

Undecentralized Personal Digital Data

Alon Halevy

Meta, Reality Labs Research

August 27, 2023

Why Care About Personal Data?

We create personal digital data throughout our day.

- Significant attention on: what are *corporations* doing with this data?
- Our interest: what are *we* doing with our data?

- The ideas on personal data date back to Vannevar Bush in 1945.
 - MyLifeBits (Gordon Bell @MSR) and many other projects since.
- Our life is even more digital now:
 - Always-on perception (smart glasses) may be coming.
 - Personal AI assistants may be coming.

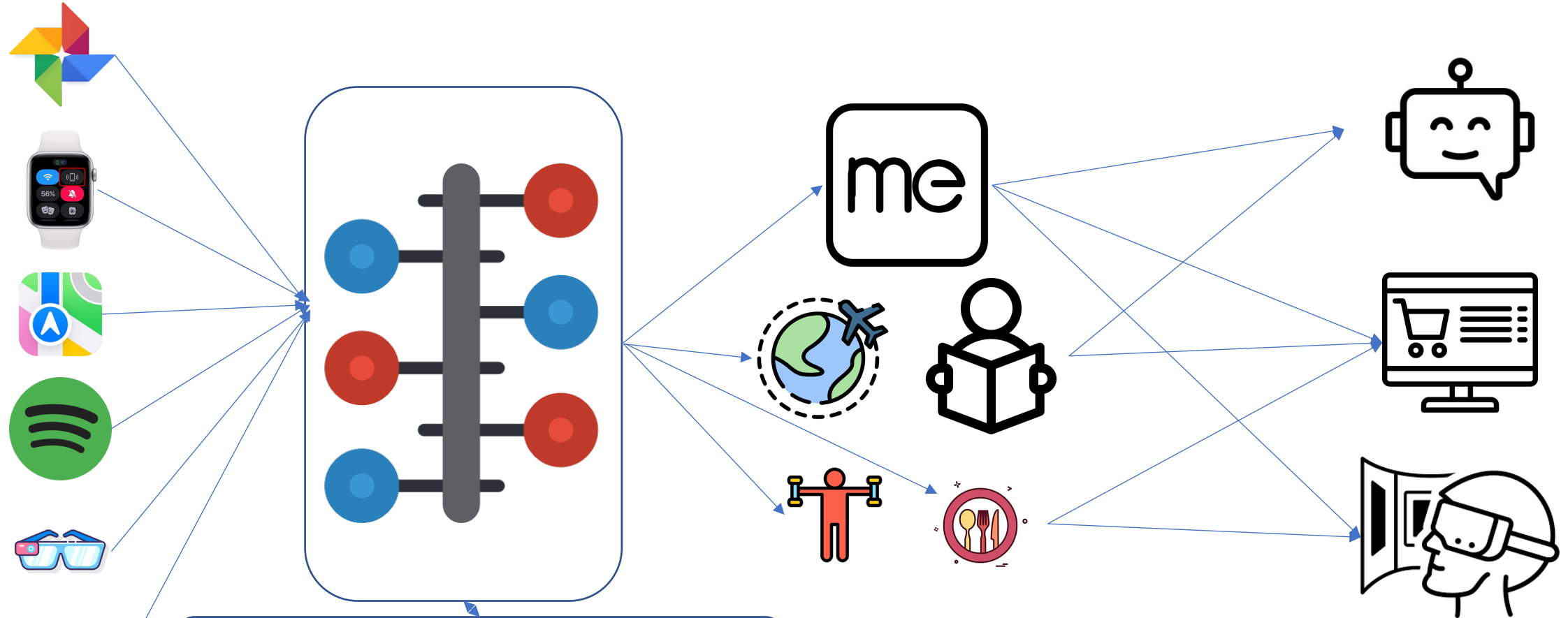
Imagine an Eco-System of Personal Data

Personal digital data

Personal timeline

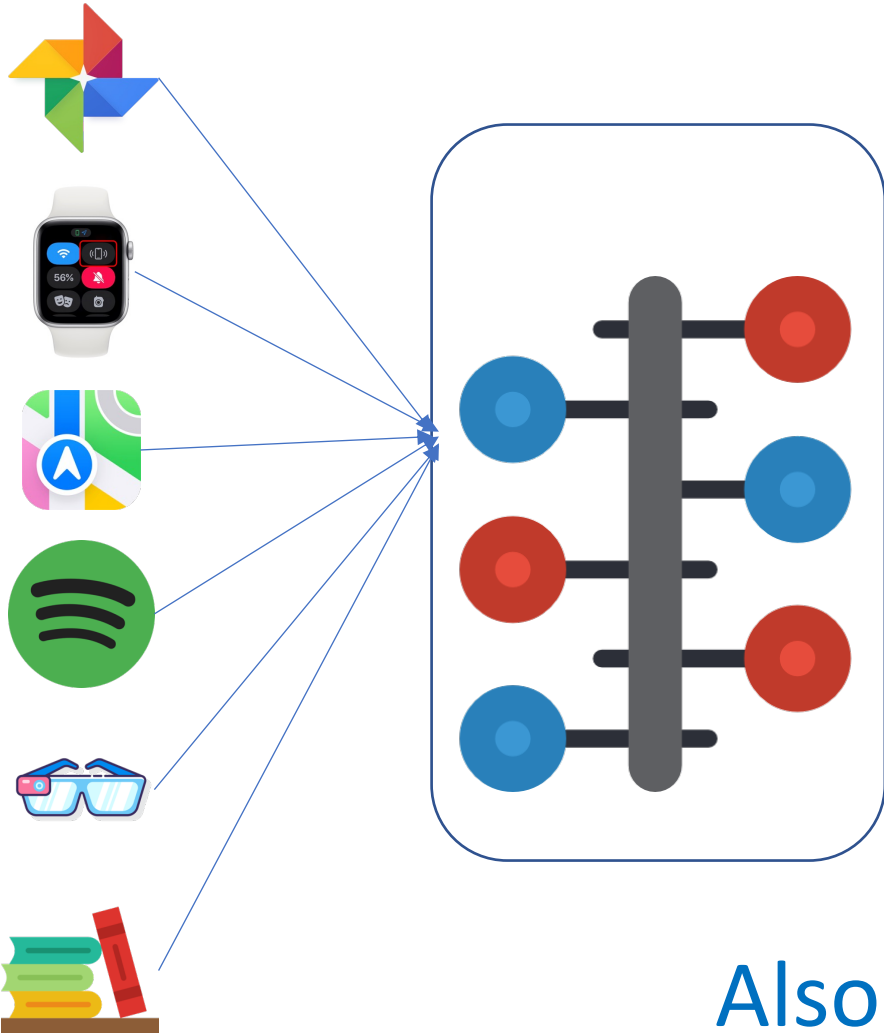
Smaller summaries of me

Interactions in the digital world



When did I go to Tokyo? Where did I run when I was in Tel-Aviv? How many books did I read last year?

Importing Data



- Possible, thanks to GDPR. DMA will make it easier.
- We don't need to import all the data
- New toys will give us more kinds of data:
 - Sleep trackers
 - Smart clothing
 - Brain-computer interfaces

Also include “Projectories and Subjectories”

TimelineBuilder

github.com/facebookresearch/personal-timeline

Alon's April
15, 2019, in
Vancouver.

Started the day with a run **Attended a conference**
Enjoyed coffee

Went for a show
in the evening

05:57: running 26 minutes

Mon Apr 15 2019 05:57:30 GMT-0700 (Pacific Daylight Time)

[More details](#)

Brewing up a fresh cup of coffee.

Mon Apr 15 2019 14:19:43 GMT-0700 (Pacific Daylight Time)



[More details](#)

"Ready for the show!"

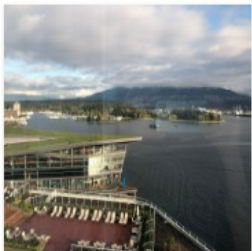
Mon Apr 15 2019 18:30:48 GMT-0700 (Pacific Daylight Time)



[More details](#)

"A sea of possibilities."

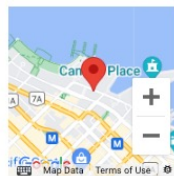
Mon Apr 15 2019 07:58:30 GMT-0700 (Pacific Daylight Time)



[More details](#)

Vancouver Convention Centre West

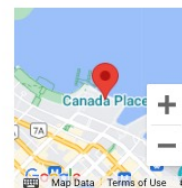
Mon Apr 15 2019 14:19:45 GMT-0700 (Pacific Daylight Time)



[More details](#)

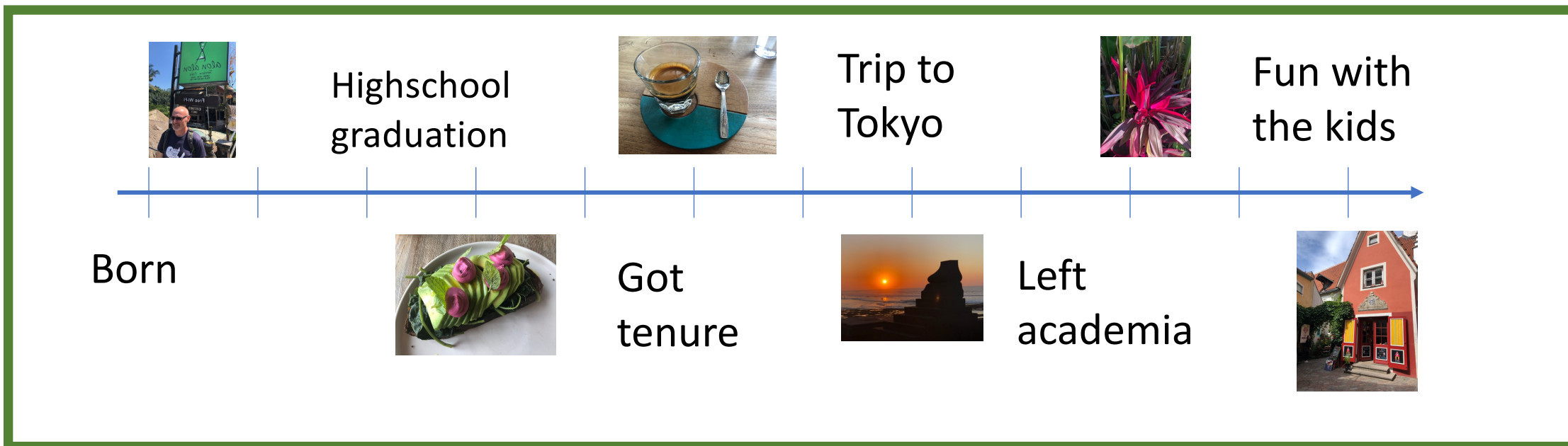
Vancouver Convention Centre West

Mon Apr 15 2019 18:31:20 GMT-0700 (Pacific Daylight Time)



[More details](#)

Fusing Data to a Timeline: a database of Episodes



- Every episode has a start/end time and location.
 - Important to compute the *experienced time*.
 - No other structure required.
- Why is fusion important? Because data sources complement each other.

Some of the Technical Challenges

- Deriving episodes -- the raw data doesn't tell the whole story
 - Lots of photos from Tokyo → but the real itinerary is not explicit
 - Photos and shopping receipts → but it was really a birthday party (and prep)
 - A video of your day → ???
- Querying timelines:
 - “Where did I take my mom when she came to Seattle?”
 - Will LLMs be able to answer these questions alone? (not if you ask a database person!)
 - TimelineQA: a synthetic benchmark to start investigating QA [Tan et al., ACL 2023]
- How can we train models to answer these queries?

How do we develop this vision securely?

- The data should be stored securely and available only to *you*.
- Sharing slices of the data should be on the fly and in language:
 - Share my eating preferences with this waiter for the next 10 minutes
 - Share the itinerary of my trip to Hong Kong with Julia
- What happens with shared experiences? (e.g., couples, families)
- How do we protect users from overreach by institutions?