischen Bedarfs

Bell's Law: More Integration

E_{IT}

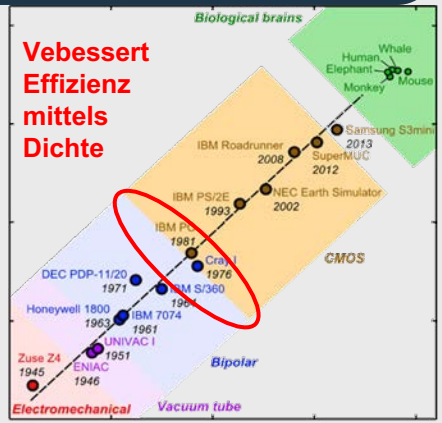
- Alle 12-15 Jahre Neustart Generation
- Hardware Kostenanteil schrumpft von 100% (Zentralcomputer) auf <10% mit Zusatzfunktionen
- Messung und Übermittlung miniaturisiert
- Tiefer Widerstand ermöglicht höhere Dichte
- Messung und Rechnung in tragbaren Sensoren
- **Dichte verbessert Effizienz!**
- **Effizienz und tiefe Kosten durch Bell's Sches Gesetz**



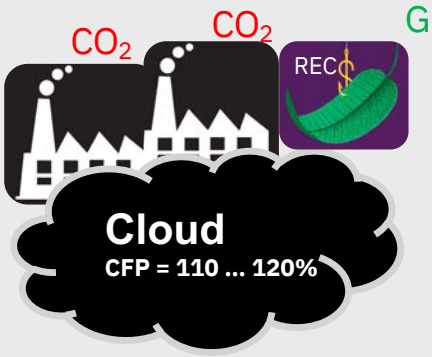
Volumen-Dichte Skalierung

- **5'000x weniger Leistung**
- 50'000'000x dichter
- Skalierbar bis zeta

P. Ruch et al., IBM J. Res. Develop. 55, 15:1-15:13 (2011).



Konzept für Emmissionsfreie Cloud



Grünfärberei

Rack Wärme-Tauscher
best practice: Kalt/ warm Gang etc.
Reduziert PUE=ERE = ~1.1
Wenige sind schon hier

Warm Wasser Kühlung
Power Usage Effectiveness =1.1
Energie Nutzung Effektivität = 0.7
Energieverkauf (bessere RoI)
Summit in Oak Ridge

Wiederverwendung mit Heiss-Wasser
Power Usage Effectiveness =1.1
Energie Nutzung Effektivität = 0.1
Kohlenstoff Fussabdruck Reduktion = 10 ... 20%

Wärmenutzung + Erneuerbare Energie
ERE = 0 → CFP = 0
Physikalischer Offset mit Erneuerbaren ergibt negative Emission
Braucht eine neue Messgrösse!
SuperMuc

Integrierte Energie und Kühlung
Reduziert Energieverbrauch (E_{IT}) by 10 ... 10'000

Cloud könnte hierher springen
CFP = 0% (ohne RECs)
Himmel

Cloud springt hierher
Fussabdruck = 0%
Negativer Fussabdruck
(ohne Zertifikate)

$$\mathbf{CFP} = E_{IT} \times ERE \times CI$$

Schlechte Luftzirkulation
PUE > 1.5
Cloud RZs sind meist da

Hölle

2. Schrittweise Einführung von Chip Heisswasser-Kühlung und Wiederverwendung in modifizierten Rechenzentren

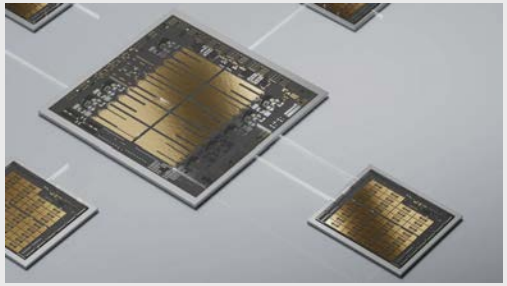
..... Treppe ins (Effizienz) Paradies .....

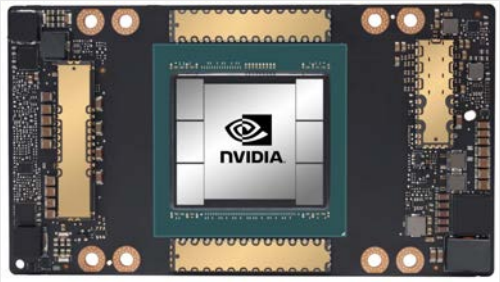
Reduktion des Datentransports E_{move}

- Architekturen mit Parameter Wiederverwendung
- Eingangs-Aktivierungen limitieren Leistung.
  3.3x Verbesserung mit 8 statt 32-bit Quantisierung
- FPGA-Akzelerator on PYNQ-Z1 vs. Nvidia Jetson Nano.
  - 2.27x Reduktion von Inferenz Latenz auf GPU
  - 1.36x Reduktion auf FPGA (3.87x on Konvolution Lagen)

z16 Telum 32-fach erweitert, als Spyre «AI-Unit» mit 25.6 Mia 5nm Transistoren. 256 Kerne in IBM Z für generative KI. Matrixbeschleuniger mit int4 und int8 Formaten, um KI-Modelle effizienter und weniger speicherintensiv zu machen.

Blackwell aus zwei mit 5 Tb/s gekoppelten Chips mit 208 Mia 4nm Transistoren (+30%). 8 HBM3e liefern 192 GB, schnellen Speicher. BG200 25-Mal effizienter/schneller als Vorläufer. Transformer Engine mischt FP16, FP8 und FP4.

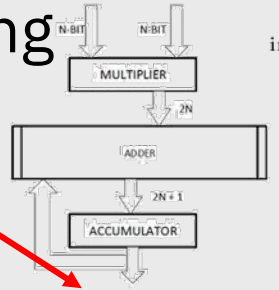




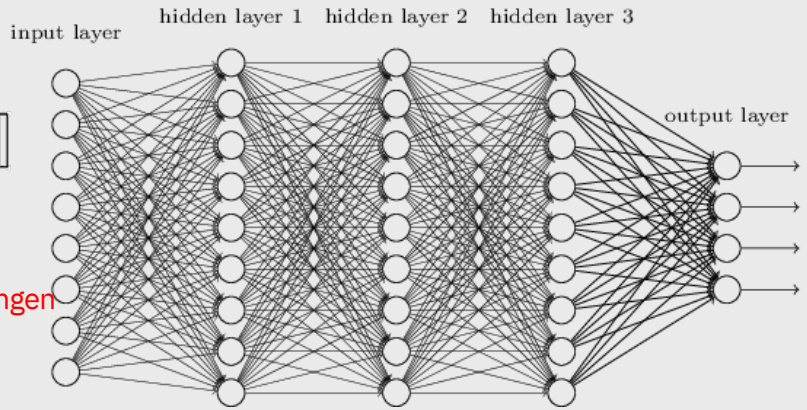
Plattformen Spyre und GPU

Skalierung von Deep Learning

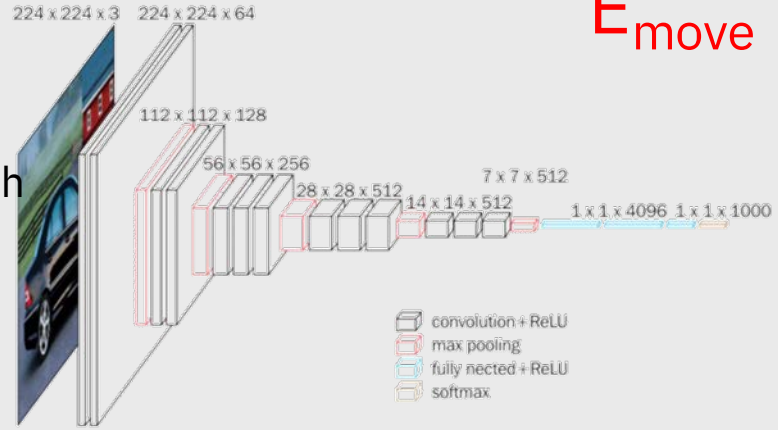
- Branching factor (BF) N^2 mit N von
  - 224*224*64 = branching factor 3.2 Millionen *2
  - 112*112*128 = branching factor 1.6 Millionen * 3
  - 56*56*256 = branching factor 0.8 Millionen * 4
  - 28*28*512 = branching factor 0.4 Millionen * 4
  - 14*14*512 = branching factor 0.1 Millionen *4
- Branching factor KI vs. Klassische Architektur
  - BF = Rent Exponent GPU 0.4 (Reduziert Datentransport)
  - BF = Rent Exponent RISC-Maschine 0.5
  - BF = Rent Exponent CISC-Maschine 0.8
- Modell mit 2x Elementen braucht 4x mehr Leistung und pro Element mehr Training (➔ bis 8x mehr)
- Systeme erreichen Sättigung ➔ Mehr Energieverbrauch
  - Braucht Hierarchie von kleineren Modellen



1000 Verbindungen



E_{move}

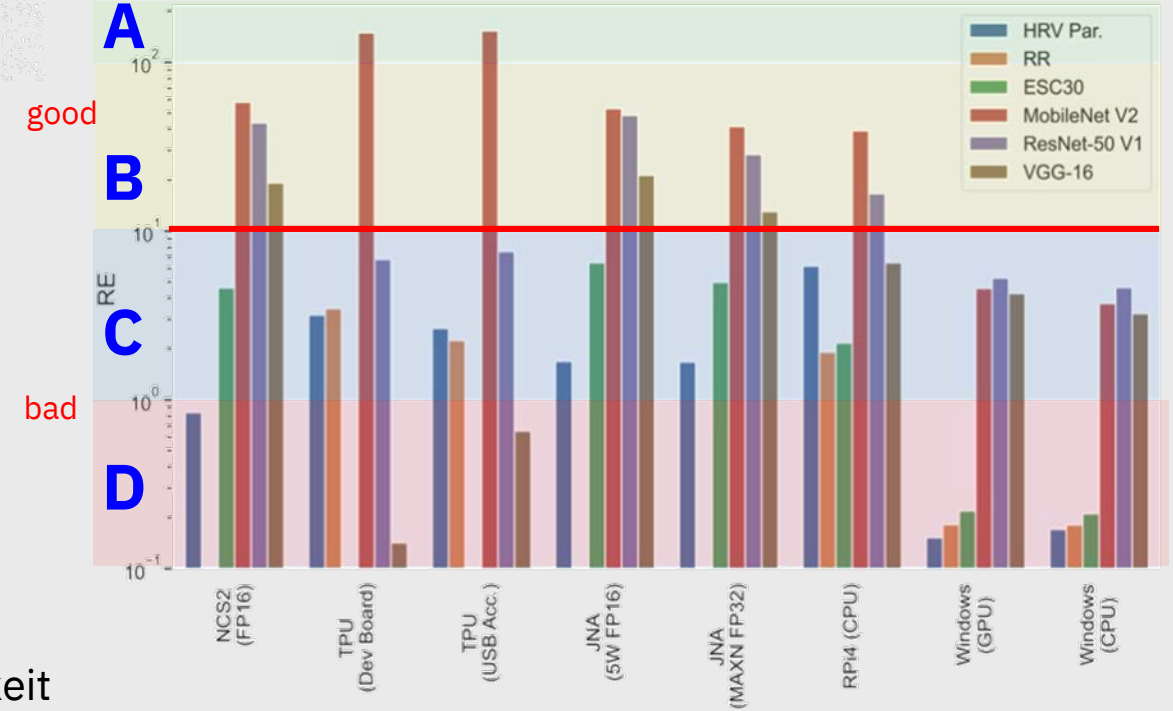


Recognition Efficiency

$$RE = \frac{RECgrad \times Cl}{\sqrt[2]{Einf}}$$

- Einfache Bewertung A-D
  - A Bewertung 100 < RE
  - B Bewertung 10 < RE < 100
  - C Bewertung 1 < RE < 10
  - D Bewertung RE < 1
- Einfache Herleitung
- Ausgeglichener Einfluss von Genauigkeit, Energie, und Komplexität
- Immer basierend auf Top 1
- Benutzt mit REC_{grad} und Genauigkeit

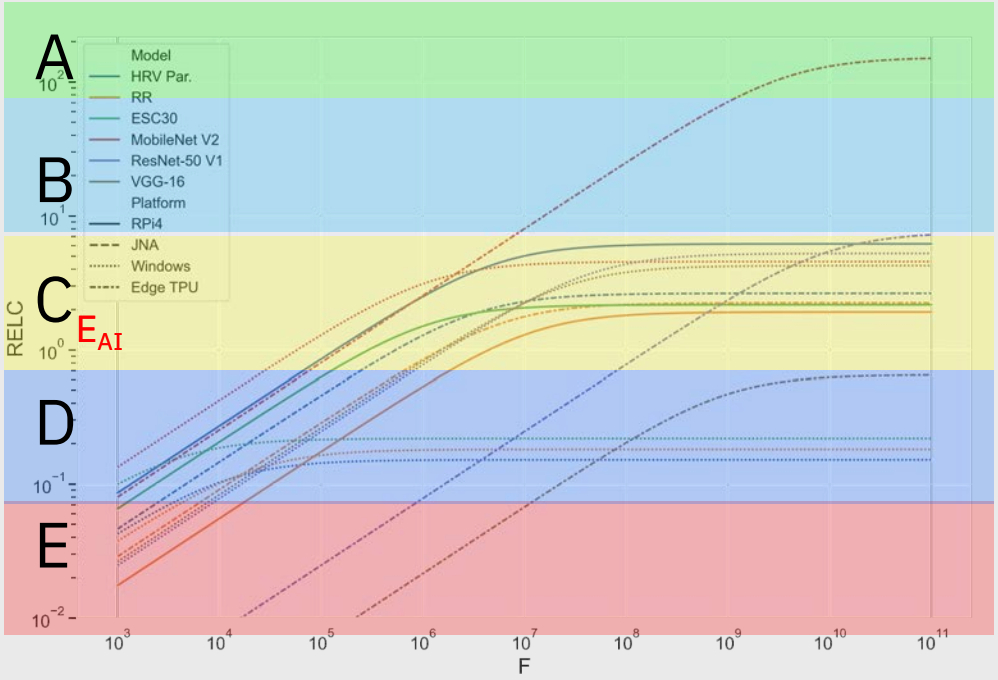
$$E_{IT} = E_{KI} \quad (+E_{other}+E_{comm})$$



Lifecycle Efficiency

- Lifecycle recognition efficiency für verschiedene Modelle (Farbe) und Plattformen (Linientyp)
- Braucht viele Nutzungen des Modells (F) um die Energie zu amortisieren Gewinner: mobile net auf TPU
- Cloud “leidet” unter Energieverbrauch für Datentransport
- Übergang zu Training Inferenz Dominanz bei mehr als 1 Million Nutzungen

$$RE_{LC} = \frac{RECgrad * Cl}{\sqrt[2]{Einf + Etrain/F}}$$



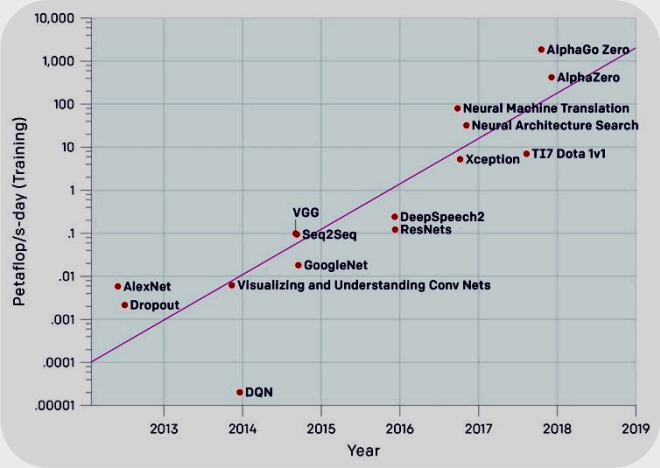
E_{KI}

17

Grüne Rechenzentren - Grüne KI

KI Energie Problem

Deep Learning Energie verdoppelt in
3 Monaten; 300,000x Zunahme 2012-19



Schritt 1

Universelle Messung:
Recognition Efficiency (RE)

Mann kann nur verbessern, was man misst!

– RE = Genauigkeit * Cl / $(E_{inf})^{0.5}$
– Messung definiert grüne oder rote KI

Schritt 2

Modell-Vergleich auf gleichem System
RE variiert >50x für KI-Modelle

– Modell-und Bibliothek-Optimierung

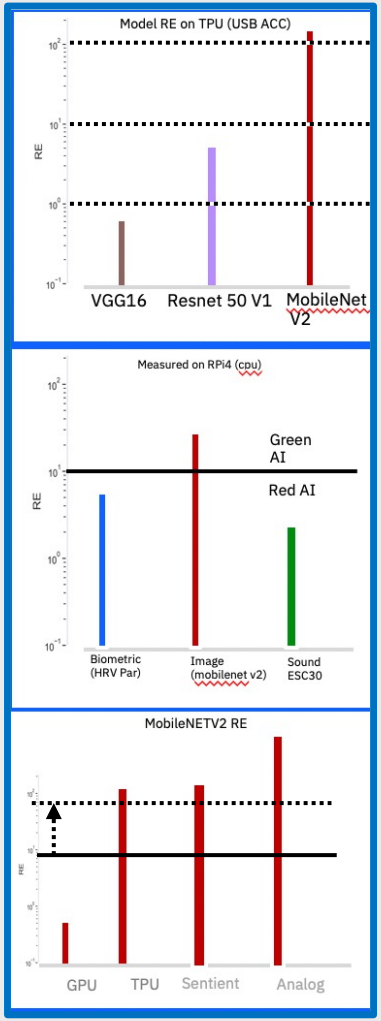
Schritt 3

Modell-Vergleich mit anderer Komplexität
Biometrie, Bild, Ton wie Energy STAR

– Standardisierte Messung von Grüner KI
– Beurteilung der Modell Reife und Qualität

Schritt 4

Vergleich von Plattformen
– Beschleuniger und effiziente Bibliotheken
– ASICs 8-2 bit
– Analoge Beschleuniger



Effizienz Steigerung in eine Nachhaltige Zukunft

- Explodierender Energieverbrauch
- **KI «versaut» Cloud Fussabdruck**
- Computer Architektur >35 Jahre alt
- GPUs neu umdefiniert für KI
- Bis dato keine Effizienzmessung!
- Training mit schlechten Daten
- Schlechte Architektur

Heutiger Stand 1x
➔ Rote KI

Legacy
Explodierender Energie Verbrauch
GPUs

Besser Modell Architektur
Fokus auf Leistung
CNN, RNN, Hybrid
1st Gen. Beschleuniger

Effizientere Modelle bevorzugt
Recognition Efficiency
$RE = Acc. * CI / (E_{inf})^{0.5}$
Beschleuniger und Modellauswahl

Messungen, Modelle, und Trainings Daten
100 x in 0.5 - 2 Jahren

Effizientere Models entwickelt
Eltern Filter und Literatur Durchsuchung
Wenig Trainingsvoluimen
Mehr Genauigkeit

Effizientere Beschleuniger
Parameter wiederverw.
Weniger Datenbewegung
Angepasste Genauigkeit
2nd Gen. Accelerators

Effizientere Rechner
Kompakte Bauweise
Edge Inferenz

1000x besser in 2 - 6 Jahren

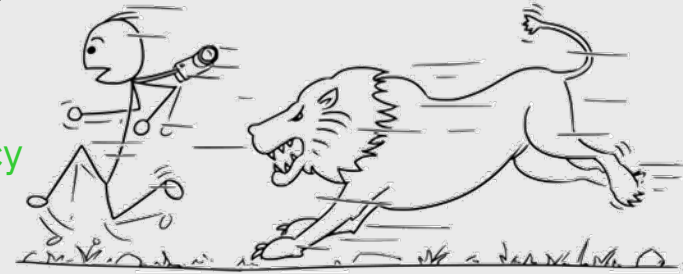
- **10'000x kleinerer CFP**
- Recognition Efficiency
- Training mit Eltern Filter
- Aktuelle Literatursuche
- Bessere Beschleuniger
- Micro-Rechenzentren
- Edge Inferenz wichtig

Beschleuniger und Plattform
➔ Grüne KI

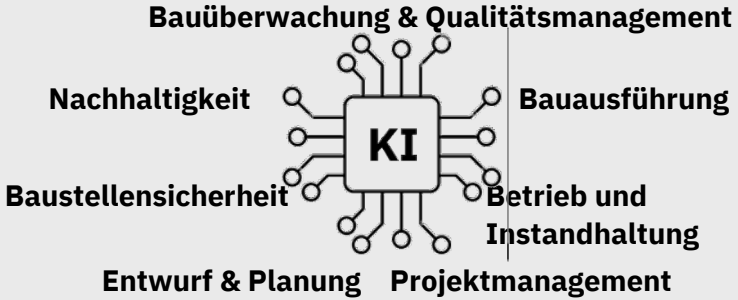
bmi@zurich.ibm.com

Vergleich Mensch - Maschine

- 20 J Cloud Inferenz wie 10 W, LED-Lampe für 2 s
  - Menschen unterscheiden Bilder mit 4 J ➔ 5x so effizient wie VGG-16
  - MobileNet TPU = 4 mJ für 1000 Objekte, Menschen unterscheiden >30’000 Objekte
- Menschliche Reflexe: Effiziente Entscheidungen ohne Neocortex
  - Biologie nutzt Hierarchie! Zuunterst sind die schnellen, effizienten Reflexe
  - Seltene Fälle erlauben einen höheren Energieverbrauch
- Chatbot Antwort braucht 1000-10’000-fache Energie
  Service ist gratis weil wir mit unseren Daten »bezahlen»
- DL flexibel, ineffizient – braucht effiziente HW für Nachhaltigkeit
- Energie optimierte Netzwerke sind genügsam bei Genauigkeit
- Menschen sind genügsam, «schneiden» Kurven; gut genug reicht
  - Den perfekten Pfad zur Flucht vor Löwen zu finden hätte zu lange gedauert.
    Wir wären gefressen worden bevor wir die Rechnung abgeschlossen hätten
- Messung von genügsamen Entscheidungen: Recognition efficiency RE, und Life cycle recognition efficiency RELC



KI-Anwendungen im Gebäude



KI ist aber einer gewissen Variabilität besser als regelbasierte Multiparameter-Optimierung.

KI-Anwendungen sind aufwendig, weil sie auf grossen Mengen guter Daten basieren.

Saisonales Energiemanagement profitiert von KI weil saisonale Energiespeicher sehr teuer sind, und bestmöglich eingesetzt werden müssen.

Die Baubranche ist ein digitaler «Nachzügler». Entwicklungen sind KI-basiertes Informations-management, Planungsunterstützung, Terminkontrolle, und Risikomanagement.

KI verbreitet sich schneller mit digitalen Zwillingen, IoT-Sensoren, und kontinuierlicher Bilderkennung. Dafür brauchen Baustellen bessere Kommunikationsinfrastruktur.

Digitale Tools wie openBIM (Building Integration Modeling) müssen mit besseren Schnittstellen vorangetrieben und Dokumente besser integriert werden.

Leuchtturmprojekt mit saisonalen Speichern in Obersays, Graubünden

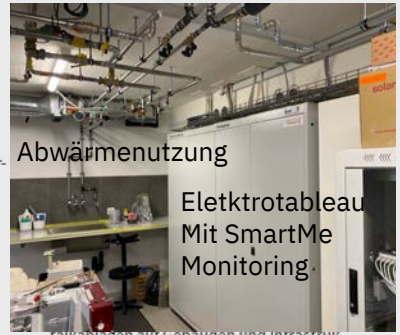
Vorreiter der Solartechnologie

Durch innovative Ansätze wie ästhetische Fassaden-PV, PVT auf dem Dach und saisonale Speicher wird mit einem aktuell im Bau befindlichen Projekt in Obersays ob Trimmis GR ein wegweisendes Modell für die Energiezukunft alpiner Regionen geschaffen.

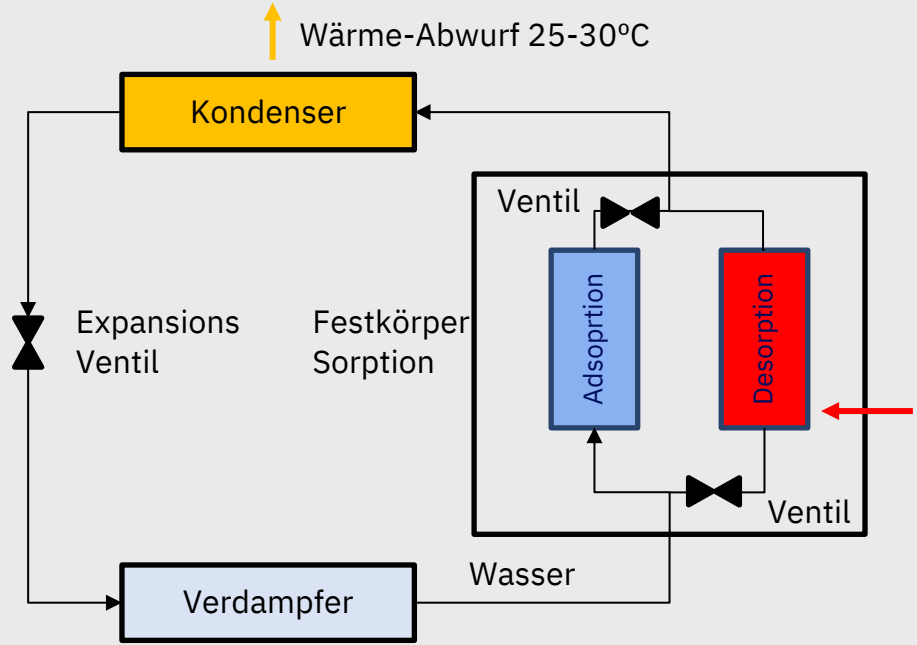
Gebäudetechnik 5 · 24

Bauherr Bruno Michel, Nachhaltigkeitsexperte bei IBM, will mit seinem persönlichen Bauprojekt die nachhaltige Entwicklung alpiner Gemeinden vorantreiben.

Architekt dieses Bauprojekts ist Roman Hug, der auch politisch als Nationalrat aktiv ist.



Umwandlung von GPU-Abwärme in Kälte



Festkörper Sorption kann niedergradige Wärme von GPUs nutzen

Geplante Kommerzialisierung
bmi@thermaltransformer.ch

KI: Grundlagen, Entwicklungen, und Anwendungen

KI-Grundlagen und Entwicklungen

Teile und Löse Nachhaltigkeits-Problem von KI...

Cloud und Edge Rechenzentrums-Effizienz

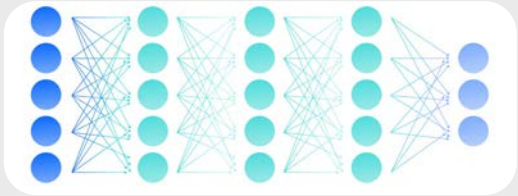
Trend Umkehr von «Roter KI» zu «Grüner KI»

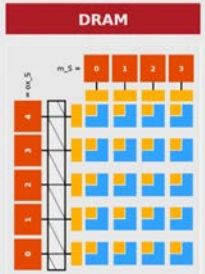
Entscheidungs-Effizienz and Lifecycle-Effizienz

Bedarf für universelle Messung (wie Energy Star)

Skalierbarkeit natürliche gegen künstliche Intelligenz

Wie geht es weiter? ➔ Druck auf Effizienz, Grüne Effi KI





Vielen Dank für Ihre Aufmerksamkeit

Kontakt: bruno.michel.bmi@bluewin.ch

Bruno Michel, Winkel 8, 7202 Says

Firma zur Kommerzialisierung von Wärmegetriebenen Wärmepumpen

bmi@thermaltransformer.ch

Webseiten:

Haus der Zukunft und saisonales Energiemanagement

https://glaskreationen.ch

https://seasonal-energymanager.com