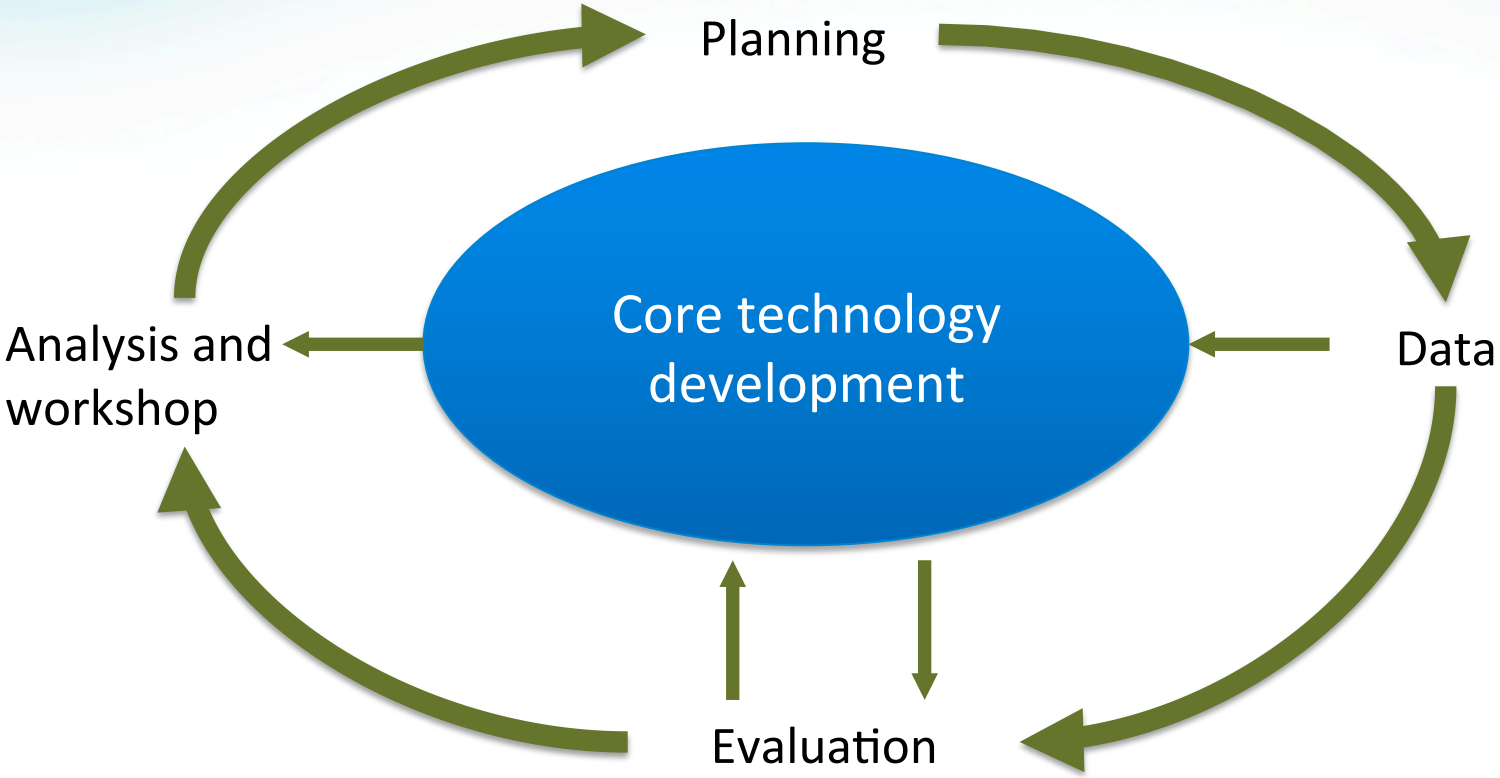
The NIST IAD Data Science Evaluation(DSE)

Craig Greenberg
March 17-18, 2016

Importance of Measurement

Evaluation Driven Research



Efficiency of Evaluation

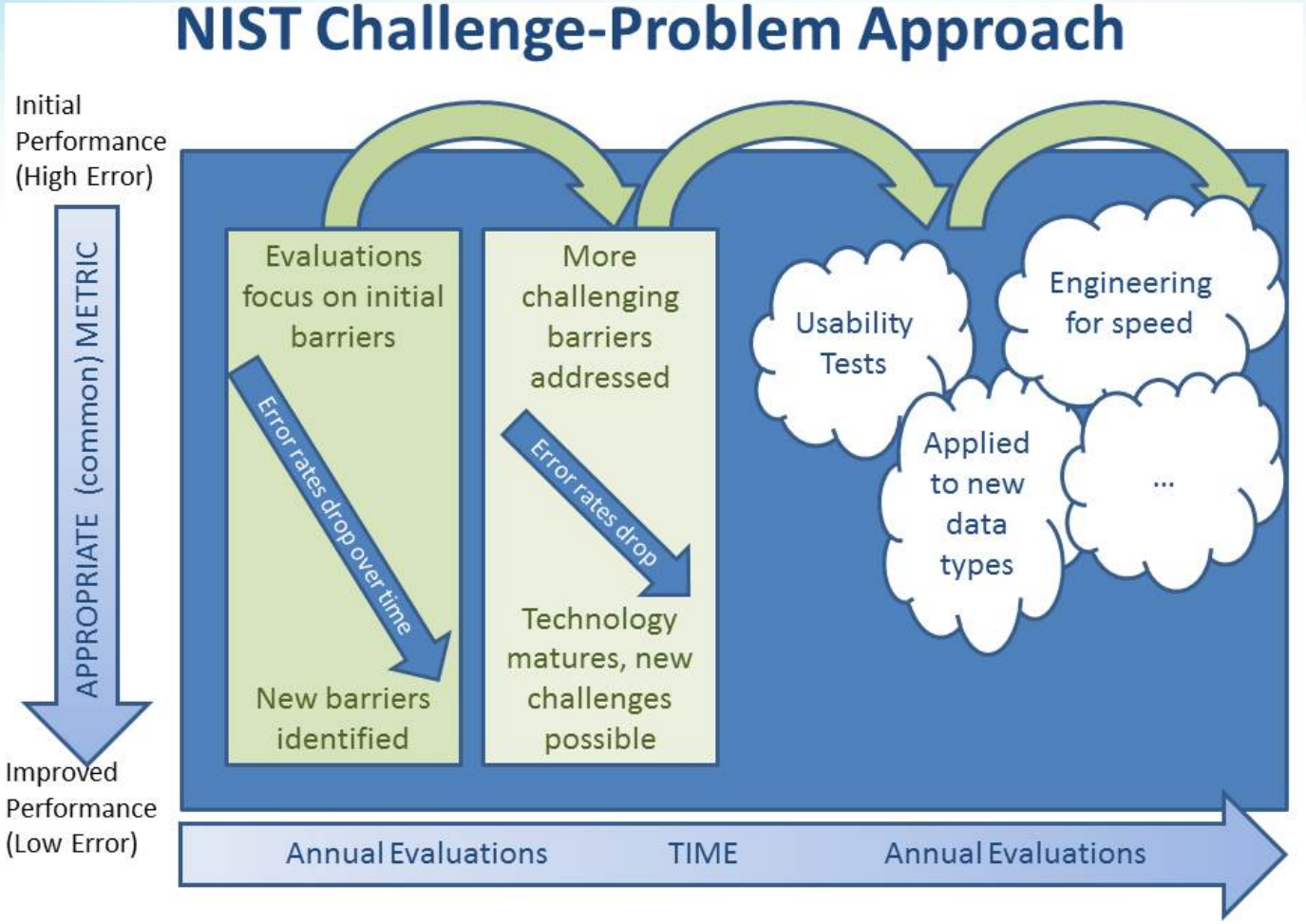
Well designed challenge problems, datasets, and metrics **facilitate research progress**

- Reduces spin-up time and general overhead
- Provides a common framework for sharing and understanding approaches and results
- Fosters collaboration

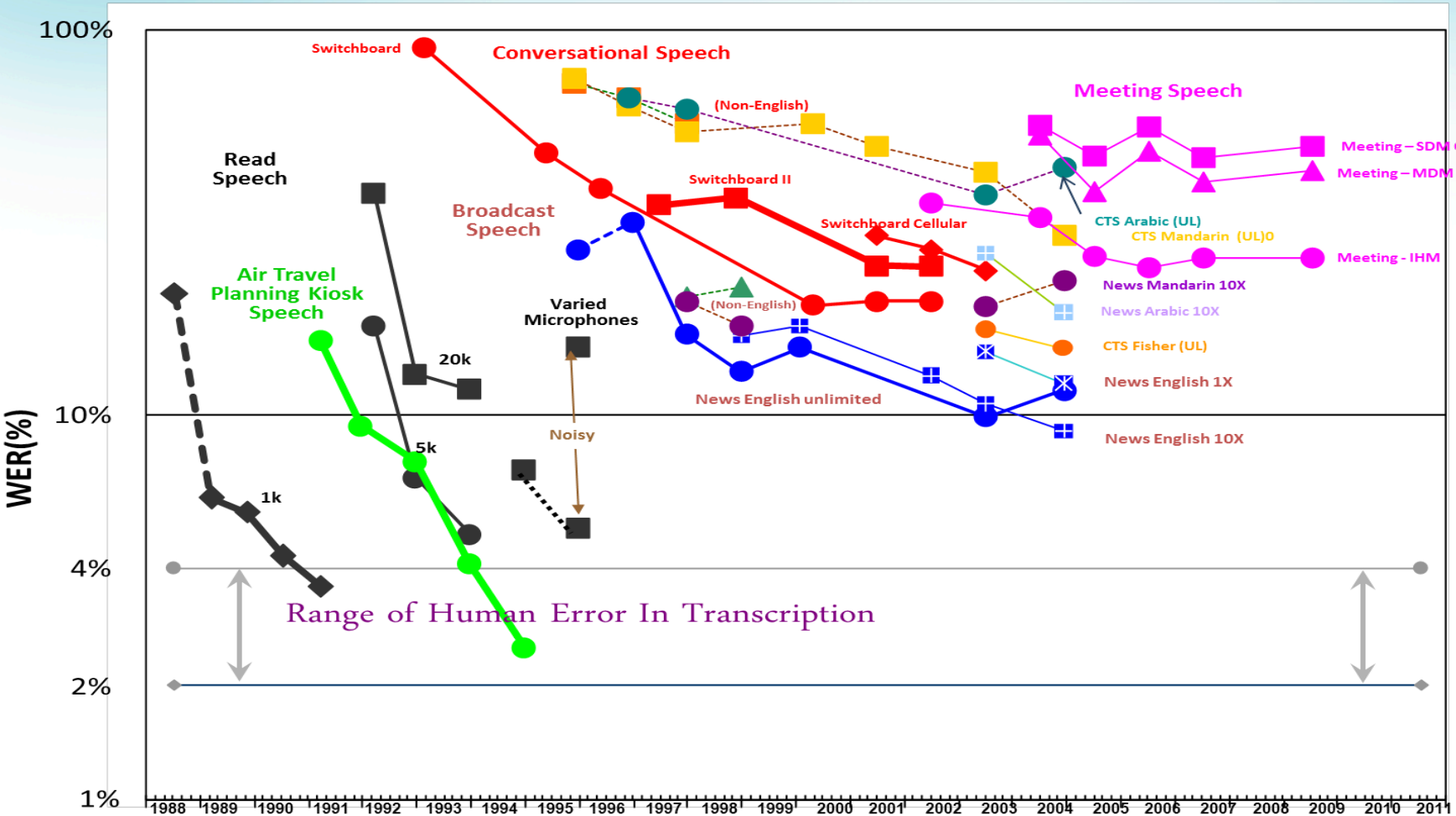
To be effective, evaluation must be

- Goal driven
- Systematic
- Rigorous

How Evaluation Drives Progress



Evaluation for ASR at NIST-IAD



Data Science Evaluation Goals

Apply measurement methods for data science systems

Measure the state-of-the-art and drive progress

Research measurement methods for data science

- General measurement/evaluation methods
- Effective use of "found" data
- Large datasets
- Workflows (component and end-to-end performance)
- Human involvement
- System benchmarking
- Mixed measurements (e.g., accuracy as a function of runtime)

Data Science Evaluation Hurdles

Goal	Hurdle
Found data	Data licensing / rights, privacy
Workflows	Structure of communities
Large datasets	Logistical, cost
System benchmarking	Is difficult, requires hardware
Human involvement	Requires labor & IRB, varied

Data Science Evaluation Plan

DARPA XDATA

- Identify Hurdles

Pre-pilot Evaluation

- Overcome Hurdles on Small Scale

Pilot Evaluation

- Overcome Hurdles on Large(r) Scale

Annual Evaluation Series with Multiple Tracks

- Join Measurement and Core Technology Research

Local Private Cloud

- Address benchmarking and technical challenges of running systems at NIST

Data Science Evaluation Schedule

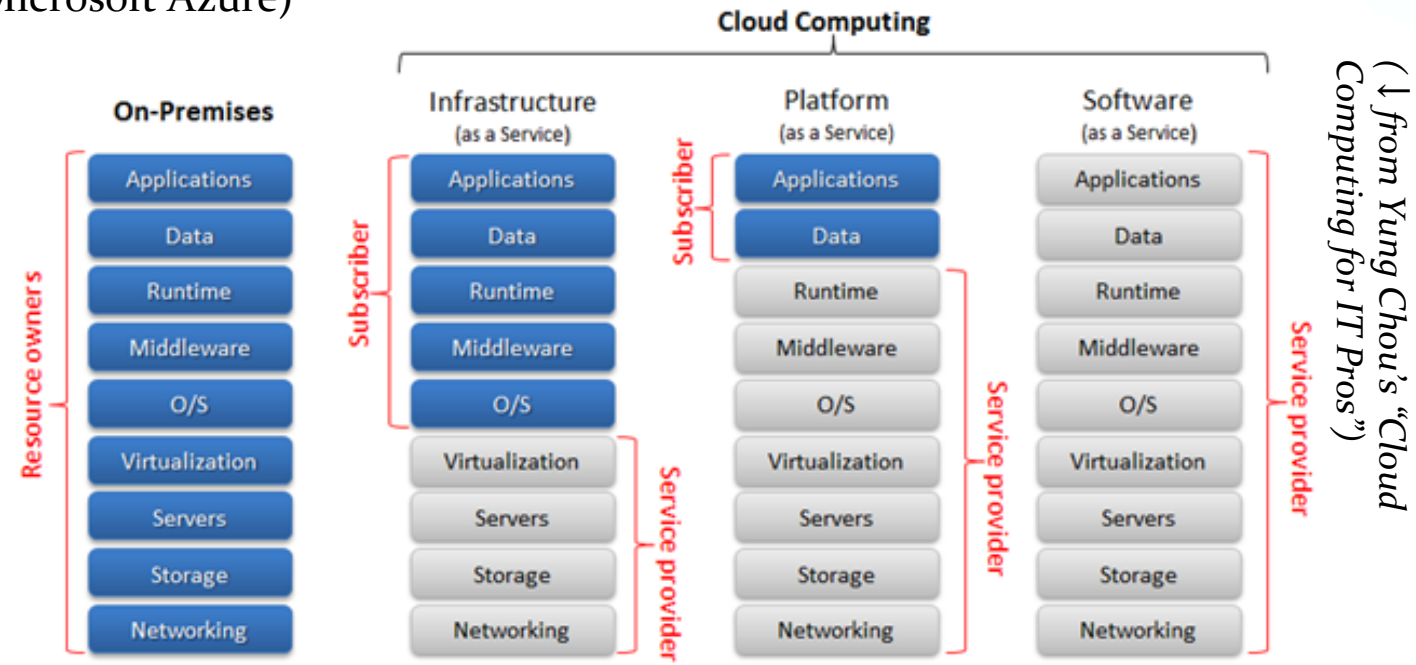
2014	2015	2016	2017
XDATA	**Pre-Pilot**	**Pilot**	**Full-Scale Evaluation**
		NIST Cloud	
	Single-Track	Single-Track	Multiple Tracks
Closed	Invitation-Only	Open to Everyone	Open to Everyone
Domain: Multiple	**Domain:** Traffic	**Domain:** Traffic	**Domain:** Multiple
✔	✔		

Evaluation Management System

Set up as a Private IaaS (aka Cloud infrastructure services): IT environment with an ability for a subscriber to provision infrastructure on demand (ex: Amazon Web Service, Google Compute Engine, Microsoft Azure)

<u>Hardware:</u>

- 496x CPU (Cores)
- 467.8TB HDD
- 2.1TB RAM
- 2x Tesla K80
- 4x Intel Phi 5100



CentOS Community ENTerprise Operating System

docker

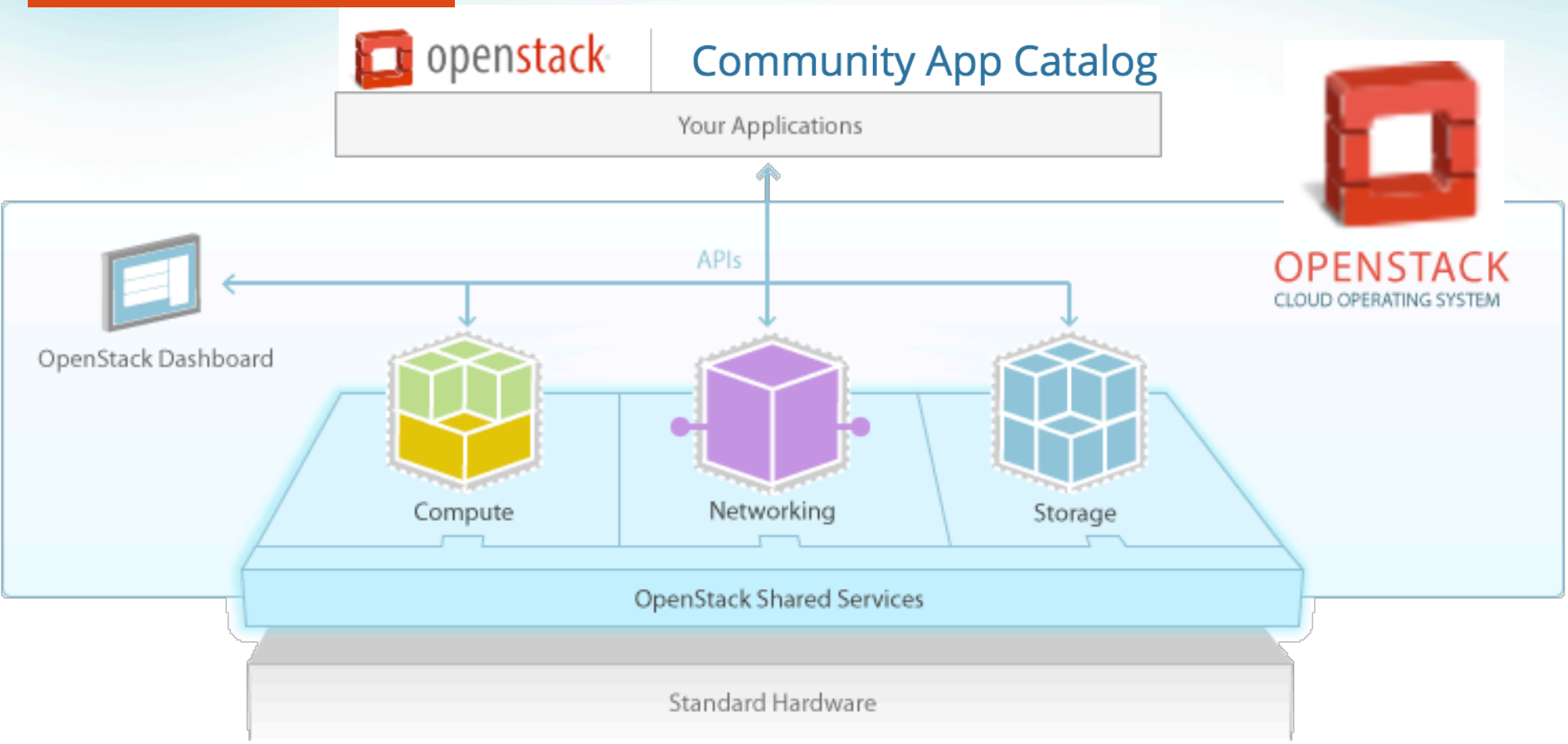
hadoop

ubuntu

kubernetes

Spark

and more ...



OPENSTACK
CLOUD OPERATING SYSTEM

TMP SLIDE: outline

- Importance of measurement / what's EDR (2 min)
- Indeed, look at successes: ASR, TREC, SRE (1 min)
- We plan to do similar for DS (10-15 min)
  - Specific goals: general measurement/eval methods, workflows, large datasets, system benchmarking, human involvement
  - Challenges: logistical (size of data, structure of communities), data rights, privacy, eval design/implementation (also size of data, domain knowledge, benchmarking requires hardware and is hard)
  - Plan/Schedule: feet wet in XDATA; EMS; pre-pilot, pilot, eval
- Transition to next talk (1 min)